Supplementary 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

Cambridge RNA Technologies Ltd and MorphoKinetics Limited, B25 Bioinnovation Centre, Cambridge Science Park, Cambridge CB4 0FW, United Kingdom

This document holds Supplementary Notes N1 to N56. Supplementary Tables S1 to S14 are supplied separately as the workbook Supplementary_Tables_S1_S14.xlsx, with one sheet per table and a README naming the analysis output behind each. Every value in those tables is the output of the named analysis; where the manuscript reports a subset, the flag column naming that subset is given so the reported figure can be reproduced from the sheet.

Table S1. Every locus called by the six-frame search, one row per locus, with accession, organism, coordinates, network read, reading-frame state, exon count and identity. The complete_network and in_chromophore_census columns mark the subsets the Results report on. 396 rows.

Table S2. The 77 distinct carrying units, collapsed from the 132 assemblies that hold a complete-network locus, with taxon, rank, host-or-passenger assignment and the accessions behind each. 77 rows.

Table S3. The fungal candidate loci by reading-frame state, the set behind the fungal reading-frame counts reported in the Results. 144 rows.

Table S4. Lesioned loci put to the reads of their own BioSample, with platform, layout, mean and minimum depth over the gene, and the lesion called from the assembly and from the read consensus. 71 rows.

Table S5. Every cross-species locus pair attempted for dN/dS. The in_final_analysis column marks the 145 pairs behind the reported median, and exclusion_reason gives the ground for each of the rest, including the sixteen pairs on the assembly that holds two fungi. 610 rows.

Table S6. The chromophore-equivalent residue at every locus reading the complete network, with the panel controls flagged so the 131-locus census is reproducible from the sheet. 140 rows.

Table S7. Approximately unbiased tests across the eight topology sets built on the 53, 71, 75, 77 and 85-tip alignments, one row per arrangement, with log likelihood, the difference from the best tree, the p-value and whether the arrangement is rejected. The 77-tip prokaryotic set appears three times, under LG+G4, scored under LG+C20+F+G on those topologies, and searched under it by PMSF. 58 rows

across eight topology sets. The 430-taxon tests quoted in the Results are not in this sheet and are held in the repository with the run that produced them.

Table S8. Mobile-element counts per 40 kb window at every carrier locus and its matched controls, by feature class, across all eleven sweeps and not only the prokaryotic ones. 259 rows.

Table S9. Every assembly searched at -max_target_seqs 20 and again at 500, with HSP, locus and subject counts at both settings, the most subjects any single query returned, and whether the two outputs are identical. 2,348 rows.

Note N1. What the choice of frame-calling rule reaches, and what it cannot reach. Where a locus has no spliced model, its network is read by translating all six frames and keeping one, and two rules for choosing that frame were compared, one scoring a frame by identity to the reference multiplied by aligned columns and one taking the longest open reading frame. The two pick a different frame at 55 of the 190 loci carrying more than one candidate frame, and 53 of those 55 are among the 396 called loci. None of the 53 reads a complete network and none is a carrier unit, so the census, the 140 complete-network loci and the 77 carrying units are identical under either rule. Reading-frame state is not called from this choice at all: a stop or a frameshift is called against a spliced model, and all 82 lesioned loci carry one. Two of the 36 lesioned fungal loci are among the 53, and in both the lesion is a frameshift read from the model while the six-frame read reaches one and three of the five positions respectively; neither reads the complete network under either rule. Sixteen of the 66 fungal loci of unassigned reading-frame state are among the 53, which is where a frame choice can move a network string without changing how the locus is counted. results/20_genomic_tblastn/orf_network/.

Note N2. The founder's nr hit list at the published parameters, 51 hits across 36 organisms. The list holds 51 hits across 36 distinct organisms, with Mucilaginibacter at 46.4%, a Bacteroidota record at 44.9% and Winogradskyella at 43.0%, matching the three identities named in the original description to the decimal. The identity ranking in that list is not a ranking of relatedness, for the reason given in Note N5.

Note N3. The nearest relative of reference sequence 18, at 32.9 to 69.4% identity. The sequence is the barrel on the *Chamaecostus congestiflorus* carrier contig, whose provenance resolves as bacterial and unnamed. Its nearest relative among the sequences compared here is MEO7688339, a *Sphingomonas* sp. metagenome-assembled genome from Antarctic soil, at 69.4% identity over 219 residues with 99% of the query covered, and the two are a reciprocal best pair. That figure is 36.5 points above the 32.9% recorded against A0A931S450, which was the closest comparison available before MEO7688339 entered the set. The range in the title is therefore the effect of the comparison set and not a measure of uncertainty, and the provenance is no better resolved for it, since a nearest relative within a 28-sequence comparison set is not a nearest relative in the databases.

Note N4. Pontellid copepod records are fold-level matches and not candidates for the family. Five Pontella mimocerami records and one Labidocera aestiva, at 26 to 27% identity over about 172 aligned residues.

Note N5. Eleven of the 22 previously reported barrels cannot be identified from the publication. The criterion is identification by accession. Three accessions appear in their main text and eight in the per-member provenance of their Supplementary Table 2, giving eleven. Of the other eleven, three are cultured isolates named by organism alone, which locates the organism but not the sequence, and eight are named only by taxon as representatives of the seed profile. Their Supplementary Table 1, which the main text cites for the full membership, is absent from the supplementary file as supplied. Two of the 22 resolve to one gene, at 97.6% identity over 253 aligned columns, so the 22 is probably 21. The eight named only by taxon were assigned accessions against the reproduced search of their founder against nr at their published parameters, which returns 51 hits. Three are settled by the identity their text quotes, matching to the decimal place quoted, at 46.377, 44.928 and 42.995%, and each is also the only hit of its taxon. Four are the only hit of their taxon. The Gaiellaceae entry had two candidates and is settled by the order their own sentence lists its further hits in, which reproduces as descending bit score and not as descending identity. Each of the eight resolves to a single assembly record, and neither those assemblies nor any other genome assigned to those taxa appears anywhere in the assembly lists of this work. That check was made against every row of all 68 lists, 194,862 distinct accessions before the searched filter and the GenBank-RefSeq collapse are applied, which is deliberately wider than the 193,342 searched assemblies the Results report, because the question is whether the organism was ever listed rather than whether it was searched. Two things follow from their list. It orders by bit score while comparing by identity, and on bit score their Acropora hit at 170.2 sits above the Winogradskyella hit at 169.9. This is why the comparison in the main text is made on what each method reaches rather than on totals. Eight of the nine assemblies behind those taxa are metagenome-assembled genomes from soil, seepage and marine metagenomes. The comparison rests on a single unrefereed version 1 preprint, so the sentences in the main text that describe it would change with it.

Note N6. GCA_019827475.1 carries sequence from two fungi, which bears on how its loci are counted. The assembly is submitted as *Penicillium* cf. *chrysogenum* strain S/N-308-OC-P1 and holds 68,790,985 bp in 1,424 scaffolds. Current assemblies of *Penicillium chrysogenum* have a median of 31.5 Mb and those of *Aspergillus niger* 35.7 Mb, so this assembly is larger than the largest of either species and close to their sum. Its single barrel locus, 253 residues reading REYRE on an intact frame on JACVQR010000162.1, is 99.6% identical at the protein level to the *Aspergillus niger* locus across 15 kb of flanking sequence diverging by 0.17%. Those are within-species values and not the divergence expected between two genera separated by on the order of a hundred million years, so the assembly is a co-assembly of two fungi and not a mislabelled or singly contaminated one. The consequence for the constraint sweep is that a pair drawing one partner from this assembly compares two *Aspergillus*

loci under two species names. Every such pair is therefore excluded: 33 were attempted and 16 were informative, and removing those 16 takes the informative set from 161 to the 145 reported. Table S5 marks them in the in_final_analysis column together with the reason.

Note N7. The 71 lesioned loci against the reads of their own BioSample, per locus. Table S4 carries the outcome, the mean and minimum depth over the gene and the lesion called from each of the assembly and the read consensus, per locus. These are 71 of the 82 loci called with a lesion; the other eleven either share an assembly with a locus that was checked or fall outside the set this comparison was run over. Four loci reproduce part of their lesion with the remainder uncovered, and the table reports them separately from outright agreements.

Note N8. The three wide-span loci are an extraction fault and not a property of the loci. Each is a single two-exon copy whose span is an intron of 1.0 to 6.3 kb, and two of the three carry internal stops.

Note N9. The *Lepraria* lesion is present in its own reads at 0.7-fold mean depth. The reads cover over 39% of the gene at 0.7-fold mean depth and carry the lesion. It is one of the five runs below 5-fold mean depth (Note N38).

Note N10. The cnidarian tandem arrays are underpowered at the available read depth. The arrays are underpowered at the available depth and no call is returned for them.

Note N11. Four measurements locate the methionine disagreement in the circularity at the family node. Four independent measurements bear on the methionine reconstructed at the family's internal node in the original description. They are set out in Notes N12 to N15, and together they locate the disagreement in the circularity of the tip selection rather than in a misfitting model.

Note N12. The two substitution models agree at the chromophore column and at the arginine at 182. On one topology the two models agree at both of those columns and disagree about position 168. They agree on tyrosine at the chromophore column at the vertex outside the family, where both read valine at 182, and on arginine at 182 at the family's own vertex, where both read methionine at the chromophore column. No vertex reconstructs tyrosine and that arginine together, which is the point Note N13 makes of the extant tips.

Note N13. No extant sequence orders the chromophore substitution against the arginine at 182. None of the 422 readable tips carries tyrosine at the chromophore together with arginine at 182. Three of the five positions are near-static across the same tips.

Note N14. All 432 reconstructed states are identical on an independently searched topology. Re-run on an independently searched topology, all 432 called states are identical. Confidences move where the two trees disagree.

Note N15. The chromophore column's states moved when the taxon sampling was rebuilt. On the earlier 44-tip family set valine and phenylalanine were effectively tied at 20 and 21 tips, with threonine at one. On the rebuilt 89-tip set they read 44, 26 and 12. The 0.756 phenylalanine posterior was computed over the 44-tip sampling.

Note N16. The family is monophyletic at bootstrap 100 under the mixture model's own searched topology. Searched under LG+C20+F+G, the family is monophyletic at ultrafast bootstrap 100. Fixing the topology instead costs the mixture 9.80 log units.

Note N17. What the de-replicated set holds, and the clustering rule behind it. The set is 430 tips: 186 of the 607 canonical fluorescent proteins, 127 of the 278 entries with no measured quantum yield, the 89 family tips, 21 engineered-dark sequences and seven non-fluorescent ones. Clustering is greedy centroid at 95% identity and small classes are kept whole, so the threshold means what it says.

Note N18. W152 is invariant across the sequences read here and S67 is not, although both are invariant across the published 22. Neither position had been read by this project before the published description was examined, and reading both across the 26 sequences then available separates them. W152 is invariant at 26 of 26, across sequences spanning archaea, alveolates, sponges and unresolved bacteria, which are lineages absent from the published 22. S67 is not, reading 21 of 26 here against 22 of 22 there. P64, which the published description treats as environment rather than as invariant, is 26 of 26. Neither W152 nor P64 enters the membership criterion used here, which gates on the five buried positions alone.

Note N19. The glycine scaffold is numbered two ways in the original description, and both are read. Glycine 62, the residue that performs the cyclisation reaction in fluorescent proteins, is retained in a family that has no chromophore to cyclise, which makes its retention a property of the barrel structure.

Note N20. MEO7688339 returns 13 strands in four of four ColabFold models at pLDDT 79. Four ColabFold models return 13 strands in every one, 51 to 54% beta content and pLDDT 79.

Note N21. W152 and P64 were tested as candidates for the definition and rejected. W152 and P64 are retained at 93% in intact copies, which makes them as invariant among sequences as the five positions on which the definition is built, and on sequence alone this would constitute a reason to include them. In dead copies they behave in the manner of the structural group rather than in the manner of 182 and 199. A protein therefore retains them after the gene has ceased to function, and they cannot distinguish this family from any other sequence carrying this barrel. A position invariant in intact copies and variable in dead copies was treated as a candidate, and a position invariant in both was treated as a property of the structure.

Note N22. No tyrosine-retaining sequence in the superfamily carries arginine at 182. Across 994 readable sequences of the superfamily alignment, none of the 823 sequences that retain tyrosine at

the chromophore carries arginine at 182. The two therefore do not occur apart anywhere in that alignment. They do occur apart in the locus census: one of the two intact Edwardsia elegans loci on GCA_042464825.1 reads the complete network, and so arginine at 182, together with tyrosine at the chromophore site (Note N41). A protein carrying both would be excluded from the family on the basis of its tyrosine but would nevertheless be present in this alignment and would be counted, so the absence is not an artefact of the membership rule. No order between the two changes follows from this observation, since the database was assembled by searching for one of them.

Note N23. Reading the five positions from a sequence alignment does not recover the family. Of 6,336 UniProt proteins assigned to this barrel superfamily, 722 fall within the length range occupied by the family, and 30 of these carry the complete arrangement when the positions are read from the alignment. When the positions were reassigned by structural superposition, which is the frame in which the geometry is defined, the 30 were reduced to seven, and those seven are the family. The alignment had been pairing the second arginine of the family with an arginine or lysine two to three residues along, which most proteins in this superfamily carry. Of the 23 sequences admitted by the alignment and rejected by the superposition, seventeen are laboratory fluorescent proteins at 0.929 to 1.000 identity to deposited constructs and the remainder are accounted for. A reader implementing the criterion on an alignment will therefore recover the superfamily rather than the family. Across the sequences satisfying the other four positions, the residues occurring at the fifth that cannot exchange a proton were counted: isoleucine in eight sequences, histidine in five, threonine and alanine in four or five each, and seven further residues in ones and twos. Aspartate, which occurs in four or five, is not among them: it is a carboxylate and the criterion admits it at this position. Locating the fifth position in a distant sequence gives three reads that disagree. Under the pairwise alignment with which the locus tables were produced, two of the panel sequences enter, MGYG000455962 reading aspartate and a *Caballeronia* isolate reading glutamine; under either structural superposition or the 1,002-sequence multiple alignment, three sequences read glutamine and MGYG000455962 reads the reference glutamate.

How far the pairwise read carries the count was measured against structural superposition over 194 committed models. The two rules disagree below about 30% identity to the reference and essentially not above it: of 110 readable models at or below 29.8% they name different residues in 97, and of the 30 above it, in one. A sample drawn before any structure was predicted tests that directly. Forty-eight loci were drawn ten per identity band, 45 were modelled and read under both rules, and three of the 28 that read five of five by pairwise alignment do not under superposition. One of the three is a contradiction, at 26.6% identity where the pairwise read gives glutamate at position 199 and the superposition threonine; in the other two the superposition could not place position 168 at all, at TM-scores of 0.884 and 0.870, where the barrel is confidently superposed.

Note N24. The two controls separating a host gene from a passenger. Two controls were performed in order to separate a host gene from sequence belonging to another organism within the same

assembly. In the first, the protein behind every locus was recovered rather than only those considered suspicious, and foreign DNA was identified within one eukaryotic assembly, which is the Petrosia ficiformis locus described below. In the second, base composition at each locus was compared against its own contig and against a 100 kb window, and the two sponge records reported above were found to lie within chromosome-scale sequence at host base composition. Of the 77 units, 45 were placed within the stated organism, four were identified as passengers within assemblies of other labelled organisms, 23 were located on contigs too short to resolve, and five were not reached by any of the checks.

Note N25. The environmental metagenome survey. Environmental sequence was searched separately. Of 56,757 metagenome-assembled genomes in MGnify, eleven returned a locus and two of those are members of the family. A near-absence across environmental bins is a statement about what has been assembled and deposited, and not an estimate of how common the family is in nature, and the per-bin counts are given in the supplementary material.

Note N26. How the wide-span cnidarian loci were re-modelled. Seventeen cnidarian loci were called with a genomic span at least three times the coding length implied by their protein. This pattern is produced by a locus clusterer when adjacent copies are merged, since loci whose spans overlap on one contig are combined. That behaviour is correct for a locus reported twice by two query panels and incorrect for tandem copies lying close enough to overlap. Each span was therefore re-modelled, and all seventeen were resolved. Nine of the seventeen resolved as a single copy, and in three of those nine the wide span is an intron of 1.0 to 6.3 kb while the other six are single-exon models whose span the clusterer had overstated, so a wide span requires re-modelling before a copy number can be called. The remaining eight are tandem arrays. Six species were found to hold 29 copies where the locus table records 12: Plumapathes pennacea eleven, Paragorgia arborea six, Hemicorallium imperiale four, Ricordea florida four, Corallium rubrum two and Paragorgia pacifica two.

Note N27. The grounds for calling the Plumapathes array an amplification. The Plumapathes array was assessed as an amplification rather than an assembly duplication on three grounds that an artefact would not satisfy together. A blastn self-comparison of each of its two regions places the off-diagonal alignments at exact integer multiples of a single period, 2,731 bp in one region and 2,755 bp in the other. The units are 95.1 to 99.5 per cent identical to one another, rather than showing the near-identity expected of a duplicated consensus. Two of the eleven copies carry an internal stop, which an assembly error would not generate in two copies of a sequence written down twice. Within the phylum in which the gene is being lost, one lineage has therefore amplified it, and the amplification carries two pseudogenes of its own. This observation does not reduce the losses described above and does not constitute a competing explanation of them. It indicates that the fate of the gene is not uniform within a phylum, as the transcribed dinoflagellate copies indicate that it is not uniform within the family. Cnidarian copy counts, and the denominator of the cnidarian loss rate, should therefore be treated as floors.

Note N28. The membership criterion, stated in full. Definition of membership. A sequence was included if all five buried positions were read correctly against the reference and the position equivalent to avGFP Tyr66 was not matched (Figures 1 and 2). No member of this family has been expressed, and none is therefore known to carry a chromophore. That position is referred to as the chromophore position throughout on the basis of its location in the barrel rather than any pigment it is known to form.

Note N29. How each of the five positions is satisfied. At each of the five positions, included sequences were required to carry the reference residue, another residue of the same formal charge at physiological pH, or the amide of the reference acid. Histidine was refused at the arginine positions, since partial protonation alters the chemistry within the barrel and does not conserve charge. The fifth position is the only position at which a non-conserving residue is common among sequences satisfying the other four, and the treatment of that position is considered in the Discussion.

Note N30. The fifth position and why it is treated separately. The fifth position is the one position where a non-conserving residue is common across sequences that satisfy the other four, and the Discussion sets out what the criterion does with it.

Note N31. What the unassigned reading-frame state is, and what it is not. That state reflects a property of the data rather than of the gene. Homology at these loci is established from the DNA sequence through the six-frame hit, and what is missing is a protein model over it, not evidence that the locus belongs to the family. In these cases either no spliced model could be constructed, so that homology rests on the DNA alignment alone, or the alignment did not extend to the five positions. Neither case establishes whether the locus is an intact gene. In the locus tables the reading-frame column carries intact for these loci by default, so a frame state is recorded for all 396 loci and determined for the 299 carrying a model; counts of intact genes are taken on the intersection of the frame state with the position read and not on that column alone. Of the 58 prokaryotic loci, 33 carry a model and 25 do not, and none of the 25 produced a protein. One of the 25, MGYG000494800, carries a stop inside the segment its six-frame hit aligns to, which is why exonerate could build nothing across it, so the absence of lesioned prokaryotic copies is reported over the 33 the test reaches.

Note N32. The archaeal record examined before acceptance. Because a barrel located on a contig within a metagenome-assembled genome may belong to a contaminating organism, the archaeal record was examined before it was accepted. The locus is situated on a 1.81 Mb contig that constitutes 97.8% of a four-contig assembly. A 16S ribosomal RNA gene was identified 105 kb from the barrel on the same contig. This gene is 98.4% identical over 1,469 bp to Nitrosopumilus maritimus SCM1, produces no blastn alignment against Escherichia coli or Bacillus cereus 16S, and carries the archaeal start motif TTGATCCTGCCGGA rather than the bacterial motif TTGATCCTGGCTCAG. The nearest domain-diagnostic gene is ribosomal protein S26e at 22 kb, a 30S subunit protein that is archaeal and eukaryal and has no bacterial counterpart, followed by an archaeal RNA polymerase

subunit K at 32 kb. A chimeric join would be required to fall between the barrel and these genes. To test for such a join, a sliding scan of 181 windows of 10 kb was performed, and no window was found to lie more than three standard deviations from the contig mean. These observations indicate that the locus resides within an archaeal genome.

Note N33. What the archaeal locus context does and does not show about route. The route by which the gene reached this genome is not resolved by the genomic context. No signature of recent mobilisation was detected, four mobile-element genes being identified across 1.81 Mb with none within 100 kb of the barrel. Mobility machinery decays over time, so this absence is equally consistent with an ancient transfer. The locus is compositionally equilibrated with its host, at z = +0.48 against 78 windows of its own neighbourhood, which is the pattern expected of long residence. The closest relatives of the protein are nevertheless the bacterial members, at 44.5% and 41.6% identity over 173 residues against 20 to 26% for the eukaryotic members, so presence within an archaeon does not establish vertical inheritance. Two limitations apply to this record. No sequencing run has been deposited for the BioSample, so the read-level test that would exclude a chimeric join outright could not be performed. The coding sequence is an ab initio GeneMarkS-2+ prediction [22], as are 18 of the 21 genes flanking it, so the network read rests on the DNA sequence rather than on the annotation.

Note N34. The six archaeal loci tested and rejected. Six further archaeal loci were recovered by the same sweep and none was accepted as a member. Each of the sixteen panel queries was aligned against each of the six, giving 96 query-locus pairs, and no pair produced a spliced model. The highest score obtained among the six is 46 bits, against a floor of 65 bits for every query that has modelled the locus of the archaeal member itself. Three of the six lie on genes annotated by the depositors, of 216, 218 and 965 codons, each labelled hypothetical protein, and no mobile element was identified nearby. The six carry open reading frames of 226 to 978 residues reading one to three of the five positions, so they fall among the loci too divergent to model rather than among the decayed copies of the family. The kingdom-matched query does not reach five of the six at all.

Note N35. The Chamaecostus plant-assembly locus in full. The third is a plant assembly carrying a bacterial sequence. In Chamaecostus congestiflorus the carrier contig is 1,004 bp within an assembly of 190,262 contigs, and a spliced model of 187 residues in one exon was constructed over it with no internal stop, reading the complete network with threonine at the chromophore position. Base composition places that contig 31.2 orders of magnitude outside the distribution of the assembly it was deposited in. Relatives of the sequence were counted separately in curated and in environmental databases, returning eight entries in UniRef100, of which five are members of this family and none is a plant protein. A barrel from a Zixibacteria bin returns 29 UniRef relatives under the same search, so a low count in the curated set is a property of the sequence rather than of the method.

Note N36. The nearest known relative of the Chamaecostus sequence. The source of the Chamaecostus sequence has since been narrowed to a bacterial family. Its nearest known relative is a

Sphingomonadaceae protein from an Antarctic soil metagenome, at 69.4% identity over 219 residues with the whole query covered, at E = 2.8 x 10⁻¹¹⁶, and the two form a reciprocal best pair. The nearest relative recorded before that sequence became available read 32.9% identity, so the provenance of this contig has moved from unresolved to a named bacterial family.

Note N37. The Petrosia sponge-assembly locus in full. The fourth locus lies within an animal assembly. A Petrosia ficiformis contig of 2,083,518 bp at 65.7% GC within a genome of 54% GC carries a locus reading the complete network over 206 residues in one exon, with threonine at the chromophore position, and base composition is flat across every 50 kb window of that contig. Were the host phylum taken as the test, the sequence would be assigned to the sponge. Its nearest relative is bacterial, at 42.3% identity over 208 residues, the nearest poriferan member is seventh at 27.1%, and the nearest prokaryote exceeds the nearest animal by 80.6 bits. A blastp search against the non-redundant database returns 17 hits, every one of them a member of this family, and a search of Swiss-Prot returns none at E ≤ 10⁻³ over 575,748 sequences, where the reference sequence finds five GFP-like chromoproteins in the same invocation. The locus is recorded as bacterial sequence within a sponge assembly and the bacterium is not identified.

Note N38. The read-validation outcome for every lesioned locus. Forty-four BioSamples were found to have no public reads and could not be assessed, and read mapping failed at two further loci. Of the 25 loci for which reads covered the gene, 20 reproduced the lesion called from the assembly outright and four reproduced part of the lesion with the remainder uncovered. One locus was read differently. In Antipathozoanthus remengesaui the assembly carries two frameshifts and in-frame stops at S113, Y153 and G162, whereas the reads carry the two frameshifts and the stops at Y153 and G162 but not the stop at S113. That locus is lesioned under either reading. No locus for which reads were informative lost its lesion, and the pseudogenes are therefore not an artefact of assembly. Depth over the gene across these 25 runs had a median of 20.8-fold over a range of 0.4 to 58.0, with five runs below 5-fold.

Note N39. Reading frame and network completeness vary together outside the family. Among the loci that do not satisfy the membership criterion, the state of the reading frame and the completeness of the network vary together. Of 46 such loci carrying an in-frame stop, 18 read four of the five positions and one reads none of them, whereas the 87 with an intact frame and a readable network are spread across the same range, 21 reading none, 35 one, 12 two, 11 three and eight four. A stop beside an almost complete network is the pattern expected where the frame has been disrupted recently and the network has not yet accumulated substitutions. Two genera supply 43 of the 82 lesioned loci, Aspergillus with 25 and Cladonia with 18, so these loci are not independent observations and the pattern is reported as a description of them.

Note N40. How the carrier count depends on which assembly is used. Among the eight species with a positive result and more than one assembly, seven were found to give discordant calls between their

own assemblies. Aspergillus niger represents the extreme case, in which seven of 123 assemblies carry an intact copy and the remainder do not. Presence and absence are therefore unstable at the assembly level. A locus can be missing from an assembly because the strain lacks it, because the region fell in a gap, or because a fragmented draft broke the gene across contigs, and a single assembly cannot separate these. A survey based on one assembly per species will under-report this family. The count that rests on these tips was resampled accordingly. Redrawing the assembly at each of the 24 tips whose species holds more than one, without replacement, over 10,000 draws returns five to eleven losses with a median of seven: the published eight occurs in 30.1% of draws and seven is the mode at 35.0%, and no draw returns fewer than five, so loss is never absent from the resampled tree. The remaining six tips are in species with one assembly each and are fixed by their own evidence, all six scored absent. Two scope rules were run, the published clade held fixed and the clade recomputed from each draw’s own carriers, and they give identical distributions, because the carrier count varies from 12 to 16 while the smallest clade containing the section Nigri carriers is the same 28 tips in every draw checked, Aspergillus heteromorphus carrying a lesioned locus outside it. Losses may still be reassigned within that clade.

Note N41. The five complete-network loci departing from the lineage pattern. Within the complete-network set, five loci depart from this pattern, and each is located within a clade whose other sequences do not. Two of these are the lichen alanines. The remaining three are cnidarian. Edwardsia elegans carries two complete-network loci, one reading threonine and one reading tyrosine, and Heteractis aurora carries a 125-residue copy with a stop at which the model deleted the chromophore position. The tyrosine-reading locus is intact and complete at the network and was excluded on the basis of its chromophore residue alone, which indicates that the criterion removes real sequence.

Note N42. Why imposing family monophyly is not an independent test. The monophyly of the family does not constitute a constraint. On the 77-tip tree, imposing it returns a topology 0.0007 log units from the unconstrained tree, which is within the numerical resolution of the likelihood, so that row records the same topology and its p-value is not usable. On the rebuilt 430-taxon tree the relationship is inverted. Imposing monophyly returns a tree 6.67 log units better than any unconstrained search recovered, and neither arrangement is rejected, the two being separated by search quality and not by evidence. The second result is that the absence of a chromophore does not represent a single inheritance. The other chromophore-less proteins in the tree do not form a clade, the family does not group with them, and it is not their sister. All three arrangements are rejected here by 53 orders of magnitude or more, at differences of 497 to 543 log units, and on the rebuilt 430-taxon tree by margins of 2,012 log units and above.

Note N43. The family recovered as monophyletic across the trees built here. Across the trees constructed in this work, the family itself was recovered as monophyletic on five, four constructed on the de-replicated set of 89 tips and one on a family side constructed from carrier assemblies. Support was ultrafast bootstrap 100 on the two trees carrying bootstrap replicates and SH-aLRT 99.2 on the

carrier-based tree. Replacing the family side with sequences assembled in a different manner does not disturb the monophyly. The outgroup used for these trees was a set of barrels verified by residue rather than by annotation, since FPbase records do not constitute a GFP-fold outgroup. Fluorescence has been engineered into at least five structurally unrelated folds, and a set assembled on the phenotype is therefore not a set assembled on the fold.

Note N44. The five prokaryotic loci other than the prophage. The remaining five loci differ from this case and from one another. In Maribacter spongiicola the locus lies between two recombinases 3.8 kb and 0.5 kb away, with an ArdA anti-restriction protein between the upstream recombinase and the gene. ArdA is carried by conjugative elements in order to survive the restriction systems of a new host, and this copy is therefore located within DNA adapted to move between genomes. Because mobile neighbourhoods are common in Bacteroidota, the controls are informative: all five windows on the same contig carry nothing in any class, and the enrichment is therefore local. In Saccharothrix australiensis the locus carries no mobile feature within 20 kb, whereas its own five control windows hold two integrase or recombinase genes, a transposase, an insertion sequence and a phage regulatory gene between them, so that this copy lies in a quieter neighbourhood than the remainder of its own chromosome. The archaeal locus carries a single toxin-antitoxin gene within the window, and its five control windows carry none, which is too thin a comparison to establish whether one exceeds the background of the genome. The two Myxococcota loci carry one and seven toxin-antitoxin genes, and for neither could a control window be placed, so those counts stand without a background.

Note N45. The two members found outside the genomic search. Two further bacterial members were identified by a blastp search of the reference sequence against the non-redundant protein database, which the genomic search does not cover. One is a Sphingomonadaceae sequence of 220 residues from an Antarctic soil metagenome, reading the complete network at the reference residues, with threonine at the chromophore position. The other is a Holophagales sequence of 235 residues from a coral-reef metagenome, reading the complete network at the reference residues, also with threonine, and it is the first Acidobacteriota record for the family. AlphaFold2 returns 12 strands and 52% beta sheet at pLDDT 87 for the Holophagales sequence, with every ranked model a barrel, and 13 strands and 54% beta sheet at pLDDT 79 for the Sphingomonadaceae sequence. Neither has an assembly behind it, so neither enters the carrier counts, and neither is present in any tree reported here.

Note N46. Why the family’s own vertex cannot answer the ancestral-state question. The ancestral vertex of the family cannot address this question under either model. The tips conditioning that vertex were selected by a rule that reads this column and excludes tyrosine, so a low posterior for tyrosine at that vertex follows from the selection rather than from the data. Vertices outside the family are not conditioned in this way and were therefore read instead. Walking outward from the family on the unrooted neighbour graph, tyrosine was recovered as the best state at 34 of 34 deep vertices on the 430-tip de-replicated set, at 41 of 41 on the 1,002-sequence three-class alignment from

which that set was cut and at 34 of 34 on a third tree whose family side was constructed from carrier assemblies rather than panel sequences. This result holds under both substitution models, with no family tip below any of these vertices, and on the 1,002-tip set no deep vertex falls below 0.980 under LG+G4. The two models disagree at a single vertex, located one edge above the stem of the family and nearest the conditioned tips. Re-running the reconstruction on an independently searched topology returned all 432 called states unchanged, and no called state therefore depends on the tree. That single disagreement is a property of the construct-heavy sampling rather than of the vertex. The tips were pruned to natural sequences and the reconstruction re-run on the same fixed topology under both models, on two definitions of the exclusion: the outgroup tips for which a natural database record exists, 122 of the 341, and one tip per natural source organism. Every deep vertex then reads tyrosine under both models, 21 of 21 and 20 of 20, and every one at a posterior of 0.95 or better, where the committed 430-tip set reached that threshold at 29 of 34 under the mixture and 32 of 34 under LG+G4. At the disputed vertex the mixture reads tyrosine at 0.999 on the natural set and at 0.995 on the one-tip-per-organism set, so the two models agree there once the deposited constructs no longer condition the reconstruction. The 341 laboratory tips derive from 87 natural source organisms, 86 of the 87 reading tyrosine at this column, and the outgroup reads tyrosine in 325 of 341 tips on the full set and in 122 of 122 once the constructs are dropped. The family’s own vertex, set aside above, reads methionine at 0.7426 under LG+G4 and 0.7685 under the mixture on natural sequences only, against 0.999 in the published description. The topology is inherited rather than re-searched, the 430-tip search having taken thirteen hours, so this tests taxon weighting and not tree search. results/_ancestral_natural/.

Note N47. Every buried pair measured across the family and the controls. Every pair among the five buried positions was measured on five family structures against up to 22 superfamily structures, giving ten pairs in total. Three pairs separate the family with no overlap between the two ranges, and all three involve the arginine at 182. The remaining seven overlap the controls and represent properties of the barrel. The contact between the chromophore site and the same arginine also separates, at 2.71 to 3.72 A in this family against 6.13 to 7.59 A in sixteen measurable controls. Per-site substitution rates rank the positions in the same order. Over 253 aligned positions the two arginines were recorded at the 2.8th percentile, the first glutamate at 4.3, the hub tyrosine at 8.3, the fifth position at 13.0 and the chromophore site at 34.4. On these measurements the discriminating component of the arrangement is one arginine together with the contacts it makes, rather than five residues acting as a unit.

Note N48. The decay comparison at the two positions, with its dependence structure. In a second approach, the same two residues were reached from sequence. Because a dead copy is under no selection to retain a site required by the protein, a position that remains invariant in a pseudogene is held by the barrel structure, while a position that decays was held by the function. Of the twenty positions reported as the fingerprint of this family, seven had never been tested for decay and were

read as a baseline, holding at 94 of 105 reads across the 15 species carrying a pseudogene readable at these positions. The arginine at 87, the glutamate at 134 and the tyrosine at 168 are retained at 38 of 45 reads. The arginine at 182 and the glutamate at 199 are retained at 17 of 30. These are counts of residue calls and not of independent observations. Twenty-one of the 36 fungal lesions fall in Aspergillus, several species contribute more than one locus, and a significance test over the pooled calls would treat one lineage as many, so no p-value is attached to them and they are reported as an observation about these loci. The structural measurement above is the independent evidence for the same two positions, and it does not depend on how the pseudogene calls are grouped.

Note N49. The duplication, transfer and loss reconciliation and its controls. A duplication, transfer and loss reconciliation was run in order to test whether the distribution of the family can be accounted for without transfer. Two programs were run against two species trees constructed by hand. Between 9 and 16 transfers were inferred depending on the run, and a model permitting no transfer costs 33.1 log units and requires 41 duplications together with 310 losses. Controls bound what this result supports. No individual transfer edge was retained under a change of rate regime, and the edges weighted highest fall within Cnidaria, which is the lineage the topology tests account for without transfer. The reconciliation supports the occurrence of transfer within the family and identifies no particular transfer.

Note N50. The transcript evidence for the fungal copies, per library. To establish whether the surviving copies are used, transcript data were examined for the fungal copies, and the intact copies were found not to be transcribed. Across eight fungal RNA-seq libraries the intact loci carry no more reads than background. In Aspergillus luchuensis, for which the question can be addressed directly, eight public libraries from four independent studies totalling 132 million mapped reads returned no reads at the intact locus, with no sense, antisense or spliced records in any library. The single exception in the set is a broken copy. In Lepraria finkii a locus carrying a frameshift and a non-canonical AT-AG intron was transcribed at 148 sense reads from 37.8 million, which is 0.18 times the average gene of that library. Transcription persisting at a disrupted locus is consistent with recent inactivation, although read-through from a neighbouring gene would produce a similar result at this depth. A search of 6,843,051 contigs across 203 fungal transcriptome sets returned no acceptable hit, whereas agmatinase, located two genes downstream and 4.8 kb away, was recovered from 122 of the same 203 sets. Outside the fungi the same measurement places the family’s extremes within lineages and never between them. The coral copy was recorded at 0.005 times the average expressed component of its assembly and in the 4th percentile, and *Gymnoxanthella radiolariae* at the 97th percentile of its own assembly and at 5.17 times the average expressed component.

Note N51. The radiolarian read panel behind the dinoflagellate provenance. Two of the three dinoflagellate records are photosymbionts of radiolarians, so host sequence is the alternative origin for the gene and it was tested against radiolarian sequence directly. Sixteen RNA-seq runs were searched, two of them positive controls. The two controls carry family reads, 13 and 29, which is what

makes a negative readable. Of the remaining 14, ten are from heterotrophic and four from photosymbiotic species; all 14 carry a working actin control and all 14 are deep enough; ten count as negatives and four fail on library selection, three on RACE and one on CAGE, both of which amplify part of the transcriptome so that a negative is not conclusive. No run carries a family read. Separately, 43 assembled rhizarian transcriptome sets were searched, 42 with a working actin control, and one carries a hit: a declared metagenomic assembly from hypoxic fjord sediment, which is not evidence about the organism it is named after. None of the 43 is radiolarian, because no assembled radiolarian transcriptome exists. Of the two carrier assemblies, one holds enough of its own 18S for the rRNA question to be asked, and in it neither host reference matches a contig other than the dinoflagellate's own 18S at same-taxon identity. results/13_origins/rhizaria/.

Note N52. The census under four alternative numbering references. The five positions are located on each candidate by pairwise alignment to the published founder A0A951HUW2, and the read was repeated under four further references in order to establish whether the census depends on that choice: A0A1M3T1J2 (*Aspergillus luchuensis*), A0A7U3V473 (Scleractinia), A0A6V0IBN8 (*Zooxanthella nutricula*) and A0A495W682 (*Saccharothrix australiensis*). Re-reading a locus requires its protein sequence, which is available for 206 of the 396 called loci, so every figure here is measured over those 206 and the behaviour of the other 190 loci is unmeasured. Those 190 include 46 that read a complete network under the committed read, so the test leaves a third of the members untested for reference-dependence. Table S14 carries every read behind this note, locus by locus. The positions were transferred into each reference by pairwise alignment and the transfer checked twice: all five references read the five reference residues at their mapped positions, and three-dimensional superposition onto the founder places the five positions and the chromophore site at the same residue numbers the alignment gives, at 30 assignments out of 30, the fungal frame's 182 to 184 offset included. Over the 206 loci, and in the order the references are listed above with the founder first, 99, 93, 99, 96 and 95 read the complete network, in 95, 89, 95, 92 and 91 assemblies and 58, 56, 57, 57 and 56 carrying units. The movement is concentrated rather than general: 87 loci read five of five under every reference and 14 of the 101 that read a complete network under any reference do not, while a further 77 move only between sub-threshold strings and 28 do not move at all. Of the 77 carrying units, 54 have an accession in this subset reading five of five under the founder, and against the founder the four alternatives lose two, one, one and three of them while none gains one. Those unit losses are not the carrying-organism losses above, which run two, one, one and two: a unit can hold several assemblies and is lost only when every one of them stops reading the complete network. In the other direction two loci in no carrying unit read a complete network under an alternative and not under the founder: GCA_044327125.1 (*Bunodosoma granuliferum*) reads REYRE under the fungal reference alone, and MGYG000495968 reads REYRQ under the fungal and cnidarian references against REYRG under the founder. results/_founder_swap/.

Note N53. Whether a reference drawn from a candidate's own lineage reads it better, and the three strings the superposition can arbitrate. The expectation behind the reference swap of Note N52 was that a bacterial anchor misplaces the positions on eukaryotic candidates, and it fails in the direction it was predicted to hold. Of the 87 Ascomycota loci in that subset, 70 read five of five under the bacterial founder and 67 under the *Aspergillus* reference, which is the joint worst of the five on them: three *Aspergillus* loci read RE-RE, a gap at the tyrosine, under their own lineage's anchor while reading REYRE under all four other references. On the 30 cnidarian loci the cnidarian reference and the founder are equal at 20. The pooled figure, 88 of 118 same-lineage pairs reading five of five against 394 of 912 other-lineage pairs, is an artefact of composition: the 82 loci in no carrying unit are almost all sub-threshold and can only ever contribute other-lineage pairs, since no reference is drawn from that group. The within-phylum rows are the test. Three loci in the subset have a structural model, which lets the superposition arbitrate the string and not only the frame. On GCA_964254835.1 and GCA_963678995.2 it supports the founder against the one reference that differs in each case. On MGYG000495968 it supports the two alternatives against the founder, reading REYRQ where the committed alignment reads REYRG, which is the same disagreement at position 199 already recorded for that locus between the superposed and aligned reads. That locus is in no carrying unit under the committed read and the decision not to switch it stands. Three models are enough to show that the founder's pairwise read is wrong on at least one locus and too few to put a rate on it. results/_founder_swap/.

Note N54. The construct screen in full, over the called loci and at the thresholds the filter applies. The screen aligns every protein and reading-frame translation held for a locus against an FPbase snapshot with blastp and flags the locus where the match reaches 90% identity over at least 50 aligned residues, the two thresholds applied together. Across every frame read, 103 of the 396 called loci trip that test, falling to 22 distinct constructs, of which EGFP accounts for 27 loci, mCherry 18 and Citrine 12. Ninety-six were removed and seven held for review, five of the seven being one Free Use GFP sequence at 1.000 identity in five *Burkholderia cenocepacia* assemblies and the other two metagenome bins. Two further loci reach 1.000 identity over an alignment of seven residues, too short for the test to fire, and are one *Aspergillus neoniger* locus held in two database copies of one assembly. Among the 140 loci reading the complete network the highest FPbase identity recorded is 0.402, a cnidarian protein against a sponge genome. The screen reaches 361 of the 396 called loci: 27 hold no frame of 50 residues or more and the remaining eight were called by the coral screens of an earlier stage, which record their own reporter match at 32.4 to 38.5% identity. None of the 77 carrying units was removed by the construct screen, while 18 hold a locus with a conservative review flag firing on taxonomy and on score against DsRed rather than on identity.

Note N55. The nidogen G2 domain beside the family, and why the two contacts are not the same pair. The more abundant of the two domain families in this fold lacks the second arginine and the mapped tyrosine. Nidogen G2 (PDB 1H4U, crystal structure [31]) reads REILL at the five superposed positions,

carrying the first arginine and the first glutamate and differing at the other three: leucine where the family places its second arginine, isoleucine at the mapped tyrosine and leucine at the fifth position. Its 2.64 Å tyrosine-glutamate contact is a local Tyr544-Glu542 pair two positions apart in a Glu-Leu-Tyr run, and it is not the couple found in this family, which runs between positions 134 and 168 and measures 2.76 Å in the *Saccharothrix* member and 2.66 Å in the *Aspergillus* one. The two distances are therefore not measurements of the same pair. A figure of four superposed structures, avGFP and the *Aspergillus* member from Figure 1 with nidogen G2 and the *Saccharothrix* member in place of the founder and the coral copy, is held in the repository at figures/paper1/fig3_nidogen.png and is produced by 27_figures/fig3_nidogen.py.

Note N56. The archaeal locus read a second way, by superposition. The archaeal search used GCF_965223225.1 as a query and recovered the locus in that same assembly, so its five positions were located at 100% identity to the query. Its identity to the founder, at 25 to 27%, falls inside the band where this work reports pairwise alignment to be unreliable. The locus also lies outside the 206 for which the reference swap of Note N52 could be run. Its positions were therefore read again by three-dimensional superposition. The predicted model was TM-aligned onto each of the two structural references and the assignment read out of the structural alignment itself rather than by a nearest-atom search. Against the founder the superposition returns TM 0.874 at RMSD 1.66 Å and places all five network positions and the chromophore site on the residues the pairwise read gives, at backbone separations of 0.30 to 0.63 Å; against the second reference it returns TM 0.720 at RMSD 1.99 Å and the same six residues, at 0.23 to 1.02 Å. Both reads give REYRE with threonine at the chromophore site. A second model of the same protein, built by a different predictor, sits at a mean pLDDT of 40 against this work's floor of 70 and superposes at TM 0.42 and 0.31 onto the two references, so it does not present the same fold and its assignment is reported rather than used: it leaves three of the five positions unaligned against the founder. The usable model is at mean pLDDT 92.0 with 94% of its residues at or above the floor. tools/archaeal_superposed_read.py, results/_archaeal_read/.