Evidence grades for protein database entries longer than titin

Ahmet Tahir Mat

Papex Labs

Correspondence: ahmet@papex.org

13 November 2026. Preprint, not peer reviewed.

Abstract

Protein databases hold hundreds of sequences longer than titin, the muscle protein often described as the largest protein. Most are gene predictions whose records do not state their evidence. We collected every sequence longer than a 35,991-aa human titin isoform from UniProtKB, UniParc and NCBI Protein and kept 510; 57 are longer than PKZILLA-1 (45,212 aa). Each was graded A–D by explicit checks, then tested with public RNA-seq (261 runs) or, for three bacterial models, with their source reads. Of the 57, 17 are contradicted (grade D) and 34 have no line of support (C). The most common pattern, a translated tandem DNA repeat, also marks real silk proteins and adhesins in our positive controls, so it counts against a model only with DNA-level evidence; RNA-seq can rarely settle these models, because almost none of their splice junctions is unique. Two fish models are consistent with merged tandem titin genes. Six sequences, at four loci, reach grade A or B, all on gene-model evidence and stable in 31 of 33 threshold settings; the longest, an 85,803-aa *Candidatus* Omnitrophota open reading frame, is supported by its source reads except across a 1.2-kb repeat. For a 39,686-aa *Candidatus* Velamenicoccus protein whose detection its authors had reported, all 60 raw proteomics files gave 182 unique peptides, none beyond its first 40 %: the N-terminal part is made, a full-length chain is not shown. Controls exposed weaknesses in three of our rules; we report every revision and release the table, rules and code.

1 Introduction

Titin, the giant protein of striated muscle, is often described as the largest protein [1, 2]; its complete gene sequence is known [3]. Human titin isoforms reach 35,991 aa (RefSeq NP_001254479.2), the length we use as the threshold here. In 2024, Fallon et al. [4] described PKZILLA-1, a polyketide synthase of the haptophyte *Prymnesium parvum* of 45,212 aa, supported by RNA-seq across the gene and by peptides that map uniquely to 7 of its 17 exons. West-Roberts et al. [2] used genome curation of metagenome assemblies to recover predicted proteins of up to 85,804 aa, most common in *Candidatus* Omnitrophota; their preprint has no journal version listed.

Protein databases hold many more entries of this size. Amaral and Devos [5] surveyed more than 41,000 UniProtKB proteins of at least 5,000 aa and discussed indirect indicators of whether such "huge proteins" are real. Giant genes in bacteria have long been described [6], and gene-model errors that merge neighbouring genes remain common in eukaryotic annotations [7]. Tools such as GeneValidator [8] and OMArk [9] assess gene models one gene or one proteome at a time.

We did not find a per-entry account of the longest database entries: which are supported as full-length gene models, which are contradicted, and by what evidence. We therefore graded every sequence longer than 35,991 aa in three databases with explicit, rerunnable checks; reanalysed public RNA-seq for the eukaryotic models; re-mapped the original sequencing reads for

Table 1. Database entries longer than titin.

Database	Snapshot	> 35,991 aa	> 45,212 aa
UniProtKB	release 2026_04	51	4
UniParc	with UniProt 2026_04	477	65
NCBI Protein	11 Nov 2026	472 (349 RefSeq)	47

Table 2. Evidence grades.

Grade	Name	Meaning
D	contradicted	At least one check contradicts the model.
C	no support line	No contradiction, but no line of support either.
B	supported in parts	No contradiction, plus at least one line of support.
A	supported full-length model	Two independent lines of support, one of them over the whole model (transcripts or reads).

three bacterial models; and re-searched public proteomics data where they exist for an entry or its organism. The result is an evidence-graded table. It is not a claim of a new longest protein, and none of the entries we grade is newly described here.

2 Results

2.1 An inventory of 552 sequences longer than titin

UniProtKB (release 2026_04) holds 51 entries longer than 35,991 aa, UniParc 477 and NCBI Protein (11 November 2026) 472 records, 349 of them RefSeq (Table 1). None of the UniProtKB entries is reviewed; their evidence levels are "inferred from homology" (37) and "predicted" (14), as is PKZILLA-1's own entry (A0AB34IYJ6). The union holds 552 unique sequences (merged by MD5 checksum), 71 of them longer than 45,212 aa. We excluded 42: 29 without a live record, 6 synthetic or patent sequences (including the two longest records, 187,500 aa) and 7 marked as partial. The remaining 510 sequences are 452 eukaryotic, 57 bacterial and 1 archaeal; 57 are longer than 45,212 aa (34 eukaryotic, 23 bacterial). The 329 RefSeq XP_ records come from only 52 genes, so the sequences fall into 228 loci.

2.2 Grading rules, and how far to trust them

A grade applies to a gene model, the sequence as annotated in its assembly, not to the biology of the gene (Table 2; every rule and threshold is in Supplementary Methods). D means that at least one check contradicts the model. C means no contradiction and no line of support. B adds at least one line of support: an orthologue of consistent length, a near-identical copy in another assembly, RNA-seq support of the intron chain, NCBI's own statement of full RNA-seq support, for bacteria support by the source reads, or unique peptides along the whole model. A requires two independent lines, one of them transcript evidence over the whole intron chain (eukaryotes) or read-level support of the assembly (bacteria).

The rules are ours and were revised seven times while the work was in progress (Supplementary Note 1). Three revisions were forced by positive and negative controls that we built to test the rules themselves; one corrected an error in how twin RefSeq and GenBank records were counted; the others replaced checks that failed on real cases. Two rules were added late: a peptide support line, written after its only test case had been seen, and a guard for one RNA-seq

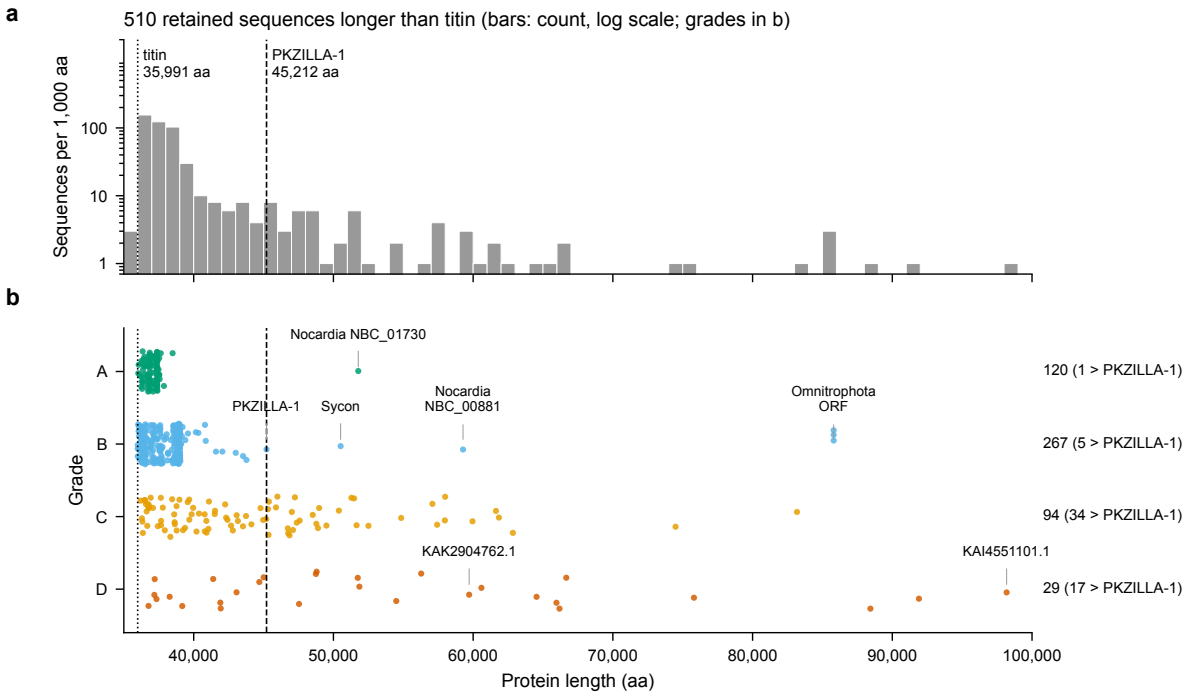


Figure 1. Lengths and final grades of the 510 retained sequences. (a) Sequences per 1,000-aa bin (log scale). (b) Each sequence as one point by final grade (jittered within a grade). Dotted line: human titin isoform NP_001254479.2, 35,991 aa; dashed line: PKZILLA-1, 45,212 aa. Counts give the sequences per grade and how many are longer than PKZILLA-1. Data: `data/locus_checks.tsv`.

test, fixed before its data existed. Every revision is reported with its effect, and every threshold was varied in turn (Supplementary Table S3). This matters for how far the grades can be trusted. They are reproducible from the released code and data, and a D or an A rests on a named check that a reader can rerun. But the rules were tuned on two controls (human titin and PKZILLA-1) and tested on a limited set of further controls, not by an outside annotator. Several of the controls changed a rule only after it had been applied, and only the repeat rule was retested on an independent set. A grade is therefore a statement about which of our checks a model passed with the data available on the snapshot dates, not a verdict on the gene.

2.3 Translated tandem repeats: frequent above 45,212 aa, but not by themselves a contradiction

Sequence and locus checks (Stage 1) graded the 510 sequences 114 A, 213 B, 154 C and 29 D; none of the 57 longer than 45,212 aa reached A (Figures 1 and 2). The most frequent signal was the translation of a tandem repeat (the D1 pattern): high protein self-identity at the best period, or most residues inside repeated 25-mers, with almost no recognisable domain. It occurs in 72 sequences, 32 of them longer than 45,212 aa (29 eukaryotic).

Positive controls showed that this pattern alone does not separate real proteins from questionable models (Figure 3a; Supplementary Note 2). Among 176 reviewed proteins with protein-level evidence, chosen by a rule fixed before scoring, it flagged 10: three of four complete spidroins, the silk fibroin heavy chain, four serine-rich bacterial adhesins and two long human proteins. None of the flagged proteins with a coding sequence showed a DNA-level contradiction. The pattern therefore counts as a contradiction only together with a non-coding Fickett signal in most coding windows (D2) or several alternative reading frames as open as the annotated one (D3). Tested on a hold-out set fixed before scoring (9,428 further reviewed proteins), the first

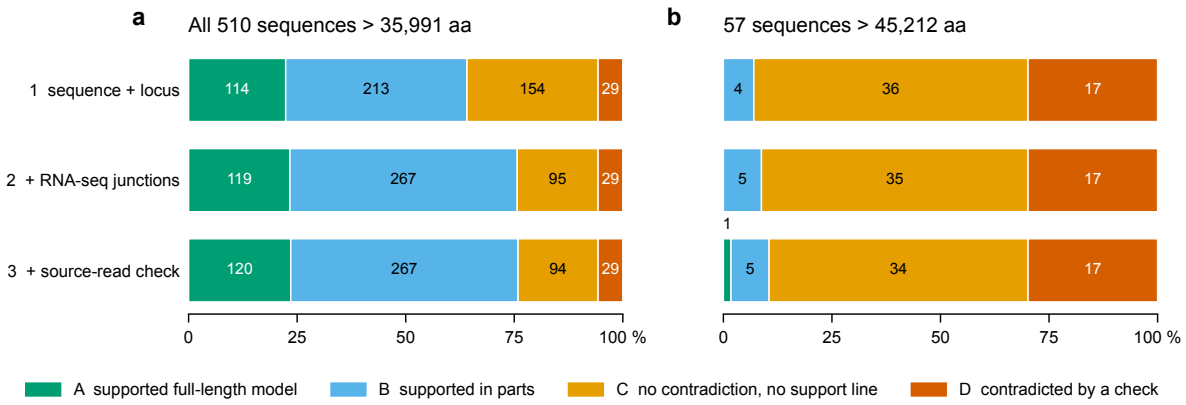


Figure 2. Grades after each evidence stage. Stage 1: sequence and locus checks; Stage 2: adds RNA-seq junction evidence for eukaryotic models; Stage 3: adds the source-read check of bacterial assemblies. (a) All 510 sequences. (b) The 57 longer than 45,212 aa. Data: `data/locus_checks.tsv`.

rule would have flagged 17, the revised rule flagged 0. The revision moved 47 audit sequences from D to C, 17 of them longer than 45,212 aa. Of the 32 patterned sequences above that length, 15 remain D and 17 are C with a repeat flag; under the first rule D would hold 34 of the 57.

The longest retained entry shows the pattern with DNA-level contradiction (Figure 4). KAI4551101.1 (98,182 aa) from an *Ovis ammon polii* × *O. aries* hybrid assembly repeats with a period of 22 aa at 0.847 identity and has no recognisable domain; its coding sequence repeats with a period of 22 bp, has a Fickett coding score of 0.465 (non-coding range), and two alternative frames are nearly free of stop codons. All 11 retained entries from this hybrid share the 22-aa period or a multiple of it and remain D. These models are consistent with open reading frames called through a GC-rich tandem array; they are not supported as protein-coding genes by the evidence examined (Supplementary Note 3).

2.4 RNA-seq supports most titin-like models but can rarely test the repeat models

For eukaryotic models we streamed public Illumina RNA-seq runs from ENA and counted spliced reads that match each annotated intron exactly (Methods); 261 run results from 71 species were kept. The controls behaved as expected: human titin had 353 of 361 introns supported and none of its 325 introns with covered flanks unsupported; PKZILLA-1 had all 16 supported, 7 of them only by multi-mapped reads inside repeated modules. Of the 173 models with a verdict from reads, 59 were “full” (at least 90 % of introns supported), 75 “partial” (50–90 %), 27 “weak” and 12 without expression in the runs available; none was “contradicted”. The stage moved 59 sequences from C to B (48 of them isoforms of one *Apostichopus japonicus* gene) and 5 from B to A. All 119 sequences at grade A after this stage are vertebrate titins. Above 45,212 aa it added one B: the sponge *Sycon ciliatum* model XP_065194412.1 (50,516 aa), with 22 of 24 introns supported by unique reads; its record notes that 2,968 transcript bases bridge a gap in the genome assembly.

The 43 models that had been graded D under the first repeat rule were tested afterwards, under rules fixed before the run (Supplementary Note 4, Table S4). Only 10 had an assembly and usable runs; the others lack RNA-seq in ENA (28) or a gene model with coordinates (5). Because multi-mapped reads cannot tell repeat copies apart, these models count only uniquely placed reads, and a correct intron inside a repeated copy cannot obtain one. In the five testable models longer than 45,212 aa, at most 8 % of introns have a junction sequence that is unique within the model, so “partial” support was out of reach before any read was counted. The

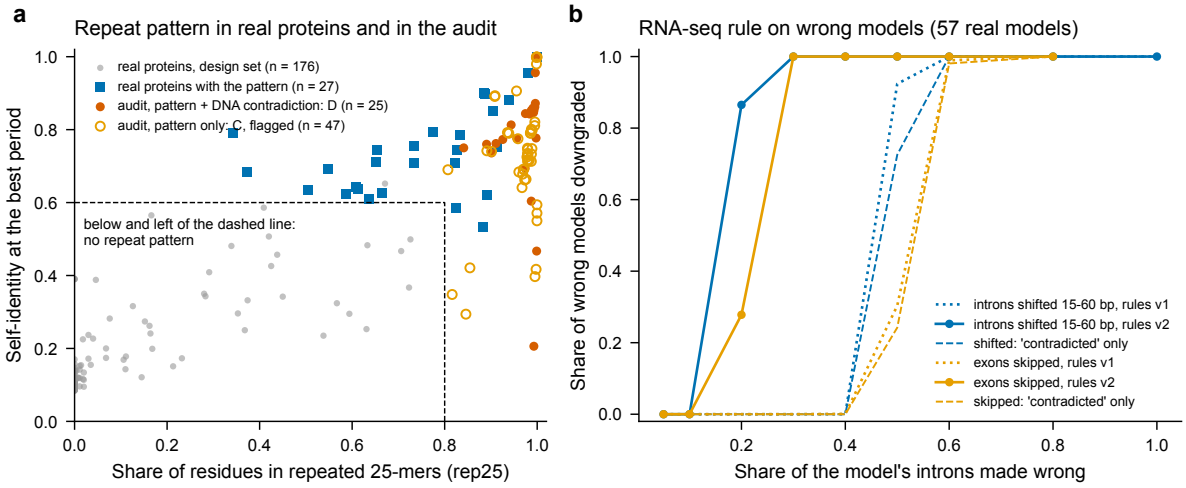


Figure 3. Controls of the grading rules. (a) Self-identity at the best period against the share of residues in repeated 25-mers. Grey: the 176 real proteins of the design set; blue squares: real proteins with the repeat pattern (design and hold-out sets); filled circles: audit sequences with the pattern plus a DNA-level contradiction (D); open circles: audit sequences with the pattern alone (C, repeat flag). The dashed line bounds the pattern. (b) Share of deliberately wrong versions of 57 real models that the RNA-seq rules downgrade, by the share of introns made wrong; solid: revised verdict order, dotted: first order, dashed: "contradicted" alone. Data: `controls/d1/`, `controls/rna_power/rna_power.tsv`.

results fit this: the two *Schistosoma haematobium* isoform models were "partial" (32 of 45 and 46 introns, none unsupported among 32 testable ones) and move from C to B; *Owenia* and *Arctogadus* models were "weak"; the six stickleback, eel and *Acnodon* models had at most 34 uniquely placed reads. No model was contradicted, and no grade above 45,212 aa changed. One test run of a *Schistosoma* model was seen before the rules were committed; the rule code predates it, and its result is not needed for the verdict (Supplementary Note 4).

2.5 Two fish titin models are consistent with merged tandem titin genes

Zebrafish has two titin genes in tandem, *ttna* and *ttnb* [10]. KAK2904762.1 (*Cirrhinus molitorella*, 59,714 aa, 454 exons) splits into two titins (Figure 5): positions 1–31,000 follow zebrafish *ttna* from its N- to its C-terminus (30 of 30 windows), and from position 31,001 the windows start again at a titin N-terminus and run to the C-terminus of *ttnb* (20 of 29). Across 6 RNA-seq runs, the intron at amino acid 30,675, right after the *ttna*-like C-terminus, has no junction read, while the five introns before it have a median of 103 reads and the five after it 2. TSR51413.1 (*Bagarius yarrelli*, 54,498 aa) shows the same two-part pattern, with too few reads to test the join. Both were graded D5 (parts of different complete proteins). We describe them as consistent with two adjacent titin genes joined in one model; gene order on the source contigs was not checked independently.

2.6 Source reads support the bacterial giants, with one untestable repeat

Bacterial models cannot be tested with RNA-seq, because no strain-level transcript data exist for these organisms. For the 85,803-aa Omnitrophota open reading frame (three assemblies) and two *Nocardia* models on complete genomes, we re-mapped all reads used to build each assembly and checked depth, read-pair continuity, clipped reads, discordant pairs and the reading frame of the read consensus, with an artificial misassembly as a negative control in every run (Methods; Supplementary Note 3).

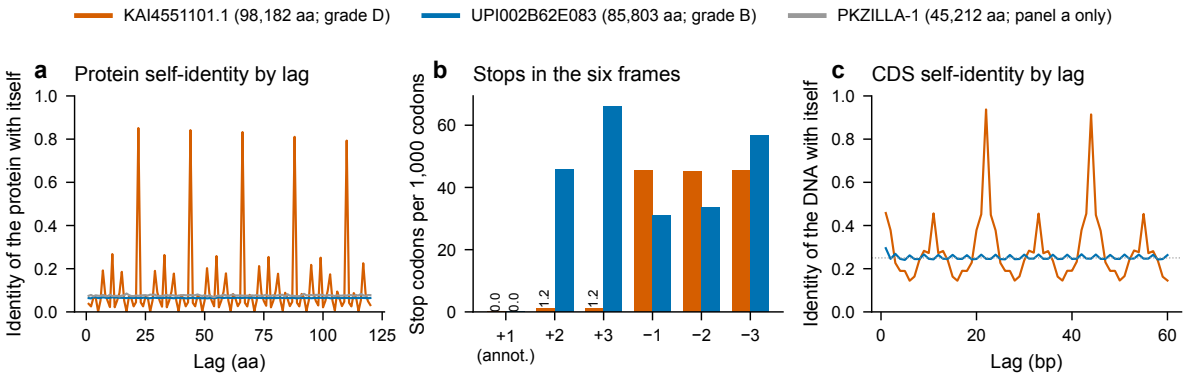


Figure 4. A translated tandem repeat. (a) Identity of each protein with itself shifted by the given lag. KAI4551101.1 (grade D) peaks at multiples of 22 aa; the 85,803-aa Omnitrophota ORF UPI002B62E083 (grade B) and PKZILLA-1 stay at background. (b) Stop codons per 1,000 codons in the six reading frames of each coding sequence. (c) Identity of the coding DNA with itself by lag (first 200 kb). Data: `data/fasta/unique_ge35992.fasta.gz`, `data/manual/KAI4551101_locus.fa.gz`, `data/manual/DAOTDT01.fasta.gz`.

Both *Nocardia* models are supported by Illumina and nanopore reads (median unique depth 272 and 252, gene-to-contig depth ratio 1.02). With an orthologue of consistent length, WP_327112628.1 (51,778 aa) reaches A; WP_406270362.1 (59,275 aa) reaches B, because the same homologue gives it a length ratio of 1.27. Neither has transcript or peptide evidence; for these bacteria, A means that the gene model is supported by the assembly, its reads and orthology, not that the protein is made.

A single 257,412-bp open reading frame of 85,803 aa occurs in three metagenome-assembled genomes from three samples (DAOTDT01, DAOPMG01, DAOOUB01), 99.96–99.98 % identical (Figure 6). Over the gene, the reads of each sample give median unique depths of 40, 52 and 45, at most 2.1 % discordant pairs per window, and no frame-breaking change. The gene carries a direct repeat: two copies of about 1.18 kb, 10,362 bp apart, that differ at 2 positions in all three assemblies. Fragments of these libraries cannot be placed uniquely inside a copy, so the middle of each copy cannot be tested; a first version of our check mistook this for a break, and the revised check asks how many spanning pairs a correct assembly would give before counting a missing pair against it. Two assemblies are "supported outside 1.3 kb of repeat" and the third is "inconclusive" at one boundary at the edge of a copy (4 pairs where about 6 are expected). The locus stays at B; whether the copies are joined in the assembled order cannot be decided with these libraries, and no long-read data exist for these samples.

2.7 Peptides cover only the N-terminal part of a 39,686-aa bacterial protein

WP_128698900.1 of *Candidatus* Velamenicoccus archaeovorus (39,686 aa) is shorter than PKZILLA-1, but its organism has a public proteomics dataset, PRIDE PXD025008 [11]. The dataset's authors list "a very large multienzyme surface protein (4,384 kDa)" among the most frequently detected proteins, whose "gene codes for 39,678 amino acids" [11]. Our analysis confirms their detection and maps where the peptides fall.

We first searched the authors' deposited peak lists and then, because those turned out to hold only 54 % of the MS/MS spectra, all 60 raw files (1,106,517 spectra) with Sage [12], against the authors' database with the organism's earlier protein set replaced by the proteins of its closed genome, plus a reversed and a shuffled copy of the giant (Methods; Supplementary Note 5, Table S5). In the raw files, 182 peptides unique to WP_128698900.1 passed 1 % FDR at the spectrum, peptide and protein levels (1,505 PSMs). 179 of them start within the first 6,000 aa;

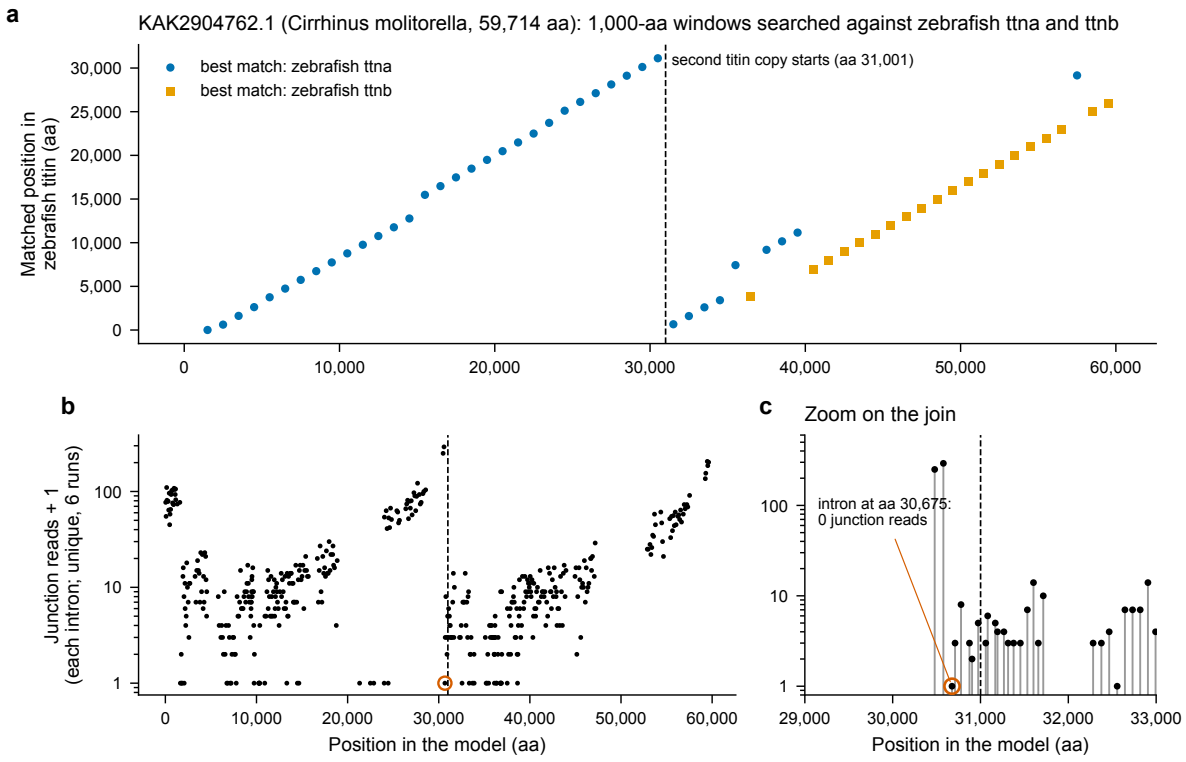


Figure 5. A model consistent with two merged titin genes. (a) KAK2904762.1 in 1,000-aa windows searched against zebrafish ttna (circles) and ttnb (squares); y is the start of the best match. The match restarts at a titin N-terminus at aa 31,001 (dashed line). (b) Uniquely mapped junction reads per intron, summed over 6 runs (reads + 1). (c) The join: the intron at aa 30,675 (open circle) has no junction read. Data: `data/followups/fish_titin_pair.tsv`, `data/rnaseq_run_results.tar.gz`.

three lie further on, at aa 8,279, 12,982, 15,831, each from a single spectrum; none maps after residue 15,840 (39.9 % of the length; Figure 7). The gap is not a property of the sequence: fully tryptic peptides of 7–40 aa occur evenly along it (41–63 per 1,000 aa), and no peptide lies inside or after the protein's internal repeat blocks (aa 17,653–25,339). The reversed and shuffled copies obtained no PSM at 1 % FDR, while elongation factor Tu of the same organism was covered to 57 % by 24 unique peptides. The peak-list search gave the same picture (186 peptides, 179 shared with the raw-file search); its one peptide beyond residue 16,000 (start 16,949, peptide $q = 0.0096$) was not found in the raw files.

These peptides show that the N-terminal part of the protein is made. They do not show that a single 39,686-aa chain is made, and their absence from the C-terminal 60 % neither supports nor rules out that part of the model. The entry stays at C: our peptide line requires unique peptides in all five fifths of the length, and the raw-file search covers two. That line was written after this result had been seen; a first draft that asked for three fifths would have counted the peak-list search as support (Supplementary Note 1).

For the 50,516-aa *Sycon ciliatum* model we examined the two PRIDE projects listed for the species. PXD031991 holds acid-extracted, chemically derivatised histones [13] and is not suited to detecting this protein. PXD060105 is a proteome of the spicule organic matrix [14]; the authors' search database contains the model's sequence as 6 transcript-derived fragments, and their report of 37 proteins does not include it. In our search of the 5 raw files, no spectrum matched the model at 1 % spectrum-level FDR, while 31 of the authors' 37 proteins were recovered; the dataset is too small to reach 1 % at peptide or protein level for any protein. A giant protein is not expected in a mineral matrix prepared by dissolving the surrounding tissue, so this absence says nothing about the model.

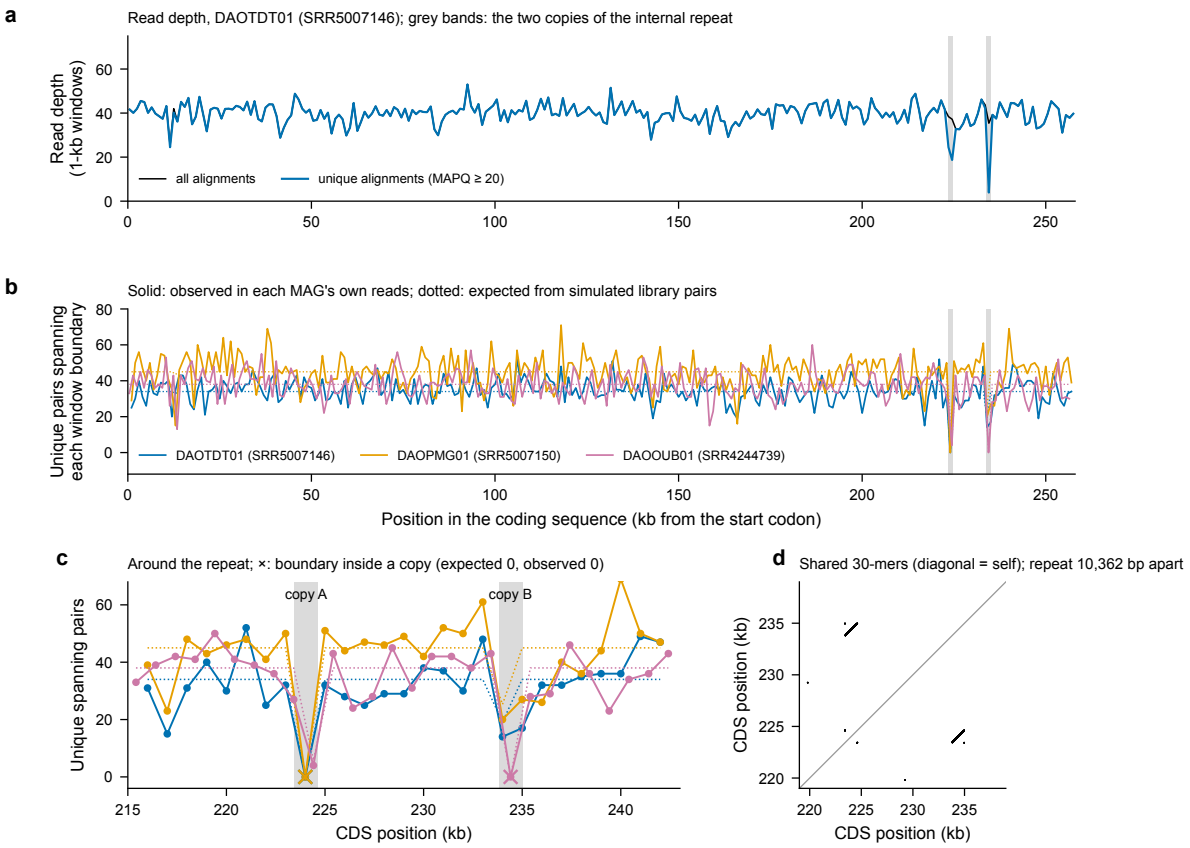


Figure 6. Source-read check of the 85,803-aa Omnitrophota ORF. (a) Read depth per 1-kb window in DAOTDT01: all alignments and uniquely placed alignments (MAPQ ≥ 20); unique depth drops only at the two repeat copies (grey bands). (b) Uniquely placed read pairs spanning each 1-kb boundary in each assembly's own reads (solid) and the number expected from simulated pairs of the same library (dotted). (c) The repeat; × marks boundaries inside a copy, where 0 pairs are expected and 0 were observed. (d) Shared 30-mers within the coding sequence from 219 to 239 kb. Data: `data/dna_checks/*.json`, `data/manual/DAOTDT01.fasta.gz`.

2.8 What remains above 45,212 aa

After all stages, the 510 sequences grade 120 A, 267 B, 94 C and 29 D. In the leaderboard (Supplementary Table S1), isoforms of one gene, near-identical copies in several assemblies and duplicate records of one gene in one genome (11 records merged) form one row each, giving 213 graded loci: 9 A, 85 B, 90 C and 29 D. Above 45,212 aa there are 55 loci: 1 A, 3 B, 34 C and 17 D (Table 3). Grade A is otherwise reached only by 8 vertebrate titin loci (36,247–38,490 aa). The A/B status of the four loci in Table 3 held in 31 of 33 threshold settings; the exceptions were counting a single open alternative frame as D3 and requiring 10 junction reads per intron (Supplementary Table S3).

3 Discussion

The longest entries in protein databases are mostly not supported as full-length gene models by the evidence we examined. Of the 57 sequences longer than PKZILLA-1, 17 are contradicted by at least one check, 34 have no line of support (17 of them with the repeat pattern), and six, at four loci, are supported as gene models.

The most frequent pattern is a protein that reads like the translation of a tandem DNA

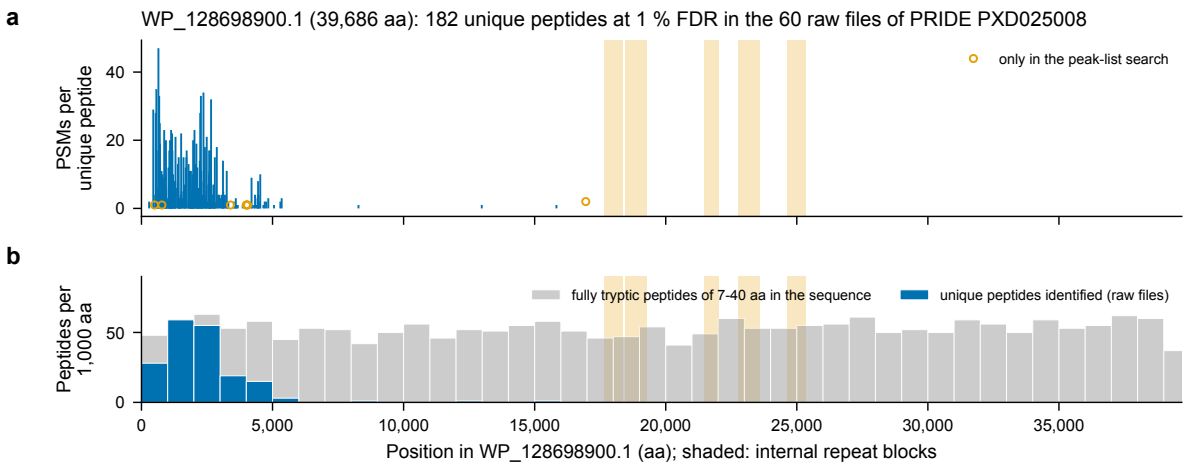


Figure 7. Peptides of WP_128698900.1 in PRIDE PXD025008. (a) Each peptide unique to the protein at 1 % FDR (spectrum, peptide and protein level) in the search of all 60 raw files, at its start position; height is its number of PSMs. Open circles: peptides found only in the search of the deposited peak lists. (b) Per 1,000 aa: unique peptides identified in the raw files (blue) and fully tryptic peptides of 7–40 aa in the sequence (grey). Shaded: internal repeat blocks (20-mers occurring more than once, blocks of at least 100 aa). Data: `proteomics/results/raw60_giant_peptides.tsv`, `proteomics/results/full6_giant_peptides.tsv`, `proteomics/results/verify_extra.json`.

repeat. Our controls show that the pattern is not, by itself, evidence against a model: real silk proteins and serine-rich adhesins share it, and at the protein level nothing we measured separates them from the audit entries. The difference lies in the coding sequence: 15 of the patterned entries above 45,212 aa have a non-coding DNA signal or several open frames, which none of the real proteins with a coding sequence showed. The other 17 remain unresolved: 11 have a weak coding signal (Fickett median 0.919–0.942), which is absence of support rather than a contradiction, 4 look like real repetitive genes by every sequence test, and 2 have no locus that could be tested. Short-read RNA-seq is a weak tool for them: in a tandem array a junction read fits every copy, so the test can show that the unique parts are spliced as annotated, but not how many copies the gene has. Long-read transcripts spanning the array, or peptides along the whole model, would decide them.

The merging of neighbouring genes, here two titin paralogues, is harder to detect without a reference proteome or transcript evidence; junction reads across the join gave a clear signal in one of the two cases. The remaining candidates are bacterial or sponge models, supported at the level of the gene model: assemblies, reads, coding signal and, for one *Nocardia* model, orthology. None has protein-level evidence: PRIDE lists no project for the Omnitrophota or *Nocardia* organisms by species name, and the one *Sycon* proteome suitable for a search is of a fraction where the protein would not be expected. By contrast, PKZILLA-1 itself has published peptide support [4]. The 85,803-aa Omnitrophota ORF falls in the length class reported by West-Roberts et al. [2] for curated Omnitrophota genomes; we could not compare sequences, because their data repository was not reachable from our environment. Our observation adds read-level support from the three source metagenomes and identifies the one stretch that short reads cannot resolve.

The one positive proteomic case shows what peptides can and cannot settle. For the 39,686-aa *Velamenicoccus* protein they confirm that the N-terminal part is made, as its authors reported [11], in the full raw data as in the deposited peak lists. They say nothing about the C-terminal 60 %, although tryptic peptides are as dense there as elsewhere. That absence is weak evidence either way: the model could be wrong beyond that point, or the C-terminal part could be cleaved off or harder to extract or detect. A grade based on peptides should therefore require them along

Table 3. Loci longer than 45,212 aa graded A or B.

Accession	Organism	Length (aa)	Grade	Supported by	Missing
UPI002B62E083 (+2 copies)	*Candidatus* Omnitrophota (3 MAGs)	85,803	B	one ORF in 3 assemblies; coding signal; source reads outside a 1.2-kb repeat	long reads across the repeat; orthologue; expression or peptides
WP_406270362.1	*Nocardia* sp. NBC_00881	59,275	B	complete genome; coding signal; Illumina and nanopore reads	orthologue of consistent length (ratio 1.27); expression; peptides
WP_327112628.1	*Nocardia* sp. NBC_01730	51,778	A	complete genome; coding signal; Illumina and nanopore reads; homologue of 46,734 aa	expression; peptides
XP_065194412.1	*Sycon ciliatum*	50,516	B	RNA-seq: 22 of 24 introns, full CDS coverage	orthologue (closest 4,063 aa); bridges a 2,968-bp assembly gap with transcript sequence; peptides

the whole model.

The table also says what would change each grade. For the repeat models, long-read transcripts that span the whole array would show whether the copies are spliced into one messenger RNA, and peptides from the repeat-flanking regions would show whether the protein continues past them. For the Omnitrophota ORF, long reads or long-insert libraries from the same samples would settle the arrangement of the repeat copies. For the *Sycon* model, a proteome of whole tissue searched against the full model, rather than a mineral fraction, would test whether the gene is translated. For the *Velamenicoccus* protein, the open question is the C-terminal 60 %: a targeted search for its peptides, or a proteome of a fraction enriched for surface proteins, could show whether it is part of the same chain. The column `still_missing` of Supplementary Table S1 lists this per locus.

A grade records which checks a model passed with the data available on the snapshot dates. It is not a statement about the biology of the gene, does not assess the work of the submitters, and should be updated as new assemblies, transcripts or peptides appear.

Limitations

1. **The rules are ours.** They were set on two controls and revised seven times during the work, three times after controls had been run; only the repeat rule was retested on an independent set, and one support line (peptides) was written after its only test case had been seen. All revisions and their effects are in Supplementary Note 1.
2. **RNA-seq is incomplete and uneven.** Runs were taken from whatever tissues ENA offered; 51 runs ended early on a dropped connection, so junction counts are lower bounds, and a "weak" or "no-expression" verdict is not evidence against a model. Repeat models can be tested only where their junctions are unique, and the power test shows that errors in fewer than 30 % of introns are often missed (Supplementary Notes 2 and 4).

3. **Little protein evidence.** Two proteomic datasets were searched and a third examined; the only peptides found concern an entry shorter than 45,212 aa. Our search limits (one missed cleavage, peptides of 7–40 aa) were set to fit a laptop's memory. Peptides show that parts of a protein are made, not that it is made as one chain.
4. **Repeats limit what short reads can test.** The 1.2-kb Omnitrophota repeat cannot be crossed by the available libraries, and our copy-number test cannot see an error shorter than 3 kb there; the three assemblies come from one study and one pipeline.
5. **Scope.** MGnify and other metagenomic catalogues were not included; most bacterial entries rest on sequence and assembly checks only; orthology was judged by length consistency of DIAMOND hits in UniProtKB, which favours well-sampled clades; every count refers to the stated snapshot dates.

Supplementary Note 6 lists these limitations in full.

4 Methods

Full methods, with every threshold, are in Supplementary Methods; code paths are given there.

Inventory. UniProtKB (release 2026_04) and UniParc were queried through the UniProt REST API [15], and NCBI Protein with the Entrez length filter `35992:100000000[SLEN]` on 11 November 2026 [16, 17]. Sequences were merged by MD5 checksum; withdrawn, patent or synthetic records and titles marked "partial" were excluded; loci were defined by shared NCBI GeneID or overlapping coordinates. All requests used a generic user agent within the providers' anonymous rate limits.

Stage 1: sequence and locus checks. For each protein we computed low-complexity and X content, the share of residues in repeated 25-mers, the best tandem period and its self-identity, and coverage by a panel of 74 Pfam profile HMMs [18] searched with PyHMMER [19, 20]. Locus DNA was fetched from INSDC/RefSeq to measure exon structure, distance to contig ends, N runs, GC by codon position, the Fickett statistic in 300-bp windows [21], stop codons in the alternative frames and the DNA tandem period. Proteins were cut into 1,000-aa windows and searched with DIAMOND 2.2.8 [22] against UniProtKB for orthologues and fusions. Contradiction rules (D1 repeat pattern, only with D2 or D3; D2 non-coding Fickett signal; D3 open alternative frames; D4 Ns or joins across contigs; D5 parts of different complete proteins) and support lines (S1 orthologue of consistent length; S2 near-identical copy in another assembly; S3 NCBI report of full RNA-seq support; S4 curated RefSeq record; S6 unique peptides in all five fifths of the length) define the grades as in Table 2. Record twins of one genome (RefSeq and GenBank or UniProtKB copies) do not count as independent.

Stage 2: RNA-seq. For eukaryotic models at grade B or C, four to six public Illumina RNA-seq runs per species (0.2–6 GB, not single-cell, different studies first, muscle first for titin-like models) were read from ENA [23], at most 60 million reads per file; runs first cut short by a read-saturation rule were repeated without it. Reads were filtered with minimap2 2.31 [24] against the coding exons and aligned with `-x splice:sr` to the whole genome, which serves as a decoy; alignments were handled with samtools [25]. An intron is supported by at least three exactly matching reads; multi-mapped reads count only for models without a repeat flag. An intron is testable when both flanking exons reach a mean unique depth of 10 in one run. Verdicts, in order: no-expression (< 50 unique reads); contradicted (≥ 10 testable introns with $\geq 50\,\%$ unsupported, or more reads on junctions leaving the gene than on annotated ones);

under-supported where covered ($\geq$ 20 % of $\geq$ 10 testable introns unsupported); full ($\geq$ 90 % supported); partial ($\geq$ 50 %); otherwise weak. For the repeat-flag stage, files were downloaded to disk with resumed retries before mapping, and a contradiction had to hold also on introns whose junction is unique within the model (`cloud/bundle_rep_rna/RULES.txt`).

Stage 3: source reads. Illumina pairs and nanopore reads from each assembly's BioSample were mapped with minimap2 to the whole assembly. In 1-kb windows over the gene we measured unique and total depth, uniquely placed pairs spanning each boundary, clip pile-ups, discordant pairs, the gene-to-contig depth ratio and frame-breaking consensus changes. A boundary is tested only where simulated error-free pairs of the same library would give at least five unique spanning pairs; stretches below that must be crossed as a whole. Since version 3.1 a copy-number test flags at least three consecutive windows whose total depth departs from the gene's depth by more than a factor of $\sqrt{2}$. Every check ran a negative control with an inverted join at the gene midpoint.

Proteomics. For PXD025008 the authors' database was kept, minus their reversed entries and with the organism's earlier genome-bin proteins replaced by the GenBank translations of its closed genome (CP019384.1); QAT16263.1 was replaced by its RefSeq version WP_128698900.1. For PXD060105, database entries of which at least half of the 10-mers occur in XP_065194412.1 were treated as fragments of the same gene and replaced by the full model. Raw files were converted with ThermoRawFileParser [26] (vendor peak picking) and searched with Sage 0.14.7 [12]: trypsin, one missed cleavage, peptides of 7–40 aa, 10 ppm precursor tolerance, fragment tolerance 0.5 Da (ion trap) or 20 ppm (Orbitrap), variable methionine oxidation, fixed carbamidomethylation only where the authors used it, reversed decoys. Identifications were kept at spectrum, peptide and protein $q \leq 0.01$; a peptide is unique if its sequence (I = L) occurs in no other entry of the database. Each database replaced partial or earlier versions of the model by the full model and added a reversed and a shuffled copy as controls.

Controls and sensitivity. The repeat-rule design and hold-out sets were defined by UniProtKB queries fixed before scoring. The RNA-seq power test scored deliberately wrong versions of 57 real models (introns shifted by 15–60 bp or exons removed, 20 repeats per setting) against the stored junction counts. Known misassemblies were built from the finished *Nocardia* sp. NBC_01730 genome and checked with its own reads. Thresholds were varied one at a time by rerunning the pipeline in a copy of the repository; the published setting reproduced the released grades exactly.

Numbers and figures. Every number in this manuscript is printed from `paper/numbers.tex`, which `paper/make_numbers.py` computes from the repository's files; `paper/NUMBERS.md` maps the text to its sources. Figures were drawn with matplotlib by `paper/make_figures.py`.

Data and code availability

All scripts, intermediate tables, per-run RNA-seq results, source-read check results, proteomics results and the code that computes the numbers, tables and figures of this manuscript are public at https://github.com/papexlabs/giant-proteins (code under the MIT licence, data under CC BY 4.0) and archived at Zenodo (https://doi.org/10.5281/zenodo.24201672). Every input is public and identified by accession in Supplementary Table S1 and the data files. Large inputs (reads, raw spectra, genomes) are downloaded by the scripts and listed with their sizes in the repository's data manifest.

Statements

Author contributions. A.T.M. is the sole author. He chose the question, set the scope and the budget, approved each stage before it ran, decided which additional controls and analyses to run and which claims to keep or withdraw, and reviewed the manuscript. He is responsible for its content.

Use of AI tools. Most of the hands-on work in this study was done by an AI system under the author's direction: Claude (Anthropic; model Claude Fable 5.2), used through the Claude Code command-line tool between 11 and 13 November 2026. The AI system searched the literature and the databases, wrote all analysis code, ran it on a laptop and on rented cloud machines, proposed the grading rules and their revisions, drew the figures, compiled the tables and wrote the first draft of this text, which the author then reviewed and edited. The author did not write the code and did not recompute the results by hand. For that reason every number in the text is printed directly from the data files by a script, every rule change is listed in Supplementary Note 1, and the repository reproduces the grades from the stored results. An AI system cannot take responsibility for a paper and is not an author; errors are the author's responsibility.

Competing interests. A.T.M. is the founder of Papex Labs, which runs papex.org, a service for recording research provenance. He has no financial interest in any protein, database or sequencing product discussed here.

Funding. This work received no external funding. Cloud computing (about US$5) was paid for by Papex Labs.

Acknowledgements. This study depends entirely on open resources: UniProt, NCBI (Protein, RefSeq, GenBank, SRA), the European Nucleotide Archive, PRIDE and Pfam, and on the researchers who deposited the assemblies, reads and spectra reused here.

References

[1] Siegfried Labeit and Bernhard Kolmerer. Titins: Giant Proteins in Charge of Muscle Ultrastructure and Elasticity. *Science*, 270(5234):293–296, 1995. doi: 10.1126/science.270.5234.293.

[2] Jacob West-Roberts, Luis Valentin-Alvarado, Susan Mullen, Rohan Sachdeva, Justin Smith, Laura A. Hug, et al. Giant genes are rare but implicated in cell wall degradation by predatory bacteria. *bioRxiv*, 2023. doi: 10.1101/2023.11.21.568195. Preprint.

[3] Marie-Louise Bang, Thomas Centner, Friderike Fornoff, Adam J. Geach, Michael Gotthardt, Mark McNabb, et al. The Complete Gene Sequence of Titin, Expression of an Unusual ≈700-kDa Titin Isoform, and Its Interaction With Obscurin Identify a Novel Z-Line to I-Band Linking System. *Circulation Research*, 89(11):1065–1072, 2001. doi: 10.1161/hh2301.100981.

[4] Timothy R. Fallon, Vikram V. Shende, Igor H. Wierzbicki, Amanda L. Pendleton, Nathan F. Watervoot, Robert P. Auber, et al. Giant polyketide synthase enzymes in the biosynthesis of giant marine polyether toxins. *Science*, 385(6709):671–678, 2024. doi: 10.1126/science.ado3290.

[5] Anibal S. Amaral and Damien P. Devos. The neglected giants: Uncovering the prevalence and functional groups of huge proteins in proteomes. *PLOS Computational Biology*, 20(9):e1012459, 2024. doi: 10.1371/journal.pcbi.1012459.

[6] Oleg Reva and Burkhard Tümmler. Think big – giant genes in bacteria. *Environmental Microbiology*, 10(3):768–777, 2008. doi: 10.1111/j.1462-2920.2007.01500.x.

[7] Andreas Bachler, Thomas K. Walsh, Rahul V. Rane, and Gunjan Pandey. Chimeric misannotations of genes remain pervasive in eukaryotic non-model organisms. *BMC Genomics*, 26(1):630, 2025. doi: 10.1186/s12864-025-11765-w.

[8] Monica-Andreea Drăgan, Ismail Moghul, Anurag Priyam, Claudio Bustos, and Yannick Wurm. GeneValidator: identify problems with protein-coding gene predictions. *Bioinformatics*, 32(10):1559–1561, 2016. doi: 10.1093/bioinformatics/btw015.

[9] Yannis Nevers, Alex Warwick Vesztrocy, Victor Rossier, Clément-Marie Train, Adrian Altenhoff, Christophe Dessimoz, et al. Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*, 43(1):124–133, 2025. doi: 10.1038/s41587-024-02147-w.

[10] Michael Seeley, Wei Huang, Zhenyue Chen, William Oscar Wolff, Xueying Lin, and Xiaolei Xu. Depletion of Zebrafish Titin Reduces Cardiac Contractility by Disrupting the Assembly of Z-Discs and A-Bands. *Circulation Research*, 100(2):238–245, 2007. doi: 10.1161/01.RES.0000255758.69821.b5.

[11] Jana Kizina, Sebastian F. A. Jordan, Gerrit Alexander Martens, Almud Lonsing, Christina Probian, Androniki Kolovou, et al. *Methanosaeta* and "*Candidatus* Velamenicoccus archaeovorus". *Applied and Environmental Microbiology*, 88(7):e02407–21, 2022. doi: 10.1128/aem.02407-21.

[12] Michael R. Lazear. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *Journal of Proteome Research*, 22(11):3652–3659, 2023. doi: 10.1021/acs.jproteome.3c00486.

[13] Xavier Grau-Bové, Cristina Navarrete, Cristina Chiva, Thomas Pribasnig, Meritxell Antó, Guifré Torruella, et al. A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. *Nature Ecology & Evolution*, 6(7):1007–1023, 2022. doi: 10.1038/s41559-022-01771-6.

[14] Oliver Voigt, Magdalena V Wilde, Thomas Fröhlich, Benedetta Fradusco, Sergio Vargas, and Gert Wörheide. Genetic parallels in biomineralization of the calcareous sponge *Sycon ciliatum* and stony corals. *eLife*, 14:RP106239, 2025. doi: 10.7554/elife.106239.

[15] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

[16] Eric W Sayers, Evan E Bolton, Anna M Fine, Christopher Kelly, Sunghwan Kim, Melissa Landrum, et al. Database resources of the National Center for Biotechnology Information in 2026. *Nucleic Acids Research*, 54(D1):D20–D27, 2026. doi: 10.1093/nar/gkaf1060.

[17] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016. doi: 10.1093/nar/gkv1189.

[18] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, et al. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

[19] Sean R. Eddy. Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10):e1002195, 2011. doi: 10.1371/journal.pcbi.1002195.

[20] Martin Larralde and Georg Zeller. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics*, 39(5):btad214, 2023. doi: 10.1093/bioinformatics/btad214.

[21] James W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10(17):5303–5318, 1982. doi: 10.1093/nar/10.17.5303.

[22] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, 2021. doi: 10.1038/s41592-021-01101-x.

[23] David Yuan, Alisha Ahamed, Awais Athar, Rajkumar Devaraj, Dipayan Gupta, Muhammad Haseeb, et al. The European Nucleotide Archive in 2025. *Nucleic Acids Research*, 54 (D1):D120–D127, 2026. doi: 10.1093/nar/gkaf1295.

[24] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, 2018. doi: 10.1093/bioinformatics/bty191.

[25] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, et al. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, 2021. doi: 10.1093/gigascience/giab008.

[26] Niels Hulstaert, Jim Shofstahl, Timo Sachsenberg, Mathias Walzer, Harald Barsnes, Lennart Martens, et al. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *Journal of Proteome Research*, 19(1):537–542, 2020. doi: 10.1021/acs.jproteome.9b00328.