Supplementary Information for

Ancient proteins resolve controversy over the identity of *Genyornis* eggshell

Beatrice Demarchi[1]*, Josefin Stiller[2], Alicia Grealy[3‡], Meaghan Mackie[4,5], Yuan Deng[2,6], Tom Gilbert[7], Julia Clarke[8], Lucas J. Legendre[8], Rosa Boano[1], Thomas Sicheritz-Pontén[4,7,13], John Magee[9], Guojie Zhang[2,6,10,11], Michael Bunce[3], Matthew James Collins[4,12], Gifford Miller[14].

[1] Department of Life Sciences and Systems Biology, Palaeoproteomics Laboratory, University of Turin, 10123 Turin, Italy
[2] Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
[3] School of Molecular and Life Science, Trace and Environmental DNA (TrEnD) laboratory, Curtin University, Bentley, Western Australia 6102, Australia
[4] Section for Evolutionary Genomics, The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[5] The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[6] China National GeneBank, BGI-Shenzhen, 518083 Shenzhen, China
[7] Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, 2100 Copenhagen, Denmark
[8] Department of Geological Sciences, The University of Texas at Austin, TX 78712, USA
[9] Research School Earth Sciences, The Australian National University, 2601 Acton, Australian Capital Territory, Australia
[10] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 65000 Kunming, China
[11] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, 65000 Kunming, China
[12] McDonald Institute for Archaeological Research, University of Cambridge, CB2 3ER Cambridge, United Kingdom
[13] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia
[14] INSTAAR and Department of Geological Sciences, University of Colorado Boulder, Boulder, CO 80309, USA

This PDF file includes:

Supplementary text

Figures S1 to S12

Tables S1 to S6

Legend for Dataset S1

SI References

Other supplementary materials for this manuscript include the following: Dataset S1

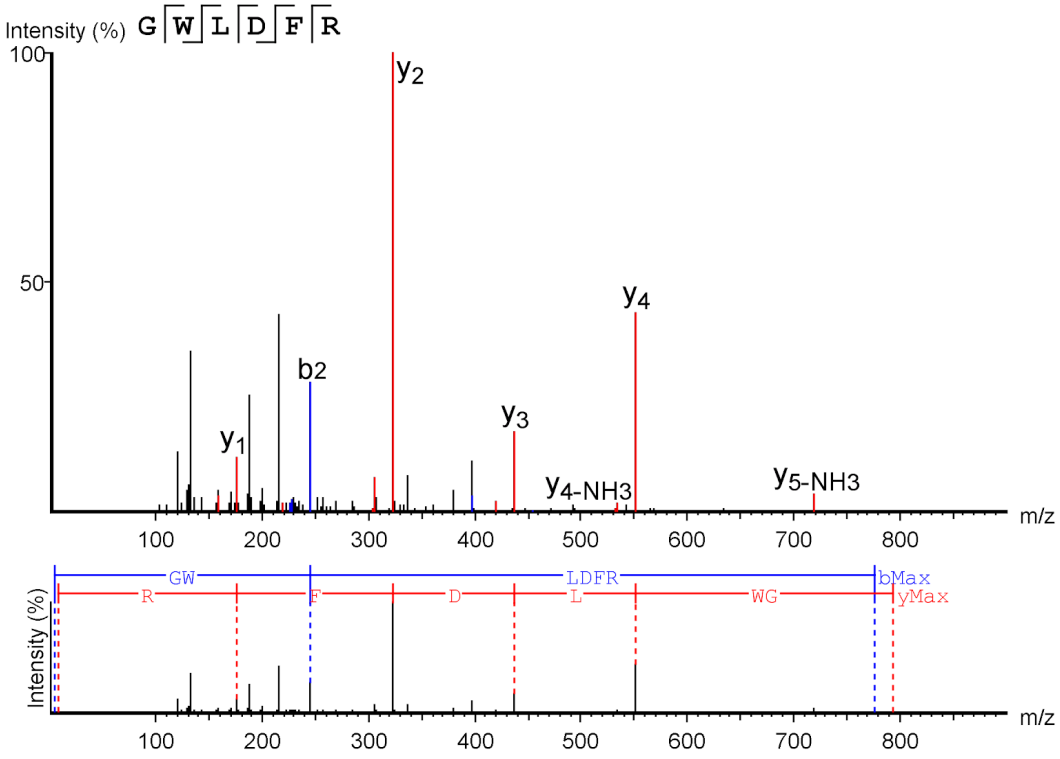
Contents

SI I Selection of annotated tandem mass spectra supporting the reconstruction of UO protein sequences

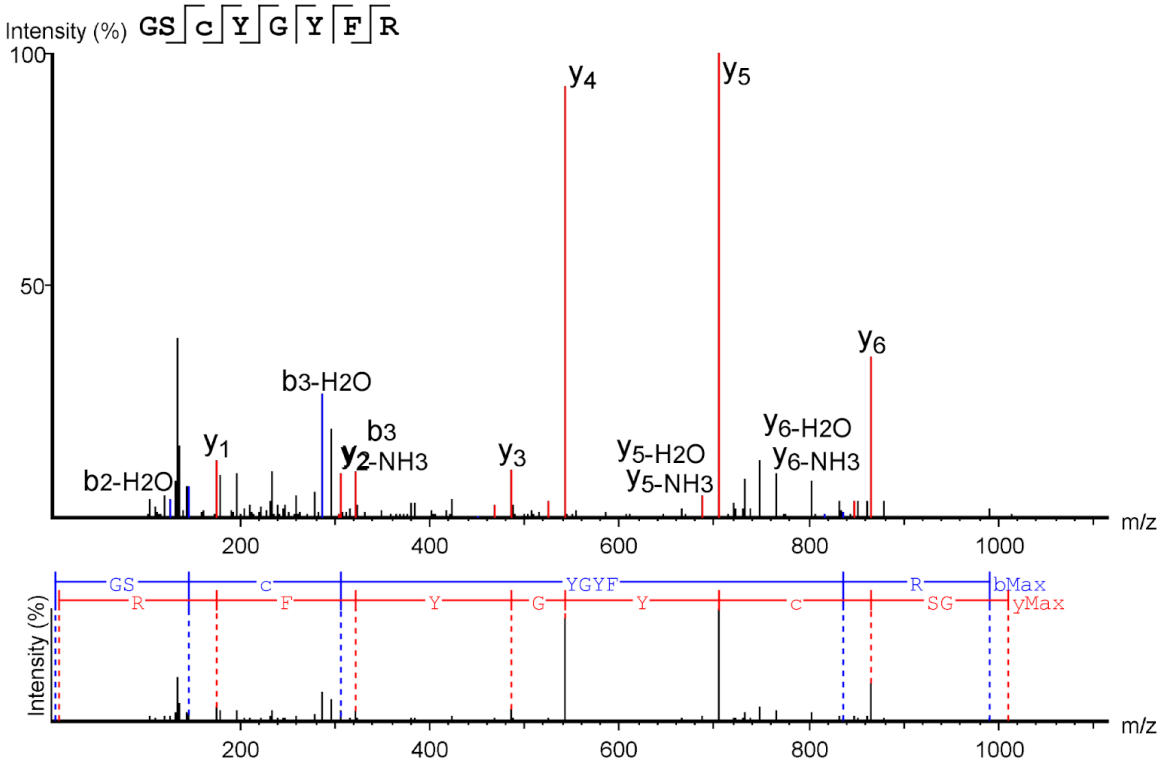
Annotated product ion spectra reported here are not unique, therefore we report each spectrum indicating one of the reference sequences they were detected from. Furthermore, the same portion of protein was generally supported by multiple peptides and peptide-spectrum matches - which are not reported in full here - full list of peptide sequences can be found in Dataset S1.

I.1 XCA-1

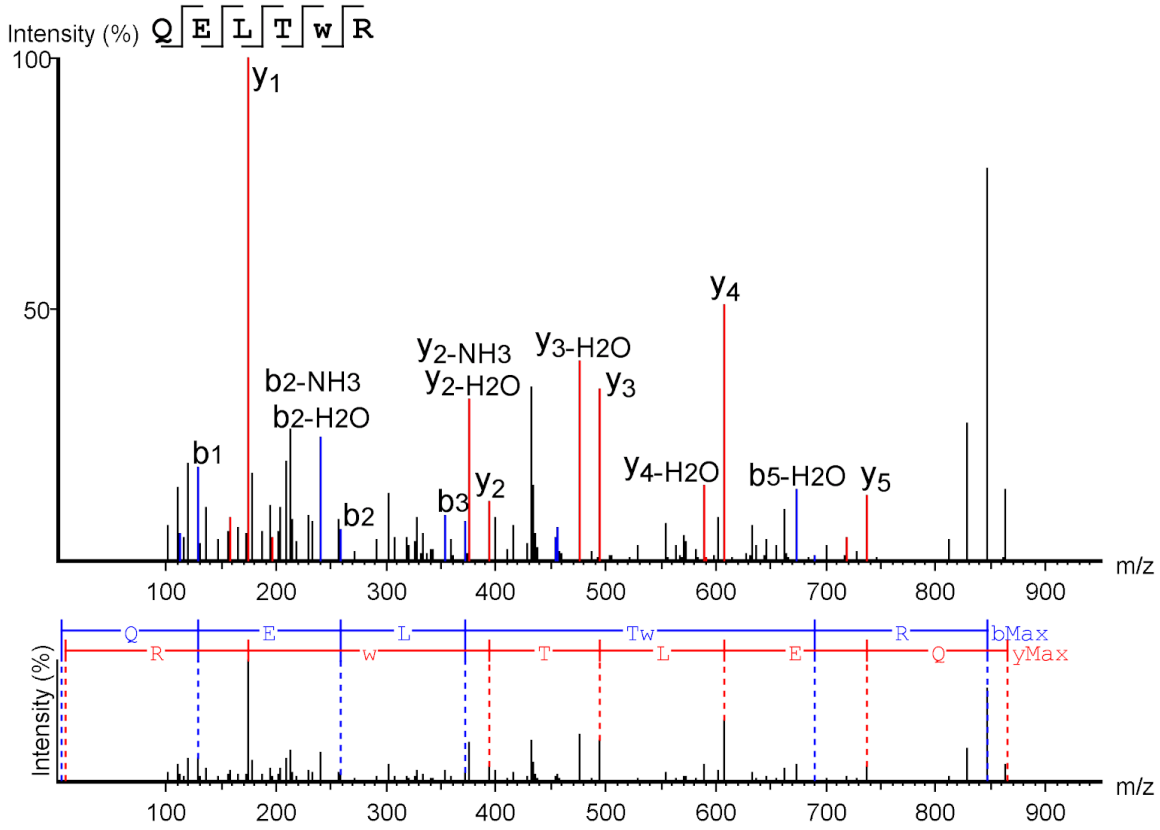
Source reference sequence = *Anseranas semipalmata*, *m/z* = 397.2051, score -lgP = 30.15, ppm = 4.7, PSM = 1



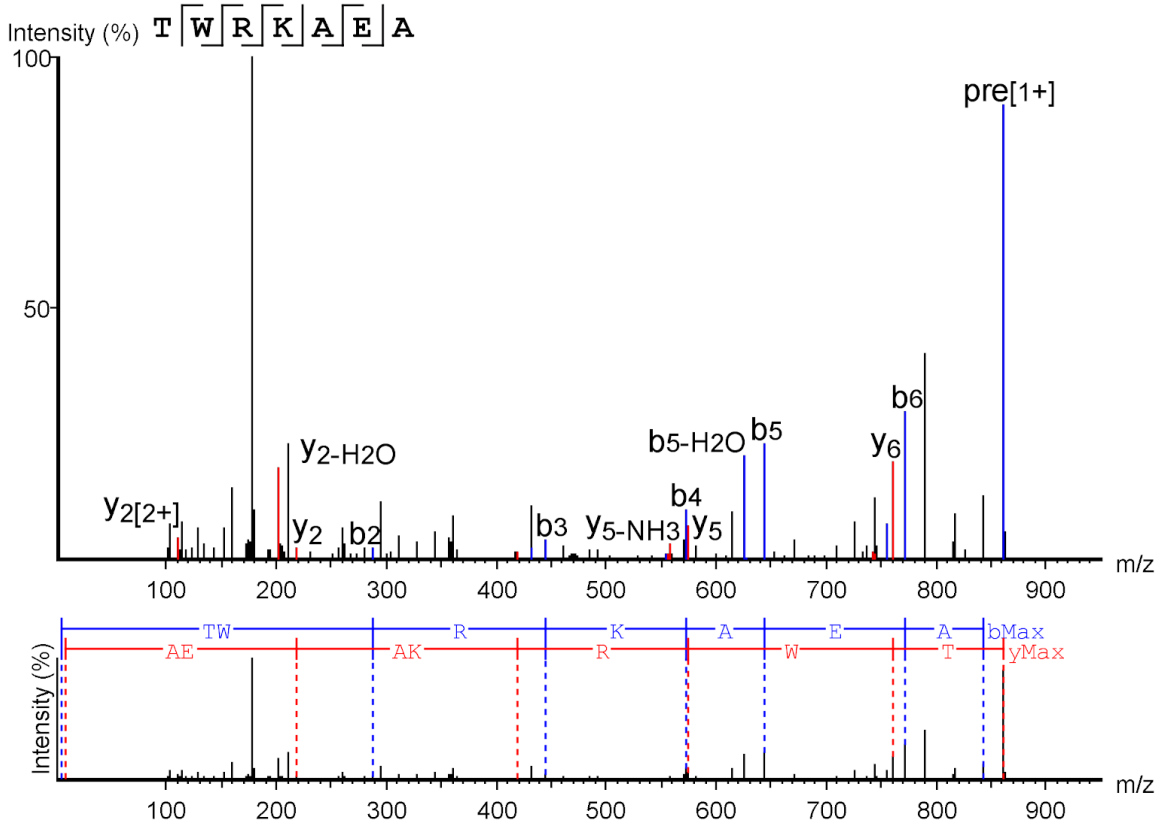
Source reference sequence = *Anseranas semipalmata*, *m/z* = 505.2150, score -lgP = 44.48, ppm = 3.1, PSM = 11



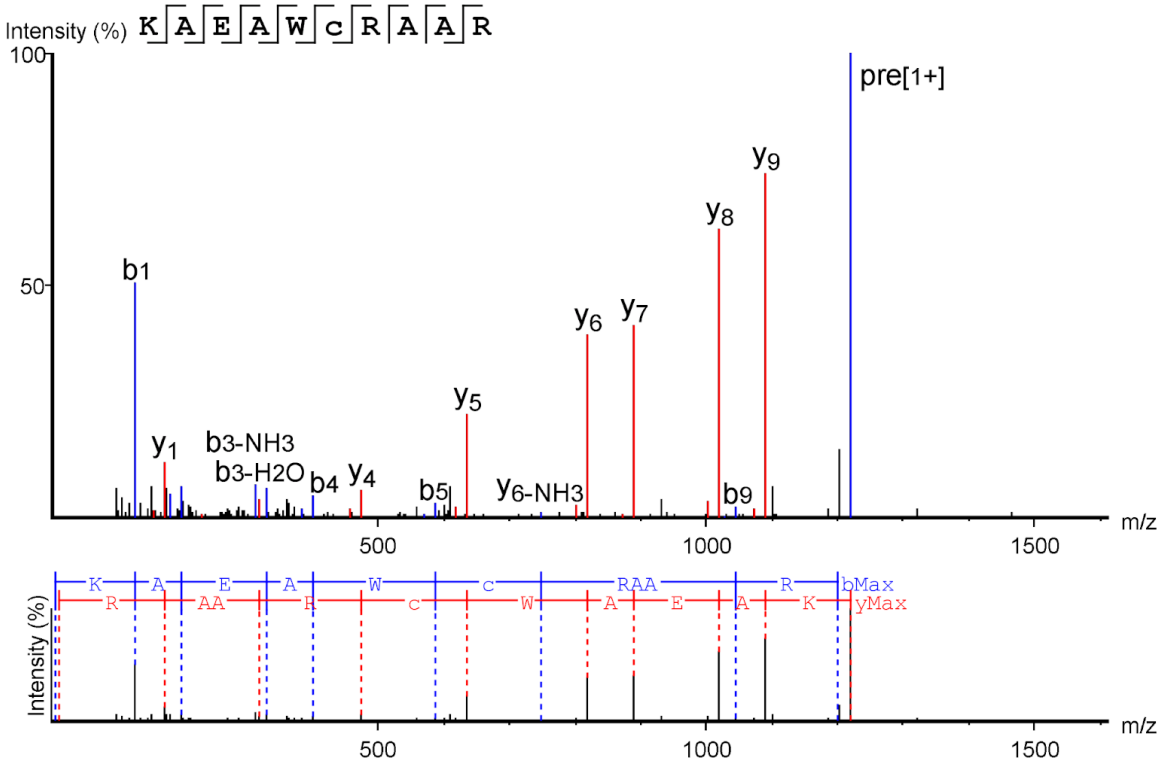
Source reference sequence = *Anseranas semipalmata*, *m/z* = 432.7148, score -lgP = 33.56, ppm = 1.6, PSM = 2



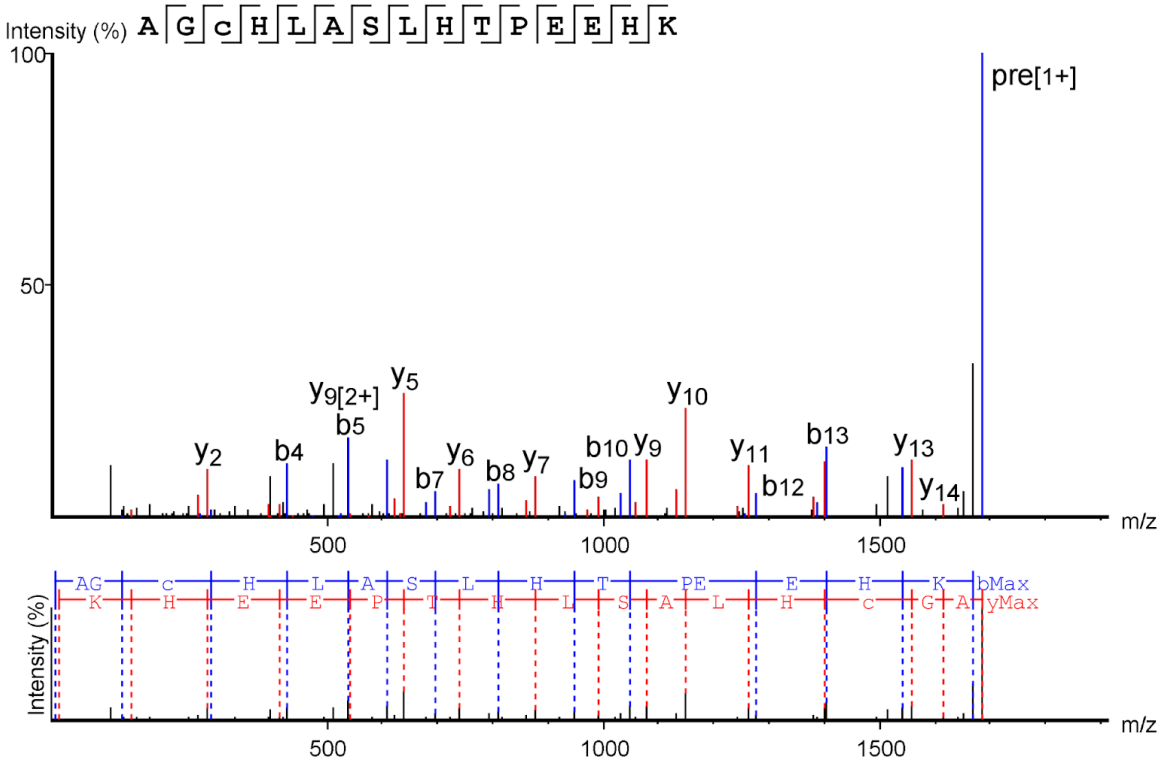
Source reference sequence = *Anseranas semipalmata*, *m/z* = 431.2334, score -lgP = 38.73, ppm = 2.0, PSM = 2



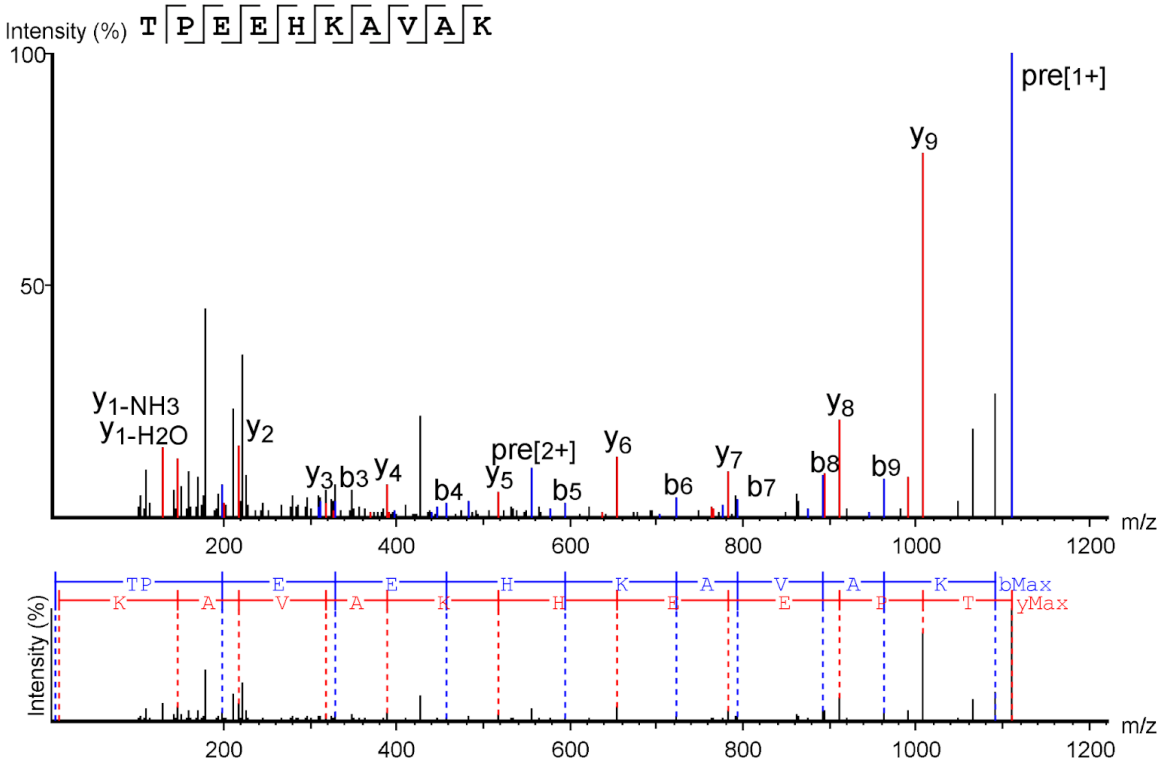
Source reference sequence = *Phalacrocorax pelagicus, m/z* = 609.8079, score -lgP = 42.82, ppm = -6.1, PSM = 7



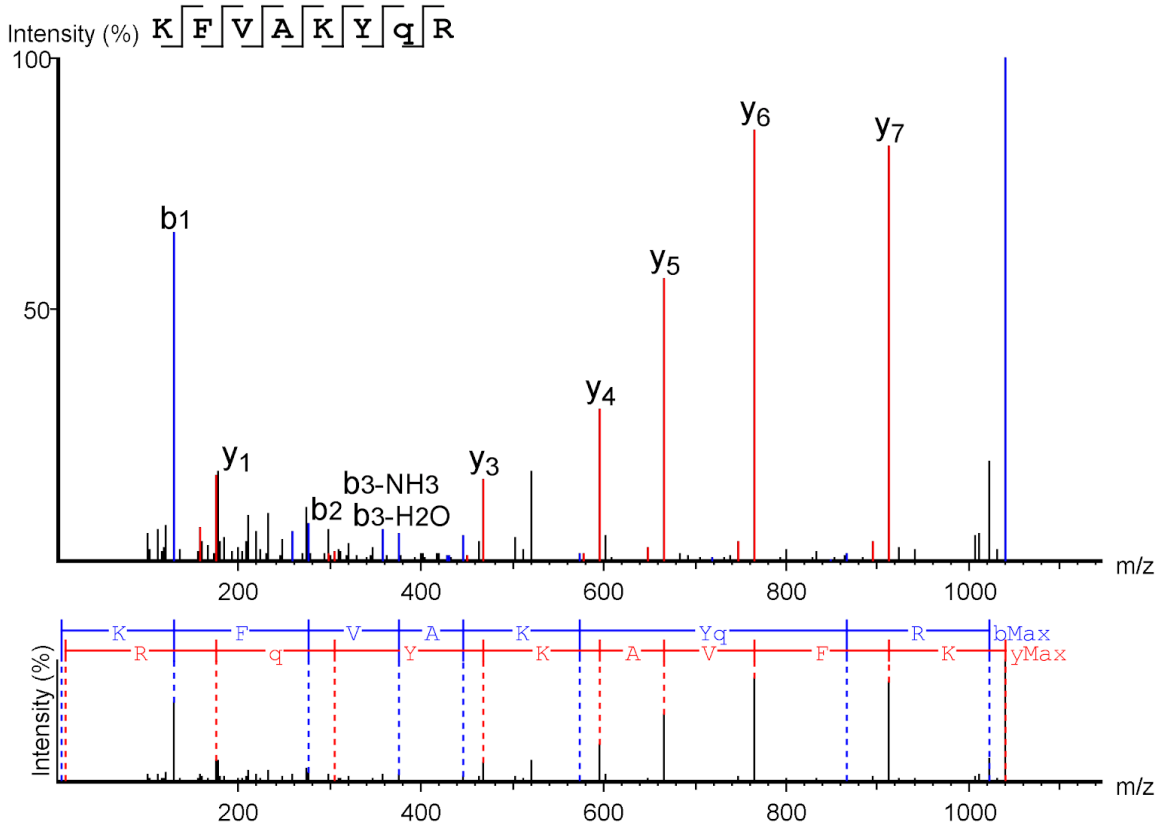
Source reference sequence = *Anseranas semipalmata*, *m/z* = 843.9058, score -lgP = 71.16, ppm = 1.6, PSM = 38



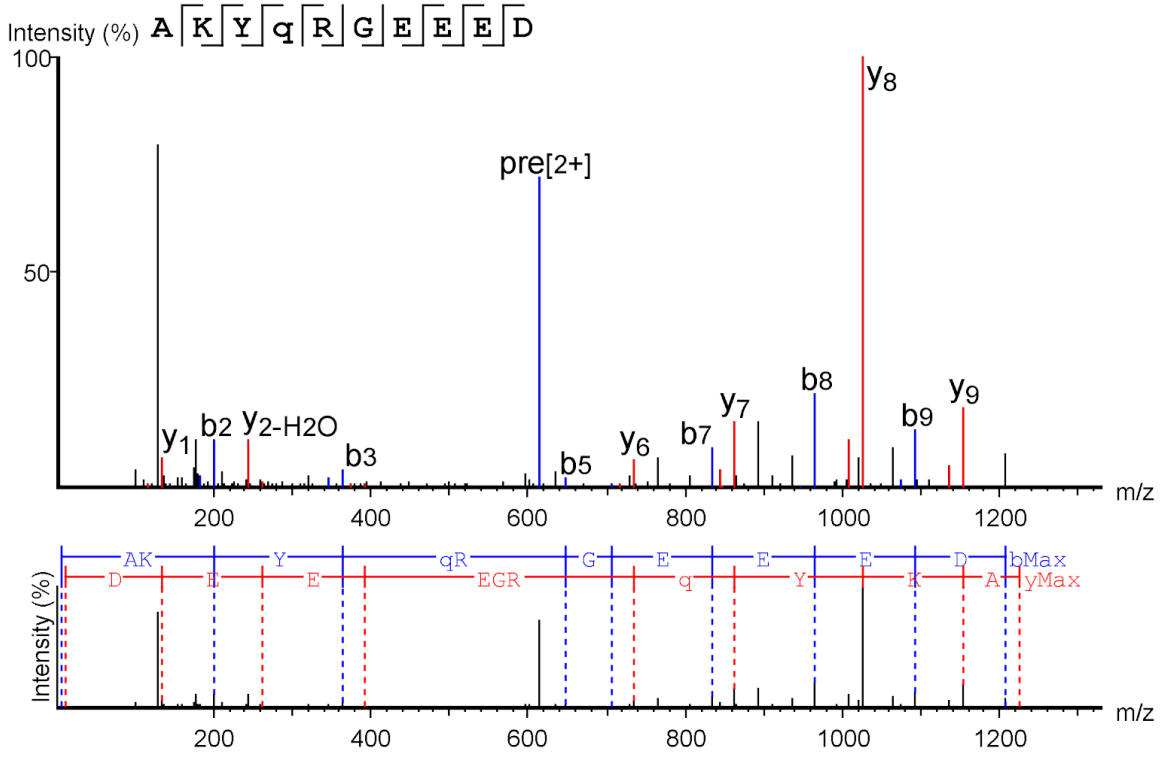
Source reference sequence = *Anseranas semipalmata*, *m/z* = 555.3030, score -lgP = 39.53, ppm = 3.5, PSM = 2



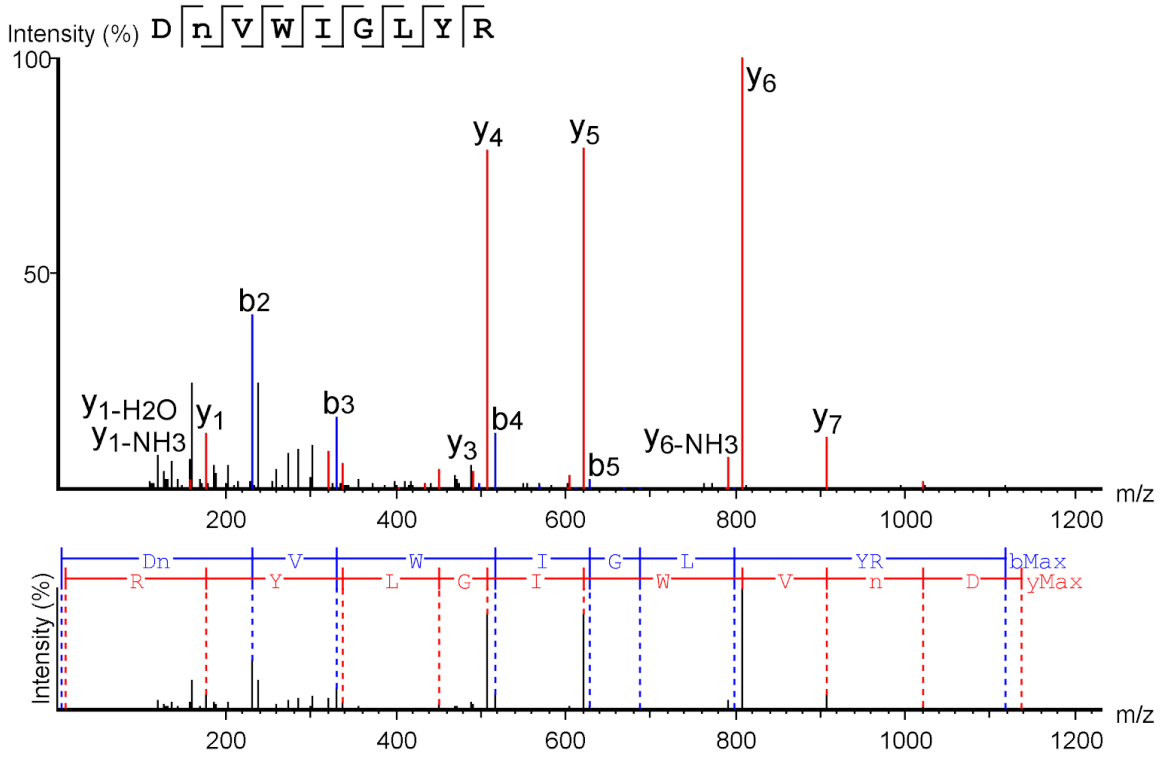
Source reference sequence = *Anseranas semipalmata*, *m/z* = 520.7990, score -lgP = 45.06, ppm = 1.9, PSM = 4



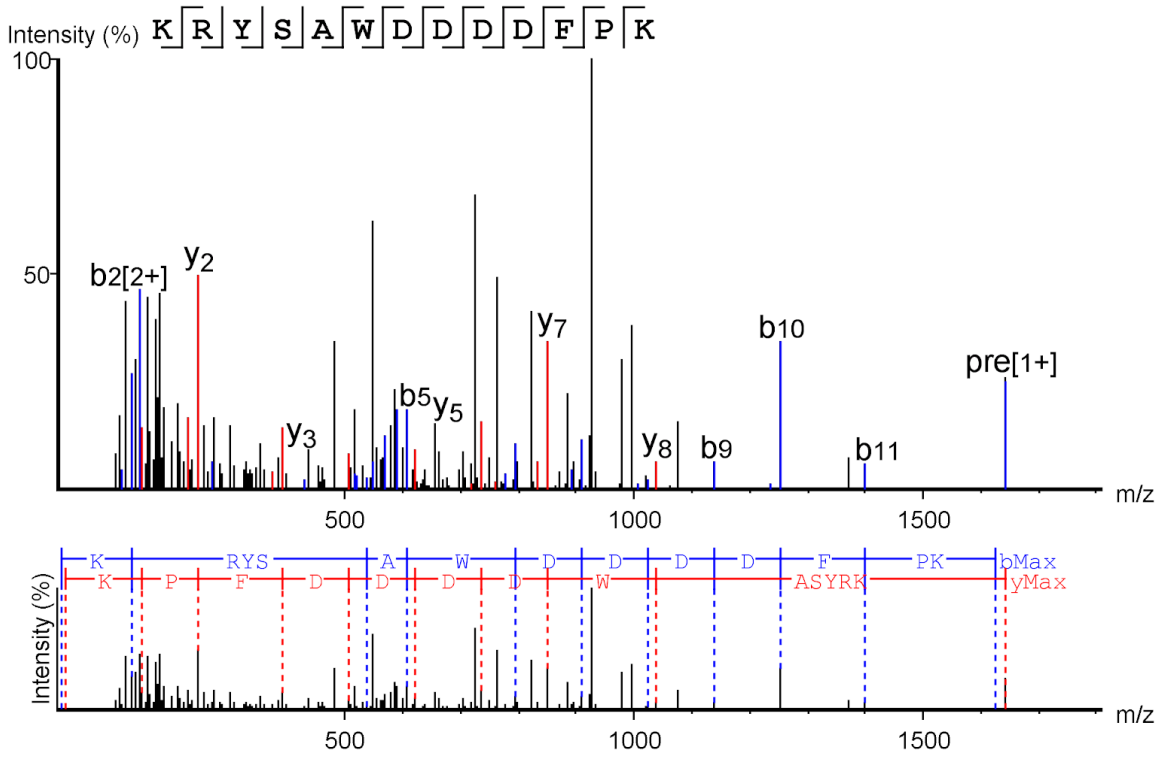
Source reference sequence = *Chauna torquata*, *m/z* = 613.2712, score -lgP = 45.85, ppm = 1.7, PSM = 3



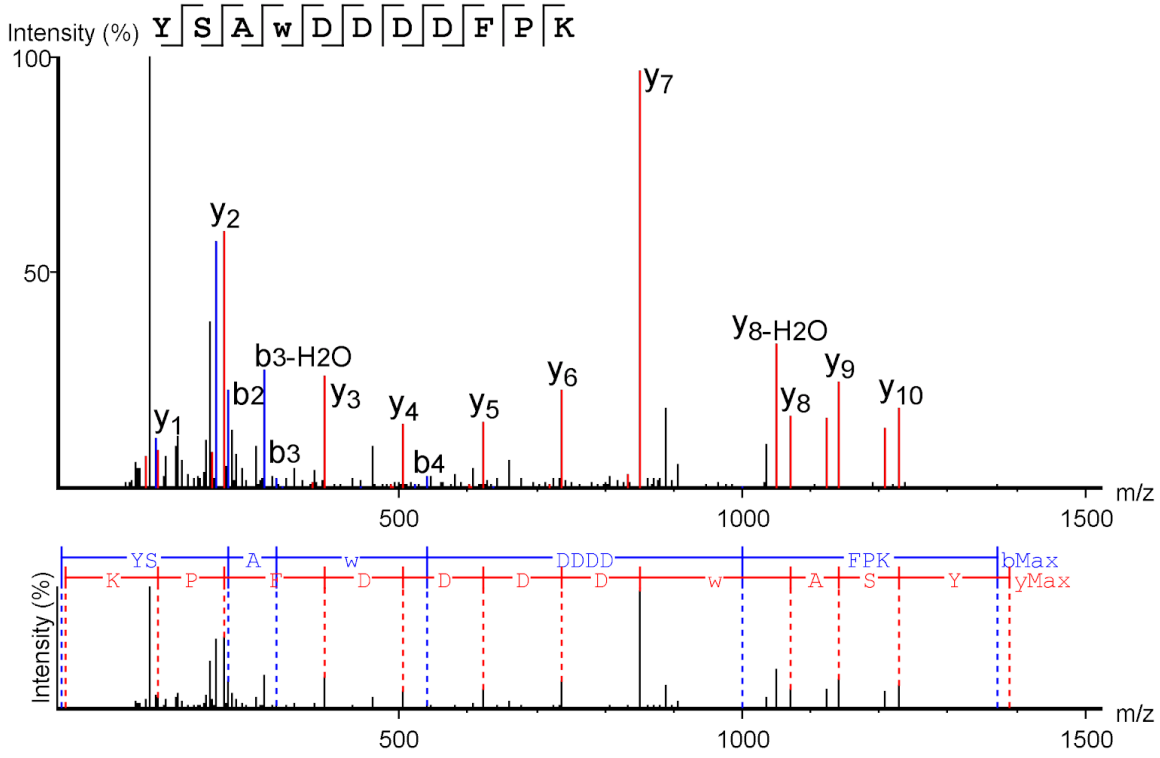
Source reference sequence = *Chauna torquata*, *m/z* = 568.7939, score -lgP = 54.57, ppm = 6.3, PSM = 24



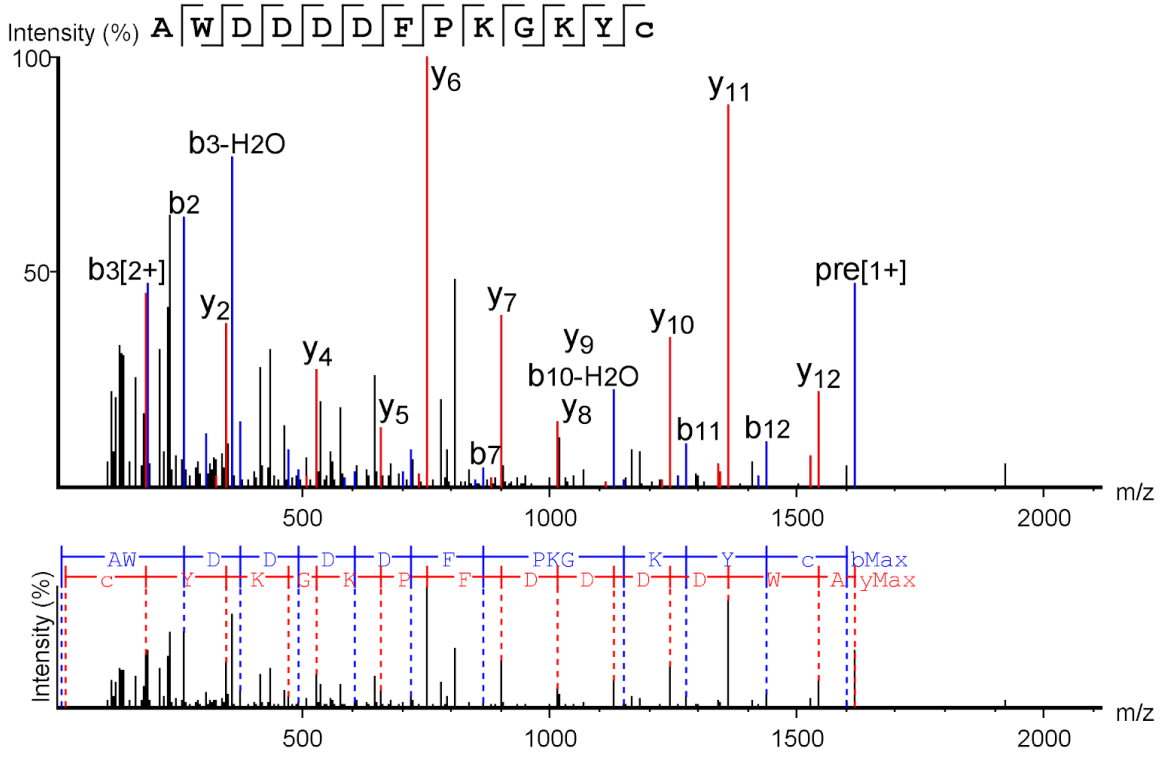
Source reference sequence = *Chauna torquata*, *m/z* = 548.2573, score -lgP = 47.47, ppm = 4.8, PSM = 1



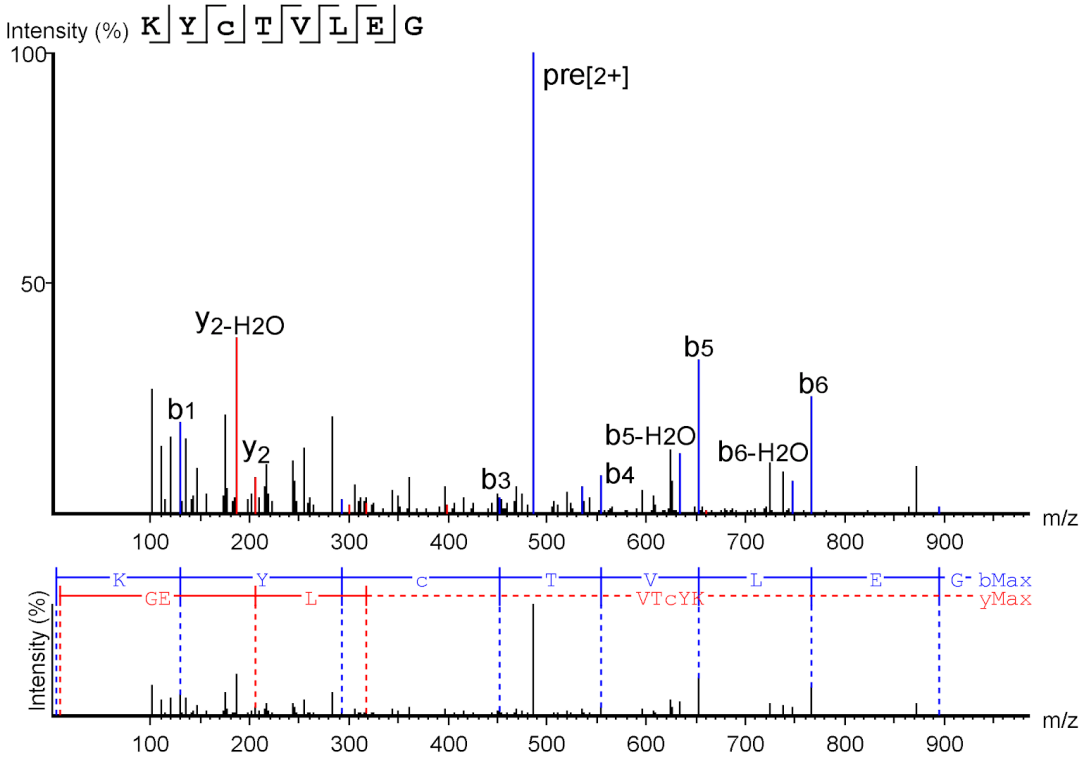
Source reference sequence = *Chauna torquata*, *m/z* = 695.7792, score -lgP = 50.46, ppm = 5.6, PSM = 13



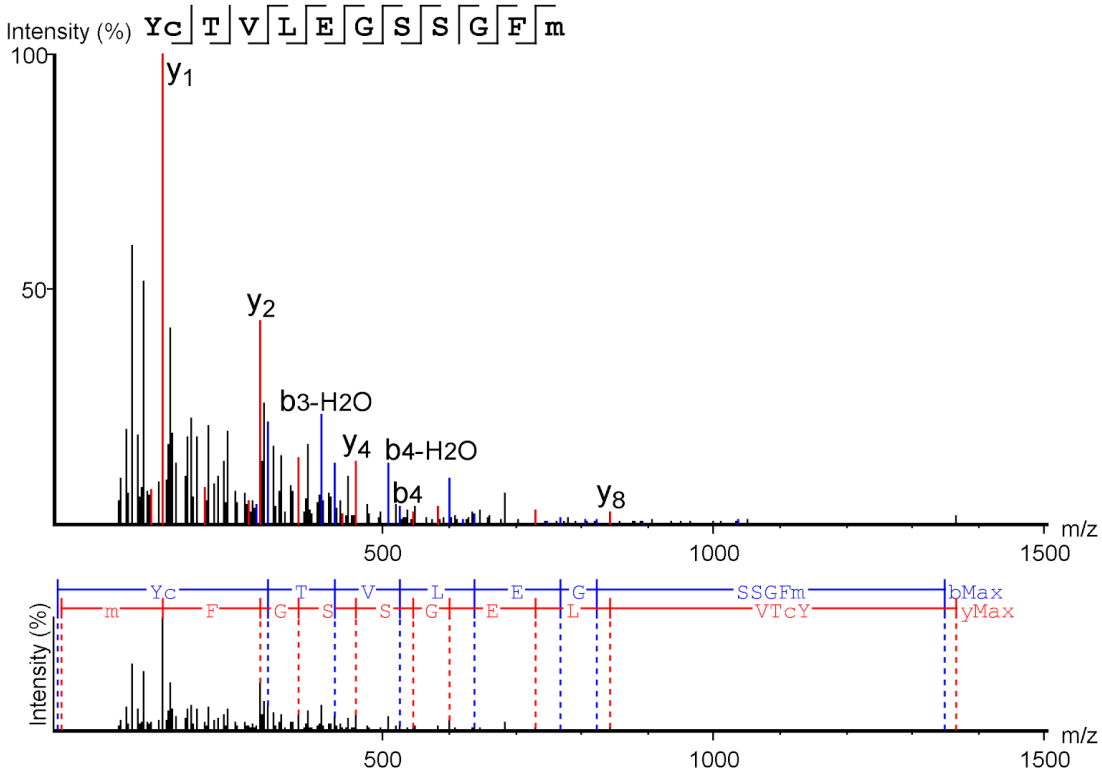
Source reference sequence = *Anseranas semipalmata*, *m/z* = 808.8416, score -lgP = 59.47, ppm = 4.6, PSM = 4



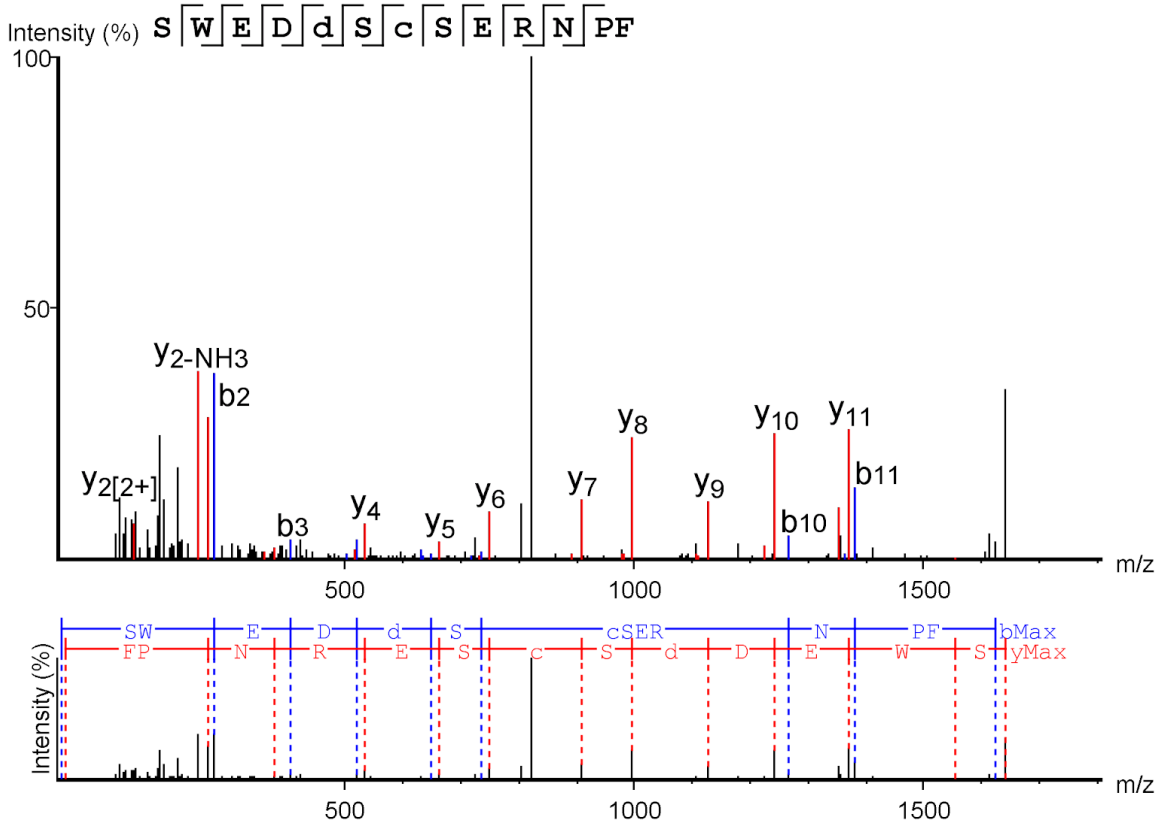
Source reference sequence = *Anseranas semipalmata*, *m/z* = 485.2398, score -lgP = 36.17, ppm = 1.5, PSM = 6



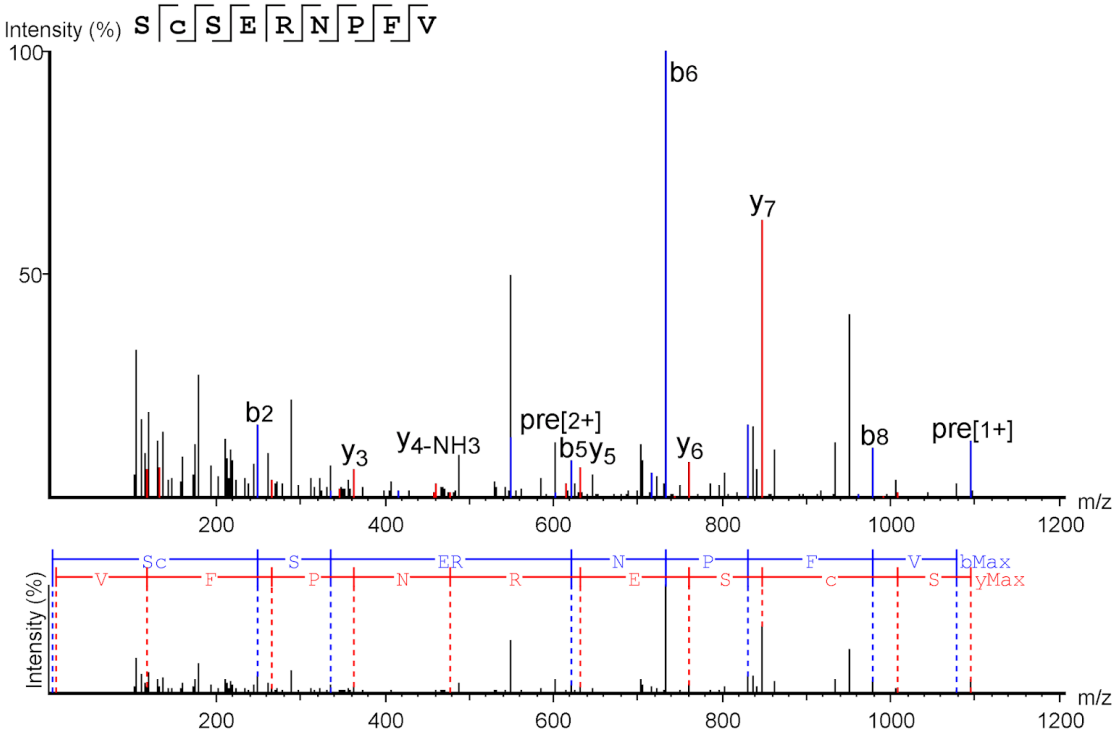
Source reference sequence = *Anseranas semipalmata*, *m/z* = 683.7900, score -lgP = 46.09, ppm = 5.3, PSM = 3



Source reference sequence = *Anseranas semipalmata*, *m/z* = 821.8307, score -lgP = 51.16, ppm = 6.3, PSM = 3

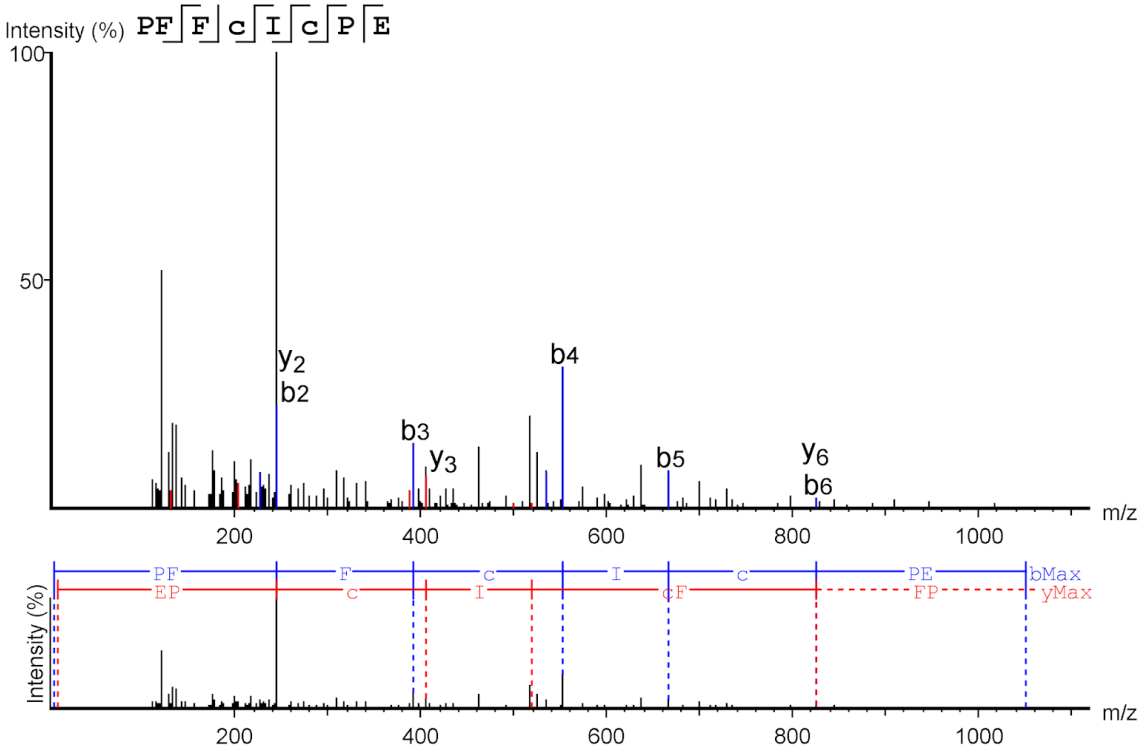


Source reference sequence = *Anseranas semipalmata*, *m/z* = 548.2497, *de novo* ALC score 87%, ppm = 3.1, PSM = 3

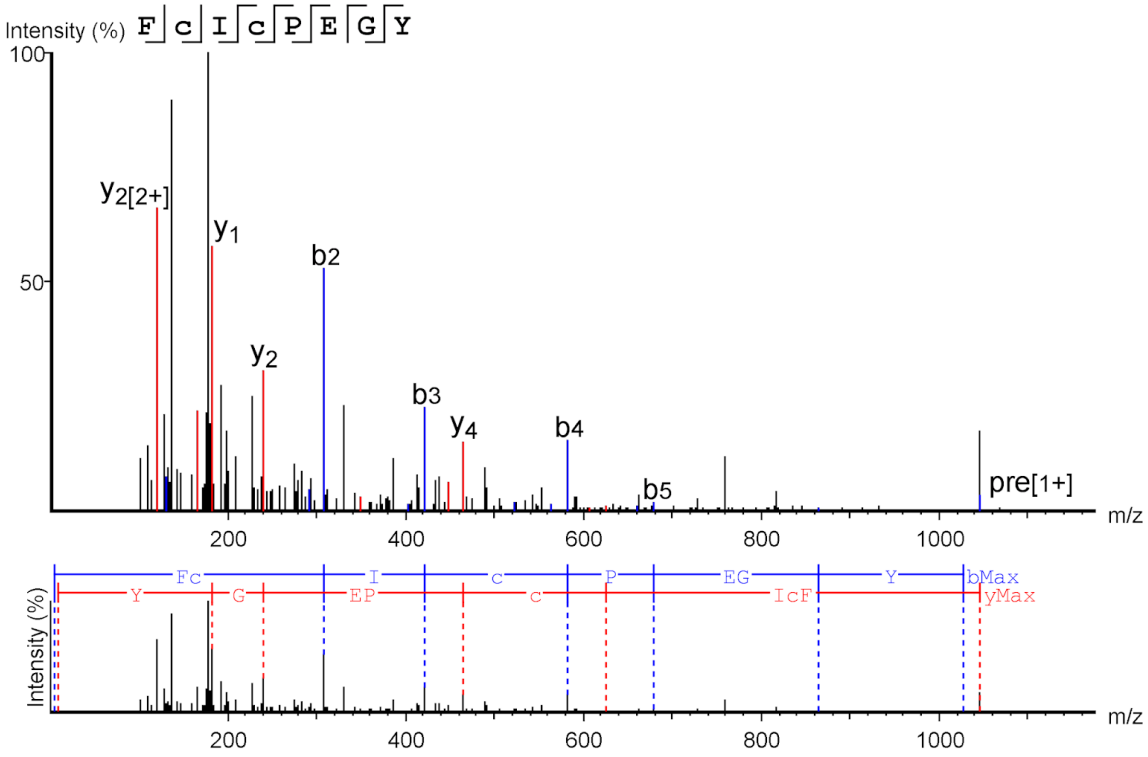


I.2 Lactadherin

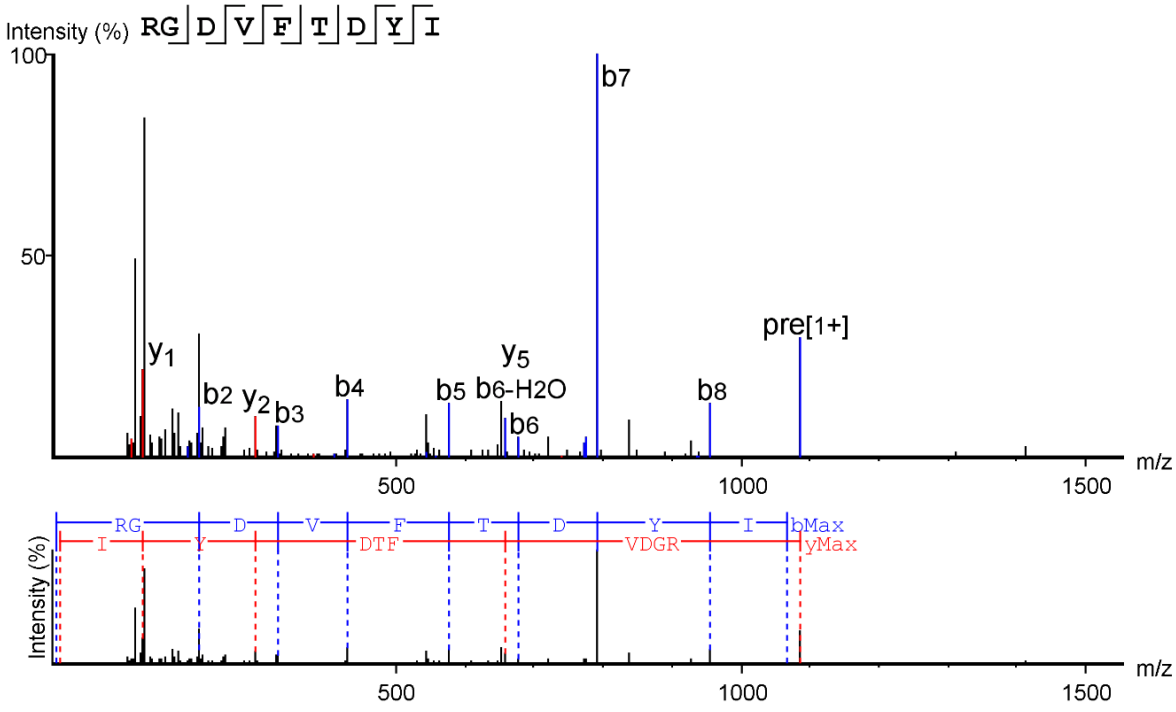
Source reference sequence = *Anseranas semipalmata*, *m/z* = 535.2316, score -lgP = 34.68, ppm = 7.4, PSM = 2



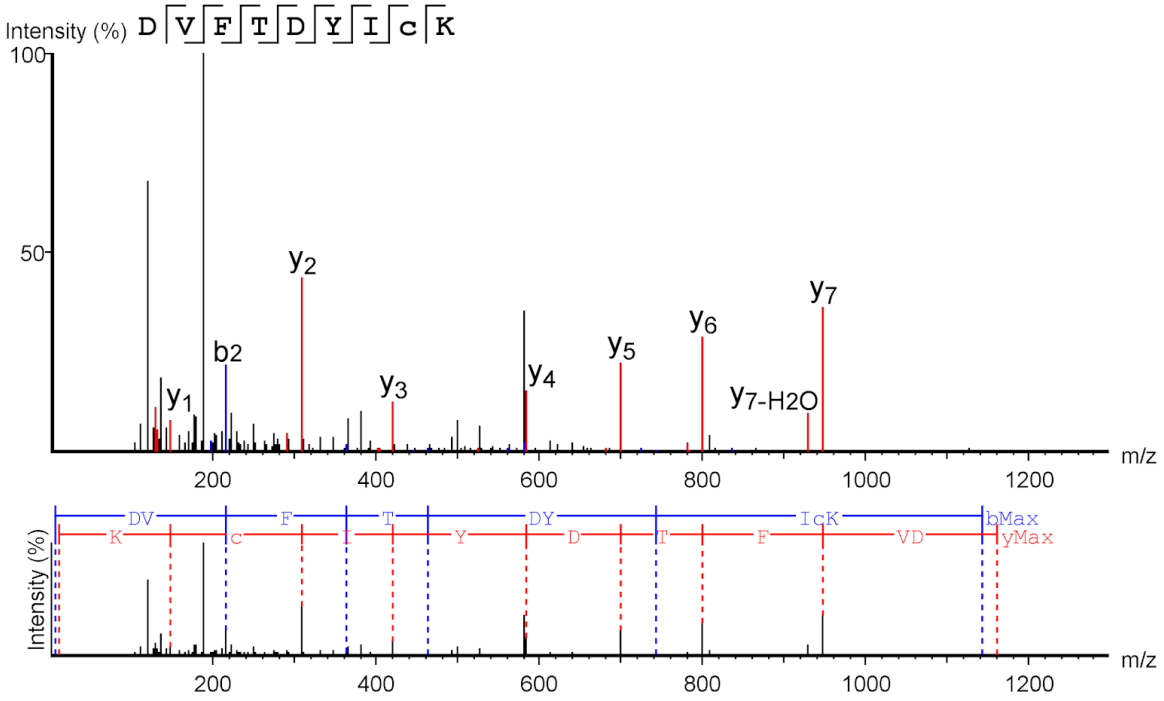
Source reference sequence = *Anseranas semipalmata*, *m/z* = 523.2126, score -lgP = 43.91, ppm = 6.1, PSM = 3



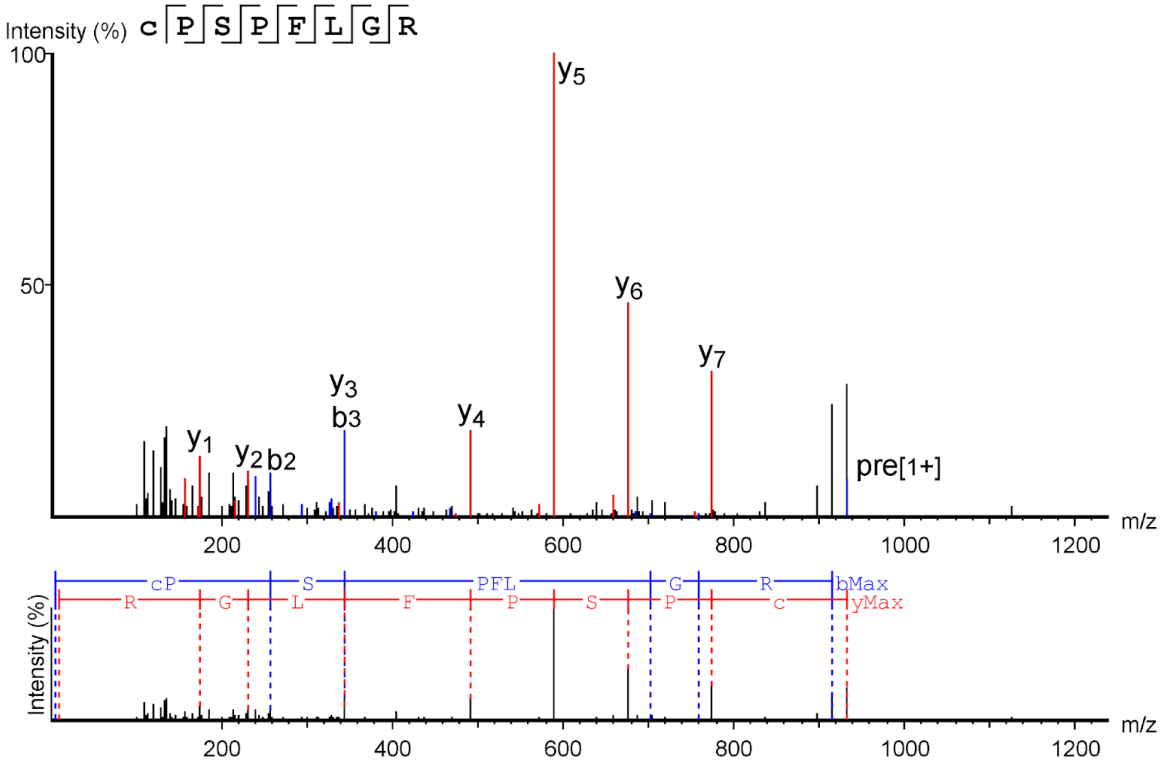
Source reference sequence = *Penelope pileata*, *m/z* = 543.2703, score -lgP = 45.45, ppm = 6.6, PSM = 2



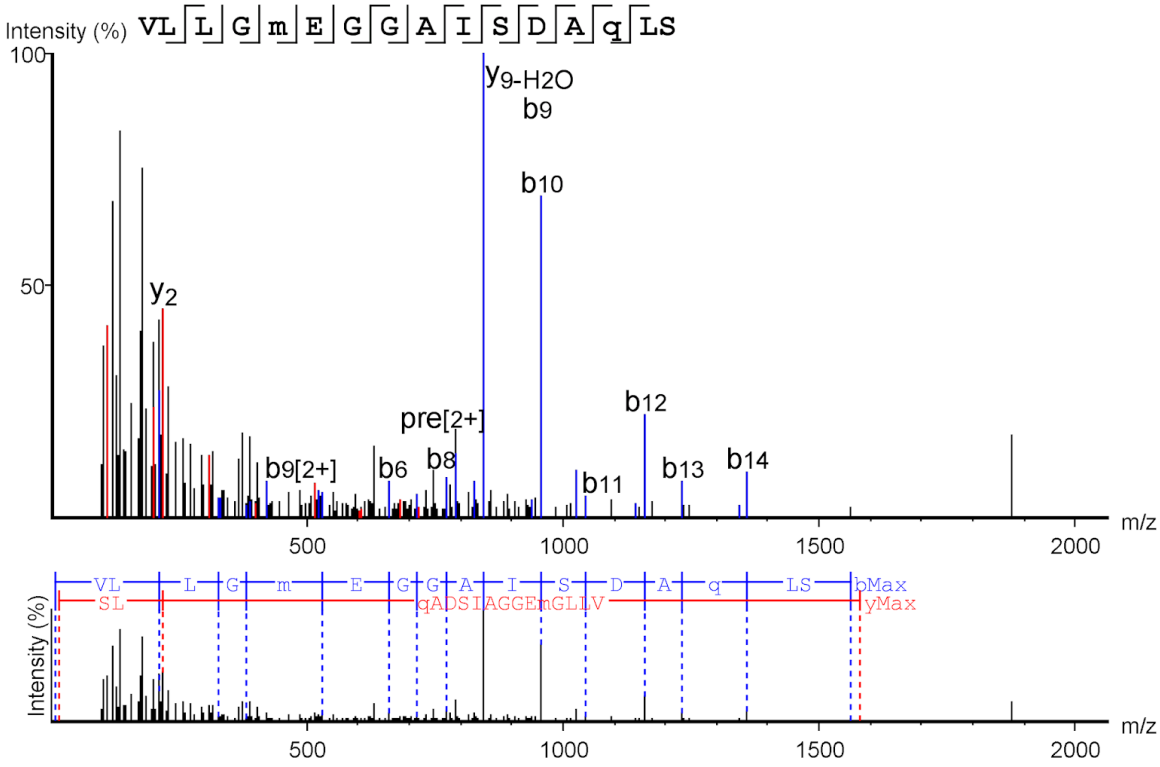
Source reference sequence = *Penelope pileata*, *m/z* = 580.7718, score -lgP = 44.89, ppm = 6.1, PSM = 7



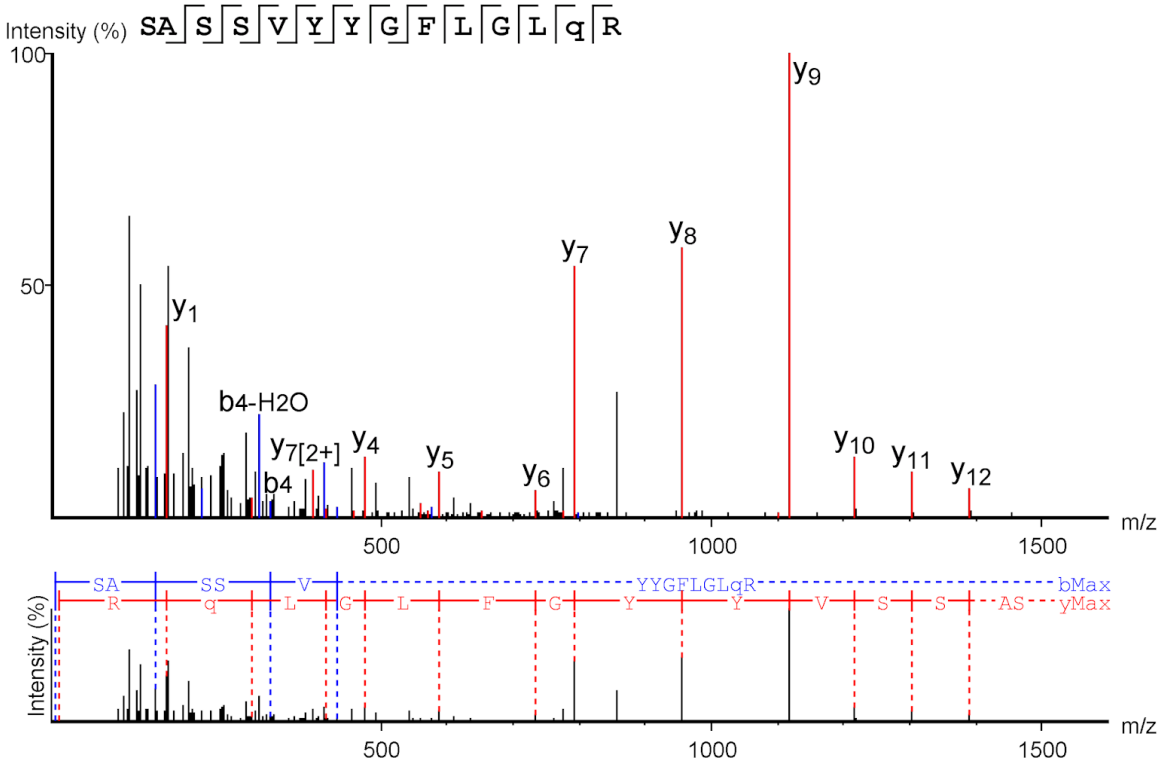
Source reference sequence = *Anseranas semipalmata*, *m/z* = 467.2347, score -lgP = 45.74, ppm = 1.2, PSM = 4



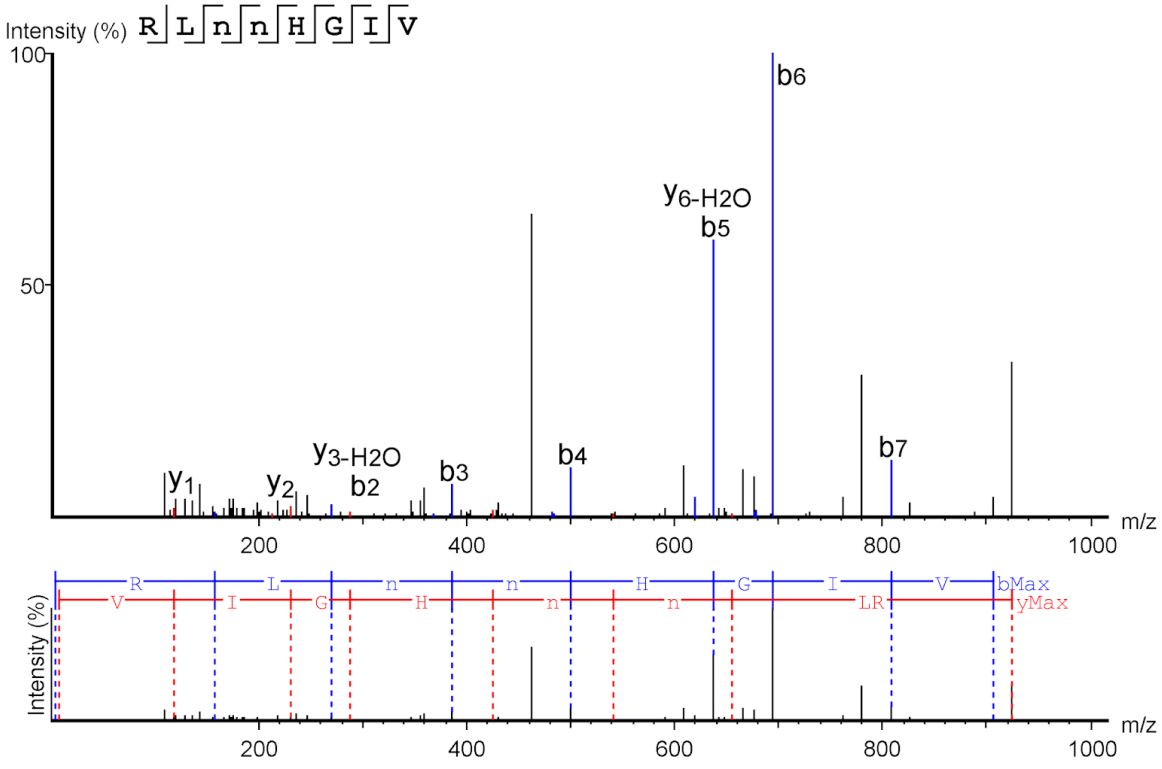
Source reference sequence = *Columba livia*, *m/z* = 789.3946, score -lgP = 40.33, ppm = 5.9, PSM = 1



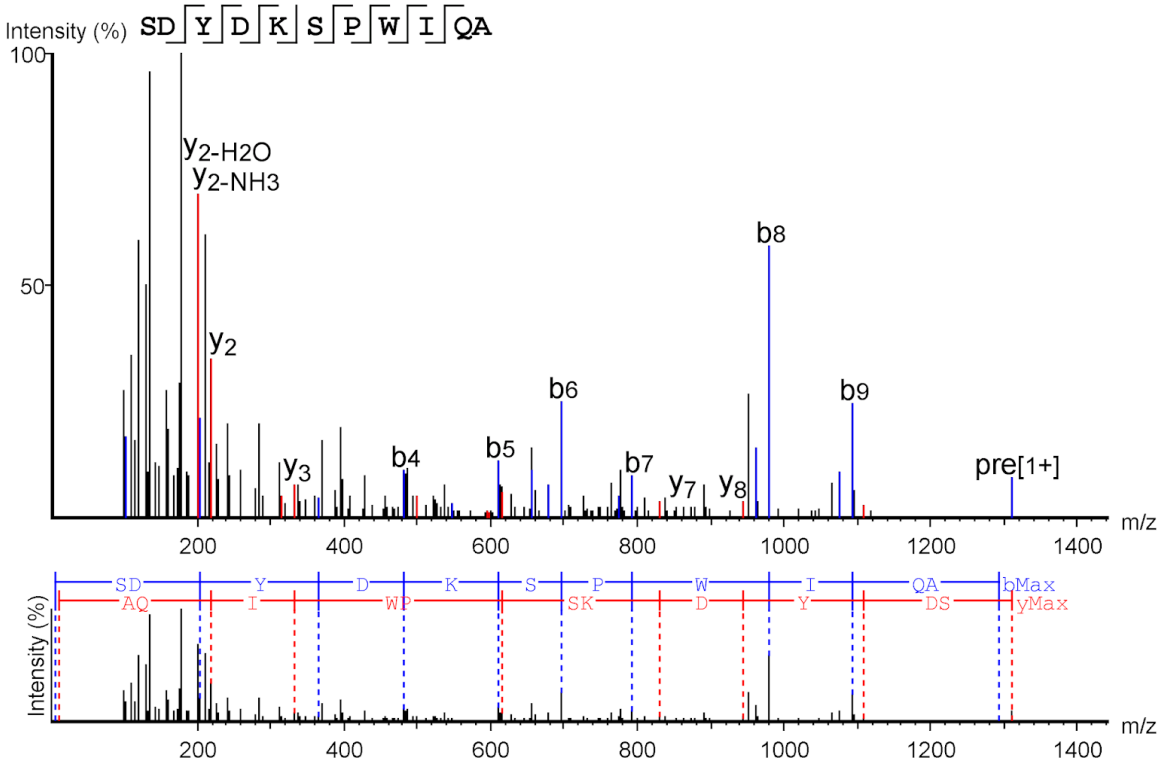
Source reference sequence = *Columba livia*, *m/z* = 774.8927, score -lgP = 44.34, ppm = 5.7, PSM = 2



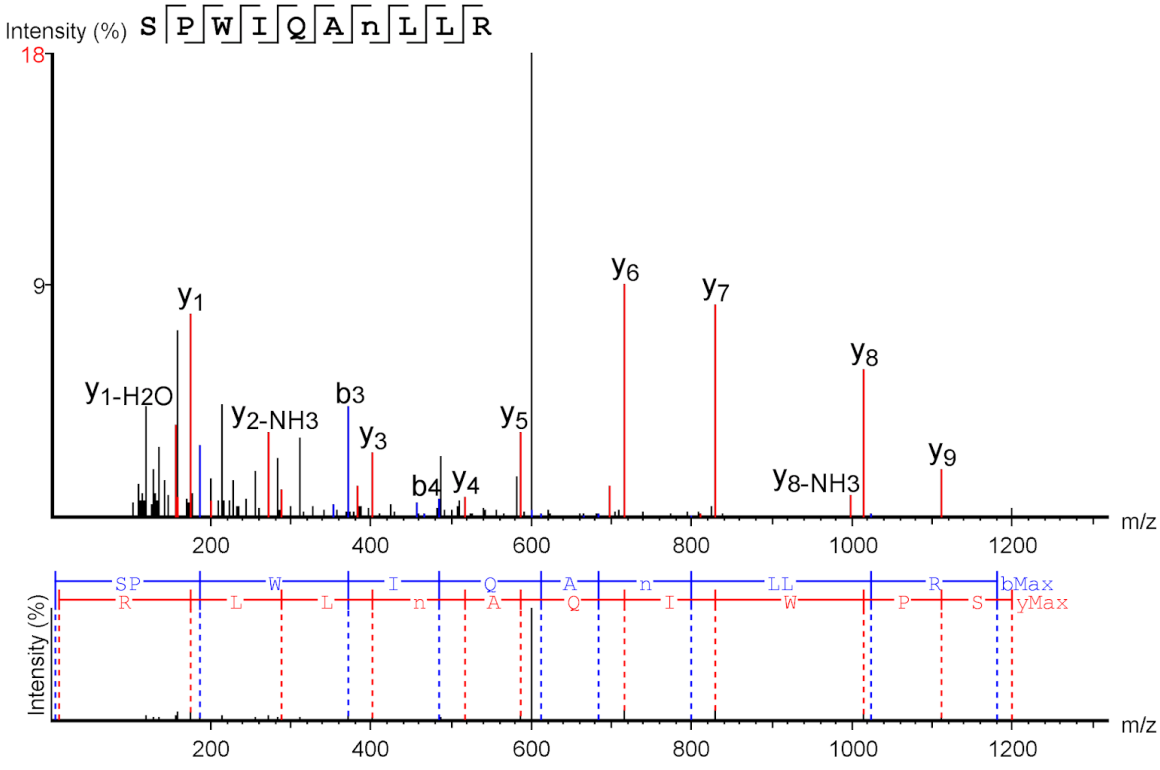
Source reference sequence = *Columba livia*, *m/z* = 462.7504, score -lgP = 43.71, ppm = 4.1, PSM = 7



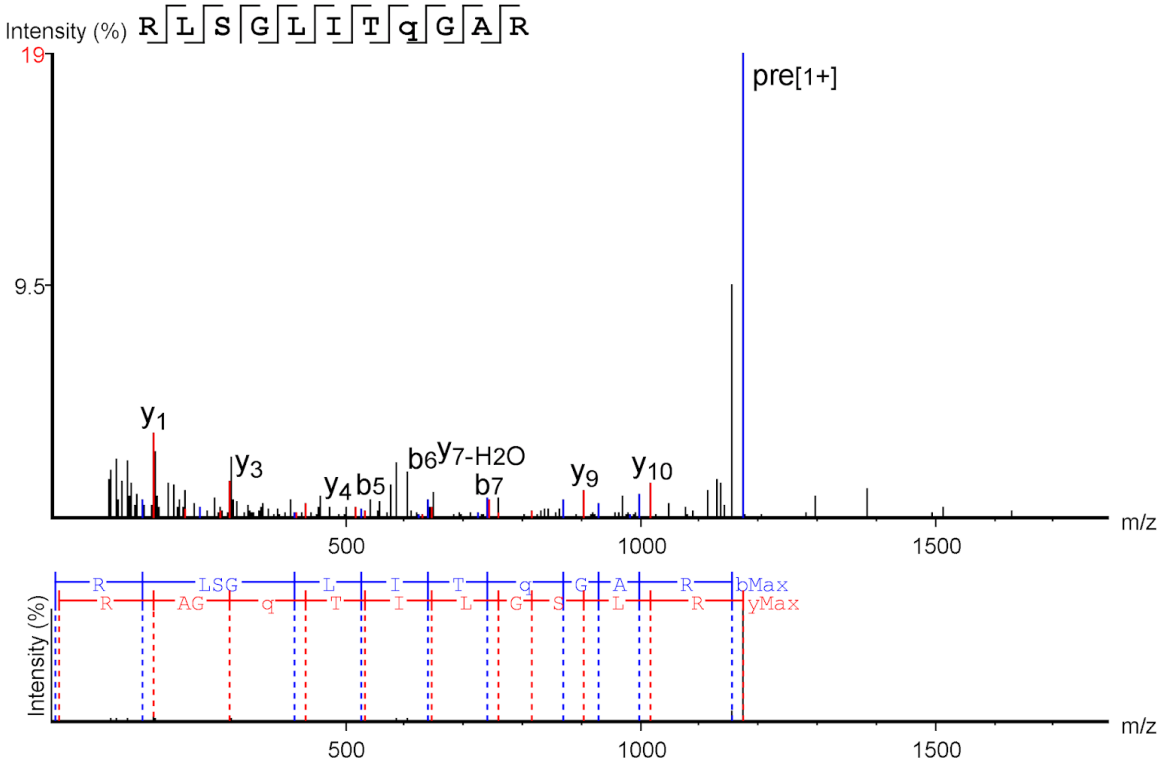
Source reference sequence = *Columba livia*, *m/z* = 655.3097, score -lgP = 44.27, ppm = 4.7, PSM = 1



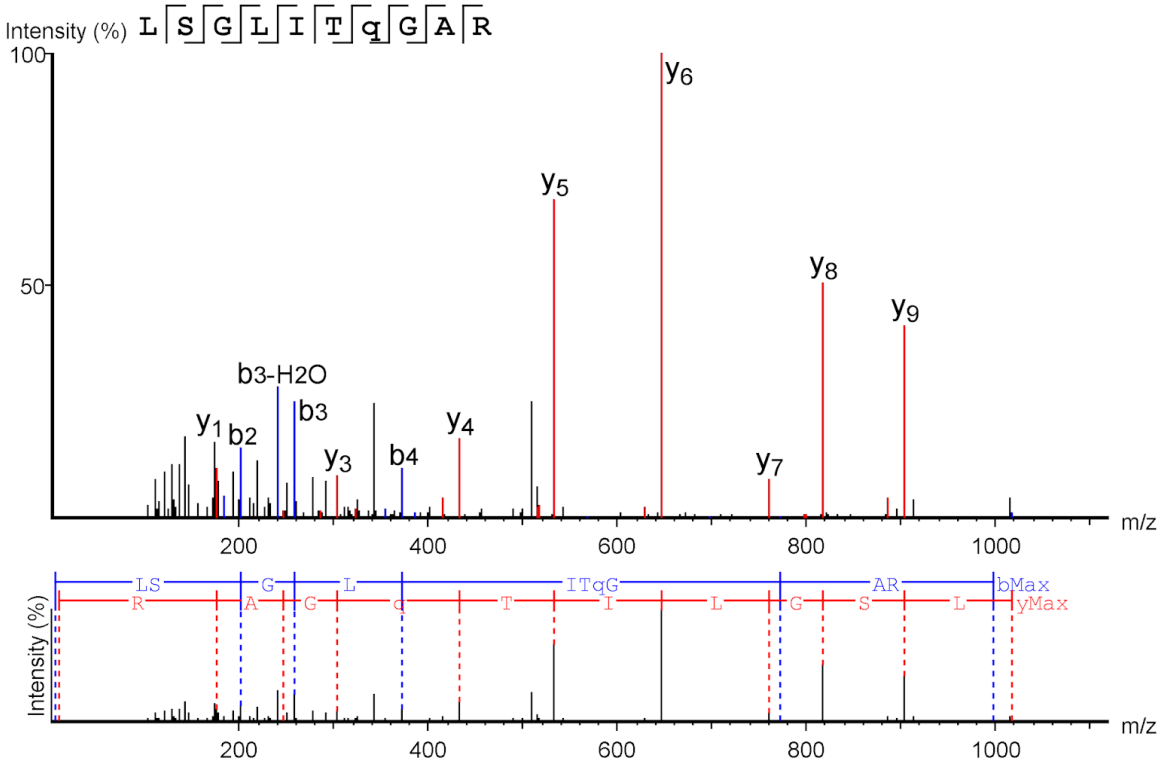
Source reference sequence = *Columba livia*, *m/z* = 599.8362, score -lgP = 45.29, ppm = 6.1, PSM = 3



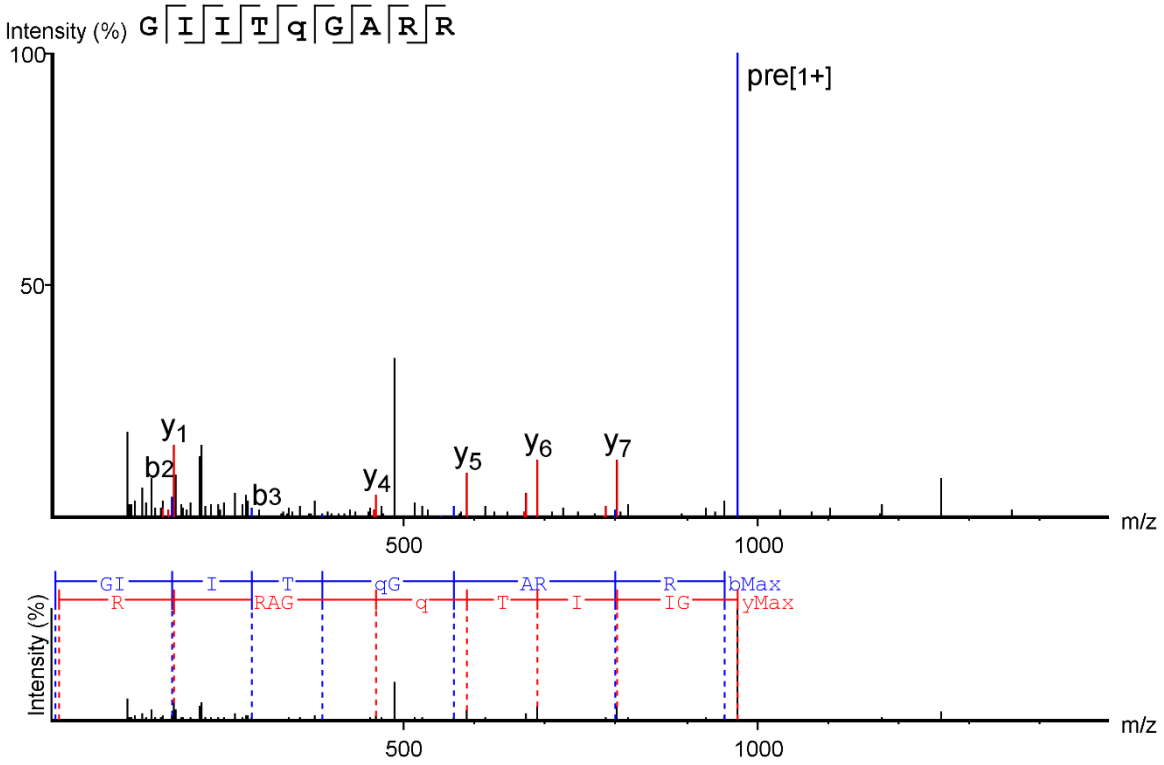
Source reference sequence = *Columba livia*, *m/z* = 586.8418, score -lgP = 37.11, ppm = 1.5, PSM = 1



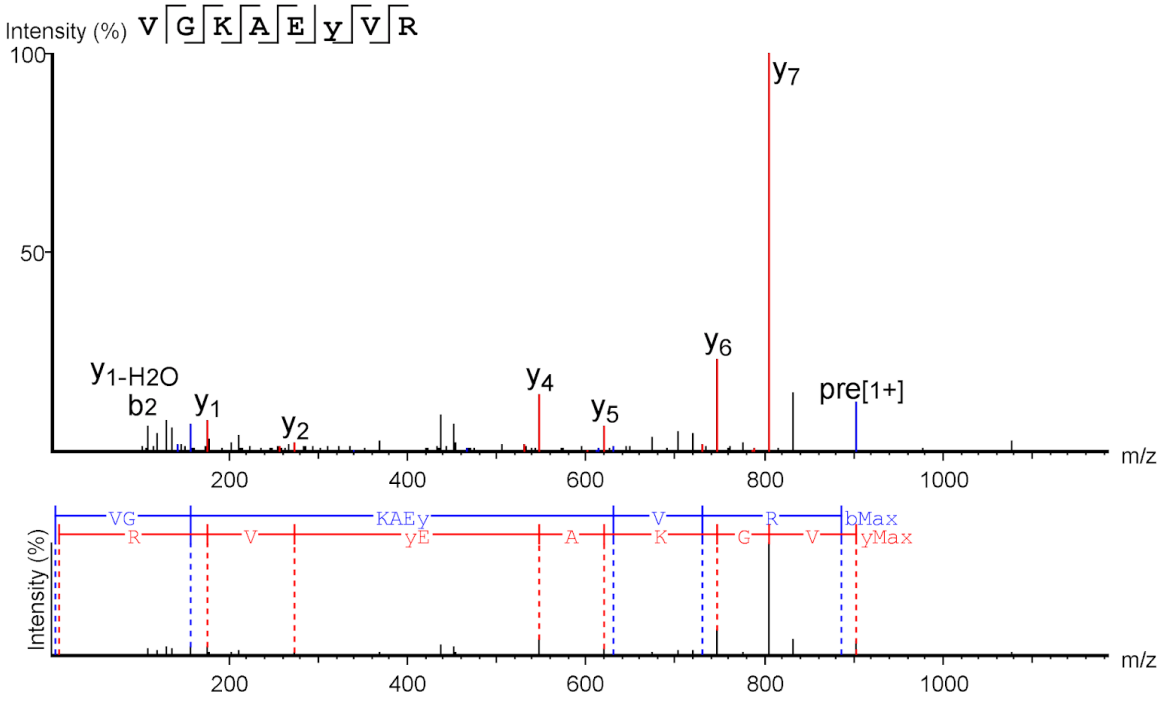
Source reference sequence = *Columba livia*, *m/z* = 508.7923, score -lgP = 38.45, ppm = 3.8, PSM = 14



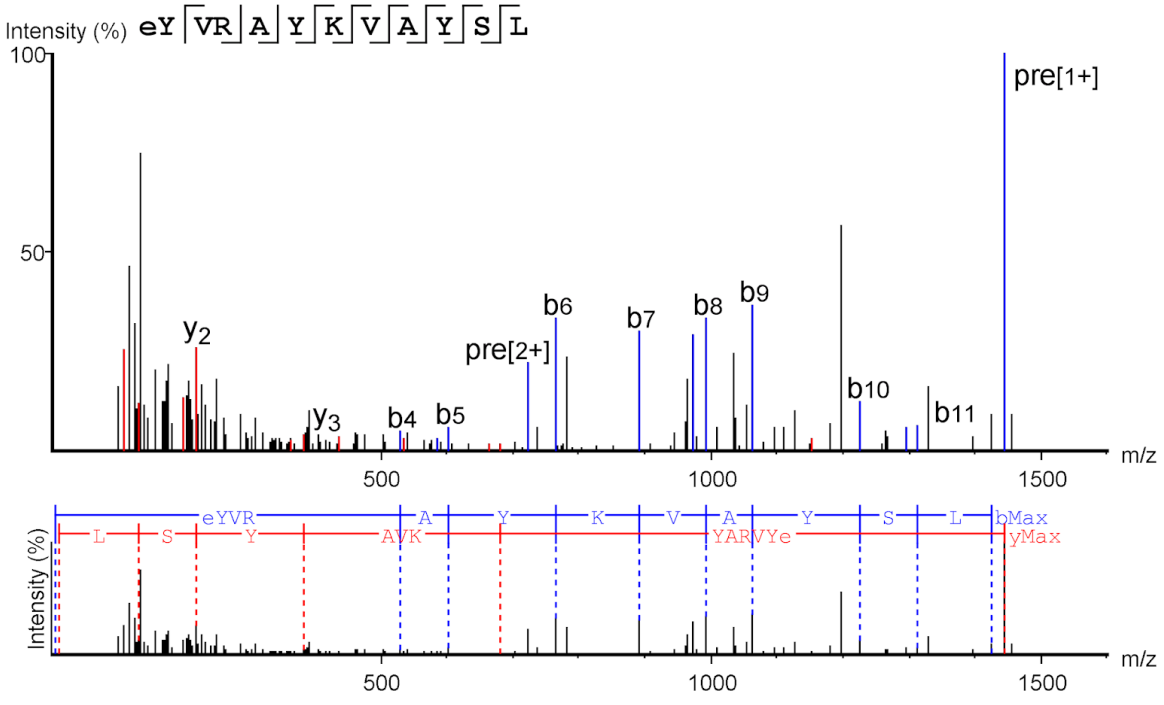
Source reference sequence = *Columba livia*, *m/z* = 486.7835, score -lgP = 34.63, ppm = 1.3, PSM = 1



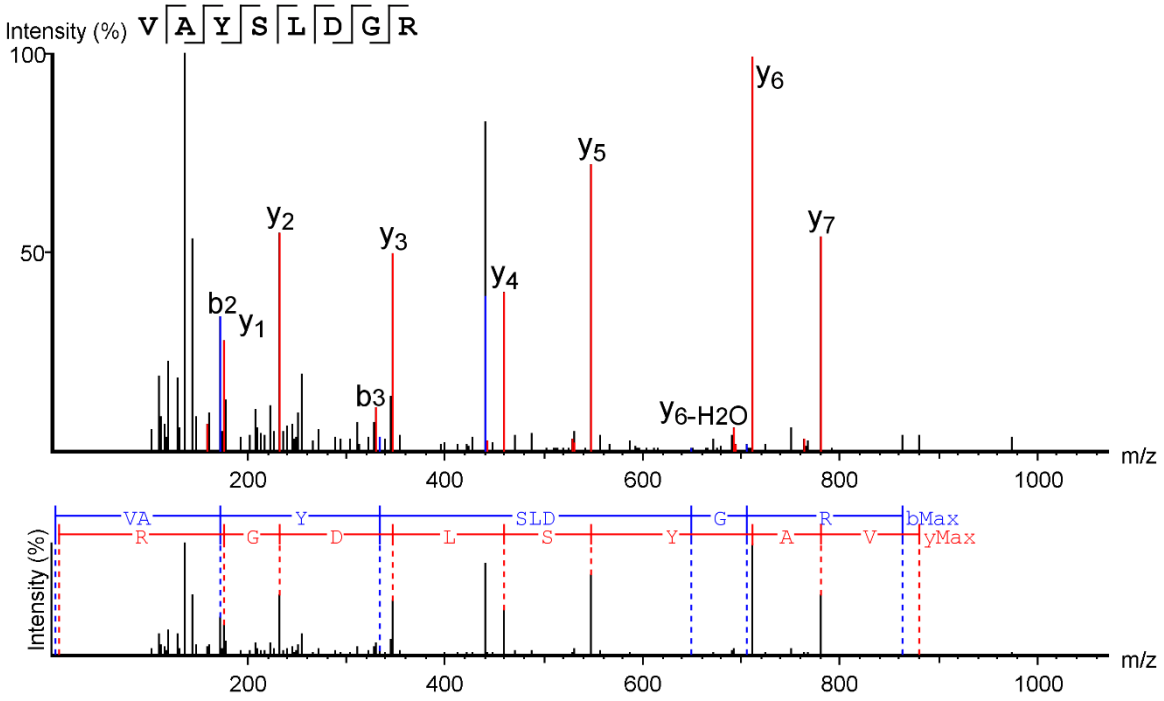
Source reference sequence = *Phasianus colchicus*, *m/z* = 464.2628, score -lgP = 38.42, ppm = 3.3, PSM = 3



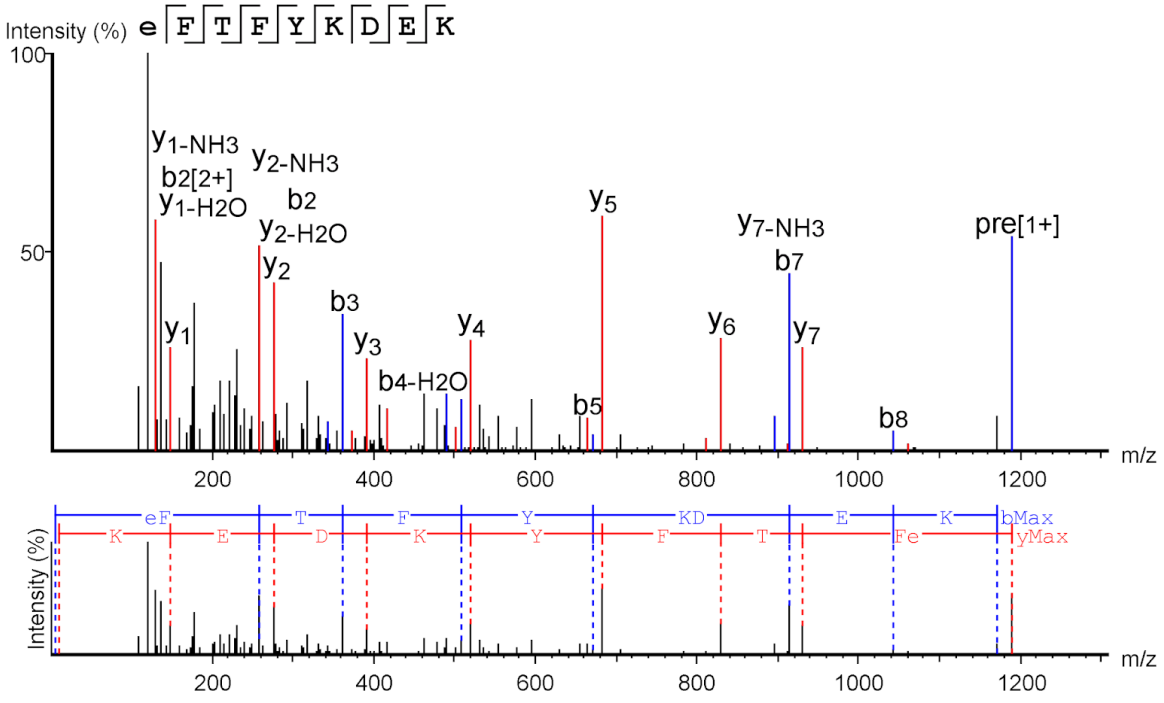
Source reference sequence = *Columba livia*, *m/z* = 722.389, score -lgP = 43.24, ppm = 5.4, PSM = 2



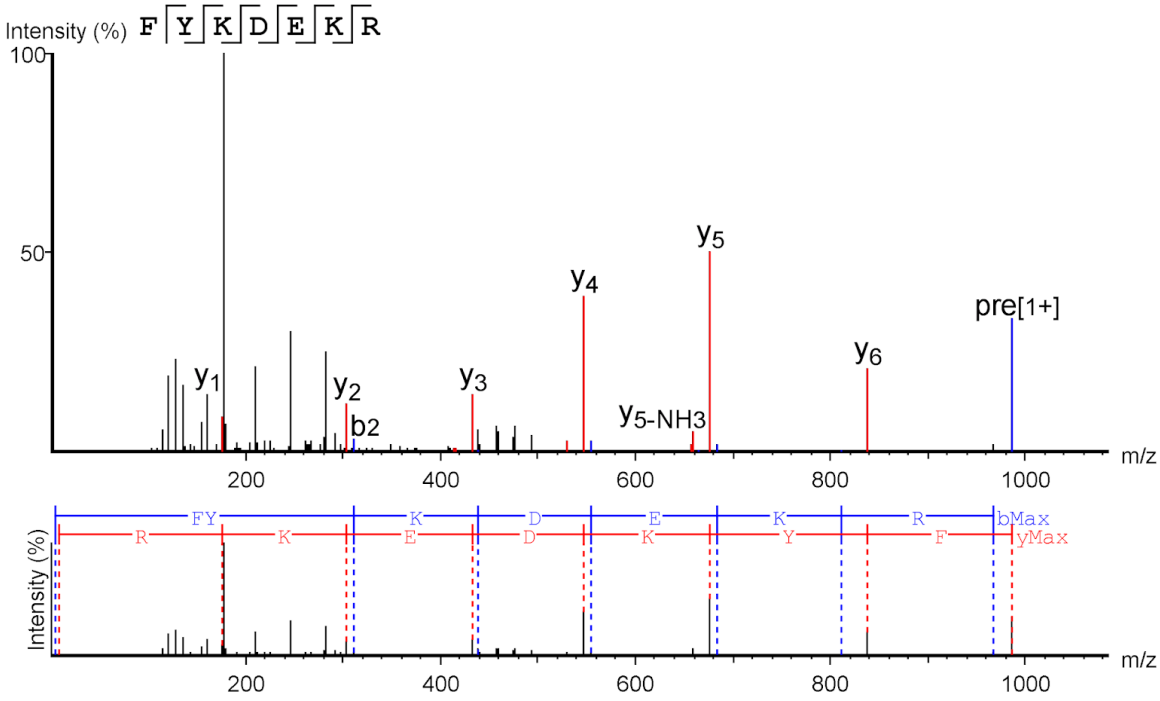
Source reference sequence = *Columba livia*, *m/z* = 440.7313, score -lgP = 39.13, ppm = 3.4, PSM = 13



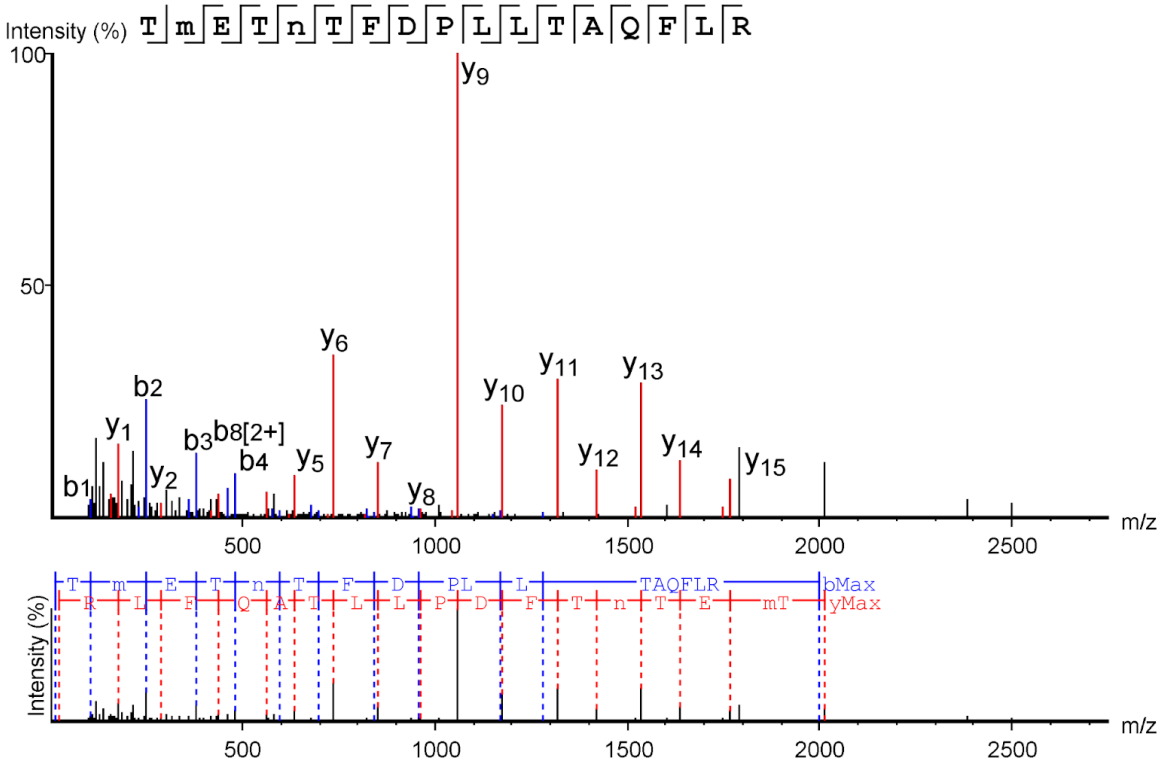
Source reference sequence = *Columba livia*, *m/z* = 603.7903, score -lgP = 45.03, ppm = 4.6, PSM = 11



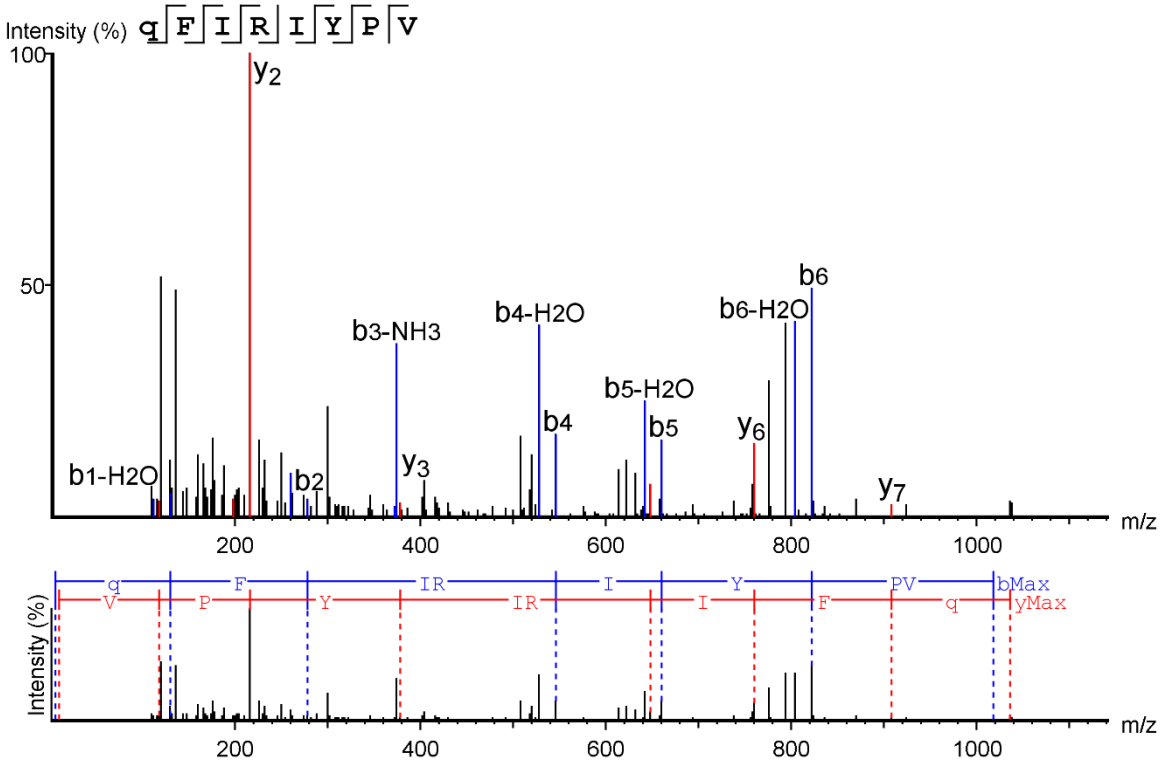
Source reference sequence = *Columba livia*, *m/z* = 415.2090, score -lgP = 37.46, ppm = 2.1, PSM = 2



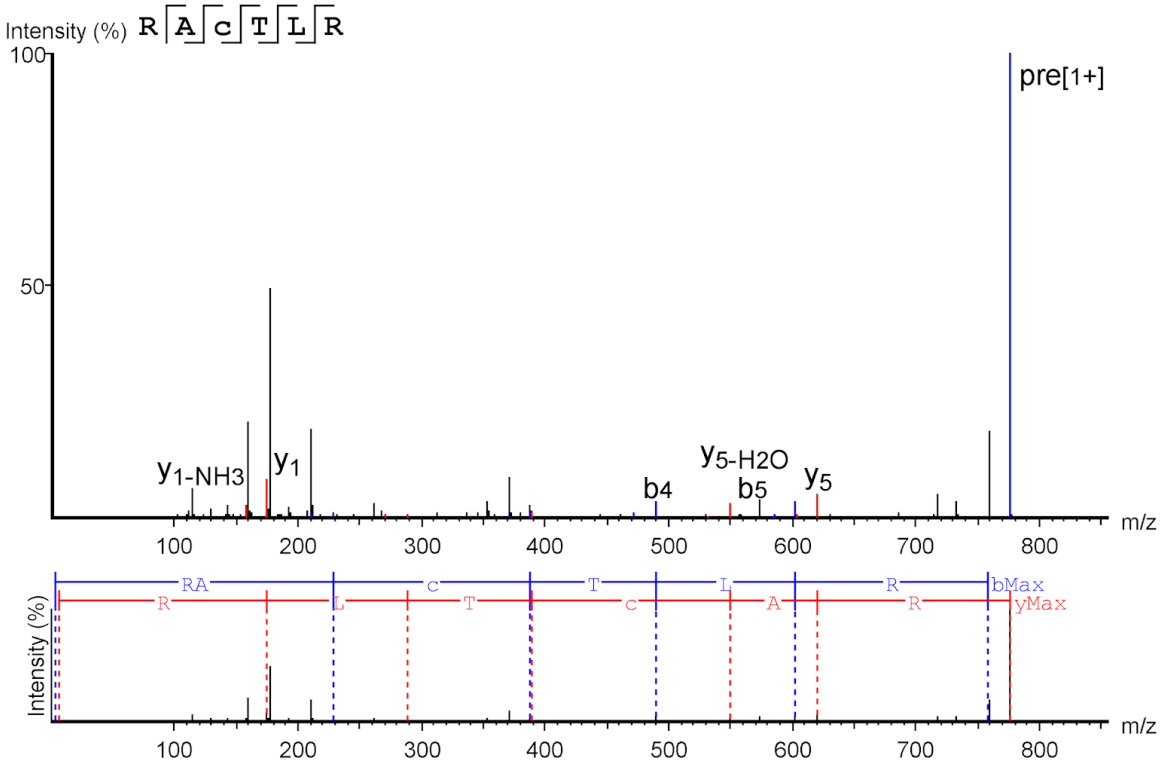
De novo sequence, *m/z* = 1008.0029, ALC score 90%, ppm = 9.7, PSM = 5



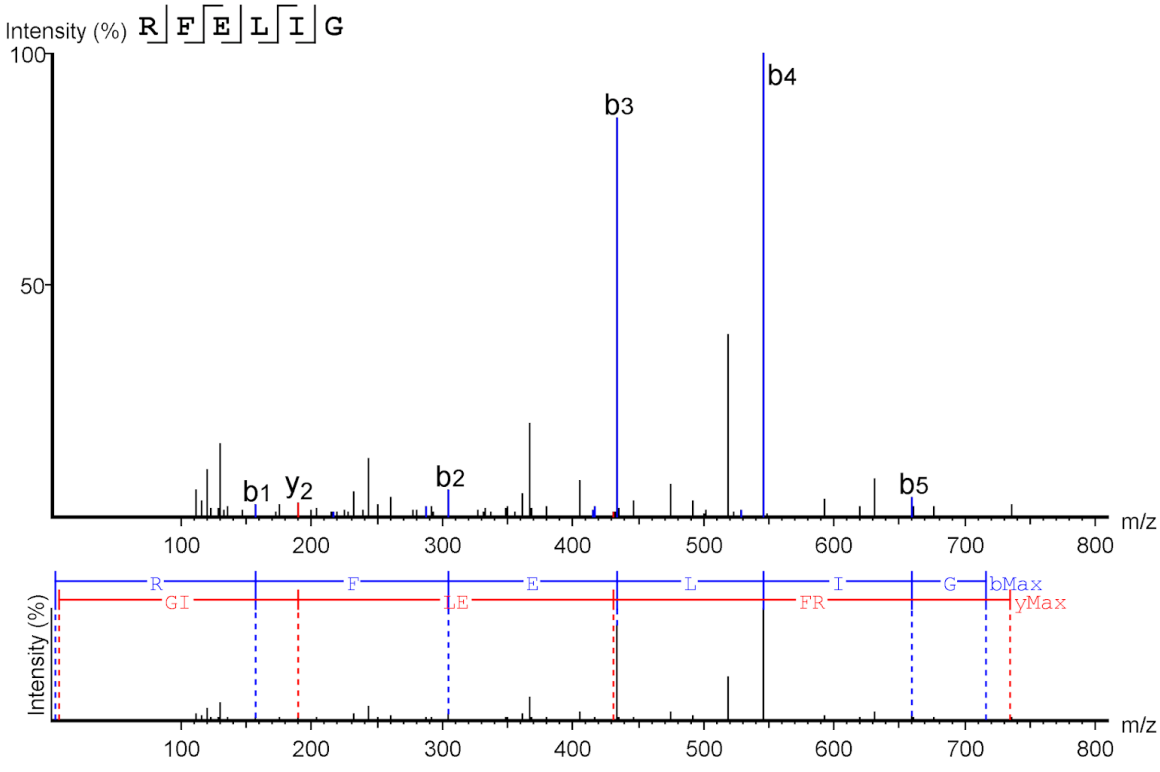
Source reference sequence = *Columba livia*, *m/z* = 518.7981, score -lgP = 43.46, ppm = 6.1, PSM = 6



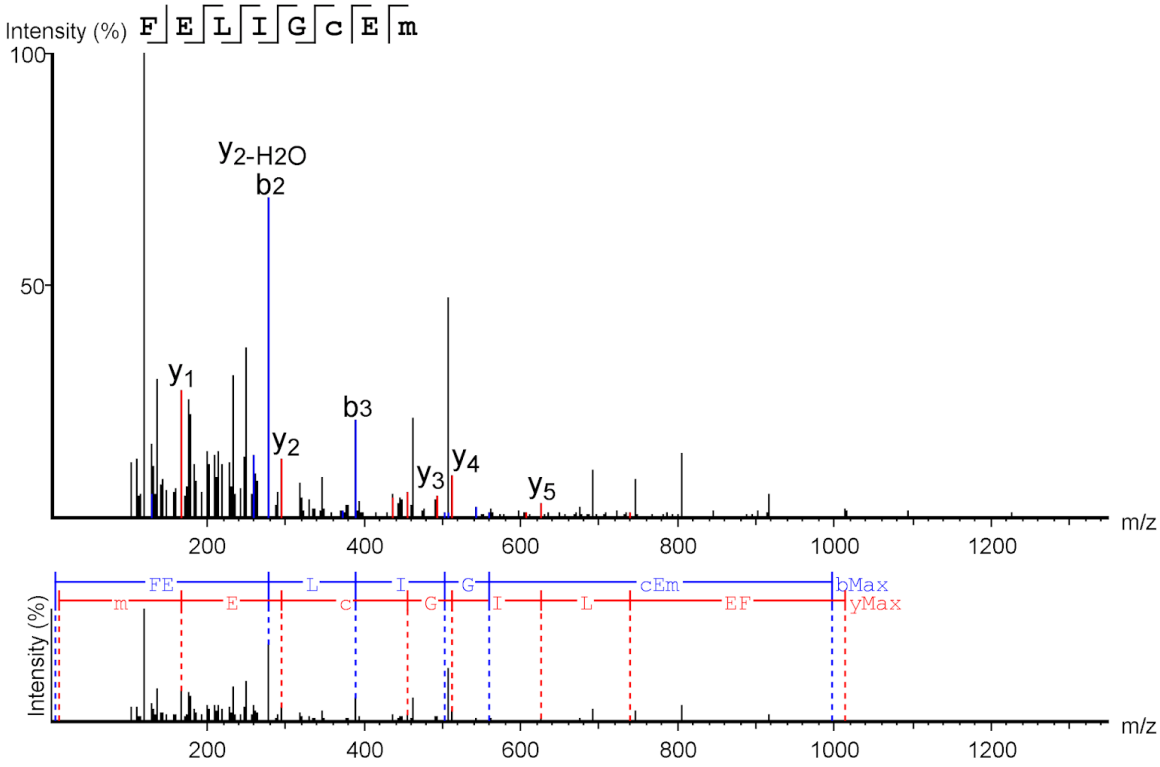
Source reference sequence = *Leptomus discolor*, *m/z* = 388.7139, score -lgP = 34.31, ppm = 1.1, PSM = 3



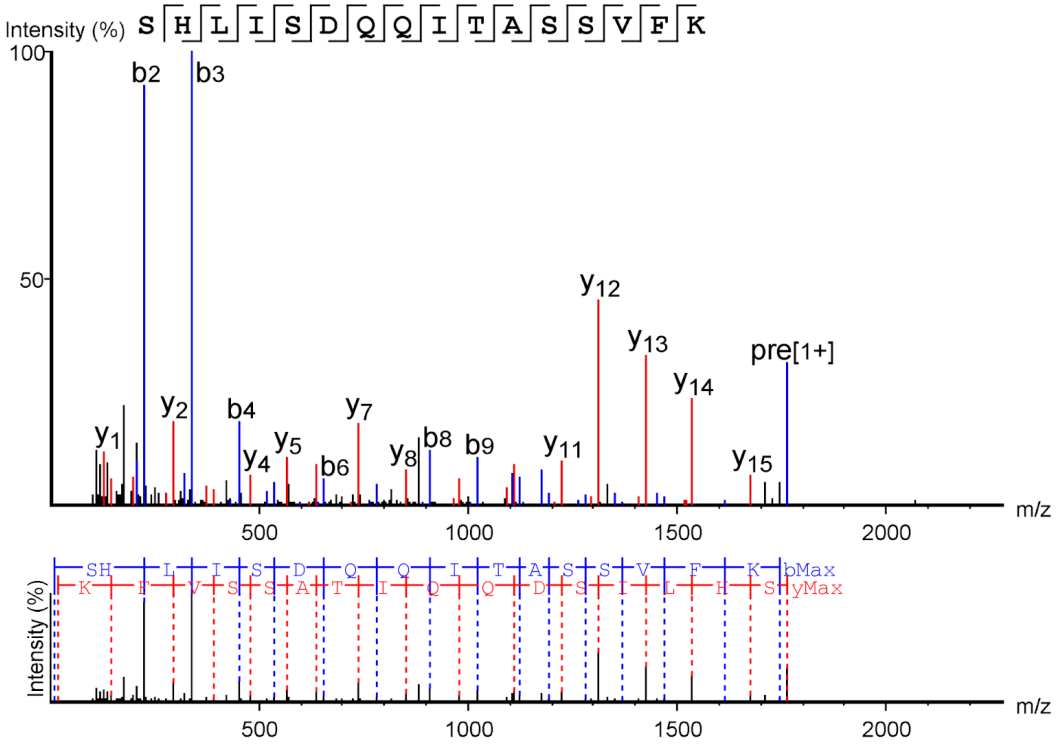
Source reference sequence = *Leptomus discolor*, *m/z* = 367.7149, score -lgP = 32.62, ppm = 4.0, PSM = 1



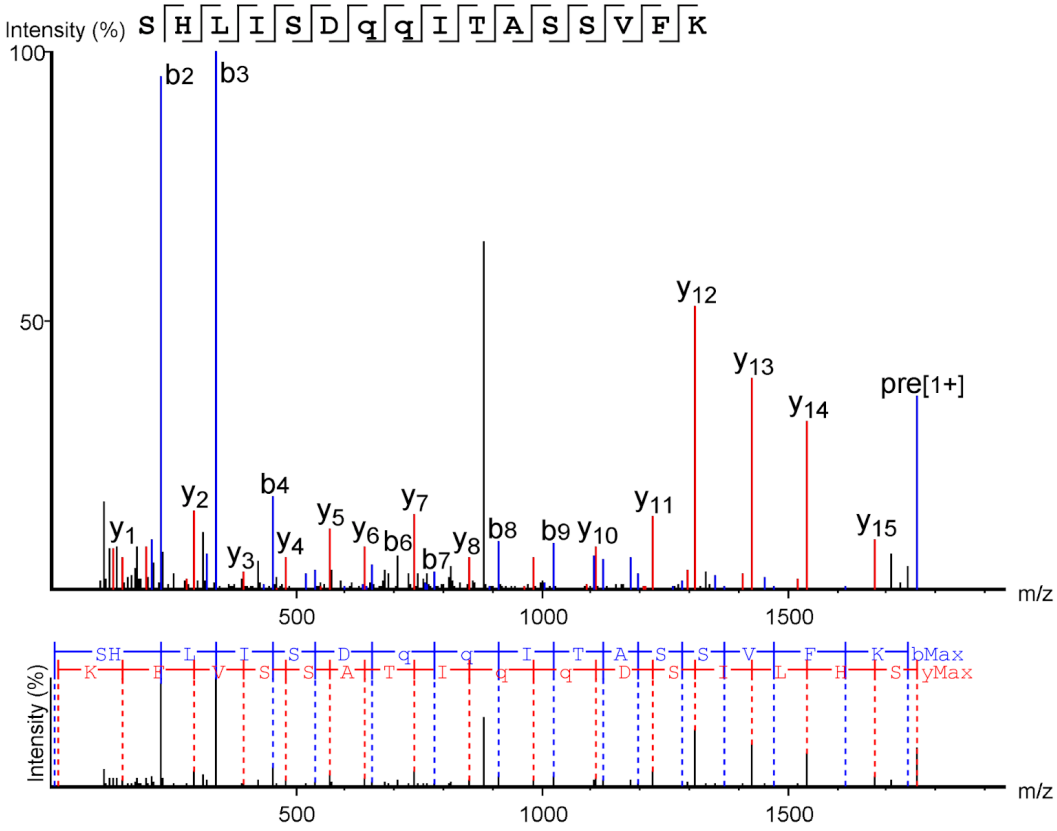
Source reference sequence = *Leptomus discolor*, *m/z* = 475.7208, score -lgP = 36.57, ppm = 5.8, PSM = 3



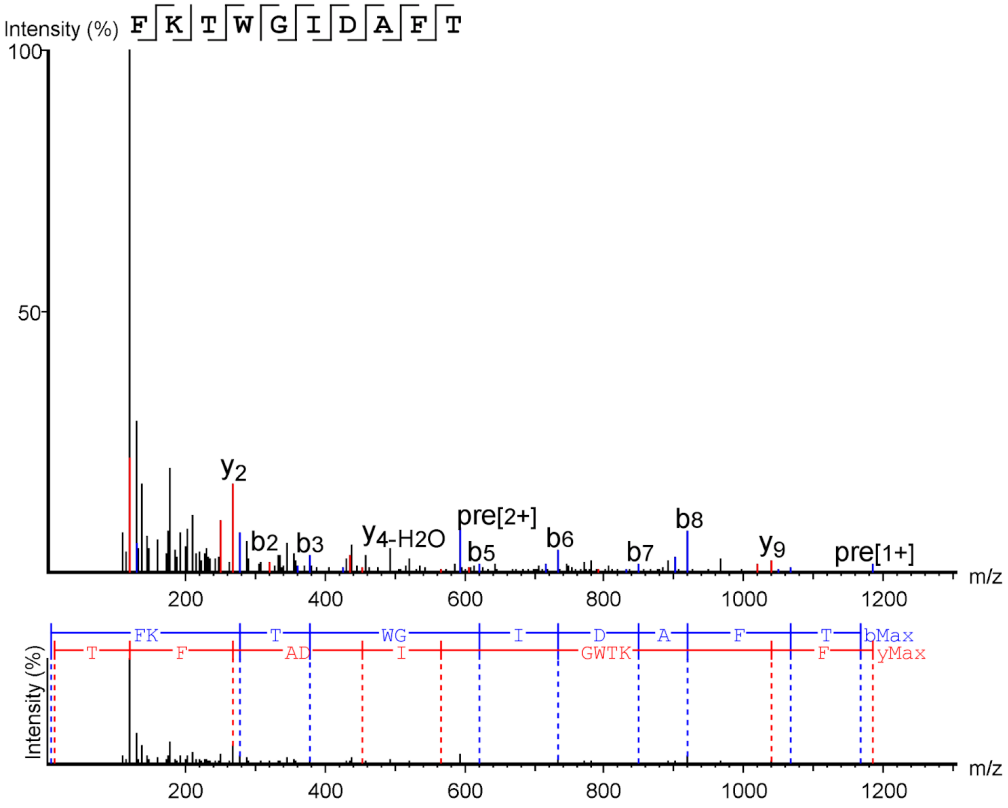
Source reference sequence = *Columba livia*, *m/z* = 880.9683, score -lgP = 74.10, ppm = 6.6, PSM = 5



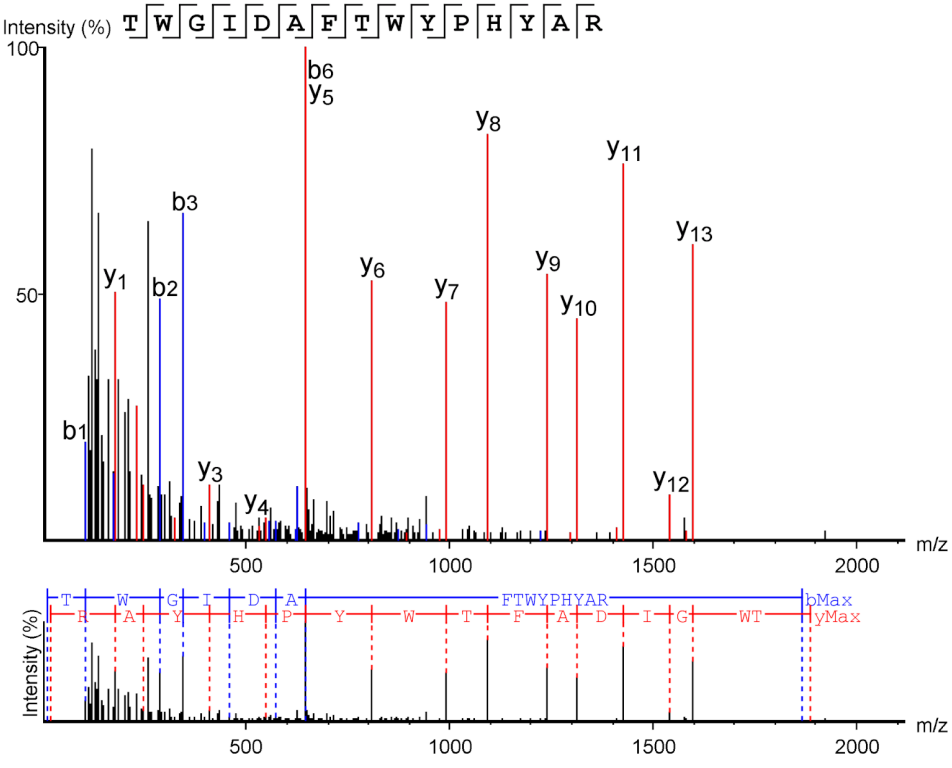
Source reference sequence = *Columba livia*, *m/z* = 881.9504, score -lgP = 66.29, ppm = 4.7, PSM = 8



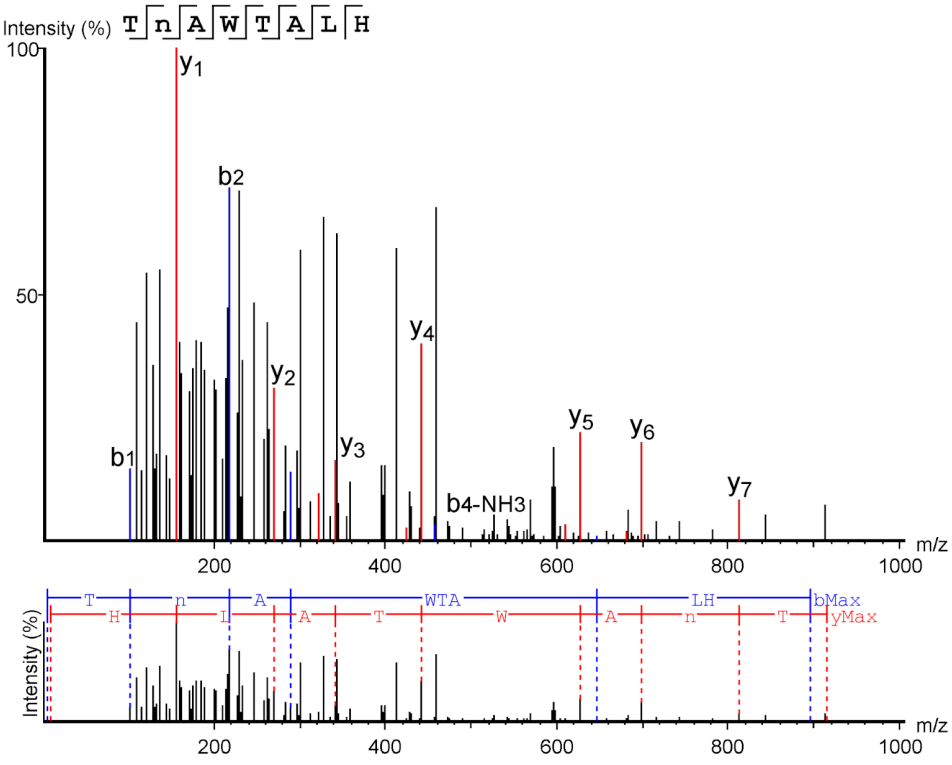
Source reference sequence = *Columba livia*, *m/z* = 593.3040, score -lgP = 42.61, ppm = 5.5, PSM = 2



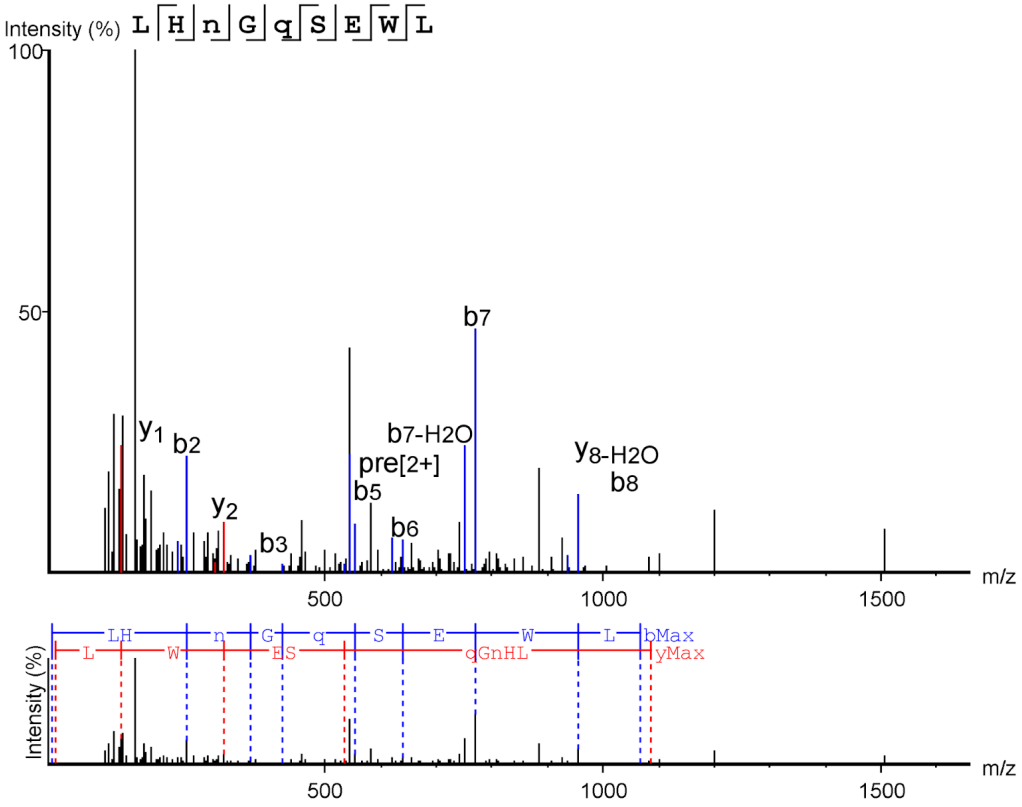
Source reference sequence =*Crotophaga sulcirostris*, *m/z* = 942.4535, score -lgP = 49.88, ppm = 7.2, PSM = 12



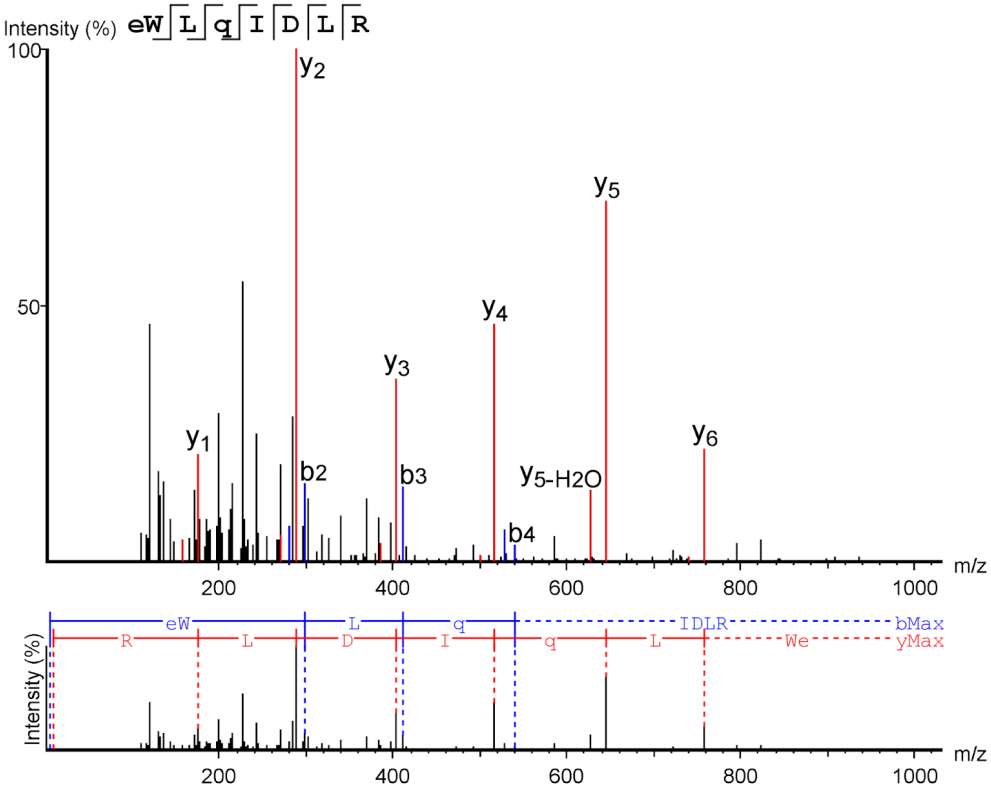
Source reference sequence = *Columba livia*, *m/z* = 457.7243, score -lgP = 40.67, ppm = 5.1, PSM = 1



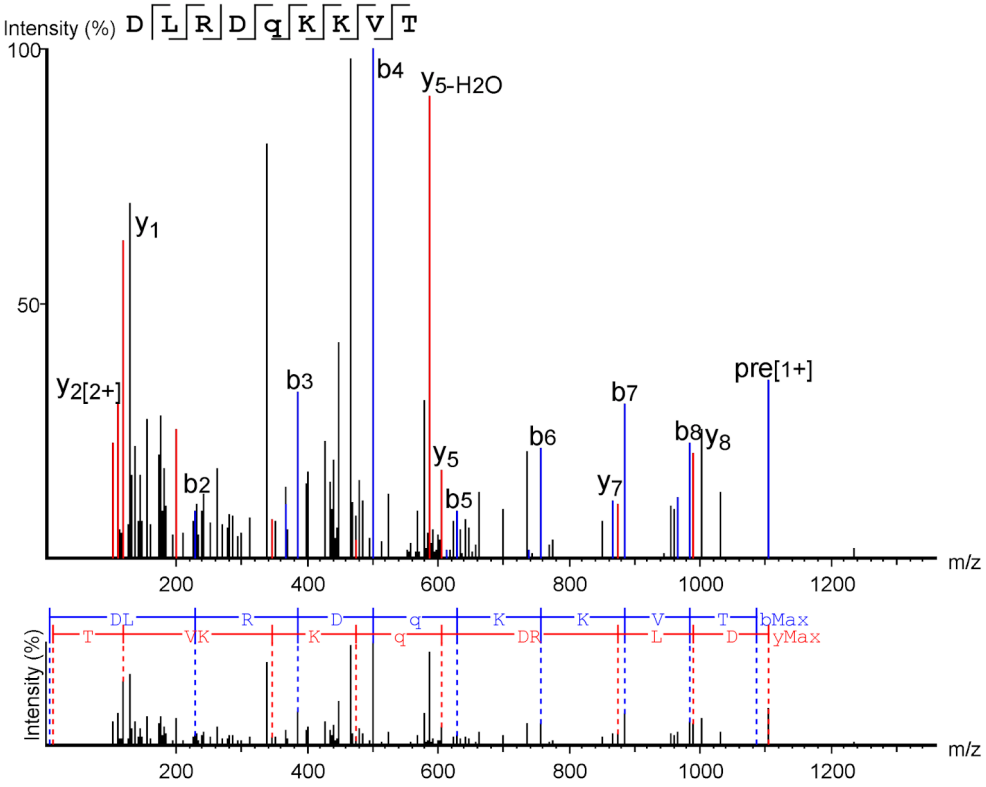
Source reference sequence = *Columba livia*, *m/z* = 543.2518, score -lgP = 31.06, ppm = 5.9, PSM = 2



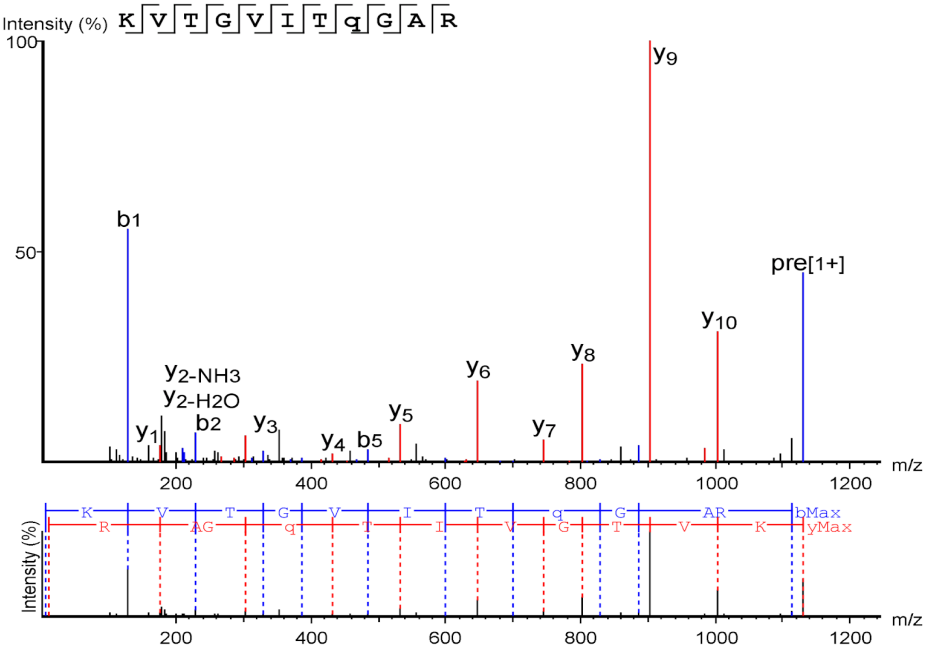
Source reference sequence = *Columba livia*, *m/z* = 528.2831, score -lgP = 32.60, ppm = 6.5, PSM = 1



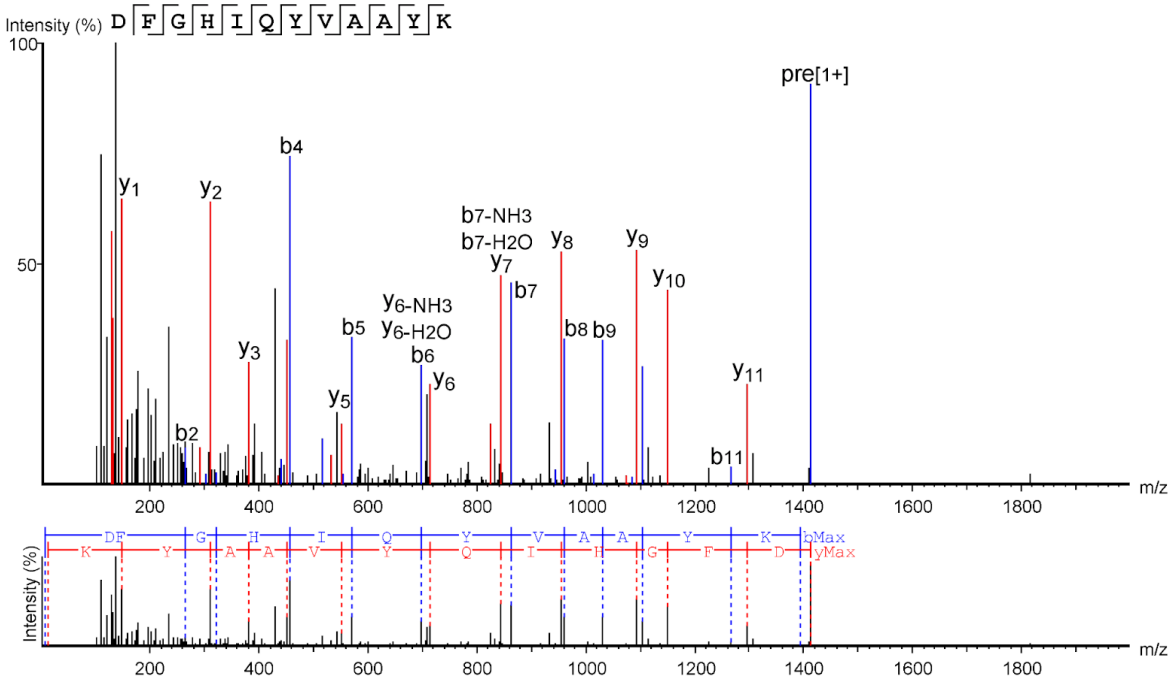
Source reference sequence = *Columba livia*, *m/z* = 368.5406, score -lgP = 42.51, ppm = 1.8, PSM = 3



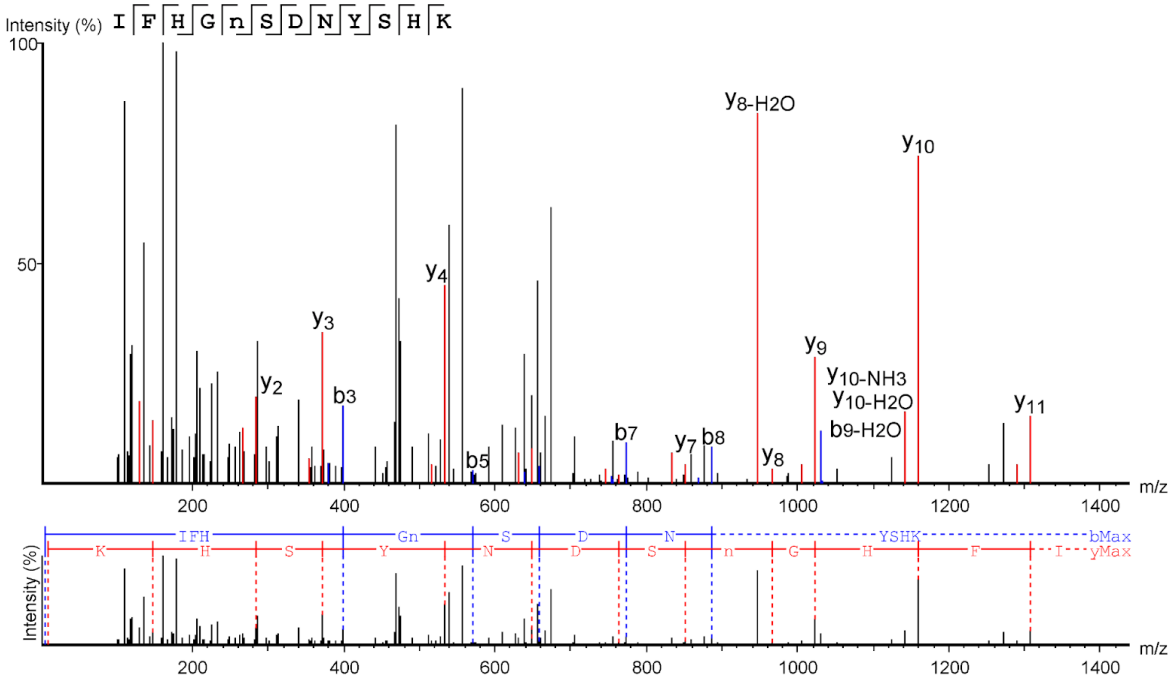
Source reference sequence = *Onychorhynchus coronatus*, *m/z* = 565.8321, score -lgP = 43.52, ppm = 3.7, PSM = 7



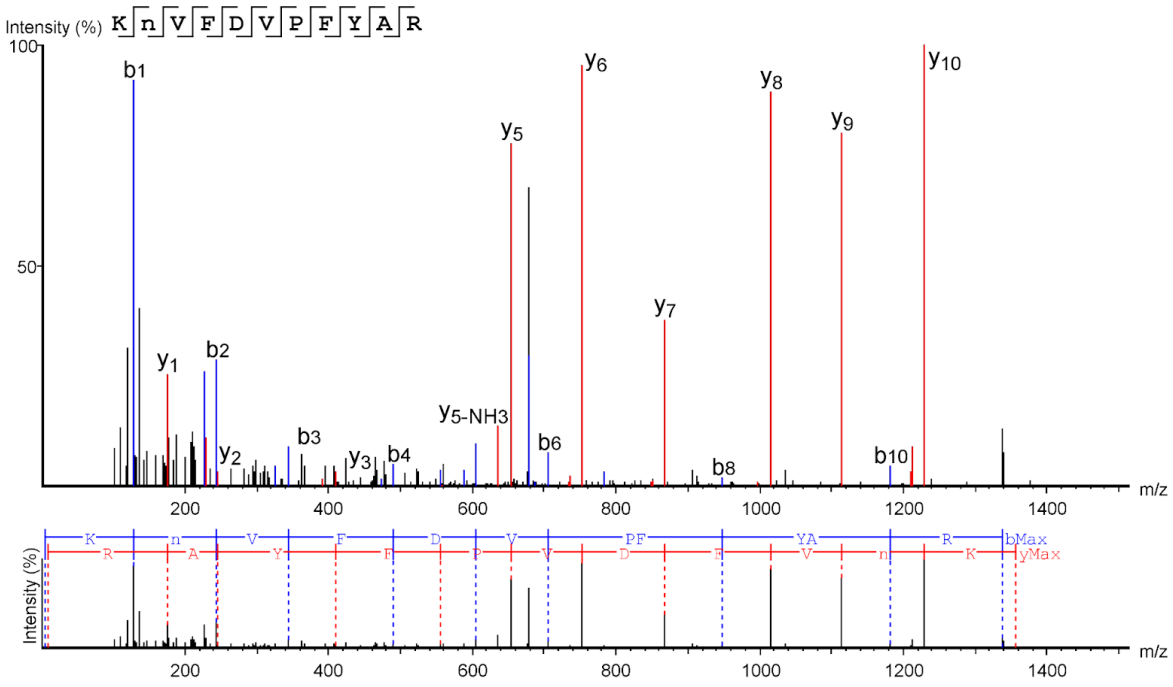
Source reference sequence = *Onychorhynchus coronatus*, *m/z* = 706.3588, score -lgP = 63.08, ppm = 7.1, PSM = 8



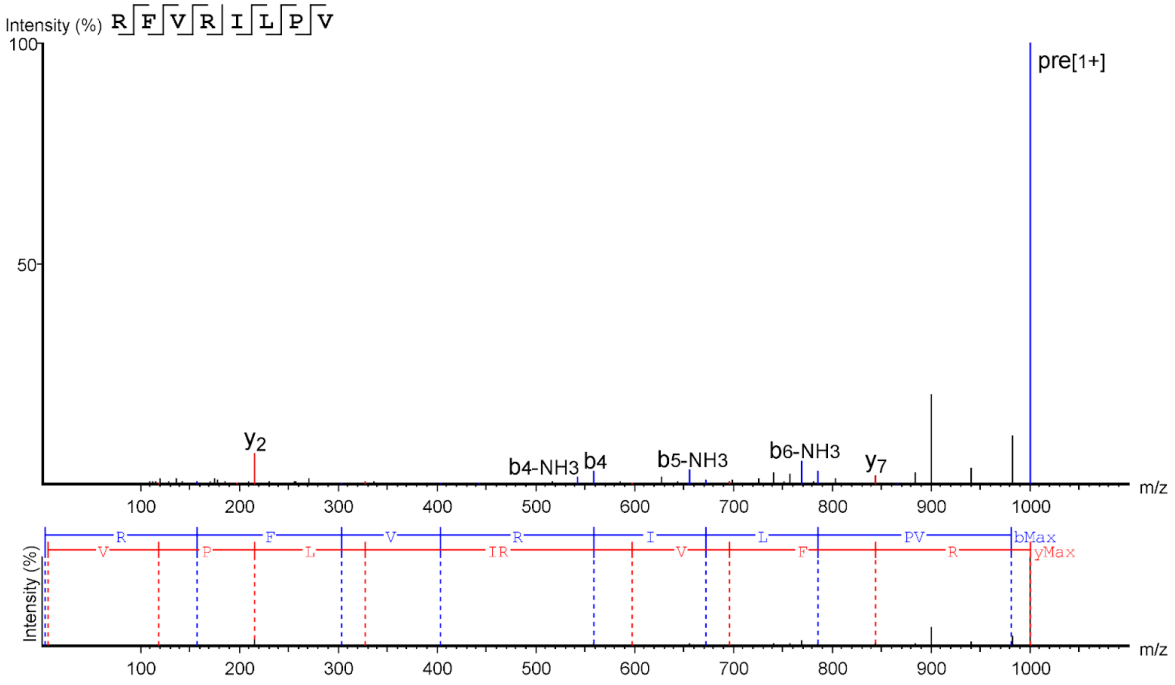
Source reference sequence = *Columba livia*, *m/z* = 473.8817, score -lgP = 46.56, ppm = 1.3, PSM = 2



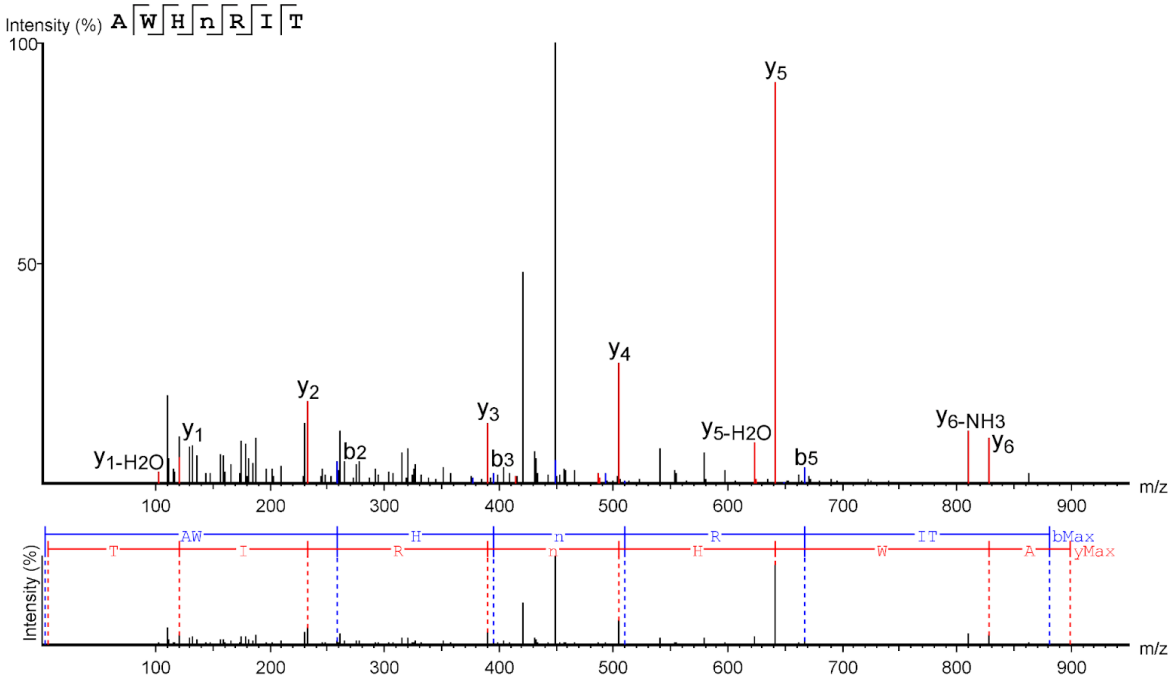
Source reference sequence = *Columba livia*, *m/z* = 678.8552, score -lgP = 49.22, ppm = 6.1, PSM = 19



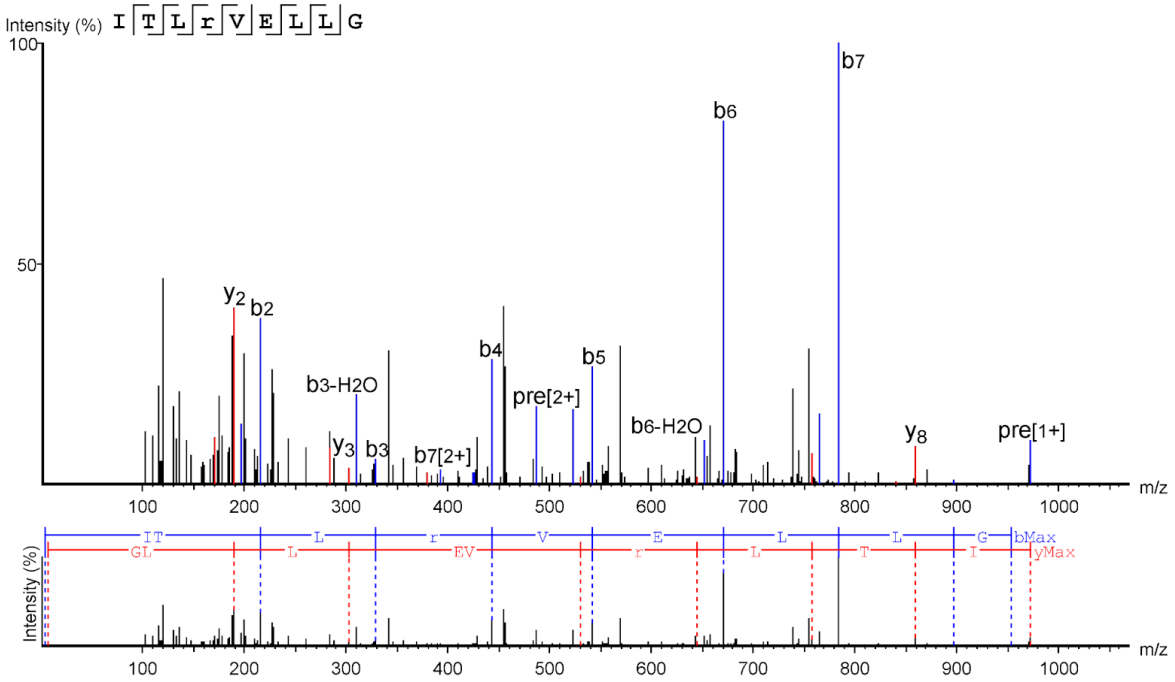
Source reference sequence = *Columba livia*, *m/z* = 500.3297, score -lgP = 38.54, ppm = 5.7, PSM = 3



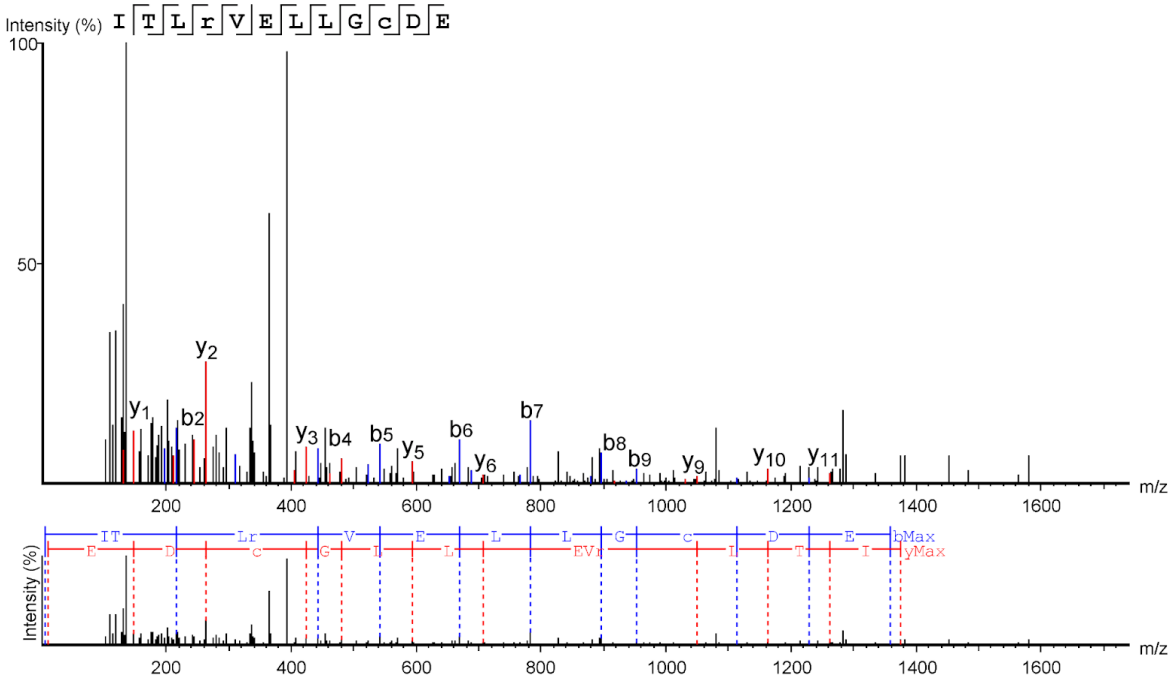
Source reference sequence = *Columba livia*, *m/z* = 449.7315, score -lgP = 38.55, ppm = 3.0, PSM = 9



Source reference sequence = *Columba livia*, *m/z* = 486.3126, score -lgP = 34.62, ppm = 4.6, PSM = 1



Source reference sequence = *Columba livia*, *m/z* =688.3644, score -lgP = 49.82, ppm = 5.8, PSM = 3



SI II Supplementary Figures

II.1 Protein sequences alignments

Figures S1 and S2 were created using Geneious version 2021.2 created by Biomatters. Available from https://www.geneious.com.

Alignments highlight all disagreements to the consensus sequence.

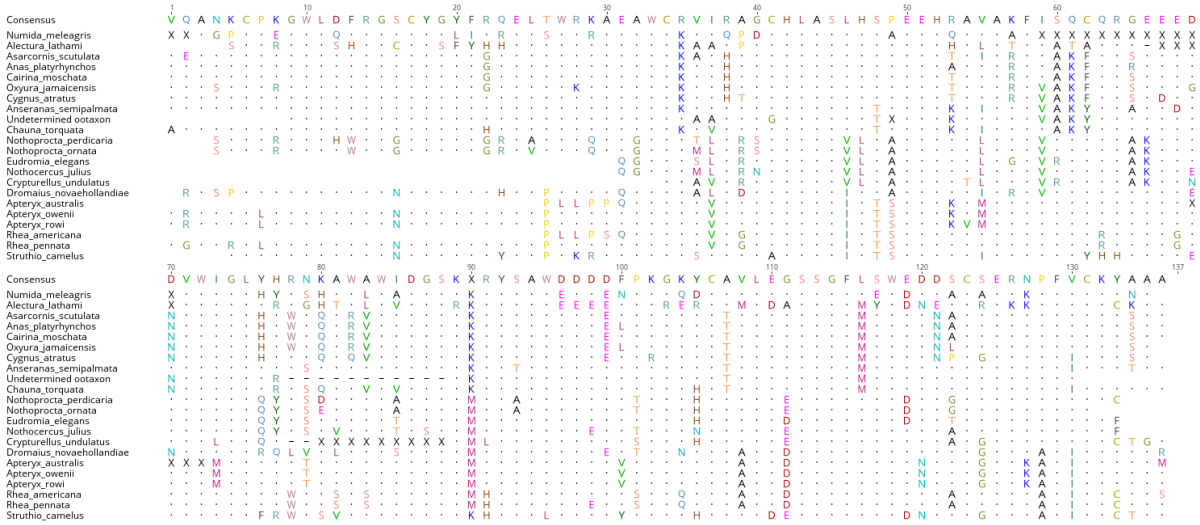


Fig. S1: Alignment of XCA-1 sequences including 21 B10K references and UO.

Fig. S2: Alignment of lactadherin sequences including 40 B10K references and UO.

II.2 Predicted XCA-1 structure

Missing regions of the XCA-1 sequence (underlined) were inferred from modern data.

NKCPKGWLDFRGSCYGYFRQELTWRKAEAWCRAARAGGHLASLHTPEEHKAVAKFVAKYQRGEEE
DNVWIGLYRRNKAWAWIDGSKKRYSAWDDDDFPKGKYCTVLEGSSGFMSWEDDSCSERNPFVVCK
CAAA

The missing data comprises the first and last three residues of the sequence and amid region. A structure for XCA-1 was inferred for this sequence using the AlphaFold2 (1) as described (https://github.com/deepmind/alphafold) on the Danish National Supercomputer Computerome2 and compared with struthiocalcin-1 (PDB 4uww) (Fig. S3A). The missing region was predicted to be an exposed loop connecting two beta sheets.All three are areas which are predicted to have low values in a Local Distance Difference Test (lDDT). In order to highlight the ability of this region to accommodate variable sequences we substituted sequences from other Anatidae and Palaeognathae (Fig. S3B).

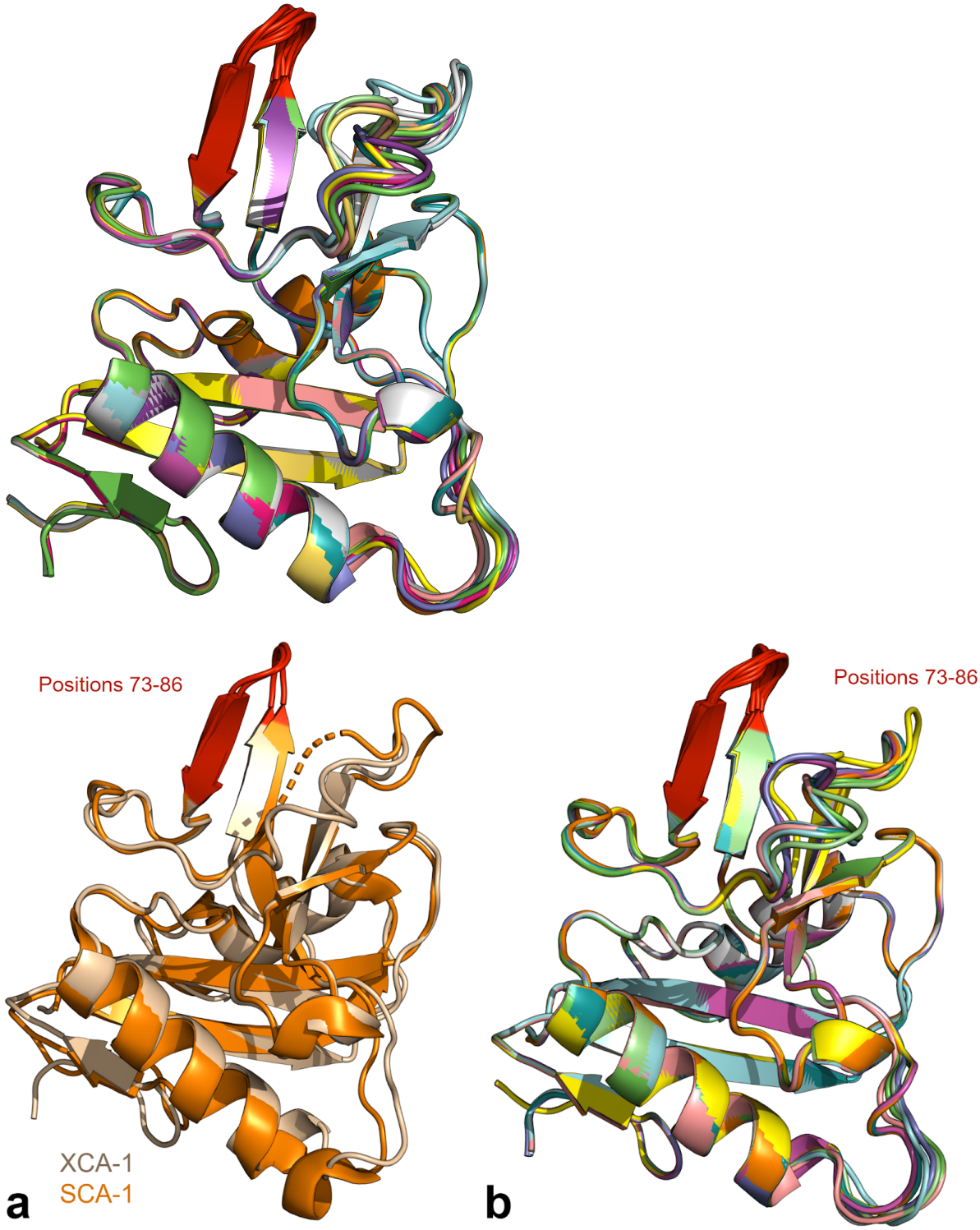


Fig.S3 XCA-1 structure prediction by AlphaFold (2) compared with struthiocalcin-1 (SCA-1) (3) (orange), showing the variable region (76-83) (red) which was not recovered. Running models of XCA1 in which this is replaced by other sequences taken from Anatidae and Palaeognathae highlights that this exposed loop can accommodate multiple different amino acid substitutions.

Modifications used in Fig S3

>XCA-1_w_Alectura_lathami

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRGHTWL
WVDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Anas_platyrhynchos__Asarcornis_scutulata__Cairina_moschata__Oxyura_jama
icensis

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRWNQARV
WIDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Anseranas_semipalmata

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSKAWA
WIDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Apteryx_australis__Apteryx_owenii__Apteryx_rowi

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRTKAWA
WIDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Chauna_torquata

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSQAWV
WVDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Cygnus_atratus

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRNQAQV
WIDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Dromaius_novaehollandiae

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRLVKLWA
WSDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Eudromia_elegans

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSKAWA
WTDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Nothocercus_ulius

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSKVWA
WTDSSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Nothoprocta_ornata

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSEAWA
WADGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Nothoprocta_perdicaria

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSDAWA
WADGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Numida_meleagris

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRRSHAW
LWADGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Rhea_americana__Rhea_pennata

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRWNKSW
SWIDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaaa

>XCA-1_w_Struthio_camelus

nkcpkgwldfrgscygyfrqeltwrkaeawcraaragghlaslhtpeehkavakfvakyqrgeeednvwiglYRWNSVW
AWIDGSKKrysawddddfpkgkyctvlegssgfmsweddscsernpfvvckcaa

II.3 Phylogenetic trees

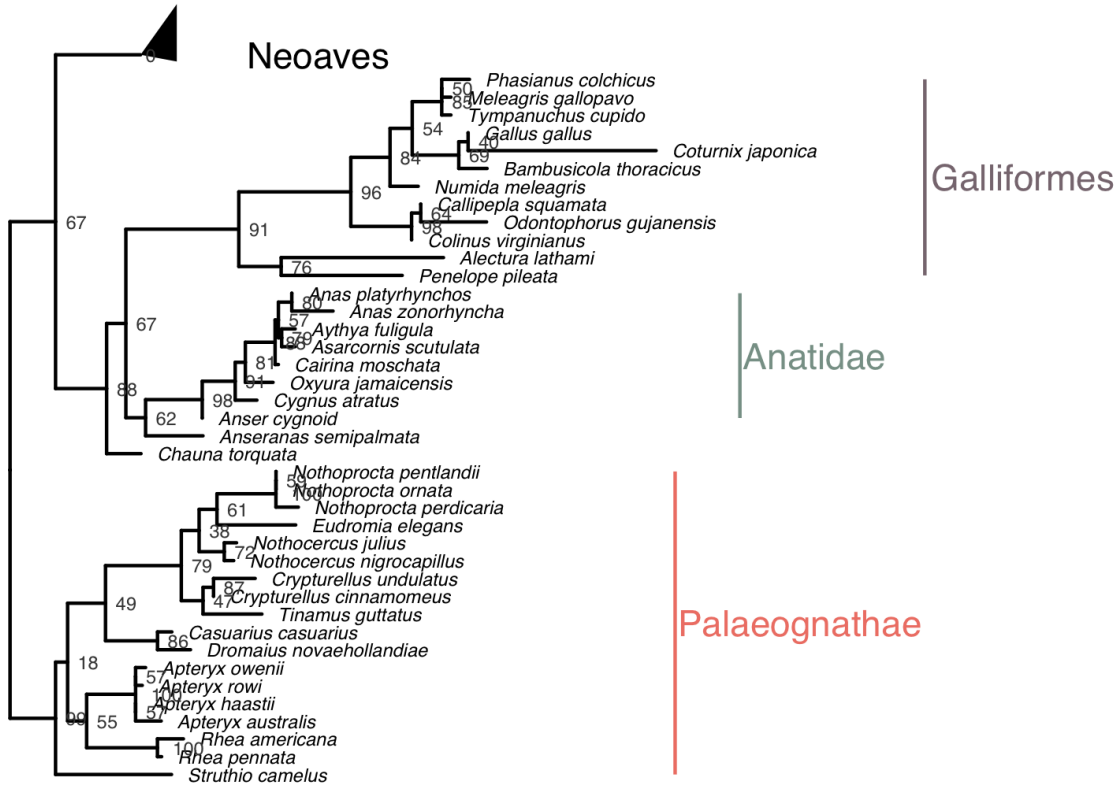


Fig. S4: Phylogenetic relationships from analysis of amino acid data of both genes concatenated for the extant sequences without the undetermined ootaxon. Neoaves are collapsed. The full tree file with all taxa can be found on FigShare. Node numbers are bootstrap support.

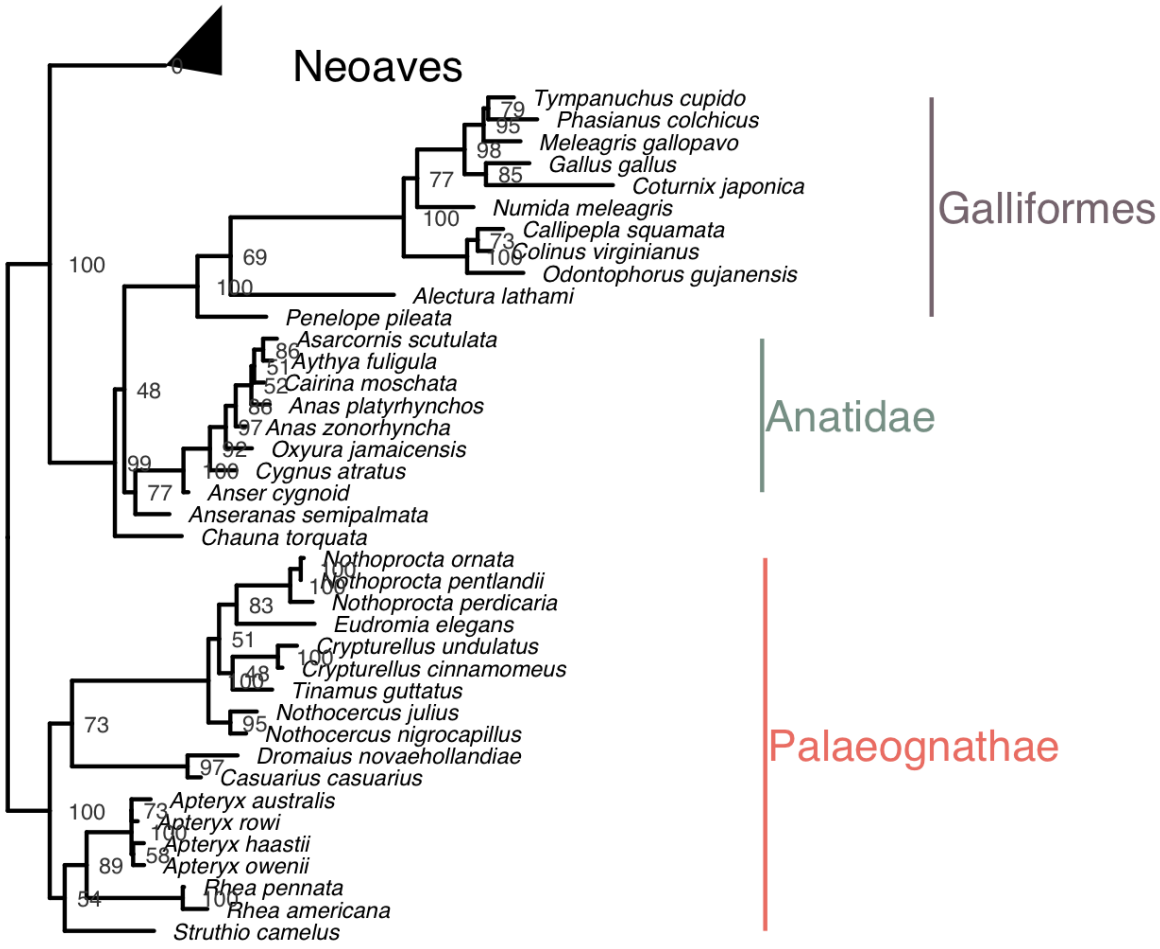


Fig. S5: Phylogenetic relationships from analysis of nucleotide data of both genes concatenated for the extant sequences without the undetermined ootaxon. Neoaves are collapsed. The full tree file with all taxa can be found on FigShare.

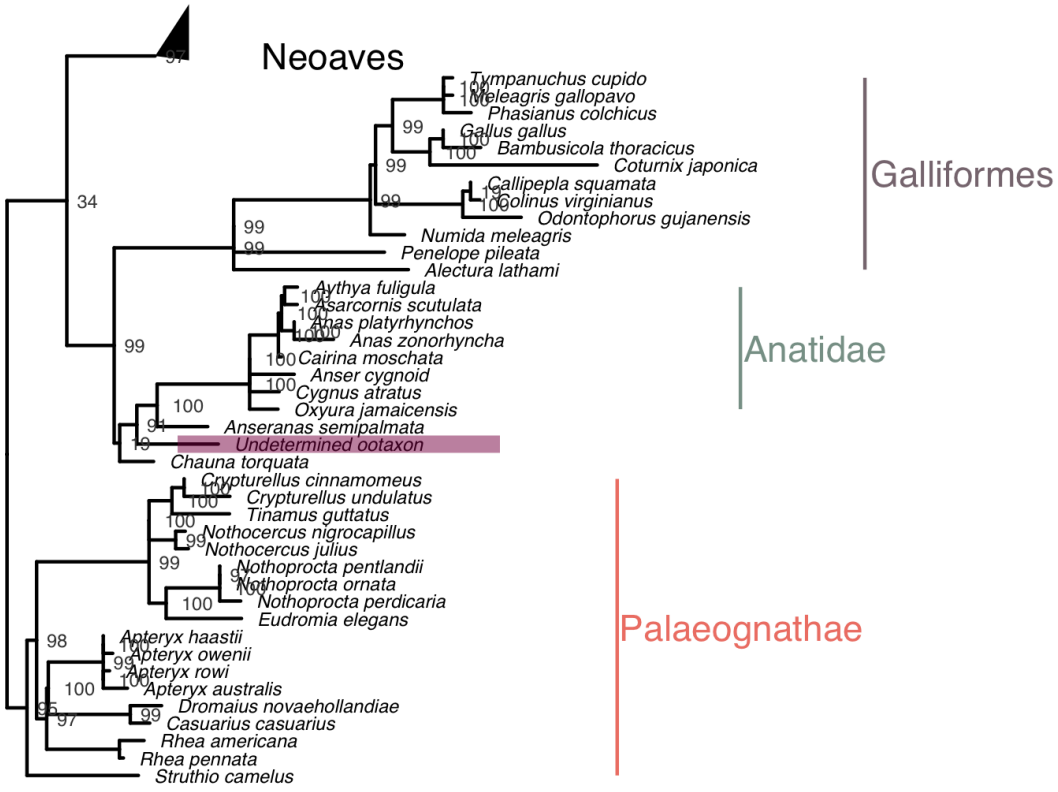


Fig. S6: Phylogenetic relationships from analysis of amino acid sequences of both genes concatenated when a topological constraint is applied to fix the relationships among living species. Neoaves are collapsed. The full tree file with all taxa, in addition to the constraint tree file can be found on FigShare.

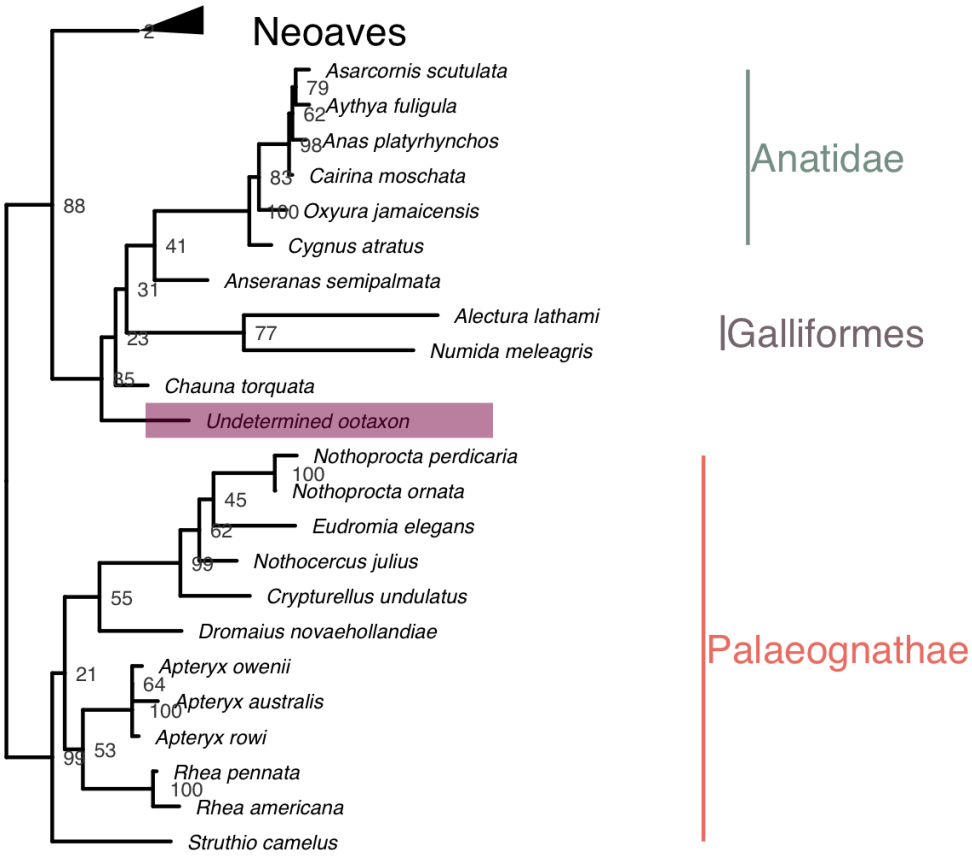


Fig. S7: Phylogenetic relationships from analysis of amino acid sequences of both genes concatenated for species that have both genes represented (106 species + the undetermined ootaxon). Neoaves are collapsed. The full tree file with all taxa can be found on FigShare.

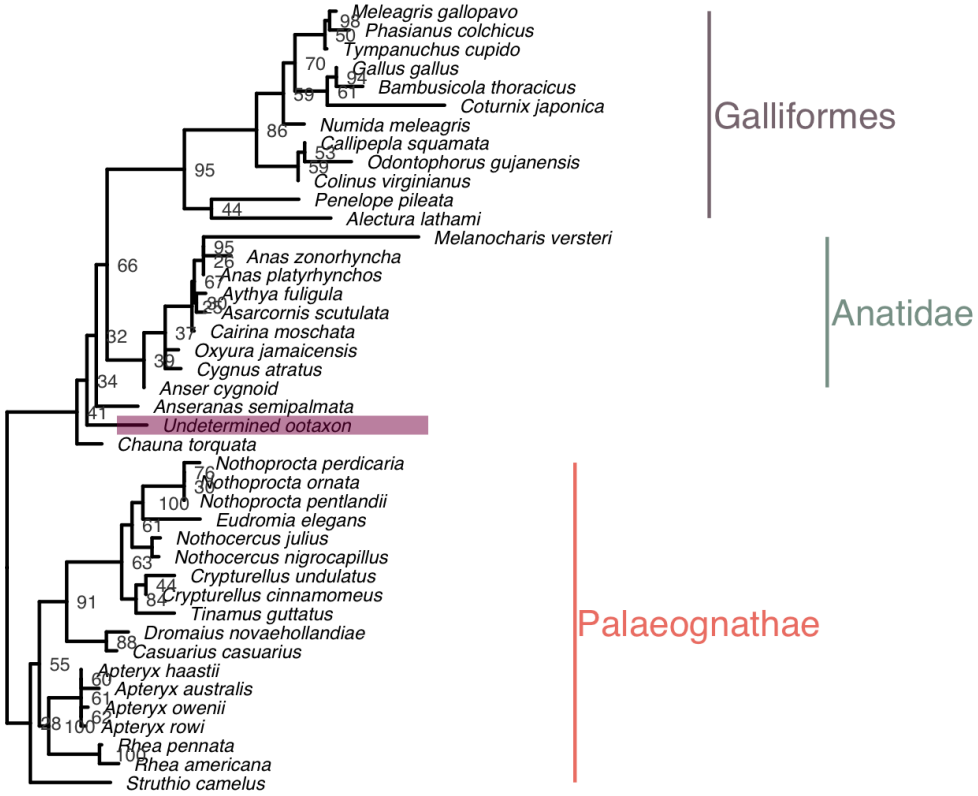


Fig. S8: Phylogenetic relationships from analysis of amino acid sequences of both genes concatenated for species in Palaeognathae and Galloanseres (41 species + the undetermined ootaxon). Neoaves are collapsed. The full tree file with all taxa can be found on FigShare.

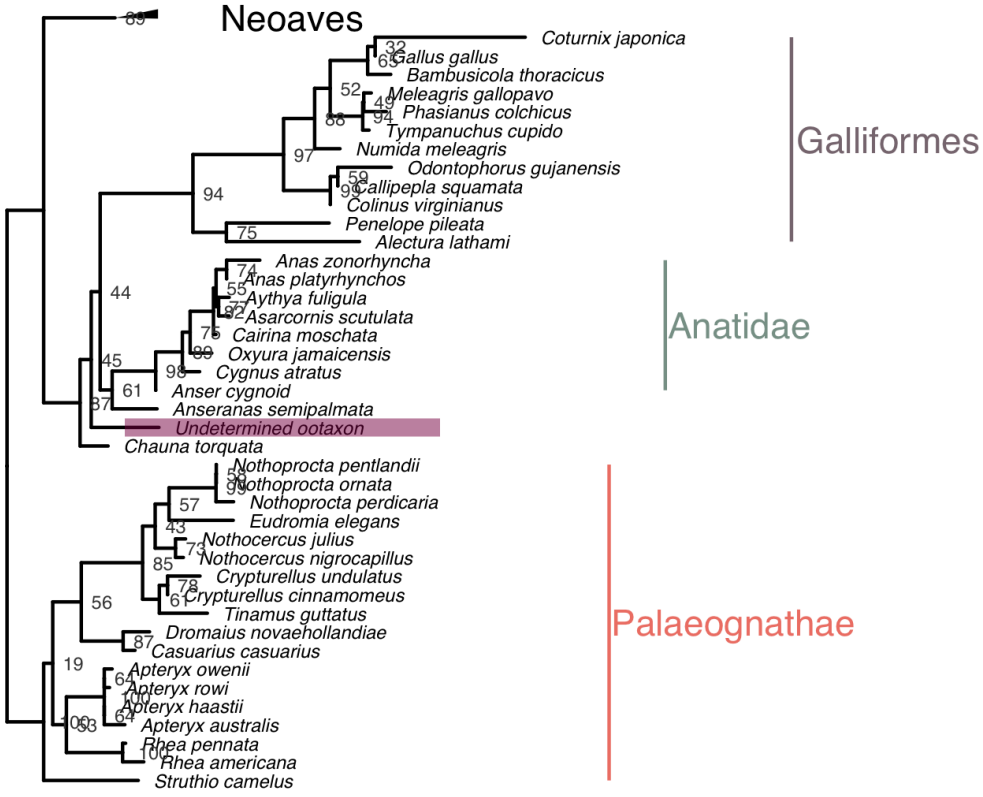


Fig. S9: Phylogenetic relationships from analysis of amino acid sequences of both genes concatenated for species that occur in Australia and adjacent islands (92 species + the undetermined ootaxon). Neoaves are collapsed. The full tree file with all taxa can be found on FigShare.

SI III Ancient DNA supplementary methods

Table S1 Details of the eggshell samples that underwent DNA extraction and sequencing. Asterisked * samples were sent for protein sequencing; ** samples where protein data were obtained.

Ancient DNA ID	Field ID	Locality	DNA extraction ID	Amplification of *12S* rDNA	Shotgun library ID	Enriched library ID
AD2077	M14-A023	Spencer Gulf North -33.3471 137.8990	MB3681	DEC2017; no amplification		
AD2078**	M14-A031	Spencer Gulf North -33.3475 137.8988	MB3621	SEP2017; no amplification		
AD2079**	M14-A035	Spencer Gulf North -33.8842 137.6110	MB3622	SEP2017; no amplification	SS3622	CSS3622
AD2080**	M14-A036	Spencer Gulf North -33.8841 137.6110	MB2564 MB3680	MAR2015, DEC2017; no amplification		
AD2081	M14-A124	Arcoona Station, Woomera -31.0985 136.9873	MB3624	SEP2017; no amplification		
AD2082*	M14-A130	Arcoona Station, Woomera -31.0869 137.0002	MB3625	SEP2017; no amplification		
AD2083*	M14-A138	Arcoona Station, Woomera -31.0865 137.0028	MB3626	SEP2017; no amplification		
AD2084*	M14-A147	Arcoona Station, Woomera -31.0817 136.9966	MB3627	SEP2017; initial amplification, no amplification on indexing	SS3627	Not captured; library was similar to the

							no-template controls
AD2085*	M14-A154	Arcoona Woomera 136.9873	Station, -31.0858	MB3628	SEP2017; initial amplification, no amplification on indexing	SS3628	Not captured; library was similar to the no-template controls
AD2086*	M14-A163	Arcoona Woomera -31.1040 137.0068	Station,	MB3629	SEP2017; no amplification		
AD2087	M14-A168	Arcoona Woomera -31.0974 137.0132	Station,	MB3630	SEP2017; no amplification		
AD2088*	M14-A179	Arcoona Woomera 137.0080	Station, -31.1238	MB3617 MB3620	SEP2017; no amplification		
AD2089*	M14-A194	Arcoona Woomera 137.0011	Station, -31.1228	MB3536-3539 MB3618 MB3686	JAN2017, SEP2017, DEC2017; no amplification	SS2089	CSS2089
AD2090*	M14-A227	Arcoona Woomera -31.1372 136.9896	Station,	MB3541-3544 MB3619 MB3623	JAN2017, SEP2017; amplification of MB3542 only--barcode 100% ID to *Gallus gallus*	SS2090	CSS2090

The surface of all eggshell specimens was cleaned with 10% household bleach followed by 70% ethanol. The surfaces were then drilled off using a Dremel tool, and discarded. Eggshell was ground in a fine powder using a Retsch planetary ball mill (PM200) at 400 rpm. 200 mg of powder was added to 1.66 ml of 0.25 mg/ml Proteinase K in 0.5 M EDTA and pre-digested for 30 minutes at 55°C with rotation at 10 rpm, after which the samples were centrifuged for three minutes at maximum speed in a microcentrifuge. The supernatant was discarded and replaced by an additional 1.66 ml of digest buffer. Samples were then

incubated as above for another 24 hours, after which they were centrifuged for 10 min at maximum speed in a bech-top centrifuge (Eppendorf) to pellet the debris.

DNA was extracted from the supernatant following the method described by Dabney et al. (4), with minor changes outlined in Grealy et al. (5). DNA was eluted in 15 ul EB buffer (QIAGEN). Several extractions were performed for a few of the most promising samples (Table S1) and extracts were combined and concentrated to 30 µl in a Vivaspin-500 MWCO 3000 column.

SI III.1 Amplification of the 12S rDNA mini-barcode

We followed the methods described in SI 1.0 of Grealy et al. ((6)) and SI 1.4 of Grealy et al. ((7)) to amplify the *12S* rDNA mini-barcode. PCR reaction set-up was carried out in a designated ultra-clean facility at Curtin University (WA, Australia) in a designated UV hood inside a physically separate room from the DNA extraction and sample preparation room. 12 µl qPCR reactions were prepared consisting of final concentrations of: 1.2 mg/ml Bovine Serum Albumin (*Fisher*), 1X PCR Gold Buffer (*Applied Biosystems*), 2.5 mM $MgCl_2$ (*Applied Biosystems*), 0.4 µM forward primer (*IDT*), 0.4 µM reverse primer (*IDT*), 0.25 mM dNTPs (*Bioline*), 1.25 U AmpliTaq Gold DNA polymerase (*Applied Biosystems*), 0.12X SYBR Green (*ThermoFisher Scientific*), and 2 µl of DNA, up to volume with HPLC-grade water (*Invitrogen*). PCR thermal cycling conditions were as follows: 95 °C for 5 min, 50 cycles of: 95 °C for 30 sec, 54°C for 30 sec, 72 °C for 45 sec, followed by a final extension at 72 °C for 10 min. DNA free PCR negative controls were also included.

For the one extract (MB3542) where amplification was successful, the above PCRs were repeated in quadruplicate, using fusion primers structured as follows (where NNNNNNNN represents a unique multiplex identifier sequence):

Forward: 5'- Illumina P5 (AATGATACGGCGACCACCGAGATCTACAC) - Custom Sequencing adapter (TGACGACATGGTTCTACA) - UMI (NNNNNNNN) - Gene-specific primer 12SA (CTGGGATTAGATACCCCACTAT) -3'

Reverse: 5'- Illumina P7 (CAAGCAGAAGACGGCATACGAGAT) -UMI (NNNNNNNN) - Gene-specific primer 12SC (GTTTTAAGCGTTTGTGCTCG) -3'

Replicate reactions were pooled and purified using a Qiaquick PCR purification kit (*QIAGEN*) as per the manufacturer's instructions, and were sequenced as described in the main manuscript.

SI III.2 Shotgun library preparation

Shotgun sequencing libraries were prepared following (8) with minor changes described in SI 3.0 of Grealy et al. (6). Libraries were prepared in an ultra-clean environment at Curtin University. An extraction control, no-template (water) control, and CL104 positive control were also included in the library building process. Further modifications are described below.

Purification of adapters, and dephosphorylation, heat denaturation and ligation of the first adapter was performed in low-bind 0.2 ml 8-well PCR strip tubes. 6 µl of DNA extract was

used, and Afu UDG and Endonuclease VIII were replaced by ultrapure water during the dephosphorylation and heat denaturation step (i.e., 8 µl of 10X T4 RNA ligation buffer, 2 µl of 2% Tween-20, 1 µl of 1 U/µl FastAP, 6 µl of DNA were added to the reaction, which was topped up to a total reaction volume of 46 µl with Ultrapure water). The reaction was incubated for 10 min at 37°C in a thermal cycler, followed by 2 min at 95°C, then placed immediately in an ice water bath. Ligation of the first splinter adapter (see Table S2 (6)) proceeded as per Gansauge et al. ((8)), as did the remaining steps aside from the modifications below.

A thermal cycler was used to anneal CL130 for 2 min at 65°C as opposed to a thermal shaker. A thermal cycler was used to elute the final library. All other steps were performed in a thermal shaker. At step 25, the supernatant was stored in a 0.5 ml Lo-Bind *Eppendorf* tube at -20°C. After quantitation of the library via qPCR (in a physically separated post-PCR laboratory), 10 µl of PCR product was combined with 0.5 µl of 6X loading dye (*QIAGEN*) and run alongside 3 µl of 50 bp DNA ladder (GeneRuler, *Fermentas*) for 1.5 hours at 85 V on a 2% agarose gel electrophoresis (2.2 g agarose, 110 ml 1X TAE buffer, 8 µl GelRed, 1x20-well comb) that was visualised and photographed using a *BioRad* transilluminator, in order to confirm the library building process was successful. Libraries that appeared to not have been successful (i.e., were similar to the negative control library) were not captured or sequenced.

The libraries were amplified in five replicate 25-µl qPCR reactions with unique fusion-tag indexing primers suitable for the *Illumina* sequencing platforms. The PCR reaction contained reagents in final concentrations of: 0.4 mg/ml BSA, 1X PCR buffer, 2.5 mM MgCl2, 0.25 mM dNTPs, 0.05 U/ul Amplitaq Gold DNA polymerase, 0.25X SYBR Green, 0.4 µM *IDT* forward primer (1 µl of a 10 µM stock); 0.4 µM *IDT* reverse primer (1 µl of a 10 µM stock); 5 µl of neat library, made up to a total of 25 µl final volume with HPLC-grade water (10.9 µl). In a separate, designated post-PCR laboratory, thermal cycling was performed under the following conditions: 95°C for 10 min, 13 cycles of 95°C for 30 sec, 60°C for 30 sec, 72°C for 30 sec. Replicate reactions were combined and vortexed to mix. Libraries were concentrated and purified using a *QIAGEN* MinElute PCR purification kit, following the manufacturer's instructions and eluting in 20 µl EB buffer. 2 µl was quantitated on the Qiaxpert spectrophotometer following the manufacturer's instructions.

SI III.3 Hybridisation capture

Capture reagents were prepared in an ultra-clean environment at Curtin University but capture was performed in a designated post-PCR laboratory.

In a 0.2-ml Lo-bind PCR tube (round capped), a solution was prepared (per reaction) containing final concentrations of 0.45 µg/µl Chicken Cot-1 (2.5 µl of a 1 µg/ul stock), 9.1 µM IDT forward blocking primer (Table S2; 0.25 µl of a 200 µM stock), 9.1 µM IDT reverse blocking primer (Table S2; 0.25 µl of a 200 uµM stock). In a separate 0.2-ml tube, a solution was prepared (per reaction) containing final concentrations of 9X Hyb#1 (i.e., 9 µl of 20X SSPE), 0.0125 M Hyb#2 (i.e., 0.5 µl of 0.5 M EDTA, pH 8.0), 8.75X Hyb#3 (i.e., 3.5 µl of 50X Denhardt's solution), 0.25% Hyb#4 (i.e., 0.5 µl of 10% SDS), 1 U/µl Rnase Block (SUPERase; i.e., 1 µl of a 20 U/µl solution of RNase block), and 5.5 ul of baits for a final volume of 20 µl. Solutions were gently vortexed to mix and briefly spun in a bench-top

microcentrifuge to collect the solution. In a post-PCR environment, 5 µl of the first solution was added to 7 µl of the library for a total of 12 µl.

Table S2 Oligos used for hybridisation capture and reamplification.

Oligos	Working concentration	Synthesis scale	Purification	Sequence (5'-3')
Ganmey_SSBlockF	200 uM	100 nmole	HPLC	AATGATACGGCGACCACCGAGATCTACACIIIIIIII ACACTCTTTCCCTACACGACGCTCTT/3InvdT /
Ganmey_SSBlockR	200 uM	100 nmole	HPLC	CAAGCAGAAGACGGCATACGAGATIIIIIIIIGTGAC TGGAGTTCAGACGTGTGCTCT/3InvdT /
P5	10 uM	25 nmole	Desalt	AATGATACGGCGACCACCGAGATCTACAC
P7	10 uM	25 nmole	Desalt	CAAGCAGAAGACGGCATACGAGAT

The first solution was placed in a thermal cycler and incubated for 5 min at 95°C. The second solution was then placed in the thermal cycler and both solutions were incubated for 5 min at 55°C, after which 18 µl of the second (baits) solution was added to the first (library) solution. The solution was gently vortexed to mix. 10 µl of mineral oil was added on top of the reaction to prevent evaporation, and briefly supon in a microcentrifuge to collect the liquid. The solution was incubated in a thermal cycler with a heated lid for 40 hours at 55°C. The remainder of the MYbaits protocol v.3 (*MYcroarray*) was performed as per the manufacturer's instructions between steps 2A.1 and 3.1. Captured libraries were denatured from the baits in 30 µl of 10 mM Tris-HCL,0.05% Tween-20 at 95°C for 5 min.

The libraries were amplified in six replicate 25-µl qPCR reactions with Illumina P5/P7 reamplification primers (Table S2). The PCR reaction contained reagents in final concentrations of: 0.4 mg/ml BSA, 1X PCR buffer, 2.5 mM MgCl2, 0.25 mM dNTPs, 0.05 U/ul Amplitaq Gold DNA polymerase, 0.25X SYBR Green, 0.4 µM *IDT* forward primer (1 µl of a 10 µM stock); 0.4 µM *IDT* reverse primer (1 µl of a 10 µM stock); 5 µl of neat library, made up to a total of 25 µl final volume with HPLC-grade water (10.9 µl). In a separate, designated post-PCR laboratory, thermal cycling was performed under the following conditions: 95°C for 10 min, 16 cycles of 95°C for 30 sec, 60°C for 30 sec, 72°C for 30 sec. Replicate reactions were combined and vortexed to mix. Libraries were concentrated and purified using a *QIAGEN* MinElute PCR purification kit, following the manufacturer's instructions and eluting in 32 µl EB buffer. Fragments between 155-400 bp were selected from 30 µl of library using a PippinPrep Prep (*Sage Science*) ethidium bromide eGel cassette (2%), following the manufacturer's instructions. The size-selected libraries were purified again using a *QIAGEN* MinElute PCR purification kit eluting in 32 µl EB buffer (note that after the addition of PB buffer, 3 µl of 3 M sodium acetate was added to adjust the pH). 2 µl was quantitated on the Qiaxpert spectrophotometer following the manufacturer's instructions. 10 ul was run on a Qiaxcel fragment analyser with the OL500 program (alignment marker 15 bp-3 kb, marker 50 bp-800 bp) to gauge the library fragment length. Libraries were combined equimolarly, concentrated using a Vivaspin 500 MWCO 30,000, and quantitated using a Qubit fluorometer following the manufacturer's instructions. The final sequencing library was

diluted to 2 nM in EB buffer and sequenced on a MiSeq standard flowcell paired-end 300-cycle V2 using custom sequencing primer CL72 (8).

SI III.4 Bioinformatics

For analysis of the 12S rDNA mini-barcode, the methods described in SI 2.0 of Grealy et al. (6) and SI 1.5 of Grealy et al. (7) were followed (i.e., adapters were trimmed in Geneious Prime ((9); *Biomatters*) and quality control was conducted using Usearch v.8 and v.9 (Edgar 2010) by: (1) filtering by quality (fastq_filter, maxee 0.5), (2) dereplicating (derep_fulllength), (3) denoising (unoise), and (4) abundance filtering). Filtered, unique sequences were aligned to NCBI's GenBank reference database (10) using the blastn algorithm with the default parameters.

For analysis of the captured libraries, the methods described in SI 7.0 of Grealy et al. (6) were followed (i.e., reads were quality filtered using Usearch v.8 (11) by: (1) filtering by quality (fastq_filter, maxee_rate 0.01), (2) dereplicating (derep_fulllength), and (3) chimera filtering (uchime_denovo). Reads from the extraction control and no-template control libraries were combined. Reads from the sample libraries were mapped to the control libraries in Geneious Prime 2020.2.4 (Geneious mapper, Low Sensitivity/Fastest with no iterations) and unmapped reads were used for downstream analysis (i.e., any reads present in the controls were removed from the sample libraries). The three sample libraries and combined control libraries were then individually mapped to each of the following reference mitochondrial genomes in Geneious with two stringency settings (Low Sensitivity/Fastest with no iterations and Medium/low Sensitivity/Fast with 10 iterations): NC_007227 (*Alectura lathami*), KJ637997 (*Anas platyrhynchos*), KJ778617 (*Gallus gallus*), NC002785 (*Struthio camelus*), and NC031869 (*Caloenas nicobarica*). Mapped reads were aligned to NCBI's GenBank reference database (10) using the blastn algorithm (12):

```
blastn -task blastn -db /data/bioref/blast/ncbi/nt -query ${input_folder}/${input_file} -out ${output_folder}/${output_file} -perc_identity 80 -qcov_hsp_perc 80 -max_target_seqs 5 -max_hsps 5
```

Blast files were imported into MEGAN 6 (13) with default LCA parameters, and all reads assigned to class Aves were extracted. These reads were then combined by sample, deduplicated, and mapped to a 75% consensus mitochondrial avian genome generated from an alignment of the above bird species (High sensitivity / Medium, 25 iterations), and the 50% consensus mitochondrial genome was extracted. Where there was no coverage, a ? was called. The minimum accepted coverage per base was 1X as because so few reads mapped. Mapping was also repeated excluding reads that were assigned to *Gallus gallus*, assuming them to be contamination.

Consensus sequences were then separately aligned with 40 other avian mitochondrial genomes (see Fig. S11), using *Struthio camelus* (ostrich) as the outgroup, using *MAFFT v7.450* (14) with default parameters implemented in Geneious Prime. Alignments were refined using *MUSCLE 3.8.425* (15) default parameters implemented in Geneious Prime. Note that all three samples could not be combined into one phylogeny because the same loci were not recovered by each sample, causing a combined analysis to have too much missing data.

All protein-coding genes (ATP6, ATP8, CO1, CO2, CO3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, and ND6) and RNA genes (rRNAs and tRNAs) were individually extracted from the alignment using the 'Extract' function in Geneious v.2020.2.4. Protein coding genes were

translated to ensure the ORFs were in-frame. Each protein-coding gene was partitioned by codon position in Geneious using the 'Tools/Mask alignment/Save a copy with sites stripped/Codon position 1 and 2' (will generate a file containing codon position 3). Codon positions for each gene were concatenated in Geneious using the 'Tools/Concatenate sequences or alignments' function, such that three files were generated, one containing the 1st codon positions for all protein-coding genes (m1), one containing the 2nd codon positions for all protein-coding genes (m2), and one containing the 3rd codon positions for all protein-coding genes (m3). rRNA and tRNA genes were partitioned into stems and loops as per Grealy et al. (39) which were concatenated and treated as separate partitions for a total of five partitions. Any ambiguous sites and gaps were stripped from each alignment.

To find the best substitution model for each partition, ModelFinder ((16) in iqtree 1.6.12 (17) was used:

```
 > iqtree -s  m1.phy -nt 20 -m TESTONLY
```

The best scoring models as determined through the Bayesian Information Criterion (BIC) score (see Table S6). Models were then specified per partition for each phylogeny in a .nex parition file, e.g.:

```
#nexus

begin sets;
        charset m1 = 1-689;
        charset m2 = 690-1381;
        charset m3 = 1382-2058;
        charset stems = 2059-2498;
        charset loops = 2499-3031;
        charpartition mine = SYM+I+G4:m1, TIM3+F+I+G4:m2, TIM3+F+I+G4:m3, TIM2e+I+G4:stems,
TIM+F+I+G4:loops;

end;
```

A maximum-likelihood tree (ML) was generated in iqtree 1.6.12 ((17, 18):

```
> iqtree -s  m1.phy -p part.nex -o NC_002785_Struthio_camelus_Palaeognathae -nt 20 -b 500 -m
TEST
```

Finally, used reads were re-mapped onto the consensus sequence from the previous iteration (High sensitivity / Medium, 25 iterations) in Geneious, and the contig file and reference was exported in .sam format. MapDamage 2.2.1 (19) was run:

```
> mapDamage –i INPUT.sam –r REFERENCE.fasta
```

Nucleotide misincorporation plots were examined for a higher proportion of C to T misincorporations at the 5' and 3' terminus of reads that are indicative of authentic ancient DNA (Fig. S12).

SI IV Ancient DNA supplementary results

SI IV.1 12S rDNA mini-barcode

The *12S* rDNA mini-barcode 12SAC was only able to be amplified and sequenced from one extract. None of the extraction controls or PCR negative controls amplified. 13,307 raw reads were sequenced from MB3542 (AD2090). After quality filtering, 13,298 reads remained, and after deduplication, 128 unique reads remained. After denoising, one unique read with 13,068 copies:

5'-GCCTAGCCCTAAATCTAGATACCTCCCATCACACATGTATCCGCCTGAGAACTA-3'

This sequence matched with 100% identity across its entire length to *Gallus gallus* isolate 11053 mitochondrion, complete genome (GenBank accession # MT800504.1), with an e-value of 5e-18.

Although seven extracts were prepared from AD2090, MB3542 was the only one where the barcode could be amplified. As such, and because the identity to *Gallus gallus* is 100%, it is likely that this sequence has arisen from sporadic contamination; even if the closest relative of the UO was domestic chicken, we would not expect the sequence to be identical at this locus.

SI IV.2 Captured libraries

All reads mapped from the controls were human in origin.

Medium stringency mapping did not typically result in any more avian reads being mapped, just more contamination (see Figure S7).

Choice of reference does make a small difference in taxonomic profile of mapped reads; for instance, more reads assigned to *Gallus gallus* map when the *Gallus gallus* reference is used; more reads assigned to *Anas platyrhynchos* map when the *Anas platyrhynchos* reference is used. But, we do not find any reads assigned to *Alectura*, *Struthio* or *Caloenas* no matter which reference is used. All reads assigned to species within Galloanserae (*Gallus gallus, Anas platyrhynchos,* and *Meleagris gallopavo*) could be the result of contamination as these are all introduced, non-native species. Phylogenies reconstructed including including reads assigned to *Gallus* reads show all three samples placed sister to *Gallus gallus*, indicating that chicken is a likely contaminant: if the samples were truly Galliformes, we would expect them to fall within Galliformes but not immediately sister to chicken (Fig. S11 Ac, C, E). However, when reads assigned to *Gallus* are excluded from mapping, we find CSS2089 is placed within Galliformes but not sister to *Gallus* or Megapodidae (*Alectura*), while CSS2090 and CSS3622 are both placed within Anseriformes. If the Anseriform reads were likely to be contamination by domestic duck (or mapping bias), we would predict the samples to fall sister to *Anas platyrhynchos*, but they do not. A higher proportion of C to T nucleotide misincorporations at both the 5' and 3' ends of mapped reads was not observed (Fig. S12); however, the coverage was so low that this was expected. As such, we cannot be confident the mapped reads are truly ancient in origin. However, it is curious that the topology of the phylogenetic tree for two independent samples (CSS2090 and CSS3622) are very similar to that obtained from the protein analysis (Fig. S11 D,F): CSS3622 was sequenced for protein, but CS2090 was not. While Anseriformes is not

resolved as a monophyletic group, this is likely because the overall number of sites in the final alignment of these samples was too low for deep phylogenetic resolution. However, despite Anseriformes being paraphyletic, the UOs fall within the well-supported Anseriform clade (79-87% bootstrap support). As for CSS2089, it is placed at the base of Galliformes with 89% bootstrap support (Fig. S11 B). Because of this conflict, and because we cannot be sure that DNA is authentically ancient, the aDNA analysis is inconclusive. Deeper sequencing would be required to obtain higher coverage and determine whether any endogenous DNA survived at all, and given the age and preservation conditions, this is probably unlikely.

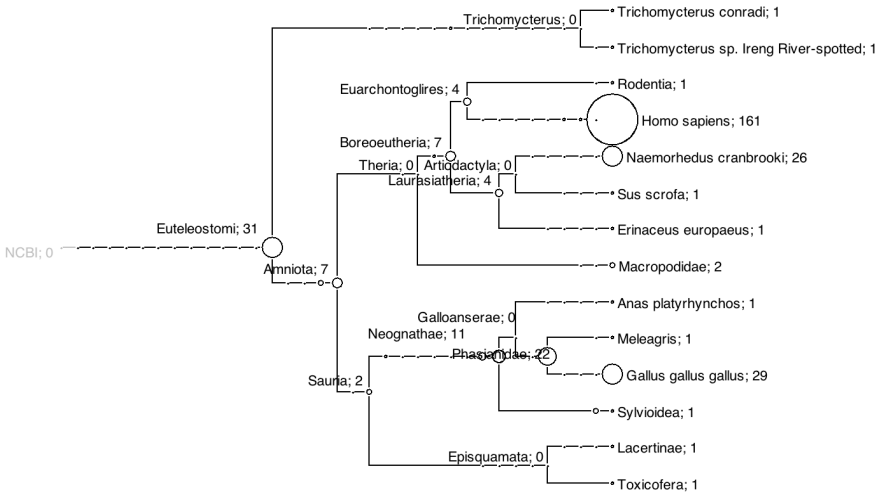
Table S3 Number of reads mapped to various mitochondrial references with the settings: Medium/Low sensitivity / Fast mapping with 10 iterations.

					Reads mapped				
Library	**# reads**	**Filtered # (-controls)**	**Reads Read length (bp)**	**St Dev**	***Gallus***	***Alectura***	***Anas***	***Struthio***	***Caloenas***
CSS2089	4979143	4941877	57.3	24.7	3815	6469	10205	9511	9096
CSS2090	1108872	1096709	69	38.2	2593	4234	6185	5432	5590
CSS3622	2905839	2889718	61.6	31.6	1181	1545	3898	1078	1349
Controls	98353		56.2	23.6	3	1	1	0	2

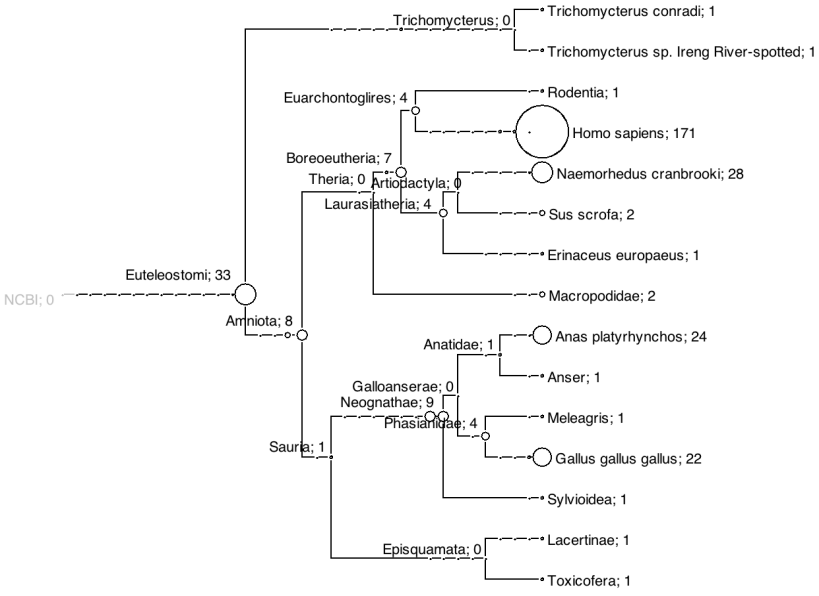
Table S4 Number of reads mapped to various mitochondrial references with the settings: Low sensitivity / Fastest mapping with no iterations.

					Reads mapped				
Library	**# reads**	**Filtered # (-controls)**	**Reads Read length (bp)**	**St Dev**	***Gallus***	***Alectura***	***Anas***	***Struthio***	***Caloenas***
CSS2089	4979143	4941877	57.3	24.7	728	331	344	340	291
CSS2090	1108872	1096709	69	38.2	2470	117	142	5596	5008
CSS3622	2905839	2889718	61.6	31.6	192	44	59	53	39
Controls	98353		56.2	23.6	3	1	1	0	2

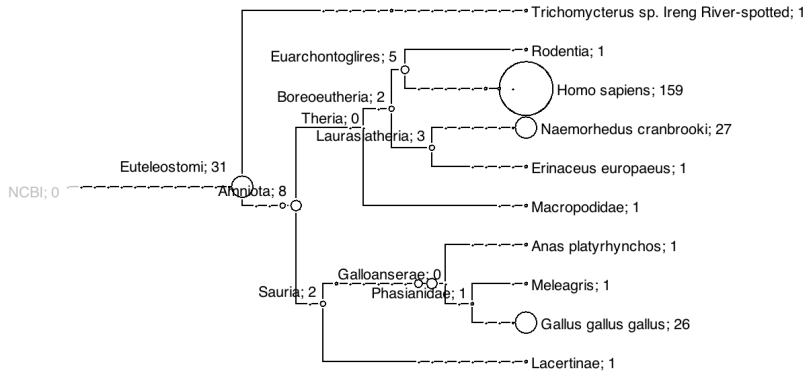
A CSS2089 mapped to *Alectura* with low stringency



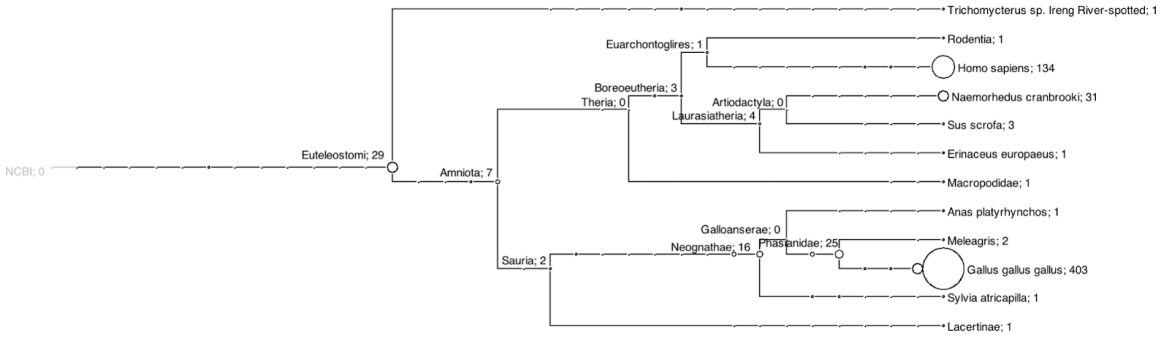
B CSS2089 mapped to *Anas* with low stringency



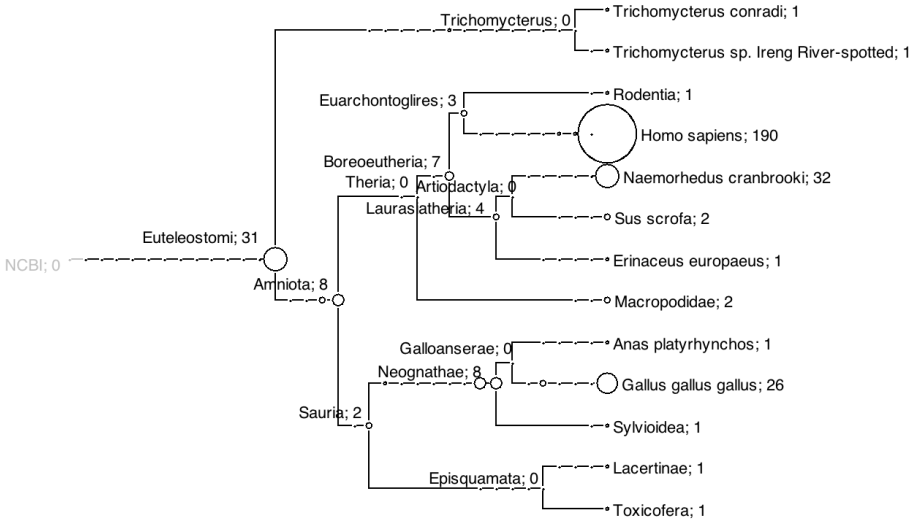
C CSS2089 mapped to *Caloenas* with low stringency



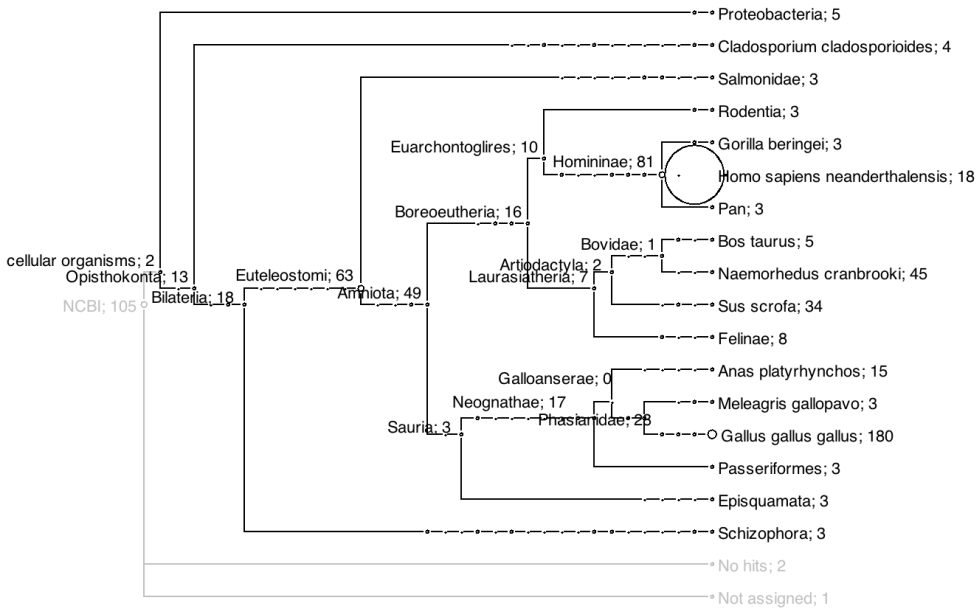
D CSS2089 mapped to *Gallus* with low stringency



E CSS2089 mapped to *Struthio* with low stringency

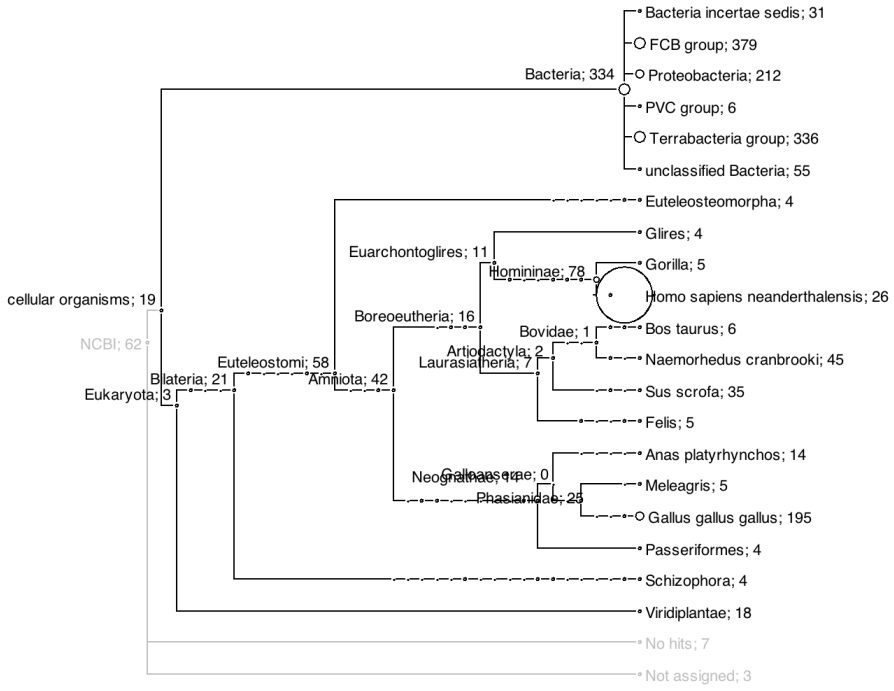


F CSS2089 mapped to *Alectura* with medium stringency

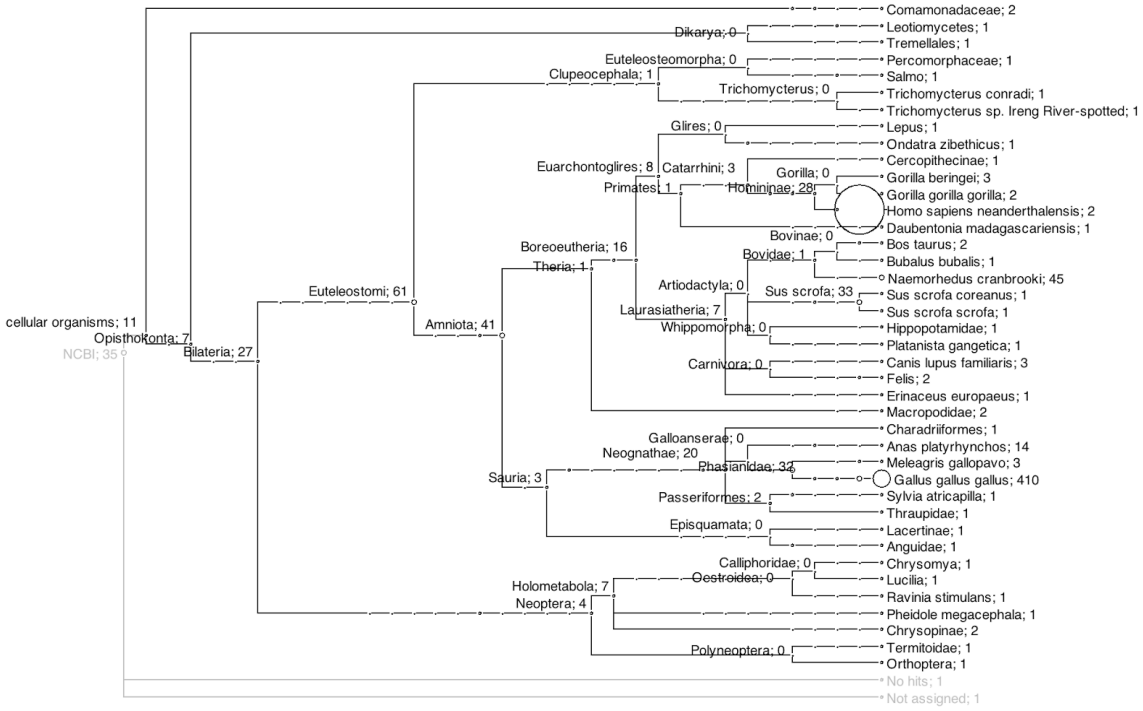


G CSS2089 mapped to *Anas* with medium stringency

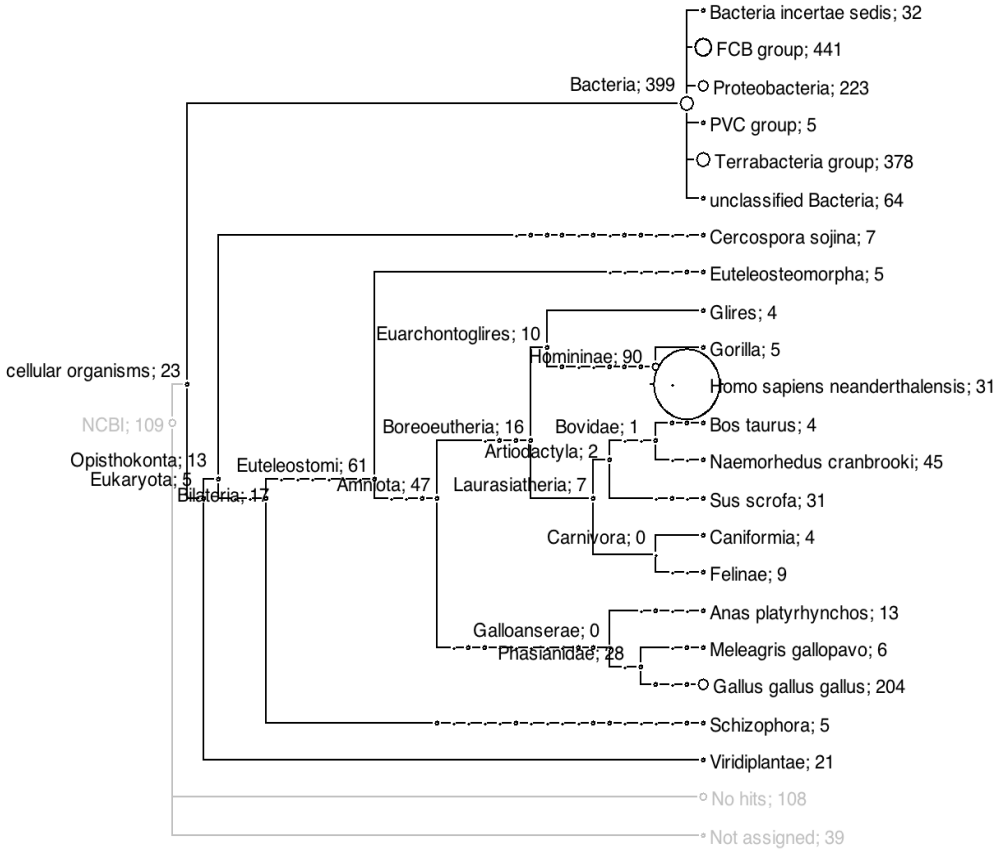
H CSS2089 mapped to *Caloenas* with medium stringency



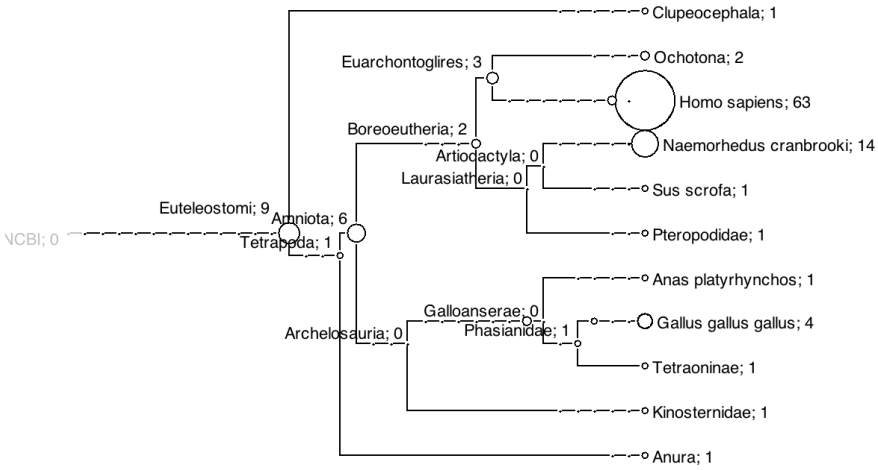
I CSS2089 mapped to *Gallus* with medium stringency



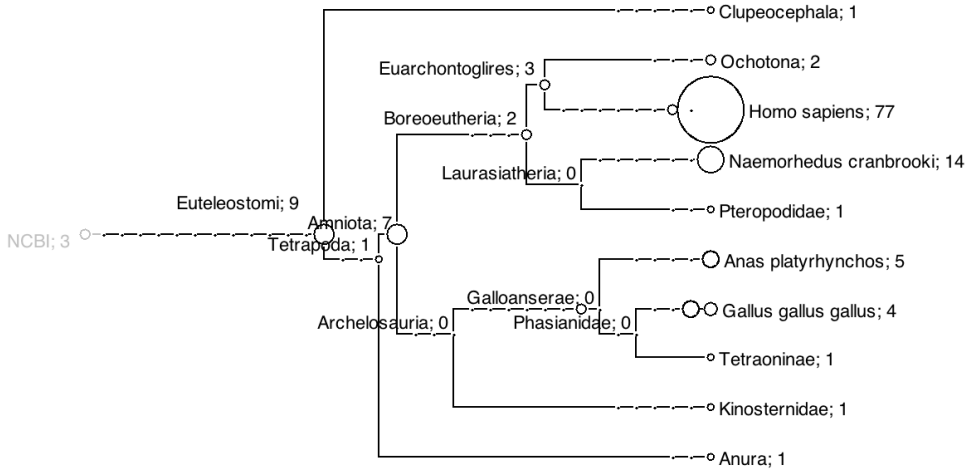
J CSS2089 mapped to *Struthio* with medium stringency



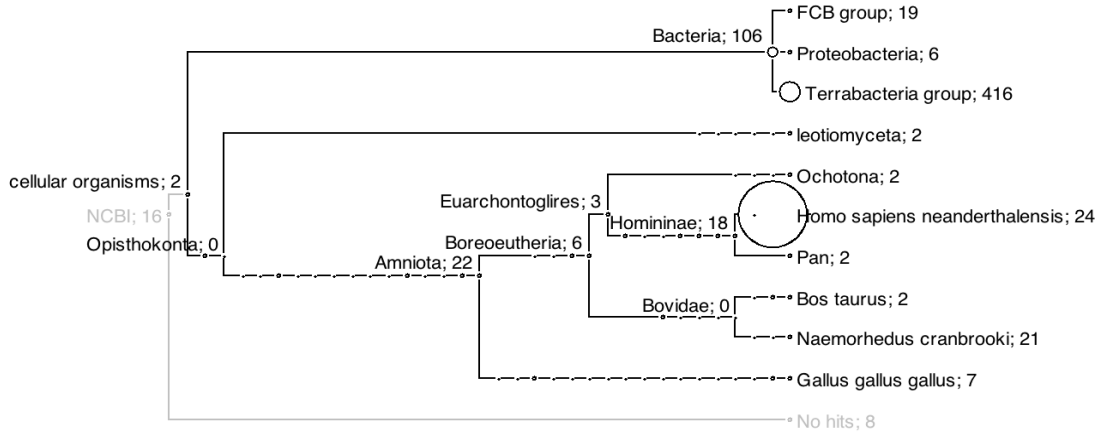
K CSS2090 mapped to *Alectura* with low stringency



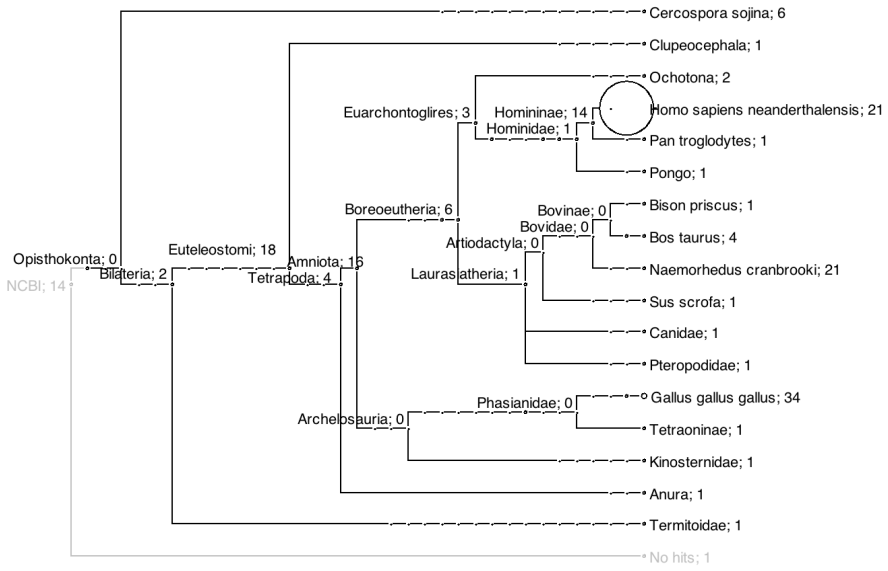
L CSS2090 mapped to *Anas* with low stringency



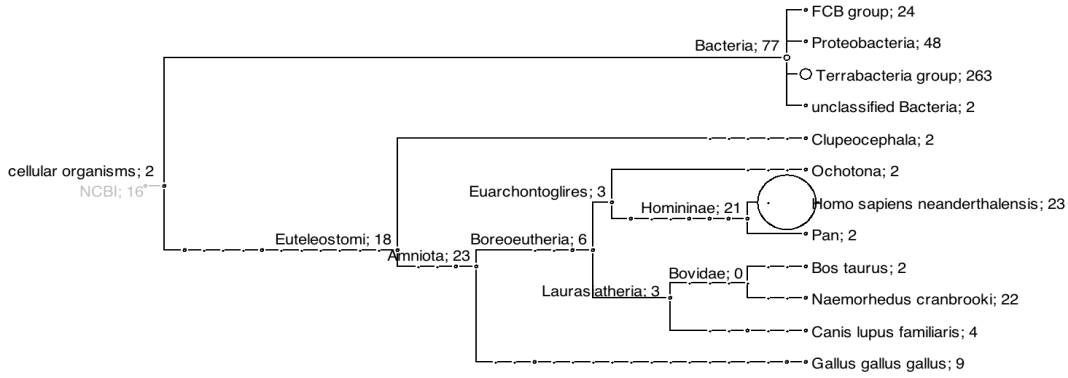
M CSS2090 mapped to *Caloenas* with low stringency



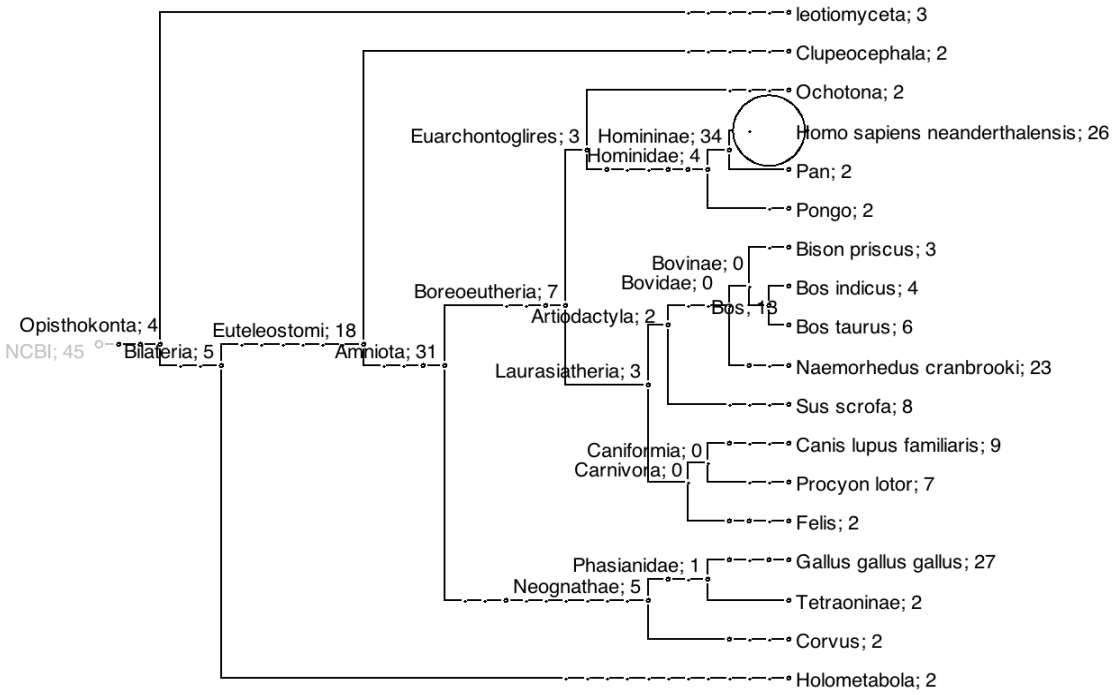
N CSS2090 mapped to *Gallus* with low stringency



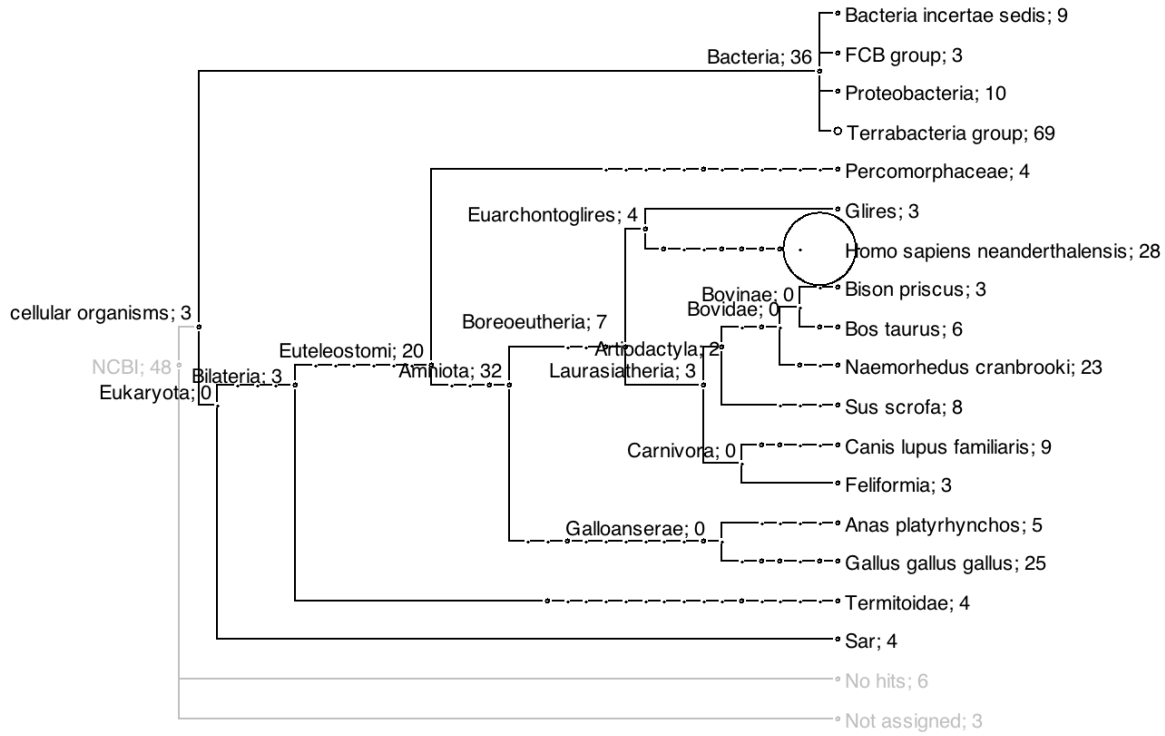
O CSS2090 mapped to *Struthio* with low stringency



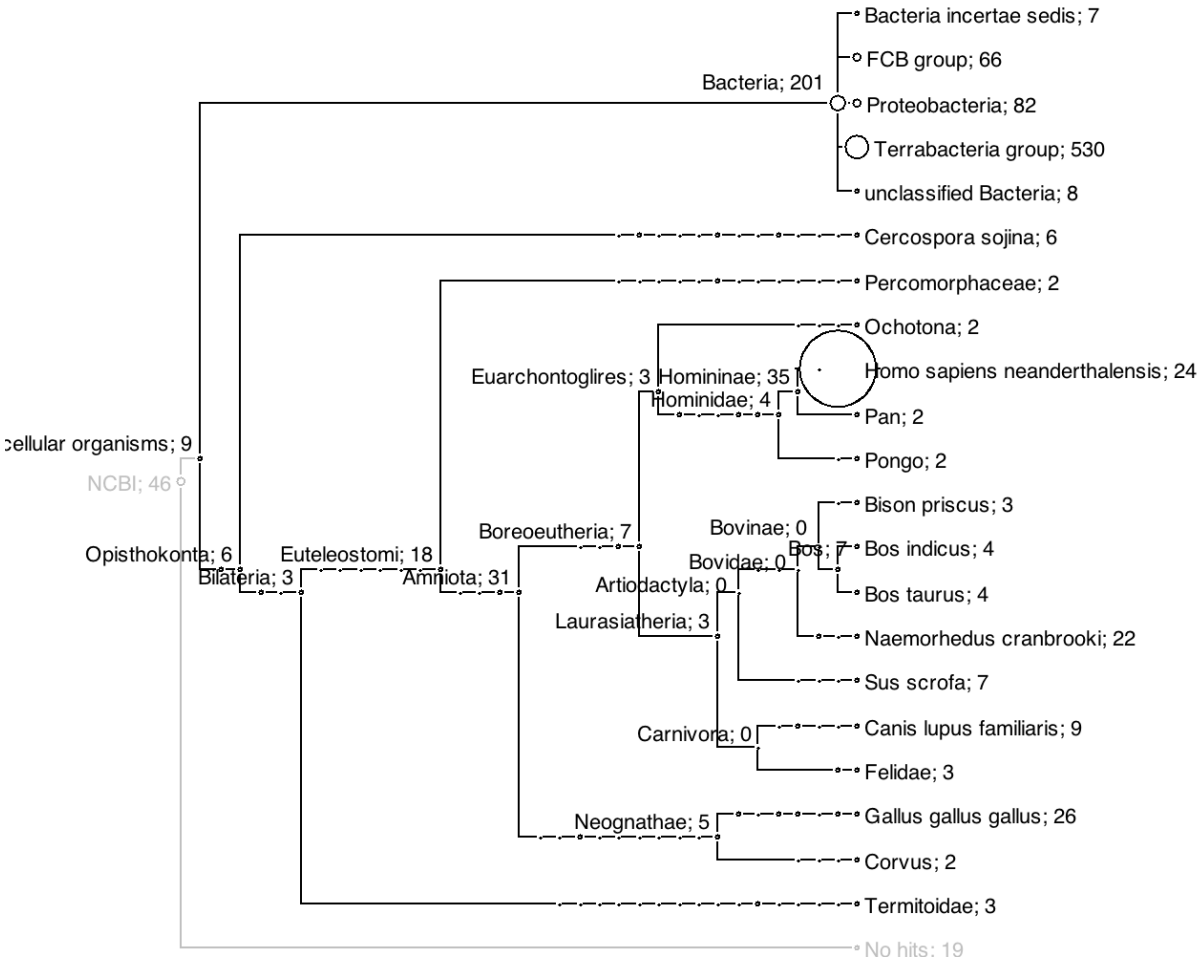
P CSS2090 mapped to *Alectura* with medium stringency



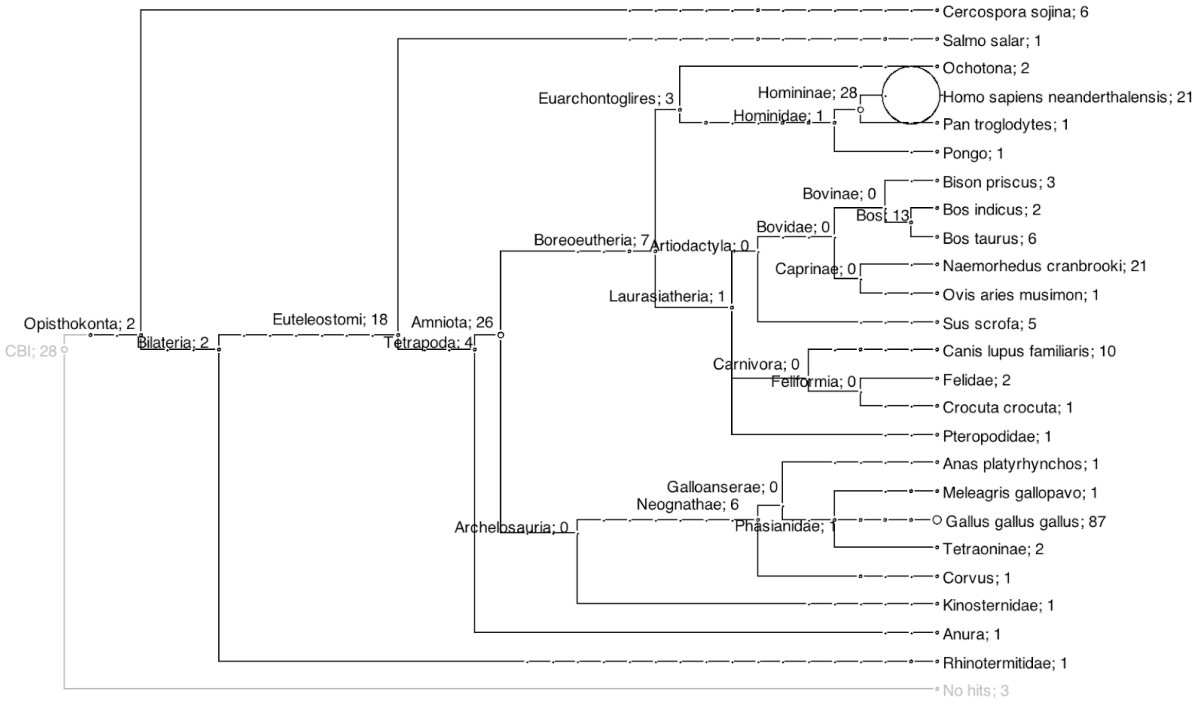
Q CSS2090 mapped to *Anas* with medium stringency



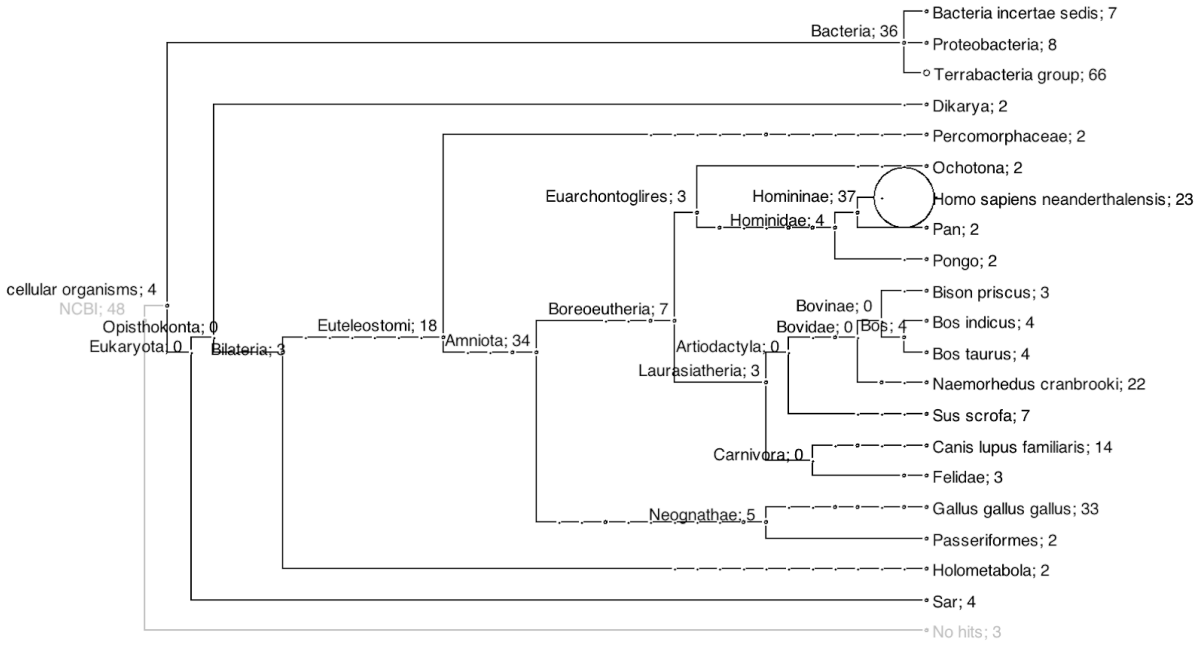
R CSS2090 mapped to *Caloenas* with medium stringency



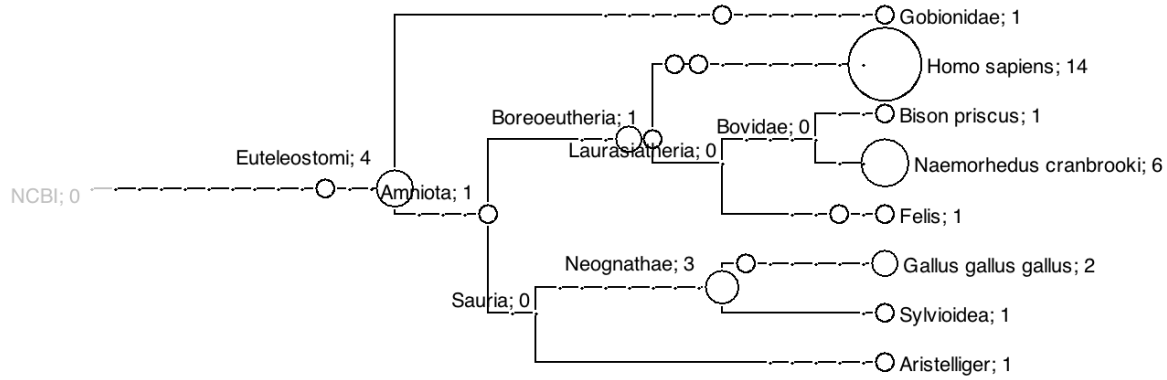
S CSS2090 mapped to *Gallus* with medium stringency



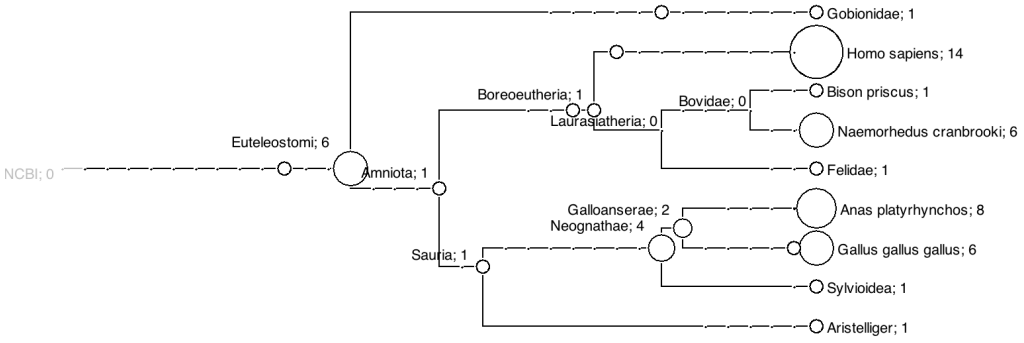
T CSS2090 mapped to *Struthio* with medium stringency



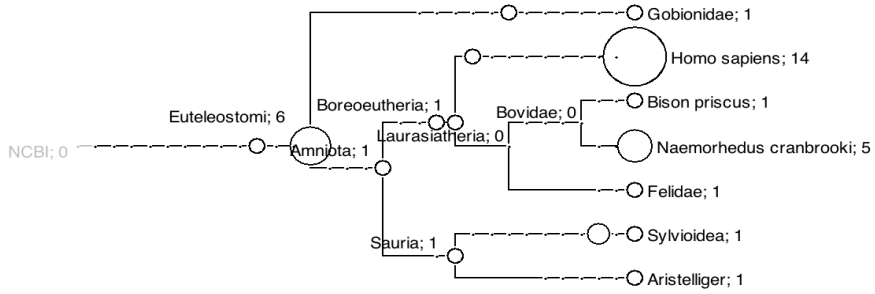
U CSS3622 mapped to *Alectura* with low stringency



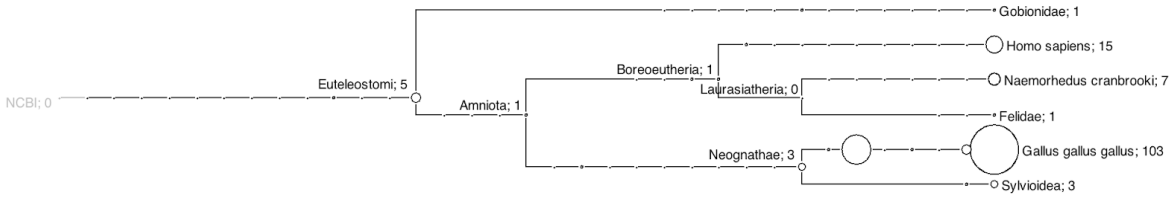
V CSS3622 mapped to *Anas* with low stringency



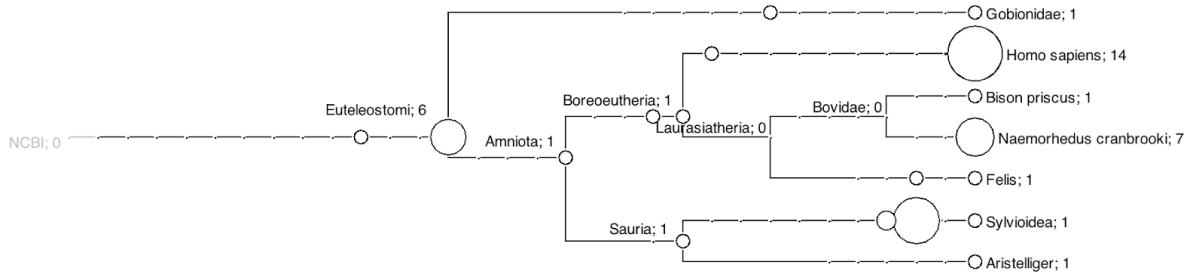
W CSS3622 mapped to *Caloenas* with low stringency



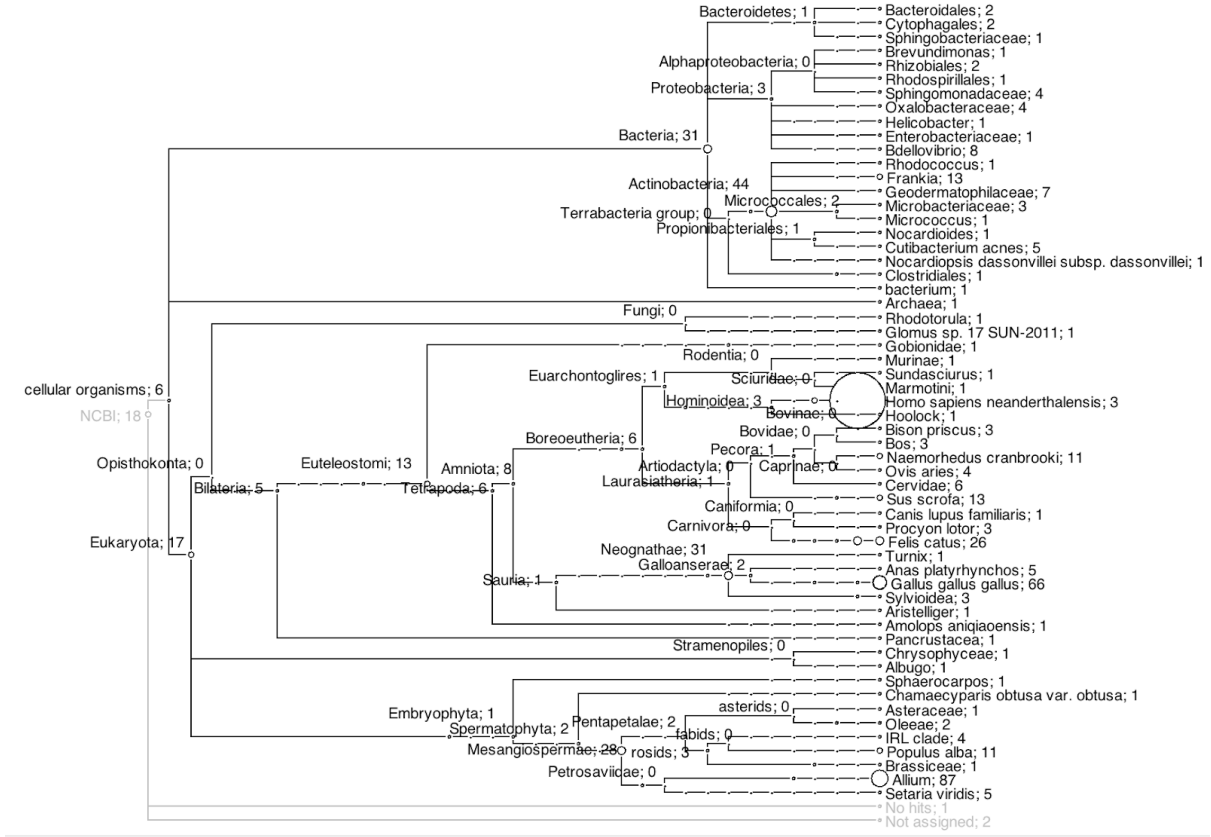
X CSS3622 mapped to *Gallus* with low stringency



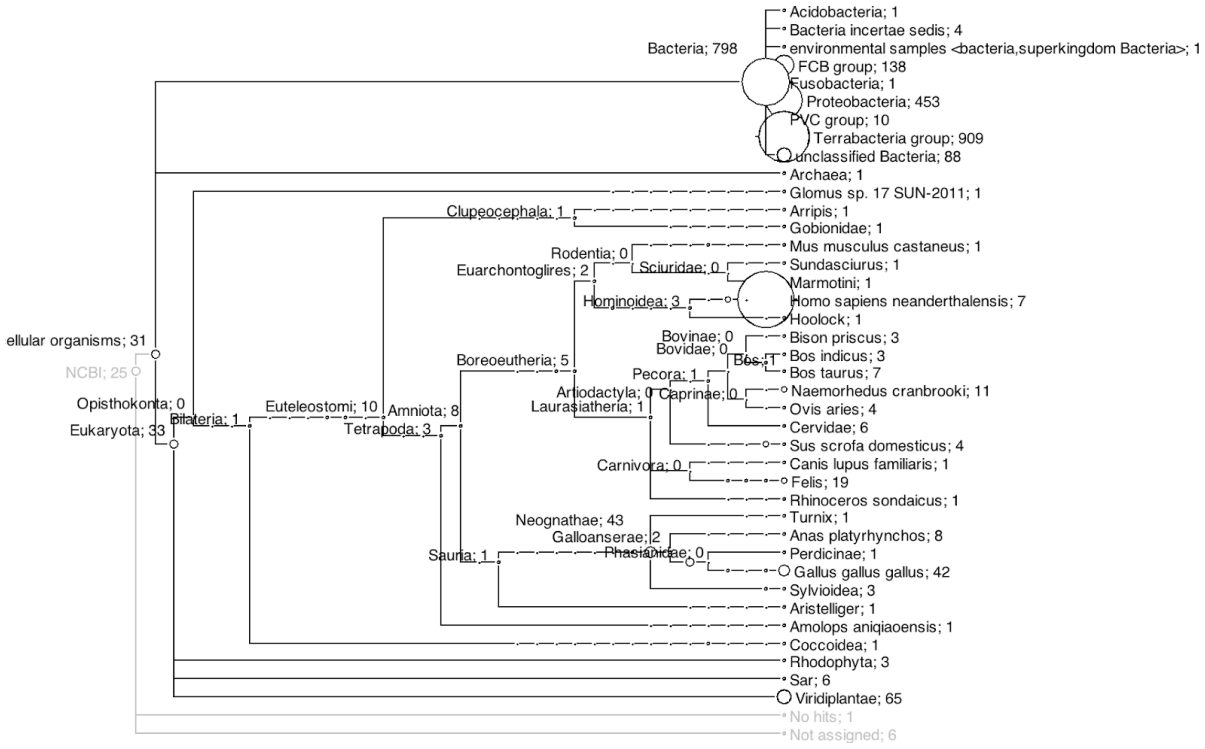
Y CSS3622 mapped to *Struthio* with low stringency



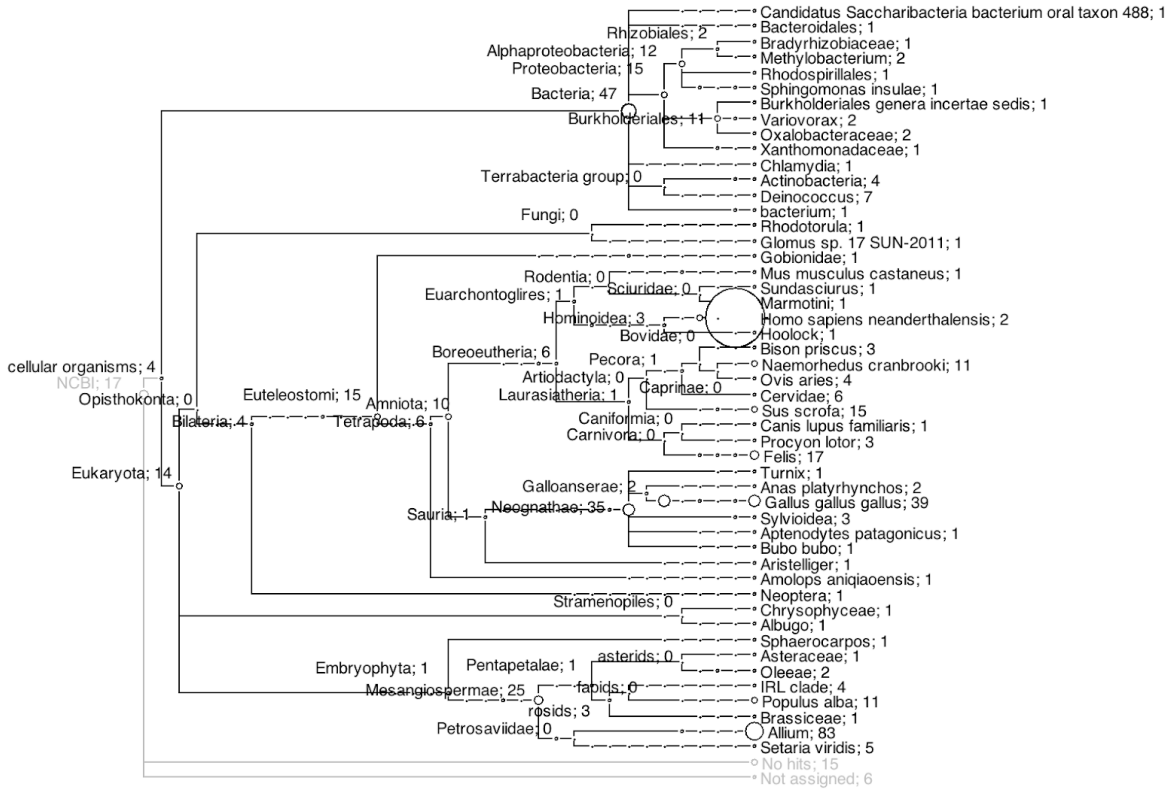
Z CSS3622 mapped to *Alectura* with medium stringency



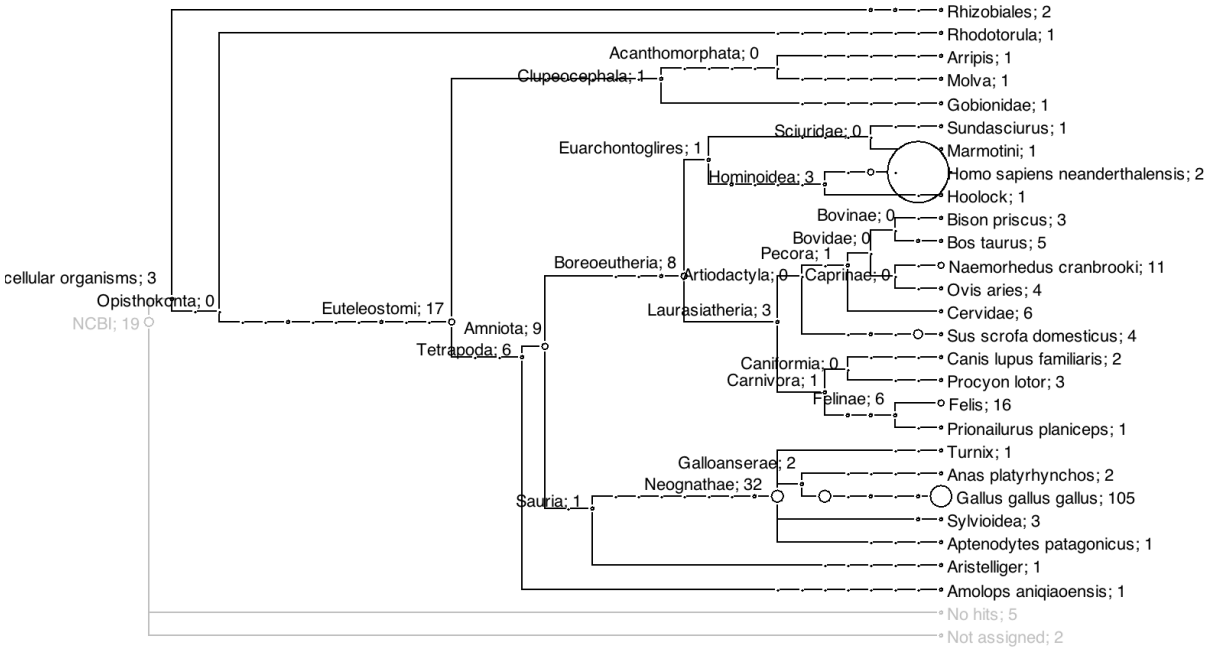
AA CSS3622 mapped to *Anas* with medium stringency



BB CSS3622 mapped to *Caloenas* with medium stringency



CC CSS3622 mapped to *Gallus* with medium stringency



DD CSS3622 mapped to *Struthio* with medium stringency

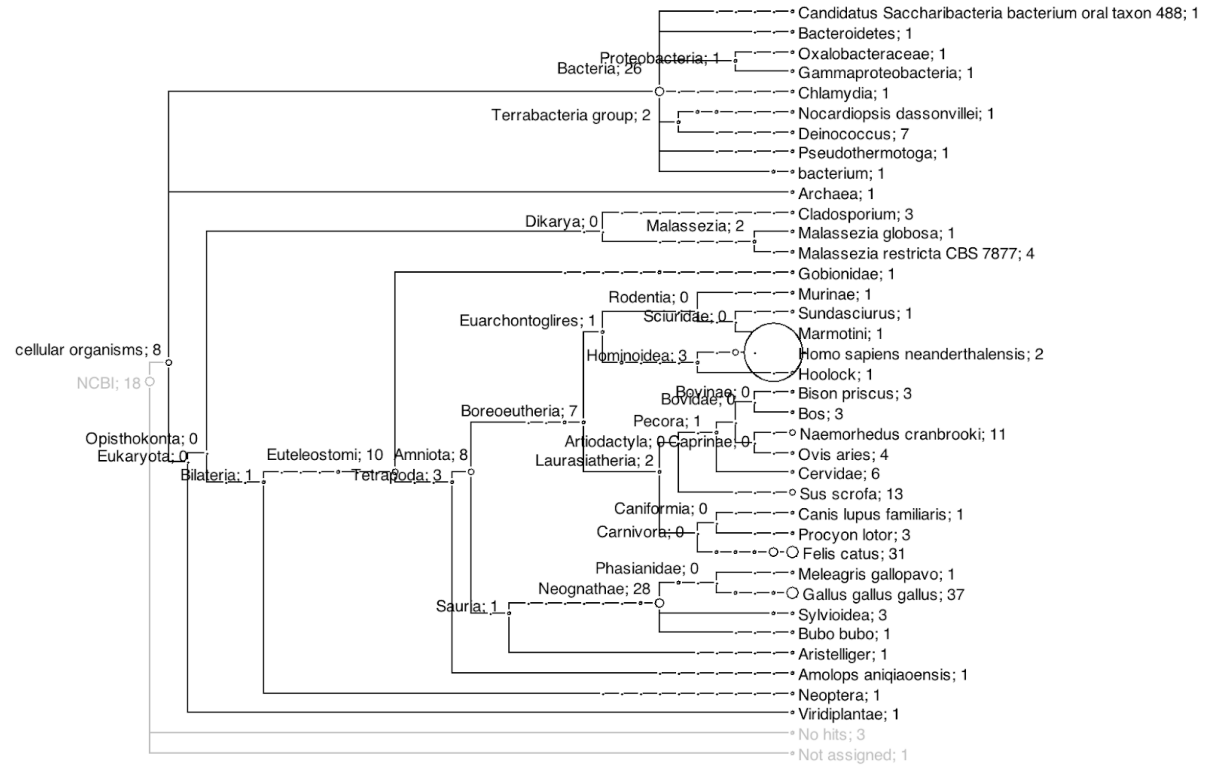


Fig. S10 A-DD: Taxonomic profiles from mapped reads (blast files imported to MEGAN).

Table S5 Number of avian reads including and excluding reads assigned to *Gallus*. This is the total number of unique reads from combining the results from the various analyses (see Fig. S11).

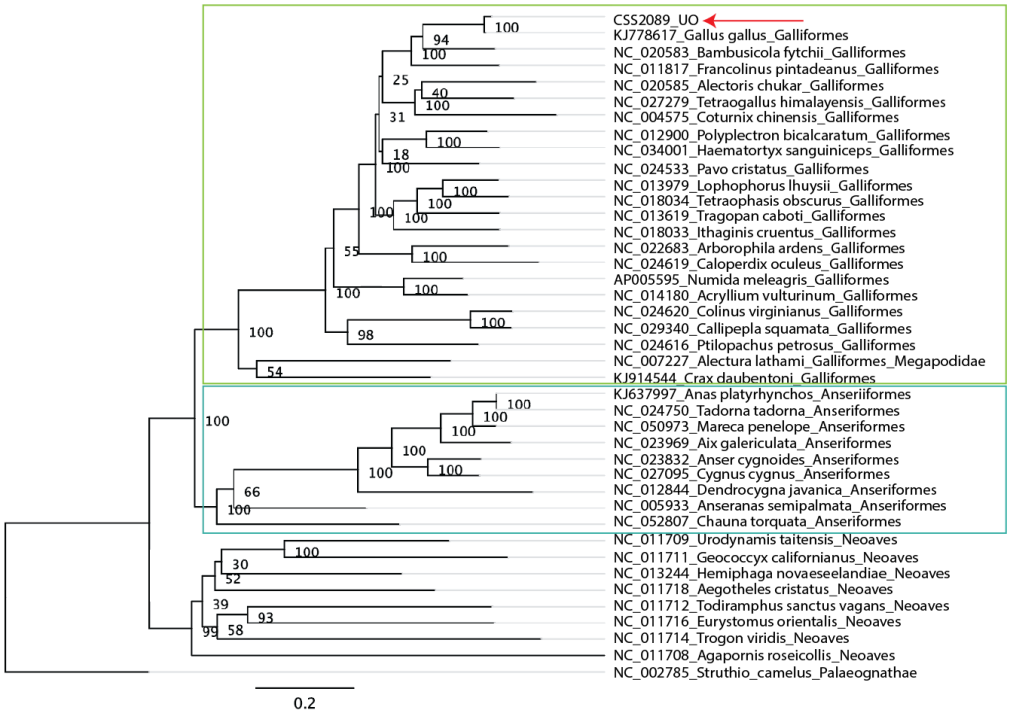
Library	# Avian reads mapped	Average coverage	# Avian reads mapped (excluding *Gallus*)	Average Coverage	Total bp avian (bp) in alignment	Total bp avian (excluding *Gallus*) in alignment
CSS2089	567	2.2X	124	0.5X	11,197	3,031
CSS2090	279	1.1X	127	0.5X	6,130	1,725
CSS3622	228	0.9X	101	0.4X	3,637	1,265

Table S6 Substitution models used for each partition in each ML analysis performed.

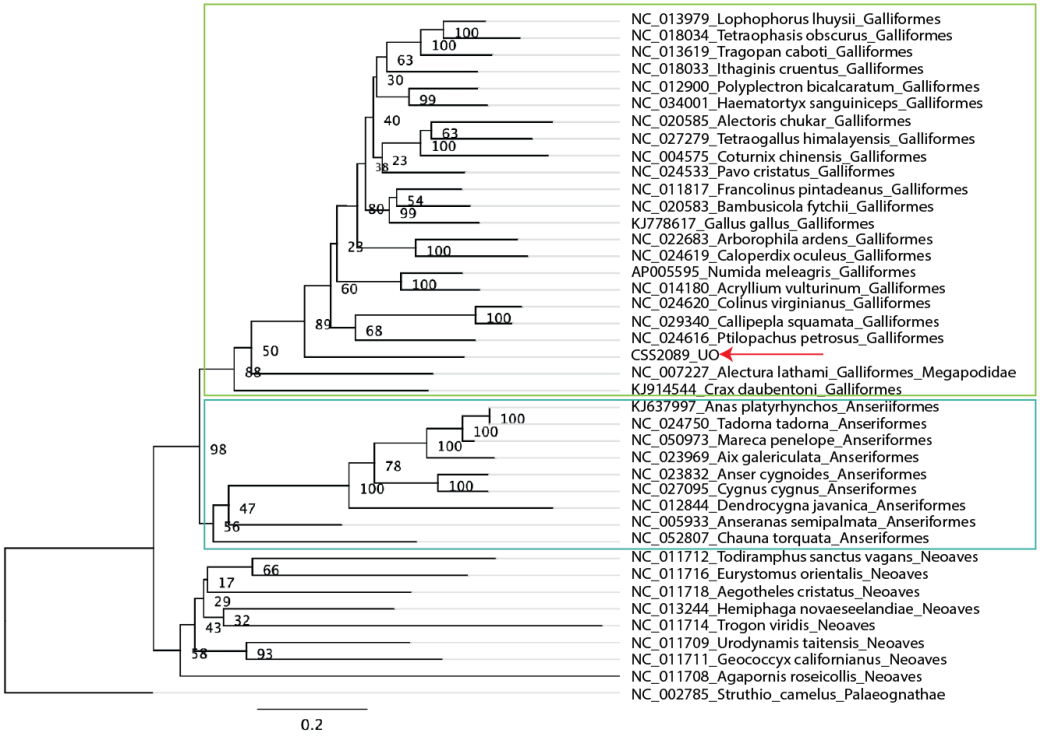
		Partition				
Library	**Alignment**	**m1**	**m2**	**m3**	**stems**	**loops**
CSS2089	Avian reads	GTR+F+I+G4	TVM+F+I+G4	GTR+F+I+G4	TIM2e+I+G4	GTR+F+I+G4
	Avian reads (excluding *Gallus*)	SYM+I+G4	TIM3+F+I+G4	TIM3+F+I+G4	TIM2e+I+G4	TIM+F+I+G4
CSS2090	Avian reads	GTR+F+I+G4	TPM3+F+I+G4	TVM+F+I+G4	TIM2e+I+G4	GTR+F+I+G4
	Avian reads (excluding *Gallus*)	TVMe+G4	TIM3+F+I+G4	TIM3+F+I+G4	TIM2e+I+G4	TVM+F+I+G4
CSS3622	Avian reads	GTR+F+I+G4	TPM3+F+I+G4	TVM+F+I+G4	TIM2e+I+G4	TVM+F+I+G4

Avian reads (excluding *Gallus*)	TIM2e+I+G4	HKY+F+I+G4	TPM3+F+I+G4	TIM2e+I+G4	TPM3u+F+I+G4

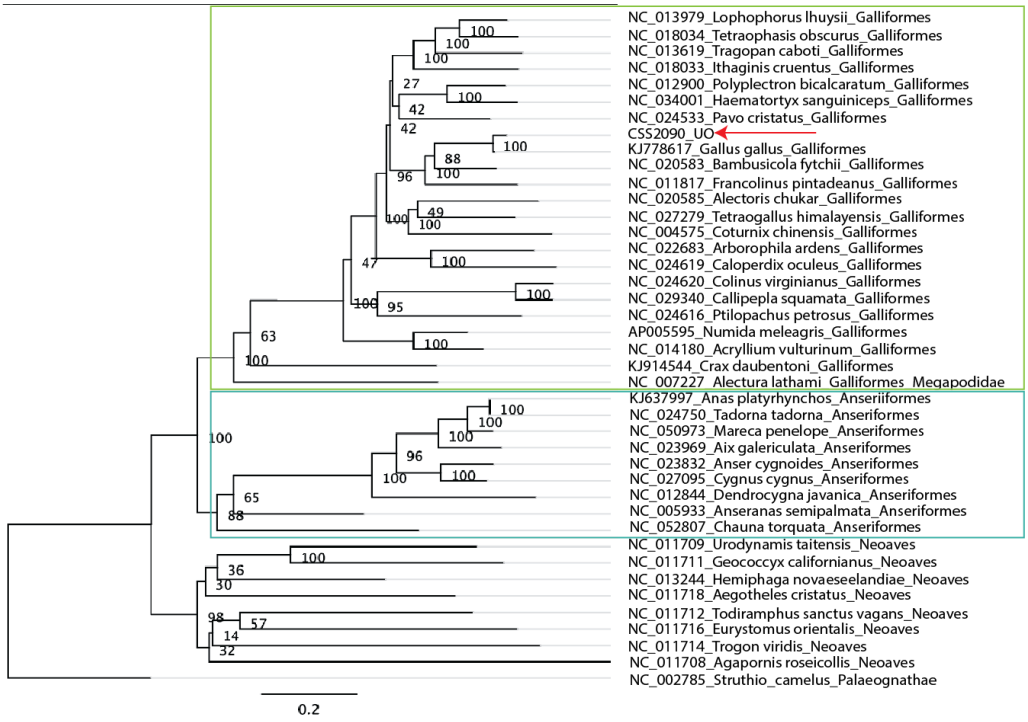
A ML tree CSS2089 (avian reads)



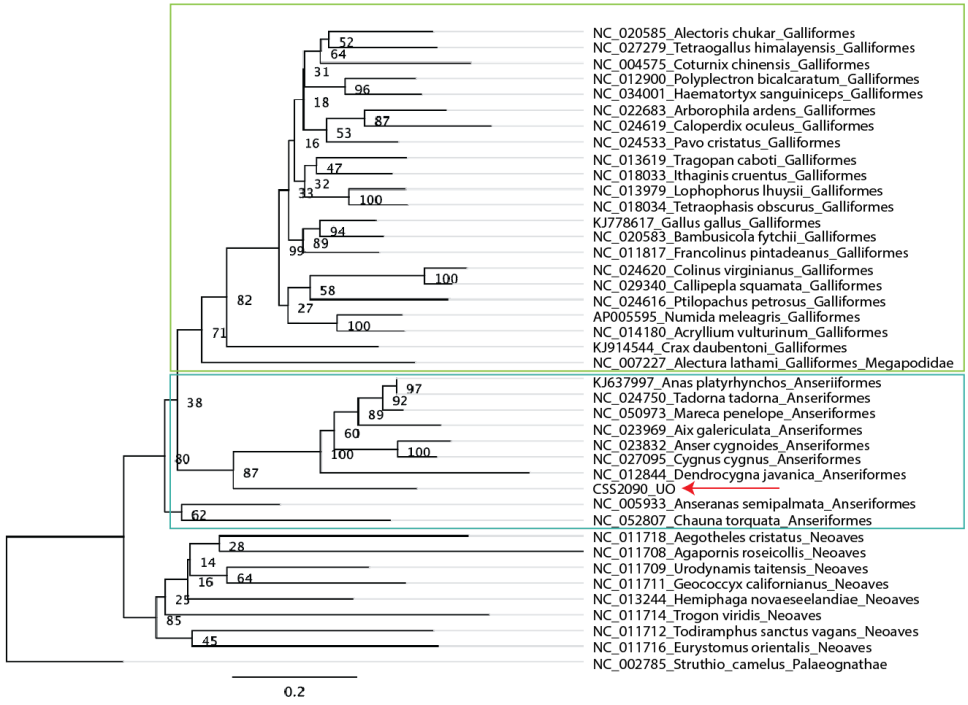
B ML tree CSS2089 (avian reads excluding *Gallus*)



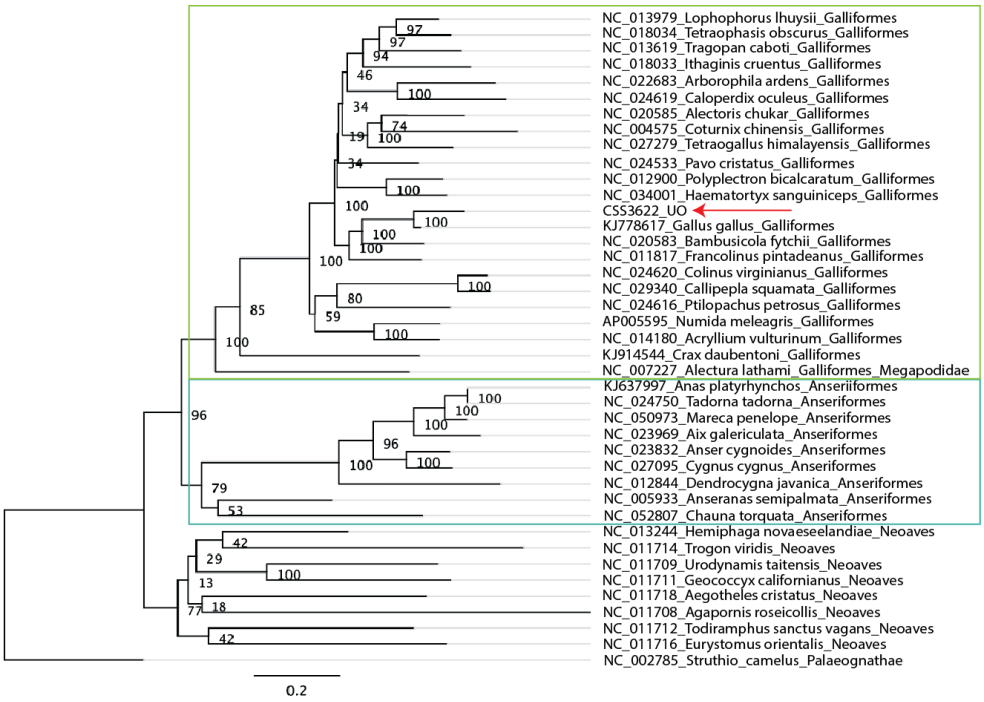
C ML tree CSS2090 (avian reads)



D ML tree CSS2090 (avian reads excluding *Gallus*)



E ML tree CSS3622 (avian reads)



F ML tree CSS3622 (avian reads excluding *Gallus*)

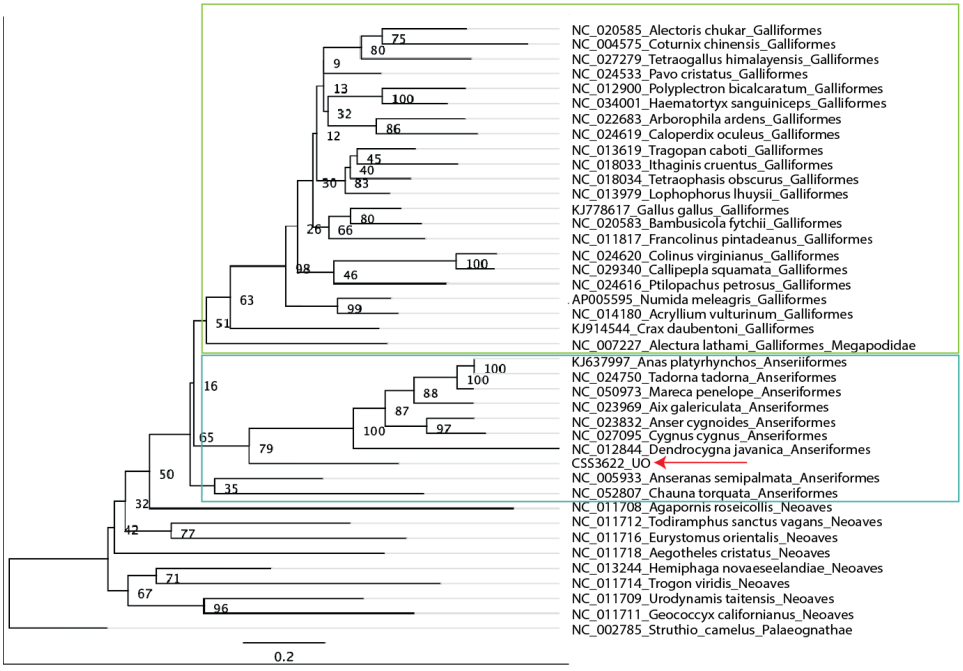
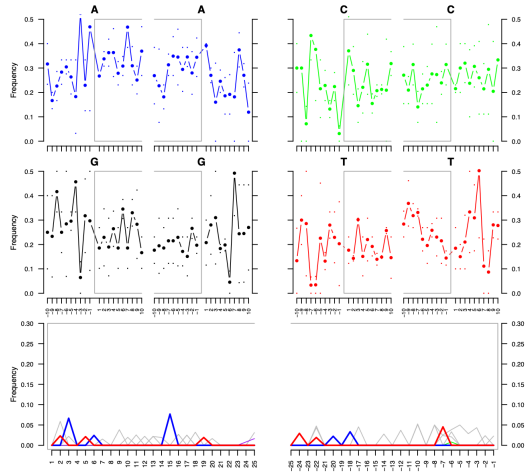
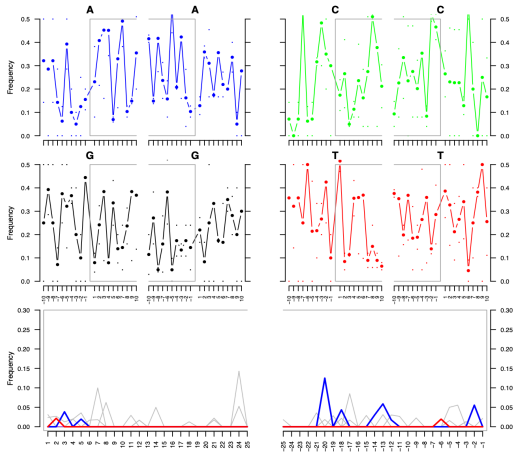


Fig. S11 A-F. Maximum likelihood phylogenetic trees (500 bootstraps) generated from alignments for each sample that either included or excluded *Gallus* reads. Bootstrap support % is indicated above the nodes.

A


A
A
C
C
G
G
T
T
Frequency

B


A
A
C
C
G
G
T
T
Frequency

C

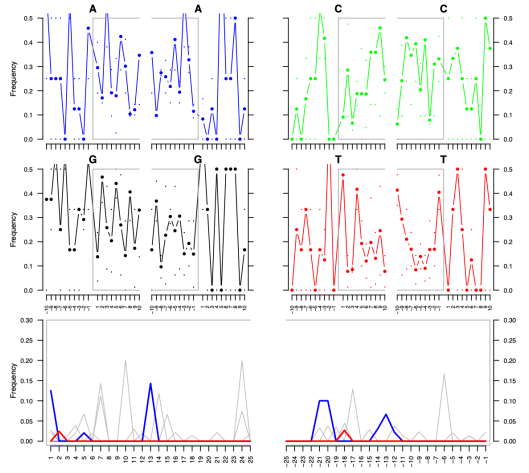

A
A
C
C
G
G
T
T
Frequency

Fig. S12 Damage profiles of reads used to reconstruct partial mitochondrial genomes from **A** CSS2089, **B** CSS2090, and **C** CSS3622, excluding reads assigned to *Gallus*. A higher proportion of C to T nucleotide misincorporations (red) at both the 5' and 3' ends of mapped reads (bottom panel) was not observed. Figure generated by MapDamage 2.0 (19).

Legend for Dataset S1

Full proteomics dataset. This Excel file reports all the peptide and protein data for UO eggshell samples

SI References

1. M. Steinegger, S. Ovchinnikov, M. Mirdita, AlphaFold2 w/ MMseqs2. *Google Colaboratory, 2021* (July 30, 2021).

2. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

3. R. R. Ruiz-Arellano, F. J. Medrano, A. Moreno, A. Romero, Structure of struthiocalcin-1, an intramineral protein from *Struthio camelus* eggshell, in two crystal forms. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 809–818 (2015).

4. J. Dabney, *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15758–15763 (2013).

5. A. Grealy, *et al.*, Eggshell palaeogenomics: Palaeognath evolutionary history revealed through ancient nuclear and mitochondrial DNA from Madagascan elephant bird (*Aepyornis* sp.) eggshell. *Mol. Phylogenet. Evol.* **109**, 151–163 (2017).

6. A. Grealy, M. Bunce, C. E. Holleley, Avian mitochondrial genomes retrieved from museum eggshell. *Mol. Ecol. Resour.* **19**, 1052–1062 (2019).

7. A. Grealy, N. E. Langmore, L. Joseph, C. E. Holleley, Genetic barcoding of museum eggshell improves data integrity of avian biological collections. *Sci. Rep.* **11**, 1605 (2021).

8. M.-T. Gansauge, *et al.*, Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* **45**, e79 (2017).

9. M. Kearse, *et al.*, Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

10. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, GenBank. *Nucleic acids research* (2006).

11. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

12. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

13. D. H. Huson, A. F. Auch, J. Qi, S. C. Schuster, MEGAN analysis of metagenomic data. *Genome Res.* (2007).

14. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

15. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

16. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

17. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

18. O. Chernomor, A. von Haeseler, B. Q. Minh, Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).

19. H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).