The relationship between intron-richness and complex multicellularity has been decoupled repeatedly from genome size evolution

Irma Lozada-Chávez\(^1,2,3,^*\), Peter F. Stadler\(^1,2,3,4,5,6,7\), and Sonja J. Prohaska\(^1,2\)

\(^1\)Department of Computer Science, University of Leipzig, HärTELstraβe 16-18, D-04107 Leipzig, Germany.
\(^2\)Interdisciplinary Center for Bioinformatics, University of Leipzig, HärTELstraβe 16-18, D-04107 Leipzig, Germany.
\(^3\)Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.
\(^4\)Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany.
\(^5\)Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria.
\(^6\)Santa Fe Institute, 1399 Hyde Park Rd., 87501 Santa Fe NM, USA.
\(^7\)Center for non-coding RNA in Technology and Health.

\(^*\)Corresponding author: ilozada@bioinf.uni-leipzig.de

Compiled January 22, 2017

Text S1. Overview of the work flow applied in the GenomeContent pipeline
Text S2. Density, average size and content calculations as implemented in GenomeContent
Text S3. Intron features are evolving independently across eukaryotes
Text S4. Higher non-repetitive intron contents occur in some unicellular and multicellular organisms with small genomes
Text S5. Defining complex multicellularity: beyond the number of cell types

1. METHODS

S1. Overview of the work flow applied in the GenomeContent pipeline

GenomeContent was written in Perl to calculate global statistics and sequence-based estimators of genome features in five major steps (see Figure 2). First, the processing of gene annotations focuses on identifying protein-coding gene coordinates, as well as filtering and checking intron annotations. Second, the coordinates are taken then as “reference sets” for introns, exons and intergenic regions to directly estimate several statistic descriptors, such as intron size, density and number, according to the definitions described in supplementary text S2. Third, the “reference sets” sets are projected onto the genome sequence in both strands, so that genome-feature contents are calculated according to the equations S2, S3 and S4 described in supplementary text S2. Finally, all statistic descriptors obtained with the program are provided as text files and exploratory figures (see also Figure S2). GenomeContent runs on an entire genome in few minutes or hours, depending on genome size and number of annotated genes, and is available with detailed documentation at the link: http://www.bioinf.uni-leipzig.de/publications/supplements/16-014/

The first step of GenomeContent is a filtering process. We excluded introns and exons smaller than 15 nucleotides (nts), although small exons and introns (of \(\sim\) 20 nts) have been reported in several eukaryotic organisms [1–7]. Alternative splice variants were kept in the data. To avoid redundant/overestimated data, exons with partial or full matching boundaries to exons of other transcripts were overlapped; the same rule was applied to introns. In both cases, their coordinates were joined or replaced accordingly; thus, every exon and intron is only counted once. We did not include features occurring at the UTR regions [8, 9] because there are very few genome projects fully annotated at the 5’ and 3’ boundaries of genes [10, 11]; thus, these regions were taken as part of the intergenic ncDNA. For the purpose of this study, we call this filtered set of protein-coding gene coordinates for every genome as the “reference gene set”.

Although the collected gene annotations are based on transcript evidence, we implemented the approach proposed by Roy and Penny (2007) in GenomeContent to further estimate systematic errors in protein-coding gene annotations by identifying the excess/deficit of the intron-length distributions modulo 3. Since introns are not expected to respect the coding frame, intron lengths \(3n, 3n + 1\), and \(3n + 2\) should appear in similar fractions \(p_{3n} \approx p_{3n+1} \approx p_{3n+2}\). As stated by Roy and Penny (2007), large values of the “3n excess”, \(E_3 = p_{3n} - (p_{3n+1} + p_{3n+2})/2\), suggest that a considerable fraction of internal exons could have been incorrectly predicted as introns or that there are several “intron retention” events. On the other hand, a deficit of \(3n\) introns, i.e., \(E_3 \ll 0\), may suggest that a considerable fraction of \(3n\) introns—lacking of stop codons—may have been mistaken for exons. With exception of intron-poor species (e.g., parasites and endosymbiots), most of gene annotations collected for this
study shown values of the 3n excess close to 0. (see Figure S1 and Table S1S). Few genomes were excluded from the present study because their corresponding intron-length distributions showed an extreme 3n excess (between 0.4 and 0.7).

S2. Density, average size and content calculations as implemented in GenomeContent

Several statistical estimators for every genome feature were obtained with GenomeContent. Genome size is defined as the net length of nucleotides and placeholders of all sequences conforming the nuclear genome. Exon density and intron density were estimated as the mean number of exons in genes, or introns per intron-bearing genes, respectively, per genome (“absolute-ID”). We employ the absolute-ID, instead of the number of introns per sequence region (which depends on genome size) or from the total number of pc-genes (“normalized-ID”), which depends on the number of pc-genes), because we actually want to know how many introns those pc-genes containing introns within a genome have on average, regardless of genome size and the total number of pc-genes. Other features associated with intron-richness were measured independently; for instance, the fraction of pc-genes harboring introns from the total number of pc-genes, which might reflect complete intron loss from the pc-gene structures.

For instance, the four species listed below (Table S1) exhibit similar normalized-IDs (around 5 introns per pc-gene), although they have a different fraction of pc-genes containing introns (from 70% to 96%). Clearly, the bias is produced by their different number of pc-genes. The absolute-ID, instead, clearly shows that X. laevis and B. ﬂoridae have indeed more introns per pc-gene, within a smaller fraction of pc-genes containing introns in comparison to the other species, and regardless of having a different number of pc-genes. The absolute-ID is, thus, more appropriated to understand even small changes of intron-richness.

Also, an absolute-ID might buffer dramatic changes among updated genome releases and across genomes, owing possibly to considerable under/overestimation of pc-gene models or correction of gene annotation (e.g., pseudogenes). Accordingly, an absolute intron density correlates very weakly with the total number of CDS (PGLS adj.R² = 0.0205, p < 0.001, n = 461) and the fraction of CDS containing introns (PGLS adj.R² = 0.2306, p < 0.001, n = 461).

Furthermore, the average feature size (e.g., intron size and exon size) of genes per genome was calculated in two ways. The straight average (A_feature) depends on the number of data points from the whole sample (i.e., gene models with introns), which equally contribute to the final average regardless of which gene they belong to. On the other hand, the weighted average (A_feature) depends on the gene-structure of the genome, and thus it samples more broadly the data points that contribute, in the case of introns, to the very well known high-skewed length distribution. We used the weighted average (A_feature) for the statistics analyses of this study because it helps to normalize the exon and intron lengths of genes with different number of exons and introns [12, 13].

Accordingly, the straight average is calculated as the total length of all feature sequences (exons or introns) in a genome (L_feature) divided by the total number of all feature sequences (exons or introns, respectively) in a genome (N_feature):

\[ A_{\text{feature}} = L_{\text{feature}} / N_{\text{feature}} \]

where \( A_{\text{feature}} \) is the average length of the respective feature (exon or intron) within single gene (\( A_{\text{feature}} = l_{\text{feature}} / n_{\text{feature}} \)), then the weighted average, \( \bar{A}_{\text{feature}} \), is calculated as the mean of the \( A_{\text{feature}} \) values of the respective feature (exons or introns) in a genome, according to [14]:

\[ \bar{A}_{\text{feature}} = \frac{1}{n} \sum_{i=1}^{n} A_{\text{feature}} \]  (S1)

where \( n \) represents the total number of genes in a genome when calculating \( A_{\text{exon}} \), or the total number of intron-bearing protein-coding genes in a genome when calculating \( A_{\text{intron}} \).

Finally, we calculated the “feature content”, i.e., the fraction or proportion of nucleotides of the respective genome feature (intron, exon, intergenic) that contributes to genome size. By using the “reference gene set” of every genome project, we projected the sets of genomic intervals for all exons and introns from the genes located in the plus strand (setA), the minus strand (setB), and of the isomers (setC). Since the protein-coding capability of a genomic region could be argued as derived even from one strand only, we use the following rule to estimate the dominance of gene features when calculating their contribution to genome size: Exon > Intron > Intergenic-region. Figure 2 shows a schematic overview of the genomic intervals and their projections to calculate the genome-feature content for every genome. Accordingly, we calculated the exon content (EX_content) in Megabases (Mb) and as percentage (%) of the total genome size from the net nucleotide length of the corresponding union intervals of all pc-exons found in sets A, B and C (E pc_\text{content}) as follows:

\[ \text{EX}_{\text{content}} = \sum_{i=1}^{g} E_{\text{pc}_i} \]  (S2)

where \( g \) correspond to the total number of genes in a genome. We then calculated the intron content (IN_content) in Megabases (Mb) and as percentage (%) of the total genome size from the net nucleotide length of the corresponding union intervals of all introns flanked by pc-exons (Ipc_\text{content}) that do not overlap exonic nucleotides (Ipc_\text{content} = Ipc_\text{content} - Ipc_\text{content}), according to:

\[ \text{IN}_{\text{content}} = \sum_{i=1}^{g} l_{\text{pc}_i} \]  (S3)

where \( g \) correspond to the total number of intron-bearing genes, and \( l_{\text{pc}_i} \) represents the symmetric difference among \( l_{\text{pc}_i} \) intervals without the intersection of all those intronic nucleotides overlapping exonic nucleotides (Ipc_\text{content} = Ipc_\text{content} ∩ EX_content). Similarly, the intergenic content in Megabases (Mb) and as percentage (%) of the total genome size was calculated from the net nucleotide length of all union intervals of intergenic regions (IR_\text{content}) that do not overlap the exonic and intronic contents (IR_\text{content} = IR_\text{content} - (EX_content + IN_content)) as described by:

\[ \text{IR}_{\text{content}} = \sum_{i=1}^{n} l_{\text{r}_i} \]  (S4)

It is important to note that the new coordinates generated to calculate the nucleotide content of exons (E pc_\text{content}), introns (Ipc_\text{content})
and intergenic regions ($I_{IR}$) were also used to create a “reference feature-content set” that was required to calculate the contribution of repeats content to every of these genome features.

Finally, by taking into account four major descriptors of intron features, GenomeContent estimates the Intron-richness Index (IRI) of a genome as:

$$IRI = \frac{1}{4}(ID + IP + IC + SL_{IS})$$  \hspace{1cm} (S5)

where $ID$, $IP$ and $IC$ correspond to, respectively: intron density, the fraction of CDS with introns, and the genome fraction of intron content, according to the definitions described previously. Because introns exhibit high-skewed length distributions, $SL_{IS}$ measures the fraction of introns from the total number that are longer than twice the median size.

### 2. RESULTS

#### S3. Intron features are evolving independently across eukaryotes

**Fungi.** With some important exceptions, genes in Fungi exhibit the shortest intron sizes of Eukarya, ~75% of their intron population is ≤100 nts (Figure 3a). Nevertheless, three global trends can be distinguished within this supergroup. With an average genome size of 47.3 Mbs, Basidiomycetes exhibit the highest fractions of genes harboring introns (81.8%) as well as higher intron densities (4.7), although with small intron sizes of ~90 nts (Table 4). Despite the fact that the sequenced species from Zygomycota, Chytridiomycota and Pezizomycotina have genome sizes similar to Basidiomycota (Table 4), they show a reduction in the fraction of genes harboring introns (65-76%) and also in their density (2.9-3.8), but their intron sizes are larger (~140 nts). Finally, the compact genomes in Saccharomycotina (Ascomycota) exhibit an increase of intron sizes (Figure 3a: ~50% of their introns are between 250 and 1,000 nts) as well as a clear intron loss due to the drastic reduction of intron density (1.2) and the fraction of genes with introns (16.3%).

**Protists.** As expected, the genome-wide variation of intron features across the protists genomes is very heterogeneous (Figure 3b). With some remarkable exceptions, however, ~80% of the introns in protists have a size ≤250 nts, around 60% of their genes harbor introns, with a density between one and four introns per gene. A dramatic intron loss can be observed in most parasitic species from Excavata, Apicomplexa and Entamoebidae (Table 4) owing to the reduction in the number of introns per genome, of intron density (1-2) and the low fraction of genes with introns (5-30%). Remarkably, some unicellular, multicellular and even parasitic protists such as B. natans, E. siliculosus, S. minutum, N. caninum, T. gondii and H. hammondi show high intron densities (7-16), large intron sizes (500-800 nts) and most of their genes (80-95%) harbor introns, which resemble to those values found in some land plants and vertebrates. Likewise, the small genomes from Choanozoa (61.6 Mbs), in particular Sphaeroforma arctica and S. rosetta, not only present large intron sizes (315.3 Mbs), but they also show higher intron densities (5.9) and high fractions of genes with introns (80.7%) in comparison to the global trends of protists. Chlorophytes, in particular Volvox carteri and Asterochloris sp, also exhibit unexpected intron patterns. Most of the green algae genes harbor introns (92.6%), in a very high density (7.76) and with sizes comparable to those of land plants (305.5 nts) (Figure 4b and Tables S2 and S9). By contrast, the sequenced members of Rhodophyta and Prasinophytes (Chlorophyta) exhibit a dramatic intron loss, as supported by a low fraction of genes harboring introns (~30%), one or two introns per gene (Tables 4 and S2), and intron sizes from 50 nts (in the unicellular Galardia sulphuraria) to 350 nts (in multicellular Pyropia yezoensis).

**Viridiplantae.** Despite the huge variation of genome size in Viridiplantae (13.3 Mbs to 22.5 Gbs), ~75% of the introns in most sequenced plants have a length of ≤500 nts (Figure 4b), and ~75% of their genes harbor introns, with a density of 4 to 6 introns per gene (Table 4). The relative constancy of intron features in plants highlights the considerable changes in intron sizes observed at specific lineages. In Streptophyta, for instance, the bryophyte Physcomitrella patens, the lycophyto- phyte Selaginella moellendorfii, the two carnivorous plants from Lamiaceas and members from Brassicales (with exception of Carega papaya) show a strong reduction of intron sizes (Figure 4b: ~75% of their introns are ≤250 nts), although their other intron features do not differ from the global plant trends (Table 4). By contrast, the single species sequenced from some basal tracheophytes such as Pinus taeda, Amborella trichopoda, Phalaenopsis equestris, Elaeis guineensis, Nelumbo nucifera and Actinidia chinensis show the largest intron sizes of Viridiplantae, between 1,500 to 3,500 nts, and their other intron features do not differ from the global plant trends either (Table 4). The increase of intron size in these plants is probably the consequence of the contribution of repeats to genome size as suggested by the estimations shown in Table 4.

**Metazoa.** The genome size in Metazoa ranges from 27.97 Mbs to 6.5 Gbs, most of their genes harbor introns (87.0%), with a density of 7.5 introns per gene and large intron sizes (2,643.5 nts). However, the within-genome variation of intron features is not uniform across the major metazoan clades (Figure 4a). The first difference is found between Protostomes and Deuterostomes, and within each lineage, a further distinction can be observed between slow-growing organisms (like verte-
brates and butterflies) and fast-growing species (like fish and insects). Within the eumetazoan species, *M. leidyi* and the cnidarian genomes present considerable large intron sizes (1,511 nts), comparable to those found in some teleosts genomes, although with a reduced fraction of genes with introns (70%). As originally reported [15], *L. migratoria* (6.5 Gbs) exhibits the largest intron sizes of Eukarya, 13,440.6 nts (indeed, ~75% of its introns are between 5,000 and 50,000 nts). Surprisingly, intron density (5.73) and the fraction of genes with introns (81.68) in the locust genome have not increased with respect to the average values shown in other protostome clades.

The sequenced protostomes have genomes from 63.53 Mbs to 6.5 Gbs. Although the fraction of genes harboring introns (~85%) and the intron density (5.2) in protostomes is homogeneous among lineages, the variability around the average intron size (1,535 nts) is quite heterogeneous. For instance, the small genomes in Nematoda (116.7 Mbs) have also reduced intron sizes (424.5 nts) in comparison to the large introns present in most of the sequenced genomes in Lophotrochozoa (765.6 Mbs), with platyhelmints exhibiting intron lengths (2,902 nts) similar to those found in vertebrates. A remarkable exception is the microscopic and asexually bdellid rotifer *Adineta vaga* (244.5 Mbs): it has undergone a drastic reduction of intron size (90% of the introns are <100 nts), although its other intron features do not differ from the global trends. Within Arthropoda, Diptera exhibits a reduced intron density (3.5) and displays a very narrow skewed intron-size distribution; this is, although intron size seems large on average (1,651 nts), ~50% of introns exhibit small sizes (<100 nts). By contrast, around half of the intron population in Lepidoptera exhibit sizes between 500 and 1,000 nts; so that, moths and butterflies have on average larger intron sizes (1,577.7 nts). The within-genome variation of intron size in Hymenoptera (1,267 nts on avg.) seems to be more stable across species, ~50-60% of the introns in bees and ants are ≤250 nts.

On the other hand, deuterostomes have genomes sizes between 70.47 Mbs and 3.6 Gbs. Similar to protostomes, the fraction of genes with introns (90.1%) and the intron density (9.3) barely vary across clades. By contrast, intron size reflects the major phylogenetic clades within Chordata. For instance, around 50% of the introns in the few sequenced genomes of Echinodermata, Hemichordata and Cephalochordata have a size between 500 and 1,000 nts. Surprisingly, the genome-wide variation of intron size in the small Percomorpha genomes (774.5 Mbs) is similar to that observed in bees and ants. Indeed, ~50% of the introns in ray-finned fishes are below 250 nts (Figure 4a), although the average intron size in other teleosts genomes is slightly larger (1,662 nts).

The genome of the ocean coelacanth *Latimeria chalumnae* (2.8 Gbs) clearly exhibits the shift to the emergence of the largest intron sizes in Deuterostomia (4,566 nts). Although the three amphibian genomes have similar genomes sizes (2.4 Gb) to the ocean coelacanth, they show a consistent reduction of intron features: only 81.1% of their genes have introns, with a density of 8.79, and intron sizes of 2,385.7 nts (Table S9). Furthermore, two clear patterns can be differentiated within Sauropsida: reptiles and birds (Figure 4a, Table 4). As previously reported [16, 17], birds exhibit not only a constrained genome size (1.2 Gbs), but they also have a reduced intron size (3,342 nts) in comparison to reptiles and mammals. However, birds show a slight increase in the fraction of genes harboring introns (92.9%) and intron density (10.07). On the contrary, the few sequenced genomes from Crocodylia, Turtles, Serpentes and Iguana are not only larger on average (3,668.7 Gbs), but their intron sizes are similar to mammals or even larger (4,982.6 nts). Finally, the mammalian genomes (2.7 Gbs) show an slight reduction in the fraction of genes harboring introns (86.8%), although their intron density and size (4,938 nts) do not considerably differ from those found in the coelacanth and reptilian genomes.

**S4. Higher non-repetitive intron contents in some unicellular and multicellular organisms with small genomes**

**Protists.** On remarkable occasions, intron content covers from one quarter up to one third of the genome size in protists like the filamentous brown alga *Ectocarpus siliculosus* (195.8 Mbs), the three obligate intracellular parasites *Neospora caninum* (59.1 Mbs), *Toxoplasma gondii* (63.71 Mbs) and *Hammondia hammondi* (67.67 Mbs), the free-living chlorarachniophyte *Bigelowiella natans* (94.7 Mbs), the unicellular dinoflagellate *Symbourdidiun minutum* (1.5 Gbp) and the colonial choanoflagellate *Salpingoeca rosetta* (55 Mbs) (Figure 3b and Table 4).

**Viridiplantae.** Within land plants (Figure 4b), the very small genomes of the complex carnivorous plants *Ultracearicia gibba* (81.88 Mbp) and *Genlisea aurea* (63.6 Mbp) have 18.2% (repeats: 4.9) and 13.1% (repeats: 6.1%) as intron content, respectively. These proportions are not only almost free of repeats but they are also similar or in some cases even higher than the fractions observed in their sequenced close relatives within Lamiales and from most plants with larger genome sizes (Tables 4 and S2). For instance, the large genomes of the loblolly pine *Pinus taeda* (22 Gbp), the tobacco *Nicotiana tabacum* (5 Gbp) and the sequenced species from Triticaceae (4-5 Gbp) reach barely ~3% as intronic sequences (Table S2). Our intron content calculation partially disagree with the initial estimation reported for *U. gibba* by [18], which was probably underestimated owing to the low content of repetitive ncDNA (3%) and the short intron sizes found in this genome. Aside from land plants, the most striking highest intron contents in Viridiplantae are observed for most of the small genomes from Chlorophytes (50-150 Mbs) (Table 4), with exception of the non-photosynthetic and pathogenic *Heliosporidium sp.* Green algae harbor between 25 and 40% of their genome size as intronic sequence, with variable but still minor contributions from repeats (1.6% to 19.6%) (Figure 4b).

**Metazoa.** In some metazoan lineages and species, intron content can explain between 25% and 40% of genome size (Figure 4a); however, these fractions are mainly shaped by the differential contribution of repeats among the major clades. Within invertebrates (Table 4), for instance, intron contents with a minor contribution of repeats (5-15%) are observed in: the eusocial and solitary bees from Apoidea (291.8 Mbs on avg.), the tentaculate ctenophore *Mnemiopsis leidyi* (303.2 Mbs), the placozoan *Trichoplax adhaerens* (39.1 Mbs) and the nematode *Ascaris suum* (244.5 Mbs). On the other hand, invertebrates with similar high intron contents, but with a major contribution from repeats (25-50%), include: the butterflies and moths in Ditrysia (~2.4 Gbs), the platyhelmints *Clonorchis sinensis* (644 Mbs) and *Schistosoma mansoni* (254.3 Mbs), the nematode Caenorhabditis elegans (92.9 Mbs) and the migratory locust *Locusta migratoria*
Within the genomes of Deuterostomia (500 Mbs to 4 Gbp), at least one third of the high intron contents (25-40%) observed in basal deuterostomes (Ptychodera flava, Saccoglossus kowalevskii, Stronglylocentrotus purpuratus and Branchiostoma floridanum), Teleostei, Reptiles and Mammalia is repetitive sequences (Figures 4a and S3a, Tables 4, S2 and S9). However, the fractions of intronic repeats in basal deuterostomes (39%) and mammals (34.1%) are higher in comparison to the proportion observed in ray-finned fishes (19.1%), but similar to the intronic repeat content (30.5%) from the few sequenced genomes of frogs, S. cerevisiae. Likewise, unicellular organisms may experiment development. For instance, some single-celled organisms may have several differentiated state cells during their life time cycle, e.g., S. cerevisiae. Likewise, unicellular organisms may experimentally develop simple multicellularity –as long as the selective pressure is maintained– in organisms naturally living as unicellular, such as S. cerevisiae and C. reinhardtii [23, 24]. Furthermore, some unicellular and simple multicellular organisms may develop into a multinucleated mass of cytoplasm that is either partially separated by cell membranes or not at all separated into UCTs, e.g., syncytium and coenocyte sensu amplio, plasmodium, pseudomycelium, among others [25, 26]. Examples include Dictyostelium [27], Caulerpa [28] and Neurospora [29].

By following previous work [30-35], thus, we created here three “working definitions” embracing three basic but distinctive aspects of cellular development and life cycle in order to distinguish complex multicellularity (CM) from simple multicellularity (SM) and unicellularity:

(1) The number of differentiated state cells (i.e., unique cell types) at once across a life cycle.
(2) The duration of the global and coordinated organization of the cell types within a life cycle.
(3) The reversion to any of the other cellular organismal states during the life cycle.

Accordingly, we distinguish:

**Unicellular**: is an organism exhibiting a single differentiated state cell at once across its life cycle. Single-celled organisms using a pseudohypha with very few spores to reproduce or having a filamentous development experimentally induced by stress are also included in this category.

**Simple multicellular (SM)**: is an organism exhibiting either several aggregated cells (with differentiated cell types or not) or a syncytium/coenocyte sensu amplio, arranged in a reproducible spatial pattern during a transient time of its life cycle. Reversion to unicellularity may occur.

**Complex multicellular (CM)**: is an organism exhibiting irreversible spatio-temporal organization formed by several differentiated cell types, arranged in a global and coordinated pattern during most of its life cycle. Reversion to unicellularity or SM life style does not occur.

Based on phenotype, the differences between unicellular and multicellular life-styles entail basically two hallmarks: the aggregation of several cells (either the same or different types) and their transition in individuality (during a transient or longer stage of a life cycle) [32, 33]. By contrast, the transition from SM to CM is not that obvious [30, 31]. Beside the fact of having a larger number of UCTs [36], CM is genetically constrained (at least) by the spatial and temporal differentiation of several cell types through clonal-unitary development [32, 37].

As a consequence of the genetic and environmental contingent events leading to cell differentiation from a common cell-line ancestor, complex multicellularity is irreversible, regardless the duration of such stage within the organism’s life cycle. Multicellular reversion should not be confused here with the “unicellular-multicellular transitions” that all multicellular organisms, either simple or complex, undergo by means of reproductive processes through the life cycle [34]. Likewise, defectors previously reported in vertebrates –reviewed in [38]– do not count as reversals because they do not involve irreversible differentiation of the whole organism, but just a few “mutant-selfish” cell lineages, which are still dependent on the hierarchical structure of the host. In contrast, SM can originate from diverse genetic mechanisms and developmental pathways beyond clonal-unitary development, such as filamentation and aggregative development, which may have facilitated not only its independent emergence several times across the Tree of Life, but also the reversion of the whole organism to unicellular states, as suggested by some experiments [23, 24]. Through these definitions, thus, we attempt to test our hypothesis regarding the existence of pervasive ncDNA factors shaping the emergence of CM [31].

Our approach was particularly useful to define the intricate cellular state of several non-model organisms. Among the species in our dataset that represent independent origins of CM, land plants and eumetazoan animals have a body plan characterized by embryonic development [35, 39], while complex multicellular fungi and the filamentous brown alga E. siliculosus has some sort of early embryonic stages [35, 40]. However, it still represents a challenge to distinguish between SM and CM in species within Fungi, Protists and Parazoa, whose multicellular body plans are generated from a few unique cell types or during very transient stages of their life cycles. For instance, the “basidia” (commonly known as the mushroom fruiting body) and the “ascocarp” basically represent the CM state in fungi, this is, the sexual reproductive form and usually the final state of the life cycle in Basidiomycota and Ascomycota, respectively. Other controversial cases are found in Porifera. Despite the fact that most sponge embryos and larvae have an anterior-posterior axis with radial symmetry, the adult demosponge loses such elementary spatio-temporal organization at metamorphosis to assume its sessile and differentiated
but highly plastic body form [41, 42], which allows the dispersal of UTCs and re-aggregation into a new body sponge [42]. Since this is the first attempt to formally define CM versus SM, we acknowledge that our definitions and the classification for the most controversial cases in this study are not free of future improvements and corrections.

REFERENCES


25. V. Coneva and D. H. Chitwood, “Plant architecture without multicellularity: quandaries over patterning and the