Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An AI-assisted genomic census of 193,342 assemblies finds a GFP-barrel family in unreported lineages, carrying pseudogenes in both fungi and cnidarians

Adam Sidaway

adamsidaway@hotmail.co.uk

Cambridge RNA Technologies/Morphokinetics

Research Article

Keywords: GFP, Barrel Proteins, Beta barrels, Fluorescent Proteins, Chromoproteins, Chromophore

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11569155/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The Author is Director of Cambridge RNA Technologies Ltd and MorphoKinetics Ltd, neither of which are connected to the work contained in this pre-print

An AI-assisted genomic census of 193,342 assemblies finds a GFP-barrel family in unreported lineages, carrying pseudogenes in both fungi and cnidarians

Adam Sidaway[1,2]

[1] Cambridge RNA Technologies Ltd, B25 Bioinnovation Centre, Cambridge Science Park, Cambridge CB4 0FW, United Kingdom

[2] MorphoKinetics Limited, B25 Bioinnovation Centre, Cambridge Science Park, Cambridge CB4 0FW, United Kingdom

Correspondence: adam@cambridgernatechnologies.com, adam@morphokinetics.com

Abstract

Green fluorescent protein and its homologues share an eleven-stranded beta-barrel enclosing a central helix, on which three consecutive residues cyclise to form the chromophore. A family of these barrels carrying a buried network of two arginines, two glutamates and a tyrosine, with a hydrophobic residue in place of the chromophore tyrosine, was recently assembled from protein records by Schneider and Zimmer. Here 193,342 genome assemblies were searched in all six reading frames, returning 396 loci, of which 140 read the complete buried network and collapse to 77 distinct carrying organisms. The charge network family reaches two sponge genera, a coronate jellyfish, three dinoflagellate transcriptome records and an archaeal genome, none of them represented in any proteome. In fungi and in cnidarians a substantial minority of loci carry an in-frame stop or a frameshift: 36 of 144 fungal loci, and 18 of 66 cnidarian loci across eleven genera. A Dollo reconstruction over a 28-genome Aspergillus clade requires two to six independent losses at the species level. The copies that survive are under purifying selection, at a median ratio of non-synonymous to synonymous substitution of 0.146 per species pair over 39 such pairs, and in one anthozoan lineage the gene is instead tandemly amplified to eleven copies. Two novel bacterial copies sit in mobile DNA, one inside a phiE-type prophage in Burkholderia pseudomallei and one beside an IS4 family transposase within a toxin-antitoxin array in a Myxococcota genome. A family that protein records show as intact is, in genomic sequence, intact in some lineages and breaking down in others.

Introduction

Green fluorescent protein, the staple laboratory reporter, was found during the purification of aequorin from *Aequorea* [1] and was cloned and sequenced from Aequorea victoria thirty years later [2]. It folds into a distinct eleven-stranded beta-barrel enclosing a single central helix [3, 4]. Three consecutive residues on that helix, Ser65-Tyr66-Gly67, cyclise and oxidise after folding to form the p-hydroxybenzylidene-imidazolinone chromophore that gives the protein its absorbance and emission. The reaction needs no cofactor and no enzyme, only molecular oxygen, and it proceeds in any host that will fold the barrel [5]. Homologues of this barrel occur throughout the anthozoan corals and the hydrozoan jellyfish, and more sparsely in copepods, lancelets and ctenophores [6, 7, 8]. Several have been developed into the fluorescent reporters in routine laboratory use, and mutation of the barrel shifts the emission across the visible spectrum [9].

Not every member of the superfamily fluoresces, and the structures deposited for it include members with no fluorescent function [33]. Chromoproteins absorb strongly and emit weakly or not at all, releasing the absorbed energy non-radiatively through twisting of the chromophore instead of as light. The residues that distinguish them from their fluorescent relatives sit *around* the chromophore, and fluorescent and non-fluorescent states have been interconverted by mutagenesis at four such positions [10]. Within the anthozoan radiation the non-fluorescent purple-blue chromoproteins form one paralogous lineage retained across every sampled scleractinian family, with two pigments, one in *Stylophora pistillata* and one in *Echinopora forskaliana*, reported as having arisen independently of the rest [7]. Two cases bear directly on what follows. An orthologous group present in every ctenophore examined lacks the residues chromophore formation needs, reading glutamine-threonine-asparagine or similar in place of the chromophore tripeptide, and losing both the arginine and the glutamate equivalent to avGFP Arg96 and Glu222. Francis and colleagues propose that these proteins do something other than fluoresce, and name no specific biochemical function [11]. In the hydromedusa *Aequorea coerulescens* a colourless homologue retains the chromophore tyrosine associated with fluorescence but carries glutamate at position 222 and does not fluoresce; an E222G substitution recovered by random mutagenesis restores green fluorescence [12]. In that colourless state the backbone has cyclised while Tyr66 remains unoxidised, which a crystal structure resolves as a trapped intermediate of chromophore formation [13].

Schneider and Zimmer [14] recently described a group of barrels that departs further still, keeping both Arg96 and Glu222 while losing the chromophore. They asked whether the AlphaFold database holds GFP barrels that keep the catalytic machinery without being annotated as fluorescent proteins, and searched it with the *Aequorea victoria* green fluorescent protein, avGFP, as a structural query. That search returned one sequence, a 218-residue protein from an Alphaproteobacterial metagenome-assembled genome, superimposing closely on avGFP at 17% sequence identity, and it

is the founder their family is built on and the reference its positions are numbered against. From it they assembled 22 complete barrels spanning bacteria, fungi and corals in which the chromophore-forming tyrosine is replaced by a hydrophobic residue, and in which five buried positions carry two arginines, two glutamates and a tyrosine. They report that this network is more conserved than the protein surface, that it occupies the most strongly constrained sites in the alignment, and that ancestral reconstruction places a non-tyrosine residue at the chromophore position early in the lineage. A barrel that holds a buried charge network where a chromophore would sit is a third state for the barrel family, alongside the fluorescent proteins and the chromoproteins, and what it does is not known.

That study was assembled from protein records, so whether this charge network family's genes are intact, how far the barrel reaches into sequence that no proteome represents, and how the lineages carrying it are related to one another all remain open.

This work searched genomic sequence directly. Reading genomes in six frames to recover what a protein record cannot hold is long-established practice for pseudogenes [15, 16] and is how chemoreceptor regression has recently been read across vertebrates [17]. Oral-anatomy regression has been read in myrmecophagous mammals on the same principle at a narrower scope, over a named set of orthologues in a defined clade [18]. The search returned 36 fungal pseudogenes and eight organisms from lineages that had not been reported to carry the charge network family. It also finds the gene lost repeatedly within one fungal section, amplified to eleven copies in one anthozoan, under purifying selection where it survives, and sitting in mobile DNA in two new bacterial phyla.

Results

A six-frame search of genomic sequence recovers copies that protein records do not hold

To recover copies of the charge network family absent from the protein records, genomic sequence was searched directly in all six reading frames across 193,342 assemblies, and a spliced gene model was built over each hit that supported one. A total of 396 loci were called, 299 of them with a model. Across the 24,479 fungal assemblies, 144 candidate loci were returned, against 46 fungal protein records in UniProt. A protein record corresponds to a single gene call, whereas a locus is a region of genomic sequence that aligns to the family, so a tandem array can yield several loci where a proteome yields one. The assembly lists, the inclusion criteria that differ

between them, and the 26 listed assemblies that could not be searched are set out in the Methods and in Tables S1 and S13.

The reading frame was read for each of the 144 fungal loci (Table S3). Forty-two carry an intact network with no stop and no frameshift, and 36 carry an in-frame stop or a frameshift within the modelled alignment. Those 36 are pseudogenes, four of them reading the complete buried network alongside the lesion, and decay of this gene has not been reported before. The remaining 66 could not be assigned to either state, because no spliced model could be built over them or because the five network positions could not all be read. Schneider and Zimmer [14] report 22 intact sequences and no pseudogenes, so decay of this gene has not been described before. Decay elsewhere in the fluorescent barrel superfamily has: the *Branchiostoma floridae* repertoire carries two pseudogenes beside thirteen functional genes [19], within a family of GFP-like proteins of differing spectral properties [20], and across cephalochordates the duplicated genes are held under purifying selection with that constraint relaxed in the most expanded clades [21], which is the pattern of amplification with constraint reported below for the octocorals.

The charge network barrel occurs in eight organisms from lineages not previously reported to carry it

The complete network was read at 140 loci, across 132 assemblies. Of the 140, 136 carry the reference residue at all five positions; three of the remaining four read a residue of the same formal charge and one reads the amide of the reference acid. Because one organism is often represented by several assemblies, the assemblies were collapsed to distinct carrying units, giving 77 of them of which 65 are named species, seven are unnamed species, four sit above species rank and one carries no taxonomic assignment below phylum (Table S2). Within the broad taxonomic searches a complete-network locus was recovered from 46 of the 24,479 fungal assemblies searched, 0.19%, against 22 of 340 cnidarian assemblies, 6.5%, two of 23,562 bacterial assemblies and one of 45,478 archaeal assemblies. The other 59 assemblies holding a complete-network locus were recovered by the targeted Burkholderiaceae, Myxococcota, Porifera, Streptophyta and coral searches and by the control search, each run over its own assembly list (Table S1). The survey of 56,757 environmental metagenome-assembled genomes returned eleven loci, two of which read the complete

network (Note N25). Because an assembly can also contain sequence from an organism other than the one it is named for, each unit was examined with the available contig-level instruments: four were identified as passengers within assemblies of other labelled organisms, and 28 were not placed, 23 of them lying on contigs too short to resolve and five not reached by any instrument (Note N24).

Eight organisms from lineages not previously reported to carry the charge network family were recovered: an archaeon, three dinoflagellates of the class Dinophyceae, two sponges, a coronate scyphozoan and a Thiomonas isolate. Four of the eight were returned by the genomic search, the archaeon, the two sponges and the scyphozoan, and their loci and carrier units are recorded in Tables S1 and S2. The three dinoflagellate records are transcriptome assemblies with no genome behind them, and are treated below. The Thiomonas isolate is a RefSeq protein record whose locus the genomic search did not recover. Seven of the eight are in no reference proteome. At the level of the higher groups these extend the family to one new domain, Archaea, and to two new phyla within Eukaryota, Porifera and Alveolata. Within Cnidaria, already a well known carrier phylum, they extend it to the Coronatae, in which it has not been reported. The cnidarian members of this barrel family described to date are fluorescent proteins and chromoproteins, so with the Coronatae record all three character states are now known from the same phylum.

The alveolate records are dinoflagellates that live as photosynthetic symbionts within radiolarians and foraminifera. Those three records are transcriptome assemblies with no genome behind them, so the five provenance checks applied to the 132 assemblies (Methods) cannot be run on them, and two of the three organisms live inside radiolarians, where host sequence is the obvious alternative origin. That alternative was tested against radiolarian sequence directly: no radiolarian RNA-seq run carries a family read, in ten informative libraries that each carry a working actin control, while two positive controls do. A radiolarian origin would put the gene in those libraries and it is not in them. The test is thinnest where it matters most, since nine of the ten informative libraries are from heterotrophic species and the dinoflagellates in question are themselves photosymbionts. The foraminiferan alternative, which applies to one of the three records, was tested over the assembled arm: of the 43 assembled rhizarian transcriptome sets searched, 20 are foraminiferan and all 20 carry a working actin control, none carries the charge network family except one declared metagenomic assembly from hypoxic fjord sediment. Only one of those 20 bears dinoflagellate symbionts, so that arm is thin where it matters in the same way the radiolarian one is. The records place family-like sequence in an alveolate lineage on three transcriptome assemblies, with the organismal attribution resting on that panel and on an 18S check rather than on a genome (Note N51).

Two further bacterial sequences were identified outside the genomic search, by blastp of the reference against the non-redundant database: a Sphingomonadaceae sequence from Antarctic soil and a Holophagales sequence from a coral reef, the first Acidobacteriota record for the charge network family. A third sequence carried in the query panel reads the complete network with methionine at the chromophore site, and the RefSeq assembly behind it is suppressed for an unverified source organism. None of these has a searched assembly behind it, so none enters the carrier counts or any tree reported here. Eight such sequences exist in all, including the one the published family was built outward from, and counted beside the 77 carrier units they give 85 organisms in which the family is identified in this work (Note N45).

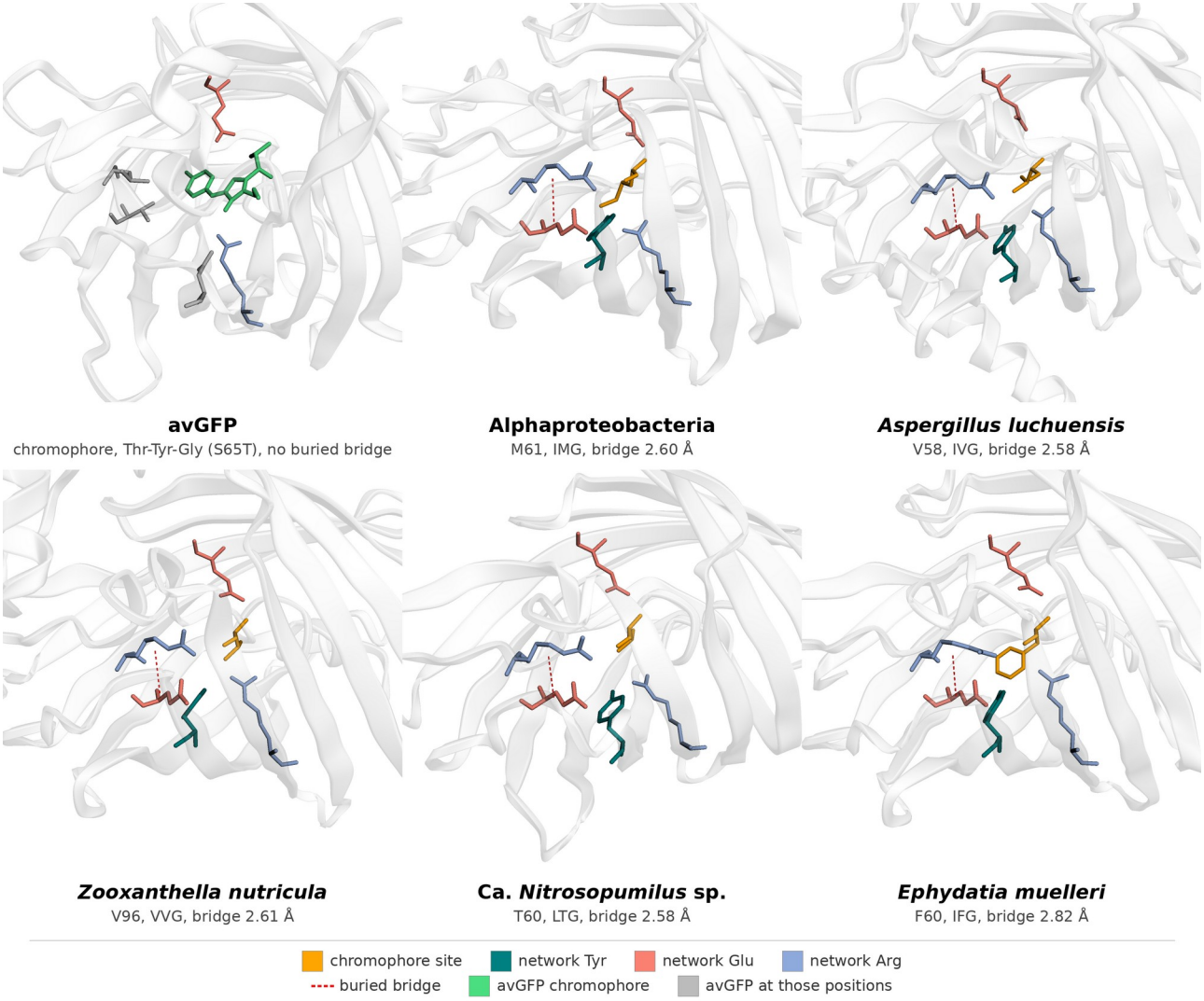


Figure 1. The chromophore site and the buried network across six lineages. Each panel shows the barrel lumen after superposition on a common frame, backbone in pale ribbon, with residues drawn at the five superposed network positions and at the chromophore site; the key below the panels gives the colouring, including the buried bridge between the first glutamate and the second arginine. The avGFP panel is the

S65T variant, in which the tripeptide reads Thr-Tyr-Gly against Ser-Tyr-Gly in the wild type, and it carries no buried bridge. The other five all carry the complete network with a bridge of 2.58 to 2.82 Å and a non-tyrosine residue at the chromophore site: methionine in the Alphaproteobacteria panel, valine in Aspergillus luchuensis and in Zooxanthella nutricula, threonine in the archaeon and phenylalanine in Ephydatia muelleri. The last three are the lineages reported here for the first time. Zooxanthella nutricula is the accepted name, under which Brandtodinium nutricula is a synonym. avGFP is the 1EMA crystal structure; the Alphaproteobacteria, Aspergillus and Zooxanthella panels are AlphaFold Protein Structure Database entries, and the Nitrosopumilus and Ephydatia models were predicted for this study with ColabFold 1.6.3 under AlphaFold2-ptm.

A barrel on a contig within a metagenome-assembled genome may belong to a contaminating organism. The carrier contig behind what is suggested here as the first archaeal occurrence holds a 16S gene 105 kb from the barrel that is 98.4% identical to the archaeon Nitrosopumilus maritimus, with no bacterial alignment at all, carries archaeal ribosomal protein S26e at 22 kb, and shows no composition break across 181 windows of 10 kb, so the locus likely, though not conclusively, resides within an archaeal genome (Note N32). The same search recovered six further archaeal loci, and none of them reads the network: no panel query builds a model over any of the six, and the best score among them reaches 46 bits against a 65-bit floor for the archaeal locus itself, so the archaeal occurrence is one locus and not a group of them (Note N34).

Among the newly recovered lineages, the sponge and jellyfish records place the family on both sides of the deepest splits within two animal phyla, which bears on when it was first present in animals. Barrels of this superfamily are well characterised in anthozoan corals and hydrozoan jellyfish as fluorescent proteins and chromoproteins. *Nausithoe racemosa* is a coronate scyphozoan, on the far side of the deepest split in Cnidaria from the anthozoans, and it carries two loci reading the complete charge network, one intact and one carrying a stop. Porifera had not previously been reported to carry the charge network family. The buried bridge between the second arginine and the first glutamate is present in all three new lineages that carry a model, at 2.58 Å in the archaeon, 2.61 Å in Zooxanthella nutricula and 2.82 Å in Ephydatia muelleri, both at the charge network family's own spacing of 45 to 55 residues, against 1 of 15 full-network fluorescent proteins measured in the same run. Each of the three records was recovered independently from more than one assembly, Phakellia ventilabrum from three and the other two from two, the two Nausithoe assemblies each carrying both of its loci.

Four loci belong to organisms other than the assemblies that carry them

Four of the 132 assemblies hold the locus on sequence belonging to another organism, and for each of the four the source lineage was established. Two are unannotated plant

scaffold submissions carrying a fungal sequence, *Corymbia citriodora* over 246 residues and *Eperua duckeana* over 254, both with valine at the chromophore position, which is the residue carried by the fungal and bacterial copies. The other two carry bacterial sequence: a 1,004 bp contig in *Chamaecostus congestiflorus* whose 187-residue model reads the complete network with threonine at the chromophore position and whose nearest known relative is a Sphingomonadaceae protein from an Antarctic soil metagenome at 69.4% identity, and a 2,083,518 bp contig at 65.7% GC in *Petrosia ficiformis* whose 206-residue locus has a bacterial nearest relative at 42.3% while the nearest poriferan copy is seventh at 27.1% (Notes N35 to N37). None of the four is present in the proteome of the organism its assembly is named for, so a search of protein records would not have reached any of them, and all four are excluded from the taxonomic counts reported above, because the source organism cannot be identified from the sequence recovered. A separate screen against FPbase removed 96 further loci as deposited laboratory constructs and held seven for review; among the 140 loci reading the complete network the highest identity to any FPbase entry is 0.402.

Lesioned copies occur in fungi and in cnidarians, and the gene is lost repeatedly in one Aspergillus clade

On inspection of the modelled alignment, each of the 36 lesioned fungal loci carries an in-frame stop or a frameshift. Four of them read the complete network alongside that lesion. In a further 22 all five network positions are readable but at least one departs from the reference, ten of these reading histidine at the arginine of 182, and in the remaining ten only two to four of the positions can be read at all. A locus that does not read the complete network is placed in the family on its identity to the kingdom-matched query that recovered it together with the positions that can be read, and Table S3 carries both for every locus. Across the 32 of these in the fungal set that identity runs from 55.4 to 96.0%, median 91.7%, and none falls below 50%, so homology to the family is not in question at any of them. A copy whose reading frame has been disrupted is not expected to retain all five positions (Note N39). Because a lesion called from an assembly may represent a sequencing or assembly error instead of a feature of the gene, 82 loci were called with an in-frame stop or a frameshift, and 71 of these, including 35 of the 36 fungal loci, were compared against the reads of their own BioSample (Table S4). The eleven not covered either share an assembly with a locus that was checked or fall outside the set that comparison was run over. Reads were mapped across each locus, a consensus was called, and that consensus was aligned with the same query used for the assembly.

Of the 71 loci put to the reads, 25 had reads covering the gene; 20 reproduced the assembly verdict outright, four reproduced part of the lesion and one differed at a single stop while remaining lesioned (Note N38; Table S4).

The 36 lesions do not represent 36 independent events. Sixteen of the lesioned loci carry an in-frame stop at the same codon. All sixteen are black aspergilli: ten in *Aspergillus niger*, five in *A. welwitschiae* and one in A. awamori. A count of lesions therefore overstates the number of times the gene was disrupted. Decay also reaches the fifth network position itself. Glutamine occurs there in one intact copy among the 77 carrying units and in 15 lesioned loci falling in three species, so a glutamine at that position can be a state a broken copy has reached through a single mutation rather than a residue the protein uses.

No lesioned prokaryotic locus was recovered. None of those 33 loci carries an in-frame stop or a frameshift, against 82 of the 266 non-prokaryotic loci carrying a model, a rate of 30.8%. Seventeen of the 33 are *Burkholderia pseudomallei* carrying one prophage insertion, described below, so they are likely one event observed seventeen times rather than seventeen observations. Collapsed to independent occurrences the test is over 17 loci, where 17 without a lesion arise by chance at 1.9×10^{-3} and a Fisher exact test gives $p = 4.2 \times 10^{-3}$. Over all 33 loci, treating each as independent, the same two tests give 5×10^{-6} and 1.6×10^{-5}. The other 25 prokaryotic loci lie outside the test: no gene caller emits a protein for any of them, and in one a stop was found inside the aligned segment by six-frame alignment without a model being buildable across it, so the absence is reported over the 33 and not over the 58 (Note N31). Bacterial genomes generally remove unselected sequence instead of retaining it in a readable state, so an absence of lesioned prokaryotic copies is equally consistent with faster loss than in the eukaryotic lineages and cannot be used to conclude an altered decay rate.

Whether the gene was lost once or many times was reconstructed over a tree (Figure 2A), as has been done for the coral fluorescent proteins [23]. Over the smallest clade containing the carriers recovered from the Aspergillus niger aggregate and its immediate relatives, 28 of the 30 genomes, Dollo parsimony requires seven or eight independent losses, the ambiguity being whether two strains sharing a stop at the same codon inherited one lesion or acquired two. Each tip is read from a single assembly of its species, so the count was resampled by redrawing that assembly at each of the 24 tips whose species holds more than one: over 10,000 redraws it spans five to eleven losses, median seven, 95% interval five to ten, and no draw returns fewer than five (Note N40). That interval is a lower bound, since more assemblies exist for the carrier genera than were searched.

At the species level the corresponding figure is two to six: two of the eight dissolve, because the species they rest on retain an intact copy in another assembly, two shrink, and two contain species the fungal search never covered, which cannot be called absent. Of the eight species with a positive result and more than one assembly, seven give discordant calls between their own assemblies, and seven of the 123 *Aspergillus niger* assemblies carry an intact copy. The intact *Aspergillus luchuensis* copy returns

no gene-body reads in any of the four RNA-seq libraries, which bears on the same question; the four CAGE libraries tag start sites rather than gene bodies and three of them carry one, three and one tags at the locus. Loss has therefore occurred repeatedly within one section of one genus, which is consistent with the gene being dispensable in those lineages.

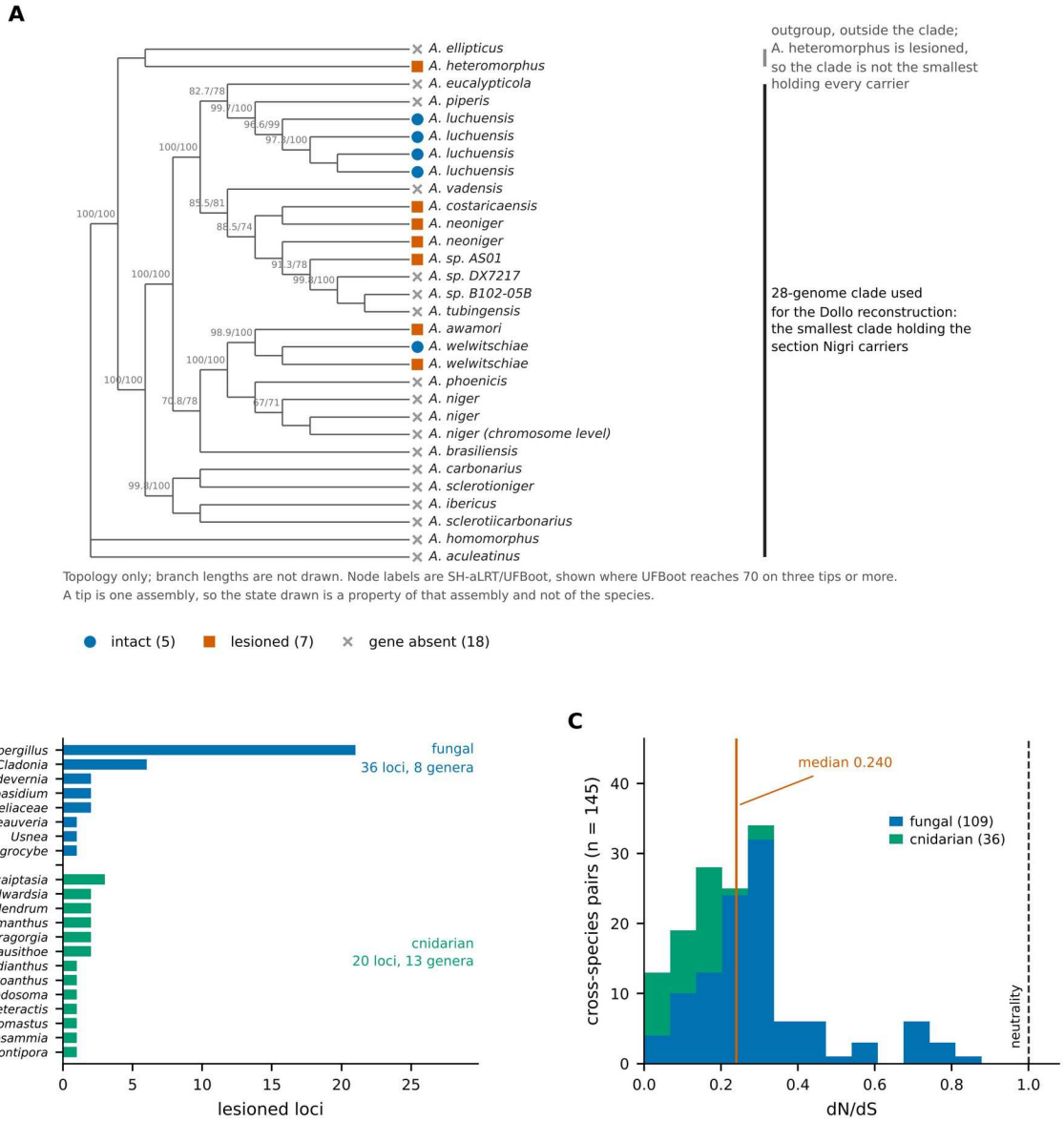


Figure 2. The gene is lost repeatedly in one fungal section, decays across many cnidarian genera, and remains under purifying selection where it survives. (A) Maximum-likelihood species tree of Aspergillus section Nigri and its outgroups, 30 genomes, with the state of the locus at each tip: intact reading frame with the complete buried network (5 genomes), a locus carrying an in-frame stop or a frameshift (7), and no locus recovered (18). The bracket marks the 28-genome clade

over which the Dollo reconstruction was run, the smallest clade containing the carriers of the A. niger aggregate and its immediate relatives; Aspergillus ellipticus and A. heteromorphus lie outside it as the outgroup, and A. heteromorphus carries a lesioned locus, so the clade does not hold every carrier of section Nigri. Node labels are SH-aLRT/UFBoot, drawn where UFBoot reaches 70 on a node of three tips or more; branch lengths are not drawn. A tip is one assembly, so each state is a property of that assembly rather than of the species. (B) Lesioned loci by genus. The two sides are counted over different sets. The fungal bars are the 36 lesioned loci of the fungal search across eight groups, one of them a Parmeliaceae isolate named only to family. The cnidarian bars pool the cnidarian search with the coral survey, giving 20 lesioned loci across 13 genera, the two genera beyond the cnidarian search being Duncanopsammia and Montipora. The lesion-rate comparison in the text is made on the cnidarian search alone, at 18 of 66 loci across eleven genera, so that each side carries one denominator. Cnidarian counts are floors, because loci spanning a tandem array were merged. (C) dN/dS over the 145 informative cross-species locus pairs, 109 fungal and 36 cnidarian, stacked by lineage. The solid line marks the median at 0.240 and the dashed line the neutral expectation at 1.0.

The cnidarian copies are also being lost, at a rate the fungal one cannot be distinguished from: eighteen of 66 cnidarian loci carry a stop or a frameshift against 36 of 144 fungal loci (odds ratio 1.13, 95% confidence interval 0.58 to 2.18; Fisher exact p = 0.74). Three of the eighteen read the complete network alongside the lesion, and across the other fifteen the identity to the kingdom-matched query runs from 51.8 to 87.9%, median 77.6%. What differs is where the lesions sit. Twenty-one of the 36 fungal lesions are in *Aspergillus*, while the cnidarian lesions fall across eleven genera with no genus holding more than three (Figure 2B), so the fungal decay is concentrated in one genus and the cnidarian decay spread through a clade. The same Dollo reconstruction over 237 cnidarian species returns sixteen to eighteen losses, which is not comparable with the fungal seven or eight, resting on a taxonomy backbone in place of a supported topology and on far more taxa. The cnidarian copy counts understate the true number for a second reason: seventeen loci were called with a genomic span at least three times the coding length, and re-modelling them recovers more copies than the locus table records, which the gene amplification section below sets out.

A census of stops cannot distinguish a gene that is drifting out of use throughout its range from a gene under constraint that has been lost in some lineages. Among the informative pairs, the most informative single comparison is provided by the lichen pair. *Tuckermannopsis ciliaris* and *Flavocetraria cucullata* carry the same 224 residues without a single difference, across three introns that have diverged at 11% and 8,691 bp of co-linear flanking sequence that has diverged at 7.5%. Eighteen codons differ between the two sequences and all eighteen differences are silent. The coding sequence carries 529.7 non-synonymous and 142.3 synonymous sites under the Nei-

Gojobori count [36], against which eighteen synonymous differences and none non-synonymous stand against a neutral expectation of about 67 non-synonymous differences, and a Fisher exact test on those site-classified counts returns p = 2.9 × 10⁻¹³. The two proteins were already known to be identical when this pair was picked out. The figure above therefore describes the pair and is not an independent test. Mutation has therefore had the opportunity to alter this protein in two lineages, and none of that mutation has been retained.

Transcript data were examined for the fungal copies and the intact copies were found not to be transcribed: eight public libraries from four studies totalling 132 million mapped reads returned no reads at the intact *Aspergillus luchuensis* locus, while a broken copy in *Lepraria finkii* is transcribed at 0.18 times the average gene of its library. A search of 6,843,051 contigs across 203 fungal transcriptome sets returned no acceptable hit, where agmatinase, two genes downstream, was recovered from 122 of the same sets (Note N50).

Where two species carry the same locus, the synonymous sites and the surrounding non-coding sequence record how much mutation has accumulated. Five such pairs are in this set. Three are uninformative, diverging by 0.01 to 0.15 per cent over 17 kb of flanking sequence and not at all at synonymous sites, so no amino-acid change had been offered to be declined. The remaining two are informative, and both show purifying selection (Table 1).

Table 1. The two cross-species pairs where the comparison can be made on coding and on non-coding sequence. Synonymous and non-synonymous divergence over the coding sequence, beside the divergence of the introns and flanking sequence across the same window. Neither pair is a sample. Table S5 carries a flanking measurement for every pair in the final analysis; these two are given because they are the co-linear pairs that carry synonymous divergence, the other three having none.

pair	clade	dS	dN	dN/dS	non-coding divergence
Tuckermannopsis ciliaris / *Flavocetraria cucullata*	Ascomycota	0.139	0	0	11.0% in three introns, 7.5% in 8.7 kb of flank
Paragorgia pacifica / *Hemicorallium imperiale*	Cnidaria	0.037	0.004	0.11	3.9% in 7.4 kb of flank

The same measurement was extended to every pair the data permit. Over 145 cross-species locus pairs with non-zero synonymous divergence below saturation, the median dN/dS is 0.240 per locus pair (Figure 2C; Table S5). Those pairs are not independent, being drawn from 42 loci in 39 species pairs, so the median was recomputed one species pair at a time: the median of the per-pair medians is 0.146 over 39 units, and a cluster bootstrap over those units puts the 95% interval at 0.111 to 0.163, entirely below one. A ratio this far below one indicates purifying selection, so amino-acid changes are being removed from the copies that survive.

The two lineages differ, at 0.274 over the 109 fungal pairs against 0.124 over the 36 cnidarian pairs, and the cnidarian figure is a floor. Its five species are four octocorals and one antipatharian, all of them tandem-array carriers, so every cnidarian pair is drawn between two arrays and gene conversion within an array depresses divergence. No pair in the set exceeds 0.84. Three limits bound this subsection: the losses are counted in one section of one genus and say nothing about dispensability elsewhere, the cnidarian counts are floors because loci spanning an array were merged, and the selection estimate was measured on the copies that survive and not inside the lineages carrying the lesions.

In one anthozoan lineage the gene is tandemly amplified

Seventeen cnidarian loci were called with a genomic span at least three times the coding length, and re-modelling them returned 29 copies across six species where the locus table had recorded 12 of these wide-span loci among them. Most are tandem arrays. Tandem arrangement of these genes is established in anthozoans: the cyan and green genes of *Acropora digitifera* lie clustered within about 80 kb and are attributed to tandem duplication [24], fifteen *Acropora* species carry nine to eighteen fluorescent-protein genes each with tandem duplication evident in synteny [23], *Montastraea cavernosa* expresses its set from at least four and possibly seven loci [25], and in *Acropora millepora* pigment concentration tracks the number of copies carrying a particular promoter type [26]. Notes N26 and N27 give the re-modelling procedure and the three grounds for calling this an amplification and not an assembly duplication. *Hemicorallium imperiale* carries four copies within 56 kb of one contig, two of them a tandem pair and the other two separate regions, and all four read the complete network. *Paragorgia arborea* carries six, all six reading the complete network, five of them within 49 kb of one contig and the sixth in a second assembly of the same animal, the two being the principal and alternate pseudohaplotypes of one diploid specimen, so that count is of two assemblies and not of one genome. The largest is in *Plumapathes pennacea*, a black coral, which carries eleven copies in one array at periods of 2,731 and 2,755 bp, with the units 95.1 to 99.5% identical to one another and two of the eleven carrying an internal stop. Larger sets are known in this superfamily, 31 to 35 fluorescent-protein copies having been counted by quantitative PCR in four *Acropora*

digitifera colonies [27], so what is new here is the lineage, and that the duplication is reaching the charge network family and not only the fluorescent proteins.

Six residues occupy the chromophore position, and valine and phenylalanine divide by lineage

The family contributes 89 tips to a 1,002-sequence alignment built so that this column is read in one frame across all three groups, 63 of them loci and 26 panel sequences. Across those 89, the chromophore position carries six residues. Valine occurs in 44, phenylalanine in 26, threonine in 12, cysteine in four, methionine in two and alanine in one. The glycine that cyclisation requires at the next position is present throughout. The two methionines are protein records with no assembly behind them, and the single alanine is one of the lichen loci described above, a residue absent from the 22 sequences of the original description. The other two groups in that alignment read differently: tyrosine occupies the position in 543 of 607 canonical fluorescent proteins, and in all 21 engineered-dark and seven non-fluorescent sequences.

The same position was read across the 131 loci that recover the complete network from genomic sequence. Valine occurs in 88, phenylalanine in 33, threonine in four, alanine and cysteine in two each and tyrosine in one, with one further locus where the gene model deleted the position. These are counts of loci and not of organisms: nine loci belong to sequences carried through the screens as controls and are set aside, and six of those organisms hold no other complete-network locus, the archaeon among them. An *Edwardsia elegans* locus carries the complete buried network together with a tyrosine at the chromophore position, which is the residue the fluorescent proteins and chromoproteins carry there (Note N41).

Every valine occurs in fungal or bacterial sequence, 69 and 17 respectively, with two further loci on plant assemblies whose sequence is fungal. The 17 bacterial valines are all *Burkholderia pseudomallei*, one locus in each of 17 assemblies carrying the same prophage, so they are one event seen 17 times and carry no weight of their own. Every phenylalanine occurs in an animal, 30 cnidarian and three poriferan across 19 genera, and no animal locus reads valine. A single phenylalanine outside animals sits in a Verrucomicrobiota metagenome-assembled bin, which has no searched assembly behind it and falls outside the 131 for that reason. The substitution is therefore fixed separately in the animal and the fungal lineages and is not one state inherited from their common ancestor. The valine count exceeds the 42 intact fungal loci because a locus can carry all five network positions and a lesion elsewhere, and is counted in both places.

Ancestral states were reconstructed at the chromophore position on a fixed topology under both substitution models. No sequence below the family's own vertex carries tyrosine at this column, so a posterior for tyrosine there follows from the tips and not from the reconstruction. It is given here against the published description, with no ancestral state claimed from it (Note N46). At that vertex the two models agree on which non-aromatic residue is favoured: methionine at 0.584 and 0.627, threonine second at 0.21 to 0.26, so the vertex reads methionine or threonine at 0.83 to 0.84. What moves that column is the taxon set. The earlier 44-sequence set returned phenylalanine at 0.756, and rebuilding it to 89 moves the column by 0.499, where the substitution model is worth 0.036 and the topology 0.007.

That posterior rests on two sequences. Methionine occurs in 2 of the 89 family sequences against valine in 44 and phenylalanine in 26, and those two sit closer to the vertex than any other, 3 and 4 edges away where the median is 10. Five tests were applied (Note N15). Pruning both drops methionine to 0.005 and the vertex reads tyrosine at 0.50 and phenylalanine at 0.39. A jackknife dropping a tenth of the sequences over 200 replicates spreads the methionine-or-threonine posterior from 0.23 to 0.85, and the spread turns on whether a replicate happened to drop one of the two: of the 37 that did, the best state is aromatic in 24. Fitch parsimony, which ignores branch lengths, returns threonine and tyrosine at the vertex and no methionine. Constraining the vertex to methionine against threonine separates them by 1.75 log-likelihood units, inside the two that would tell them apart.

Permuting which sequence carries which residue over 200 replicates, holding the tree, the model and every other column fixed, puts methionine between 0.0009 and 0.059, so the observed 0.584 sits an order of magnitude outside that null. The permutation holds the family's residue multiset fixed and that multiset carries no tyrosine, so the null bears on the placement of methionine among the residues the family does carry and not on tyrosine.

Outside the charge network family, tyrosine is favoured at every deep vertex reached, 34, 41 and 34 of them on three taxon samplings under both models. Pruning the deposited laboratory constructs and reconstructing again leaves all of them reading tyrosine, 21 of 21 and 20 of 20 under two definitions of the exclusion, each at 0.95 or better. The three samplings are not three independent topologies: the 1,002-sequence search was stopped at 169 iterations where the other two ran to completion.

That result holds only for the outgroup it was computed against. All 25 outgroup sequences are fluorescent proteins and chromoproteins, and all 25 read tyrosine. Across the 722 superfamily sequences falling within the family's length range the figure is 485; among the 475 of those that are fluorescent proteins it is 448, and among the other 247 it is 37, with leucine the commonest residue instead. Fluorescence requires an X-Tyr-Gly chromophore, so an outgroup of fluorescent proteins already contains the answer being reconstructed.

Neither reconstruction reaches the node at which this family joins the rest of the superfamily, and that node cannot be read from these data either: the sequences below it carry no tyrosine at this column, so an outgroup drawn from them would settle the question before it was asked. The comparison itself does not depend on any reconstruction. Of the fluorescent members of the barrel family, 448 of 475 carry the chromophore tyrosine; the charge network family carries it in none.

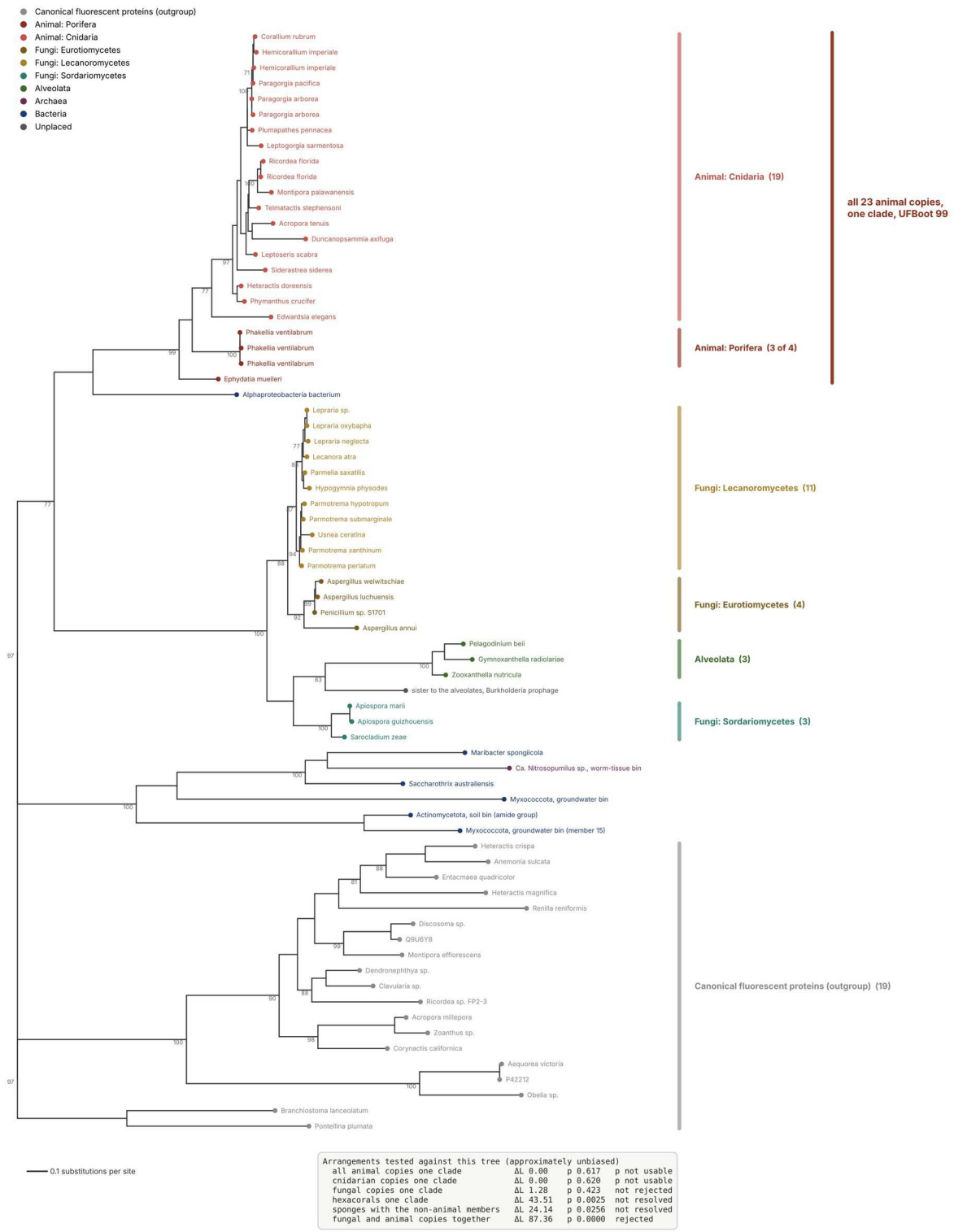


Figure 3. The animal copies form one clade. Consensus tree of 71 tips under LG+G4 from 1,000 ultrafast bootstrap replicates. Support is drawn only where it reaches 70 on a node of three tips or more, so an unlabelled node is below one of those thresholds

and not thereby unsupported. The tips are 19 canonical fluorescent proteins as the verified outgroup, 23 animal copies of which 19 are cnidarian and four poriferan, 18 fungal copies, three alveolate copies, seven prokaryotic copies of which six are bacterial and one archaeal, and one sequence lying within a Burkholderia prophage. That sequence is drawn as unplaced rather than as bacterial because the colouring records the group a sequence is assigned to and this one has none: its contig is Burkholderia, but on this tree it falls beside the alveolate clade and moving it to the prokaryotic side costs a further 25.44 log units. The tree is rooted on the edge subtending the canonical fluorescent proteins, which separates all 19 of them from the other 52 tips at ultrafast bootstrap 97 here and 98 under LG+C20+F+G, so the family is monophyletic on this tree as it is on the five of Note N43. The 23 animal copies are recovered as a single clade with no non-animal tip inside it, at ultrafast bootstrap 99 here and 99 under LG+C20+F+G. A bracket reading (n of m) marks a group that is not monophyletic on this tree, spanning the largest node whose tips all belong to it: the sponges are the one such group here, with Ephydatia muelleri outside the Phakellia group. The arrangements in the box were tested against this 71-tip tree and not against the 77-tip three-class tree (Methods). The prokaryotic relationships are not read from this tree: the six-member grouping is recovered here at 100, but the published description's founder sequence falls sister to the animal clade at 60 here and 50 under LG+C20+F+G, both below the drawing threshold, against 95 and 93 on the 77-tip alignment that the prokaryotic section rests on.

Topology and genomic context of the animal and prokaryotic copies

The 71-tip tree of Figure 3 carries 23 animal copies, nineteen cnidarian and four poriferan. They were recovered as a single clade under both LG+G4 and LG+C20+F+G, at ultrafast bootstrap 99 under each model. Competing arrangements were imposed as topology constraints and scored against the best tree found by the approximately unbiased test (Figure 3 for six of them; Table S7 for every arrangement tested). Neither the sponges nor the fungal Eurotiomycetes are monophyletic within that tree, each having one tip outside the group its other three form. A single arrival in animals before the sponge-cnidarian divergence, followed by vertical descent, is sufficient to account for the topology, and the sponge and coronate records described above place that arrival before the deepest splits within both phyla. The gene tree cannot exclude every alternative history, and two arrivals from closely related donors remain open. The approximately unbiased test returns p-AU 0.026 for separate sponge arrivals, 0.0025 for hexacoral monophyly and 0.0042 for a sponge grouping with the hexacorals. They sit among 58 arrangements tested across eight topology sets built with different tip counts, models and search strategies (Table S7), where a Bonferroni threshold would be 0.00086 and none of the three reaches it, and the same machinery

is shown in the Discussion rejecting an unconstrained topology and marking an arrangement identical to the best tree as rejected.

On the 77-tip three-class tree, which carries all sixteen reference sequences against a hand-assembled outgroup verified by residue, the unconstrained topology is not rejected at p-AU 0.872; this is the three_class_tree_16 run of Table S7, and the prokaryotic arrangements below were tested on separate 77-tip runs named there. Imposing monophyly on the non-fluorescent chromoproteins costs 497.90 log units, grouping the charge network family together with them 499.44, placing the family sister to them 543.25 and imposing monophyly on the canonical fluorescent proteins 884.49, each rejected at p-AU below 10^{-4}. That value is the resolution of the replicate count and is given in place of the raw p-AU, which across these four rows is not monotonic in the likelihood difference; Table S7 carries the raw values. Forcing the family itself into one clade costs 0.0007 log units and returns the unconstrained tree, so no p-value can be read from that row. The rebuilt 430-taxon tree rejects the same three arrangements by larger margins, 2,012 log units and more, its weakest raw p-AU being 6.4×10^{-40}.

The seven prokaryotic copies do not resolve as one group, and no test that forces them together rejects that arrangement. Six of the seven were recovered as a clade at ultrafast bootstrap 100 under both LG+G4 and LG+C20+F+G, while the seventh, which is the sequence the residue numbering is anchored on, falls sister to the clade of 23 animal copies, at 60 and 50 under the two models on the 71-tip tree, both below its drawing threshold (Figure 3), and at 95 and 93 on the 77-tip alignment the prokaryotic section rests on. The other six sit on the far side of the canonical fluorescent-protein outgroup. Forcing all seven costs 5.15 log units at p-AU 0.127 on the 53-tip tree, and on the 77-tip alignment 7.68 at p-AU 0.21 under LG+G4, 4.04 at 0.286 scored under the mixture model and 3.63 at 0.19 searched under it by the posterior mean site frequency approximation, PMSF [30], each arrangement searched from four independent starting trees (Table S7). Removing the archaeon from the bacterial copies costs 89.02 log units and is rejected, although which copy it sits beside within the group is not resolved: the 77-tip alignment places it beside the *Saccharothrix* copy, a 32-tip alignment beside the *Maribacter* copy at ultrafast bootstrap 91, and pairwise identity places the *Maribacter* copy nearest at 44.5% against 41.6%.

Grouping the eukaryotic copies to the exclusion of the prokaryotic ones costs 70.20 log units and is rejected. The sequence responsible is the unplaced one lying within a *Burkholderia* prophage, which falls beside the alveolate clade at ultrafast bootstrap 99 and costs a further 25.44 log units to move to the prokaryotic side. It is reported here as unplaced and not as bacterial.

The searches do not converge on one tree. Four of the five arrangements returned different topologies from different starting trees, and the unconstrained search spanned 0.28 log units. Two arrangements fall within 0.0001 log units of the best tree,

which is that tree with different branch lengths, so the test divides probability between them arbitrarily and their p-AU values are unusable.

Each search covered a different part of the prokaryotic tree. Two of the 23,562 reference genomes of the broad bacterial search carry a complete-network locus, *Maribacter spongiicola* in the Bacteroidota and *Saccharothrix australiensis* in the Actinomycetota, against one of 45,478 archaeal assemblies, two of 7,998 Myxococcota assemblies and 18 of 15,928 Burkholderiaceae assemblies, one of which is a panel sequence recovered in its own assembly and set aside on that ground. The remaining 17 resolve to a single species. All 17 are Burkholderia pseudomallei, of which 3,764 assemblies were searched, and all 17 carry the same prophage insertion. A gene inherited vertically within one species would not be confined to 17 of its 3,764 sequenced assemblies without several thousand independent losses.

The genomic context of each prokaryotic locus was examined instead, as a line of evidence independent of the trees (Table S8). Because a context cannot be interpreted without a background, five matched control windows of the same size were placed on the same contig for each locus, at least 20 kb from the locus and from the contig ends (Figure 4). That comparison is reported separately for each search, for the reason given with Figure 4D. Where a carrier row and its controls are both near zero, as they are for the fungal loci, the comparison records thin annotation as much as an empty flank.

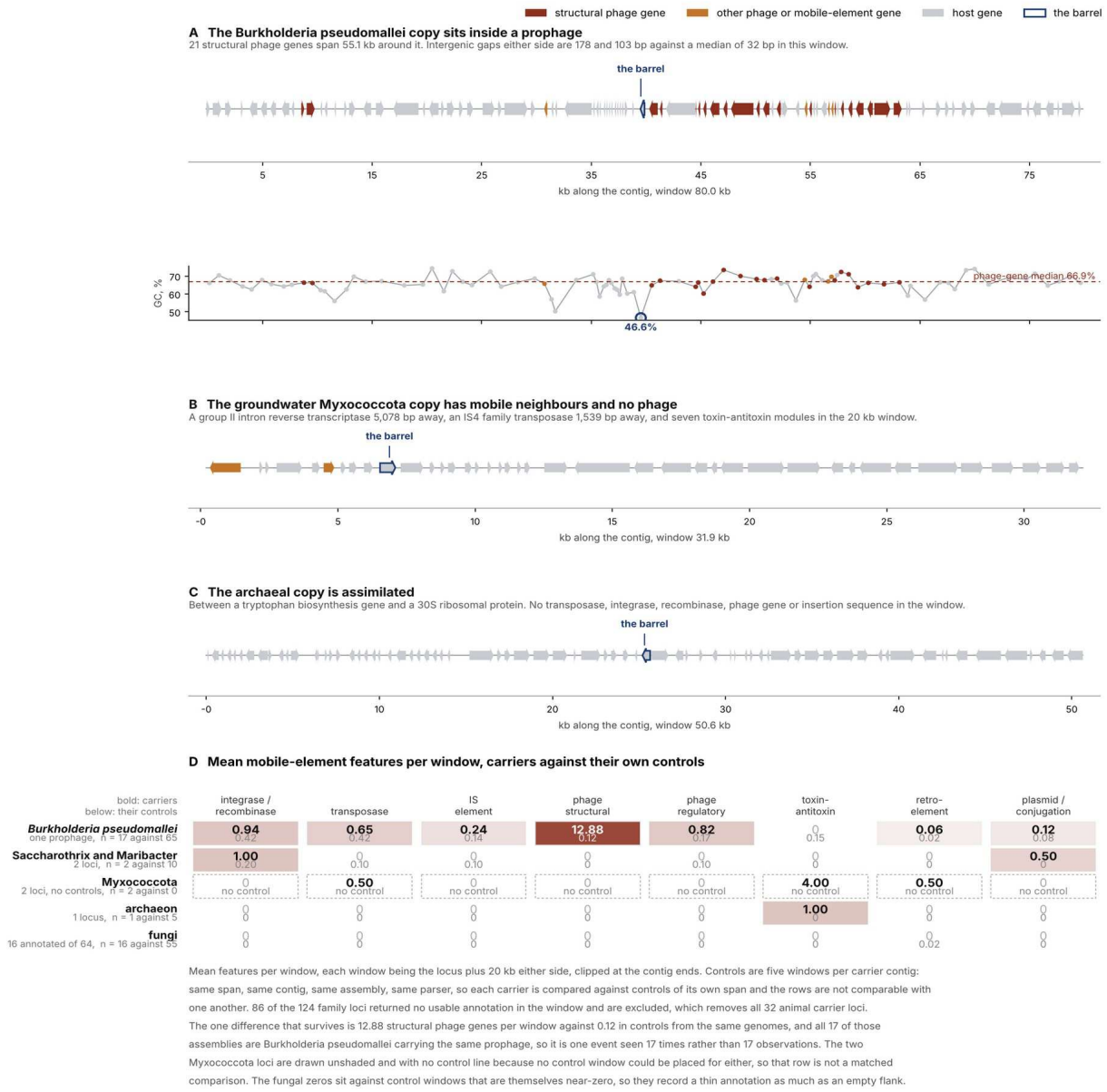


Figure 4. The genomic context of three prokaryotic loci, drawn to scale, with a matched-control comparison stratified by search. (A) The Burkholderia pseudomallei locus in an 80.0 kb display window: 21 structural phage genes span 55.1 kb around the barrel, and the intergenic gaps either side of it are 178 and 103 bp against a median of 32 bp in the window. The trace beneath gives per-gene GC. (B) The groundwater Myxococcota locus in a 31.9 kb display window; the seven toxin-antitoxin modules counted in the panel annotation lie within 20 kb of the locus. (C) The archaeal locus in a 50.6 kb window, between a tryptophan biosynthesis gene and a 30S ribosomal protein. Genes are coloured as structural phage, as other phage or mobile-element, or as host, with the barrel drawn in outline. The Maribacter and Saccharothrix loci are described in the text and in Table S8 and carry no panel here. (D) Mean features per window, carriers in bold with their own controls beneath, each control being five

windows of the same span on the same contig and assembly, parsed identically. Panels A to C are drawn over display windows and panel D is computed over the analysis windows, which run from 20.5 to 58.0 kb, so each carrier is compared against controls of its own span and the rows are not comparable with one another. The two Myxococcota loci carry their own row, drawn unshaded and with no control line, because no control window could be placed for either, so that row is not a matched comparison; the two bacterial loci that do carry controls are shown separately above it. The rows are stratified and not pooled, because 86 of the 124 family loci with a context window returned no usable annotation inside it and are excluded, which removes all 32 animal carrier loci, whereas each control window carries annotation by construction. The fungal zeros sit against control windows that are themselves near zero.

In *Burkholderia pseudomallei* the locus sits inside a prophage, within its tail-structural operon. The phage tail gene immediately beside the barrel matches *Burkholderia* phage phiE52237 at E = 2.1×10^{-238}, and the same region carries a P2-family major capsid protein, an Ogr/Delta-like regulator and an Rz/Rz1 lysis pair, so the element is a phiE-type, P2-family prophage. *Burkholderia* appears nowhere in the published description, whose founder is an alphaproteobacterium, so this association is reported here for the first time. Across 17 carrier assemblies and the 65 control windows that could be placed, carrier windows average 12.88 phage structural genes against 0.12 in the controls, a 105-fold excess (Figure 4A). The barrel reads 46.6% GC against a median of 66.9% across the phage genes of the same window, so it is a compositional outlier within the element carrying it. A census of 6,258 assemblies of Burkholderia pseudomallei, B. mallei and B. thailandensis, searched by blastn for the copy itself and for the phage tail gene beside it, finds the tail gene at nine tenths of its length or more in 2,709 B. pseudomallei assemblies and the barrel at full length, at 98.1 to 100% identity, in 33. Those 33 are the same 17 assemblies as above once GenBank and RefSeq copies of one assembly are collapsed on the numeric stem of the accession, which this census does not do. The barrel is therefore occasional cargo in a common prophage rather than part of its core. The 17 carriers resolve to 15 strain labels and carry one insertion, not 17 independent events.

The remaining five loci differ from the prophage and from one another. The groundwater Myxococcota copy MBI3245128.1 lies 1,539 bp from an IS4 family transposase and 5,078 bp from a group II intron reverse transcriptase and maturase, inside a run of seven type II toxin-antitoxin and addiction modules that spans it on both sides (Figure 4B). Both of the two mobile-element products annotated on that 106,316 bp contig, and seven of its eight toxin-antitoxin genes, fall within 5.1 kb of the barrel, while the remaining 80 annotated features on the contig carry one between them. The architecture is that of a defence island: no relaxase, origin of transfer, *tra* or *mob* gene, or type IV secretion component occurs anywhere on the contig, so no conjugative element can be named. Neither Myxococcota copy appears among the twelve

neighbourhoods Schneider and Zimmer analysed [14], and the placeholder taxon is not among their bacterial seeds, so this association is reported here for the first time. Two things bound it: the contig begins 6,537 bp before the locus, so the upstream boundary of the cluster lies off the end of the assembly, and no matched control window could be placed for either Myxococcota locus.

The second Myxococcota copy comes from the same placeholder taxon and the same sampling programme, and it sits in ordinary chromosome between a methyltransferase and an SRPBCC family protein, with one toxin-antitoxin gene 3,086 bp away and nothing else in any of the eight feature classes across 34 features.

Maribacter spongiicola sits between two recombinases, 3,720 bp upstream and 429 bp downstream, with an ArdA anti-restriction gene 3,160 bp upstream, which is the architecture of an integrative or mobilisable element. That arrangement is already reported by Schneider and Zimmer [14]; what is added here is the comparison against a background. Its five matched control windows on the same scaffold carry nothing in any of the eight classes across 428 features, although seven of their ten pairings overlap, so those five windows are closer to one measurement than to five. The *Saccharothrix australiensis* locus carries no feature in any of the eight classes across 66 annotated features, where its five matched control windows carry one per window on average. One observation of zero against a background rate of one has a probability of 0.37, and two of those five control windows are themselves empty, so what this records is a locus with no mobile-element feature beside a chromosome that scatters them thinly. The archaeal locus carries a single toxin-antitoxin gene 17.5 kb away (Figure 4C), and its control windows read zero under one draw of the five and two under another, so that context does not hold across draws (Note N44).

No two of these six contexts are alike, across four phyla and two domains. Two of them place the charge network family in mobile DNA: the *Burkholderia* copy inside a named prophage, and the Myxococcota copy beside an insertion sequence and a retroelement within a toxin-antitoxin array. Those are different vehicles in different phyla, so they are not one acquisition observed twice. The *Saccharothrix* and archaeal copies carry no mobilisation signature and sit compositionally equilibrated with their hosts, which bears on when any movement occurred rather than on whether it did, since mobility machinery and compositional signal both decay with residence time (Notes N33, N44).

A duplication, transfer and loss reconciliation over two species trees infers between 9 and 16 transfers, with a no-transfer model costing 33.1 log units and requiring 41 duplications and 310 losses. No individual transfer edge survives a change of rate regime, so the result supports that transfer occurred and identifies no particular transfer (Note N49).

Sequence identity constrains how recent any of these transfers can have been. The sponge and *Maribacter* copies are 26.1% and 23.9% identical to one another, which is

lower than the identity of either sponge copy to the fungal reference. A recent exchange between these lineages would have left the two copies considerably more similar, so whatever movement occurred was not recent and likely independent.

Three of the ten buried distances separate the family, and all three run to the arginine at 182

All ten distances between the five buried positions were measured on the predicted models. Three of the ten separate the charge network family from the superfamily controls with no overlap between the ranges, and all three run to the arginine at 182, as does the separately measured contact between that arginine and the chromophore site (Note N47).

avGFP carries this family’s residue at two of the five positions. The arginine at 87 and the glutamate at 199 answer to its own Arg96 and Glu222, the pair through which chromophore formation is catalysed, where substitution of either stalls the protein before cyclisation: the time constant rises from 1 h in EGFP to 32 h in E222Q and about 7,500 h in R96M [34]. At 134, 168 and 182 avGFP carries valine, asparagine and leucine. The residue that separates this family from the rest of the superfamily is therefore one avGFP does not have (Note N47).

Discussion

In the previous description of the charge network family [14], 22 complete barrels were reported from a search of protein records built on a ten-member bacterial seed set. The search reported here used sixteen queries matched to the kingdom being searched, against genomic DNA in six frames. Of those 22 records, fourteen are bacterial and seven eukaryotic; of the 77 genomic carrier units here, nine are prokaryotic and 68 eukaryotic. The two recover the same eukaryotic core and not the same bacterial one. Every named eukaryotic relative of their founder reported there is in the census here, and none of the eight bacterial relatives is, no assembly of those taxa having been within the scope of this search (Note N5).

What a search recovers tracks the composition of the sequences it is seeded with. The published eukaryotic records enter at the edge of a bacterial profile’s reach, and on the archaeal locus a bacterial query reads the complete network at 41.6% identity where three fungal queries build nothing. The prokaryotic and archaeal counts here therefore depend on the query panel in a way the fungal and cnidarian counts do not. The boundaries also differ in kind: a protein-record census reports only what a gene caller has already emitted, which is why the eight lineages added here and the fungal pseudogenes are inaccessible to one, while a six-frame search reaches similarity no gene caller has emitted a protein for.

In animals a single arrival before the sponge-cnidarian divergence followed by vertical descent accounts for the topology under both substitution models, with separate sponge arrivals, hexacoral monophyly, and a sponge grouping with the hexacorals each disfavoured. The animal copies are also more similar to one another than any of them is to a prokaryotic copy: across the 33 animal loci the pairwise identities run from 50.7% upward, while the highest identity between an animal locus and a prokaryotic one is 45.2%. In fungi the gene is being lost repeatedly within one section of one genus while remaining under purifying selection where it survives. Both lineages show inheritance with loss. Why the two vertically inherited groups are a sponge-cnidarian one and a fungal one is not resolved by these data; one transfer in either direction, followed by divergence under continued selection, could account for it.

The gene is being lost in fungi and in cnidarians, it is held under purifying selection in the copies that survive, and in one anthozoan lineage it has been amplified. A gene may be dispensable in one lineage and maintained in another, and the dN/dS median indicates that amino-acid changes are removed where it survives. The two lineages are not losing the same thing. In the black aspergilli the loss runs to absence, most of the genomes returning no locus at all. Much of the cnidarian decay sits in lineages carrying several copies, where a lesioned copy need not cost the organism whatever the protein does, although that does not cover all of it: in five cnidarian species the lesioned locus is the only locus recovered (Table S4). Constraint was measured on the surviving copies and not inside the lineages carrying lesions.

The transcript evidence and the selection estimate do not sit together straightforwardly. The intact fungal copies returned no reads across eight libraries from four studies, and no acceptable hit came from 203 fungal transcriptome sets, while across the same family the cross-species locus pairs give a median dN/dS of 0.240 per locus pair, and 0.146 per species pair. A gene that is never transcribed is under no selection to retain its reading frame.

The two measurements are drawn from different samples. The transcript evidence comes from one species, *Aspergillus luchuensis*, with whatever conditions the public sets happen to cover, while the selection estimate spans 24 species across two phyla, so selection acting in species whose expression has not been tested is consistent with both. That the loci were ever genes does not rest on transcript in any case: the lichen pair carries eighteen synonymous differences and none non-synonymous across its coding sequence, against substantial divergence in its introns and co-linear flank, a pattern untranslated sequence does not produce. Transcription and an intact reading frame are not coupled here either, a lesioned *Lepraria* locus being transcribed while the intact loci are not. Expression of a representative copy across varied conditions is the measurement that would settle the state of the maintained copies.

The prokaryotic occurrences are not resolved to that standard, and the pattern they form is difficult to produce by inheritance: two phyla out of the 23,562 bacterial

reference genomes, confinement within one species to the assemblies carrying a single prophage insertion, two of the six loci with a context sitting in DNA built to move, and a reconciliation that forbids transfer costing 33.1 log units. Transfer into the prokaryotic lineages is the simplest account of those together, offered here as a hypothesis the data suggest rather than a result they establish.

Two of those contexts also bear on the published description, which reports its founder locus beside phage-associated and toxin-antitoxin genes and finds no significant enrichment for mobile-element proximity in its own matched controls. For that founder locus geNomad, VirSorter2 and CheckV each call no provirus. *Burkholderia* appears nowhere in that work, and there the same comparison comes out positive, with the element identifiable as a phiE-type, P2-family prophage. The Myxococcota copy adds a vehicle of a different class in a phylum those 22 records do not reach. On toxin-antitoxin features the *Burkholderia* carriers agree with the published null.

The trees do not independently support multiple prokaryotic origins. Prokaryotic monophyly is rejected by no test on the 77-tip alignment, no donor-to-recipient edge is inferred consistently when the rates are re-estimated, and the edges weighted highest fall within Cnidaria, the lineage the topology tests account for without transfer, so the reconciliation establishes that transfer occurred somewhere and locates none of it (Note N49). The *Saccharothrix* and archaeal copies are compositionally equilibrated with their hosts, which puts any movement far enough back for the signal to have decayed. The eukaryotic carriers have no matched context comparison of their own, most of the loci with a context window returning no usable annotation, so the contrast between the two halves rests on the trees and the occupancy, with no context comparison made the same way in both.

The published reconstruction places methionine at the charge network family’s internal node at posterior 0.999 [14]; the reconstruction here returns the same residue at 0.58 to 0.63, and most of the disagreement between the two accounts sits in that distance. Both reconstructions rest on the same tips, none of which carries tyrosine at this column, so the comparison is of two readings of one vertex, and which residue it reads is not settled by these data. A permutation null puts methionine far below the observed posterior, and no complete-network locus recovered from a genome reads methionine. Ancestral reconstruction has been applied to this superfamily with a stronger instrument, coral pigments having been reconstructed and expressed [28, 29]. What is reconstructed here is one column, and resolving the ancestry will need the barrel’s variants mapped across more lineages than are currently sequenced.

The two residues that catalyse chromophore formation in avGFP [34] are the two this work reads at positions 87 and 199, so their presence follows from the definition; their conservation across the wider barrel family does not. Over the 1,929 sequences on which per-site rates were measured, the arginine at 87 lies at the 2.8th percentile of substitution rate among 253 aligned positions and the glutamate at 199 at the 13th,

while the chromophore position is ordinary. It is the residue at the chromophore site, rather than the absence of the catalytic pair, that separates these proteins from a maturing fluorescent protein.

No copy has been expressed, so the consequence of that substitution cannot be measured here. For an aliphatic residue non-fluorescence is supported: the Y66L variant of EGFP is colourless, and its crystal structure traps an oxidised intermediate in which the backbone has cyclised while the leaving group remains attached [37]. Valine, the most common residue recorded here, is aliphatic. For phenylalanine the inference does not hold. In the chromoprotein Rtms5 a phenylalanine at the equivalent position gives green emission at 508 nm, and a second substitution raises the quantum yield from 0.11 to 0.75, both variants crystallised with a mature chromophore [38]. Every phenylalanine locus is cnidarian or poriferan, so if any copy in the charge network family is fluorescent it is an animal one. *Acropora tenuis* is the candidate for expression, its tripeptide Ser-Phe-Gly differing from the Ser-Tyr-Gly of wild-type avGFP at the chromophore-forming residue alone.

The published description [14] treats the five positions as a single functional unit, and the evidence for that is substantial: all five are invariant across its 22 sequences, and on the per-site rates measured here all five lie within the slowest part of the protein. The measurements here point at one of them. The distances separate this family from the superfamily only at pairs involving the arginine at 182; the dead copies lose 182 and 199 first; and that arginine and the chromophore substitution occur together at every complete-network locus but one, an *Edwardsia elegans* locus carrying the complete buried network together with the tyrosine the fluorescent proteins and chromoproteins carry at the chromophore site. None of the three is decisive, and a residue may be essential to a protein without being diagnostic of its family. Substituting that arginine alone in an expressed protein would test it.

Several limits bound what is reported. Every functional statement here is an inference from sequence and from predicted models, and every distance reported was measured on a model. The predictors agree on the backbone and disagree on side chains, ESMFold returning a different fold at a TM-score of 0.34 to its own AlphaFold2 model of the bacterial reference, so no claim here rests on a single side-chain distance.

The fungal losses are counted in one section of one genus, and the cnidarian counts are floors, loci spanning a tandem array having been merged. Pairwise alignment to the reference is unreliable at or below 29.8% identity, where eight of the 95 complete-network loci whose reference identity can be read fall; each of those is read instead through the kingdom-matched query that found it, at a higher identity. Fourteen of the fifteen searches predate the download guards that detect truncated assemblies, so an absence is a limit on what was searched, and the two searches underpinning the zero results argued from here were rerun under the guards and recovered their full target sets. Where a lineage or an association is described here as not previously reported,

the basis is that searches of PubMed, bioRxiv, Crossref and the open web returned no prior report.

No sequence in the family carries tyrosine at the chromophore column, so the ancestral state reconstructed at the family’s own vertex follows from the tips below it and is set aside (Note N46). On the trees, the approximately unbiased test rejects its own unconstrained topology at p-AU 0.00287 on the 77-tip run searched under the mixture model by the posterior mean site frequency approximation, which bounds what a p-AU from that run can be read to mean; the three_class_tree_16 and LG+G4 runs on 77 tips do not reject their own unconstrained topologies (Table S7). How the five positions are read, their numbering and the reference sequence are all inherited from a single unrefereed version 1 preprint [14], and a revision of it would move the carrier counts here with it (Note N5).

Constraint in one lineage alongside loss in another is most simply read as the protein being used in some cells and not in others. Two experiments would answer more than further sequencing. Expressing the Acropora tenuis copy, whose tripeptide differs from wild-type avGFP at the chromophore-forming residue alone, would show whether any copy in this family matures a chromophore. Substituting the arginine at 182 alone would test whether it carries the distinction the distances point to.

Methods

Sequence sources

Eight fungal entries seeded the search, with provenance in the repository. Two have since been removed from UniProtKB as not part of a reference proteome and were recovered from UniParc, byte-identical to the NCBI records, which are the identifiers to cite. The eight are not independent: two are identical over their shared length, five sit at 93% identity or above, and de-replication at 90% leaves four lineages. Two of the six active entries are translations of broken genes, one with an in-frame stop at the codon for the glutamate at 199 and one a four-exon model spliced around a stop and frameshifts; every gene model downstream is therefore re-derived from genomic DNA.

Databases: UniProtKB [39] at PF01353 and IPR011584, UniRef50, InterPro [32] and FPbase [40]; the AlphaFold Protein Structure Database; NCBI Datasets [41], RefSeq viral, the Gut Phage Database and the MGnify genome catalogues [42]; NCBI nt, TSA and dbEST with the ENA read archives.

Network positions and their reading

Positions are numbered against A0A951HUW2, the founder sequence of the previously published description [14], which also specifies the five buried positions. That accession has been removed from UniProtKB; the identifiers are UniParc UPI001DADDAB1 and NCBI MBV9931914.1, a 218-residue protein on contig

JAFAZW010000077.1. The three numbering schemes in use are set side by side in Table S12.

Positions were transferred from the reference into each of the sixteen queries by pairwise alignment and checked against a three-dimensional superposition onto the reference, 30 assignments of 30. Each candidate was then aligned to the kingdom-matched query that recovered it, so a locus is read at its identity to that query rather than to the reference; Table S1 carries both identities for every locus.

A position is satisfied by the reference residue, by another residue of the same formal charge at physiological pH, or by the amide of the reference acid: lysine for arginine at 87 and 182, aspartate or glutamine for glutamate at 134 and 199. Tyrosine at 168 is strict and histidine is refused throughout. The rule is applied in a single function.

Reference dependence was tested by repeating the read under four further references, one fungal, one cnidarian, one alveolate and a second bacterium, over the 206 of 396 loci carrying a protein sequence on disk (Notes N52 and N53; Tables S10, S11 and S14). No change of reference was adopted.

Protein-level searching

MMseqs2 [43] clustering over barrel-range sequences of 180 to 300 residues at 95, 70, 50 and 30% identity. A profile hidden Markov model built from the charge network family, plus jackhmmer [44] against UniProtKB, monitored for drift into engineered fluorescent proteins. Single-seed searches use phmmer [45] at the EBI HMMER API with identical thresholds for every seed (E = domE = incE = incdomE = 1e-5). Structural searching used Foldseek [46] against the AlphaFold database, afdb50 and MGnify mgnify_esm30, with a transitive structural closure promoting a sequence on the network readout rather than on E-value.

Six-frame genomic searching

193,342 of the 193,368 listed assemblies were searched. The lists were retrieved from NCBI Datasets between 29 September and 12 October 2026, each search taking its list on the date in the run log committed beside it: 5 October for the cnidarian and poriferan lists, 6 October for the fungal, broad bacterial and Burkholderiaceae lists, 7 October for the archaeal and Myxococcota lists, and 8 October for the metagenome-assembled genomes. The environmental survey used nineteen MGnify genome catalogues at the versions named in the repository, among them Unified Human Gastrointestinal Genome v2.0.2, Marine v2.0 and Soil v1.0. The list is the union of 68 NCBI Datasets queries grouping into the 17 named searches of Table S1, one search being one taxonomic target searched with one query panel. The searches were built to different inclusion criteria, so each denominator bounds its own search and the counts in the Results are per search unless a list is named. The 26 unsearched assemblies are itemised in Table S13: two failed to download, 21 failed at database construction or

indexing, and three built databases covering 0.8 to 70.2% of their expected bases. The two download failures were retrieved later and searched at both settings of the cap test, indexing 350 contigs each and returning no hit.

The query panel was sixteen sequences, matched to the kingdom being searched and never pooled: four fungal, six bacterial, one archaeal (GCF_965223225.1, the archaeal carrier's own assembly), one animal (A0A7U3V473, Scleractinia), three dinoflagellate (A0A6V0IBN8 from *Zooxanthella nutricula* and two transcript records) and one of unresolved domain. The two broken-gene models were not used as queries. The panel was widened from seven queries to fourteen and then to sixteen as loci were added; every figure reported here is from the sixteen-query search, with the seven- and fourteen-query runs retained as baselines.

Search was tblastn from BLAST+ 2.12.0 [47] over six frames against genomic DNA, -evalue 1e-5 -max_target_seqs 20. The subject cap was tested at -max_target_seqs 500 over the 2,348 assemblies in which a change could move a reported number, panel and parameters otherwise fixed: the output is identical at both settings, no query returned as many as twenty distinct subjects, and no locus was gained or lost (Table S9).

Hits within 10 kb were merged into one locus and each locus re-modelled with exonerate [48] --model protein2genome over the locus plus 2 kb either side, so exon count, splice sites, in-frame stops, frameshifts and network state are read from one object. Where exonerate builds no model the locus is recorded as such, counted separately from a failure.

GenBank and RefSeq copies of one assembly were collapsed on the numeric stem of the accession, keeping the higher version and preferring RefSeq at equal version, so every assembly count here is of distinct assemblies. The Burkholderiaceae and Myxococcota searches nest inside the bacterial one and the control search shares genomes with the fungal one, so 447 assemblies were searched twice and none three times, 30 of them also held under both a GenBank and a RefSeq accession.

Controls on the search

A control task re-searches the 58 assemblies whose 60 loci the protein-level survey had found and fails if it does not recover them; 60 of 60 were recovered with no unreported loci. That control cannot measure a miss, so misses were characterised separately by re-searching every assembly that disagreed with another assembly of the same species.

Separating host sequence from passengers

Sponge and coral assemblies carry symbiont and dietary DNA, so five checks were applied to every locus:

1. contig length, and the locus's rank and position in the assembly;

2. GC at the locus against both the assembly mean and a 100 kb window;
3. the 20 kb either side, scanned for bacterial or fungal content;
4. N content of the 40 kb neighbourhood;
5. the GC profile along the whole contig.

A transferred gene reads as a local GC island with host-like sequence either side, while a foreign contig is flat. A locus flagged on contig length alone is reported as unplaceable, a verdict distinct from passenger. A separate construct screen aligns every protein and reading-frame translation held for a locus against an FPbase snapshot with blastp [47], flagging the locus where the match reaches 90% identity over at least 50 aligned residues; it reaches the 361 of 396 called loci carrying a frame of 50 residues or more to align (Note N54).

Transcript evidence

Eight public runs from four studies, four RNA-seq and four CAGE, were aligned with HISAT2 [49] to the *Aspergillus luchuensis* mut. kawachii IFO 4308 genome and the gene at that locus read out for gene-body reads, start-site tags, splice junctions and antisense signal, each library’s strandedness calibrated independently and every gene in the surrounding 45 kb read out alongside it. The six active candidates were searched by tblastn against NCBI fungal TSA, dbEST and the mRNA records in nt, every hit screened for laboratory constructs; the web TSA search reaches none of the modern TSA sets, so the search was repeated locally against every fungal TSA set in ENA. Lesions called from an assembly were re-checked against the reads of the same BioSample, mapped with minimap2 [50] and read out as a consensus with samtools 1.19.2 [51]. Intactness was re-checked for every copy with exonerate 2.4.0.

Phylogenetics

Alignment with MAFFT v7.505 [52]. Trees with IQ-TREE [53] under LG+G4 [54] and LG+C20+F+G [55], models selected by ModelFinder [56], with 1,000 ultrafast bootstrap replicates [57] and the Shimodaira-Hasegawa-like approximate likelihood ratio test [58, 59]. Competing arrangements were tested as topology constraints by the approximately unbiased test [60] with 10,000 replicates under resampling of estimated log-likelihoods, each constraint searched from two starting trees.

Copies identical over the alignment’s columns were collapsed to one tip with the representative named. A locus entered a tree only where its model covered at least 60% of avGFP’s 238 aligned columns, so the family side of a tree is smaller than the number of loci: the same 52 family sequences carry the 71-tip and 77-tip trees, twelve of them panel sequences standing for carriers with no protein from a searched assembly. The 1,002-sequence alignment applies no length or coverage floor, admitting a sequence on the barrel screen instead, and its family side is 89 tips over 63 loci and 26 panel sequences. Clade labels come from NCBI taxonomy by taxid and never from a description string.

Rooting used four methods: an outgroup of canonical fluorescent proteins verified by residue and not by annotation, minimal ancestor deviation [61], MinVar [62], and a non-reversible model with rootstrap support [63]. Ancestral states were reconstructed under both substitution models with the full posterior reported at each node. The duplication, transfer and loss reconciliation used GeneRax [64] against two species trees built by hand, and gene presence was reconstructed under Dollo parsimony [65]. The loss analysis maps each copy’s state onto an *Aspergillus* subgenus Circumdati tree built from calmodulin, beta-tubulin and RPB2, with strain-level relationships in section *Nigri* from the same three markers, so that a shared lesion can be tested for inheritance.

Structure and geometry

Superpositions were made against reference A0A1M3T1J2, in which the five network positions read 87, 137, 170, 184 and 202 and the chromophore position reads 58. Distances are intramolecular closest approaches between side chains and so do not depend on the superposition frame, which decides only which residue in a target answers to each reference position; the charge network family and the control sets were mapped against the same reference. Predicted models, built with AlphaFold2 [66] through ColabFold [67] and with ESMFold [68], were compared against crystal structures over the superfamily proteins holding both, taking the best-resolution X-ray entry for each accession.

Statistics

Synonymous and non-synonymous divergence was estimated by the method of Nei and Gojobori [36] with the Jukes-Cantor correction [35]. Counts were compared by Fisher’s exact test, two-sided throughout, with odds ratios and 95% confidence intervals where a comparison carries a conclusion, and the unit of independence is stated wherever it affects a result. Several loci in these tables belong to one lineage, so the positional-retention counts carry no significance test, and the fungal against cnidarian lesion comparison is made on loci, its p-value to be read against that dependence. No interval is quoted on the pairwise dN/dS median, the pairs overlapping so that a bootstrap across them would treat one locus as several independent observations. The species-pair median is a different case, the species pair being already the unit, so it carries a cluster bootstrap over those units at 20,000 resamples under seed 20260928. Control-window draws are seeded on the locus, so a re-run cannot move a figure it did not touch.

Reproducibility

Analyses were scripted in Python with Biopython [69]. Every run is captured with its inputs, outputs and environment, and the counts in this paper are recomputed by a test that reads this manuscript and fails when a reported number and the table behind it disagree. It runs against the analysis tables committed in the repository, which are the

sources Supplementary Tables S1 to S14 are drawn from, and never against the workbook; figures quoted here that the workbook does not carry are named in the supplementary notes with the repository file holding them.

Use of AI assistance

The analyses reported here were carried out by the author working with Claude Code instances running Claude Opus 5, and Claude Opus 5.5 on a small number of early runs, which wrote and ran the analysis scripts, queried the sequence databases and drafted the manuscript text under the author's direction. Four failure modes recurred often enough that each is now guarded in code. Reaching for the nearest committed table when a better-powered one sat beside it is the standing one, and `audit_numbers.py` is re-run after any change to the draft because of it. A lookup that succeeds while answering a different question than the one asked is the second, and it is treated as open rather than closed. Reading a sample-level value as an assembly-level state is the third, which `test_provenance_states.py` documents and tests. The fourth is a substituted sequence that leaves a readout the right shape while shifting every position in it, against which `structural_closure.py` refuses to run unless the reference returns 218 residues reading R, E, Y, R and E at positions 87, 134, 168, 182 and 199. One instance is on the record in full: `GCA_003184835.1` and `GCF_003184835.1` are the same assembly and were counted as two loci, which is why every assembly count here is of distinct assemblies after the GenBank and RefSeq collapse. The author verified every reported figure against the committed tables.

Software versions and seeds

Versions are recorded for MAFFT v7.505, BLAST+ 2.12.0, exonerate 2.4.0, samtools 1.19.2 and two IQ-TREE releases. The split between the latter is clean: the non-reversible rootings, root tests, rootstrap and per-site rate tests ran under 2.3.6 and every other tree under 2.0.7, with the version in the header of each output file. For runs predating the version-capture policy the version is not recoverable, which covers MMseqs2, Foldseek, HISAT2, minimap2, GeneRax, ColabFold, AlphaFold2, ESMFold and the Python and Biopython releases. Three FPbase snapshots of different vintages are in use, holding 949, 994 and 992 entries, each named by its own file where it is used so that no percentage here mixes them. Seeds are recorded for the control-window draws, the background and open-reading-frame controls, the UniRef subsampling, and for the ancestral reconstructions and tree searches rerun on the 89-tip and 430-tip sets, which use 20261024. No seed was kept for the original IQ-TREE searches or their bootstrap replicates; the 77-tip unconstrained search spanned 0.28 log units across starting trees.

Data and code availability

Supplementary Tables S1 to S14 accompany this manuscript as Supplementary_Tables_S1_S14.xlsx and Supplementary Notes N1 to N56 as

Supplementary_Information.docx. All sequence accessions, locus tables, alignments, trees, structural models and the scripts that produced every figure and table are held in the project repository, organised by analysis stage, together with the manuscript, its supplementary information, its table workbook and the test that recomputes every number in them. Each supplementary note below names the file holding the result it describes. The repository is not yet public. It will be released at https://github.com/AdamSidaway/charge-network-barrel-search, and until then it is available from the author on request.

Funding

This work received no external funding. It is entirely computational and was carried out by the author.

Competing interests

The author is a director of Cambridge RNA Technologies Ltd and of MorphoKinetics Limited. Neither company relates to the work reported here.

Author contributions

The author is the sole contributor, set the research questions, made every inclusion and exclusion decision including how the network positions are read, performed the analysis, wrote the manuscript and takes responsibility for the content.

References

1. Shimomura O., Johnson F.H. & Saiga Y. (1962) Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan Aequorea. J. Cell. Comp. Physiol. 59:223-239. doi:10.1002/jcp.1030590302.
2. Prasher D.C., Eckenrode V.K., Ward W.W., Prendergast F.G. & Cormier M.J. (1992) Primary structure of the Aequorea victoria green-fluorescent protein. Gene 111:229-233. doi:10.1016/0378-1119(92)90691-h.
3. Ormö M. et al. (1996) Crystal structure of the Aequorea victoria green fluorescent protein. *Science* 273:1392-1395. doi:10.1126/science.273.5280.1392.
4. Yang F., Moss L.G. & Phillips G.N. (1996) The molecular structure of green fluorescent protein. *Nat. Biotechnol.* 14:1246-1251. doi:10.1038/nbt1096-1246.
5. Heim R., Prasher D.C. & Tsien R.Y. (1994) Wavelength mutations and posttranslational autoxidation of green fluorescent protein. Proc. Natl. Acad. Sci. USA 91:12501-12504. doi:10.1073/pnas.91.26.12501.
6. Shagin D.A., Barsova E.V., Yanushevich Y.G., Fradkov A.F., Lukyanov K.A., Labas Y.A., Semenova T.N., Ugalde J.A., Meyers A., Nunez J.M., Widder E.A., Lukyanov

S.A. & Matz M.V. (2004) GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. Mol. Biol. Evol. 21:841-850. doi:10.1093/molbev/msh079.

7. Alieva N.O., Konzen K.A., Field S.F., Meleshkevitch E.A., Hunt M.E., Beltran-Ramirez V., Miller D.J., Wiedenmann J., Salih A. & Matz M.V. (2008) Diversity and evolution of coral fluorescent proteins. PLoS ONE 3:e2680. doi:10.1371/journal.pone.0002680.
8. Labas Y.A., Gurskaya N.G., Yanushevich Y.G., Fradkov A.F., Lukyanov K.A., Lukyanov S.A. & Matz M.V. (2002) Diversity and evolution of the green fluorescent protein family. Proc. Natl. Acad. Sci. USA 99:4256-4261. doi:10.1073/pnas.062552299.
9. Chudakov D.M., Matz M.V., Lukyanov S. & Lukyanov K.A. (2010) Fluorescent proteins and their applications in imaging living cells and tissues. Physiol. Rev. 90:1103-1163. doi:10.1152/physrev.00038.2009.
10. Bulina M.E., Chudakov D.M., Mudrik N.N. & Lukyanov K.A. (2002) Interconversion of Anthozoa GFP-like fluorescent and non-fluorescent proteins by mutagenesis. BMC Biochem. 3:7. doi:10.1186/1471-2091-3-7.
11. Francis W.R., Christianson L.M., Powers M.L., Schnitzler C.E. & Haddock S.H.D. (2016) Non-excitable fluorescent protein orthologs found in ctenophores. BMC Evol. Biol. 16:167. doi:10.1186/s12862-016-0738-5.
12. Gurskaya N.G., Fradkov A.F., Pounkova N.I., Staroverov D.B., Bulina M.E., Yanushevich Y.G., Labas Y.A., Lukyanov S. & Lukyanov K.A. (2003) A colourless green fluorescent protein homologue from the non-fluorescent hydromedusa Aequorea coerulescens and its fluorescent mutants. Biochem. J. 373:403-408. doi:10.1042/bj20021966.
13. Pletneva N.V., Pletnev V.Z., Lukyanov K.A., Gurskaya N.G., Goryacheva E.A., Martynov V.I., Wlodawer A., Dauter Z. & Pletnev S. (2010) Structural evidence for a dehydrated intermediate in green fluorescent protein chromophore biosynthesis. J. Biol. Chem. 285:15978-15984. doi:10.1074/jbc.M109.092320.
14. Schneider T.L. & Zimmer M. (2026) Discovery of a buried charge-network GFP-fold family spanning prokaryotes and eukaryotes. bioRxiv, posted 2 August 2026. doi:10.64898/2026.07.29.741452.
15. Harrison P.M., Echols N. & Gerstein M.B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res. 29:818-830. doi:10.1093/nar/29.3.818.
16. Pei B. et al. (2012) The GENCODE pseudogene resource. Genome Biol. 13:R51. doi:10.1186/gb-2012-13-9-r51.
17. Policarpo M., Baldwin M.W., Casane D. & Salzburger W. (2024) Diversity and evolution of the vertebrate chemoreceptor gene repertoire. Nat. Commun. 15:1421. doi:10.1038/s41467-024-45500-y.

18. Emerling C.A., Teullet S., Allio R., Gatesy J., Springer M.S. & Delsuc F. (2026) Pseudogenes document protracted parallel regression of oral anatomy in myrmecophagous mammals. Mol. Biol. Evol. 43:msag009. doi:10.1093/molbev/msag009.
19. Li G., Zhang Q.J., Zhong J. & Wang Y.Q. (2009) Evolutionary and functional diversity of green fluorescent proteins in cephalochordates. Gene 446:41-49. doi:10.1016/j.gene.2009.07.003.
20. Baumann D., Cook M., Ma L., Mushegian A., Sanders E., Schwartz J. & Yu C.R. (2008) A family of GFP-like proteins with different spectral properties in lancelet Branchiostoma floridae. Biol. Direct 3:28. doi:10.1186/1745-6150-3-28.
21. Yue J.X., Holland N.D., Holland L.Z. & Deheyn D.D. (2016) The evolution of genes encoding for green fluorescent proteins: insights from cephalochordates (amphioxus). Sci. Rep. 6:28350. doi:10.1038/srep28350.
22. Lomsadze A. et al. (2018) Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.* 28:1079-1089. doi:10.1101/gr.230615.117.
23. Kashimoto R., Hisata K., Shinzato C., Satoh N. & Shoguchi E. (2021) Expansion and diversification of fluorescent protein genes in fifteen Acropora species during the evolution of acroporid corals. Genes 12:397. doi:10.3390/genes12030397.
24. Shinzato C., Shoguchi E., Tanaka M. & Satoh N. (2012) Fluorescent protein candidate genes in the coral Acropora digitifera genome. Zool. Sci. 29:260-264. doi:10.2108/zsj.29.260.
25. Kelmanson I.V. & Matz M.V. (2003) Molecular basis and evolutionary origins of color diversity in great star coral Montastraea cavernosa (Scleractinia: Faviida). Mol. Biol. Evol. 20:1125-1133. doi:10.1093/molbev/msg130.
26. Gittins J.R., D'Angelo C., Oswald F., Edwards R.J. & Wiedenmann J. (2015) Fluorescent protein-mediated colour polymorphism in reef corals: multicopy genes extend the adaptation/acclimatization potential to variable light environments. Mol. Ecol. 24:453-465. doi:10.1111/mec.13041.
27. Takahashi-Kariyazono S., Gojobori J., Satta Y., Sakai K. & Terai Y. (2016) *Acropora digitifera* encodes the largest known family of fluorescent proteins that has persisted during the evolution of *Acropora* species. Genome Biol. Evol. 8:3271-3283. doi:10.1093/gbe/evw265.
28. Ugalde J.A., Chang B.S.W. & Matz M.V. (2004) Evolution of coral pigments recreated. Science 305:1433. doi:10.1126/science.1099597.
29. Field S.F. & Matz M.V. (2010) Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. Mol. Biol. Evol. 27:225-233. doi:10.1093/molbev/msp230.

30. Wang H.-C., Minh B.Q., Susko E. & Roger A.J. (2018) Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67:216-235. doi:10.1093/sysbio/syx068.
31. Hopf M., Göhring W., Ries A., Timpl R. & Hohenester E. (2001) Crystal structure and mutational analysis of a perlecan-binding fragment of nidogen-1. Nat. Struct. Biol. 8:634-640. doi:10.1038/89683.
32. Blum M. et al. (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49:D344-D354. doi:10.1093/nar/gkaa977.
33. Ong W.J., Alvarez S., Leroux I.E., Shahid R.S., Samma A.A., Peshkepija P., Morgan A.L., Mulcahy S. & Zimmer M. (2011) Function and structure of GFP-like proteins in the protein data bank. Mol. BioSyst. 7:984-992. doi:10.1039/c1mb05012e.
34. *Sniegowski J.A., Lappe J.W., Patel H.N., Huffman H.A. & Wachter R.M. (2005) Base catalysis of chromophore formation in Arg96 and Glu222 variants of green fluorescent protein. J. Biol. Chem. 280:26248-26255. doi:10.1074/jbc.M412327200.*
35. Jukes T.H. & Cantor C.R. (1969) Evolution of protein molecules. In Munro H.N. (ed.) Mammalian Protein Metabolism, Vol. 3, pp. 21-132. Academic Press, New York.
36. Nei M. & Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426. doi:10.1093/oxfordjournals.molbev.a040410.
37. Rosenow M.A., Huffman H.A., Phail M.E. & Wachter R.M. (2004) The crystal structure of the Y66L variant of green fluorescent protein supports a cyclization-oxidation-dehydration mechanism for chromophore maturation. Biochemistry 43:4464-4472. doi:10.1021/bi0361315.
38. Battad J.M., Traore D.A.K., Byres E., Rossjohn J., Devenish R.J., Olsen S., Wilce M.C.J. & Prescott M. (2012) A green fluorescent protein containing a QFG tri-peptide chromophore: optical properties and X-ray crystal structure. PLoS ONE 7:e47331. doi:10.1371/journal.pone.0047331.
39. Bateman A. et al. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51:D523-D531. doi:10.1093/nar/gkac1052.
40. Lambert T.J. (2019) FPbase: a community-editable fluorescent protein database. *Nat. Methods* 16:277-278. doi:10.1038/s41592-019-0352-8.
41. O'Leary N.A. et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44:D733-D745. doi:10.1093/nar/gkv1189.
42. Richardson L. et al. (2023) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* 51:D753-D759. doi:10.1093/nar/gkac1080.

43. Steinegger M. & Söding J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35:1026-1028. doi:10.1038/nbt.3988.
44. Johnson L.S., Eddy S.R. & Portugaly E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11:431. doi:10.1186/1471-2105-11-431.
45. Eddy S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi:10.1371/journal.pcbi.1002195.
46. van Kempen M. et al. (2024) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42:243-246. doi:10.1038/s41587-023-01773-0.
47. Camacho C. et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi:10.1186/1471-2105-10-421.
48. Slater G.S.C. & Birney E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. doi:10.1186/1471-2105-6-31.
49. Kim D. et al. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37:907-915. doi:10.1038/s41587-019-0201-4.
50. Li H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094-3100. doi:10.1093/bioinformatics/bty191.
51. Danecek P. et al. (2021) Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008. doi:10.1093/gigascience/giab008.
52. Katoh K. & Standley D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772-780. doi:10.1093/molbev/mst010.
53. Minh B.Q. et al. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37:1530-1534. doi:10.1093/molbev/msaa015.
54. Le S.Q. & Gascuel O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25:1307-1320. doi:10.1093/molbev/msn067.
55. Le S.Q., Gascuel O. & Lartillot N. (2008) Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317-2323. doi:10.1093/bioinformatics/btn445.
56. Kalyaanamoorthy S. et al. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587-589. doi:10.1038/nmeth.4285.
57. Hoang D.T. et al. (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35:518-522. doi:10.1093/molbev/msx281.
58. Guindon S. et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307-321. doi:10.1093/sysbio/syq010.

59. Anisimova M. & Gascuel O. (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55:539-552. doi:10.1080/10635150600755453.
60. Shimodaira H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51:492-508. doi:10.1080/10635150290069913.
61. Tria F.D.K., Landan G. & Dagan T. (2017) Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* 1:0193. doi:10.1038/s41559-017-0193.
62. Mai U., Sayyari E. & Mirarab S. (2017) Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLoS ONE* 12:e0182238. doi:10.1371/journal.pone.0182238.
63. Naser-Khdour S., Minh B.Q. & Lanfear R. (2022) Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. *Syst. Biol.* 71:959-972. doi:10.1093/sysbio/syab067.
64. Morel B. et al. (2020) GeneRax: a tool for species-tree-aware maximum-likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* 37:2763-2774. doi:10.1093/molbev/msaa141.
65. Farris J.S. (1977) Phylogenetic analysis under Dollo's law. *Syst. Zool.* 26:77-88. doi:10.2307/2412867.
66. Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583-589. doi:10.1038/s41586-021-03819-2.
67. Mirdita M. et al. (2022) ColabFold: making protein folding accessible to all. *Nat. Methods* 19:679-682. doi:10.1038/s41592-022-01488-1.
68. Lin Z. et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379:1123-1130. doi:10.1126/science.ade2574.
69. Cock P.J.A. et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422-1423. doi:10.1093/bioinformatics/btp163.

Numbered in order of first appearance. Every entry is cited in the manuscript or in the supplementary notes.

Supplementary material

Supplementary Notes N1 to N56 are supplied as Supplementary_Information.docx. Supplementary Tables S1 to S14 are supplied as the workbook Supplementary_Tables_S1_S14.xlsx, with one sheet per table and a README naming the analysis output behind each.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SidawayGFPbarrelcensussupplementaryinformation1.pdf
- SidawayGFPbarrelcensussupplementarytables.xlsx