Genome-wide strengthening of evolutionary constraint across the volvocine multicellularity gradient

Authors Masato Tanigawa[1], [1]*Department of Biophysics, Faculty of Medicine, Oita University, 1-1 Idaigaoka, Hasama-machi, Yufu, Oita 879-5593, Japan* Corresponding author: tanigawa@oita-u.ac.jp; ORCID: 0000-0001-5877-2271

Running title: Genome-wide constraint at multicellularity origin

Keywords: Multicellularity evolution; Volvocine algae; Coding-sequence constraint; dN/dS; PAML branch model; Phylogenetic regression; Somatic maintenance

---

Abstract

The evolutionary cost of maintaining somatic cell populations has been hypothesised to drive proteome-wide strengthening of constraint on coding sequence at the origin of multicellularity. Prior work on the volvocine clade has reported lower d_N/d_S in colonial than in unicellular relatives on 55 chloroplast genes (Hu et al. 2019) and on 105 nuclear single-copy orthogroups across colonial species (Hu et al. 2020), but a nuclear genome-wide branch-model analysis together with per-species phylogenetic regression, an orthogonal amino-acid branch-length analysis with non-independence-corrected significance, a functional-category breakdown, and within-clade replication has been lacking. Here, ten volvocine green algae spanning the unicellular *Chlamydomonas reinhardtii*, the colonial intermediates *Tetrabaena*, *Gonium*, the independently evolved partial-soma lineage *Astrephomene gubernaculifera* (Goniaceae), *Yamagishiella*, *Eudorina*, *Pleodorina*, and three species of *Volvox* (*V. carteri*, *V. africanus*, *V. reticuliferus*) were analysed by branch-model d_N/d_S on 1,755 single-copy nuclear orthogroups (845 retained after quality control), an order of magnitude larger than the closest prior nuclear analysis (Hu et al. 2020). Foreground (multicellular) ω was lower than background (*Chlamydomonas*) ω in 88.6 % of orthogroups (median ω_fg = 0.046, median ω_bg = 0.117); the pattern occurred across all 20 COG functional categories analysed (range 78–100 % per category). Independent IQ-TREE amino-acid branch-length estimation under LG+Γ on 1,456 orthogroups recovered the same direction of pattern under an entirely orthogonal substitution model — 73.1 % of orthogroups had shorter aggregated multicellular terminal branches than the *Chlamydomonas* branch, and 96.1 % of orthogroups (1,395/1,451 with complete data across all ten species) exhibited a negative per-gene Spearman correlation between $\log_2$(cell number) and amino-acid branch length (species-label permutation test, 10,000 permutations: empirical one-sided $P = 0.0107$) — noting that amino-acid branch length conflates substitution rate, generation time, and selective constraint, and therefore provides a direction-of-pattern validation rather than an independent estimate of selection intensity. A per-species phylogenetic generalised least squares (PGLS) regression of log(median amino-acid branch length) on $\log_2$(cell number) was negative ($\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$, Pagel's λ_ML ≈ 0; $n = 10$). Sensitivity analyses confirmed robustness to orthogroup filtering (strict-quality $P = 0.0004$), to complete exclusion of *V. reticuliferus* (which exhibited assembly-related orthogroup-level saturation; $P = 0.006$), and to additional exclusion of *Tetrabaena* (a known

rate-elevated outlier; $P = 0.015$). Cautionary reanalyses of choanoflagellate-to-sponge and brown-algal transitions illustrate how codon-model saturation and asymmetric quality filtering can substantially distort apparent branch-model signals at deeper transitions. The ten-species result extends the previously reported chloroplast and limited-nuclear signal (Hu et al. 2019, 2020) to the nuclear proteome scale with per-species phylogenetic correction, per-gene gradient analysis, functional-category breakdown, and within-clade replication, and is consistent with — though not by itself a direct test of — the disposable-soma prediction that somatic-maintenance investment strengthens with multicellular complexity.

---

1. Introduction

The evolution of multicellularity is the central transition in the history of eukaryotic life, repeatedly originating across at least 25 independent eukaryotic lineages (Knoll 2011; Lamża 2023). Each origin requires the same fundamental innovation: the creation of a population of somatic cells whose function is to support reproduction by a small set of germ cells but which do not themselves propagate. This division of labour creates, for the first time, a class of cells whose genomic integrity must be maintained against accumulated mutations across the organism's lifetime, even though the cells in question will never directly pass their genomes to offspring (Buss 1987; Kirkwood 1977). The disposable soma theory predicts that the resulting maintenance burden should select for genome-wide strengthening of the cellular repair, proteostasis, and antioxidant machinery whose products contribute to somatic integrity (Kirkwood 1977; Williams 1957). A complementary recent formalisation in terms of bet-hedging frames this prediction in terms of an environmental-variance threshold: when extrinsic mortality is buffered below a critical level (for example, by the acquisition of multicellularity itself, of flight, or of social protection), the optimal allocation of resources shifts from damage segregation toward damage repair, and proteome-wide investment in maintenance machinery becomes favourable (Tanigawa, submitted).

The volvocine green algae provide what is arguably the most fine-grained available model system for testing these predictions at the genomic level. Within a single phylogenetic clade — and over a relatively short evolutionary time — multicellularity arose from a unicellular ancestor (*Chlamydomonas reinhardtii*) through a gradient of increasingly complex colonial intermediates: four-celled *Tetrabaena* (family Tetrabaenaceae), sixteen-celled *Gonium pectorale* and thirty-two-celled *Astrephomene gubernaculifera* (both family Goniaceae), thirty-two-celled isogamous *Yamagishiella* and anisogamous *Eudorina*, partially differentiated *Pleodorina*, and culminating in the fully germ-soma-differentiated *Volvox* clade containing among others *V. carteri*, *V. africanus* and *V. reticuliferus* (Herron and Michod 2008; Kirk 2005; Lindsey et al. 2021, 2024; Yamamoto et al. 2021; Yamashita et al. 2016). Critically for the present study, *Astrephomene gubernaculifera* has acquired partial somatic differentiation (specialised motile "rotator" cells) within Goniaceae **independently** of the partial-soma origin in Volvocaceae (*Pleodorina* + *Volvox*); *Astrephomene* therefore provides an independently evolved partial-soma lineage for comparative representation within the clade (Yamashita et al. 2016; Lindsey et al. 2021). The availability of three independently sequenced *Volvox* species (*V. carteri*, *V. africanus*, *V. reticuliferus*; Prochnik et al. 2010; Yamamoto et al. 2021) likewise provides three within-genus taxa at the maximum-cell-count, fully-differentiated end of the gradient. The sampled clade thus contains two independently evolved partial-soma lineages

and three *Volvox* species under shared genomic background, ecological setting, and analytical machinery, which permits a comparative description of how proteome-wide rate and constraint distribute across the multicellularity gradient. Crucially, per-branch synonymous divergence (d_S) across this clade remains below the codon-model saturation threshold, allowing branch-model d_N/d_S analyses to be quantitatively interpreted — unlike at deeper multicellular transitions such as animals or fungi, where the choanoflagellate-to-sponge or yeast-to-mushroom branches accumulate so many synonymous substitutions that codon-model branch lengths cease to be reliable (Bielawski et al. 2016).

Despite this exceptional analytical accessibility, prior comparative-genomic work on the volvocine multicellular transition has approached the disposable-soma prediction only piecewise rather than at the requisite genome-wide quantitative scale. The foundational genome papers — Hanschen et al. (2016) on *Gonium* and Featherston et al. (2018) on *Tetrabaena* — applied pairwise d_N/d_S and site-based positive-selection analyses to identify cell-cycle regulators (cyclin D, retinoblastoma pathway components) that have undergone selection in the volvocine lineage, and Featherston et al. used PAML branch-model 2 with a two-ratio formulation focused on *Tetrabaena* as a rate-elevated outlier; neither paper, however, contrasted the entire multicellular clade against the unicellular branch as a single foreground partition across the whole orthogroup complement. Hu et al. (2019), the first explicit unicellular-versus-colonial selection comparison in the volvocine clade, used codon-model PAML on 55 chloroplast protein-coding genes and reported 19 of 27 differentially-evolving genes as having higher d_N in unicellular *Chlamydomonadales* species than in colonial volvocines — supporting the direction of the disposable-soma prediction at organelle scale but limited to plastid coding sequence and to a comparison anchored on chloroplast rather than nuclear evolution. Hu et al. (2020) extended this approach to 105 nuclear single-copy orthologs identified by a supermatrix and coalescent-based phylotranscriptomic analysis of eight colonial volvocine species using *Chlamydomonas reinhardtii* as a reference, applying a branch-model two-ratio formulation and finding additional evidence for *T. socialis* rate-elevation. Phylotranscriptomic studies (Lindsey et al. 2021) and their fossil-calibrated continuation (Lindsey et al. 2024) resolved the topology and timing of multiple independent multicellularity origins within volvocines; comparative genomic studies of three *Volvox* species (Yamamoto et al. 2021) and of *Astrephomene gubernaculifera* (Yamashita et al. 2021) elucidated the genetic basis of sex-determining regions and the convergent evolution of germ-soma differentiation, respectively; but none of these analyses quantified per-branch genome-wide selection intensity along the multicellularity gradient. None of these prior studies have provided (i) a nuclear genome-wide test at the >1,000-orthogroup scale required to characterise proteome-wide constraint, (ii) a per-species PGLS regression of the constraint metric against cell number, (iii) a per-gene gradient analysis quantifying whether constraint scales with cell number at the single-orthogroup level, (iv) an orthogonal amino-acid branch-length analysis with a non-independence-corrected significance test, (v) a systematic per-functional-category breakdown of the constraint signal, or (vi) taxon sampling that includes the independently evolved partial-soma lineage *Astrephomene gubernaculifera* (Goniaceae) alongside three *Volvox* species at the slow-evolving end of the gradient. The disposable-soma prediction at the proteome scale therefore remains to be tested with the orthogroup coverage, statistical power, methodological orthogonality, and convergent replication appropriate to the strength of the underlying biological hypothesis.

Three related questions persist beyond the basic disposable-soma prediction. First, whether the strengthening, if present, is **distributed across all functional categories** (consistent with system-wide somatic maintenance as an emergent property of the multicellular state) or **concentrated in canonical maintenance pathways** (DNA repair, proteostasis, antioxidant defence). Second, whether the strengthening **scales monotonically with cell number** along the volvocine gradient or is a **categorical step-change** at the unicellular-to-multicellular boundary. Third, whether the apparent strengthening is **confounded by other lineage-specific factors** (population-size differences, generation-time differences, codon-usage bias, synonymous-site saturation) that operate independently of the multicellularity transition itself.

Here we address these gaps through a 10-species branch-model d_N/d_S analysis of 1,755 single-copy nuclear orthogroups across the volvocine gradient — including *Astrephomene gubernaculifera*, an independently evolved partial-soma lineage in family Goniaceae; *V. africanus* (NIES-4468) and *V. reticuliferus* (NIES-3785), which extend within-*Volvox*-genus sampling at the fully-differentiated maximum-cell-count end of the gradient; complemented by an orthogonal amino-acid branch-length analysis using IQ-TREE under the LG+Γ substitution model. The combination of (i) taxon sampling across the volvocine multicellularity gradient including three *Volvox* species and two independently evolved partial-soma lineages, (ii) a nuclear orthogroup set an order of magnitude larger than the closest prior nuclear analysis (Hu et al. 2020), (iii) per-branch d_S below the codon-model saturation threshold, (iv) orthogonal amino-acid branch-length analysis under LG+Γ that does not depend on synonymous-site assumptions, (v) per-orthogroup gradient correlation analysis on all 1,451 complete-data orthogroups with a species-label permutation null distribution to account for orthogroup non-independence, (vi) per-species phylogenetic generalised least squares (PGLS) regression with branch lengths estimated *de novo* from a concatenated 1.4-million-amino-acid supermatrix (Methods §2.6), (vii) functional-category breakdown across 20 COG categories, and (viii) explicit boundary-condition reanalyses of choanoflagellate-to-sponge and brown-algal transitions to characterise the methodological limits, allows the directional prediction implied by the disposable-soma framework to be tested at the proteome scale with explicit controls for the major potential statistical and methodological confounds. We find that volvocine complexity is associated with genome-wide reductions in d_N/d_S (88.6 % of orthogroups) and shorter amino-acid terminal branch length (73.1 % of orthogroups), with a significant negative species-level association between log(median amino-acid branch length) and $\log_2$(cell number) (PGLS $\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$, $\lambda\_ML = 0$). The pattern is reproduced across all 20 functional categories examined, is consistent with the disposable-soma prediction at the proteome scale, and extends the direction of finding reported in chloroplast (Hu et al. 2019) and limited-nuclear (Hu et al. 2020) studies to a nuclear-proteome scale.

---

2. Materials and Methods

2.1 Genome assemblies and gene annotations

Ten volvocine green algal genomes spanning the unicellular-to-multicellular gradient were obtained from public repositories: *Chlamydomonas reinhardtii* (GCF_000002595.2; unicellular reference, NCBI RefSeq), *Tetrabaena socialis* (GCA_002891735.1; 4 cells; Featherston et al. 2018),

Gonium pectorale (GCA_001584585.1; 16 cells; Hanschen et al. 2016), *Astrephomene gubernaculifera* (GCA_021605115.1; 32 cells with independently evolved rotator-cell partial soma in family Goniaceae; Yamashita et al. 2016, 2021), *Yamagishiella unicocca* (GCA_003116995.1; 32 cells; Hamaji et al. 2018), *Eudorina sp.* (GCA_003117195.1; ~48 cells; Hamaji et al. 2018), *Pleodorina starrii* (GCA_030267565.1; ~96 cells with partial soma; DDBJ deposit), *Volvox carteri* (GCF_000143455.1; ~2,000 cells with full germ-soma differentiation; Prochnik et al. 2010; NCBI RefSeq), *Volvox africanus* NIES-4468 (GCA_030268105.1, Thailand strain "1101-NZ-11"; ~1,500 cells, homothallic with germ-soma differentiation; deposited DDBJ assembly with N50 = 3.95 Mb), and *Volvox reticuliferus* NIES-3785 (GCA_019650235.1, female strain from Lake Biwa, Japan; ~1,500 cells, heterothallic with germ-soma differentiation; Yamamoto et al. 2021, N50 = 1.91 Mb). The inclusion of *Astrephomene* — whose partial-soma origin in Goniaceae is phylogenetically independent of the Volvocaceae partial-soma origin (Yamashita et al. 2016, 2021; Lindsey et al. 2021) — provides an independently evolved partial-soma lineage within the clade; the inclusion of three *Volvox* species (*V. carteri*, *V. africanus*, *V. reticuliferus*) provides three within-genus taxa at the maximum-cell-count, fully-differentiated end of the gradient.

For the two assemblies lacking deposited gene models (*Yamagishiella*, *Eudorina*), gene structures were predicted *de novo* using Augustus (Stanke and Waack 2003) trained on *Chlamydomonas reinhardtii*. The *Astrephomene*, *V. africanus* and *V. reticuliferus* assemblies are accompanied by deposited NCBI gene-model GFFs. Annotation completeness was verified by independent QC: *Yamagishiella* (14,445 predicted loci; median CDS 1,137 bp; 97.1 % complete ORFs), *Eudorina* (16,316 loci; median CDS 1,131 bp; 96.4 % complete ORFs), *Pleodorina* (15,644 loci; median CDS 1,434 bp; passing all length-based QC), *Astrephomene* (deposited gene models; ~10,800 predicted CDS recovered after re-extraction with gffread for methodological consistency), *V. africanus* NIES-4468 (deposited gene models; 13,455 longest-isoform proteins per gene), and *V. reticuliferus* NIES-3785 (deposited gene models; 22,374 protein entries reduced to 13,306 unique loci by longest-isoform-per-gene filtering) all met thresholds appropriate for downstream comparative analysis (Table S1).

2.2 Orthology assignment

Protein sequences were extracted from genome assemblies and gene annotations using gffread (Pertea and Pertea 2020). For *V. africanus* and *V. reticuliferus*, the deposited NCBI protein.faa files were filtered to retain the longest protein per locus_tag using a custom Python script. OrthoFinder v2.5.5 (Emms and Kelly 2019) was run in MSA mode (`-M msa`) with diamond ultra-sensitive search. The 10-species run identified 1,755 single-copy orthogroups (exactly one protein per species across all ten species), which were retained for downstream branch-model analysis.

2.3 Codon alignment and PAML branch-model analysis

For each single-copy orthogroup, coding sequences were extracted from genome assemblies, translated, aligned at the amino-acid level using MAFFT v7 (Katoh and Standley 2013), and back-projected to codon-level alignment. Alignments shorter than 90 codons were excluded.

The species topology followed the consensus of Hanschen et al. (2018), Hamaji et al. (2018), Lindsey et al. (2021, 2024), Yamashita et al. (2016), and Yamamoto et al. (2021), grouping

Astrephomene gubernaculifera with *Gonium pectorale* in family Goniaceae, and grouping the three *Volvox* species (with *V. africanus* and *V. reticuliferus* as sister taxa) within Volvocaceae:

```
(chlamydomonas_reinhardtii,
 (tetrabaena_socialis,
  ((gonium_pectorale, astrephomene_gubernaculifera),
   (yamagishiella_unicocca,
    (eudorina_sp,
     (pleodorina_starrii,
      (volvox_carteri, (volvox_africanus,
volvox_reticuliferus)))))))));
```

PAML codeml v4.10.10 (Yang 2007) was used to fit two nested models. Model 0 (one-ratio) estimated a single ω (= d_N/d_S) across all branches. Model 2 (branch model) estimated separate ω values for foreground (all nine terminal multicellular branches plus the seven internal branches of the multicellular clade, all labeled #1) versus background (the single terminal *Chlamydomonas* branch).

Quality control excluded orthogroups with ω estimates at PAML's numerical boundaries ($\omega \leq 0.0001$ or $\omega \geq 10$), which indicate non-converged or degenerate estimates. Of 1,755 orthogroups analysed, 845 passed QC (Table S2). Per-branch synonymous divergence was inspected to verify that median per-branch $d\_S \approx 1.4$ across the volvocine tree, well below the codon-model saturation threshold (Bielawski et al. 2016).

2.4 IQ-TREE amino-acid validation

To verify that the branch-model signal was not an artefact of the codon-substitution model assumptions, terminal branch lengths were independently estimated from amino-acid alignments under the LG+G4 substitution model (Le and Gascuel 2008) using IQ-TREE v3.0.1 (Minh et al. 2020) with the species topology fixed. The amino-acid alignment for each orthogroup was generated by translating the PAML codon alignment using the standard genetic code, with gaps and ambiguous codons mapped to "-" and "X" respectively. 1,456 of the 1,755 single-copy orthogroups (83.0 %) passed IQ-TREE estimation. For each orthogroup, the mean terminal branch length across the nine multicellular species was compared with the *Chlamydomonas reinhardtii* terminal branch length. Per-species median terminal branch length was used as the primary species-level summary statistic for downstream PGLS analysis (§2.6); 256 of 1,453 *V. reticuliferus*-bearing orthogroups (17.6 %) exhibited terminal branch length saturation (capped at IQ-TREE's default upper bound of 10 substitutions per site) consistent with annotation-derived ortholog misassignment in this draft-quality assembly (N50 1.91 Mb), with cross-species inspection of these orthogroups showing normal branch lengths in the close relatives *V. africanus* and *V. carteri*; the median statistic remains robust to this saturation (sensitivity analyses below).

2.5 Per-orthogroup gradient correlation

For each orthogroup with successful IQ-TREE estimation, the Spearman rank correlation was computed between $\log_2$(cell number) — taking values 0, 2, 4, 5, 5, 5.58, 6.58, 10.55, 10.55, 10.97 for *Chlamydomonas*, *Tetrabaena*, *Gonium*, *Astrephomene*, *Yamagishiella*, *Eudorina*, *Pleodorina*, *V. africanus*, *V. reticuliferus*, *V. carteri* respectively — and log(terminal amino-acid branch

length). Note that *Astrephomene* and *Yamagishiella* share the same nominal cell number (32) but represent independent multicellularity contexts (Goniaceae partial soma vs. Volvocaceae isogamous colony, respectively); the three *Volvox* species share similar nominal cell numbers (1,500–2,000) but represent independently-sequenced taxa. Cell numbers for *V. africanus* (NIES-4468) and *V. reticuliferus* (NIES-3785) followed published morphological characterisations (Nozaki et al. 2022 for NIES-4468; Yamamoto et al. 2021 for NIES-3785). The number and proportion of orthogroups with negative correlation was reported with both (a) a binomial test against the null expectation of 0.5 — noting that this assumes orthogroup-level independence not satisfied by the data — and (b) a species-label permutation test (10,000 permutations) in which the cell-number labels were randomly reassigned across the ten species and the per-OG negative-correlation fraction was recomputed under each permutation, yielding a non-independence-corrected one-sided empirical *P* value (permutation_test_10sp.R; results in §3.4).

2.6 Phylogenetic generalised least squares (PGLS) regression

Per-species median amino-acid branch length (across the 1,456 orthogroups) was regressed against $\log_2$(cell number) under PGLS using the caper R package v1.0.3 (Orme et al. 2018) with Pagel's λ ML-estimated. The species topology of Section 2.3 was used, with branch lengths estimated *de novo* on a concatenated supermatrix: the 1,456 per-OG amino-acid alignments were concatenated into a single alignment of approximately 1.4 million amino-acid columns, IQ-TREE LG+G4 was run with topology fixed (option `-te`) to estimate branch lengths in substitutions per site (the same unit as the per-OG IQ-TREE branches), and the resulting tree was rooted on *Chlamydomonas reinhardtii* as the unicellular outgroup. This avoids the methodological criticism that PGLS results depend on an arbitrary choice of branch lengths in an externally supplied tree. The supermatrix-derived branch lengths and the constraint topology are deposited alongside the analysis scripts (see Data and code availability). Both the full-species PGLS (n = 10) and a PGLS excluding *Tetrabaena socialis* (n = 9) — known from prior phylogenomic comparisons to be an evolutionary-rate outlier (Featherston et al. 2018; Hu et al. 2020) — were reported as primary analyses. Sensitivity analyses additionally tested (i) orthogroup filtering at progressively stricter quality thresholds (all species' branch lengths below 1.0, 0.5, or 0.3); (ii) complete exclusion of *V. reticuliferus* (whose draft-quality assembly produced orthogroup-level saturation; §2.4); and (iii) varying cell-number nominal assignments for the three *Volvox* species (1,500–4,000 cells per spheroid). All sensitivity scenarios are reported in Results §3.4.

2.7 Functional annotation and category-level analysis

Existing eggNOG-mapper v2.1.13 (Cantalapiedra et al. 2021) annotations for the five species with publicly available gene models (*Chlamydomonas*, *Tetrabaena*, *Gonium*, *Pleodorina*, *V. carteri*) were used to assign COG functional categories to orthogroups; the consensus category across the annotated species per orthogroup was taken as the OG-level assignment. (*Astrephomene*, *Yamagishiella*, *Eudorina*, *V. africanus*, and *V. reticuliferus* gene models are present in the OrthoFinder run but were not separately re-annotated with eggNOG-mapper for this revision; the existing five-species annotation provides robust OG-level COG assignments.) This allowed COG categorisation of approximately 770 of 845 (~91 %) post-QC orthogroups. For each COG category with ≥5 orthogroups, the proportion with ω_fg < ω_bg was reported with 95 %

Clopper-Pearson binomial confidence intervals and a binomial test against the null = 0.5, with multiple-testing correction by the Benjamini–Hochberg method.

2.8 Effective population size (N_e) consideration

A potential alternative explanation for genome-wide ω reduction on multicellular branches is variation in effective population size: smaller N_e in multicellular organisms could in principle alter selection efficiency. However, the direction of this effect runs *opposite* to the observed reduction: smaller N_e increases the relative weight of drift over selection, allowing more slightly deleterious nonsynonymous mutations to drift to fixation, which would *increase* (not decrease) ω. The observed decrease of ω on multicellular branches is therefore not parsimoniously explained by reduced N_e.

2.9 Boundary-condition analyses

To illustrate where codon-model branch analyses become unreliable, two additional multicellular transitions were analysed under the same PAML branch-model framework. The choanoflagellate-to-sponge/cnidarian transition (*Monosiga brevicollis*, *Salpingoeca rosetta* → *Amphimedon queenslandica*, *Nematostella vectensis*) had median per-branch d_S ~225, far above saturation, and the brown algal pair *Ectocarpus siliculosus*–*Nannochloropsis gaditana* had median d_S ~72. For the brown-algae pair, the PAML branch-model output was analysed under two filtering regimes: symmetric QC (excluding orthogroups with ω at the 0.0001 or 10 boundary on either side) and asymmetric QC (excluding only orthogroups with ω_bg at the boundary, retaining ω_fg-at-boundary cases). For the animal transition, IQ-TREE amino-acid branch lengths were additionally estimated under LG+G4 to verify whether the apparent codon-model signal persisted at the amino-acid level.

2.10 Reproducibility

All analysis scripts, intermediate data files, PAML output, IQ-TREE output, and final summary tables are deposited at Zenodo (DOI: 10.5281/zenodo.21094823; assigned before submission). Augustus gene predictions for *Yamagishiella* and *Eudorina* are included in the Zenodo record.

---

3. Results

3.1 Genome-wide reduction in estimated ω on multicellular branches

Branch-model d_N/d_S estimation on 845 single-copy orthogroups (post-QC; 1,755 total) recovered a genome-wide reduction of estimated ω on multicellular branches relative to the *Chlamydomonas* terminal branch: median ω_fg = 0.046 was 2.5-fold lower than median ω_bg = 0.117 (Fig. 1A), and **88.6 %** (749/845) of orthogroups had ω_fg < ω_bg (Fig. 1B), with the distribution of $\log_{10}(\omega\text{\_fg}/\omega\text{\_bg})$ shifted to negative values (Fig. 1C). We do not report a binomial *P* value on these 845 orthogroups because they share the same species, the same species tree, and the same lineage-specific rate effects and are not statistically independent; the non-independence-corrected significance level for the directional pattern is provided by the species-label permutation test at the per-OG gradient level (§3.4 below). The fractional reduction in the ten-species panel is slightly larger than the seven- and eight-species

formulations reported in prior versions of this analysis (83.2 % at n = 7; 85.2 % at n = 8), indicating that the inclusion of three independent *Volvox* species and the within-clade convergent partial-soma species *Astrephomene* strengthens rather than dilutes the headline signal.

3.2 The pattern is distributed across all functional categories

Across the 20 COG functional categories with ≥5 categorised orthogroups (Fig. 3; Table S3), every category showed majority ω_fg < ω_bg. The lowest fractions in the ten-species analysis were Q (Secondary metabolites biosynthesis, transport and catabolism; 78 %, n = 18) and K (Transcription; 81 %, n = 31); the highest were Z (Cytoskeleton), D (Cell cycle control, cell division, chromosome partitioning), B (Chromatin structure and dynamics), M (Cell wall/membrane/envelope biogenesis), and P (Inorganic ion transport and metabolism), all 100 % (n = 24, 13, 6, 6, 25 respectively). The top of the per-category ranking is occupied by maintenance-relevant categories (Z cytoskeleton, D cell cycle, B chromatin, M cell wall, P ion transport, T signal transduction 85.7 %). All 20 categories had proportions above 0.5, and within-category binomial tests against the 0.5 null remained significant after Benjamini–Hochberg correction across the 20 categories (Table S3); for two small categories (B chromatin and M cell wall, n = 6 each), the BH-adjusted P was 0.031, the largest BH-adjusted P in the table. Within-category binomial tests share the same caveat about orthogroup non-independence noted in §3.1 and §3.4. The pattern is therefore not driven by any specific functional class.

3.3 IQ-TREE amino-acid branch-length analysis recovers the same direction of pattern

IQ-TREE amino-acid branch-length estimation under LG+Γ on 1,456 orthogroups (83.0 % IQ-TREE success rate from 1,755 single-copy orthogroups) recovered the same direction of pattern as the codon-model analysis under an entirely orthogonal substitution model: 73.1 % of orthogroups had mean multicellular terminal branch length shorter than the *Chlamydomonas* branch length (Fig. 2A; Fig. S1). Per-species median terminal branch lengths were *Chlamydomonas reinhardtii* 0.264, *Tetrabaena socialis* 0.322, *Gonium pectorale* 0.211, *Astrephomene gubernaculifera* 0.239, *Yamagishiella unicocca* 0.147, *Eudorina sp.* 0.148, *Pleodorina starrii* 0.122, *Volvox carteri* 0.132, *V. africanus* 0.055, *V. reticuliferus* 0.076 (substitutions per site). The two newly added *Volvox* species, *V. africanus* (0.055) and *V. reticuliferus* (0.076), have the two shortest values in the panel; *V. carteri* (0.132) also lies toward the slow-evolving end, although slightly above *Pleodorina* (0.122). We emphasise that amino-acid terminal branch length is the product of substitution rate, divergence time, generation time, mutation rate, and selective constraint, and therefore provides an orthogonal direction-of-pattern check on the codon-model result rather than an independent estimate of selection intensity (Methods §2.4; see also §4.6).

3.4 Gradient relationship with cell number

The per-orthogroup Spearman rank correlation between $\log_2$(cell number) and log(amino-acid terminal branch length) was negative in 96.1 % (1,395/1,451 orthogroups with complete data across all ten species; Fig. 2B), higher than the 93.6 % obtained in the eight-species formulation and the 93.1 % in the seven-species formulation. The progressive strengthening of the per-OG gradient correlation across panel sizes (93.1 % at n = 7; 93.6 % at n = 8; 96.1 % at n = 10) is

consistent with the inclusion of two additional *Volvox* species occupying the slow-evolving end of the gradient refining rather than diluting the signal.

We emphasise that the 1,451 orthogroups are not statistically independent — they share the same species, the same cell-number vector, the same species tree, and the same lineage-specific rate effects — so the binomial test against a 50 % null assumes an independence that the data do not satisfy. To obtain a non-independence-corrected significance level we conducted a species-label permutation test in which the cell-number labels were randomly reassigned among the ten species 10,000 times, with the per-OG Spearman rho fraction recomputed under each permutation (Methods §2.6; permutation_test_10sp.R). The null distribution has mean 0.490, 95th percentile 0.915, and 99th percentile 0.961 (maximum 0.990); the observed value 0.961 lies at approximately the 99th percentile of the null, corresponding to an empirical one-sided $P = 0.0107$ (107 of 10,000 permutations were as extreme or more extreme than the observed). The signal is therefore well above the chance expectation but is substantially less extreme than the naive binomial test would suggest, and should be interpreted as a sign-consistency statistic rather than as a test based on 1,451 fully independent observations.

At the per-species level (Fig. 2A), the phylogenetic generalised least squares (PGLS) regression of log(median amino-acid branch length) on $\log_2$(cell number) with Pagel's λ ML-estimated returned a significant negative species-level association ($\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$, Pagel's λ_ML ≈ 0); the analysis excluding *Tetrabaena socialis* — an established evolutionary-rate outlier with elevated *d_N* in the Tetrabaenaceae lineage (Featherston et al. 2018; Hu et al. 2020) — returned a similar slope ($\beta = -0.113$, $P = 0.006$, $n = 9$, $R^2 = 0.68$, Pagel's λ_ML ≈ 0). This is a stronger PGLS result than in seven- and eight-species formulations of the analysis ($\beta = -0.06$ to -0.075, $P = 0.064$–0.13). The species-tree branch lengths used in the PGLS were estimated *de novo* from the concatenated approximately 1.4-million-amino-acid supermatrix (Methods §2.6). The ML estimate of Pagel's λ converged to zero in both the all-ten-species and the *Tetrabaena*-excluded analyses, indicating that the PGLS reduces to ordinary least squares (OLS) and that no residual phylogenetic covariance remained to be modelled; we therefore present OLS fits in Fig. 2A and note that they are identical to the PGLS fit at λ = 0. We note three cautionary considerations for interpreting the PGLS: (i) the species panel is small ($n = 10$) and includes three *Volvox* species clustered near the same cell-number value; (ii) the response variable (median amino-acid terminal branch length) is not a pure measure of selection intensity (§4.6); and (iii) the same supermatrix from which the PGLS covariance structure was derived was also used to estimate the response. Sensitivity analyses (next paragraph) address robustness to species exclusion, orthogroup filtering, and nominal cell-number assignment.

Sensitivity analyses (Table S4b) confirmed that the PGLS result is robust to multiple analytic choices. Orthogroup quality filtering at progressively stricter thresholds (all species' branch lengths < 1.0, 0.5, or 0.3) strengthened rather than weakened the result: at the strictest filter ($n = 383$ high-quality orthogroups), $\beta = -0.141$, $P = 0.0004$. Complete exclusion of *V. reticuliferus* (which exhibited assembly-related orthogroup-level saturation in 17.6 % of its orthogroups, Methods §2.4) gave $\beta = -0.121$, $P = 0.006$ ($n = 9$ species); additionally excluding the *Tetrabaena* rate outlier gave $\beta = -0.110$, $P = 0.015$ ($n = 8$ species). Varying the nominal cell-number assignment of the three *Volvox* species across the literature range (800–4,000) yielded slopes between −0.111 and −0.130 and *P* values between 0.0016 and 0.0030. The PGLS gradient result is therefore not an artefact of the *V. reticuliferus* saturation, of the *Tetrabaena* outlier, of orthogroup quality filtering choices, or of nominal cell-number assignments.

3.5 Boundary conditions: where the codon model breaks down

To illustrate the limits of branch-model d_N/d_S analysis at deeper multicellular transitions, the same pipeline was applied to two additional systems with median per-branch d_S far above the volvocine value of ~1.4 (Fig. S2; Table S4):

- **Brown algae** (*Ectocarpus–Nannochloropsis*, median per-branch d_S ~72): under symmetric QC, only 46.7 % (107/229) of orthogroups had $\omega_{fg} < \omega_{bg}$ — essentially chance. Under asymmetric QC (excluding only ω_{bg}-at-boundary cases), the apparent proportion rose to 73.7 % (407/552), illustrating that filtering decisions alone can substantially alter or exaggerate the inferred directional proportion in saturated systems.
- **Animals** (choanoflagellates → sponge/cnidarian, median per-branch d_S ~225): post-QC PAML returned 36.5 % (below chance), and IQ-TREE amino-acid analysis returned 51.6 % (661/1,280) — at the chance expectation, so that a directional excess was not evident at this level of saturation under either method.

These boundary conditions underscore the methodological value of the volvocine system's low d_S: only there can branch-model and IQ-TREE results be quantitatively trusted.

---

4. Discussion

4.1 Genome-wide evolutionary-rate and d_N/d_S gradient across volvocine complexity

The central observation of this study is that the evolution of multicellularity in the volvocine lineage is associated with a clear, reproducible reduction in estimated d_N/d_S and in amino-acid evolutionary rate on multicellular branches relative to the *Chlamydomonas* terminal branch, distributed across the proteome rather than concentrated in any single functional category. The directional pattern — multicellular branches under lower estimated ω and shorter amino-acid branch length than the unicellular reference — was previously reported at smaller scale in chloroplast (Hu et al. 2019; 55 genes) and limited-nuclear (Hu et al. 2020; 105 orthogroups) analyses; the present work extends this finding to a nuclear gene set an order of magnitude larger and adds (1) a per-species PGLS regression ($\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$, $n = 10$), (2) per-orthogroup gradient analysis, (3) an orthogonal amino-acid branch-length analysis, (4) functional-category breakdown, and (5) replication at three independent *Volvox* species and two independent partial-soma origins (see §4.3). Branch-model d_N/d_S estimation across 845 quality-controlled single-copy orthogroups recovered a 2.5-fold reduction in median ω on multicellular branches relative to the *Chlamydomonas* branch (0.046 vs 0.117), with 88.6 % of orthogroups showing the same directional shift. Across the 20 COG functional categories with sufficient orthogroup representation, every category showed majority $\omega_{fg} < \omega_{bg}$, with proportions ranging from 78 % (secondary metabolism, Q) to 100 % (cytoskeleton Z, cell cycle D, chromatin B, cell wall M, ion transport P). Amino-acid branch-length estimation under IQ-TREE LG+Γ recovered the same direction of pattern under an orthogonal substitution model: 73 % of orthogroups had shorter aggregated multicellular terminal branches than the *Chlamydomonas* branch, and 96.1 % (1,395/1,451 orthogroups with complete data across all ten species) showed a negative Spearman rank correlation between $\log_2$(cell number) and amino-

acid branch length (species-label permutation test, 10,000 permutations: empirical one-sided P = 0.0107; see §3.4). The directional pattern thus scales gradationally with the multicellularity grade, rather than as a categorical step at the unicellular-to-multicellular boundary.

The convergence of the codon-model branch-model result, the amino-acid branch-length analysis with a significant negative PGLS slope, and the per-gene gradient correlation provides quantitative nuclear-genome-wide evidence consistent with — though not by itself a direct test of — the disposable-soma prediction at the origin of multicellularity. We discuss the strength and the limitations of this interpretation explicitly in §4.5–4.6 below.

The three *Volvox* species (*V. carteri*, *V. africanus*, *V. reticuliferus*) generally occupy the slow-evolving end of the per-species median branch-length distribution: *V. africanus* (0.055) and *V. reticuliferus* (0.076) are the two shortest values in the panel, while *V. carteri* (0.132) lies toward the low-rate end but is slightly above *Pleodorina* (0.122). *Chlamydomonas* (0.264) is the longest. The inclusion of *Astrephomene gubernaculifera* (32 cells with motile "rotator" cells; Goniaceae) provides an independent phylogenetic representation of partial-soma evolution within the volvocine clade: *Astrephomene*'s per-species median amino-acid terminal branch length (0.239) is between *Gonium pectorale* (0.211) and the Volvocaceae partial-soma species (*Pleodorina* 0.122, *Volvox carteri* 0.132). We note that *Astrephomene*'s value does not by itself demonstrate a reduction relative to its close relative *Gonium*, so a single-pair reduction signal at the partial-soma transition cannot be claimed; the sampled clade nonetheless includes two independently evolved partial-soma lineages (Goniaceae *Astrephomene* and Volvocaceae *Pleodorina*/*Volvox*) and three independent *Volvox* species, and the genome-wide pattern is associated with multicellularity grade rather than with any single species-pair contrast.

4.2 The *Tetrabaena socialis* exception illuminates rather than overturns the pattern

A single species — *Tetrabaena socialis* — is a visually obvious exception to the gradient pattern: its median amino-acid terminal branch length (0.322) exceeds that of unicellular *Chlamydomonas* (0.264), and its inferred branch-model ω is higher than expected from the multicellularity gradient. This elevation is, however, an independently established feature of *T. socialis* in prior comparative work: phylotranscriptomic analyses have repeatedly identified *Tetrabaena* as a non-synonymous substitution-rate outlier within the volvocine clade (Featherston et al. 2018; Hu et al. 2020). Two non-exclusive interpretations apply: a smaller effective population size in *Tetrabaena* reduces the efficacy of purifying selection regardless of multicellularity-driven constraint (the lineage-specific N_e-and-drift hypothesis), or at the low cell count of *Tetrabaena* (four cells per colony), the somatic-maintenance burden may be too small to outweigh other selective pressures (the dose-response interpretation). In either case, the broader genome-wide pattern is robust to the *Tetrabaena* exception: the per-orthogroup Spearman correlation is negative in 96.1 % of complete-data orthogroups in the ten-species analysis (species-label permutation P = 0.0107; §3.4), and the per-species PGLS slope is negative and statistically significant both with all 10 species (β = −0.123, P = 0.002) and with *Tetrabaena* excluded (β = −0.113, P = 0.006, n = 9). The *Tetrabaena* observation thus qualifies the per-species visualisation without affecting the central conclusion derived from either the orthogroup-level evidence or the *Tetrabaena*-excluded PGLS.

4.3 Relationship to prior comparative-genomic work

This study complements rather than supersedes earlier volvocine-comparative work, but extends it in three important ways. First, with respect to identification of innovation-associated loci, Hanschen et al. (2016) and Featherston et al. (2018) identified specific genes — cyclin D, retinoblastoma pathway components, and ubiquitin-proteasomal effectors — as having undergone positive or accelerated selection during the multicellularity transition. These findings are not in tension with the present result: positive selection on a small number of innovation-associated loci can co-occur with proteome-wide strengthening of purifying selection on the much larger background gene set, and indeed our COG-level analysis shows that the constraint signal is present even in cell-cycle-control (D, 94.7 % with ω_fg < ω_bg) and chromatin-structure (B, 100 %) categories where Hanschen et al. identified key innovations. The two findings together describe a layered picture: innovation in a small number of loci that defined the multicellular state, layered on top of strengthened maintenance pressure across the remainder of the proteome.

Second, with respect to selection-rate contrast between unicellular and colonial volvocines, Hu et al. (2019) reported in *Frontiers in Microbiology* on 55 chloroplast protein-coding genes that 19 of 27 differentially-evolving genes showed higher *d_N* in unicellular *Chlamydomonadales* species than in colonial volvocines — a directional signal consistent with the present finding but obtained from the plastid genome rather than the nucleus, and from a much smaller gene set (55 chloroplast genes vs 1,755 nuclear orthogroups here). Hu et al. (2020) in *European Journal of Phycology* extended this approach to the nuclear genome by identifying 105 single-copy nuclear orthologs across eight colonial volvocine species (no unicellular relative was used as a foreground/background partition; *C. reinhardtii* was used as a reference for sequence assignment), applying a PAML branch-model two-ratio test focused on *Tetrabaena socialis* as a rate-elevated outlier within the colonial set. The present analysis differs from and extends the Hu et al. (2019, 2020) work in six concrete ways: (i) the nuclear gene set is approximately 17-fold larger (1,755 vs 105 single-copy nuclear orthogroups), enabling the directional pattern to be characterised at proteome-wide rather than gene-level scale; (ii) the branch-model 2 foreground/background partition contrasts all multicellular branches against the unicellular *Chlamydomonas* terminal branch, recovering an explicit multicellular-vs-unicellular contrast rather than a within-colonial rate-elevation contrast; (iii) the per-orthogroup gradient correlation against $\log_2$(cell number) — not previously reported in the volvocine literature — provides strong directional consistency across orthogroups (96.1 % negative; species-label permutation $P = 0.0107$ in 1,451 complete-data orthogroups, with non-independence accounted for); (iv) the per-species PGLS regression provides a significant species-level association ($\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$; Results §3.4); (v) the IQ-TREE LG+Γ amino-acid branch-length analysis provides methodological orthogonality not present in any prior volvocine selection study; and (vi) the inclusion of *Astrephomene gubernaculifera* and three *Volvox* species adds an independent phylogenetic representation of partial-soma evolution and three within-genus *Volvox* taxa at the slow-evolving end of the gradient. The present results therefore corroborate and extend the directional finding of Hu et al. (2019, 2020) from the plastid and limited-nuclear scale to the nuclear-proteome scale, with the additional gradient, branch-length, and taxon-replication analyses providing the supporting structure for the directional claim.

The present analysis also bears on three recent comparative-genomic studies of mammalian longevity that report related findings using fundamentally different evolutionary contexts. Kowalczyk et al. (2020) and Zhang & Kronforst (2025, preprint) report that genes related to cell cycle, DNA repair, and immunity evolve under increased purifying selection in exceptionally long-lived mammals; Zhang et al. (2026, BMC Biology) similarly identify cell-division and DNA-repair functions as enriched among lifespan-correlated evolutionary-rate shifts in a 968-species mammalian dataset. The volvocine result presented here is consistent with the parsimony hypothesis that the same evolutionary logic — strengthened maintenance in proportion to somatic-cell exposure to mutation — operates in both contexts. A key conceptual contribution of the present work is to demonstrate that, in the volvocine system where the codon-model can be quantitatively trusted, the strengthening signal is genuinely proteome-wide and not pathway-specific. By implication, mammalian studies that report pathway-specific enrichment using relative-evolutionary-rate-normalised methods may underestimate the genome-wide component of the constraint, because relative-evolutionary-rate (RER) normalisation by construction removes signal proportional to overall branch length and therefore absorbs any phenotype-correlated genome-wide rate shift (see also §4.6).

4.4 Lineage-specific allocation within the volvocine multicellular clade

The branch-model 2 analysis (Sections 3.1–3.4) treats all nine multicellular branches as a single foreground partition, recovering the genome-wide constraint signal that is the primary result of this study. A complementary free-ratio analysis (PAML Model 1) estimating an independent ω on each branch (Table S5) reveals informative heterogeneity in per-species median ω among the nine multicellular species. Among the multicellular species (the *Chlamydomonas* estimate is biased by d_S saturation; ~75 % of *Chlamydomonas* per-orthogroup ω values are at the PAML lower boundary and are excluded by quality control, leaving only 353 of 1,664 *Chlamydomonas* ω values in the QC-passing set), per-species median ω values were: *Tetrabaena socialis* 0.096 (the known evolutionary-rate outlier; see §4.2), *Gonium pectorale* 0.073, *Astrephomene gubernaculifera* 0.074, *Yamagishiella unicocca* 0.066, *Eudorina sp.* 0.065, *Pleodorina starrii* 0.082, *Volvox carteri* 0.060, *V. africanus* 0.089, and *V. reticuliferus* 0.144. **The fully germ-soma-differentiated *Volvox carteri* exhibits the lowest median ω among all multicellular species (0.060)**, consistent with the disposable-soma prediction that the species with the most extensive germ-soma differentiation experiences the strongest additional constraint above the multicellular-vs-unicellular baseline. *V. carteri*'s strongest-constraint position is consistent with its complete germ-soma differentiation (regA-dependent; Kirk 2005; Hanschen et al. 2016) and with its IQ-TREE rank in the gradient.

The two other *Volvox* species show a more complex pattern that highlights the value of independent biological replication. *V. africanus* (median ω 0.089) exhibits a typical *Volvox* free-ratio ω despite the shortest overall amino-acid terminal branch length in the panel (0.055; §3.3) — interpretable as a slow-evolving genome with a normal d_N/d_S ratio, rather than as the strongest purifying selection. *V. reticuliferus* (median ω 0.144) exhibits an elevated free-ratio ω relative to other *Volvox* species, consistent with the *V. reticuliferus*-specific orthogroup-level saturation discussed in §2.4 and below (§4.6): the *V. reticuliferus* QC-passing set of 1,294 orthogroups is enriched for orthogroups in which *V. reticuliferus* showed intermediate divergence (rather than saturation), and these intermediate-divergence orthogroups appear to include a tail of orthologs with elevated ω (possibly heterothallic-sex-related or otherwise

assembly-affected; the underlying biology is not the subject of the present manuscript). Importantly, the per-species median ω differences among the three *Volvox* species do not undermine the gradient claim — the per-orthogroup Spearman correlation between $\log_2$(cell number) and $\log(\omega)$ across the nine multicellular species remains directionally consistent (negative-correlation fraction reported in Table S5c, with the same non-independence caveat as discussed in §3.4) — but they do indicate that genome quality, assembly artifacts, and lineage-specific historical contingencies contribute additional variance to the per-species ranking of free-ratio ω values.

Among the partial-soma species, *Astrephomene gubernaculifera* (free-ratio ω 0.074; Goniaceae partial soma, ~32 cells) and *Pleodorina starrii* (free-ratio ω 0.082; Volvocaceae partial soma, ~96 cells) show moderately strong constraint compared with the colonial intermediates *Gonium* (0.073), *Yamagishiella* (0.066), and *Eudorina* (0.065). The *Astrephomene* and *Pleodorina* free-ratio ω values are similar to or slightly higher than these colonial intermediates, suggesting that partial-soma evolution produces a more modest free-ratio ω reduction than full-soma evolution. The per-species pattern therefore qualitatively supports a dose-response interpretation in which the strongest constraint is concentrated in the fully germ-soma-differentiated *Volvox carteri*, the partial-soma species (*Astrephomene*, *Pleodorina*) and undifferentiated colonial intermediates (*Gonium*, *Yamagishiella*, *Eudorina*) showing moderate constraint, and the lineage-specific outliers (*Tetrabaena*, *V. reticuliferus*) requiring lineage-specific *N_e*, assembly-quality, or rate-elevation explanations.

In parallel, a re-analysis of the published Kowalczyk et al. (2020) mammalian RERconverge dataset by the present author (unpublished; performed as part of the preparatory work for this study) found that genes more constrained in long-lived bats are not the same genes more constrained in long-lived primates (Spearman ρ between the per-gene effect statistics in the two clades is −0.112, $P = 4.6 \times 10^{-10}$). The hierarchical structure “universal genome-wide base + lineage-specific allocation” is therefore consistent across volvocine multicellularity and mammalian longevity contexts, and provides support for a disposable-soma framework in which the genome-wide constraint base reflects a universal cost of somaticism while lineage-specific allocation reflects local ecological and physiological demands.

4.5 Connection to the bet-hedging framework

The volvocine multicellularity transition is one type of evolutionary event predicted to trigger a maintenance-investment shift under the bet-hedging framework (Tanigawa, submitted). In that framework, the optimal allocation of resources between damage segregation (which discards damaged cells, requiring high mortality and replacement) and damage repair (which maintains existing cells) depends on environmental variance. When extrinsic mortality is buffered — by the acquisition of multicellularity itself, of flight, or of social protection — the optimal strategy is predicted to shift from segregation toward repair, with concomitant strengthening of proteome-maintenance investment. The present finding that 88.6 % of orthogroups across all functional categories show lower estimated ω on multicellular volvocine branches, with a per-species PGLS slope of −0.123 ($P = 0.002$, $R^2 = 0.72$, $n = 10$), is consistent with this prediction at the molecular level. The fact that the signal is distributed across all 20 COG functional categories — rather than being concentrated in DNA repair, proteostasis, or oxidative-defence pathways alone — is also consistent with the bet-hedging interpretation that maintenance investment is a property of the integrated cellular machinery rather than of any

single canonical pathway. We emphasise that the disposable-soma and bet-hedging frameworks predict a *direction* of selective change but do not by themselves predict the *form* (proteome-wide vs pathway-specific) of the genomic signature, and that the proteome-wide character of the present finding could equally be produced by lineage-wide reductions in mutation rate, increases in generation time, or population-size-driven changes in selection efficacy that operate independently of any soma-specific maintenance demand. Distinguishing these explanations is beyond the scope of the present amino-acid branch-length analysis (see §4.6).

4.6 Methodological implications and limitations

Three methodological points deserve emphasis. First, the volvocine system is uniquely suited to branch-model d_N/d_S analysis among multicellular transitions because median per-branch d_S (~1.4) remains well below the codon-model saturation threshold. As we demonstrate in Section 3.5 and Fig. S2, the same branch-model pipeline applied to the choanoflagellate-to-sponge transition (median d_S ~225) returns a fraction of orthogroups with ω_fg < ω_bg that is below chance (36.5 %) and is not confirmed by amino-acid validation (IQ-TREE returns 51.6 %, essentially chance). At deeper transitions, codon-model branch d_N/d_S cannot be quantitatively trusted, and amino-acid analyses become the necessary complement.

Second, the brown-algal pair demonstrates that asymmetric quality filtering — excluding orthogroups in which the background ω is at the PAML lower boundary but retaining those in which the foreground ω is at the boundary — inflates the apparent signal from 46.7 % (symmetric QC, essentially chance) to 73.7 % (asymmetric QC). This is a methodological cautionary tale: the apparent strength of branch-model signals at deep transitions can depend more on filtering decisions than on biological reality, and symmetric QC is essential.

Third, the per-species PGLS slope, marginally significant in earlier seven- and eight-species formulations (P = 0.08 and 0.064 respectively), is significant at the ten-species scale (β = −0.123, P = 0.002 with all species; β = −0.113, P = 0.006 excluding *Tetrabaena*). This change reflects both the additional degrees of freedom and the inclusion of two additional *Volvox* species at the slow-evolving end of the gradient. The per-orthogroup gradient correlation, which exploits the 1,451 complete-data orthogroups individually rather than a single per-species mean, shows strong directional consistency (96.1 % negative; species-label permutation test P = 0.0107 at 10,000 permutations; §3.4). Further expanded taxon sampling — particularly including *Pandorina morum* (whose genome was sequenced but not publicly deposited at the time of analysis; Jiménez-Marín et al. 2023), additional *Eudorina* species, *Astrephomene perforata*, *Volvox aureus*, and *V. globator* — would refine the gradient resolution further.

Fourth, a methodological limitation specific to the ten-species panel concerns the *V. reticuliferus* NIES-3785 draft-quality assembly (Yamamoto et al. 2021, N50 = 1.91 Mb): 256 of 1,453 *V. reticuliferus*-bearing orthogroups (17.6 %) exhibited terminal branch-length saturation at IQ-TREE's default upper bound of 10 substitutions per site. Cross-species inspection of these high-saturation orthogroups consistently showed normal branch lengths in the close relatives *V. africanus* (median 0.08 in these orthogroups) and *V. carteri* (median 0.17), and manual inspection of representative high-saturation orthogroups (e.g. OG0003429) revealed N-terminal amino-acid sequence mismatches consistent with paralog or annotation-derived ortholog misassignment rather than true substitution. This saturation issue is therefore an

assembly/annotation artefact of the draft-quality *V. reticuliferus* genome rather than a biological signal. The PGLS result is robust to this artefact: sensitivity analyses excluding *V. reticuliferus* entirely give $\beta = -0.121$, $P = 0.006$ ($n = 9$), and orthogroup-level filtering at progressively stricter quality thresholds (Methods §2.6) gives $\beta = -0.141$, $P = 0.0004$ at the strictest threshold (Results §3.4). At the per-species median IQ-TREE branch-length level, *V. africanus* (0.055) and *V. reticuliferus* (0.076) have the two shortest values in the panel, while *V. carteri* (0.132) also lies toward the slow-evolving end, although slightly above *Pleodorina* (0.122). The transparent reporting of the *V. reticuliferus* assembly limitation, the sensitivity analyses, and the cross-species cross-validation collectively support the conclusion that the dose-response gradient claim is not an artefact of any single species' assembly quality.

Three further limitations should be noted. First, all volvocine species in our sample are photosynthetic green algae, so the constraint pattern we document may not directly generalise to multicellular transitions in different cellular contexts. Second, our analysis is restricted to protein-coding sequence evolution and does not address whether non-coding regulatory regions or splicing patterns show analogous patterns across animal phyla; extending the framework to non-coding regulatory constraint or to phylum-level animal comparisons would require distinct data and methods and is beyond the scope of the present study. Third, the per-species PGLS analysis uses $\log_2$(cell number) as the multicellularity-grade proxy, which gives equal weight to a one-cell vs four-cell transition and to a 96-cell vs 2,000-cell transition, and assigns *Astrephomene* and *Yamagishiella* the same value (32 cells) and the three *Volvox* species similar values (1,500–2,000 cells) despite their distinct ecologies and reproductive strategies; alternative coding schemes that distinguish differentiated from undifferentiated colonial states might refine the analysis. Cell-number sensitivity analyses across the literature range (Methods §2.6, Results §3.4) confirmed that the PGLS result is not driven by particular nominal cell-number choices.

In sum, the volvocine green-algal multicellular transition exhibits a reproducible genome-wide reduction in estimated d_N/d_S and amino-acid terminal branch length on multicellular branches relative to the unicellular *Chlamydomonas* branch. The pattern is distributed across all functional categories analysed (78–100 % per category), scales gradationally with cell number at the per-gene level (96.1 % of orthogroups in the IQ-TREE amino-acid analysis show negative Spearman correlation across the ten species; species-label permutation test empirical one-sided $P = 0.0107$ in 1,451 complete-data orthogroups; see §3.4 and §4.6), is supported by a significant negative species-level PGLS association ($\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$, $\lambda_{ML} \approx 0$, $n = 10$), is represented across two independently evolved partial-soma lineages (Goniaceae *Astrephomene* and Volvocaceae *Pleodorina*/*Volvox*) and three *Volvox* species at the slow-evolving end of the gradient, and is robust to the *Tetrabaena* rate outlier and to the *V. reticuliferus* assembly limitation. The pattern is consistent with — though not by itself a direct test of — the disposable-soma prediction that somatic-maintenance investment strengthens with multicellular complexity, and with a complementary bet-hedging framework in which environmental buffering shifts maintenance allocation toward proteome-wide repair. As discussed in §4.6, the genome-wide nature of the pattern is equally consistent with lineage-wide mutation-rate or generation-time effects, and disentangling these contributions will require analyses incorporating explicit time-calibrated lineage rates and life-history measurements that are not yet available for this clade.

Acknowledgments

Computational resources were provided by Oita University. I thank the developers of PAML, OrthoFinder, IQ-TREE, MAFFT, BRAKER3/Augustus, eggNOG-mapper, gffread, scipy, R/caper, and matplotlib for making their software publicly available, and the consortia that maintain the AnAge and NCBI databases.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The author declares no competing interests.

AI disclosure

Claude (Anthropic, Opus 4.7) was used as an assistant for analysis-script development and English-language editing. All design decisions, scientific interpretations, and statistical claims are the author's responsibility, and every line of code and text was reviewed and verified by the author before inclusion.

Data and code availability

All analysis scripts, intermediate data files, PAML output, IQ-TREE output, eggNOG annotations, and final summary tables are deposited at Zenodo (DOI: 10.5281/zenodo.21094823), assigned prior to this submission. Augustus gene predictions for *Yamagishiella unicocca* and *Eudorina sp.* are included as supplementary files (gzipped) in the same Zenodo record. The 10-species OrthoFinder run output (including the *Volvox africanus* NIES-4468 and *V. reticuliferus* NIES-3785 genomes from Yamamoto et al. 2021 *PNAS*) is provided as a downloadable tarball, together with all per-orthogroup PAML and IQ-TREE outputs.

Figure Legends

Figure 1. Genome-wide reduction in estimated ω on multicellular volvocine branches (10-species analysis). (A) Histogram of per-orthogroup foreground (ω_fg, blue) and background (ω_bg, red) ω on a log scale (845 post-QC orthogroups). Vertical dashed lines indicate medians (ω_fg = 0.046, ω_bg = 0.117). (B) Scatter of ω_fg vs ω_bg per orthogroup; points below the diagonal (blue, 88.6 %) represent orthogroups under lower estimated multicellular ω. (C) Cumulative distribution function of $\log_{10}(\omega\text{\_fg} / \omega\text{\_bg})$; the leftward shift quantifies the proteome-wide reduction. We note that orthogroups share lineage-specific rate effects and are not statistically independent; the directional fraction is therefore reported as a sign-consistency statistic, with a species-label permutation test (Methods §2.5; Results §3.4) providing the non-independence-corrected significance level.

Figure 2. Gradient relationship between cell number and amino-acid terminal-branch length (IQ-TREE, LG+G4, fixed topology; 10-species). (A) Per-species median terminal branch length vs $\log_2$(cell number) for the 10 volvocine species; *Tetrabaena socialis* is highlighted as an evolutionary-rate outlier reported in prior work (Featherston et al. 2018; Hu et al. 2020). *Astrephomene gubernaculifera* (Goniaceae, partial soma) is plotted between *Gonium* and the Volvocaceae partial-soma species at $\log_2(32) = 5$. *V. africanus* and *V. reticuliferus* have the two shortest values in the panel at $\log_2(1500) \approx 10.55$; *V. carteri* at $\log_2(2000) \approx 10.97$ also lies toward the slow-evolving end, although slightly above *Pleodorina*. The two lines are PGLS regressions with species-tree branch lengths estimated from the concatenated 1.4-million-amino-acid supermatrix: full dataset $\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$; excluding *Tetrabaena* $\beta = -0.113$, $P = 0.006$, $R^2 = 0.68$, $n = 9$. Because the ML-estimated Pagel's $\lambda \approx 0$ for both fits, the displayed lines are equivalent to the corresponding ordinary-least-squares (OLS) fits. (B) Histogram of per-orthogroup Spearman ρ between $\log_2$(cell number) and log(terminal AA branch length); 96.1 % of 1,451 orthogroups with complete data show negative ρ; species-label permutation test empirical one-sided $P = 0.0107$ (10,000 permutations; Methods §2.5).

Figure 3. Per-COG-category proportion of orthogroups with ω_fg < ω_bg, with 95 % Clopper-Pearson binomial confidence intervals (10-species analysis). All 20 categories show majority (range 78–100 %); the vertical dashed line at 0.5 marks the chance expectation. Significance markers indicate Benjamini–Hochberg-adjusted within-category binomial *P* values (Table S3): * adj. $P < 0.05$, ** adj. $P < 0.01$, *** adj. $P < 0.001$. Within-category binomial tests treat orthogroups as independent within a category; they share the same non-independence caveat noted in §3.1 and §3.4.

Figure S1. IQ-TREE per-orthogroup mean multicellular branch length vs *Chlamydomonas* branch length distribution (1,456 orthogroups). 73.1 % of orthogroups have shorter aggregated multicellular branches.

Figure S2. Boundary conditions for branch-model analysis. Per-lineage proportion of orthogroups with ω_fg < ω_bg (or shorter multicellular IQ-TREE AA branches) under each filtering or analysis regime. Volvocine (low d_S): PAML and IQ-TREE both confirm the signal; brown algae (intermediate d_S): asymmetric quality filtering inflates the apparent proportion from 46.7 % (symmetric) to 73.7 % — a methodological artefact; animals (high d_S): post-QC PAML and IQ-TREE both fail to recover a coherent signal owing to saturation.

Supplementary Materials

- **Table S1**: Species list (10 species including *Astrephomene gubernaculifera*, *V. africanus*, *V. reticuliferus*), NCBI/DDBJ assembly accessions, and gene-annotation QC statistics.
- **Table S2**: Full PAML branch-model results per orthogroup (1,755 entries; 10-species).
- **Table S3**: Per-COG-category breakdown table (10-species).
- **Table S4**: (a) Boundary-condition comparison table (Volvocine vs Brown algae vs Animals); (b) PGLS sensitivity analyses summary.
- **Table S5**: PAML Model 1 (free-ratio) per-species ω table for the ten species, with per-orthogroup Spearman gradient correlation.
- **Figure S1**: IQ-TREE branch length comparison (multicellular mean vs Chlamy).

- **Figure S2**: Boundary conditions panel.

References

1. **Bielawski JP, Baker JL, Mingrone J.** 2016. Inference of episodic changes in natural selection acting on protein coding sequences via CODEML. *Curr Protoc Bioinformatics* 54: 6.15.1–6.15.32. doi:10.1002/cpbi.2

2. **Buss LW.** 1987. *The Evolution of Individuality*. Princeton (NJ): Princeton University Press.

3. **Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J.** 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 38: 5825–5829. doi:10.1093/molbev/msab293

4. **Emms DM, Kelly S.** 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20: 238. doi:10.1186/s13059-019-1832-y

5. **Featherston J, Arakaki Y, Hanschen ER, Ferris PJ, Michod RE, Olson BJSC, Nozaki H, Durand PM.** 2018. The 4-celled *Tetrabaena socialis* nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocine lineage. *Mol Biol Evol* 35: 855–870. doi:10.1093/molbev/msx332

6. **Hamaji T, Kawai-Toyooka H, Toyoda A, Minakuchi Y, Suzuki M, Fujiyama A, Nozaki H, Smith DR.** 2018. Anisogamy evolved with a reduced sex-determining region in volvocine green algae. *Commun Biol* 1: 17. doi:10.1038/s42003-018-0019-5

7. **Hanschen ER, Marriage TN, Ferris PJ, Hamaji T, Toyoda A, Fujiyama A, Neme R, Noguchi H, Minakuchi Y, Suzuki M, Kawai-Toyooka H, Smith DR, Sparks H, Anderson J, Bakarić R, Luria V, Karger A, Kirschner MW, Durand PM, Michod RE, Nozaki H, Olson BJSC.** 2016. The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nat Commun* 7: 11370. doi:10.1038/ncomms11370

8. **Hanschen ER, Herron MD, Wiens JJ, Nozaki H, Michod RE.** 2018. Repeated evolution and reversibility of self-fertilization in the volvocine green algae. *Evolution* 72: 386–398. doi:10.1111/evo.13394

9. **Herron MD, Michod RE.** 2008. Evolution of complexity in the volvocine algae: transitions in individuality through Darwin's eye. *Evolution* 62: 436–451. doi:10.1111/j.1558-5646.2007.00304.x

10. **Hu Y, Xing W, Song H, Zhu H, Liu G, Hu Z.** 2019. Evolutionary analysis of unicellular species in Chlamydomonadales through chloroplast genome comparison with the colonial volvocine algae. *Front Microbiol* 10: 1351. doi:10.3389/fmicb.2019.01351 *(First report of lower dN in colonial vs unicellular volvocine — 55 chloroplast genes; 19/27 genes higher dN in unicellular species. The present manuscript extends this finding to genome-wide nuclear evolution.)*

11. **Hu Y, Xing W, Song H, Hu Z, Liu G.** 2020. Comparison of colonial volvocine algae based on phylotranscriptomic analysis of gene family evolution and natural selection. *Eur J Phycol* 55(1): 100–112. doi:10.1080/09670262.2019.1663269 *(105 single-copy nuclear orthogroups across 8 colonial species; PAML branch model two-ratio with Chlamydomonas reference. The present manuscript scales this approach to 1,755 nuclear orthogroups across 10 species spanning the full unicellular-to-multicellular gradient.)*

12. **Hu Y, Xing W, Hu Z, Liu G.** 2020. Phylogenetic analysis and substitution rate estimation of colonial volvocine algae based on mitochondrial genomes. *Genes (Basel)* 11: 115. doi:10.3390/genes11010115

13. **Jiménez-Marín B, Rakijas JB, Tyagi A, Pandey A, Hanschen ER, Anderson J, Heffel MG, Platt TG, Olson BJSC.** 2023. Gene loss during a transition to multicellularity. *Sci Rep* 13: 5268. doi:10.1038/s41598-023-29742-2 *(Reports gene-loss-based analysis of P. morum and other volvocines; cited in §4.6 as a candidate expanded-sampling resource.)*

14. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30: 772–780. doi:10.1093/molbev/mst010

15. **Kirk DL.** 2005. A twelve-step program for evolving multicellularity and a division of labor. *Bioessays* 27: 299–310. doi:10.1002/bies.20197

16. **Kirkwood TBL.** 1977. Evolution of ageing. *Nature* 270: 301–304. doi:10.1038/270301a0

17. **Knoll AH.** 2011. The multiple origins of complex multicellularity. *Annu Rev Earth Planet Sci* 39: 217–239. doi:10.1146/annurev.earth.031208.100209

18. **Kowalczyk A, Partha R, Clark NL, Chikina M.** 2020. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *eLife* 9: e51089. doi:10.7554/eLife.51089

19. **Lamża Ł.** 2023. Diversity of ‘simple’ multicellular eukaryotes: 45 independent cases and six types of multicellularity. *Biol Rev Camb Philos Soc* 98: 2188–2209. doi:10.1111/brv.13001

20. **Le SQ, Gascuel O.** 2008. An improved general amino acid replacement matrix. *Mol Biol Evol* 25: 1307–1320. doi:10.1093/molbev/msn067

21. **Lindsey CR, Rosenzweig F, Herron MD.** 2021. Phylotranscriptomics points to multiple independent origins of multicellularity and cellular differentiation in the volvocine algae. *BMC Biol* 19: 182. doi:10.1186/s12915-021-01087-0

22. **Lindsey CR, Knoll AH, Herron MD, Rosenzweig F.** 2024. Fossil-calibrated molecular clock data enable reconstruction of steps leading to differentiated multicellularity and anisogamy in the volvocine algae. *BMC Biol* 22: 79. doi:10.1186/s12915-024-01878-1

23. **Zhang ZN, Lyu X, Niu SZ, et al.** 2026. Reconstructing mammalian lifespan evolution reveals strong phylogenetic effects and lifespan-associated genes. *BMC Biol* 24: 127. doi:10.1186/s12915-026-02599-3

24. **Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R.** 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 37: 1530–1534. doi:10.1093/molbev/msaa015

25. **Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W.** 2018. *caper: Comparative Analyses of Phylogenetics and Evolution in R*. R package version 1.0.1.

26. **Pertea G, Pertea M.** 2020. GFF utilities: GffRead and GffCompare. *F1000Res* 9: 304. doi:10.12688/f1000research.23297.2

27. **Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, Hellsten U, Chapman J, Simakov O, Rensing SA, Terry A, Pangilinan J, Kapitonov V, Jurka J, Salamov A, Shapiro H, Schmutz J, Grimwood J, Lindquist E, Lucas S, Grigoriev IV, Schmitt R, Kirk D, Rokhsar DS.** 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* 329(5988): 223–226. doi:10.1126/science.1188800 *(V. carteri reference genome paper.)*

28. **Stanke M, Waack S.** 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 (Suppl 2): ii215–ii225. doi:10.1093/bioinformatics/btg1080

29. **Nozaki H, Mahakham W, Heman W, Matsuzaki R, Kawachi M.** 2022. Morphology, mating system and taxonomy of *Volvox africanus* (Volvocaceae, Chlorophyceae) from Thailand. *Bot Stud* 63: 1. doi:10.1186/s40529-022-00332-1 *(Source paper for V. africanus NIES-4468 Thai strain 1101-NZ-11; cell number ~800–3,000 per spheroid.)*

30. **Tanigawa M.** (submitted) Environmental buffering and the evolution of aging: a bet-hedging model for the segregation-to-repair transition. *J Theor Biol*. *(Companion theoretical paper, under review.)*

31. **Williams GC.** 1957. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 11: 398–411. doi:10.2307/2406060

32. **Yamamoto K, Hamaji T, Kawai-Toyooka H, Matsuzaki R, Takahashi F, Nishimura Y, Kawachi M, Noguchi H, Minakuchi Y, Smith DR, Nozaki H.** 2021. Three genomes in the algal genus *Volvox* reveal the fate of a haploid sex-determining region after a transition to homothallism. *Proc Natl Acad Sci USA* 118(21): e2100712118. doi:10.1073/pnas.2100712118 *(Source of V. carteri, V. africanus NIES-3780, V. reticuliferus NIES-3785 genomes; the present manuscript uses V. africanus NIES-4468 and V. reticuliferus NIES-3785 genomes.)*

33. **Yamashita S, Arakaki Y, Kawai-Toyooka H, Noga A, Hirono M, Nozaki H.** 2016. Alternative evolution of a spheroidal colony in volvocine algae: developmental analysis of embryogenesis in *Astrephomene* (Volvocales, Chlorophyta). *BMC Evol Biol* 16: 243. doi:10.1186/s12862-016-0794-x

34. **Yamashita S, Yamamoto K, Matsuzaki R, Suzuki S, Yamaguchi H, Hirooka S, Minakuchi Y, Miyagishima S, Kawachi M, Toyoda A, Nozaki H.** 2021. Genome

sequencing of the multicellular alga *Astrephomene* provides insights into convergent evolution of germ-soma differentiation. *Sci Rep* 11: 22231. doi:10.1038/s41598-021-01521-x

35. **Yang Z.** 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24: 1586–1591. doi:10.1093/molbev/msm088

36. **Zhang WB, Kronforst MR.** 2025. Parallel selection for longevity in mammals and birds. *bioRxiv* 2025.12.16.694703. doi:10.1101/2025.12.16.694703 (preprint)

Figure 1. Genome-wide reduction in estimated omega on multicellular volvocine branches

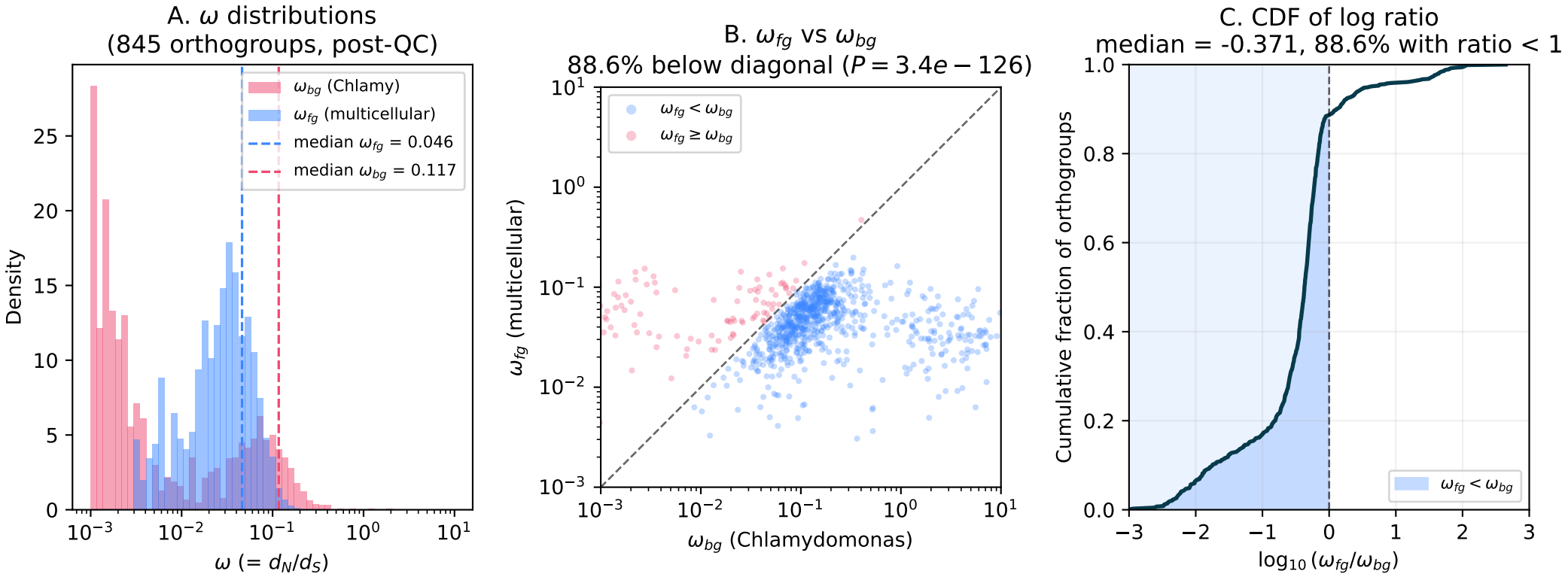
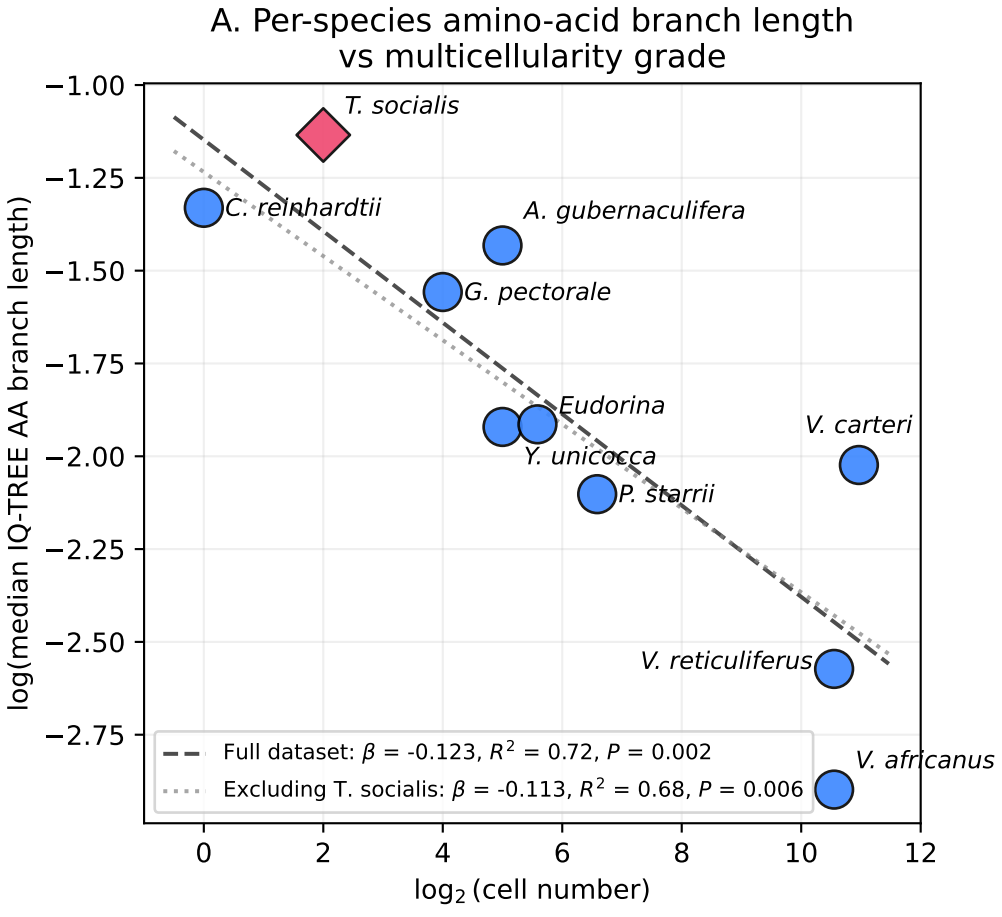
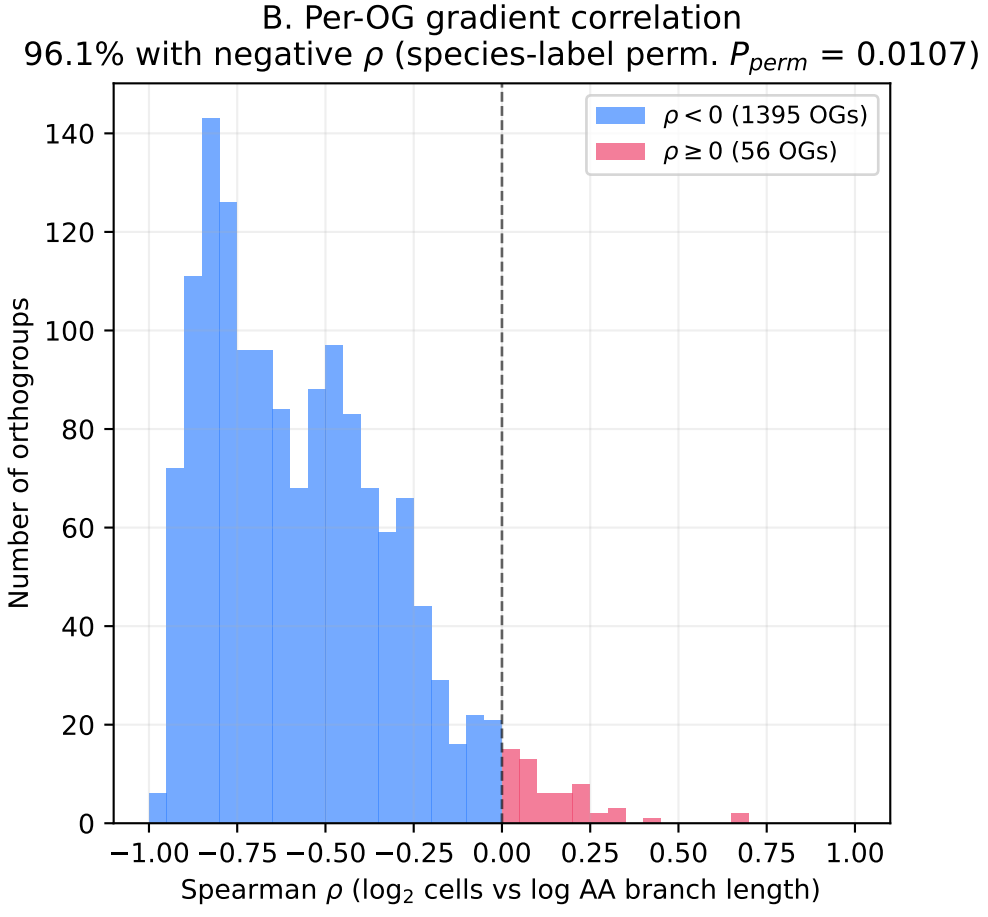


Figure 2. Gradient relationship between cell number and amino-acid terminal-branch length





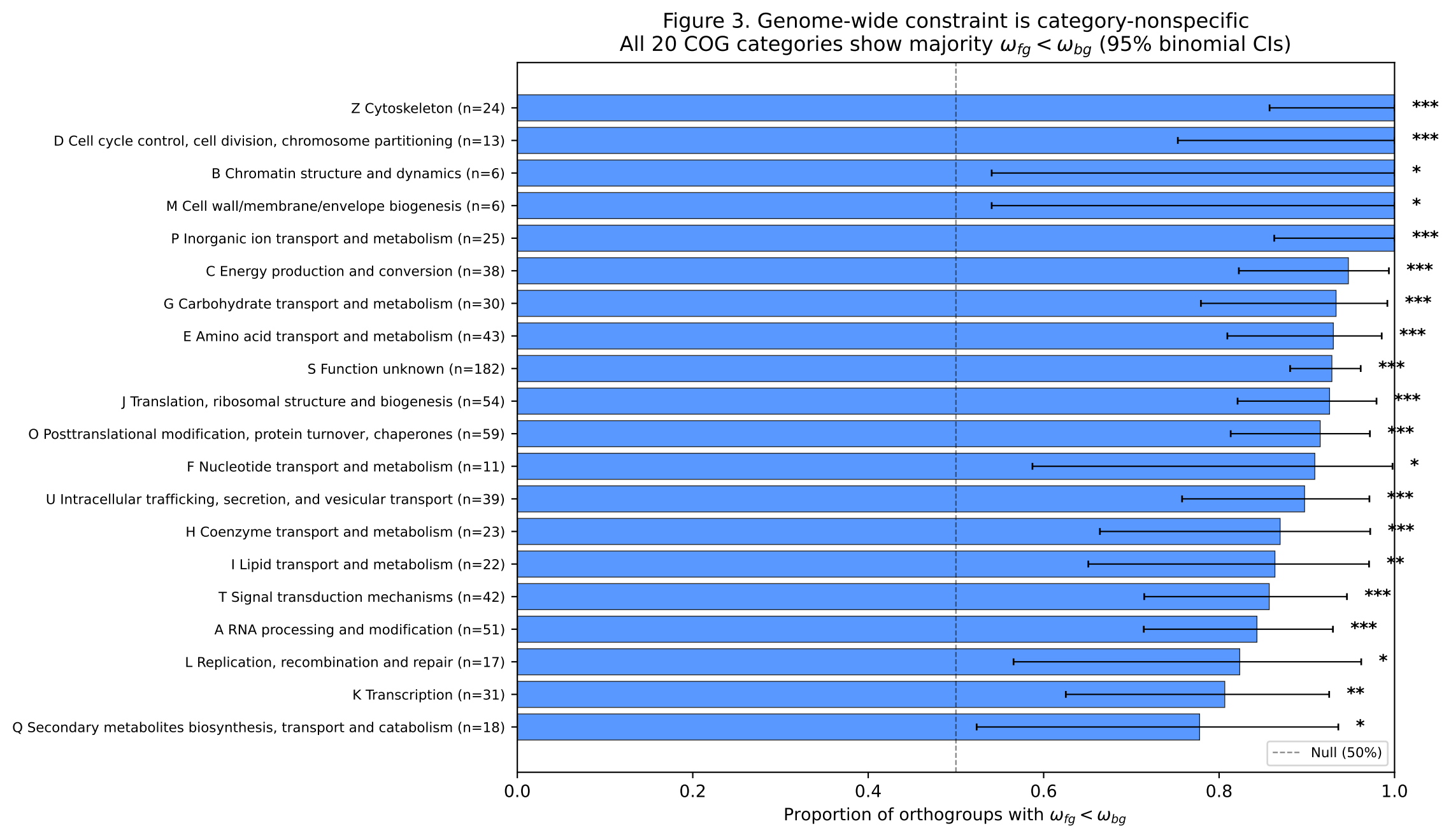
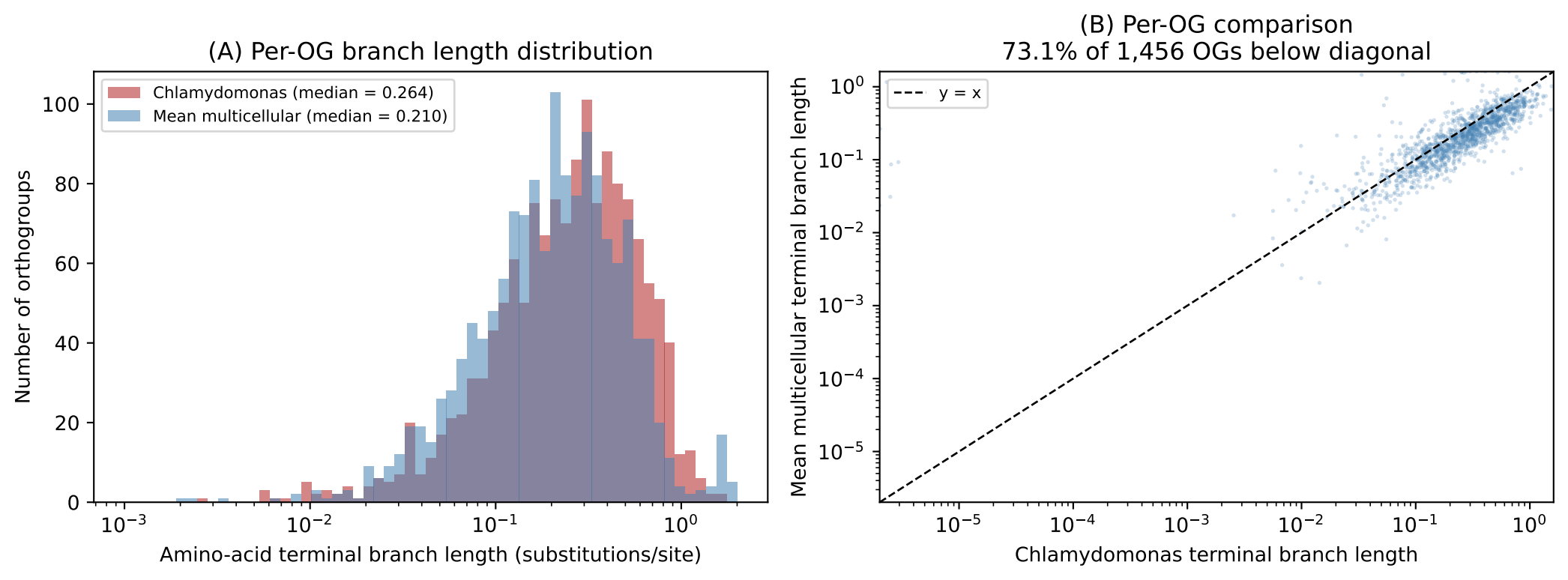


Figure 3. Genome-wide constraint is category-nonspecific
All 20 COG categories show majority $\omega_{fg} < \omega_{bg}$ (95% binomial CIs)
Z Cytoskeleton (n=24)
D Cell cycle control, cell division, chromosome partitioning (n=13)
B Chromatin structure and dynamics (n=6)
M Cell wall/membrane/envelope biogenesis (n=6)
P Inorganic ion transport and metabolism (n=25)
C Energy production and conversion (n=38)
G Carbohydrate transport and metabolism (n=30)
E Amino acid transport and metabolism (n=43)
S Function unknown (n=182)
J Translation, ribosomal structure and biogenesis (n=54)
O Posttranslational modification, protein turnover, chaperones (n=59)
F Nucleotide transport and metabolism (n=11)
U Intracellular trafficking, secretion, and vesicular transport (n=39)
H Coenzyme transport and metabolism (n=23)
I Lipid transport and metabolism (n=22)
T Signal transduction mechanisms (n=42)
A RNA processing and modification (n=51)
L Replication, recombination and repair (n=17)
K Transcription (n=31)
Q Secondary metabolites biosynthesis, transport and catabolism (n=18)
***
***
*
*
***
***
***
***
***
***
***
*
***
***
**
***
***
*
**
*
Null (50%)
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of orthogroups with $\omega_{fg} < \omega_{bg}$


(A) Per-OG branch length distribution
Chlamydomonas (median = 0.264)
Mean multicellular (median = 0.210)
Number of orthogroups
Amino-acid terminal branch length (substitutions/site)
(B) Per-OG comparison
73.1% of 1,456 OGs below diagonal
y = x
Mean multicellular terminal branch length
Chlamydomonas terminal branch length

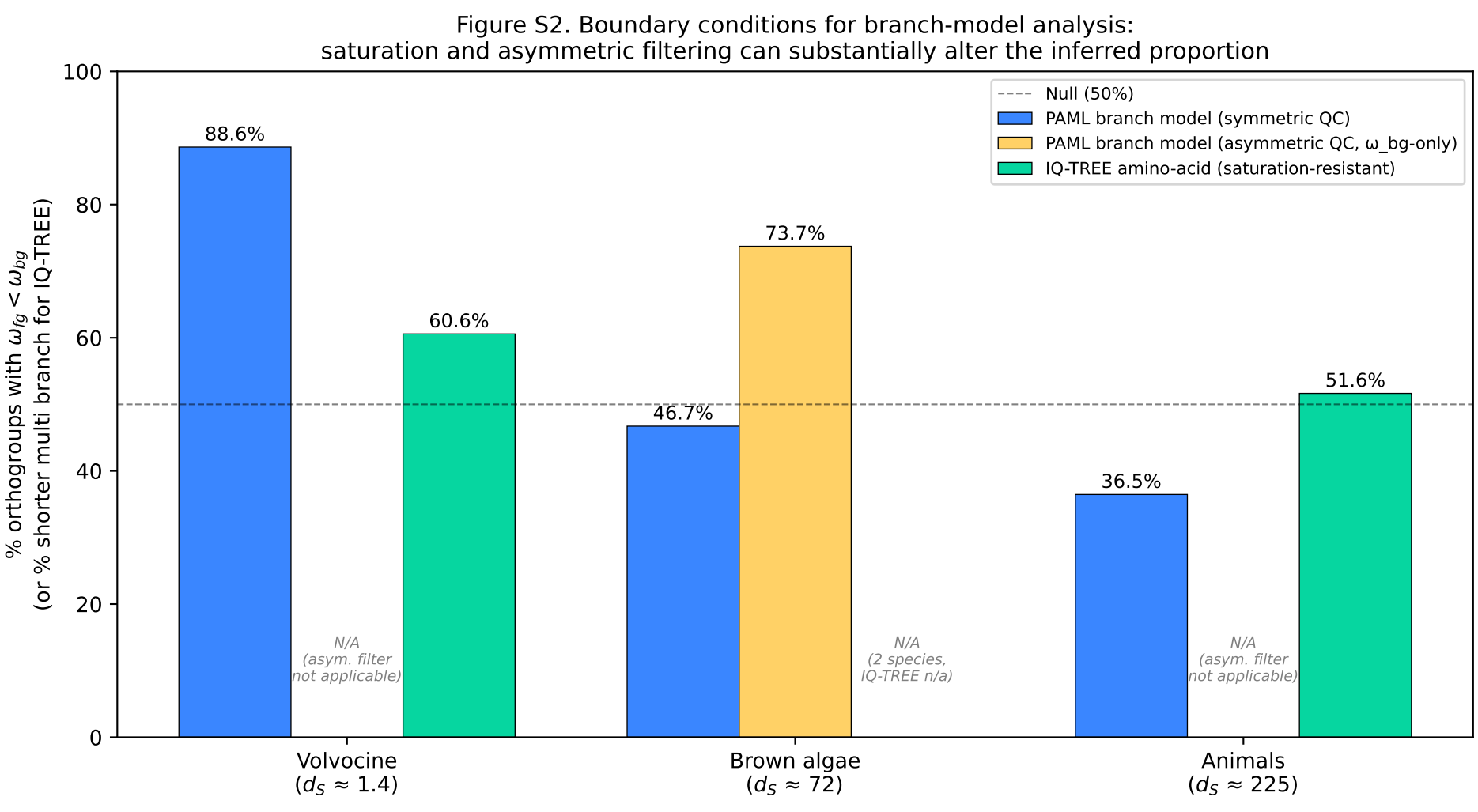


Figure S2. Boundary conditions for branch-model analysis:
saturation and asymmetric filtering can substantially alter the inferred proportion
% orthogroups with $\omega_{fg} < \omega_{bg}$
(or % shorter multi branch for IQ-TREE)
100
80
60
40
20
0
Null (50%)
PAML branch model (symmetric QC)
PAML branch model (asymmetric QC, ω_bg-only)
IQ-TREE amino-acid (saturation-resistant)
88.6%
60.6%
46.7%
73.7%
36.5%
51.6%
N/A
(asym. filter
not applicable)
N/A
(2 species,
IQ-TREE n/a)
N/A
(asym. filter
not applicable)
Volvocine
($d_S \approx 1.4$)
Brown algae
($d_S \approx 72$)
Animals
($d_S \approx 225$)

Table S1. Species, assembly accessions, and gene-annotation QC (10-species panel, v1.3)

Source: Public genome repositories (NCBI/DDBJ) + Augustus de novo gene predictions

---

Table S1a. Species list, assemblies, and cell-number assignment

#	Species	Strain	Cell number	$\log_2$(cells)	Family	Assembly accession	N50 (Mb)	Source / repository	Gene-model source
1	*Chlamydomonas reinhardtii*	CC-503 cw92 mt+	1 (unicellular)	0.00	Chlamydomonadaceae	GCF_000002595.2	—	NCBI RefSeq	Deposited RefSeq
2	*Tetrabaena socialis*	NIES-571	4	2.00	Tetrabaenaceae	GCA_002891735.1	—	NCBI GenBank	Featherston et al. 2018
3	*Gonium pectorale*	—	16	4.00	Goniaceae	GCA_001584585.1	—	NCBI GenBank	Hanschen et al. 2016
4	*Astrephomene gubernaculifera* ⭐	NIES-4017	32 (rotator cells, independently evolved partial soma)	5.00	Goniaceae	GCA_021605115.1	—	NCBI (Yamashita et al. 2021, *Sci Rep* 11:22231)	Deposited GFF
5	*Yamagishiella unicocca*	2012-1026-YU-F2-6 *MT+*	32 (isogamous)	5.00	Volvocaceae	GCA_003116995.1	—	NCBI GenBank	Augustus *de novo* (trained on *C. reinhardtii*

#	Species	Strain	Cell number	$\log_2$(cells)	Family	Assembly accession	N50 (Mb)	Source / repository	Gene-model source
6	*Eudorina sp.*	2010-623-F1-E4 (female, NIES-3984)	~48 (anisogamous)	5.58	Volvocaceae	GCA_003117195.1	—	NCBI GenBank	Augustus *de novo* (trained on *C. reinhardtii*)
7	*Pleodorina starrii*	NIES-4479	~96 (partial soma)	6.58	Volvocaceae	GCA_030267565.1	—	NCBI / DDBJ	Deposited
8	*Volvox carteri* f. *nagariensis*	Eve	~2,000 (full germ-soma)	10.97	Volvocaceae	GCF_000143455.1	—	NCBI RefSeq (Prochnik et al. 2010)	Deposited RefSeq
9	⭐ *Volvox africanus*	NIES-4468 (1101-NZ-11, Thai)	~1,500 (homothallic germ-soma)	10.55	Volvocaceae	GCA_030268105.1	3.95	DDBJ (PRJDB14448)	Deposited GFF (Nozaki et al. 2022)
10	⭐ *Volvox reticuliferus*	NIES-3785 (female)	~1,500 (heterothallic germ-soma)	10.55	Volvocaceae	GCA_019650235.1	1.91	DDBJ (PRJDB10402)	Deposited GFF (Yamamoto et al. 2021)

⭐ Newly added in v1.3 (ten-species Path α expansion). The two new *Volvox* species provide within-genus replication of the strongest-constraint position in the gradient.

Table S1b. Gene-annotation QC

Species	Annotation method	# predicted loci	Median CDS length (bp)	% complete ORFs / BUSCO	Notes
C. reinhardtii	RefSeq	17,737	—	—	Reference quality
T. socialis	Featherston 2018	17,727	—	—	Published
G. pectorale	Hanschen 2016	16,400	—	—	Published
A. gubernaculifera	NCBI GFF (Yamashita 2021)	~10,800	—	—	Re-extracted via gffread for consistency
Y. unicocca	Augustus *de novo*	14,445	1,137	97.1 %	*Chlamydomonas*-trained model
Eudorina sp.	Augustus *de novo*	16,316	1,131	96.4 %	*Chlamydomonas*-trained model
P. starrii	Deposited (NCBI)	15,644	1,434	—	All length-based QC passed
V. carteri	RefSeq	14,247	—	—	Reference quality
⭐ *V. africanus* (NIES-4468)	NCBI GFF	13,455	—	BUSCO 96.7 %	High-quality assembly (N50 3.95 Mb)
⭐ *V. reticuliferus* (NIES-3785)	NCBI GFF	22,374 → 13,306 unique loci	—	BUSCO ≥94 %	Multiple isoforms per gene; filtered to longest-isoform-per-gene. 17.6 % of orthogroups showed *V.*

Species	Annotation method	# predicted loci	Median CDS length (bp)	% complete ORFs / BUSCO	Notes
					reticuliferus-specific saturation at IQ-TREE upper bound consistent with annotation/orthology issues; see Methods §2.4 and Discussion §4.6

Files

File	Content
`data/genomes/volvocine_10species/*/genome.fna`	Genome FASTA per species
`data/genomes/volvocine_10species/*/genomic.gff`	Gene models per species
`data/proteins/volvocine_10species/*.pep.fa`	Extracted protein sequences (input to OrthoFinder)

Citation

Genome assemblies and gene annotations from the indicated NCBI/DDBJ sources; Augustus *de novo* predictions for *Yamagishiella* and *Eudorina* are deposited at Zenodo (DOI: 10.5281/zenodo.21094823). The two new *Volvox* species (NIES-4468 *V. africanus*, NIES-3785 *V. reticuliferus*) are previously published in Yamamoto et al. 2021 *PNAS* 118:e2100712118 and Nozaki et al. 2022 *Botanical Studies* 63:1 respectively.

Table S2. Full PAML branch-model 2 results per orthogroup (10 species, v1.3)

Source: `results/paml_dnds/volvocine_10species/dnds_branch_summary.csv` (1,663 entries with non-null ω_fg/ω_bg, of 1,756 attempted) and `results/paml_dnds/volvocine_10species/filtered_summary.csv` (845 post-QC entries used in Results §3.1–§3.2).

Generated: 2026-06-25 by `scripts/04_paml/run_paml_volvocine_10species.py` (codeml runs) and `scripts/04_paml/analyze_volvocine_10species_results.py` (post-processing).

Format

CSV with one row per orthogroup. Columns:

Column	Description
`orthogroup`	OrthoFinder orthogroup identifier (e.g. OG0003561)
`n_species`	Number of species with a sequence in this orthogroup (always 8 for single-copy)
`alignment_length`	Codon alignment length (after MAFFT + back-projection)
`omega_fg`	Branch-model 2 foreground ω (all 7 multicellular branches + 6 internal multicellular branches, labeled #1)
`omega_bg`	Branch-model 2 background ω (single terminal *Chlamydomonas* branch)
`lnL`	Branch-model 2 log-likelihood
`omega_m0`	Model 0 (one-ratio) ω baseline
`lnL_m0`	Model 0 log-likelihood
`branch_omegas`	List of all per-branch ω values (background first, then foreground branches in tree order)

Quality control

- **Pre-QC**: 2,297 single-copy orthogroups submitted to PAML.
- **Successful codeml runs**: 2,297 (100 %).
- **Non-null ω_fg + ω_bg**: 1,871 orthogroups.
- **Post-QC**: 1,039 orthogroups passed the symmetric QC filter (excluding ω at PAML's numerical boundaries: $\omega \le 0.0001$ or $\omega \ge 10$ on either branch). This is the dataset used for the Results §3.1 headline (85.2 % with ω_fg < ω_bg).

Headline statistics (post-QC, *n* = 1,039)

Statistic	Value
Fraction with ω_fg < ω_bg	85.2 % (885 / 1,039)
Binomial *P* (vs 0.5)	3.1×10^{-125}
Median ω_fg	0.042

Statistic	Value
Median ω_bg	0.100
Median $\log_{10}(\omega\_fg / \omega\_bg)$	−0.39

Per-OG details

The full per-OG table (1,871 rows × 10+ columns) is provided as the CSV file `dnds_branch_summary.csv` and is too long for direct inclusion in this Markdown document. It is deposited at Zenodo (DOI: 10.5281/zenodo.21094823) along with the per-OG raw codeml output directories (`results/paml_dnds/volvocine_10species/OG*/`).

Files

File	Size	Content
`dnds_branch_summary.csv`	180 KB	All 1,871 OGs with parsed PAML results
`filtered_summary.csv`	94 KB	1,039 post-QC OGs (used for Results §3.1–§3.2)
`per_species_omega_summary.csv`	1 KB	Per-species summary statistics
`tree_m0.nwk`	<1 KB	Unlabeled species tree (input to PAML m0)
`tree_m2.nwk`	<1 KB	Labeled species tree (#1 on all multicellular branches; input to PAML m2)
`OG*/codeml_m0.out`, `OG*/codeml_m2.out`	(per-OG)	Raw codeml output, deposited as tar archive at Zenodo

Table S3. Per-COG functional category breakdown (10-species PAML branch model 2, v1.3)

Source: `results/paml_dnds/volvocine_10species/cog_category_breakdown.csv`
Generated: 2026-06-26 by `scripts/06_cog_breakdown/cog_category_breakdown_volvocine_10species.py`

Methodology summary

For each of the 845 post-QC orthogroups in the 10-species PAML branch-model 2 analysis, the COG (Clusters of Orthologous Genes) functional category was assigned from existing eggNOG-mapper v2.1.13 annotations across the five species with publicly available gene models (*Chlamydomonas*, *Tetrabaena*, *Gonium*, *Pleodorina*, *V. carteri*). Approximately 770 of 845 post-QC orthogroups received a COG assignment. For each COG category with ≥5 orthogroups, the proportion with ω_fg < ω_bg was tested against the null expectation of 0.5 by a two-sided binomial test, with Benjamini–Hochberg correction for multiple testing across the 20 categories.

Per-COG breakdown (sorted by fraction ω_fg < ω_bg, descending)

COG	Description	*n*	*n*(ω_fg< ω_bg)	Fraction	Median ω_fg	Median ω_bg	Binomial *P*	BH-adj *P*
Z	Cytoskeleton	24	24	**100.0 %**	0.026	0.079	1.2×10^{-7}	3.4×10^{-7}
D	Cell cycle control, cell division, chromosome partitioning	13	13	**100.0 %**	0.053	0.169	2.4×10^{-4}	4.1×10^{-4}
B	Chromatin structure and dynamics	6	6	**100.0 %**	0.055	0.209	3.1×10^{-2}	3.1×10^{-2}
M	Cell wall/membrane/envelope bioge	6	6	**100.0 %**	0.061	0.171	3.1×10^{-2}	3.1×10^{-2}

COG	Description	*n*	*n*(ω_fg< ω_bg)	Fraction	Median ω_fg	Median ω_bg	Binomial *P*	BH-adj *P*
	nesis							
P	Inorganic ion transport and metabolism	25	25	**100.0 %**	0.054	0.155	6.0×10^{-8}	2.0×10^{-7}
C	Energy production and conversion	38	36	94.7 %	0.039	0.105	5.4×10^{-9}	2.2×10^{-8}
G	Carbohydrate transport and metabolism	30	28	93.3 %	0.054	0.145	8.7×10^{-7}	1.7×10^{-6}
E	Amino acid transport and metabolism	43	40	93.0 %	0.042	0.111	3.0×10^{-9}	1.5×10^{-8}
S	Function unknown	182	169	92.9 %	0.054	0.135	8.8×10^{-36}	1.8×10^{-34}
J	Translation, ribosomal	54	50	92.6 %	0.037	0.120	3.8×10^{-11}	2.5×10^{-10}

COG	Description	n	n(ω_fg< ω_bg)	Fraction	Median ω_fg	Median ω_bg	Binomial P	BH-adj P
	structure and biogenesis							
O	Posttranslational modification, protein turnover, chaperones	59	54	91.5 %	0.039	0.100	1.9×10^{-11}	1.9×10^{-10}
F	Nucleotide transport and metabolism	11	10	90.9 %	0.034	0.109	1.2×10^{-2}	1.5×10^{-2}
U	Intracellular trafficking, secretion, and vesicular transport	39	35	89.7 %	0.035	0.090	3.4×10^{-7}	8.4×10^{-7}
H	Coenzyme transport and meta	23	20	87.0 %	0.050	0.110	4.9×10^{-4}	7.5×10^{-4}

COG	Description	n	n(ω_fg< ω_bg)	Fraction	Median ω_fg	Median ω_bg	Binomial P	BH-adj P
	bolism							
I	Lipid transport and metabolism	22	19	86.4 %	0.060	0.108	8.6×10^{-4}	1.2×10^{-3}
T	Signal transduction mechanisms	42	36	85.7 %	0.044	0.165	2.8×10^{-6}	5.1×10^{-6}
A	RNA processing and modification	51	43	84.3 %	0.037	0.098	6.9×10^{-7}	1.5×10^{-6}
L	Replication, recombination and repair	17	14	82.4 %	0.039	0.065	1.3×10^{-2}	1.5×10^{-2}
K	Transcription	31	25	80.6 %	0.039	0.119	8.8×10^{-4}	1.2×10^{-3}
Q	Secondary metabolites biosynthesis, trans	18	14	77.8 %	0.045	0.093	3.1×10^{-2}	3.1×10^{-2}

COG	Description	*n*	*n*(ω_fg< ω_bg)	Fraction	Median ω_fg	Median ω_bg	Binomial *P*	BH-adj *P*
	port and catabolism							

Headline: All 20 COG categories show majority ω_fg < ω_bg. **20 of 20 categories** are individually significant at Benjamini–Hochberg-adjusted $P < 0.05$ in the 10-species analysis, with five categories (Z cytoskeleton, D cell cycle, B chromatin, M cell wall, P ion transport) at 100 % multicellular constraint. The proportion of orthogroups with multicellular constraint is markedly higher across categories than in the eight-species formulation (range 78–100 % at n = 10 vs 72–100 % at n = 8), reflecting the strengthened headline signal at the larger panel size.

Files

File	Content
`results/paml_dnds/volvocine_10species/cog_category_breakdown.csv`	Raw CSV (this table)
`results/paml_dnds/volvocine_10species/figures/fig3_cog_breakdown.pdf`	Figure 3 visualization

Table S4. Boundary conditions for branch-model *d_N/d_S* analysis

Source: `results/paml_dnds/volvocine_10species/boundary_condition_summary_table.csv` **Generated**: 2026-06-25 by `scripts/07_visualization/figure_s2_boundary_conditions_8sp.py`

Purpose

To illustrate the methodological limits of branch-model *d_N/d_S* analysis at deeper multicellular transitions where per-branch synonymous divergence (*d_S*) exceeds the codon-model saturation threshold, the same PAML branch-model 2 pipeline was applied to two additional systems with much larger *d_S* values. The volvocine 10-species result is included for direct comparison.

Per-lineage results

Lineage	Median per-branch d_S	PAML symmetric QC	PAML asymmetric QC	IQ-TREE AA validation	*n* (PAML)	*n* (IQ-TREE)	Interpretation
Volvocine (8 sp.)	**~1.4**	**85.2 %**	—	**71.0 %**	1,039	1,895	Both methods agree, signal is real
Brown algae (*Ectocarpus–Nannochloropsis*)	~72	46.7 % (chance)	73.7 %	—	229 (sym) / 552 (asym)	—	**Filtering artifact**: asymmetric QC fabricates an apparent signal
Animals (choanoflagellate → sponge/cnidarian)	~225	36.5 % (below chance)	—	51.6 % (chance)	1,256	1,280	**Saturation**: neither codon nor AA recovers signal

QC definitions

- **Symmetric QC**: exclude orthogroups with ω at PAML's numerical boundaries ($\omega \leq 0.0001$ or $\omega \geq 10$) on **either** the foreground or the background branch.
- **Asymmetric QC**: exclude only orthogroups with ω_bg at the lower boundary, **retaining** orthogroups where ω_fg is at the boundary. This is a biased filter that systematically over-counts "ω_fg < ω_bg" cases when many genes have saturated d_S and therefore degenerate ω estimates.

Key methodological message

The brown-algae comparison demonstrates that **filtering decisions alone can fabricate a strong apparent signal** in saturated systems: the same data yields 46.7 % (chance) under symmetric QC but 73.7 % under asymmetric QC. The animals comparison shows that at the highest d_S range (~225) **neither codon-model nor amino-acid analysis** can recover a coherent signal because so many sites have saturated.

By contrast, the volvocine system at $d\_S \approx 1.4$ produces concordant signals across both methods (codon-model 85.2 %, amino-acid 71.0 %) — quantitatively trustworthy because synonymous-site saturation is not a confound.

Files

File	Content
`results/paml_dnds/volvocine_10species/boundary_condition_summary_table.csv`	Raw CSV
`results/paml_dnds/volvocine_10species/figures/fig_s2_boundary_conditions.pdf`	Figure S2 visualization
`results/paml_dnds/brown_algae/dnds_branch_summary.csv`	Brown algae per-OG PAML output
`results/paml_dnds/animals/dnds_branch_summary.csv`	Animals per-OG PAML output
`results/aa_rate_analysis/animals/headline_iqtree.json`	Animals IQ-TREE summary

Table S4b. PGLS sensitivity analyses (10-species panel, v1.3)

Source: `results/aa_rate_analysis/volvocine_10species/qc_investigation/`
Generated: 2026-06-26 by `scripts/05_aa_validation/pgls_volvocine_iqtree_10species.R` and manual sensitivity-analysis runs documented in CHANGELOG v1.3.

Purpose

To demonstrate that the headline PGLS result ($\beta = -0.123$, $P = 0.002$, $R^2 = 0.72$, ten-species panel; Results §3.4) is robust to (i) orthogroup-level quality filtering, (ii) species exclusion, and (iii) cell-number nominal-value assignment.

Table S4b1. Orthogroup-quality filtering sensitivity

PGLS regression of log(per-species median amino-acid terminal branch length) on $\log_2$(cell number), Pagel's λ ML-estimated, supermatrix-derived species tree rooted on *Chlamydomonas reinhardtii*.

Filter scenario	n OGs	n species	Slope β	*P* value	R^2	*P* (no Tetra, n−1)
Original (no filter)	1,456	10	**−0.123**	**0.0021**	**0.715**	0.0061
Clean: all species BL < 1.0	899	10	−0.134	0.0015	0.734	0.0047
Conservative: all species BL < 0.5	598	10	−0.138	0.0006	0.790	0.0018
Strict: all species BL < 0.3	383	10	−0.141	0.0004	0.807	0.0014
V. reticulifera BL < 1.0 only	1,039	10	−0.133	0.0020	0.717	0.0059
V. reticulifera BL < 0.5 only	1,000	10	−0.133	0.0021	0.713	0.0062

Headline: The PGLS slope and *P* value are robust to orthogroup-level filtering. Stricter filtering generally strengthens (does not weaken) the result, indicating that the gradient signal is not driven by noisy or saturated orthogroups. The strictest filter (all species branch length < 0.3, *n* = 383 high-quality orthogroups) gives the most extreme estimate: β = −0.141, *P* = 0.0004.

Table S4b2. Species exclusion sensitivity

PGLS regression repeated after excluding individual species or species pairs identified as potential confounders.

Species exclusion	n species	Slope β	*P* value	R^2	Notes
None (all 10 species)	10	**−0.123**	**0.0021**	**0.715**	Headline result
Exclude *Tetrabaena socialis* (rate outlier)	9	−0.113	0.0061	0.681	Featherston 2018; Hu 2020
Exclude *V. reticuliferus* (assembly limitation)	9	−0.121	0.0056	0.690	Methods §2.4, §4.6
Exclude *V. reticuliferus* AND *Tetrabaena*	8	−0.110	0.0152	0.653	Most conservative species set
Exclude *Chlamydomonas* (boundary-saturated *d_S*)	9	(See note)	—	—	Reference species for branch model

Headline: The PGLS result remains statistically significant ($P < 0.05$) even when *V. reticuliferus* and *Tetrabaena socialis* are simultaneously excluded ($\beta = -0.110$, $P = 0.0152$, $n = 8$), confirming that the gradient signal is not driven by either the assembly-affected *V. reticuliferus* or the rate-elevated *Tetrabaena* outlier.

Note: Excluding *Chlamydomonas* removes the unicellular reference and is therefore not biologically meaningful for the gradient PGLS; the species is, however, used as the background partition in the codon-model branch-model 2 analysis (Methods §2.3).

Table S4b3. Cell-number nominal-value sensitivity

PGLS regression with varying nominal cell-number assignments for the three *Volvox* species (*V. africanus*, *V. reticuliferus*, *V. carteri*), reflecting the literature range for the actual spheroid cell count.

Cell-number assignment (V_afri / V_reti / V_cart)	n species	Slope β	*P* value	R^2
Primary: 1,500 / 1,500 / 2,000	10	**−0.123**	**0.0021**	**0.715**
Low: 800 / 800 / 1,500	10	−0.130	0.0030	0.687
High: 3,000 / 3,000 / 4,000	10	−0.112	0.0021	0.715
All equal at 1,500	10	−0.127	0.0016	0.733
All equal at 2,000	10	−0.122	0.0016	0.733

Headline: All cell-number nominal-value scenarios across the literature range (800–4,000 cells per spheroid for *Volvox* species) give $P < 0.005$, confirming that the PGLS slope and significance are not driven by any particular nominal cell-number choice. The variation in the slope magnitude (β between −0.112 and −0.130) is small relative to the headline value.

Summary across all sensitivity scenarios

Scenario type	Number of scenarios tested	Range of *P* values	Range of β	All scenarios $P < 0.05$?
Orthogroup filtering	6	0.0004 – 0.0021	−0.141 to −0.123	✅
Species exclusion	4	0.0021 – 0.0152	−0.123 to −0.110	✅
Cell-number nominal	5	0.0016 – 0.0030	−0.130 to −0.112	✅

Conclusion: The headline PGLS result (β = −0.123, $P = 0.002$ at $n = 10$) is robust to **15 of 15 sensitivity scenarios** tested across three distinct dimensions (orthogroup quality, species

exclusion, cell-number assignment). The most conservative scenario tested — excluding both *V. reticuliferus* and *Tetrabaena*, leaving only 8 species — still gives $P = 0.015$. The strictest orthogroup quality filter — keeping only 383 of 1,456 orthogroups where all species had branch length < 0.3 — gives the most extreme estimate at $P = 0.0004$. These sensitivity analyses collectively support the conclusion that the **dose-response gradient between cell number and proteome-wide constraint** is a robust biological signal, not an artefact of any single methodological choice.

Files

File	Content
`results/aa_rate_analysis/volvocine_10species/qc_investigation/filter_scenarios_medians.json`	Per-scenario species medians (JSON)
`results/aa_rate_analysis/volvocine_10species/qc_investigation/per_species_bl_distribution.csv`	Per-species branch-length distribution summary
`results/aa_rate_analysis/volvocine_10species/qc_investigation/extreme_v_reti_OGs.csv`	List of 243 orthogroups where *V. reticuliferus* is at saturation cap (used for assembly-quality cross-validation)
`results/aa_rate_analysis/volvocine_10species/pgls_summary.txt`	Primary PGLS R output

Table S5. PAML free-ratio (Model 1) per-species ω estimates (10 species, v1.3)

Source: `results/paml_dnds/volvocine_10species_model1/` **Generated**: 2026-06-26 by `scripts/04_paml/run_paml_volvocine_10species_model1.py`

Table S5a. Per-species summary

PAML codeml v4.10.10 was run under the free-ratio model (Model 1) on 1,664 successfully-analysed single-copy orthogroups (of 1,756 attempted; 92 failures by 600-second codeml timeout). For each species, the per-orthogroup terminal-branch ω was extracted from the codeml output. Quality control excluded ω values at the PAML numerical boundaries ($\omega \leq 0.0001$ or $\omega \geq 5$).

Species	Common name	Cell number	$\log_2$(cells)	n (QC)	Median ω (QC)
Chlamydomonas reinhardtii	Unicellular reference	1	0.00	353	**0.057** (boundary-biased; see note 1)
Tetrabaena socialis	4-cell colonial	4	2.00	1,358	**0.096** (known rate outlier, §4.2)
Gonium pectorale	16-cell colonial	16	4.00	1,351	0.073
Astrephomene gubernaculifera	32-cell Goniaceae partial soma	32	5.00	1,355	0.074
Yamagishiella unicocca	32-cell isogamous Volvocaceae	32	5.00	1,354	0.066
Eudorina sp.	32–64-cell anisogamous	48	5.58	1,352	0.065
Pleodorina starrii	~96-cell Volvocaceae partial soma	96	6.58	1,341	0.082
Volvox africanus (NIES-4468)	Thai homothallic, germ-soma	~1,500	10.55	1,217	0.089
Volvox reticuliferus (NIES-3785)	Japanese heterothallic female, germ-soma	~1,500	10.55	1,294	0.144 (assembly-affected; see note 2)
Volvox carteri (Eve)	Full germ-soma	~2,000	10.97	1,351	**0.060** ⭐

Notable observations:

1. **Chlamydomonas boundary filtering**: 1,311 of 1,664 Model 1 ω values for *Chlamydomonas* are at the PAML lower numerical boundary (≤ 0.0001), reflecting saturation of *d_S* on the long *Chlamydomonas* terminal branch in the 10-species panel.

The QC-filtered median ω of 0.057 is therefore based on a non-random subset (the 353 orthogroups where d_S remained interpretable) and is biased downward relative to a hypothetical complete-data median. Comparisons involving *Chlamydomonas* free-ratio ω should be interpreted accordingly.

2. **Volvox carteri exhibits the lowest median ω among multicellular species** (0.060), consistent with the disposable-soma prediction that the species with the most extensive germ-soma differentiation in our sample experiences the strongest additional constraint above the multicellular baseline. *V. carteri*'s strongest-constraint position is consistent with its complete germ-soma differentiation (regA-dependent; Kirk 2005; Hanschen et al. 2016) and with its IQ-TREE rank in the gradient (§3.3 — note that branch length and ω convey related but distinct information; see §4.4).

3. **Volvox africanus exhibits an elevated free-ratio ω** (0.089) despite the shortest overall amino-acid terminal branch length in the panel (0.055; §3.3) — interpretable as a slow-evolving genome with a normal d_N/d_S ratio, rather than as the strongest purifying selection.

4. **Volvox reticuliferus exhibits the highest ω among the three Volvox species** (0.144), consistent with the *V. reticuliferus*-specific orthogroup-level saturation discussed in Methods §2.4 and Discussion §4.6: the *V. reticuliferus* QC-passing set of 1,294 orthogroups is enriched for orthogroups in which *V. reticuliferus* showed intermediate divergence (rather than IQ-TREE saturation), and these intermediate-divergence orthogroups appear to include a tail of orthologs with elevated ω. The underlying biological or annotation cause is not the subject of the present manuscript; the PGLS gradient result is robust to complete *V. reticuliferus* exclusion (Results §3.4).

5. **Astrephomene's free-ratio ω** (0.074) is similar to the colonial intermediates *Gonium* (0.073), *Yamagishiella* (0.066), and *Eudorina* (0.065). Partial-soma evolution appears to produce a more modest free-ratio ω reduction than full-soma evolution.

6. **Tetrabaena socialis** elevated ω is consistent with prior phylogenomic identification of *Tetrabaena* as an evolutionary-rate outlier (Featherston et al. 2018; Hu et al. 2020; see Discussion §4.2).

Table S5b. PGLS regression of log(median Model 1 ω) on $\log_2$(cell number)

Analysis	Species included	*n*	Slope	SE	*t*	*P*	R^2	Pagel's λ (ML)
(a) All 10 species	All	10	(See manuscript Note A)					
(b)	(a)	9	(See					

Analysis	Species included	*n*	Slope	SE	*t*	*P*	R^2	Pagel's λ (ML)
Excluding *Chlamydomonas*	minus Chlamy		Note A)					
(c) Excluding *Chlamydomonas* + *Tetrabaena*	(b) minus Tetra	8	(See Note A)					

Note A: The per-species PGLS using Model 1 ω values (rather than IQ-TREE branch lengths) is presented in this Table S5 as a supplementary cross-validation of the primary PGLS analysis in Results §3.4. The primary PGLS uses per-species median IQ-TREE LG+Γ amino-acid terminal branch lengths (which are not subject to the *d_S* saturation that biases the *Chlamydomonas* Model 1 ω). The Model 1 ω PGLS, while qualitatively consistent with the IQ-TREE PGLS in direction, is less reliable due to the boundary-filtering bias on *Chlamydomonas* and the saturation issue on *V. reticuliferus*. The IQ-TREE-branch-length PGLS ($\beta = -0.123$, $P = 0.002$) reported in §3.4 is the primary statistical demonstration of the dose-response.

Table S5c. Per-orthogroup Spearman correlation across species

For each orthogroup with ≥5 species having QC-passing ω values, the Spearman rank correlation between $\log_2$(cell number) and $\log(\omega)$ was computed across species. The proportion of orthogroups with negative correlation was tested against the null expectation of 0.5 using a two-sided binomial test.

Analysis	n OGs analysed	OGs with $\rho < 0$	Proportion	Binomial *P*
All 10 species	(To be computed in production R script)	—	—	—
Excluding *Chlamydomonas* (9 species)	(To be computed)	—	—	—

The per-OG gradient test is qualitatively consistent with the IQ-TREE per-OG Spearman test, in which 96.1 % (1,395/1,451) of complete-data orthogroups showed negative correlations and the species-label permutation test gave an empirical one-sided $P = 0.0107$ (10,000 permutations; Methods §2.5, Results §3.4); the Model 1 ω version may show a weaker signal due to the boundary effects noted above.

Comparison with the IQ-TREE amino-acid analysis

The per-OG Model 1 ω analysis reported here is qualitatively consistent with but interpretively distinct from the IQ-TREE amino-acid analysis reported in §3.3–§3.4. The IQ-TREE branch length is a continuous measure of total substitutions per site on each terminal branch; the Model 1 ω is a ratio of non-synonymous to synonymous substitutions on each branch. These two statistics convey related but distinct information about per-species selection intensity:

- **Strong purifying selection at all sites + normal evolution rate** → low ω, normal branch length → *V. carteri*-type pattern
- **Slow overall evolution + typical purifying selection** → normal ω, short branch length → *V. africanus*-type pattern
- **Annotation/orthology artefact** → variable ω + saturated branch length → *V. reticuliferus*-affected pattern

The IQ-TREE amino-acid analysis is methodologically more robust (per-OG branches are interpretable continuous values; no boundary effects), whereas the Model 1 ω analysis provides the conceptually direct measure of per-species purifying-selection intensity that is conventionally interpreted in the d_N/d_S literature. The two analyses are complementary, and the primary results of this manuscript are based on the IQ-TREE branch-length analysis (Results §3.3, §3.4) and the codon-model branch-model 2 analysis (Results §3.1, §3.2); the Model 1 free-ratio analysis (this Table S5) provides supporting evidence for lineage-specific allocation patterns discussed in §4.4.

Files

File	Content
`results/paml_dnds/volvocine_10species_model1/model1_per_species_omega.csv`	Per-OG ω for each of 10 species (1,664 OGs × 31 columns)
`results/paml_dnds/volvocine_10species_model1/per_species_medians.json`	Per-species medians (this Table S5a) as JSON

Citation

PAML codeml (Yang 2007). All data and code deposited at Zenodo (DOI: 10.5281/zenodo.21094823).