bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genome-wide features of introns are evolutionary decoupled among themselves and from genome size throughout Eukarya

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, Härtelstraße 16-18, D-04107 Leipzig, Germany.*
[2] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 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***

ABSTRACT. The impact of spliceosomal introns on genome and organismal evolution remains puzzling. Here, we investigated the correlative associations among genome-wide features of introns from protein-coding genes (*e.g.*, size, density, genome-content, repeats), genome size and multicellular complexity on 461 eukaryotes. Thus, we formally distinguished simple from complex multicellular organisms (CMOs), and developed the program `GenomeContent` to systematically estimate genomic traits. We performed robust phylogenetic controlled analyses, by taking into account significant uncertainties in the tree of eukaryotes and variation in genome size estimates. We found that changes in the variation of some intron features (such as size and repeat composition) are only weakly, while other features measuring intron abundance (within and across genes) are not, scaling with changes in genome size at the broadest phylogenetic scale. Accordingly, the strength of these associations fluctuates at the lineage-specific level, and changes in the length and abundance of introns within a genome are found to be largely evolving independently throughout Eukarya. Thereby, our findings are in disagreement with previous estimations claiming a concerted evolution between genome size and introns across eukaryotes. We also observe that intron features vary homogeneously (with low repetitive composition) within fungi, plants and stramenophiles; but they vary dramatically (with higher repetitive composition) within holozoans, chlorophytes, alveolates and amoebozoans. We also found that CMOs and their closest ancestral relatives are characterized by high intron-richness, regardless their genome size. These patterns contrast the narrow distribution of exon features found across eukaryotes. Collectively, our findings unveil spliceosomal introns as a dynamically evolving non-coding DNA class and strongly argue against both, a particular intron feature as key determinant of eukaryotic gene architecture, as well as a major mechanism (adaptive or non-adaptive) behind the evolutionary dynamics of introns over a large phylogenetic scale. We hypothesize that intron-richness is a pre-condition to evolve complex multicellularity.

INTRODUCTION

Spliceosomal introns are not only germane to eukaryote origins, they also represent an evolutionary innovation on the way in which protein-coding genes have been stored, expressed and inherited throughout Life's history on Earth. Spliceosomal introns (hereafter "introns") form a class of non-coding DNA (ncDNA) sequences that interrupt exons within a gene. Thus, they have to be removed from the primary transcript by the splicing machinery to form a mature messenger RNA (mRNA), and hence, a functional RNA or protein molecule. Although introns are ubiquitous sequences along eukaryotes, their genome-wide features (such as length and abundance within and across genes) differ among species, and their the origins are still under intense debate. Yet, some large scale patterns of intron evolution appear to be reaching a consensus. For instance, the high conservation of intron-positions found in orthologous genes throughout eukaryotes suggests the existence of intron-rich ancestors [1, 2]. Also, intron loss has been found to be more frequent than intron gain in most lineages [1, 3–7], although episodes of rapid and extensive intron gain are also observed across eukaryotes [8–13]. Other evolutionary and functional aspects of introns remain amongst the longest-abiding puzzles, such as the phenotypic consequences of harboring intron-rich genes and their evolutionary relationship with genomic and multicellular complexity.

Because introns are less evolutionarily constrained than coding sequences, they usually evolve at high rates as do 4-fold degenerate sites and other non-coding regions [14, 15]. Nevertheless, a variable proportion of intron sites has been found to be under selective constraints in mammals [15–17], some invertebrates [18, 19], fungi[20], algae and plants [21–23]. Also, the energetic and time costs to transcribe and splice introns can be significant enough to influence the organism's phenotype [24–27]. This is expected, in part, because eukaryotic genomes are pervasively transcribed [28, 29] and intronic RNAs constitute a major fraction of the transcribed non-coding sequences [30]. For instance, the transcription of a large gene, as the one encoding human dystrophin (2.3 Mbs), can still take up to 10 hrs. at an *in vivo* RNA pol-II elongation rate of 3.8 kb/min [31] because

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

99% of its length is intronic [32]. A substantial delay of gene expression owing to transcription and splicing of long and/or numerous introns, a phenomenon termed *"intron delay"* [33, 34], has turned out to be essential for cells with short mitotic cycles and for timing mechanisms during early body segmentation [35, 36]. Also, levels of gene expression (either housekeeping or tissue specific) are often associated to particular intron features [37]. For instance, some highly expressed genes are found under strong selection to remain intron-poor for transcriptional efficiency [25, 38, 39], whereas other genes are found to have longer and numerous introns to enhance expression [40–44].

Although co-transcriptional splicing depends on many parameters [45, 46], the exon-intron structure of genes is also found to have an impact on the mode of splice-site recognition and the efficiency of splicing [45, 47–51]. For instance, it was recently found that intron-containing genes and intron-rich genomes are best protected against R-loop accumulation, and subsequent transcription-associated genetic instability, by favoring spliceosome recruitment [52]. Studies across phylogenetically distant eukaryotes have also found that the length of introns and exons exerts an important influence on the likelihood of an exon to be constitutively or alternatively spliced [47–51, 53–55]. For instance, canonical splicing errors produced by splice-site recognition across small introns are more likely to result in *"Intron Retention"* or unspliced mRNA [54]. Whereas *"Exon Skipping"* or inclusion of alternative exons in the mRNA is thought to comply best with splice-site recognition across small exons [54]. Remarkably, these canonical and other non-canonical splicing errors can account for the major portion of the alternative splicing (AS) events in some eukaryotes and cell types [53, 56–58]. The expansion of AS events is thought to be key in the emergence of multicellular complexity, by creating proteome diversity and by regulating gene expression post-transcriptionally through RNA surveillance pathways [2, 59]. Accordingly, species with more tissues and cell types tend to have more alternatively spliced genes [53, 60, 61].

It remains to be fully understood, however, to what extent intron-richness (and potentially increased AS events) is coupled to the evolution of complex multicellularity (as defined in Appendix 1), genome size and of other ncDNA classes across eukaryotes. In earlier studies, a number of strong positive correlations over large evolutionary scales was found among genome size and particular ncDNA classes [62–64], including the average size, total number and nucleotide content of introns in the genome [62, 64–66]. These results have fueled the suggestion that changes in the genomic features of any ncDNA class are scaling uniformly with changes in genome size, leading to the premise that larger genomes tend to harbor more and longer introns, and *vice versa* for smaller genomes. However, correlative associations over large evolutionary scales are prone to result in biases due to low phylogenetic diversity and the lack of both phylogeny-controlled statistics [67, 68] and systematically obtained datasets. Challenging these results are also the studies showing that the average number of introns per gene (*i.e., intron density*, see Appendix 2) appears to inversely correlate with generation time [26, 69] and gene expression levels (as reviewed in [34, 37]). Accordingly, genome-wide changes of intron density are found to vary widely across eukaryotic lineages [2, 7, 70], with no clear association to genome size or another intron feature [6, 7]. Furthermore, it has been largely presumed that repeats –in particular, transposable elements (TEs)– are strongly driving the evolution of some intron features [66, 71, 72]. However, the strong contribution of repeats to intron size, for instance, is supported by studies on a few model species, particular clades or repeat families [73–76].

Some of the findings and correlative associations described previously have been interpreted as evidence for either adaptive or non-adaptive forces being the major determinant of the intron-richness complexity observed across eukaryotes. And to argue, consequently, on whether the functionality of introns can be mainly explained by the effect for which it was selected for (*i.e., selected-effects*) or by the effect of causal-role activities [29, 71, 72, 77]. For instance, the "genomic design" model postulates that the length and number of introns is determined by selection for gene function and the necessity to preserve conserved intronic elements for complex regulation [78]. Likewise, the "selection for economy" model proposes that decreases in genic size are the results of selected mutations (mainly on the length and number of introns) to reduce the time and energetic costs of transcription [25, 38, 39]. By contrast, the "mutational bias" model states that the abundance and length of introns in certain chromosomal regions is driven by different recombination rates and/or transcription-associated mutational biases [38, 70, 79]. Alternatively, the "mutation-hazard" model suggests that variation in the hazardous accumulation of introns –along with other ncDNA sequences– is primarily the outcome of increased genetic drift when effective population sizes (N_e) remain small for an extended period of time [62, 63, 80]. So that, for instance, larger genomes would have more and larger introns owing to insufficient purifying selection to remove them in species with lower values of N_e, a condition expected to occur in multicellular organisms particularly. The strong correlations reported among genome size, $Ne\mu$ (as a proxy of N_e) and some ncDNA classes –including some genome-wide features of introns– appear to support this hypothesis [62, 63, 65, 66, 68, 81]. Several controversies have emerged, however, from the contradicting evidence and arguments supporting all previous hypotheses, as discussed in [37, 67].

The discrepancy between the current observations and the evolutionary models have raised a conundrum: are the genome-wide features of introns within protein-coding genes (such as their length, abundance and repetitive composition) evolving throughout Eukarya in either a concerted or an independent way among themselves, with genome size and multicellular complexity? Our study contributes to clarify this conundrum by investigating the correlative associations among these organism traits over 461 eukaryotes. To that end, we formally distinguish simple from complex multicellular organisms (CMOs) (see Appendix 1), and developed the program `GenomeContent` to systematically estimate genomic traits (see Appendix 2). We then estimated correlations under phylogenetic controlled analyses, taking into account significant uncertainties in the tree of eukaryotes and variation in genome size estimates. We found that intron features are weakly correlated among themselves and with genome size at the broadest phylogenetic scale, revealing different associations between those features estimating intron abundance across genes and those measuring intron length and repeat composition. We also found that CMOs and their closest ancestral relatives are characterized by high intron-richness, regardless their genome size. These patterns contrast

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the narrow distribution of exon features found across eukaryotes. Our findings are thus in disagreement with previous estimations claiming a concerted evolution between genome size and introns over a large phylogenetic scale. They also argue strongly against both a particular intron feature as key determinant of eukaryotic gene architecture, as well as a major mechanism (adaptive or non-adaptive) behind the evolutionary dynamics of introns over a large phylogenetic scale. Here, spliceosomal introns are unveiled as a dynamically evolving ncDNA class, whose relationships with genome and organismal complexity are better explained by the influence of numerous life-history factors and evolutionary forces. We argue why intron-rich lineages are more likely to evolve complex multicellularity.

RESULTS

Phylogenetic signal is strong in genome-based traits and robust to uncertainties in the tree of eukaryotes

Phylogenetic signal is the "tendency of related species to resemble each other more than species drawn at random from the same tree" [82, 83]. Because the genome traits of the 461 eukaryotes analyzed here share an evolutionary history, we first evaluated the strength of their phylogenetic signal (*i.e.*, statistical dependence) with four alternative "species trees", each of which has been extensively used in the literature. Figure 1 shows the comparison of the tree topologies: a tree based on literature consensus from sequence-based phylogenies (Figure 1a); two NCBI taxonomy-based trees, one with no polytomies (Figure 1b), while another one with polytomies (Figure 1c); and a protein domain-based tree corrected for protein content biases derived from differences in genome size and lifestyles (Figure 1d). The literature consensus-based tree was selected as the "reference tree for eukaryotes" (Figure 1a) to present the results throughout the article. However, we do not attempt to single out any particular tree topology as the best or the correct species tree of eukaryotes. Instead, the goal is to test the robustness of the comparative analyses to alternative phylogenetic assumptions [67, 82].

Accordingly, dissimilarity metrics for the four tree topologies can be observed in Figure 1. The symmetric difference (RF) and the tree aligment (Align) metrics measure the number of clades not shared between two trees from either the total number of their partitions or the best alignment of their branches, respectively. Notably, the highest phylogenetic inconsistencies are observed in the protein domain-based tree: between 1.28 and 1.48 partitions per species when compared to the other trees (see table on Figure 1). This is because the protein domain-based phylogeny exhibits a *long branch attraction* (LBA) problem of several species that had undergone massive protein-domain loss, in spite of implementing a correcting factor for protein-domain content. Examples of LBA on Figure 1d include the myxosporean *Thelohanellus kitauei*, the trematode *Schistosoma mansoni*, the bdelloid rotifer *Adineta vaga* and the green algae *Helicosporidium sp*. In contrast, we observe a similar magnitude of branch dissimilarities between the literature-based tree and the NCBI-based trees, even with different polytomy resolutions: between 0.66 and 0.86 partitions per species. On closer inspection, however, the phylogenetic resolution at the level of species is not only different across the four trees, but also known conflicting hypotheses are observed for the phylogenetic positions of Rhodophyta, Rhizaria, Excavata, Stramenopiles, Alveolata, among others [84, 85]. As summarized above, the four alternative eukaryotic trees exhibit significant phylogenetic inconsistencies from one another. This allow us to incorporate adequate phylogenetic uncertainty into our comparative analyses to evaluate their sensitivity.

Table 1 presents estimates of strong phylogenetic signal for all 30 sequence-based genome traits analyzed here, as indicated by their Pagel's λ values close to 1.0 and significantly > 0. Notably, the λ values are significantly robust to the phylogenetic disagreements shown by the alternative tree topologies. Likewise, λ values are significantly robust to different estimations of genome sizes and genome contents based on two sources: genome assemblies and experimental estimations (see Supplementary Table S4). The latter robustness is expected because, as also reported by Elliott and Gregory [64], the correlation between the assembled and estimated genome sizes is strong at the broadest phylogenetic scale: $r = 0.958$ (see Table 3). Consistent with previous studies [67, 76, 81, 86], these results indicate that genome traits are not statisticallly independent when compared among species. Therefore, correction for phylogenetic signal is accounted for any comparative analyses in this study.

Genome size correlates weakly with genome-wide intron features at the broadest phylogenetic scale

Table 2 shows the estimated log Bayes Factors (log BF) and coefficients of determination (r^2) used here as the criterion to assess both the "strength of the evidence" and the "explanatory power" of the correlative associations between two traits (X and Y), respectively. The "explanatory power" of the r^2 values should be understood as means of a statistical range to associate the variation observed between X and Y, with no implication as to the evolutionary mechanism that might cause (or not) such associations nor the primary trait (X or Y) subject to this action.

When the data is analyzed as phylogenetically independent with the OLS model, strong associations are observed on Table 2 among genome size and most intron features (r^2 between 0.6 and 0.9). In particular, the regression between genome and intron sizes ($r^2 = 0.76$) is consistent with previous estimations performed over a broad evolutionary range by Vinogradov ($r^2 = 0.792$, $n = 27$) [65] and by Lynch and Conery ($r^2 = 0.641$, $n = 30$) [62]. However, the strength of such associations substantially dropped after phylogenetic corrected regressions were performed with both PGLS and PICs models (see Table 2). For instance, the correlations among genome size and the features estimating intron length and genomic content have robust evidence for a positive association but a weak explanatory power at the broadest phylogenetic scale: $r^2 = 0.382$ ($logBF = 80.1$) for intron size, $r^2 = 0.447$ ($logBF = 95.1$) for intron content, and $r^2 = 0.485$ ($logBF = 119.6$) for the repetitive-intronic content of the genome. On the other hand, none or no simple associations were found among genome size and those features measuring the abundance of introns within and across genes: $r^2 = 0.068$ ($logBF = 1.3$) for intron density, $r^2 = 0.022$ ($logBF = 1.7$) for the fraction (and total number) of intron-containing genes per genome, and $r^2 = 0.184$ ($logBF = 7.8$) for the total number of introns per genome. Therefore, only reduced and differential fractions of the variation observed among intron features (from 2% to 45%) can be associated (directly or indirectly) to the ~2,200-fold variation of the genome size observed over the 461 eukaryotes analyzed here.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The strenght of these correlations, either through log BF or r^2 values, is not only similar regardless which genome trait (*e.g.*, genome size) is associate to the X and Y variables, but it is also consistent with both PGLS and PIC models (see Supplementary Tables S5-S10). Nonetheless, the PGLS model appears to fit the data significantly better than the OLS and PIC models do, as shown by its consistent lower AIC values and higher ML values (see Table 2). We further investigated potential discrepancies in our correlations owing to the differences to estimate tree topologies, genome traits, species number and diversity (see Supplementary Tables S5-S10). As shown in Table 2, the weak associations found among genome size and intron features also proved to be robust to the phylogenetic inconsistencies of the alternate trees. A slighter strenght should be notice in the correlations performed with the protein-domain content tree, which exhibits the highest topological and branch length dissimilarities in comparison to the other trees. We also found no major discrepancies in the PGLS correlations performed with estimates of intron features based on (and against) the two sources of genome size. The PGLS correlations described in Table 3 further show that the weak associations found at the broadest phylogenetic scale, significant r^2 values from 0.02 to 0.5, are robust to randomly reduced datasets of 231 and 116 species sampled from the original 461 species (see also Supplementary Table S11). These results support the robustness of the phylogenetic diversity of our species dataset. By contrast, we found overestimated correlations among intron features and genome size when smaller and less diverse datasets are used to cover large phylogenetic scales. This is observed in Table 3 for two further randomly reduced datasets of 58 and 29 species, as well as for the species datasets from Lynch and Conery [62] ($n = 26$) and from Wu and Hurst [68] ($n = 30$) (see also Supplementary Table S12). These contrasting results show that phylogenetic diversity and phylogenetic corrected correlations over large evolutionary scales are strongly affected by very small and biased datasets.

Under the "replicated co-distribution" approach [87], we also tested the decoupled association between intron features and genome size across multiple independent clades. Table 3 describes the PGLS correlations performed over 18 lineage-specific datasets compiled from the original dataset of 461 eukaryotes (see Methods and Supplementary Table S11). According to our previous results, the r^2 values show that the strength of the associations among genome size and intron features is indeed different at the local phylogenetic scale. For instance, the genome-intron size relationship varies from $r^2 = 0.003$ in stramenopiles up to $r^2 = 0.856$ in teleosts, whereas the association between intron density and genome size varies from $r^2 = 0.077$ in deuterostomes up to $r^2 = 0.807$ in chlorophytes. Likewise, a differing correlative association among intron features and genome size is observed in Hymenoptera, Aves, Monocots and Ascomycota when compared to the correlations observed in their corresponding close relatives.

Nevertheless, some of the correlations on Table 3 also highlight the impact that differences in the estimations of genome features have at the local evolutionary scale. For instance, the correlation obtained for genome size and intron density in Ascomycota is consistent with Kelkar and Ochman [81], regardless both the tree topology or the source for the genome size estimates used to perform the regressions. This is expected because the assembled and estimated genome sizes are highly correlated in this group ($r^2 = 0.964$, $p < 0.001$, Table 3). By contrast, our correlations for genome and intron size in amniotes are not consistent with those reported by Zhang and Edwards [76] when the estimated genome sizes are used to perform the regressions ($r^2 = 0.184$, $p < 0.001$, Table 3), but they are when the regressions are performed with the assembled genome sizes ($r^2 = 0.339$, $p < 0.001$, Supplementary Table S11). Such discrepancy might be the consequence of the low correlation found between the two sources of genome size estimates in amniotes: $r^2 = 0.567$ ($p < 0.001$, see Table 3). On the other hand, some of the regressions obtained across specific lineages (such as monocots and aves) or for particular intron features (such as intron density and the fraction intron-containing genes) do not reach statistical significance. While most of these correlations reveal none or no simple associations among intron features and genome size, caution should be taken in the correlations obtained for those lineages with few sequenced genomes ($n < 15$). With few exceptions, the PGLS correlations reported previously, and in the following sections, do not pose significant changes when they are performed with different tree topologies and assembled genome sizes (see Supplementary Tables S5-S13).

Changes of intronic content across lineages are not strongly associated to one particular genome-wide intron feature

As observed on Figure 2 and Table 3, the net nucleotide coverage of introns in the eukaryotic genome, *i.e.* "intron content", represents on average 25.8% of the genome size in Choanozoa, 21.9% in Metazoa, 11.8% in Viridiplantae, 11.1% in Alveolata, 10.4% in Amoebozoa, 7.5% in Stramenopiles, 7.0% in Fungi, and 0.9% in Excavata. Our results show that the variation of intron content observed at the broadest phylogenetic scale is positively, yet weakly correlated with genome size ($r^2 = 0.447$, $logBF = 95.1$), in contrasts to the stronger associations observed with the repetitive ($r^2 = 0.723$, $logBF = 1273.3$), non-repetitive ncDNA ($r^2 = 0.859$, $logBF = 449.2$) and protein-coding ($r^2 = 0.617$, $logBF = 98.5$) contents of the genome (see Table 2). As a consequence, low intron-contents are observed in some large and highly repetitive genomes, such as *Pinus taeda* (1.69% in 22.5 Gbs) and *Locusta migratoria* (13.61% in 6.5 Gbs). And *vice versa*, high non-repetitive intron contents can be observed in species with smaller genome sizes, either unicellular or multicellular, when compared to their corresponding close relatives. Some remarkable examples include Chlorophytes (29.2%, 70.7 Mbs), Aves (31.5%, 1.23 Gbs), teleosts (31.5%, 871.6 Mbs), bees (25.4%, 278.6 Mbs), buterflies (23.9%, 402.0 Mbs), *Bigelowiella natans* (32.05%, 94.7 Mbs), *Ectocarpus siliculosus* (36.04%, 214 Mbs), and *Utricularia gibba* (18.15%, 88 Mbs) (see Figures 2-4 and Supplementary Tables S14-S15). Consistent with this, we further found that the strength of the association between intron content and genome size is indeed different across lineages (Table 3). For instance, it is strong in Alveolata, Chlorophyta and Teleostei, while weak or absent in Pezizomycotina, Aves, Mammalia, Monocots, Hymenoptera and Lepidoptera. In agreement with a previous study [64], these findings show that intron content cannot fully account for the large variations of eukaryotic genome sizes.

Our findings also show that intron content in eukaryotes is not strongly associated to any other particular intron feature at the broadest phylogenetic scale. As observed in Table 2, changes of intron size ($r^2 = 0.336$), intron density ($r^2 = 0.392$), the frac-

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tion of genes harboring introns ($r^2 = 0.441$) and the repetitive-genome content ($r^2 = 0.314$) are similarly associated to the variation of intron content observed across eukaryotes. To a larger extent, intron content is associated with changes in the total number of introns per genome ($r^2 = 0.713$) and the repetitive-intron content ($r^2 = 0.738$). It is important to clarify, however, that the strength of the association among intron content and other genome-wide intron features is different across lineages (see Table 4 and Supplementary Table S13), as we report next.

Genome-wide intron features are decoupled among themselves at the local and broadest phylogenetic scales

The repeated decoupling of intron features from genome size evolution is also supported by two additional observations. First, the evolution of intron features appears to be differentially constrained by phylogeny rather than genome size throughout Eukarya. As shown in Figures 2-4 and also Supplementary Tables S14-S15, a homogeneous variation can be observed in the phylogenetic patterns of intron features across land plants, fungi and stramenophiles. This contrast the dramatic variation –at a much faster evolutionary pace– observed in the phylogenetic patterns of intron features within holozoans, chlorophytes, alveolates and amoebozoans. Second, the PGLS correlations in Tables 2 and 4 show that intron features associate weakly among themselves at the local and broadest phylogenetic scales (see also Supplementary Table S6-10 and S13). These associative correlations further support differing patterns between those features estimating intron abundance within and across genes (*i.e.*, intron density and the fraction of intron-containing genes, respectively) and those features measuring intron length and repeat composition. For instance, variation in intron size is weak but to a larger extent associated with the repetitive content within introns ($r^2 = 0.366$, $logBF = 588.2$), rather than with changes in intron density ($r^2 = 0.025$, $logBF = 0.2$), the number of introns per genome ($r^2 = 0.003$, $logBF = 1.6$), or the fraction of intron-containing genes ($r^2 = 0.002$, $logBF = 1.6$). Likewise, variation on intron density is weakly but to a larger extent associated with the number of introns per genome ($r^2 = 0.437$, $logBF = 161.5$) and the fraction of intron-containing genes ($r^2 = 0.199$, $logBF = 152.0$), rather than with changes in intron size, the repetitive content within introns ($r^2 = 0.210$, $logBF = 0.5$) and the total number of CDS ($r^2 = 0.024$, $logBF = 0.074$). As also observed in Table 4, the strenght of these associations fluctuates at local phylogenetic scales.

The evolutionary decoupling of intron features is also observed at the lineage and species levels (see summary statistics in Table 3 and Figures 3-4). For instance, Pezizomycotina and Eudicots show similar fractions of intron-containing genes on average (~75%), although their corresponding avg. intron size (107.2 nts and 495.6 nts) and intron density (2.3 and 5) are different. Likewise, the small genomes of bees (288.7 Mbs) harbor shorter but more abundant introns within (6.2) and across (94.7%) genes than the larger genomes of butterflies (402.0 Mbs, 5.6 and 86.5% introns within and across genes, respectively). Strikingly, introns in Aves are slightly more abundant within (10.07) and across (92.9%) genes in comparison to mammals (9.3 and 86.8% introns within and across genes, respectively) (see Table 3 and Figure 4a). Even despite the fact that birds have undergone a reduction of their average intron (3,342 nts) and genome (1.2 Gbs) sizes, as observed here and elsewhere [76, 88, 89]. Consistent with this, birds and mammals show similar fractions of AS genes and have among the highest rates of AS events per gene [60]. Within fungi, we observe that introns within Basidiomycota are shorter (92.2 nts) but more abundant within (4.7) and across (81.8%) genes when compared to the larger (141.4 nts) but less abundant introns within (1.9) and across (53.4%) genes in Ascomycota (see Table 3 and Figure 3a). Indeed, an increase in the average rate of alternative splicing from 6% in Ascomycota to 8.6% in Basidiomycota has been previously reported [61]. Noteworthy, the large genome of the locust *L. migratoria* (6.5 Gbs) exhibits the lastest intron sizes of Eukarya [90]: ~75% of its introns are between 5,000 and 50,000 nts. Surprisingly, the abundance of its introns within (5.73) and across (81.68%) genes is not higher with respect to other protostome clades. Additional examples are described in Supplementary Tables S14-S15.

Repeats differentially contribute to intron size and content across eukaryotes

We also investigated how significant is the contribution of repeats to the size and content of both introns and genomes. In partial agreement with previous studies [63, 64], we found that genome size is strongly associated at the broadest phylogenetic scale with its repetitive content ($r^2 = 0.723$, $logBF = 1273.3$), but to a much lesser extent with the repetitive-intronic content ($r^2 = 0.485$, $logBF = 119.6$) (see Table 2). As observed on Figure 2 and Table 3, repeats cover on average 13.6% of genome size in fungi, 41.6% in Viridiplantae, 26.5% in Metazoa, 17.7% in Choanozoa, 26.9% in Amoebozoa, 25.2% in Excavata, 26.7% in Stramenopiles, and 12.1% in Alveolata. When still observed at the supergroup level, the fraction of repeats covering the intronic genome content appears to mirror the previous trends: 8.8% in Fungi, 24.2% in Metazoa, 20.2% in Viridiplantae, 19.7% in Choanozoa, 29.6% in Amoebozoa, 13.0% in Excavata, 19.8% in Stramenopiles, and 14.6% in Alveolata. At the broadest phylogenetic scale, however, the repetitive content of the genome does not strongly correlate with intron size ($r^2 = 0.292$, $logBF = 56.4$) or intron content ($r^2 = 0.314$, $logBF = 65.7$), and it does not significantly associate either with intron density ($r^2 = 0.032$, $logBF = 3.1$) or the fraction of intron-containing genes genes harboring introns ($r^2 = 0.018$, $logBF = 2.9$) (see Table 2). These results show that the repetitive composition of introns is not strongly scaling with changes in genome size or the repetitive content of the genome at the broadest phylogenetic scale. Yet, fluctuations in the associations (from absent to strong) among intron features and the repetitive genome content are expected across lineages.

Figure 2 also shows whether or not the contribution of repeats to intron size is significant in every genome analyzed, according to the $p < 0.001$ obtained for the permutation tests performed on the *Jaccard index* (see Methods and Supplementary Table S16). In summary, we found no significant degree of nucleotide overlap between repeats and intronic sequences in most of the genomes analyzed in Fungi (71.0%), Viridiplantae (82.1%, with exceptions such as Prasinophytes), Aves, Oomycetes, Rhodophyta, and over half of the genomes within Mammalia. By contrast, the degree of overlap between repeats and intronic sequences was found statistically significant in most of the genomes analyzed within Metazoa (with notable exceptions), Amoebozoa (71.4%), and Alveolata (80%) (see Supplementary Table S16). Examples of species and clades with

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

considerable high repetitive-intron contents (25-50%) and significant nucleotide overlaps between repeats and introns include: the haptophyta *Emiliania huxleyi*, the sea sponge *Amphimedon queenslandica*, Dictyostelia, Prasinophytes, Cnidaria, protostomates such as Lophotrochozoa, Diptera and Lepidoptera, as well as vertebrates such as Teleosts, Anura, Turtles, Crocodylia and Mammalia. With exception of Prasinophytes, high repetitive-genome contents (25-40%) are also observed in the previous lineages. Conversely, an apparent concerted reduction in the contribution of repeats to genome size (6-15%) and intron content (3-14%) –also supported by a non-significant overlap between repeats and intronic sequences– is observed in most species within Ascomycota, Nematoda, Hymenoptera, Aves, and in *Trichoplax adhaerens*. This particular pattern requires further research. Additional examples with statistic estimators are revisited in Supplementary Tables S14-15, S18.

Genome-wide exon features are also weakly associated with genome size, but evolving steadier than intron features

We further estimated the phylogenetic patterns of exons features across eukaryotes to contrast those observed among intron features. As shown in Figures 2-4 and Table 5, around half of the genome size is covered by exonic sequences in Fungi (44.9%), Amoebozoa (53.8%), Alveolata (51.9%), Prasinophytes (81.5%) and Excavata (53.1%). In contrast, low fractions of exon content are observed in the genomes of land plants (13.8%) and metazoans (5.2%), with pinus and mammalian genomes harboring barely 1% of protein-coding nucleotides. As with intron features, we found that the variation estimated for exon features is weakly associated to the variation observed in eukaryotic genome sizes at the broadest phylogenetic scale. As observed in Table 2 and Supplementary Tables S6-S10, genome size is positive but to a lesser extent associated with: exon content ($r^2 = 0.262$, $logBF = 6.2$), the total number of both exons ($r^2 = 0.316$, $logBF = 11.3$) and CDS ($r^2 = 0.328$, $logBF = 20.3$), as well as the average length of CDS ($r^2 = 0.263$, $logBF = 43.8$). These estimations are in agreement with previous studies [64, 66]. However, none or no simple associations were found for genome size against exon density ($r^2 = 0.054$, $logBF = 0.05$), exon size ($r^2 = 0.096$, $slope = -0.114$, $logBF = 0.2$), and the repetitive-exon content ($r^2 = 0.135$, $logBF = 0.9$).

Contrasting the intron size's patterns, the distribution of exon size is tighter across eukaryotes (see Tables 3 and 4). We found that between 50% and 75% of the exon population within a genome has a narrow and small length below 250 nts across plants, green algae, Choanozoa, basiodiomycetes, and Metazoa. Accordingly, we observed that exon length distributions are less-skewed in most eukaryotic clades (Figures 3D and 4D) when compared to intron length distributions (Figures 3C and 4C). Nevertheless, large exon lengths are observed in Rhizaria (407.2 nts), Amoebozoa (600.7 nts), Stramenopiles (687.5 nts), Alveolata (880.7), Excavata (1,357.3 nts), Prasinophytes (985.9 nts), Rhodophyta (870.9), Ustilaginomycotina (1,107.1 nts) and Ascomycota (721.6 nts), particularly in Saccharomycetes (1,167.23 nts) (see Figures 3D-4D and Supplementary Figures S7-S10). As observed in Table 2, furthermore, two different correlative associations were found for exon size at the broadest phylogenetic scale. On the one hand, exon size significantly decreases as the density and total number of exons in the genome increases ($r^2 = 0.731$, $slope = -0.880$, $logBF = 97.5$ and $r^2 = 0.517$, $slope = -1.305$, $logBF = 10.5$, respectively). On the other hand, none or no simple associations were found between exon size and the number ($r^2 = 0.097$, $logBF = 0.6$) or length of CDS ($r^2 = 0.050$, $logBF = -0.01$). These results are in overall accordance with previous estimations [66, 91, 92].

As observed in Table 5 and Figure 2, the fraction of repeats covering the exonic genome content across eukaryotes is considerably small (between 2% and 8%), in comparison to introns. Although larger fractions are observed in some species within land plants (23.05%), Lophotrochozoa (15.2%), Cnidaria (26.3%), Basiodiomycota (10.4%), Excavata (17.5%) Amoebozoa (16.2%) and Stramenophiles (13.0%). In agreement with these observations, the average length of exons is not significantly associated to its repeat content ($r^2 = 0.034$, $slope = -0.564$, $logBF = 2.3$) or the genome repeat content ($r^2 = 0.094$, $slope = -1.269$, $logBF = 0.1$). Based on the results from the *Jaccard index*'s permutation tests (see Figure 2), we found no significant degree of overlap between repeats and exonic sequences across most of the eukaryotic genomes analyzed here: Amoebozoa (100%), Choanozoa (100%), Metazoa (97.1%), Fungi (92.4%), Excavata (92.4%), Viridiplantae (86.5%), Stramenophiles (82.4%) and Alveolata (80%). The noteworthy exceptions can be observed in Figure 2 and Supplementary Table S16.

Intron-richness is robustly associated to complex multicellular organisms and their closest ancestral relatives

We further investigated the relationship between genome-wide intron features and multicellular complexity. As described in Appendix 1, we developed four criteria and three definitions to distinguish the species in our dataset as: complex multicellular (CMOs: 288), simple multicellular (SMOs: 96) or unicellular (77) organisms (see Supplementary Table S2). Accordingly, CMOs are defined here as those *organisms exhibiting an irreversible transition in individuality produced by tissue-based body plans, through the developmental commitment of multiple and different cell types originated from a common cell-line ancestor*. We then used Principal Component Analyses (PCAs) with direct comparative data (compPCA) and phylogenetically independent contrasts (phyloPCA) to investigate how seven intron features are covarying among themselves (Figure 5b), with other eight genome features (Figure 5e), with complex multicellularity (Figure 5a,d), and with genome size (Figure 5c,f). Noteworthy, the findings that are going to be described next, also hold for both the compPCA analysis and the phyloPCAs performed with different tree topologies and assembled genome sizes (see Supplementary Figures S11-S15 and Tables S18-S20). As observed in Figure 5, high intron-richness does segregate CMOs (in red) from both SMOs (in blue) and unicellular organisms (in yellow), by clustering through genome-wide intron features exclusively and in conjunction with other genome features. The first two principal components (PC1 and PC2) from both phyloPCA analyses capture most of the variances in the data: 87.16% with 7 intron features and 64.6% with 15 variables. However, the association of the variables on each principal component is different in both phyloPCAs (see Table 6), as described next.

By analyzing the phyloPCA of intron features, we first observe that high intron-richness clusters CMOs together (Figure 5a), regardless of the lineage they belong to (Figure 5b) or the wide dispersion of their genome sizes (Figure 5c). We further observe that the content, repeat composition and number

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of introns are mainly contributing to PC1 (71.82%), while PC2 has a major contribution (73.24%) from the fraction of intron-containing genes, intron size and density (see Table 6). In the broarder phyloPCA analysis that includes additional genome features, we also found that intron features are leading the clustering of CMOs, while other genome features are clearly setting apart the major eukaryotic supergroups: metazoans (in pink), plants (in green), fungi (in blue) and protists (in brown). Consequently, the association between high intron-richness and CMOs is found to be in part a byproduct of other genomic features constraining the development of multicellularity in every supergroup. For instance, PC1 clearly captures the *ncDNA complexity* of the genomes, while PC2 captures their *protein-coding complexity*. Accordingly, PC1 has a major contribution (51.5%) from intron content, unique intron content, unique ncDNA, genome size and genome repeat content (see Table 6). Notably, the scaling distribution of genome sizes along PC1 in Figure 5f endorses the strong correlative association found between ncDNA and genome size at the broadest phylogenetic scale ($r^2 = 0.859$, see Table 2). Conversely, exon features (size, content, repeat composition), the number of CDS, the fraction of intron-containing genes and intron density contribute mostly to PC2 (89.8%). Hence, intron-richness in land plants is mainly associated with the high contribution of repeats to the genome, while intron-richness in metazoans is mostly related to the non-repetitive ncDNA fraction of the genome. By contrast, most unicellular and SM species (in fungi and protists) are strongly associated with several exon features, particularly with exon size and the number of CDS.

The previous patterns strongly suggest that high intron-richness is a robust genomic fingerprint of both CMOs and their closest ancestral relatives. For instance, ancestral relatives clustered among CMOs include: the choanoflagellates *Salpingoeca rosetta* and *Sphaeroforma arctica*, the sponge *A. queenslandica*, the slime mold *Fonticula alba*, the green alga *Volvox carteri*, and other chlorophytes (see Figure 5a,d). Nevertheless, there are some exceptions to this trend. First, there are also few exceptional intron-rich unicellular organisms clustered among CMOs, with no evidence of multicellularity (either simple or complex): the dinoflagellate *Symbiodinium minutum*, the chlorarachniophyte alga *B. natans*, the green alga *Chlamydomonas reinhardtii* and the apicomplexans *Toxoplasma gondii*, *Neospora caninum* and *Hammondia hammondi*. The high intron-richness of these unicellular species –comparable to the one observed in vertebrates– has been already acknowledged, and is suggested to have a role in the development of their complex life cycles [93–96]. Second, we observed a very few intron-poor CMOs, such as the red algae *Chondrus crispus* and the highly derived cnidarian parasite *T. kitauei*, that have undergone massive loss of introns and genome reduction due to extreme lifestyle conditions [44, 97]. Until more genomes of multicellular red algae are sequenced, however, it would be unclear to know whether the relative intron-poorness observed in *C. crispus* is an ancestral constraint or a derived (exceptional) trait [44]. Despite the presence of few intron-poor CMOs, the statistical robustness of the association between complex multicellularity (CM) and intron-richness is provided by most of the 288 intron-rich species classified as CMOs in this study, rather than by the five so far (out of the six) independent instances where CM has evolved. This is because intron loss overcomes intron gain across eukaryotes [1, 3, 4, 7] and intron-richness does not necessarily depend on genome size (this study). Therefore, there are no reasons to expect that high intron-richness has been maintained from intron-rich ancestors nor that it is only present in CMOs with large genome sizes. Yet, the major inconsistency found on both phyloPCA analyses (Figure 5) is the absence of a clear separation between some fungal SMOs and CMOs. This can be explained by the difficulty to determine the precise multicellular lifestyle of several fungi as simple or complex, owing to the lack of detailed life-cycle descriptions and the presence of "fuzzy" fruiting body development that challenge the evaluation of the criteria 3 and 4 in our definitions (see Appendix 1). This is an issue not only faced here, but also discussed elsewhere [98–101], and thus require further research.

DISCUSSION

Spliceosomal introns form a dynamically evolving ncDNA class, most likely under the influence of diverse life-history factors and evolutionary forces

Consistent with previous results [45, 66, 92] and with some noteworthy exceptions, our findings show that exons within protein-coding genes have remained within a narrow average size between 150 and 300 nts in Metazoa, Choanozoa, Viridiplantae and Basidiomycota. Hence, we observe that significant modifications in the structure of protein-coding genes in these lineages are basically a consequence of changes in the length and abundance of introns. Unexpectedly, these and other genome-wide features of introns are found to be repeatedly decoupled among themselves and from genome size evolution throughout Eukarya. Three findings support this observation. First, the strength of the associations among genome size and genome-wide intron features is different at the lineage-specific level, as consistent with other studies [76, 81, 102, 103]. Second, the features estimating the length and abundance of introns in a genome are weakly associated among themselves at the local and broadest phylogenetic scales. This explains the heterogeneous and contrasting patterns of intron evolution reported in the literature [2, 7, 55, 70, 76, 81]. As a consequence of the previous findings, changes of intron content cannot fully account for the large variations of eukaryotic genome sizes (in agreement with [64]), nor be strongly associated to the variation of one particular intron feature. Third, the repetitive composition of introns is not necessarily scaling with changes in genome size or the repeat content of the genome. Indeed, introns are found to be far from representing repetitive sequences in several lineages. As argued below, these results do not contradict –and even endorse in several cases– the contribution that different repeat classes have to either the origins [9, 11, 12] or length extension [73–76, 103] of introns at particular lineages. Therefore, our findings collectively unveil spliceosomal introns as a dynamically evolving ncDNA class.

Can a major mechanism (adaptive and non-adaptive) offer a unifying explanation to the highly heterogeneous patterns of intron evolution and genome complexity? Our findings suggest that this is highly unlikely. None of our correlative analyses imply causation nor offer evidence for the evolutionary mechanisms explaning the phylogenetic patterns reported here. Yet, strong (linear) correlations among several measures of genome complexity haven been often provided as evidence, by some evolutionary models [62, 65], to imply a concerted evolution

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

among genome size, ncDNA content and intron-richness across eukaryotes (also discussed in [37, 67]). We observe here that genome size, its overall ncDNA and repetitive contents are indeed strongly associated at the large evolutionary scale (see Table 2). However, we also demonstrate that changes in the variation of some intron features (such as size and repeat composition) are only weakly, while other features measuring intron abundance are not, scaling with changes in genome size at the broadest phylogenetic scale. Our findings are thus in clear disagreement with previous estimations claiming the opposite [62, 63, 65, 66, 71]. Moreover, our results show that the genome-wide features determining length and abundance of introns across protein-coding genes are largely evolving independently throughout Eukarya. Our results are thus inconsistent with both a particular intron feature as key determinant of eukaryotic gene architecture, as well as a major mechanism (adaptive or non-adaptive) underlying a concerted effect between genome size and intron-richness over a large phylogenetic scale. Instead, the repeated decoupling among intron features themselves and with genome size strongly suggests that the major genome-wide features of introns from coding genes are evolving under the influence (direct or indirect) of either different or several life-history factors and evolutionary forces.

For instance, non-adaptive mechanisms such as the long-term evolutionary dynamics of repeats –which depend on factors like methylation propensity, RNAi-mediated interference and mating system– are found to determine concerted changes of genome and intron size in certain lineages [67, 73]. Examples observed here and elsewhere include some species within red algae and plants [44, 74, 103–105], insects [90, 106], fish [75] and birds [76, 88, 89]. Other studies suggest that intron density is determined by mechanistic factors such as nonhomologous end-joining (NHEJ) of DNA segments, reverse transcriptase and transposition activities [7, 70]. Also, introner-like elements greatly contribute to episodic intron gains in some algae [9, 12] and fungi [11]. And recently, rates of spontaneous intron-creating and -deleting mutations were found to shape the intron-exon structures of several distantly related species [107]. Under adaptive forces, variations of intron size and density across several eukaryotic lineages have been associated with the action of natural selection: (a) to conserve regulatory binding sites [17, 20, 21, 23] and regulatory ncRNAs [108–110]; (b) to promote the creation of new exons in vertebrates [49, 111]; (c) to reduce splice error rates [68] and protect against transcription-associated genetic instability [52]; (d) to favor co-transcriptional splicing and nucleocytoplasmic export of highly expressed and rapidly regulated cell-cycle genes [34, 37, 112]; (e) to reduce the metabolic costs associated with either powered flight in birds [76, 88] or environmental changes of habitats in teleosts [113].

Consequently, our findings also endorse concerns [67, 114, 115] regarding how much of the content, variation and complexity of intron-richness (along with other ncDNA classes) in Eukarya can be explained by the strong action of N_e and genetic drift over a large evolutionary scale, as the "mutational-hazard" (MH) model states [62, 63, 80, 116]. In addition to the inconsistencies discussed previously, other studies [67, 68, 86] were not able to find statistically significant associations among $Ne\mu$ and several intron and genome features after removing the phylogenetic signal from the dataset of Lynch and Conery [62]. Another studies have also shown that N_e cannot exclusively or even largely explain major changes of intron and genome sizes in lineages within amniotes [76], insects [117], ascomycetes [81], and plants [67, 86]. Furthermore, reliable estimations of N_e (such as $Ne\mu$ and K_a/K_s) are still a matter of debate [67, 118], since they do not correlate well [68, 119] and are affected by several life-history traits [114, 120]. It is important to note that our findings do not dismiss the impact that N_e and genetic drift has on the accumulation of ncDNA in eukaryotes. Rather, they argue that the population genetic settings suggested by the MH model are most likely to be dominant at the local phylogenetic scale, or in particular intron features from coding genes, or during recent founder events, or over introns located at non-coding regions.

The robustness of systematic and phylogenetically controlled analyses

The results summarized above are based on a phylogenetic controlled framework over the largest and most diverse dataset of eukaryotic complete genomes to date. A major concern is, however, that phylogenetic uncertainty might affect considerably any phylogenetically controlled analysis [67, 82]. Here, we recapitulated no significant changes in our results after taking into account significant and numerous phylogenetic disagreements from four tree topologies estimated for the 461 species analyzed in this study. We could argue that the literature-based tree might reflect better the community consensus about the evolutionary history of these eukaryotes, since it is consistent with the *Open Tree of Life* [85] and with the species phylogenies reported with the complete genomes. However, the backbone of the tree of eukaryotes is still subject to deep rearrangements and competing hypothesis [84, 85]. Therefore, we can never exclude the possibility that a suggested tree is free of errors nor the existence of a better phylogenetic representation. Here, we have demonstrated that our phylogenetic controlled analyses are strongly robust to: (a) uncertainties about ancestral branches, such as Parahoxozoa in Metazoa [121, 122] or Excavata in Eukarya [84]; (b) discrepancies about the phylogenetic position of particular clades among their "peers", for instance, Microsporia within Fungi [50] and Rhizaria within SAR [84]; (c) uncertainties in poorly resolved branches, such as Arthropoda and Rhizaria [123, 124]; as well as (d) common tree reconstruction problems, such as the presence of hard polytomies and long branch attraction of species.

In fact, the most significant discrepancies of our regressions from previous estimations over a large evolutionary scale are caused by the absence of correction for phylogenetic signal, rather than by the accuracy of the topological information used to account for it. As also shown by Whitney *et al.* [67] and Wu and Hurst [68], uncorrected phylogenetic dependencies among species (which assumes a star polytomy) lead to a much stronger correlation signal, as those strong correlations reported by previous studies [62, 65, 66]. We further showed that the correlation signal can also be affected considerably under controlled phylogenetic analyses, as some of those correlative associations estimated recently [64, 68]. Some of the factors analyzed here that lead to such biases include: low phylogenetic diversity, very small species datasets (<100 species) attempting to represent the current diversity of the sequenced eukaryotes, and the lack of systematic estimations of genome traits. These problems are particularly found at the clade-specific level (as also reported in [76, 81]), and in biased estimations of intron fea-

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tures (*e.g.*, density and genomic content) that already incorporate changes of genome size or on the number of protein-coding genes (see Appendix 2). We also demonstrated that our results are robust, not only to a reduced number of species (as long as the phylogenetic diversity of the dataset is maintained), but also to different sources of genome size estimates.

Yet, two additional factors might challenge the findings of this study. First, a limited access to both larger genomes (>5 Gbps) and genomes from deep branches of the eukaryotic tree precludes us from evaluating eventual biases on phylogenetic and "C-value" diversity so far. Second, substantial errors in genome assembly and/or annotation of protein-coding genes might considerably affect estimations of genome-based features. For instance, under- and over-estimation of genomic features were recently found on the genome projects of *Branchiostoma floridae* and *Hydra magnipapillata*, respectively, after RNA-seq-based re-annotations were performed in seven holozoans [66]. It is unclear, however, which genome features are the most affected by these biases, because the method employed in such study cannot distinguish introns and exons of coding genes from non-coding genes, such as pseudogenes and ncRNAs. This distinction is relevant in our study, since coding genes have a different copy number and experience different selective pressure than non-coding genes [125–127]. Moreover, annotation of protein-coding genes is becoming increasingly reliable owing to the incorporation of unbiased RNA-seq data, as most of our annotations are supported by. Through the filtering approaches of GenomeContent, we further found that very small introns and exons ($\leq$15 nts) represent $<$ 1% from the total number located in the coding genes in all genomes. Likewise, annotations of coding genes exhibiting intron sizes with an *excess* or *deficit modulo* 3 (*i.e.*, coding regions probably mistaken as introns and *vice versa*, respectively) are also unlikely to occur in our dataset (see Supplementary Figure S1 and Table S3). We acknowledge that particular phylogenetic patterns might be challenged by the completeness of genome assemblies in those lineages with very few sequenced genomes [66] or that will undergo substantial genome size corrections [128]. However, we also demonstrated to a considerable extent that possible biases in the genome assemblies and annotations analyzed here do not significantly impact our correlations and overall findings, after testing the genome completeness of 461 projects over 200 randomly reduced datasets. In disagreement with [66], thus, we show that the genomic differences obtained from most genome projects are robust enough to evaluate biological and evolutionary large-scale patterns of genome features across eukaryotes.

Intron-richness is a suitable pre-condition to evolve complex multicellularity

Despite the many origins of multicellularity on Earth, complex multicellularity (CM) evolved only a few times in Eukarya [129–131]. We developed here a conceptual framework to define CM beyond the number of unique cell types (UCTs) (see Appendix 1). Our definition follows West and colleagues [132, 133] in acknowledging that contingent irreversibility from clonal-unitary development is key to evolve obligately multicellularity, but differs in arguing that a reinforced irreversibility of developmental commitment from multiple cell types is the main determinant in clonal multicellularity for a major transition to occur in individuality. By formally differentiating simple and CM, we thus hypothesize that CM is the outcome of *major evolutionary transitions* [133, 134] involving the presence of innovatory changes in genome structure and expression due to the differential evolution of particular ncDNA classes along the eukaryotic lineages [131]. We found here that complex multicellular organisms (CMOs) are characterized by high intron-richness; even those CMOs that have undergone strong selection to reduce several classes of ncDNA and genome size, such as carnivorous plants [135] and birds [76, 88, 89]. We show that the association between CMOs and intron-richness does not depends on changes of genome size, which in is agreement with the study of Niklas [136] indicating that increases in the number of UCTs fail to keep pace with increases in genome size. Our findings also suggest that CM origins were most likely preceded by high intron-richness, since the latter is also found on the closest unicellular and simple multicellular relatives of CM lineages. This is consistent with episodes of rapid and extensive intron gain found on the basal lineages of opisthokonts, holozoans and plants [1, 13, 137]. Furthermore, the diversity of intron-richness observed here among CMOs is not random nor homogeneous. Instead, it appears to be constrained by different factors that demand further research, including shared phylogenetic history, widely divergent selective regimes, lifestyles and generation times.

It is becoming clearer that intron-richness has important phenotypic consequences on eukaryotes, but which of these consequences can be considered indispensable to promote CM convergently? As summarized earlier, the functions –either causal roles or selected-effects– of introns promoting the emergence and evolution of CM can be very diverse. Most of these functions are, however, neither exclusive of CMOs nor universal across eukaryotes, but rather the outcome of exaptations originated on independent occasions [138]. This is partially because, as shown in this study, introns possess different characteristics throughout the major supergroups. Also, the rates of intron conservation and the molecular mechanisms responsible for intron processing vary considerably across eukaryotes [2]. Most importantly, introns appear to affect virtually every step of mRNA maturation, as described previously and reviewed in [138]. Yet, the role of exon skipping (ES) has been highlighted as the main promoter of multicellular complexity by expansion of proteome diversity [2, 59] through selection of, for instance, new transcription factor families, cell adhesion and signal transduction proteins [98, 139, 140]. However, most ES events and isoforms (mainly in low abundance) are found to be mainly the outcome of stochastic splicing errors [141, 142]. Also, transcriptome analyses from diverse tissues and cell lines reveal that most genes express one and the same dominant transcript in multiple tissues in human [143], mouse [144] and fly [145] (but see [146, 147]). It remains thus to be fully understood to what extent ES events are actually contributing to the suggested protein diversity of CMOs [59, 131].

The phenotypic diversity of CMOs largely relies on the expression of ancestral and species-specific genes coordinated in a particular spatiotemporal manner. We argue here that intron-richness has facilitated this process to a great extent by tuning the transcriptomes of an organism through intron-mediated mechanisms (IMMs) that alter the timing or kinetics of transcript expression. For instance, *intron retention* coupled to components of the RNA surveillance machinery can modulate gene

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expression post-transcriptionally by slowing splicing kinetics of those intron-containing transcripts stored at the nucleus in response to a variety of cellular signals [138, 148, 149]. Diverse modes of transcript regulation by intron retention have been found during cell type differentiation, cellular stress, circadian rhythm and early embryogenesis in plants [147, 150–152] and mammals [58, 153–156]. Alternatively, *intron delay* can coordinate the expression of genes that are sensitive to changes in their transcript length during particular stages of the metazoan development [33, 34]. The transcriptional delay caused by the presence of introns in the *Hes7* locus, for instance, controls the proper oscillatory expression of the genes involved in body segmentation during early vertebrate embryogenesis [36, 157–159]. Notably, the intron lengths of *Hes7* and of other genes involved in developmental patterning across mammals are highly conserved and even coevolving among coexpressed genes [160, 161]. These findings are consistent with studies showing that some species with long-complex life cycles and slowly regulated cell-cycle genes appear to be enriched in intron-rich gene-structures and patterning processes [33, 34, 148]. On the other hand, short life cycles and rapidly regulated cell-cycle genes (at different stages of development) tend to constrain gene-structures toward short and intron-poor genes for efficient expression [7, 26, 34, 35, 37, 69, 162, 163].

Rather than a selective (ultimate) effect, a major influence of IMMs to differentiate functional from nonsense transcripts in CMOs is expected if we consider that: (i) intronic RNAs constitute a considerable fraction of the transcriptomes in the CMOs analyzed [28–30]; (ii) canonical and non-canonical splicing errors (from which these IMMs emerge) appear to be more frequent when intron-richness increases [49, 58, 61, 141, 142]; and (iii) the spatiotemporal patterns of transcript expression derived from IMMs do have significant ecological and evolutionary consequences for cell cycle control and body plan formation [33, 34, 148, 149]. Nevertheless, some of the phenotypic consequences of intron-richness are also expected in certain life histories that did not evolve CM due to different evolutionary conditions. For instance, IMMs appear to influence the development of complex life cycles in intron-rich unicellular and simple multicellular species (as defined in Appendix 1) such as Apicomplexan parasites, which often involve multiple hosts and/or differentiation stages [96, 164]. Similar findings are expected to be found in the upcoming complete genomes of other CMOs within the red algae and *coenocytes* such as *C. taxifolia* [165]. Ultimately, major evolutionary transitions require extreme conditions for certain factors to become consistently important [133]. We argue that high intron-richness (through ES and IMMs processes) has laid the foundations for the emergence of novel mechanisms of transcriptome timing in Eukarya, which under exceptional conditions might have convergently co-opted for fate specification and commitment of cell types. It remains to be known whether common life-history traits [166, 167], molecular mechanisms [168, 169] and evolutionary forces have parallelly shaped the evolution of both intron-richness and tissue-based body plans in Eukarya.

METHODS AND MATERIALS

Genome-based data collections

We compiled the complete-sequenced nuclear genomes, protein-coding genes and gene annotation files for a total of 461 eukaryotic organisms from publicly available databases: 131 fungi, 78 species from Archaeplastida, 186 from Metazoa, 20 from Alveolata, 17 from Stramenopiles, 7 from Excavata, 7 from Amoebozoa, 4 from Choanozoa, 3 from Rhizaria, *Fonticula alba* (Fonticulidae), *Guillardia theta* (Cryptophyta), *Emiliania huxleyi* (Haptophyta), *Thecamonas trahens* (Apusozoa). A manual filter was applied to avoid redundant species (*i.e.*, same genus with similar genome sizes), sequenced genomes with <70% of the estimated genome size, and gene annotations without support from transcript data. To account for significant under and overestimations of genome contents, we also corrected our calculations with the "estimated" genome sizes based on experimental approaches collected from databases and literature. References and details of these datasets are provided in Supplementary Table S1.

Estimating intron features with `GenomeContent.pl`

As depicted in Appendix 2-Figure 1, `GenomeContent` was written in `Perl` to calculate global statistics and sequence-based estimators of genome features through six major steps: (1) identifying coordinates from protein-coding gene (CDS); (2) checking the quality of intron annotations; (3) calculating statistic descriptors for genome-wide features from "reference gene sets" (derived from 1 and 2), such as size, density and number; (4) estimating genome-feature contents from the overlapping projection of "reference gene sets" onto the genome sequence; (5) calculating statistic descriptors for genome-feature contents; (6) plotting of figures and retrieving of sequences (fasta format) and statistics (text format). According to the definitions described in Appendix 2, we measured 30 genome features with `GenomeContent` across 461 eukaryotes, including 10 associated to the intron-richness of a genome: intron size (average and "weighted", see equation 1 in Appendix 2), total number, absolute density, genomic content, number and fraction of genes containing introns. As observed in Appendix 2-Table 1, "(absolute) intron density" was measured as the average number of introns per intron-bearing gene because other estimations based on the average number of introns either per sequence region or from the total number of protein-coding genes are vulnerable to both genome size and considerable fluctuations of gene models, respectively.

Determining the repeat content of genome features and statistical tests

We focused on identifying *de novo* repeats along the genome, rather than on classifying them in specific families. The `Repeatscout` algorithm v1.0.5 [170] was employed to compare a genome sequence against itself and in the two reading directions in order to identify *de novo* repetitive sequences with the minimal *k*-mer length of 15 nts. To gain major repeat coverage, the *de novo* `Repeatscout` libraries were merged with the Repbase libraries version 20.03 (http://www.girinst.org/repbase/). Then, these merged repeat libraries were used to map the coordinates of the repeats (interspersed and short repeats, and low complexity sequences) across the complete genomes with the `RepeatMasker` program v4.0.5 (www.repeatmasker.org). The placeholders were also taken into account as an independent category of "potential undefined repeats" within each genome feature (see Figure 2). However, the proportion of such regions in a genome is not collected in an individual genome feature for

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

statistical analysis, given that they usually include the highly repetitive heterochromatic sequences that could not be unambiguously sequenced. Based on the genome coordinates obtained from this process, a nucleotide was classified as repetitive if it was covered by a repetitive sequence on either strand. Accordingly, the **repeat content** within introns, exons and the genome was determined as the total number (or percentage) of nucleotides in the genome feature that were classified as "repeat"; while the **unique content** of a particular genome feature is thus estimated from the non-repetitive nucleotides.

With a custom Perl script, we then calculated the *Jaccard index*, as $J(R, GF) = |R \cap GF| / |R \cup GF|$, to estimate the nucleotide overlap between the repeat coordinates (R) and the genome features (GFs: exons and introns) by using the "feature content coordinates" of every genome as observed in the Appendix 2-Figure 1. We tested the significant degree of observed overlap between repeats and GFs for each genome with the GenometriCorr package [171] from the R program v3.1.2 (www.r-project.org). For each genome, 1,000 permutations were allowed to shuffle the repeat coordinates along the genome sequence. Exon and intron positions were preserved as reference, while the repeat coordinates and the random sets were provided as the query. This setting provides the correct assessment of correlations (*p-value*) for the relative distance under the Kolmogorov-Smirnov criteria and for the permutation test on the *Jaccard index* [171, 172], which indicate whether the overlap is less (TRUE) or more (FALSE) significant than expected by chance. The statistical reports are available as Supplementary Data.

Construction of phylogenetic trees

We constructed four different tree topologies, each of which has been extensively used in literature, for the 461 eukaryotes analyzed in this study, in order to: (a) evaluate phylogenetic signal, (b) perform phylogenetic controlled analyses, and (c) test the robustness of our comparative analyses against some of the phylogenetic uncertainty surrounding the tree of eukaryotes, owing to phylogeny construction errors and the absence of the true but unknown species tree.

To test for sensitivity to topological and branch length errors, we constructed a phylogeny determined by protein domain content as described in [173] (Figure 1d). Accordingly, *hmmscan* from HMMER v3.1b1 [174] was used to search for the protein domain models of the Pfam-A families from the PFAM database v30.0 [175] against the 461 proteomes. We used the gathering threshold (-cut-ga) for filtering out false positives. Custom Perl scripts were used to obtain a presence/absence matrix for all pfam domains detected and to calculate a pairwise distance matrix for all analyzed species. A weighting factor was included to correct the distance between two genomes (owing to the great differences in genome size, gene content, and lifestyles), and according to the following relationship: $D = A'/(A' + AB)$, where A' is the number of unique pfam domains in one of two genomes compared: A and B, and AB is the number of pfam domains they share. Thus, "the two tendencies are acknowledged by setting the evolutionary distance equal to the ratio of the unique domains in the smaller genome (A') to its total number of domains ($A' + AB$)" [173]. The phylogeny construction was performed with the neighbor-joining method and bootstrapping with the program *neighbor* from PHYLIP v3.68 (http://evolution.genetics.washington.edu/phylip.html). We used *Trypanosoma brucei* as an outgroup, based on the supported basal phylogenetic position of Eozoa (Excavata and Euglenozoa) within Eukarya [176, 177].

Additionally, two NCBI taxonomy-based trees with no branch lengths were obtained with the species IDs collected from the "Taxonomy Browser" of NCBI (https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/, last accessed July 14, 2016, see Supplementary Table S1) and with a combination of the phyloT (www.phylot.biobyte.de) and iTOL (www.itol.embl.de/itol.cgi) tools to allow or not the use of polytomies on each tree. A fourth tree topology (unrooted and with no branch lengths) was obtained by manually "correcting" the NCBI taxonomy-based tree (not polytomies allowed) with the TreeGraph 2 v2.13.0-748 beta program [178] to fix the resolution at the genus and species level according to: (1) clade-specific phylogenies based on candidate orthologous sequences reported on literature, and (2) supertrees reconstruction for those few species that have not been incorporated into a sequence-based phylogeny yet. Polytomies were introduced in cases where phylogenetic uncertainty is not solved according to different studies. The 110 references employed to construct this consensus tree, some of which include the phylogenies reported along with the complete genome projects, are available on Supplementary Table S1. The literature consensus-based tree was selected as the "reference eukaryote tree" (Figure 1a) to present the results throughout the article. All tree topologies are available as Supplementary Data.

We quantitatively estimated the dissimilarity among the four trees with two measures reported to performed best among topology-only metrics [179]. The symmetric difference of Robinson and Foulds (RF) measures the number of different partitions (or clades not shared) between two trees [180], whereas the tree aligment metric (Align) of [181] scores the mismatches in the best alignment of the similar (and same) branches between two trees. Absolute RF distances were also divided by the total number of species in the tree to estimate the number of partitions per species. The Align and RF scores were calculated with the python scripts implemented in [179] and available at http://datadryad.org/resource/doi:10.5061/dryad.g9089.

Phylogenetically corrected analyses

We first evaluated the strength of phylogenetic signal (*i.e.*, their statistical non-independence) exhibited by the 30 genome-based features analyzed in this study with the tree topologies described previously. Thus, we calculated the Pagel's lambda (λ) transformation for all genome features analyzed with the *caper* R package (pgls) [182]. In comparison to other indices, Pagel's λ is very robust to both incompletely resolved phylogenies and suboptimal branch-length information [83, 183]. We then calculated coefficients of determination (r^2) to estimate the strength of the correlation to associate the variations observed between two traits (X and Y) with three linear regression models: Ordinary Least Squares (OLS) (*stats* R package: lm), Phylogenetically Independent Contrasts (PICs) (*ape* R package: pic) [184], and Phylogenetic Generalized Least Squares (PGLS) (*caper* R package: pgls). We also calculated log Bayes Factors (*LogBF*) to estimate the significance of evidence for the correlation between X and Y with the Markov chain Monte Carlo (MCMC) method and the PIC and PGLS models. log Bayes Factors were calculated as:

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$LogBF = 2(log\ [\text{marginal likelihood (complex model) - log marginal likelihood (simple model)}])$$

with 100 stones and 10,000 iterations per stone to estimate the marginal likelihood, as implemented in the BayesTraits program v3 (http://www.evolution.rdg.ac.uk/BayesTraitsV3/BayesTraitsV3.html).

We tested the robustness of the phylogenetically controlled regressions to discrepancies in tree topologies, phylogenetic diversity and estimations of genome features. First, OLS, PGLS and PIC regressions were performed with two sources for genome size: genome assemblies and experimental estimations. Also, the PGLS and PIC regressions were performed with the four tree topologies described previously. For the protein domain-based phylogeny, we also performed PGLS regressions with both equivalent branch lengths ($all = 1$) and lengths derived from the distance matrix. To test the influence of not fully resolved trees owing to the presence of hard polytomies, we generated three additional topologies for the NCBI-taxonomy tree (with polytomies allowed), two of them with randomly resolved polytomies using the procedure multi2di=TRUE, and one tree with a non-random procedure multi2di=FALSE, with the R package *ape* (library picante). Furthermore, we employed the "replicated co-distribution" approach [87] to test whether the association between X and Y is replicated across multiple independent clades. Thus, PGLS regressions were performed over 20 different sets compiled from the original dataset of 461 eukaryotes: 18 datasets correspond to divergent lineages, and two datasets (with 100 replicates each) were created through the random selection of 231 and 116 species, respectively. To further test the impact of phylogenetic diversity on the sensitivity of our phylogenetically controlled correlations, we performed PGLS regressions with four additional datasets (see Supplementary Table S12): two further randomly reduced datasets of 58 and 29 species, another dataset with 26 out of the 30 eukaryotes used in the study by Lynch and Conery [62], and the dataset of 30 eukaryotic genomes from the study of Wu and Hurst [68]. All genome-feature values were log10- transformed prior to analysis, except for the few genome-feature values == 0 that estimated the absence of repeats within introns in some extreme intron-poor genomes such as *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Giardia intestinalis*, *Spironucleus salmonicida*. These particular values were discarded from the corresponding analyses.

We computed the Cronbach's reliability coefficient alpha to measure the internal consistency (inter-relatedness) of the variables employed to test their relationship with multicellular complexity with the R package *psych* (alpha) [185]. Comparative and phylogenetic comparative Principal Component Analyses (PCA) were performed using the R packages *stats* (princomp) and *phytools* (phyl.pca), respectively. Branch transformations ($all = 1$) for the phylogenetically controlled PCAs and PGLS regressions were performed with *ape* (compute.brlen). Remaining statistical tests were calculated with custom R scripts. The plots were prepared with the R packages *ggplot*, *ggbiplot*, phenotypicForest v0.2 (http://chrisladroue.com/phorest/) and the software Inkscape (https://inkscape.org/en/). The results displayed throughout the paper are based on the estimated genome sizes and the "reference eukaryote tree" (Figure 1a). The results obtained with alternative tree topologies and assembled genome sizes are available in Supplementary Material.

AVAILABILITY

The GenomeContent program and data are available upon request during peer-review, and will be openly available after publication.

SUPPLEMENTARY DATA

Supplementary Data, Text, Figures and Tables are only available for peer-review at the moment, but will be openly available after publication.

ACKNOWLEDGMENTS

We gratefully acknowledge the work and policies of the consortia and research groups for making publicly available the current wealth of genome datasets. We thank Jens Steuck and Ronald Kriemann for their assistance with the IZBI and MPI clusters required for computational work, A. Nabor Lozada-Chávez for his assistance with the R package, and Cristina Bojórquez-Espinosa for proofreading. We deeply appreciate the feedback received from previous anonymous reviewers.

FUNDING

National Council for Science and Technology in Mexico (CONACyT: 185993) and Max Planck Institute for Mathematics in the Sciences to I.L-C.

CONFLICT OF INTEREST STATEMENT.

No competing interests declared.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

APPENDIX 1

Defining simple *versus* complex multicellularity

Multicellularity refers to the phenotype characterized by the self-organization of cells that undergo a transition in individuality to perform cooperative consumption of energy, survival, and ultimately, reproduction. Multicellularity has arisen multiple times during the evolution of life on Earth [129], and it can even be induced in experimental settings [186, 187]. However, multicellularity is hypothesized to unfold into two different transitions [130, 131, 167]: *simple* or *complex*. Complex multicellularity (CM) is restricted to Eukarya and has evolved independently a few times: florideophyte red algae, laminarian brown algae, viridiplantae, eumetazoan animals, basidiomycota and ascomycota fungi. Typically, the number of unique cell types (UCTs) is used as the defining feature of multicellular complexity. However, accurate estimates of UCTs are only available for a small fraction of the species, and they also fail to appropriately capture the complexity of multicellular species [136]. Thus, we have created three "working definitions" embracing four criteria that distinguish few distinctive aspects of cellular development and life cycle to recognize species in our dataset, first as unicellular or multicellular (criterion 1), and then as simple multicellular (SMO) or complex multicellular (CMO) (criteria 2-4):

Criterion 1. *Whether there are one or several differentiated state cells at once across a life cycle.* Some single-celled organisms may have several differentiated state cells but at different times during the life cycle, such as the yeast *Saccharomyces cerevisiae* (three UCTs) [188], or may develop a *pseudo-hypha* with very few spores during a transient stage of a life cycle to reproduce [189]. Also some naturally living unicellular organisms, such as *S. cerevisiae* and *C. reinhardtii*, may develop filamentous growth under particular experimental settings that induce stress, and as long as the selective pressure is mantained [186, 187].

Criterion 2. *Whether the organization of the cells in a multicellular organism falls into one of the following states: a) differentiated cell types; b) undifferentiated cell types; or c) "syncytium/coenocyte" sensu amplio*, *i.e.*, multiple nuclei (either genetically identical or distinct) distributed within one common cytoplasm, which might or not be partially separated by cell membranes. Examples of the latter include: coenocytic *Dictyostelium discoideum* [190], siphonous algae and non-septate fungi [191]. Some coenocytic organisms, such as *Caulerpa taxifolia* [165], have morphological structures equivalent to a multicellular organ but not comprised in tissues or cells (*i.e., pseudo-organs*). While SMOs undergo a transition in individuality through any of the three cellular states, CMOs only undergo a major transition in individuality through differentiated cell types.

Criterion 3. *Whether the transition in individuality, as defined in criterion 2, is facultatively or obligately replicated across generations.* Facultatively multicellular species are able to complete their life cycle as unicells and only become multicellular under certain environmental conditions [132, 133]. For example, formation of fruiting bodies in some organisms, such as *Dictyostelium*, is observed in particular generations that undergo critical conditions to increase dispersal success [192, 193]. By contrast, obligately multicellular species can only complete their life cycle as multicellular organisms, owing mainly to the high genetic relatedness of cells originated through clonal-unitary development [129, 132, 133]. For instance, the development of tissue-based fruiting bodies ("basidia" and "ascocarp") used for sexual reproduction in fungal CMOs is replicable on every generation. While SM is either facultatively or obligately replicated across generations, CM is only replicated as a whole. This criterion takes into account the temporal "unicellular transition" that all multicellular organisms, either simple or complex, undergo by means of reproductive processes through the life cycle [194].

Criterion 4. *Whether or not the transition in individuality, as defined in criterion 2, is produced by irreversible tissue-based body plans*. CMOs have tissue-based body plans that are developmentally irreversible, so that the within-group conflicts produced by "mutant-selfish" cell lineages (*defectors*) are negligible enough to avoid reversible differentiation of the whole organism [195–197]. Such irreversibility is consequence of active developmental commitment of multiple cell types that undergo fate specification and determination at particular stages of an organism life cycle. Cell type commitment is observed during the formation of: i) germ layers in metazoans [168] and eumetazoans such as cnidarians[198–200], ii) meristems in plants [201, 202] and in the CMOs within brown and red algae [203–205], and iii) in the primordium of fungal CMOs, although some cell types are still able to revert to vegetative growth *in vitro* [100, 206]. By contrast, SMOs do not develop true tissue-based body plans, owing in part to the lack of cell types with fixed identities and lineage commitment, so that dedifferentiation or transdifferentiation of cell types at any stage of development is common under the influence of certain factors. For instance, an absence of true tissue-based body plans, fate determination and stability of key cell types is observed in the sea sponge *A. queenslandica* (∼11 UCTs) [207, 208], and the sea placozoan *T. adhaerens* (∼5 UTCs) [209, 210] (but see [211, 212]). Likewise, the green algae *V. carteri* also lacks of a tissue-based body plan, since it only forms a colony of ∼2,000 cells with two UCTs [213].

According to these four criteria, we distinguish:

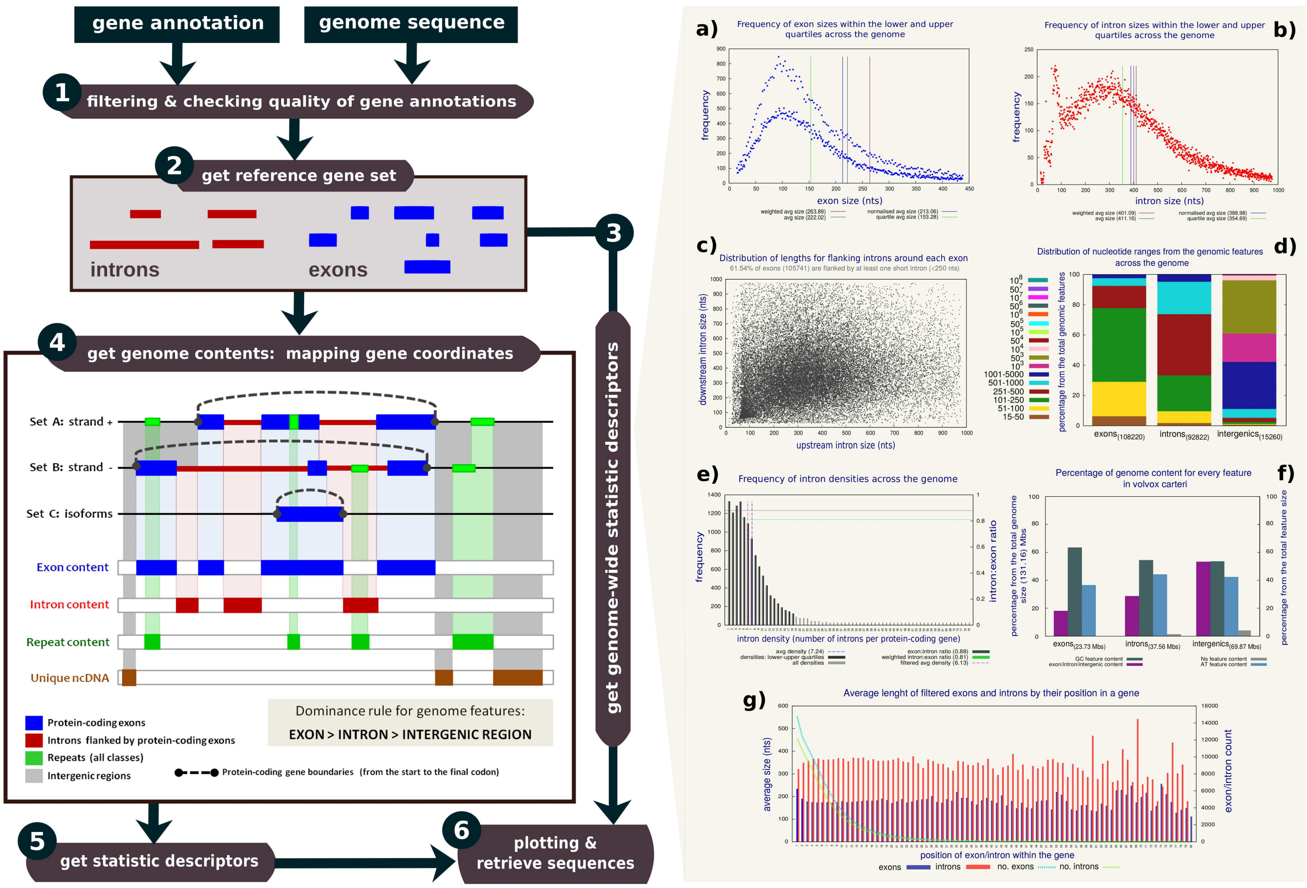
Unicelullar: is an organism exhibiting a single differentiated state cell at once across its life cycle. Single-celled organisms developing a transient *pseudo-hypha* or experimentally-driven filamentous growth are also included in this category.

Simple multicellular: is an organism exhibiting a facultatively or obligately transition in individuality through the organization of either several cells (with none or only few differentiated cell types) or a syncytium/coenocyte *sensu amplio* originated from one or more cell-line ancestors. Coenocytic and siphonous organisms structurated in *pseudo-organs* are also included in this category. Reversion to unicellularity may occur.

Complex multicellular: is an organism exhibiting an irreversible transition in individuality produced by tissue-based body plans, through the developmental commitment of multiple and different cell types originated from a common cell-line ancestor. Reversion to unicellularity or to a simple multicellularity lifestyle does not occur.

The lifestyle and body plan development of all species in our dataset were compiled from literature to evaluate the four criteria of our definitions, such information is available in Supplementary Table S2. We classified 77 species as unicellular, 96 species as SMOs, and 288 species as CMOs. This approach was useful to define the cellular state of some controversial model organisms. However, it still represents a challenge to distinguish between SM and CM in species within Fungi and Parazoa,

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.


gene annotation
genome sequence
1 filtering & checking quality of gene annotations
2 get reference gene set
introns
exons
3 get genome-wide statistic descriptors
4 get genome contents: mapping gene coordinates
Set A: strand +
Set B: strand -
Set C: isoforms
Exon content
Intron content
Repeat content
Unique ncDNA
Dominance rule for genome features:
EXON > INTRON > INTERGENIC REGION
Protein-coding exons
Introns flanked by protein-coding exons
Repeats (all classes)
Intergenic regions
Protein-coding gene boundaries (from the start to the final codon)
5 get statistic descriptors
6 plotting & retrieve sequences
a) Frequency of exon sizes within the lower and upper quartiles across the genome
b) Frequency of intron sizes within the lower and upper quartiles across the genome
c) Distribution of lengths for flanking introns around each exon
d) Distribution of nucleotide ranges from the genomic features across the genome
e) Frequency of intron densities across the genome
f) Percentage of genome content for every feature in volvox carteri
g) Average lenght of filtered exons and introns by their position in a gene
frequency
exon size (nts)
intron size (nts)
upstream intron size (nts)
downstream intron size (nts)
intron density (number of introns per protein-coding gene)
intron:exon ratio
position of exon/intron within the gene
average size (nts)
exon/intron count

Appendix 2 Figure 1. The `GenomeContent` **program.** A flowchart of the program is displayed on the left and described throughout Appendix 2. On the right, examples of exploratory figures showing some statistic descriptors for the genome of *V. carteri*, as obtained with the program.

whose multicellular body plans are generated from a few UCTs or are not well documented yet to evaluate the criteria described previously. Since this is a pioneer attempt to formally define CM and 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.

APPENDIX 2

A. Estimating intron features with `GenomeContent.pl`

`GenomeContent` was written in `Perl` to calculate global statistics and sequence-based estimators of genome features in six major steps, as shown in Appendix 2-Figure 1. First, the processing of gene annotations focuses on identifying coordinates from protein-coding gene (CDS), while the filtering process focuses on checking the quality of intron annotations. As described in next sections, the coordinates derived from both proceses are taken as the "reference gene sets" for introns, exons and intergenic regions to directly estimate several statistic descriptors, such as size, density and number. Then, the "reference gene sets" are projected onto the genome sequence in both strands, so that the nucleotide contents of each genome-feature are calculated according to the definitions described in a section below. Finally, all statistic descriptors obtained with the program are provided as text files, fasta formats and exploratory figures (see also Supplementary Figure S1). `GenomeContent` runs on an entire genome in few minutes or hours, depending on genome size and the number of annotated genes. `GenomeContent` is available upon request during peer-review, and will be openly available after publication.

B. Filtering of gene annotations

The filtering process of `GenomeContent` involves: a) identification of CDS, b) removal of small sizes, c) treatment of isoforms, and d) estimation of systematic errors in CDS. First, only genome coordinates from CDS were extracted, but their corresponding untranslated regions (UTRs) are not included because these are not fully annotated in most genome projects [214, 215]. Second, we excluded introns and exons smaller than 15 nucleotides (nts), which represent $< 1\%$ from the total number of introns and exons located within the coding genes of all genomes analyzed (as observed in Supplementary Figure S2 and Table S3). Third, alternative splice variants were kept in the data. To avoid redundant/overestimated data, however, 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 call this filtered set of protein-coding gene coordinates for every genome as the "reference gene set".

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 2 Table 1. Comparison of intron density estimations: absolute (aID) and normalized (nID).

Species name	clade	aID	nID	# CDS	% CDS	# introns	genome size
Tupaia chinensis	Mammalia	6.98	5.07	22,688	72.53 %	114,940	3,200.0 Mbs
Caenorhabditis elegans	Nematoda	5.28	5.12	20,520	96.81 %	104,966	92.9 Mbs
Gossypium raimondii	Malvales	6.00	5.20	77,267	86.86 %	402,768	880.0 Mbs
Xenopus laevis	Amphibia	7.41	5.25	43,025	70.81 %	225,747	3,110.0 Mbs
Guillardia theta	Cryptophyta	6.74	5.28	24,840	78.29 %	131,044	87.2 Mbs
Postia placenta	Basidiomycota	5.94	5.54	17,173	93.22 %	95,089	90.9 Mbs
Bombus terrestris	Arthropoda	6.25	5.99	8,334	95.80 %	49,907	274.0 Mbs

Most of the collected gene annotations are based on transcript evidence. Nevertheless, we implemented the approach proposed by Roy and Penny [216] in GenomeContent to further estimate systematic errors in CDS 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 in [216], large values of "3n excess", $E_3 = p_{3n} - (p_{3n+1} + p_{3n+2})/2$, suggest that a considerable fraction of internal exons may have been incorrectly predicted as introns or that there are several "intron retention" events. On the other hand, a deficit of 3_n introns, i.e., $E_3 \ll 0$, suggests that a considerable fraction of 3_n introns –lacking of stop codons– may have been mistaken for exons. Most gene annotations included in this study shown values of the 3n excess close to 0. Very few genomes (such as parasites and endosymbiots) were initially excluded from the present study because they exhibited high 3n excess $(0.4 - 0.7)$ (see Supplementary Figure S2 and Table S3).

C. Estimation of average sizes and density of introns

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. The average *feature size* (*e.g.*, *intron size* and *exon size*) of CDSs per genome was calculated in two ways. The **straight average size** ($A_{feature}$) 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 ($N_{feature}$): $A_{feature} = L_{feature}/N_{feature}$. The *straight average* 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.

If we now consider $a_{feature}$ to be the average length of the respective feature (exon or intron) within one single gene: $a_{feature} = l_{feature}/n_{feature}$, then the **weighted average size**, $\bar{a}_{feature}$, is calculated as the mean of the $a_{feature}$ values of the respective feature (exons or introns) in a genome, according to:

$$\bar{a}_{feature} = \frac{1}{n}\sum_{i=1}^{n} a_{feature} \qquad \textbf{(S1)}$$

where n represents the total number of CDSs in a genome when calculating $\bar{a}_{exon}$, or the total number of intron-containing CDSs in a genome when calculating $\bar{a}_{intron}$ [217]. The *weighted average* 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 well known skewed length distribution.

GenomeContent also estimates the abundance of introns within a genome with two different estimates to detect small changes of intron-richness and to buffer dramatic changes among updated genome releases and gene annotations. On the one hand, the abundance of introns across CDSs was estimated as the **fraction of intron-containing CDSs** from the total number of CDSs (% CDS), which might reflect complete intron loss from CDS structures at the genome level. On the other hand, the abundance of introns and exons within CDSs ("absolute density") was estimated as the mean number per genome of exons in CDS (**exon density**), or introns per intron-containing CDSs (**intron density**), respectively. We employ the "absolute density" because, as observed on Appendix 2-Table 1, the average number of introns either per sequence region or from the total number of genes ("normalized density") depends on both genome size and the number of CDSs, respectively.

For instance, the four species listed in Appendix 2-Table 1 exhibit around five introns per CDS when a "normalized" density is estimated from the total number of CDSs. However, the "absolute intron density" clearly shows that, for instance, *X. laevis*, *T. chinensis* and *G. theta* have indeed more introns per CDS on average than the other species, despite of having a smaller fraction of intron-containing CDSs (70.8%, 72.5% and 78.3%, respectively), and lower number of introns and CDS in some cases. Clearly, the bias observed in the "normalized intron density" is produced by larger numbers of total CDS in the genome.

D. Estimation of genome contents

GenomeContent also estimates the "feature content" of a given genome, *i.e.*, the proportion of nucleotides of the respective genome features (intron, exon, or intergenic) that contributes to genome size. Since most genome annotations only contain protein-coding regions rather than full transcript models, we count only coding exons and introns delimited by a pair of coding exons. As shown in Appendix 2-Figure 1, the program projects the sets of genomic intervals for all exons and introns from coding genes located in the plus strand (set A), the minus strand (set B), and of the isoforms (set C). Since a given nucleotide may be classified differently for different isoforms, we used the following domincance rule in order to obtain a unique classification at the genome level:

$$Exon > Intron > Intergenic\_region$$

It reflects the idea that a genomic position is exonic whenever it appears in a coding exon of at least one transcript. Thus, the **exon content** of a genome is calculated as the total number of nucleotides in the genome sequence that are classified as coding exon with respect to at least one isoform. Analogously, a position is classified as 'intronic' if it appears inside the boundaries of annotated coding exons, but it does not overlap with any coding sequence. Thus, **intron content** was determined as the total number of nucleotides of a genome that were classified as intronic. The **CDS content** of a genome is calculated

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

as the total number of nucleotides covered by the intronic and exonic positions within coding genes. Finally, the **non-coding DNA content** was computed analogously as the total number of nucleotides in a genome that are not covered by exonic and intronic positions from coding genes. Genome-feature contents are reported as: total size in Megabases (Mb), fraction (%) from the total genome size, and as genomic coordinates.

REFERENCES

1. M. Csuros, I. B. Rogozin, and E. V. Koonin, "A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes," PLoS Comput Biol **7**, e1002150 (2011).
2. M. Irimia and S. W. Roy, "Origin of spliceosomal introns and alternative splicing," Cold Spring Harb Perspect Biol **6**, a016071 (2014).
3. C. B. Nielsen, B. Friedman, B. Birren, C. B. Burge, and J. E. Galagan, "Patterns of intron gain and loss in fungi," PLoS Biol **2**, e422 (2004).
4. S. Cho, S. W. Jin, A. Cohen, and R. E. Ellis, "A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution." Genome Res **14**, 1207–20 (2004).
5. S. W. Roy and W. Gilbert, "Rates of intron loss and gain: implications for early eukaryotic evolution," Proc Natl Acad Sci U S A **102**, 5773–8 (2005).
6. J. A. Fawcett, P. Rouze, and Y. Van de Peer, "Higher intron loss rate in arabidopsis thaliana than a. lyrata is consistent with stronger selection for a smaller genome." Mol Biol Evol **29**, 849–59 (2012).
7. H. Wang, K. M. Devos, and J. L. Bennetzen, "Recurrent loss of specific introns during angiosperm evolution." PLoS Genet **10**, e1004843 (2014).
8. L. Carmel, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin, "Three distinct modes of intron dynamics in the evolution of eukaryotes." Genome Res **17**, 1034–44 (2007).
9. A. Z. Worden, J. H. Lee, T. Mock, P. Rouzé, M. P. Simmons, A. L. Aerts, A. E. Allen, M. L. Cuvelier, E. Derelle, M. V. Everett, E. Foulon, J. Grimwood, H. Gundlach, B. Henrissat, C. Napoli, S. M. McDonald, M. S. Parker, S. Rombauts, A. Salamov, P. Von Dassow, J. H. Badger, P. M. Coutinho, E. Demir, I. Dubchak, C. Gentemann, W. Eikrem, J. E. Gready, U. John, W. Lanier, E. A. Lindquist, S. Lucas, K. F. Mayer, H. Moreau, F. Not, R. Otillar, O. Panaud, J. Pangilinan, I. Paulsen, B. Piegu, A. Poliakov, S. Robbens, J. Schmutz, E. Toulza, T. Wyss, A. Zelensky, K. Zhou, E. V. Armbrust, D. Bhattacharya, U. W. Goodenough, Y. Van de Peer, and I. V. Grigoriev, "Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes micromonas." Science **324**, 268–72 (2009).
10. W. Li, A. E. Tucker, W. Sung, W. K. Thomas, and M. Lynch, "Extensive, recent intron gains in daphnia populations." Science **326**, 1260–2 (2009).
11. A. van der Burgt, E. Severing, P. J. de Wit, and J. Collemare, "Birth of new spliceosomal introns in fungi by multiplication of introner-like elements." Curr Biol **22**, 1260–5 (2012).
12. J. T. Huff, D. Zilberman, and S. W. Roy, "Mechanism for dna transposons to generate introns on genomic scales." Nature **538**, 533–536 (2016).
13. X. Grau-Bové, G. Torruella, S. Donachie, H. Suga, G. Leonard, T. A. Richards, and I. Ruiz-Trillo, "Dynamics of genomic innovation in the unicellular ancestry of animals." Elife **6** (2017).
14. J. V. Chamary and L. D. Hurst, "Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage," Mol Biol Evol **21**, 1014–23 (2004).
15. D. J. Gaffney and P. D. Keightley, "Genomic selective constraints in murid noncoding dna," PLoS Genet **2**, e204 (2006).
16. J. Ponjavic, C. P. Ponting, and G. Lunter, "Functionality or transcriptional noise? evidence for selection within long noncoding rnas," Genome Res **17**, 556–65 (2007).
17. S. G. Park, S. Hannenhalli, and S. S. Choi, "Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals." BMC Genomics **15**, 526 (2014).
18. S. A. Shabalina and A. S. Kondrashov, "Pattern of selective constraint in c. elegans and c. briggsae genomes," Genet Res **74**, 23–30 (1999).
19. P. Andolfatto, "Adaptive evolution of non-coding dna in drosophila," Nature **437**, 1149–52 (2005).
20. D. A. Skelly, J. Ronald, C. F. Connelly, and J. M. Akey, "Population genomics of intron splicing in 38 saccharomyces cerevisiae genome sequences," Genome Biol Evol **1**, 466–78 (2009).
21. R. Schmitt, "Differentiation of germinal and somatic cells in volvox carteri." Curr Opin Microbiol **6**, 608–13 (2003).
22. X. Guo, Y. Wang, P. D. Keightley, and L. Fan, "Patterns of selective constraints in noncoding dna of rice," BMC Evol Biol **7**, 208 (2007).
23. A. Haudry, A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, Z. Joly-Lopez, J. G. Steffen, K. M. Hazzouri *et al.*, "An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions," Nat Genet **45**, 891–8 (2013).
24. A. Audibert, D. Weil, and F. Dautry, "In vivo kinetics of mrna splicing and transport in mammalian cells," Mol Cell Biol **22**, 6706–18 (2002).
25. C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov, "Selection for short introns in highly expressed genes," Nat Genet **31**, 415–8 (2002).
26. D. C. Jeffares, C. J. Penkett, and J. Bahler, "Rapidly regulated genes are intron poor," Trends Genet **24**, 375–8 (2008).
27. A. Coulon, M. L. Ferguson, V. de Turris, M. Palangat, C. C. Chow, and D. R. Larson, "Kinetic competition during the transcription cycle results in stochastic rna processing," Elife **3**, e03939 (2014).
28. M. B. Clark, P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn, C. P. Ponting, P. F. Stadler, K. V. Morris, A. Morillon *et al.*, "The reality of pervasive transcription," PLoS Biol **9**, e1000625; discussion e1001102 (2011).
29. G. Liu, J. S. Mattick, and R. J. Taft, "A meta-analysis of the genomic and transcriptomic composition of complex life," Cell Cycle **12**, 2061–72 (2013).
30. G. St Laurent, D. Shtokalo, M. R. Tackett, Z. Yang, T. Eremina, C. Wahlestedt, S. Urcuqui-Inchima, B. Seilheimer, T. A. McCaffrey, and P. Kapranov, "Intronic rnas constitute the major fraction of the non-coding rna in mammalian cells," BMC Genomics **13**, 504 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

31. M. B. Ardehali and J. T. Lis, "Tracking rates of transcription and splicing in vivo," Nat Struct Mol Biol **16**, 1123–4 (2009).
32. C. N. Tennyson, H. J. Klamut, and R. G. Worton, "The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced," Nat Genet **9**, 184–90 (1995).
33. I. A. Swinburne and P. A. Silver, "Intron delays and transcriptional timing during development," Dev Cell **14**, 324–30 (2008).
34. P. Heyn, A. T. Kalinka, P. Tomancak, and K. M. Neugebauer, "Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences," Bioessays **37**, 148–54 (2015).
35. C. G. Artieri and H. B. Fraser, "Transcript length mediates developmental timing of gene expression across drosophila," Mol Biol Evol **31**, 2879–89 (2014).
36. Y. Takashima, T. Ohtsuka, A. Gonzalez, H. Miyachi, and R. Kageyama, "Intronic delay is essential for oscillatory expression in the segmentation clock," Proc Natl Acad Sci U S A **108**, 3300–5 (2011).
37. J. L. Woody and R. C. Shoemaker, "Gene expression: sizing it all up," Front Genet **2**, 70 (2011).
38. A. O. Urrutia and L. D. Hurst, "The signature of selection mediated by expression on human genes," Genome Res **13**, 2260–4 (2003).
39. U. Pozzoli, G. Menozzi, G. P. Comi, R. Cagliani, N. Bresolin, and M. Sironi, "Intron size in mammals: complexity comes to terms with economy," Trends Genet **23**, 20–4 (2007).
40. K. Juneau, M. Miranda, M. E. Hillenmeyer, C. Nislow, and R. W. Davis, "Introns regulate rna and protein abundance in yeast," Genetics **174**, 511–8 (2006).
41. X. Y. Ren, O. Vorst, M. W. Fiers, W. J. Stiekema, and J. P. Nap, "In plants, highly expressed genes are the least compact," Trends Genet **22**, 528–32 (2006).
42. W. Lanier, A. Moustafa, D. Bhattacharya, and J. M. Comeron, "Est analysis of ostreococcus lucimarinus, the most compact eukaryotic genome, shows an excess of introns in highly expressed genes," PLoS One **3**, e2171 (2008).
43. S. A. Shabalina, A. Y. Ogurtsov, A. N. Spiridonov, P. S. Novichkov, N. A. Spiridonov, and E. V. Koonin, "Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes." Mol Biol Evol **27**, 1745–9 (2010).
44. J. Collén, B. Porcel, W. Carre, S. G. Ball, C. Chaparro, T. Tonon, T. Barbeyron, G. Michel, B. Noel, K. Valentin *et al.*, "Genome structure and metabolic features in the red seaweed chondrus crispus shed light on evolution of the archaeplastida," Proc Natl Acad Sci U S A **110**, 5247–52 (2013).
45. S. Schwartz, E. Meshorer, and G. Ast, "Chromatin organization marks exon-intron structure," Nat Struct Mol Biol **16**, 990–5 (2009).
46. L. De Conti, M. Baralle, and E. Buratti, "Exon and intron definition in pre-mrna splicing," Wiley Interdiscip Rev RNA **4**, 49–60 (2013).
47. K. L. Fox-Walsh, Y. Dou, B. J. Lam, S. P. Hung, P. F. Baldi, and K. J. Hertel, "The architecture of pre-mrnas affects mechanisms of splice-site pairing," Proc Natl Acad Sci U S A **102**, 16176–81 (2005).
48. C. N. Dewey, I. B. Rogozin, and E. V. Koonin, "Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns," BMC Genomics **7**, 311 (2006).
49. M. Roy, N. Kim, Y. Xing, and C. Lee, "The effect of intron length on exon creation ratios during the evolution of mammalian genomes," RNA **14**, 2261–73 (2008).
50. P. Keeling, "Five questions about microsporidia." PLoS Pathog **5**, e1000489 (2009).
51. A. Farlow, M. Dolezal, L. Hua, and C. Schlotterer, "The genomic signature of splicing-coupled selection differs between long and short introns," Mol Biol Evol **29**, 21–4 (2012).
52. A. Bonnet, A. R. Grosso, A. Elkaoutari, E. Coleno, A. Presle, S. C. Sridhara, G. Janbon, V. Géli, S. F. de Almeida, and B. Palancade, "Introns protect eukaryotic genomes from transcription-associated genetic instability." Mol Cell **67**, 608–621.e6 (2017).
53. E. Kim, A. Magen, and G. Ast, "Different levels of alternative splicing among eukaryotes," Nucleic Acids Res **35**, 125–31 (2007).
54. A. M. McGuire, M. D. Pearson, D. E. Neafsey, and J. E. Galagan, "Cross-kingdom patterns of alternative splicing and splice recognition," Genome Biol **9**, R50 (2008).
55. V. S. Bondarenko and M. S. Gelfand, "Evolution of the exon-intron structure in ciliate genomes." PLoS One **11**, e0161476 (2016).
56. B. B. Wang and V. Brendel, "Genomewide comparative analysis of alternative splicing in plants," Proc Natl Acad Sci U S A **103**, 7175–80 (2006).
57. S. L. Fernandez-Valverde, A. D. Calcino, and B. M. Degnan, "Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge amphimedon queenslandica," BMC Genomics **16**, 387 (2015).
58. P. L. Boutz, A. Bhutkar, and P. A. Sharp, "Detained introns are a novel, widespread class of post-transcriptionally spliced introns." Genes Dev **29**, 63–80 (2015).
59. T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," Nature **463**, 457–63 (2010).
60. L. Chen, S. J. Bush, J. M. Tovar-Corona, A. Castillo-Morales, and A. O. Urrutia, "Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity," Mol Biol Evol **31**, 1402–13 (2014).
61. K. Grutzmann, K. Szafranski, M. Pohl, K. Voigt, A. Petzold, and S. Schuster, "Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study," DNA Res **21**, 27–39 (2014).
62. M. Lynch and J. S. Conery, "The origins of genome complexity," Science **302**, 1401–4 (2003).
63. M. Lynch, L. M. Bobay, F. Catania, J. F. Gout, and M. Rho, "The repatterning of eukaryotic genomes by random genetic drift," Annu Rev Genomics Hum Genet **12**, 347–66 (2011).
64. T. A. Elliott and T. R. Gregory, "What's in a genome? the c-value enigma and the evolution of eukaryotic genome content," Philos Trans R Soc Lond B Biol Sci **370**, 20140331

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2015).

65. A. E. Vinogradov, "Intron-genome size relationship on a large evolutionary scale," J Mol Evol **49**, 376–84 (1999).
66. W. R. Francis and G. Wörheide, "Similar ratios of introns to intergenic sequence across animal genomes." Genome Biol Evol **9**, 1582–1598 (2017).
67. K. D. Whitney, B. Boussau, E. J. Baack, and T. Garland, Jr, "Drift and genome complexity revisited," PLoS Genet **7**, e1002092 (2011).
68. X. Wu and L. D. Hurst, "Why selection might be stronger when populations are small: Intron size and density predict within and between-species usage of exonic splice associated cis-motifs." Mol Biol Evol **32**, 1847–61 (2015).
69. D. C. Jeffares, T. Mourier, and D. Penny, "The biology of intron gain and loss," Trends Genet **22**, 16–22 (2006).
70. A. Farlow, E. Meduri, and C. Schlötterer, "Dna double-strand break repair and the evolution of intron density." Trends Genet **27**, 1–6 (2011).
71. A. F. Palazzo and T. R. Gregory, "The case for junk dna," PLoS Genet **10**, e1004351 (2014).
72. D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik, "On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of encode," Genome Biol Evol **5**, 578–90 (2013).
73. N. Sela, E. Kim, and G. Ast, "The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates," Genome Biol **11**, R59 (2010).
74. K. Jiang and L. R. Goertzen, "Spliceosomal intron size expansion in domesticated grapevine (vitis vinifera)," BMC Res Notes **4**, 52 (2011).
75. S. P. Moss, D. A. Joyce, S. Humphries, K. J. Tindall, and D. H. Lunt, "Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage," Genome Biol Evol **3**, 1187–96 (2011).
76. Q. Zhang and S. V. Edwards, "The evolution of intron size in amniotes: a role for powered flight?" Genome Biol Evol **4**, 1033–43 (2012).
77. W. F. Doolittle, T. D. Brunet, S. Linquist, and T. R. Gregory, "Distinguishing between "function" and "effect" in genome biology," Genome Biol Evol **6**, 1234–7 (2014).
78. A. E. Vinogradov, "'genome design' model and multicellular complexity: golden middle," Nucleic Acids Res **34**, 5906–14 (2006).
79. J. M. Comeron, "Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence," Genetics **167**, 1293–304 (2004).
80. M. Lynch, "Intron evolution as a population-genetic process," Proc Natl Acad Sci U S A **99**, 6118–23 (2002).
81. Y. D. Kelkar and H. Ochman, "Causes and consequences of genome expansion in fungi," Genome Biol Evol **4**, 13–23 (2012).
82. T. Garland, Jr, A. F. Bennett, and E. L. Rezende, "Phylogenetic approaches in comparative physiology," J Exp Biol **208**, 3015–35 (2005).
83. T. Munkemuller, S. Lavergne, B. Bzeznik, S. Dray, T. Jombart, K. Schiffers, and W. Thuiller, "How to measure and test phylogenetic signal," Methods in Ecology and Evolution **3**, 743–756 (2012).
84. F. Burki, "The eukaryotic tree of life from a global phylogenomic perspective." Cold Spring Harb Perspect Biol **6**, a016147 (2014).
85. C. E. Hinchliff, S. A. Smith, J. F. Allman, J. G. Burleigh, R. Chaudhary, L. M. Coghill, K. A. Crandall, J. Deng, B. T. Drew, R. Gazis, K. Gude, D. S. Hibbett, L. A. Katz, H. D. Laughinghouse, 4th, E. J. McTavish, P. E. Midford, C. L. Owen, R. H. Ree, J. A. Rees, D. E. Soltis, T. Williams, and K. A. Cranston, "Synthesis of phylogeny and taxonomy into a comprehensive tree of life." Proc Natl Acad Sci U S A **112**, 12764–9 (2015).
86. K. D. Whitney, E. J. Baack, J. L. Hamrick, M. J. Godt, B. C. Barringer, M. D. Bennett, C. G. Eckert, C. Goodwillie, S. Kalisz, I. J. Leitch, and J. Ross-Ibarra, "A role for non-adaptive processes in plant genome size evolution?" Evolution **64**, 2097–109 (2010).
87. W. P. Maddison and R. G. FitzJohn, "The unsolved challenge to phylogenetic correlation tests for categorical characters," Syst Biol **64**, 127–36 (2015).
88. Y. S. Rao, Z. F. Wang, X. W. Chai, G. Z. Wu, M. Zhou, Q. H. Nie, and X. Q. Zhang, "Selection for the compactness of highly expressed genes in gallus gallus," Biol Direct **5**, 35 (2010).
89. A. Kapusta, A. Suh, and C. Feschotte, "Dynamics of genome size evolution in birds and mammals." Proc Natl Acad Sci U S A **114**, E1460–E1469 (2017).
90. X. Wang, X. Fang, P. Yang, X. Jiang, F. Jiang, D. Zhao, B. Li, F. Cui, J. Wei, C. Ma *et al.*, "The locust genome provides insight into swarm formation and long-distance flight," Nat Commun **5**, 2957 (2014).
91. L. Zhu, Y. Zhang, W. Zhang, S. Yang, J. Q. Chen, and D. Tian, "Patterns of exon-intron architecture variation of genes in eukaryotic genomes," BMC Genomics **10**, 47 (2009).
92. T. E. Koralewski and K. V. Krutovsky, "Evolution of exon-intron structure and alternative splicing," PLoS One **6**, e18055 (2011).
93. S. S. Merchant, S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Maréchal-Drouard, W. F. Marshall, L. H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C. L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernández, H. Fukuzawa, D. González-Ballester, D. González-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J. P. Ral, D. M. Riaño-Pachón, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C. J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, I. Dubchak,

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

D. Goodstein, L. Hornick, Y. W. Huang, J. Jhaveri, Y. Luo, D. Martínez, W. C. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Grigoriev, D. S. Rokhsar, and A. R. Grossman, "The chlamydomonas genome reveals the evolution of key animal and plant functions." Science **318**, 245–50 (2007).

94. B. A. Curtis, G. Tanifuji, F. Burki, A. Gruber, M. Irimia, S. Maruyama, M. C. Arias, S. G. Ball, G. H. Gile, Y. Hirakawa *et al.*, "Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs," Nature **492**, 59–65 (2012).

95. E. Shoguchi, C. Shinzato, T. Kawashima, F. Gyoja, S. Mungpakdee, R. Koyanagi, T. Takeuchi, K. Hisata, M. Tanaka, M. Fujiwara, M. Hamada, A. Seidi, M. Fujie, T. Usami, H. Goto, S. Yamasaki, N. Arakaki, Y. Suzuki, S. Sugano, A. Toyoda, Y. Kuroki, A. Fujiyama, M. Medina, M. A. Coffroth, D. Bhattacharya, and N. Satoh, "Draft assembly of the symbiodinium minutum nuclear genome reveals dinoflagellate gene structure." Curr Biol **23**, 1399–408 (2013).

96. M. Lunghi, F. Spano, A. Magini, C. Emiliani, V. B. Carruthers, and M. Di Cristina, "Alternative splicing mechanisms orchestrating post-transcriptional gene expression: intron retention and the intron-rich genome of apicomplexan parasites," Curr Genet **62**, 31–8 (2016).

97. Y. Yang, J. Xiong, Z. Zhou, F. Huo, W. Miao, C. Ran, Y. Liu, J. Zhang, J. Feng, M. Wang, M. Wang, L. Wang, and B. Yao, "The genome of the myxosporean thelohanellus kitauei shows adaptations to nutrient acquisition within its fish host." Genome Biol Evol **6**, 3182–98 (2014).

98. T. A. Nguyen, O. H. Cisse, J. Y. Wong, P. Zheng, D. Hewitt, M. Nowrousian, J. E. Stajich, and G. Jedd, "Innovation and constraint leading to complex multicellularity in the ascomycota." Nat Commun **8**, 14444 (2017).

99. L. G. Nagy, "Evolution: Complex multicellular life with 5,500 genes." Curr Biol **27**, R609–R612 (2017).

100. D. Moore, "Principles of mushroom developmental biology." International Journal of Medicinal Mushrooms **7**, 79–101 (2005).

101. U. Kues and M. Navarro-Gonzalez, "How do agaricomycetes shape their fruiting bodies? 1. morphological aspects of development." Fungal Biology Reviews **29**, 63–97 (2015).

102. J. F. Wendel, R. C. Cronn, I. Alvarez, B. Liu, R. L. Small, and D. S. Senchina, "Intron size and genome size in plants," Mol Biol Evol **19**, 2346–52 (2002).

103. R. Guan, Y. Zhao, H. Zhang, G. Fan, X. Liu, W. Zhou, C. Shi, J. Wang, W. Liu, X. Liang, Y. Fu, K. Ma, L. Zhao, F. Zhang, Z. Lu, S. M. Lee, X. Xu, J. Wang, H. Yang, C. Fu, S. Ge, and W. Chen, "Draft genome of the living fossil ginkgo biloba." Gigascience **5**, 49 (2016).

104. S. Lockton and B. S. Gaut, "The evolution of transposable elements in natural populations of self-fertilizing arabidopsis thaliana and its outcrossing relative arabidopsis lyrata," BMC Evol Biol **10**, 10 (2010).

105. J. Stival Sena, I. Giguere, B. Boyle, P. Rigault, I. Birol, A. Zuccolo, K. Ritland, C. Ritland, J. Bohlmann, S. Jones *et al.*, "Evolution of gene structure in the conifer picea glauca: a comparative analysis of the impact of intron size," BMC Plant Biol **14**, 95 (2014).

106. F. Maumus, A.-S. Fiston-Lavier, and H. Quesneville, "Impact of transposable elements on insect genomes and biology," Current Opinion in Insect Science **7**, 30–36 (2015).

107. S. W. Roy, "Is genome complexity a consequence of inefficient selection? evidence from intron creation in nonrecombining regions," Mol Biol Evol **33**, 3088–3094 (2016).

108. G. D. Yang, K. Yan, B. J. Wu, Y. H. Wang, Y. X. Gao, and C. C. Zheng, "Genomewide analysis of intronic micrornas in rice and arabidopsis," J Genet **91**, 313–24 (2012).

109. R. Peres da Silva, R. Puccia, M. L. Rodrigues, D. L. Oliveira, L. S. Joffe, G. V. Cesar, L. Nimrichter, S. Goldenberg, and L. R. Alves, "Extracellular vesicle-mediated export of fungal rna," Sci Rep **5**, 7763 (2015).

110. P. Ramalingam, J. K. Palanichamy, A. Singh, P. Das, M. Bhagat, M. A. Kassab, S. Sinha, and P. Chattopadhyay, "Biogenesis of intronic mirnas located in clusters by independent transcription and alternative splicing," RNA **20**, 76–87 (2014).

111. A. V. Alekseyenko, N. Kim, and C. J. Lee, "Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes," RNA **13**, 661–70 (2007).

112. J. Zhu, F. He, D. Wang, K. Liu, D. Huang, J. Xiao, J. Wu, S. Hu, and J. Yu, "A novel role for minimal introns: routing mrnas to the cytosol," PLoS One **5**, e10144 (2010).

113. A. Chaurasia, A. Tarallo, L. Berna, M. Yagi, C. Agnisola, and G. D'Onofrio, "Length and gc content variability of introns among teleostean genomes in the light of the metabolic rate hypothesis," PLoS One **9**, e103889 (2014).

114. B. Charlesworth and N. Barton, "Genome size: does bigger mean worse?" Curr Biol **14**, R233–5 (2004).

115. L. Carmel, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Evolutionarily conserved genes preferentially accumulate introns." Genome Res **17**, 1045–50 (2007).

116. M. Lynch, "The origins of eukaryotic gene structure," Mol Biol Evol **23**, 450–68 (2006).

117. G. Arnqvist, A. Sayadi, E. Immonen, C. Hotzy, D. Rankin, M. Tuda, C. E. Hjelmen, and J. S. Johnston, "Genome size correlates with reproductive fitness in seed beetles," Proc Biol Sci **282**, 20151421 (2015).

118. B. Jimenez-Mena, P. Tataru, R. F. Brøndum, G. Sahana, B. Guldbrandtsen, and T. Bataillon, "One size fits all? direct evidence for the heterogeneity of genetic drift throughout the genome." Biol Lett **12** (2016).

119. V. Daubin and N. A. Moran, "Comment on "the origins of genome complexity"," Science **306**, 978; author reply 978 (2004).

120. J. Romiguier, P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, E. Loire, J. M. Lourenco, B. Nabholz, C. Roux, G. Tsagkogeorga, A. A. Weber, L. A. Weinert, K. Belkhir, N. Bierne, S. Glémin, and N. Galtier, "Comparative population genomics in animals uncovers the determinants of genetic diversity." Nature **515**, 261–3 (2014).

121. B. Schierwater, M. Eitel, W. Jakob, H. J. Osigus, H. Hadrys, S. L. Dellaporta, S. O. Kolokotronis, and R. Desalle, "Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis." PLoS Biol **7**, e20 (2009).

122. C. W. Dunn, G. Giribet, G. D. Edgecombe, and A. Hejnol,

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

"Animal phylogeny and its evolutionary implications," Annual review of ecology, evolution, and systematics **45**, 371–395 (2014).

123. G. Giribet and G. D. Edgecombe, "Reevaluating the arthropod tree of life." Annu Rev Entomol **57**, 167–86 (2012).

124. F. Burki and P. J. Keeling, "Rhizaria." Curr Biol **24**, R103–7 (2014).

125. A. J. George, "Is the number of genes we possess limited by the presence of an adaptive immune system?" Trends Immunol **23**, 351–5 (2002).

126. A. Nitsche, D. Rose, M. Fasold, K. Reiche, and P. F. Stadler, "Comparison of splice sites reveals that long noncoding rnas are evolutionarily well conserved." RNA **21**, 801–12 (2015).

127. Z. M. Zhao, M. C. Campbell, N. Li, D. Lee, Z. Zhang, and J. P. Townsend, "Detection of regional variation in selection intensity within protein-coding genes using dna sequence polymorphism and divergence." Mol Biol Evol (2017).

128. J. Dolezel and J. Greilhuber, "Nuclear genome size: are we getting closer?" Cytometry A **77**, 635–42 (2010).

129. R. K. Grosberg and R. R. Strathmann, "The evolution of multicellularity: a minor major transition?" Annu. Rev. Ecol. Evol. Syst **38**, 621–654 (2007).

130. A. H. Knoll, "The multiple origins of complex multicellularity," Annual Review of Earth and Planetary Sciences **39**, 217–239 (2011).

131. I. Lozada-Chávez, P. F. Stadler, and S. J. Prohaska, ""hypothesis for the modern rna world": a pervasive noncoding rna-based genetic regulation is a prerequisite for the emergence of multicellular complexity," Orig Life Evol Biosph **41**, 587–607 (2011).

132. R. M. Fisher, C. K. Cornwallis, and S. A. West, "Group formation, relatedness, and the evolution of multicellularity," Curr Biol **23**, 1120–5 (2013).

133. S. A. West, R. M. Fisher, A. Gardner, and E. T. Kiers, "Major evolutionary transitions in individuality." Proc Natl Acad Sci U S A **112**, 10112–9 (2015).

134. E. Szathmary and J. M. Smith, "The major evolutionary transitions." Nature **374**, 227–32 (1995).

135. E. Ibarra-Laclette, E. Lyons, G. Hernandez-Guzman, C. A. Perez-Torres, L. Carretero-Paulet, T. H. Chang, T. Lan, A. J. Welch, M. J. Juarez, J. Simpson *et al.*, "Architecture and evolution of a minute plant genome," Nature **498**, 94–8 (2013).

136. K. J. Niklas, E. D. Cobb, and A. K. Dunker, "The number of cell types, information content, and the evolution of complex multicellularity." Acta Soc Bot Pol **83**, 337–347 (2014).

137. L. Carmel, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Patterns of intron gain and conservation in eukaryotic genes," BMC Evol Biol **7**, 192 (2007).

138. M. Chorev and L. Carmel, "The function of introns," Front Genet **3**, 55 (2012).

139. A. Rokas, "The molecular origins of multicellular transitions," Curr Opin Genet Dev **18**, 472–8 (2008).

140. D. J. Richter and N. King, "The genomic and cellular foundations of animal origins." Annu Rev Genet **47**, 509–37 (2013).

141. J. K. Pickrell, A. A. Pai, Y. Gilad, and J. K. Pritchard, "Noisy splicing drives mrna isoform diversity in human cells," PLoS Genet **6**, e1001236 (2010).

142. D. A. Bitton, S. R. Atkinson, C. Rallis, G. C. Smith, D. A. Ellis, Y. Y. Chen, M. Malecki, S. Codlin, J. F. Lemay, C. Cotobal *et al.*, "Widespread exon skipping triggers degradation by nuclear rna surveillance in fission yeast," Genome Res **25**, 884–96 (2015).

143. M. Gonzalez-Porta, A. Frankish, J. Rung, J. Harrow, and A. Brazma, "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene," Genome Biol **14**, R70 (2013).

144. B. Taneri, B. Snyder, and T. Gaasterland, "Distribution of alternatively spliced transcript isoforms within human and mouse transcriptomes," Journal of OMICS Research **1**, 1–5 (2011).

145. J. B. Brown, N. Boley, R. Eisman, G. E. May, M. H. Stoiber, M. O. Duff, B. W. Booth, J. Wen, S. Park, A. M. Suzuki *et al.*, "Diversity and dynamics of the drosophila transcriptome," Nature **512**, 393–9 (2014).

146. J. Tapial, K. Ha, T. Sterne-Weiler, A. Gohr, U. Braunschweig, A. Hermoso-Pulido, M. Quesnel-Vallières, J. Permanyer, R. Sodaei, Y. Marquez, L. Cozzuto, X. Wang, M. Gómez-Velázquez, T. Rayon, M. Manzanares, J. Ponomarenko, B. J. Blencowe, and M. Irimia, "An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms." Genome Res **27**, 1759–1768 (2017).

147. S. Li, M. Yamada, X. Han, U. Ohler, and P. N. Benfey, "High-resolution expression map of the arabidopsis root reveals alternative splicing and lincrna regulation." Dev Cell **39**, 508–522 (2016).

148. Y. Ge and B. T. Porse, "The functional consequences of intron retention: alternative splicing coupled to nmd as a regulator of gene expression," Bioessays **36**, 236–43 (2014).

149. A. G. Jacob and C. Smith, "Intron retention as a component of regulated gene expression programs." Hum Genet **136**, 1043–1057 (2017).

150. S. Perez-Santangelo, R. G. Schlaen, and M. J. Yanovsky, "Genomic analysis reveals novel connections between alternative splicing and circadian regulatory networks," Brief Funct Genomics **12**, 13–24 (2013).

151. T. C. Boothby, R. S. Zipper, C. M. van der Weele, and S. M. Wolniak, "Removal of retained introns regulates translation in the rapidly developing gametophyte of marsilea vestita," Dev Cell **24**, 517–29 (2013).

152. S. A. Filichkin, J. S. Cumbie, P. Dharmawardhana, P. Jaiswal, J. H. Chang, S. G. Palusa, A. S. Reddy, M. Megraw, and T. C. Mockler, "Environmental stresses modulate abundance and timing of alternatively spliced circadian transcripts in arabidopsis," Mol Plant **8**, 207–27 (2015).

153. K. Yap, Z. Q. Lim, P. Khandelia, B. Friedman, and E. V. Makeyev, "Coordinated regulation of neuronal mrna steady-state levels through developmentally controlled intron retention," Genes Dev **26**, 1209–23 (2012).

154. U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes transcriptomes," Genome Res **24**, 1774–86 (2014).

155. J. J. Wong, W. Ritchie, O. A. Ebner, M. Selbach, J. W. Wong,

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya *et al.*, "Orchestrated intron retention regulates normal granulocyte differentiation," Cell **154**, 583–95 (2013).

156. O. Mauger, F. Lemoine, and P. Scheiffele, "Targeted intron retention and excision for rapid gene regulation in response to neuronal activity." Neuron **92**, 1266–1278 (2016).

157. N. P. Hoyle and D. Ish-Horowicz, "Transcript processing and export kinetics are rate-limiting steps in expressing vertebrate segmentation clock genes," Proc Natl Acad Sci U S A **110**, E4316–24 (2013).

158. A. Hanisch, M. V. Holder, S. Choorapoikayil, M. Gajewski, E. M. Ozbudak, and J. Lewis, "The elongation rate of rna polymerase ii in zebrafish and its significance in the somite segmentation clock," Development **140**, 444–53 (2013).

159. Y. Harima, Y. Takashima, Y. Ueda, T. Ohtsuka, and R. Kageyama, "Accelerating the tempo of the segmentation clock by reducing the number of introns in the hes7 gene," Cell Rep **3**, 1–7 (2013).

160. C. Seoighe and P. K. Korir, "Evidence for intron length conservation in a set of mammalian genes associated with embryonic development," BMC Bioinformatics **12 Suppl 9**, S16 (2011).

161. P. A. Keane and C. Seoighe, "Intron length coevolution across mammalian genomes," Mol Biol Evol **33**, 2682–91 (2016).

162. P. Heyn, M. Kircher, A. Dahl, J. Kelso, P. Tomancak, A. T. Kalinka, and K. M. Neugebauer, "The earliest transcribed zygotic genes are short, newly evolved, and different across species," Cell Rep **6**, 285–92 (2014).

163. L. G. Guilgur, P. Prudencio, D. Sobral, D. Liszekova, A. Rosa, and R. G. Martinho, "Requirement for highly efficient pre-mrna splicing during drosophila early embryonic development," Elife **3**, e02181 (2014).

164. J. J. Wong, A. Y. Au, W. Ritchie, and J. E. Rasko, "Intron retention in mrna: No longer nonsense: Known and putative roles of intron retention in normal and disease biology." Bioessays **38**, 41–9 (2016).

165. A. Ranjan, B. T. Townsley, Y. Ichihashi, N. R. Sinha, and D. H. Chitwood, "An intracellular transcriptomic atlas of the giant coenocyte caulerpa taxifolia," PLoS Genet **11**, e1004900 (2015).

166. T. Y. James, "Why mushrooms have evolved to be so promiscuous: Insights from evolutionary and ecological patterns." Fungal Biology Reviews **29**, 167–178 (2015).

167. S. A. Rensing, "(why) does evolution favour embryogenesis?" Trends Plant Sci **21**, 562–73 (2016).

168. Z. Du, A. Santella, F. He, P. K. Shah, Y. Kamikawa, and Z. Bao, "The regulatory landscape of lineage differentiation in a metazoan embryo." Dev Cell **34**, 592–607 (2015).

169. J. Signolet and B. Hendrich, "The function of chromatin modifiers in lineage commitment and cell fate specification." FEBS J **282**, 1692–702 (2015).

170. A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," Bioinformatics **21 Suppl 1**, i351–8 (2005).

171. A. Favorov, L. Mularoni, L. M. Cope, Y. Medvedeva, A. A. Mironov, V. J. Makeev, and S. J. Wheelan, "Exploring massive, genome scale datasets with the genometricorr package," PLoS Comput Biol **8**, e1002529 (2012).

172. Y. V. Kravatsky, V. R. Chechetkin, N. A. Tchurikov, and G. I. Kravatskaya, "Genome-wide study of correlations between genomic features and their relationship with the regulation of gene expression," DNA Res **22**, 109–19 (2015).

173. S. Yang, R. F. Doolittle, and P. E. Bourne, "Phylogeny determined by protein domain content," Proc Natl Acad Sci U S A **102**, 373–8 (2005).

174. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, and M. Punta, "Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions," Nucleic Acids Res **41**, e121 (2013).

175. R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas *et al.*, "The pfam protein families database: towards a more sustainable future," Nucleic Acids Res **44**, D279–85 (2016).

176. T. Cavalier-Smith, E. E. Chao, E. A. Snell, C. Berney, A. M. Fiore-Donno, and R. Lewis, "Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and amoebozoa," Mol Phylogenet Evol **81**, 71–85 (2014).

177. D. He, O. Fiz-Palacios, C. J. Fu, J. Fehling, C. C. Tsai, and S. L. Baldauf, "An alternative root for the eukaryote tree of life." Curr Biol **24**, 465–70 (2014).

178. B. C. Stover and K. F. Muller, "Treegraph 2: combining and visualizing evidence from different phylogenetic analyses," BMC Bioinformatics **11**, 7 (2010).

179. M. K. Kuhner and J. Yamato, "Practical performance of tree comparison metrics." Syst Biol **64**, 205–14 (2015).

180. D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," Mathematical biosciences **53**, 131–147 (1981).

181. T. M. Nye, P. Liò, and W. R. Gilks, "A novel algorithm and web-based tool for comparing two alternative phylogenetic trees." Bioinformatics **22**, 117–9 (2006).

182. D. Orme, R. Freckleton, G. Thomas, T. Petzoldt, S. Fritz, N. Isaac, and W. Pearse, "caper: Comparative analyses of phylogenetics and evolution in r. r package version 0.5," Proc. R. Soc. B **283**, 7 (2012).

183. R. Molina-Venegas and M. Rodríguez, "Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information?" BMC Evol Biol **17**, 53 (2017).

184. A.-A. Popescu, K. T. Huber, and E. Paradis, "ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in r," Bioinformatics **28**, 1536–1537 (2012).

185. W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois (2017). R package version 1.7.8.

186. W. C. Ratcliff, R. F. Denison, M. Borrello, and M. Travisano, "Experimental evolution of multicellularity," Proc Natl Acad Sci U S A **109**, 1595–600 (2012).

187. W. C. Ratcliff, M. D. Herron, K. Howell, J. T. Pentz, F. Rosenzweig, and M. Travisano, "Experimental evolution of an alternating uni- and multicellular life cycle in chlamydomonas reinhardtii," Nat Commun **4**, 2742 (2013).

188. S. J. Hanson and K. H. Wolfe, "An evolutionary perspective on yeast mating-type switching." Genetics **206**, 9–32 (2017).

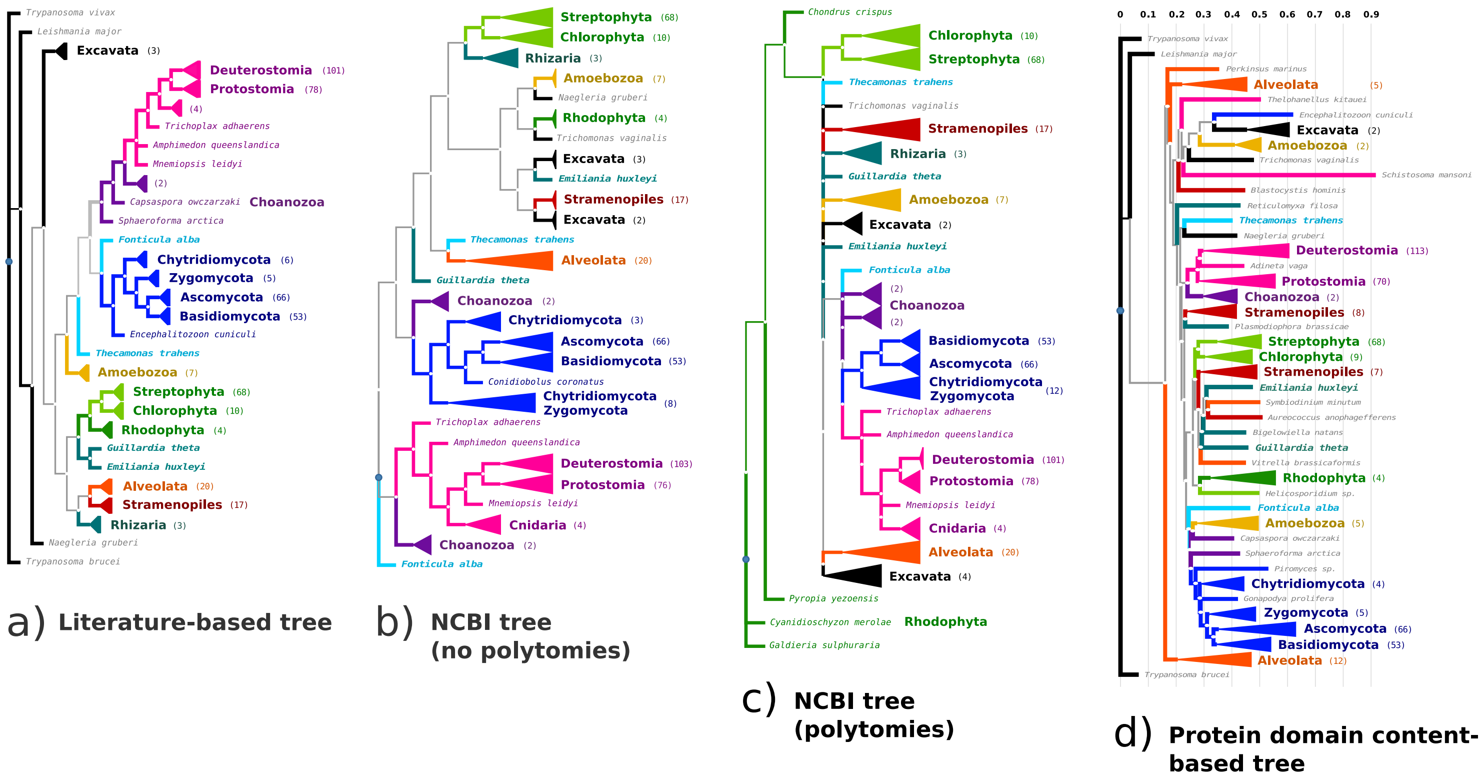
189. R. J. Bennett, "The parasexual lifestyle of candida albicans." Curr Opin Microbiol **28**, 10–7 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

190. P. Gaudet, J. G. Williams, P. Fey, and R. L. Chisholm, "An anatomy ontology to represent biological knowledge in dictyostelium discoideum," BMC Genomics **9**, 130 (2008).
191. K. J. Niklas, E. D. Cobb, and D. R. Crawford, "The evo-devo of multinucleate cells, tissues, and organisms, and an alternative route to multicellularity." Evol Dev **15**, 466–74 (2013).
192. J. E. Strassmann, Y. Zhu, and D. C. Queller, "Altruism and social cheating in the social amoeba dictyostelium discoideum." Nature **408**, 965–7 (2000).
193. N. J. Mehdiabadi, C. N. Jack, T. T. Farnham, T. G. Platt, S. E. Kalla, G. Shaulsky, D. C. Queller, and J. E. Strassmann, "Social evolution: kin preference in a social microbe." Nature **442**, 881–2 (2006).
194. J. C. Coates, Umm-E-Aiman, and B. Charrier, "Understanding "green" multicellularity: do seaweeds hold the key?" Front Plant Sci **5**, 737 (2014).
195. D. W. McShea, "The minor transitions in hierarchical evolution and the question of a directional bias," Journal of Evolutionary Biology **14**, 502–518 (2001).
196. T. Domazet-Loso, A. Klimovich, B. Anokhin, F. Anton-Erxleben, M. J. Hamm, C. Lange, and T. C. Bosch, "Naturally occurring tumours in the basal metazoan hydra." Nat Commun **5**, 4222 (2014).
197. H. Chen, F. Lin, K. Xing, and X. He, "The reverse evolution from multicellularity to unicellularity during carcinogenesis." Nat Commun **6**, 6367 (2015).
198. M. J. Layden, F. Rentzsch, and E. Rottinger, "The rise of the starlet sea anemone nematostella vectensis as a model system to investigate development and regeneration." Wiley Interdiscip Rev Dev Biol **5**, 408–28 (2016).
199. P. Steinmetz, A. Aman, J. Kraus, and U. Technau, "Gut-like ectodermal tissue in a sea anemone challenges germ layer homology." Nat Ecol Evol **1**, 1535–1542 (2017).
200. A. Kirillova, G. Genikhovich, E. Pukhlyakova, A. Demilly, Y. Kraus, and U. Technau, "Germ-layer commitment and axis formation in sea anemone embryonic cell aggregates." Proc Natl Acad Sci U S A **115**, 1813–1818 (2018).
201. A. Capron, S. Chatfield, N. Provart, and T. Berleth, "Embryogenesis: pattern formation from a single cell." Arabidopsis Book **7**, e0126 (2009).
202. E. Sparks, G. Wachsman, and P. N. Benfey, "Spatiotemporal signalling in plant development." Nat Rev Genet **14**, 631–44 (2013).
203. I. De Smet and T. Beeckman, "Asymmetric cell division in land plants and algae: the driving force for differentiation." Nat Rev Mol Cell Biol **12**, 177–88 (2011).
204. B. Charrier, A. Le Bail, and B. de Reviers, "Plant proteus: brown algal morphological plasticity and underlying developmental mechanisms," Trends Plant Sci **17**, 468–77 (2012).
205. O. Godfroy, T. Uji, C. Nagasato, A. P. Lipinska, D. Scornet, A. F. Peters, K. Avia, S. Colin, L. Mignerot, T. Motomura, J. M. Cock, and S. M. Coelho, "Distag/tbccd1 is required for basal cell fate determination in," Plant Cell **29**, 3102–3122 (2017).
206. M. Nowrousian, M. Piotrowski, and U. Kuck, "Multiple layers of temporal and spatial control regulate accumulation of the fruiting body-specific protein app in sordaria macrospora and neurospora crassa." Fungal Genet Biol **44**, 602–14 (2007).
207. M. Adamska, S. M. Degnan, K. M. Green, M. Adamski, A. Craigie, C. Larroux, and B. M. Degnan, "Wnt and tgf-beta expression in the sponge amphimedon queenslandica and the origin of metazoan embryonic patterning," PLoS One **2**, e1031 (2007).
208. N. Nakanishi, S. Sogabe, and B. M. Degnan, "Evolutionary origin of gastrulation: insights from sponge development." BMC Biol **12**, 26 (2014).
209. W. Jakob, S. Sagasser, S. Dellaporta, P. Holland, K. Kuhn, and B. Schierwater, "The trox-2 hox/parahox gene of trichoplax (placozoa) marks an epithelial boundary." Dev Genes Evol **214**, 170–5 (2004).
210. B. Schierwater, D. de Jong, and R. Desalle, "Placozoa and the evolution of metazoa and intrasomatic cell differentiation." Int J Biochem Cell Biol **41**, 370–9 (2009).
211. L. Guidi, M. Eitel, E. Cesarini, B. Schierwater, and M. Balsamo, "Ultrastructural analyses support different morphological lineages in the phylum placozoa grell, 1971." J Morphol **272**, 371–8 (2011).
212. M. Eitel, H. J. Osigus, R. DeSalle, and B. Schierwater, "Global diversity of the placozoa." PLoS One **8**, e57131 (2013).
213. D. E. Shelton, A. G. Desnitskiy, and R. E. Michod, "Distributions of reproductive and somatic cell numbers in diverse volvox (chlorophyta) species," Evol Ecol Res **14**, 707–727 (2012).
214. G. D. Stormo, "Gene-finding approaches for eukaryotes," Genome Res **10**, 394–7 (2000).
215. R. V. Davuluri, I. Grosse, and M. Q. Zhang, "Computational identification of promoters and first exons in the human genome," Nat Genet **29**, 412–7 (2001).
216. S. W. Roy and D. Penny, "Intron length distributions and gene prediction," Nucleic Acids Res **35**, 4737–42 (2007).
217. S. Gudlaugsdottir, D. R. Boswell, G. R. Wood, and J. Ma, "Exon size distribution and the origin of introns," Genetica **131**, 299–306 (2007).

FIGURES AND TABLES

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



	# tips	internal nodes	tree resolution	branch lengths	a)	b)	c)	d)
Literature-based tree	461	451	unrooted	no	0.00			
NCBI tree (no polytomies)	461	459	rooted	no	**402** (0.87) 101.02	0.00		
NCBI tree (polytomies)	461	287	rooted	no	**412** (0.89) 105.69	**314** (0.68) 89.45	0.00	
Protein domain tree	461	461	unrooted	yes	**672** (1.46) 181.77	**588** (1.28) 148.013	**690** (1.50) 187.21	0.00

Figure 1. Comparison of the four alternative tree topologies estimated for the 461 eukaryotes analyzed in this study. (a) A consensus literature-based tree was manually created according to sequence-based phylogenies and supertrees. Two NCBI taxonomy-based trees were obtained in two versions: (b) not allowing polytomies and (c) allowing polytomies. (d) A protein domain content-based tree was created and corrected for protein content biases owing to differences in genome size and lifestyles. All tree topologies include: 131 fungal species (in dark blue), 78 species from Viridiplantae (in light green), 186 from Metazoa (in pink), 20 from Alveolata (in orange), 17 from Stramenopiles (in red), 7 from Excavata (in black), 7 from Amoebozoa (in yellow), 4 from Rhodophyta (in dark green), 4 from Choanozoa (in purple), 3 from Rhizaria, Cryptophyta and Haptophyta (in cyan), Fonticulidae and Apusozoa (in light blue). Phylogenetic uncertainties and disagreements among the trees are also summarized in the table (down). Numbers highlighted in bold correspond to the absolute symmetric differences (RF), the number of RF partitions per species is coloured in blue, and the Align mismatches scored in the best alignment of the branches are coloured in red.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

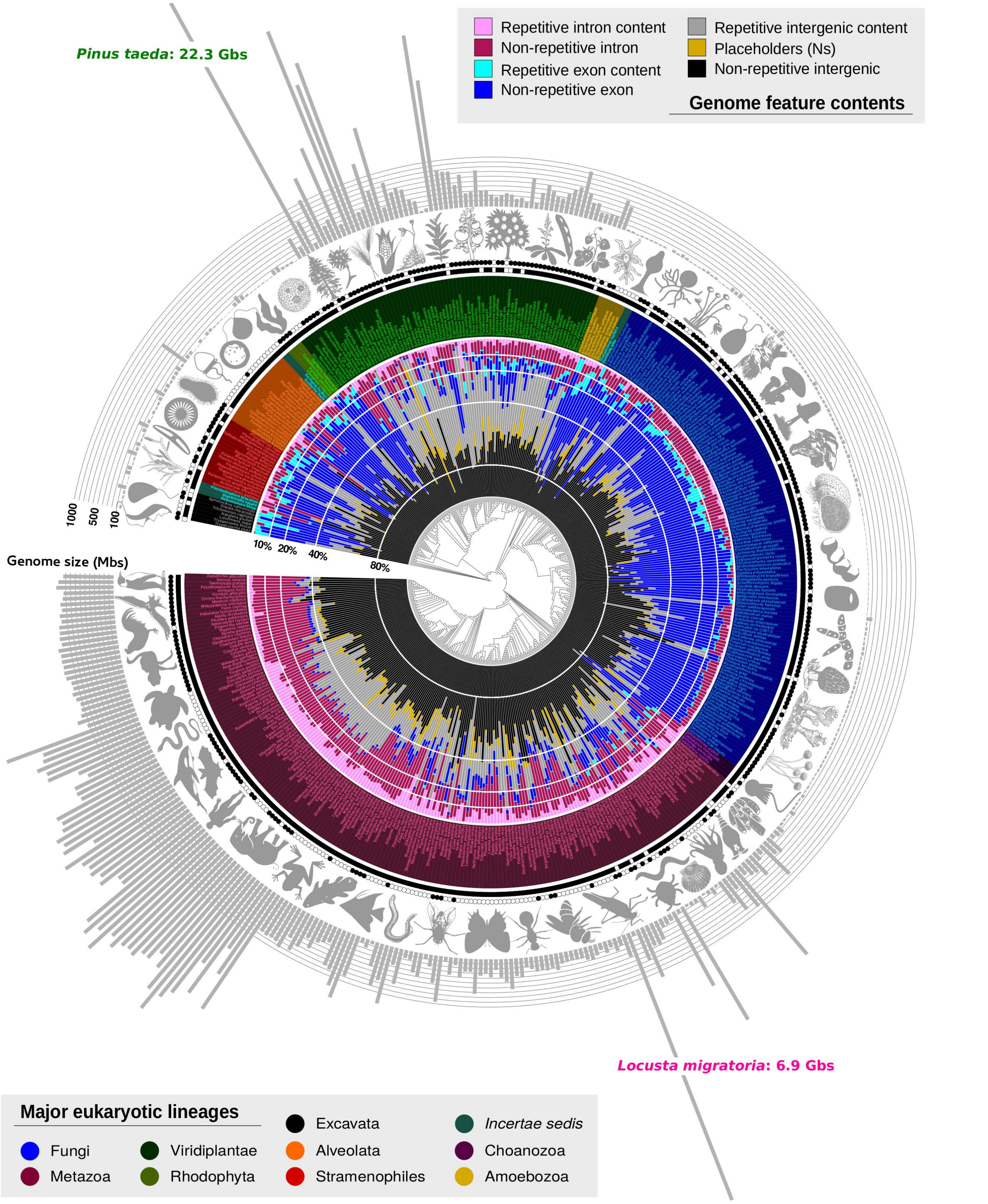


Figure 2. Approximate estimation of genome size and contents across 461 eukaryotes. The species are displayed according to the "reference tree for eukaryotes" (in Figure 1a) and coloured according to the supergroup they belong to. The fraction of genomic content (repetitive and non-repetitive) from the total assembled genome size –as calculated with the `GenomeContent` program–, for introns, exons and intergenic regions is coloured in red, blue and gray scales, respectively. The fraction de placeholders (sequences of Ns) is represented in yellow. Noteworthy, the fraction of non-repetitive intergenic regions might be smaller due to the presence of several repetitive pseudogenes and non-coding RNA families (*e.g.*, ribosomal RNAs, tRNAs) that are not fully annotated in the genome projects nor in the present study. The paired symbols for each species indicate whether the nucleotide overlap of repeats within intronic (circles) and exonic sequences (squares) is less (TRUE: filled) or more (FALSE: not filled) significant than expected by chance, according to the $p < 0.05$ estimated over 1,000 permutation tests on the *Jaccard index* for each feature and genome (see Methods and Supplementary Table S16). The assembled genome size for every species is shown in vertical bars. Data calculated from estimated genome sizes are available in Supplementary Table S3.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

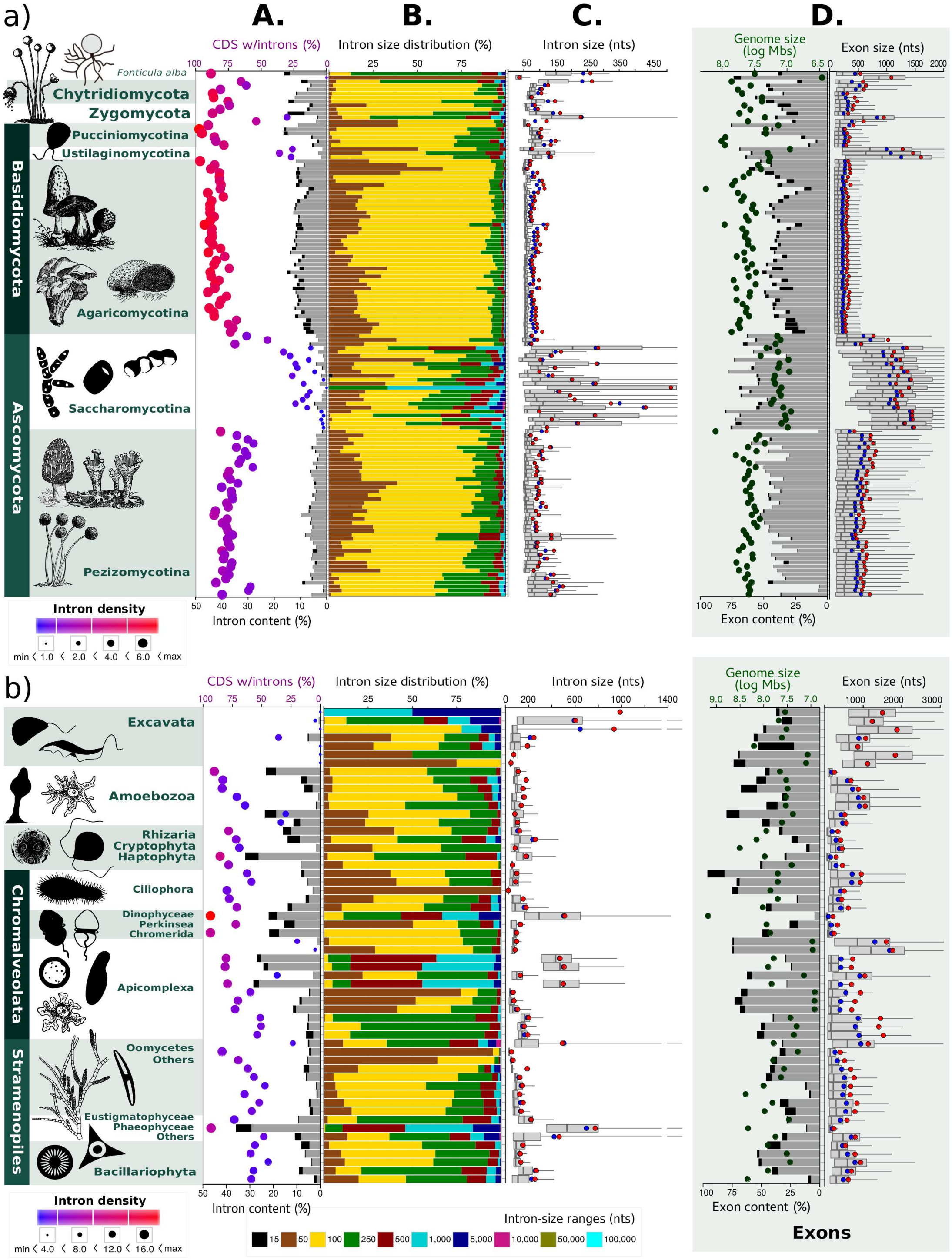


Figure 3. Phylogenetic distribution of intron and exon features across sequenced genomes in a) Fungi and b) Protists. Panel A. Distribution of intron content, density and fraction of protein-coding genes containing introns across species. Intron contents with unique and repetitive intronic sequences (based on estimated genome sizes) are shown in grey and black bars, respectively. Estimations based on the assembled genome sizes are plotted in Figures S7a-S10a. The information represented by the dots is two fold: 1) the fraction (%) of genes with introns is represented by the coordinate with respect to the top scale; 2) intron density is depicted by the size and the color of the dot, so that, a bigger dot with an intensified red color implays the presence of more introns per genes. **Panel B.** Intron size distribution within a genome is represented by the fraction (%) of introns from the total population binned in the following ranges: 15-50 nts (brown), 51-100 nts (yellow), <250 nts: introns presumably spliced by intron definition (green), >251 nts: introns presumably dismissed by exon definition. **Panel C.** Grey box-plots show the descriptors (quartiles, means and outlier-thresholds) for the intron length distribution in every genome; from right to left: Q1, median, Q3, upper-fence (line), standard average size (blue dot), and weigthed-average size (red dot). **Panel D.** Distribution of exon features. On the left side, bars show the exon content (bottom scale) with unique and repetitive exonic sequences in grey and black, respectively. Genome size (log Mbs) is represented by the coordinate with respect to the logarithmic top scale. On the right side, the exon size distribution within the genome is shown as described for introns in Panel C. Some upper-fences were cut to avoid a higher compactation of the data. The images for model species were kindly provided by silhouettesfree.com and ClipArt.com.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

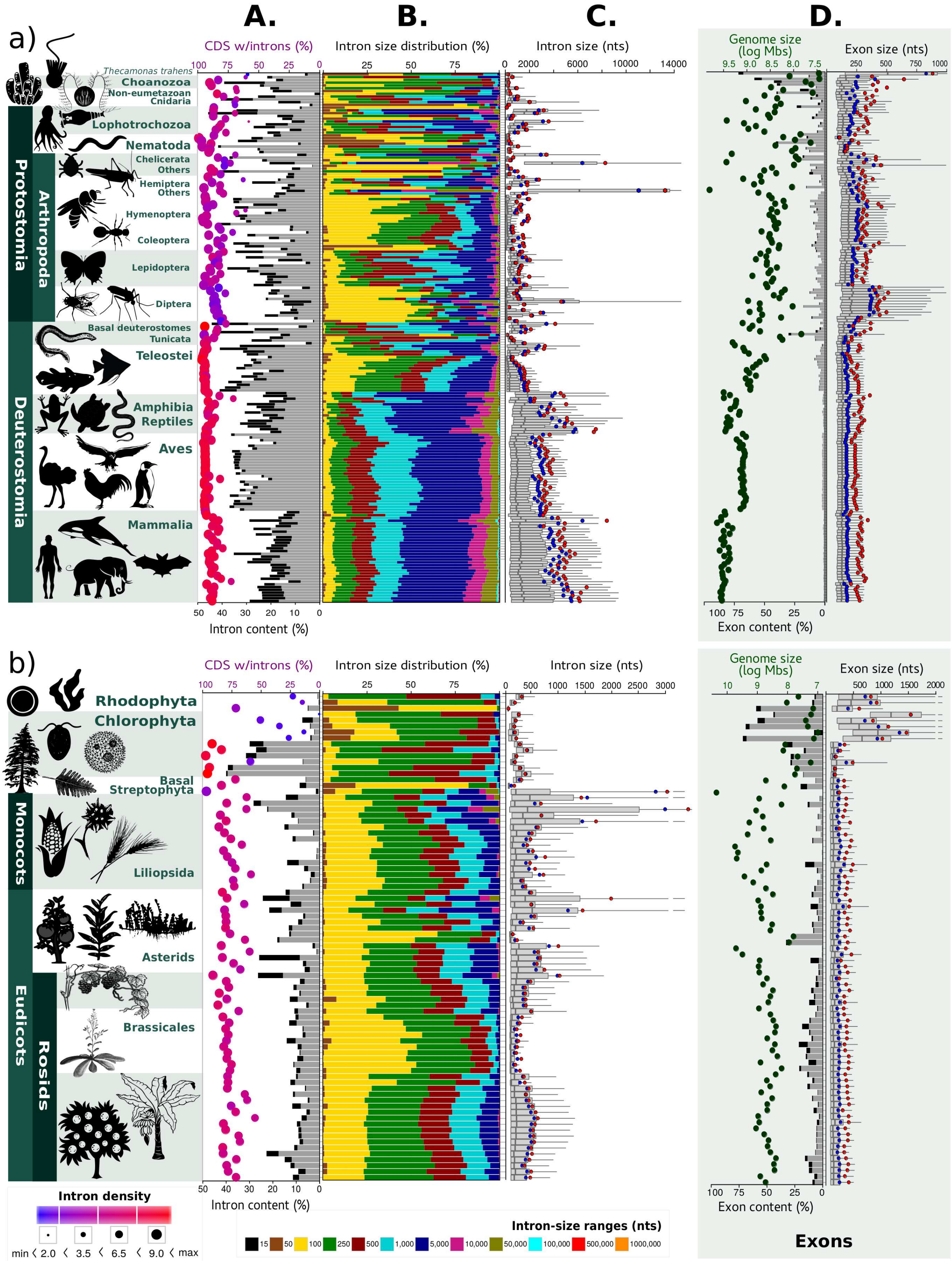


Figure 4. Phylogenetic distribution of intron and exon features across sequenced genomes in a) Metazoa and b) Archaeplastida. For panel description see Figure 3.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

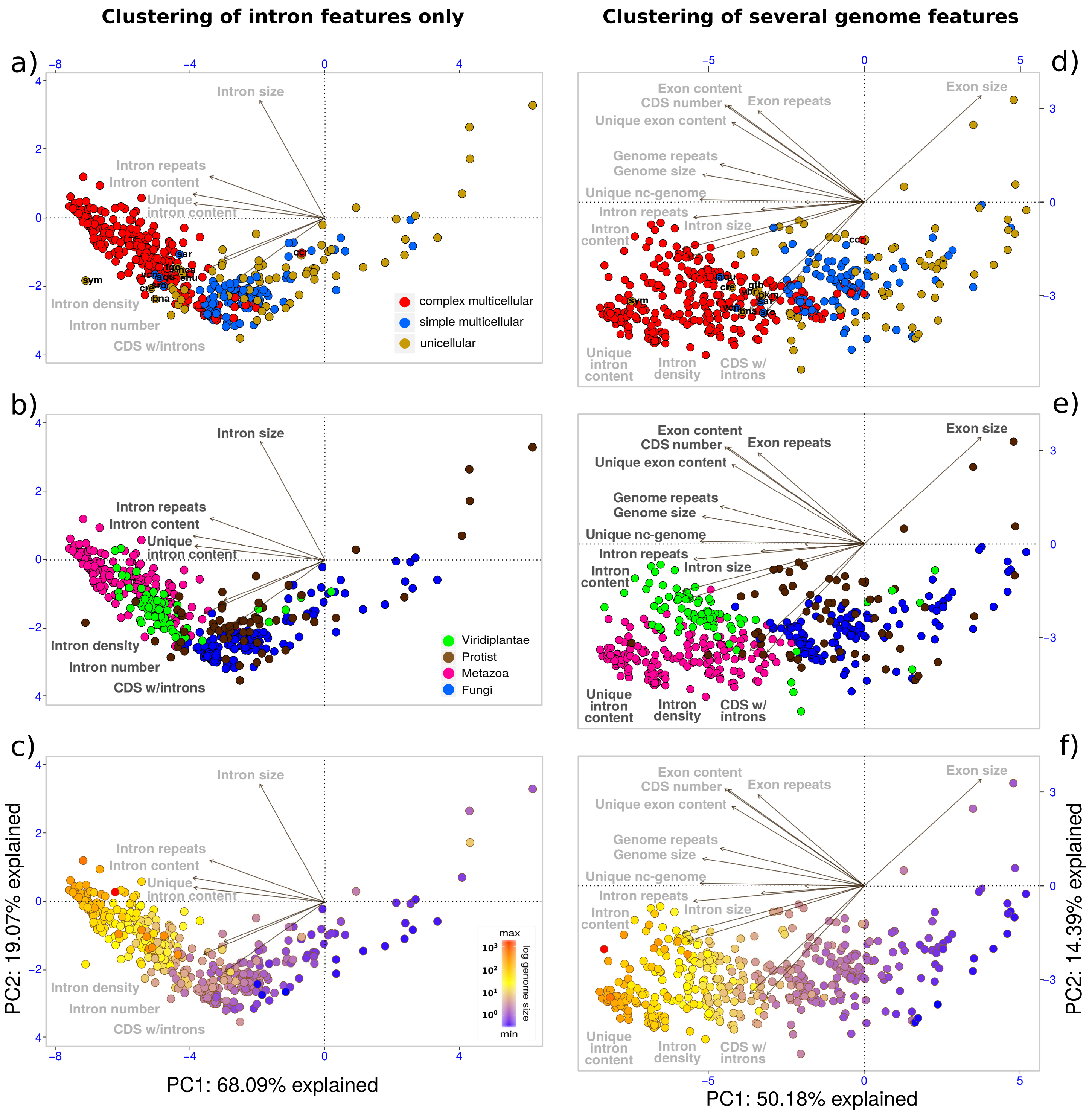


Figure 5. Phylogenetic principal component analysis (phyloPCA) of introns, other genome features and multicellular complexity. The phyloPCAs were performed with the reference tree topology (Figure 1a). The biplots depict the two first components (PC1 and PC2) inferred from the phyloPCA of 7 intron features (a-c), 14 genomic traits (d-f) and the organism complexity (a and d) estimated for 457 sequenced eukaryotes. In plots a) and d), the species are color-coded according to their organismal complexity (see Appendix 1): unicellular (yellow), simple multicellular (blue), complex multicellular (red). In plots b) and e), species are color-coded by major eukaryotic supergroups: Fungi (blue), Metazoa (pink), Viridiplantae (green), and “protists” (brown). In plots c) and f), species are color-coded by their log10- transformed genome sizes. Species are clustered in the phyloPCAs according to their dispersion along PC1 and PC2, with a confidence limit of 0.95. The dark lines radiating from (0,0) represent each variable included in the analysis; the direction of a line represents the highest correlation coefficient between the scores of the principal components and the variable, while its length is proportional to the strength of this correlation. Noteworthy, the Cronbach’s alpha coefficients obtained for the sets of 7 and 15 variables (0.88 and 0.95, respectively) indicate high internal consistency among the variables to measure the same underlying concept through phyloPCA analyses (see Methods). PhyloPCA analyses performed with alternative tree topologies and assembled genome sizes are provided in Supplementary Tables S16-S20 and Figures S11-S15.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Unvariate measures of phylogenetic signal (with the λ parameter) for the genome traits analysed in this study, by using four alternate trees for the 461 eukaryotic species (see Figure 1).

Tree topology			**Reference**		**Protein**	**NCBI**	**NCBI**	
			Estimated		**domain**	**No Poly**	**Yes Poly**	
Genome trait	λ	**ln ML**	**AIC**	**CI (95%)**	λ	λ	λ	D.F.
Estimated genome size (Mbs)	**0.938**	-129.197	260.395	(0.904, 0.960)	0.999	0.945	0.939	460
Assembled genome size (Mbs)	**0.936**	-117.203	236.406	(0.901, 0.958)	0.999	0.943	0.937	460
Genome repeat content (Mbs)	**0.898**	-362.153	726.305	(0.843, 0.936)	0.999	0.916	0.889	460
Genome repeat content (%)	**0.773**	-116.277	234.554	(0.661, 0.854)	0.965	0.786	0.757	460
Unique genome content (Mbs)	**0.939**	-41.127	84.255	(0.907, 0.960)	0.999	0.938	0.937	460
Unique genome content (%)	**0.759**	421.403	-840.806	(0.632, 0.851)	0.865	0.827	0.748	460
Unique nc-genome content (Mbs)	**0.948**	-140.983	283.965	(0.918, 0.967)	0.999	0.929	0.944	460
Unique nc-genome content (%)	**0.805**	342.896	-683.792	(0.677, 0.890)	0.971	0.679	0.755	460
CDS number	**0.914**	154.015	-306.030	(0.863, 0.947)	0.999	0.896	0.929	460
CDS avg size (nts)	**0.873**	118.854	-235.708	(0.811, 0.917)	0.999	0.924	0.891	460
CDS genome coverage (Mbs)	**0.856**	-55.893	113.786	(0.790, 0.904)	0.999	0.847	0.840	460
Exon density	**0.953**	265.697	-529.395	(0.917, 0.973)	0.999	0.937	0.940	460
Exon number	**0.954**	12.167	-22.335	(0.926, 0.972)	0.999	0.924	0.951	460
Exon avg-size (nts)	**0.987**	316.981	-631.963	(0.979, 0.992)	0.999	0.979	0.990	460
Exon content (Mbs)	**0.859**	168.034	-334.068	(0.776, 0.914)	0.999	0.835	0.885	460
Exon content (%)	**0.913**	-67.455	136.910	(0.865, 0.945)	0.999	0.924	0.887	460
Unique exon content (Mbs)	**0.880**	210.273	-418.547	(0.807, 0.926)	0.999	0.851	0.883	460
Unique exon content (%)	**0.479**	742.122	-1482.244	(0.312, 0.637)	0.743	0.500	0.634	460
Repeat exon content (Mbs)	**0.797**	-244.207	490.415	(0.688, 0.874)	0.984	0.827	0.842	460
Repeat exon content (%)	**0.776**	-141.757	285.515	(0.655, 0.863)	0.955	0.818	0.798	460
Number of CDS with introns	**0.955**	-128.310	258.620	(0.935, 0.969)	0.999	0.919	0.950	460
CDS with introns (%)	**0.949**	-0.495	2.989	(0.929, 0.964)	0.861	0.892	0.928	460
Intron number	**0.960**	-222.858	447.715	(0.940, 0.973)	0.999	0.930	0.958	460
Intron density	**0.934**	286.690	-571.380	(0.890, 0.962)	0.999	0.943	0.926	460
Intron wm-size (nts)	**0.875**	-10.518	23.035	(0.812, 0.919)	0.999	0.922	0.875	460
Intron content (Mbs)	**0.950**	-310.428	622.856	(0.924, 0.967)	0.999	0.901	0.948	460
Intron content (%)	**0.932**	-174.617	351.234	(0.895, 0.957)	0.999	0.871	0.925	460
Unique intron content (Mbs)	**0.955**	-290.605	583.211	(0.931, 0.970)	0.999	0.901	0.951	460
Unique intron content (%)	**0.616**	527.481	-1052.962	(0.447, 0.748)	0.998	0.642	0.611	460
Repeat intron content (Mbs)	**0.885**	-452.997	907.994	(0.829, 0.925)	0.999	0.926	0.873	456
Repeat intron content (%)	**0.728**	-158.590	319.180	(0.596, 0.824)	0.975	0.666	0.657	456

All genome-feature values were log10- transformed prior to analysis (see Methods). Symbology, **ln ML**: ln Max Likelihood; **AIC**: Akaike's Information Criterion; **CI**: Confidence Interval of models predicting genome traits with 95% cumulative AIC weight. Detailed results from the regressions performed with alternative tree topologies and assembled genome sizes are provided in Supplementary Table S4.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2. Correlative associations among several features measuring genome complexity and intron-richness from 461 eukaryotes under non-phylogenetic (OLS) and phylogenetic (PIC, PGLS) models.

Model type	**OLS**				**PIC**					**PGLS**						**ASS**	**NCBI†**	**NCBI‡**	**Protein**
Predictor variable	r^2	slope	ln ML	AIC	r^2	slope	ln ML	AIC	**Log BF**	r^2	slope	ln ML	AIC	λ	**Log BF**	r^2	r^2	r^2	r^2
Regressions vs genome size																			
– CDS number	**0.417**	0.220	91.7	-177.3	**0.273**	0.287	-834.2	1672.4	**20.6**	**0.328**	0.307	244.3	-484.6	0.864	**20.3**	**0.333**	**0.327**	**0.293**	**0.193**
– CDS avg size	**0.683**	0.540	-69.4	144.8	**0.382**	0.403	-876.0	1755.9	**43.9**	**0.263**	0.295	189.4	-374.9	0.850	**43.8**	**0.275**	**0.263**	**0.282**	**0.253**
– CDS genome (Mbs)	**0.885**	0.763	64.7	-123.4	**0.492**	0.693	-1022.5	2048.9	**98.7**	**0.617**	0.620	129.9	-255.9	0.455	**98.5**	**0.624**	**0.601**	**0.589**	**0.435**
– Repeat genome (Mbs)	**0.892**	1.276	-155.8	317.7	**0.614**	1.416	-1236.9	2477.8	**1273.8**	**0.723**	1.378	-91.0	185.9	0.667	**1273.3**	**0.756**	**0.721**	**0.717**	**0.677**
– Unique ncDNA (Mbs)	**0.964**	1.071	195.1	-384.2	**0.812**	0.898	-797.8	1599.7	**449.5**	**0.859**	0.958	295.9	-587.8	0.829	**449.2**	**0.887**	**0.865**	**0.854**	**0.825**
– Exon number	**0.652**	0.478	-45.4	96.7	**0.295**	0.385	-944.6	1893.2	**11.3**	**0.316**	0.427	98.8	-193.6	0.931	**11.3**	**0.316**	**0.325**	**0.292**	**0.179**
– Exon density	**0.513**	0.256	109.3	-212.6	**0.044**	0.087	-757.5	1519.0	**0.2**	**0.054**	0.101	278.5	-552.9	0.936	**0.05**	**0.053**	**0.068**	**0.059**	**0.007**
– Exon avg size	**0.514**	-0.246	129.9	-253.7	**0.102**	-0.106	-647.3	1298.6	**0.3**	**0.096**	-0.114	341.1	-678.2	0.986	**0.2**	**0.097**	**0.136**	**0.106**	**0.051**
– Exon content (Mbs)	**0.496**	0.219	166.3	-326.6	**0.226**	0.260	-846.3	1696.5	**6.1**	**0.262**	0.258	235.4	-466.9	0.746	**6.2**	**0.268**	**0.256**	**0.227**	**0.113**
– Exon repeat (Mbs)	**0.188**	0.325	-348.0	702.0	**0.117**	0.494	-1321.3	2646.7	**1.0**	**0.135**	0.457	-210.3	424.5	0.779	**0.9**	**0.142**	**0.132**	**0.116**	**0.059**
– Unique exon (Mbs)	**0.540**	0.212	222.8	-439.6	**0.220**	0.233	-804.6	1613.3	**21.6**	**0.245**	0.230	272.6	-541.2	0.787	**21.4**	**0.248**	**0.242**	**0.220**	**0.114**
– CDS w/introns (%)	**0.136**	0.224	-263.5	533.1	**0.001**∗	0.049	-1104.3	2212.7	**1.8**	**0.022**	0.117	5.0	-6.0	0.945	**1.7**	**0.021**	**0.038**	**0.022**	**0.004**∗
– CDS w/introns (#)	**0.320**	0.444	-328.3	662.6	**0.102**	0.336	-1178.2	2360.4	**35.6**	**0.175**	0.421	-84.1	172.3	0.942	**35.7**	**0.187**	**0.150**	**0.153**	**0.079**
– Intron number	**0.451**	0.704	-413.2	832.4	**0.128**	0.449	-1254.1	2512.1	**7.8**	**0.184**	0.529	-176.3	356.5	0.948	**7.8**	**0.186**	**0.165**	**0.166**	**0.078**
– Intron density	**0.540**	0.259	130.0	-254.1	**0.061**	0.100	-748.5	1501.0	**1.0**	**0.068**	0.107	303.0	-601.9	0.916	**1.3**	**0.070**	**0.074**	**0.067**	**0.009**
– Intron w-avg size	**0.757**	0.708	-109.4	224.8	**0.478**	0.592	-962.2	1928.5	**80.6**	**0.382**	0.473	101.0	-198.0	0.869	**80.1**	**0.408**	**0.386**	**0.386**	**0.371**
– Intron content (Mbs)	**0.793**	1.377	-368.7	743.4	**0.452**	1.011	-1233.5	2471.0	**95.3**	**0.447**	0.994	-174.3	352.5	0.933	**95.1**	**0.462**	**0.372**	**0.420**	**0.339**
– Intron repeat (Mbs)	**0.811**	1.629	-411.2	828.3	**0.448**	1.537	-1408.8	2821.6	**119.9**	**0.485**	1.410	-303.6	611.3	0.813	**119.6**	**0.500**	**0.422**	**0.483**	**0.375**
– Unique intron (Mbs)	**0.776**	1.322	-372.4	750.8	**0.384**	0.886	-1236.5	2477.1	**50.1**	**0.389**	0.889	-177.5	359.0	0.938	**49.9**	**0.400**	**0.328**	**0.364**	**0.275**
Regressions vs intron content																			
– Intron number	**0.777**	0.597	-205.3	416.6	**0.639**	0.663	-1051.3	2106.6	**44.9**	**0.713**	0.706	56.2	-108.4	0.905	**44.8**	**0.713**	**0.811**	**0.716**	**0.627**
– Intron w-avg size	**0.645**	0.423	-196.7	399.4	**0.365**	0.344	-1007.5	2019.1	**595.3**	**0.336**	0.302	82.7	-161.5	0.916	**593.9**	**0.336**	**0.345**	**0.337**	**0.262**
– Intron density	**0.759**	0.198	279.3	-552.6	**0.303**	0.147	-680.1	1364.2	**10.1**	**0.392**	0.169	399.4	-794.8	0.881	**10.3**	**0.392**	**0.404**	**0.389**	**0.232**
– Repeat intron (Mbs)	**0.927**	0.771	-93.9	193.8	**0.700**	0.514	-1047.2	2098.5	**806.4**	**0.738**	0.582	31.2	-58.4	0.885	**804.6**	**0.738**	**0.751**	**0.718**	**0.643**
– Repeat genome (Mbs)	**0.682**	0.722	-405.0	816.0	**0.282**	0.639	-1380.0	2764.0	**65.8**	**0.314**	0.622	-284.6	573.2	0.851	**65.7**	**0.314**	**0.243**	**0.292**	**0.246**
– CDS w/introns (%)	**0.473**	0.269	-149.5	305.1	**0.328**	0.321	-1013.3	2030.6	**0.2**	**0.441**	0.341	127.8	-251.6	0.908	**0.06**	**0.441**	**0.630**	**0.459**	**0.460**
– CDS w/introns (#)	**0.621**	0.399	-193.8	393.6	**0.556**	0.517	-1016.4	2036.9	**29.7**	**0.625**	0.537	90.4	-176.7	0.905	**29.7**	**0.625**	**0.754**	**0.625**	**0.575**
Regressions vs intron size																			
– Intron number	**0.186**	0.557	-504.2	1014.4	**-0.001**∗	0.059	-1285.6	2575.3	**1.5**	**0.003**∗	0.115	-221.6	447.1	0.960	**1.6**	**0.003**∗	**0.004**∗	**0.005**∗	**0.004**∗
– Intron density	**0.407**	0.276	71.4	-136.8	**0.026**	0.079	-756.9	1517.9	**0.07**	**0.025**	0.085	292.9	-581.8	0.937	**0.2**	**0.024**	**0.021**	**0.029**	**-0.002**∗
– Repeat intron (Mbs)	**0.658**	0.368	-184.9	375.7	**0.368**	0.228	-991.3	1986.6	**589.5**	**0.366**	0.228	98.0	-192.0	0.906	**588.2**	**0.366**	**0.299**	**0.408**	**0.350**
– Repeat genome (Mbs)	**0.611**	1.298	-451.5	908.9	**0.342**	1.238	-1359.9	2723.8	**56.8**	**0.292**	1.182	-290.8	585.5	0.900	**56.4**	**0.292**	**0.278**	**0.284**	**0.291**
– CDS w/introns (%)	**0.038**	0.149	-288.4	582.8	**0.024**	-0.16	-1099.0	2201.9	**1.6**	**0.002**∗	-0.062	0.5	3.1	0.950	**1.6**	**0.002**∗	**-0.002**∗	**-0.0001**∗	**0.015**
– CDS w/introns (#)	**0.083**	0.281	-397.3	800.5	**-0.002**	-0.010	-1203.4	2410.8	**2.7**	**-0.001**∗	0.036	-128.1	260.2	0.955	**2.7**	**-0.001**∗	**-0.002**∗	**-0.001**∗	**0.007**
Regressions vs intron density																			
– Intron number	**0.676**	2.448	-291.6	589.2	**0.351**	1.849	-1186.0	2376.0	**162.0**	**0.437**	2.012	-92.5	188.9	0.936	**161.5**	**0.437**	**0.436**	**0.416**	**0.328**
– Repeat intron (Mbs)	**0.665**	4.228	-542.5	1091.1	**0.160**	2.270	-1504.3	3012.7	**0.5**	**0.210**	2.285	-400.2	804.5	0.837	**0.5**	**0.210**	**0.219**	**0.198**	**0.073**
– Repeat genome (Mbs)	**0.443**	2.560	-534.1	1074.2	**0.022**	0.706	-1450.9	2905.7	**3.2**	**0.032**	0.761	-363.0	730.1	0.878	**3.1**	**0.032**	**0.036**	**0.030**	**0.001**∗
– CDS w/introns (%)	**0.345**	1.009	-199.7	405.4	**0.139**	0.789	-1070.2	2144.4	**152.2**	**0.199**	0.836	50.3	-96.6	0.937	**152.0**	**0.199**	**0.232**	**0.165**	**0.163**
– CDS w/introns (#)	**0.425**	0.294	78.7	-151.3	**0.113**	0.882	-1175.3	2355.0	**6.3**	**0.194**	0.175	333.6	-663.3	0.871	**6.1**	**0.194**	**0.230**	**0.188**	**0.116**
– CDS number	**0.210**	0.445	21.6	-37.2	**0.002**∗	0.093	-906.9	1817.9	**0.2**	**0.024**	0.214	159.8	-315.5	0.896	**0.074**	**0.024**	**0.023**	**0.019**	**-0.002**∗
Regressions vs CDS w/introns (%)																			
– Intron number	**0.773**	1.526	-210.2	426.4	**0.696**	1.237	-1011.8	2027.6	**43.0**	**0.705**	1.369	52.1	-100.2	0.914	**42.9**	**0.705**	**0.830**	**0.713**	**0.749**
– Intron repeat (Mbs)	**0.323**	2.139	-703.1	1412.2	**0.171**	1.557	-1501.3	3006.6	**0.1**	**0.253**	1.651	-387.1	778.2	0.843	**0.2**	**0.253**	**0.275**	**0.188**	**0.123**
– Repeat genome (Mbs)	**0.115**	0.766	-641.0	1288.0	**-0.002**∗	0.053	-1456.4	2916.8	**2.9**	**0.018**	0.306	-366.3	736.6	0.879	**2.9**	**0.018**	**0.022**	**0.020**	**0.002**∗
– CDS number	**0.107**	0.186	-6.7	19.3	**0.005**∗	0.055	-906.4	1816.7	**0.2**	**0.033**	0.131	161.8	-319.6	0.893	**0.2**	**0.033**	**0.030**	**0.021**	**0.001**
Regressions vs exon size																			
– Exon number	**0.787**	-1.533	67.8	-129.7	**0.366**	-1.305	-920.2	1844.4	**25.0**	**0.517**	-1.305	154.9	-305.8	0.759	**10.5**	**0.517**	**0.516**	**0.478**	**0.355**
– Exon density	**0.883**	-0.982	437.5	-869.1	**0.573**	-0.933	-572.3	1148.6	**97.0**	**0.731**	-0.880	548.2	-1092.3	0.743	**97.5**	**0.731**	**0.752**	**0.737**	**0.623**
– Exon genome (Mbs)	**0.326**	-0.519	99.3	-192.6	**0.023**	-0.261	-900.0	1804.1	**9.6**	**0.061**	-0.312	181.2	-358.4	0.775	**10.2**	**0.061**	**0.050**	**0.038**	**0.001**∗
– Repeat exon (Mbs)	**0.154**	-0.861	-357.4	720.7	**0.010**	-0.474	-1347.7	2699.4	**1.5**	**0.034**	-0.564	-236.3	476.6	0.744	**2.3**	**0.034**	**0.028**	**0.022**	**0.001**∗
– Repeat genome (Mbs)	**0.467**	-2.701	-524.1	1054.1	**0.061**	-1.381	-1441.6	2887.1	**0.2**	**0.094**	-1.269	-348.6	701.3	0.855	**0.1**	**0.094**	**0.099**	**0.091**	**0.039**
– CDS avg size	**0.365**	-1.155	-229.7	465.5	**0.063**	-0.505	-971.8	1947.6	**0.9**	**0.050**	-0.327	131.2	-258.4	0.875	**-0.01**	**0.050**	**0.061**	**0.054**	**0.018**
– CDS number	**0.317**	-0.560	55.0	-104.1	**0.052**	-0.389	-895.2	1794.4	**13.6**	**0.097**	-0.414	175.4	-346.8	0.853	**0.6**	**0.097**	**0.083**	**0.068**	**0.028**
– CDS genome (Mbs)	**0.533**	-1.732	-258.5	522.9	**0.093**	-0.927	-1155.9	2315.9	**2.1**	**0.151**	-0.793	-19.1	42.3	0.800	**1.9**	**0.151**	**0.137**	**0.135**	**0.049**

Note: All genome-feature values were log10- transformed prior to analysis (see Methods). Only detailed results are shown for the PGLS and PIC regressions performed with the "reference tree topology" (see Figure 1a) and genome contents (in megabases) from estimated genome sizes. Symbology, r^2: coefficients of determination, Log BF: log Bayes Factors values, ln ML: ln Max Likelihood; AIC: Akaike's Information Criterion. Only $p > 0.05$ values are shown with asterisks, the remaining p-values for r^2 have statistical significance: < 0.001. Log BF significance: weak (< 2), positive evidence (> 2), strong evidence ($5 - 10$), very strong evidence (> 10). Only r^2 values are shown for the PGLS regressions estimated with assembled genome sizes (ASS) and alternate trees: NCBI taxonomy-based trees, one with no polytomies (NCBI†), while another one with polytomies (NCBI‡), and a protein domain content-based tree (Protein). Detailed information from these regressions are provided in Supplementary Tables S5-S9.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3. Summary statistics and PGLS regressions performed between genome size and intron features for different species datasets.

Genome		**GENOME:**		**INTRONS:**					
features	n	**size** (EST vs ASS)	**repeats**	**% in CDS**	**density**	**size**	**content**	**repeats**	**number**
Taxon		r^2 [Mbs]	[%]	r^2 [%]	r^2 [no.]	r^2 [nts]	r^2 [%]	r^2 [%]	r^2 [no.]
Eukarya	461	**0.958** [—]	—	**0.022** [—]	**0.068** [—]	**0.382** [—]	**0.447** [—]	**0.485** [—]	**0.184** [—]
Random and selected datasets									
Random set 1⊕	231	**0.986** [—]	—	**0.022** [—]	**0.086** [—]	**0.343** [—]	**0.424** [—]	**0.564** [—]	**0.184** [—]
Random set 2⊕	116	**0.987** [—]	—	**0.048** [—]	**0.162** [—]	**0.209** [—]	**0.526** [—]	**0.382** [—]	**0.259** [—]
Random set 3⊕	58	**0.995** [—]	—	**0.197** [—]	**0.505** [—]	**0.687** [—]	**0.899** [—]	**0.723** [—]	**0.470** [—]
Random set 4⊕	29	**0.992** [—]	—	**0.172** [—]	**0.481** [—]	**0.808** [—]	**0.904** [—]	**0.728** [—]	**0.291** [—]
Lynch & Conery 2003†	26	—	—	**0.189** [—]	**0.102**∗ [—]	**0.683** [—]	**0.549** [—]	**0.489** [—]	**0.300** [—]
Wu & Hurst 2015‡	30	—	—	**0.214** [—]	**0.713** [—]	**0.850** [—]	**0.932** [—]	**0.895** [—]	**0.412** [—]
Local phylogenetic scale									
Protists	66	**0.961** [83.6]	21.2	**0.024**∗ [56.6]	**0.181** [3.5]	**0.271** [227.3]	**0.490** [9.6] {10.3}	**0.475** [18.2]	**0.267** [42,957]
Stramenopiles	17	**0.982** [86.0]	26.7	**-0.062**∗ [59.3]	**-0.040**∗ [2.9]	**0.003**∗ [227.8]	**0.397** [7.5] {7.7}	**0.580** [19.8]	**0.330** [29,148]
Alveolata	20	**0.977** [116.1]	12.1	**0.248** [63.4]	**0.697** [4.5]	**0.344** [193.4]	**0.803** [11.1] {13.0}	**0.704** [14.6]	**0.481** [68,920]
Fungi	131	**0.985** [38.2]	13.6	**0.080** [65.9]	**0.034** [3.2]	**0.089** [120.4]	**0.285** [7.0] {7.0}	**0.528** [8.8]	**0.220** [32,768]
Basidiomycota	53	**0.993** [47.3]	17.0	**-0.001**∗ [81.8]	**-0.020**∗ [4.7]	**0.426** [92.2]	**0.598** [10.4] {10.3}	**0.694** [11.2]	**0.569** [57,516]
Ascomycota:	66	**0.964** [30.8]	10.2	**0.489** [53.4]	**0.218** [1.9]	**-0.009**∗ [141.4]	**0.641** [4.0] {4.0}	**0.633**∗ [6.2]	**0.193** [13,311]
– Pezizomycotina	42	**0.982** [40.6]	13.4	**0.018**∗ [72.8]	**0.147** [2.3]	**0.044**∗ [107.2]	**0.107** [5.2] {5.2}	**0.182** [7.2]	**-0.012**∗ [19,740]
– Saccharomycotina	22	**0.659** [13.7]	4.6	**0.092**∗ [16.3]	**0.145** [1.3]	**0.001**∗ [211.7]	**0.248** [1.7] {1.7}	**0.361** [4.5]	**0.133**∗ [1,661]
Chlorophyta	10	**0.953** [49.3]	11.2	**0.621** [64.7]	**0.807** [4.7]	**0.629** [245.8]	**0.939** [19.3] {19.9}	**0.806** [15.2]	**0.879** [46,269]
Streptophyta	68	**0.967** [1,065.8]	41.6	**-0.015**∗ [74.9]	**0.137** [4.9]	**0.415** [562.6]	**0.435** [11.8] {13.6}	**0.299** [20.2]	**0.104** [134,777]
Monocots	15	**0.978** [1,809.8]	49.0	**-0.024**∗ [75.3]	**-0.044**∗ [4.8]	**-0.040**∗ [810.9]	**-0.035**∗ [9.4] {11.4}	**-0.077**∗ [19.5]	**-0.028**∗ [129,949]
Eudicots	49	**0.952** [641.1]	43.8	**-0.018**∗ [76.5]	**-0.016**∗ [5.0]	**0.442** [495.6]	**0.537** [11.3] {13.5}	**0.484** [20.4]	**0.214** [155,964]
Metazoa	186	**0.973** [1,260.9]	26.5	**0.048** [87.0]	**-0.001**∗ [7.5]	**0.683** [2,643.5]	**0.678** [24.6] {26.4}	**0.690** [24.2]	**0.077** [120,238]
Protostomia:	78	**0.957** [517.9]	25.7	**0.138** [84.4]	**-0.011**∗ [5.2]	**0.659** [1,535.0]	**0.606** [21.7] {22.8}	**0.693** [24.0]	**0.104** [75,745]
–Lophotrochozoa	10	**0.914** [765.6]	35.8	**0.119**∗ [79.1]	**0.293**∗ [6.2]	**0.319**∗ [1,671.4]	**0.577** [21.5] {23.6}	**0.502** [33.1]	**-0.125**∗ [126,857]
–Arthropoda:	61	**0.948** [523.3]	25.4	**0.053** [84.3]	**-0.013**∗ [4.9]	**0.684** [1,640.0]	**0.595** [21.0] {22.1}	**0.657** [23.8]	**0.080** [64,879]
– Diptera	13	**0.938** [410.1]	30.6	**-0.090**∗ [84.0]	**0.186**∗ [3.5]	**0.738** [1,484.9]	**0.482** [14.6] {16.9}	**0.754** [28.8]	**-0.090**∗ [40,508]
– Hymenoptera	18	**0.646** [278.6]	18.0	**0.154**∗ [85.4]	**-0.058**∗ [5.3]	**-0.017**∗ [1,259.4]	**-0.043**∗ [25.4] {27.8}	**-0.009**∗ [15.2]	**-0.039** [59,118]
– Lepidoptera	13	**0.975** [402.0]	29.2	**-0.089**∗ [86.5]	**0.032**∗ [5.6]	**0.346** [1,377.7]	**0.044**∗ [23.9] {23.9}	**0.412** [30.4]	**-0.038**∗ [80,139]
Deuterostomia:	101	**0.955** [1,898.6]	26.6	**0.083** [90.1]	**0.077** [9.3]	**0.729** [3,623.5]	**0.728** [26.8] {29.4}	**0.614** [23.9]	**0.044** [156,583]
–Teleostei	17	**0.928** [871.6]	20.8	**0.324** [94.2]	**-0.009**∗ [9.7]	**0.856** [1,662.4]	**0.844** [31.5] {35.1}	**0.922** [19.2]	**0.348** [200,011]
–Amniota:	72	**0.567** [2,234.2]	26.4	**0.341** [89.6]	**0.099** [9.5]	**0.184** [4,321.7]	**0.667** [26.3] {28.8}	**0.476** [23.0]	**0.112** [149,821]
– Reptiles	11	**0.806** [2,188.3]	33.9	**-0.068**∗ [89.8]	**0.036**∗ [8.8]	**0.363** [4,982.6]	**0.584** [26.5] {28.2}	**0.329** [30.6]	**-0.052**∗ [143,335]
– Aves	28	**-0.029**∗ [1,231.1]	8.5	**0.032**∗ [92.9]	**-0.036**∗ [10.1]	**-0.028**∗ [3,334.5]	**-0.023**∗ [31.5] {33.6}	**-0.028**∗ [7.0]	**0.032**∗ [139,985]
– Mammalia	33	**0.482** [3,100.6]	39.0	**-0.032**∗ [86.8]	**-0.020**∗ [9.3]	**0.092** [4,938.9]	**0.169** [21.9] {24.9}	**0.377** [34.1]	**-0.007**∗ [160,330]

Note: ⊕ Randomly reduced datasets from the original 461 species analyzed in this study, with 100 replicates each (see Methods). † Complete genome sequences are only available for 26 from the 32 metazoan species included in the original dataset (see Supplementary Table S10). ‡ This dataset has an overrepresentation of metazoan species (around two thirds), particularly of vertebrates (see Supplementary Table S10). All genome-feature values were log10- transformed prior to analysis (see Methods). The [mean value] for selected genome features is shown, and for genome contents (in % or Mbs) as calculated from estimated genome sizes. The r^2 values (highlighted in bold font) obtained from PGLS regressions were performed with estimated genome sizes and the "reference tree topology" (in Figure 1a). Unique or non-repetitive intron contents are denoted with {}. Only $p > 0.05$ values are shown with asterisks, the remaining p-values for r^2 have statistical significance: < 0.001. The regressions performed with alternative tree topologies and assembled genome sizes are provided in Supplementary Table S12. Summary statistics for additional clades are provided in Supplementary Tables S14-S15.

Table 4. PGLS regressions performed among intron features for different taxa datasets.

X-*Y* traits	*n*	**IS-ID**	**IS-CDSI**	**IS-IC**	**IS-IR**	**IS-IN**	**ID-CDSI**	**ID-IC**	**ID-IR**	**ID-IN**	**IC-CDSI**	**IC-IR**	**IC-IN**	**CDSI-IR**	**CDSI-IN**	**IR-IN**
Taxon		r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2
Eukarya	461	0.025	0.002∗	0.336	0.366	0.003∗	0.199	0.392	0.210	0.437	0.441	0.738	0.713	0.253	0.705	0.468
Protists	66	0.030∗	0.001∗	0.181	0.102	-0.015∗	0.385	0.574	0.506	0.357	0.776	0.846	0.772	0.440	0.734	0.734
Stramenopiles	17	-0.065∗	0.119∗	0.383	0.130∗	-0.066∗	0.428	0.258	-0.029∗	0.097	-0.021∗	0.719	0.570	-0.035∗	0.195	0.350
Alveolata	20	0.432	0.027∗	0.436	0.464	-0.027∗	0.409	0.876	0.801	0.425	0.612	0.937	0.712	0.607	0.741	0.681
Fungi	131	0.031	0.025	0.008∗	0.164	0.014∗	0.215	0.303	0.124	0.212	0.768	0.550	0.910	0.269	0.883	0.515
Basidiomycota	53	0.082	0.023∗	0.099	0.443	0.012∗	0.838	0.604	0.016∗	0.445	0.588	0.560	0.889	0.004∗	0.229	0.659
Ascomycota:	66	0.267	0.291	0.004∗	0.125	0.446	0.265	0.408	0.296	0.311	0.834	0.554	0.950	0.401	0.924	0.678
– Pezizomycotina	42	-0.014∗	-0.010∗	-0.025∗	0.290	0.129	0.357	0.354	0.080	0.114	0.619	0.039∗	0.804	-0.022∗	0.580	-0.020∗
– Saccharomycotina	22	0.033∗	0.100∗	-0.046∗	-0.002∗	0.091	0.471	0.543	0.542	0.525	0.870	0.843	0.893	0.730	0.993	0.759
Chlorophyta	10	0.597	0.175∗	0.632	0.350	0.352	0.752	0.942	0.559	0.677	0.830	0.739	0.863	0.597	0.854	0.844
Streptophyta	68	0.044	-0.015∗	0.767	0.627	0.004∗	0.047	0.056	0.099	0.006∗	0.006∗	0.834	0.254	-0.010∗	0.107	0.220
Monocots	15	-0.030∗	0.142∗	0.728	0.653	0.112∗	0.134∗	-0.056∗	-0.010∗	0.340	-0.035∗	0.908	-0.074∗	-0.003∗	-0.077∗	-0.040∗
Eudicots	49	-0.019∗	-0.009∗	0.689	0.586	0.022∗	0.284	-0.021∗	0.006∗	-0.019∗	0.051∗	0.884	0.368	-0.008∗	0.136	0.334
Metazoa	186	0.044	0.014∗	0.831	0.681	-0.004∗	0.087	0.154	0.042	-0.005∗	-0.002∗	0.754	0.107	0.010∗	0.022	0.095
Protostomia:	78	0.025∗	0.025∗	0.867	0.679	-0.010∗	0.034∗	0.167	0.059	0.003∗	0.011∗	0.772	0.047	0.001∗	0.013	0.059
–Lophotrochozoa	10	0.747	0.114∗	0.849	0.918	0.551	0.070∗	0.830	0.769	0.453	0.074∗	0.929	0.193∗	0.116∗	-0.106∗	0.434
–Arthropoda:	61	0.019∗	-0.014∗	0.858	0.723	-0.0001∗	0.102	0.170	0.075	0.011∗	-0.017∗	0.797	0.141	-0.016∗	0.056	0.121
– Diptera	13	-0.075∗	0.009∗	0.775	0.780	-0.032∗	-0.049∗	-0.040∗	-0.071∗	0.120∗	-0.015∗	0.791	-0.081∗	-0.069∗	-0.064∗	-0.081∗
– Hymenoptera	18	0.533	0.007∗	0.864	0.284	0.365	0.166∗	0.629	-0.053∗	0.430	0.117∗	0.321	0.156∗	0.015∗	0.072∗	-0.056∗
– Lepidoptera	13	-0.046∗	-0.010∗	0.619	0.682	0.343	-0.082∗	0.329	-0.019∗	0.186∗	-0.066∗	0.611	0.462	-0.089∗	-0.058∗	0.094∗
Deuterostomia:	101	0.027∗	0.049	0.841	0.730	0.028∗	0.205	0.101	0.006∗	0.063	0.056	0.743	0.248	0.261	0.163	0.104
–Teleostei	17	0.130∗	0.612	0.941	0.919	0.517	0.360	0.016∗	0.096∗	0.350	0.509	0.887	0.546	0.503	0.548	0.515
–Amniota:	72	0.0004∗	-0.013∗	0.809	0.408	-0.014∗	0.246	0.267	0.093	0.266	0.044	0.555	0.063	0.293	0.152	0.009∗
– Reptiles	11	-0.009∗	-0.111∗	0.295	0.211∗	-0.052∗	-0.091∗	0.455	0.380	-0.095∗	-0.107∗	0.695	-0.085∗	-0.084∗	0.117∗	-0.108∗
– Aves	28	0.176	0.032∗	0.616	0.178	0.058∗	0.386	0.447	0.153	0.430	0.193	0.182	-0.038∗	-0.035∗	0.369	0.013∗
– Mammalia	33	0.025∗	-0.018∗	0.782	0.629	0.017∗	0.150	0.267	0.055∗	0.218	-0.006∗	0.742	0.004∗	-0.032∗	-0.024∗	0.085∗

Note: All genome-feature values were log10- transformed prior to analysis (see Methods). PGLS regressions are performed with estimated genome sizes and the "reference tree topology" (in Figure 1a). Symbology, **IS**: weighted-average intron size (nts), **IN**: total number of introns, **ID**: (absolute) intron density, **CDSI**: fraction (%) of intron-containing CDSs, **IC**: intronic content of the genome in Mbs (as based on assembled sizes), **IR**: fraction (%) of repetitive content within introns. Only $p > 0.05$ values are shown with asterisks, the remaining p-values for r^2 have statistical significance: < 0.001. The regressions performed with alternative tree topologies and estimated genome sizes are provided in Supplementary Table S13.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 5. Summary statistics, mean value and (coefficient of variation), of several protein-coding features for different species datasets.

Genome			CDS:			EXONS:			
features	*n*	**Genome size**	**number**	**size**	**content**	**size**	**content**	**repeats**	**number**
Taxon		[Mbs]		[nts]	[%]	[nts]	[%]	[%]	
Protists	66	83.6 (2.23)	15,354 (0.90)	2,071 (0.66)	52.2 (0.39)	804.3 (0.51)	43.9 {44.3}(0.47)	11.8 (0.78)	58,467 (1.58)
Stramenopiles	17	86.0 (0.82)	16,190 (0.43)	1,803 (0.66)	41.4 (0.38)	687.5 (0.30)	32.9 {33.4}(0.44)	13.0 (0.62)	45,396 (0.62)
Alveolata	20	116.1 (2.82)	13,546 (0.90)	2,671 (0.75)	58.0 (0.37)	880.7 (0.52)	51.9 {51.7}(0.46)	8.1 (0.76)	82,731 (1.90)
Fungi	131	38.2 (0.61)	11,779 (0.44)	1,500 (0.13)	51.9 (0.27)	626.3 (0.59)	44.9 {45.1}(0.32)	6.4 (1.16)	44,567 (0.70)
Basidiomycota	53	47.3 (0.56)	14,626 (0.38)	1,545 (0.16)	51.9 (0.25)	373.9 (0.72)	41.5 {41.0}(0.30)	10.4 (0.82)	72,171 (0.37)
Ascomycota:	66	30.8 (0.59)	9,464 (0.38)	1,467 (0.08)	52.2 (0.29)	839.8 (0.38)	48.2 {48.7}(0.33)	2.3 (0.94)	22,788 (0.59)
– Pezizomycotina	42	40.6 (0.39)	11,519 (0.25)	1,488 (0.10)	45.4 (0.29)	623.9 (0.14)	40.3 {40.8}(0.30)	2.2 (1.10)	31,274 (0.28)
– Saccharomycotina	22	13.7 (0.29)	5,848 (0.14)	1,432 (0.04)	63.9 (0.18)	1,246.8 (0.16)	62.1 {62.8}(0.19)	2.5 (0.70)	7,520 (0.47)
Chlorophyta	10	49.3 (0.91)	10,349 (0.40)	2,603 (0.52)	67.9 (0.36)	610.7 (0.78)	47.7 {45.5}(0.72)	7.9 (0.63)	56,632 (0.93)
Streptophyta	68	1,065.8 (2.54)	35,359 (0.51)	3,053 (0.56)	27.3 (0.75)	274.7 (0.46)	13.8 {15.1}(1.34)	16.4 (0.82)	171,871 (0.55)
Monocots	15	1,809.8 (0.96)	37,055 (0.40)	3,619 (0.59)	15.0 (0.78)	385.3 (0.17)	5.2 {6.4}(0.91)	13.3 (0.82)	171,992 (0.30)
Eudicots	49	641.1 (1.23)	40,012 (0.43)	2,859 (0.49)	23.7 (0.42)	377.0 (0.11)	10.1 {12.3}(0.63)	19.5 (0.74)	197,128 (0.49)
Metazoa	186	1,261.0 (0.93)	18,413 (0.34)	17,906 (0.77)	30.6 (0.40)	287.7 (0.24)	5.2 {5.4}(1.13)	6.9 (0.89)	142,824 (0.43)
Protostomia:	78	517.9 (1.58)	17,266 (0.44)	7,223 (0.94)	30.8 (0.44)	322.3 (0.24)	7.8 {8.3}(0.72)	7.2 (0.80)	93,884 (0.53)
–Lophotrochozoa	10	765.6 (0.97)	26,623 (0.45)	8,316 (0.63)	29.6 (0.37)	308.3 (0.11)	6.8 {7.7}(0.94)	15.2 (0.38)	153,645 (0.39)
–Arthropoda:	61	523.3 (1.66)	15,608 (0.36)	7,534 (0.96)	29.2 (0.46)	340.3 (0.20)	7.0 {7.5}(0.66)	5.9 (0.80)	81,455 (0.44)
– Diptera	13	410.1 (0.66)	13,435 (0.22)	5,975 (0.64)	22.4 (0.40)	416.7 (0.17)	6.9 {7.9}(0.71)	4.4 (0.50)	54,745 (0.33)
– Hymenoptera	18	278.6 (0.27)	13,584 (0.25)	7,138 (0.48)	33.3 (0.36)	328.8 (0.16)	7.4 {7.9}(0.31)	5.2 (0.76)	74,930 (0.13)
– Lepidoptera	13	402.0 (0.37)	16,232 (0.25)	7,279 (0.30)	32.6 (0.51)	300.8 (0.14)	5.2 {5.2}(0.38)	4.0 (0.63)	96,678 (0.39)
Deuterostomia:	101	1,898.6 (0.55)	19,010 (0.24)	27,113 (0.42)	29.7 (0.28)	254.4 (0.15)	2.5 {2.8}(1.30)	5.7 (0.73)	182,585 (0.21)
–Teleostei	17	871.6 (0.39)	22,109 (0.10)	14,118 (0.34)	36.5 (0.16)	235.8 (0.14)	4.7 {5.2}(0.35)	5.1 (0.51)	235,557 (0.12)
–Amniota:	72	2,234.2 (0.43)	17,793 (0.16)	32,314 (0.25)	28.0 (0.25)	255.4 (0.12)	1.5 {1.6}(0.37)	5.0 (0.67)	173,398 (0.14)
– Reptiles	11	2,188.3 (0.24)	18,194 (0.13)	33,886 (0.29)	28.0 (0.16)	272.9 (0.08)	1.4 {1.4}(0.30)	8.4 (0.35)	163,293 (0.16)
– Aves	28	1,231.1 (0.11)	15,031 (0.09)	27,634 (0.16)	33.7 (0.15)	232.4 (0.05)	2.1 {2.2}(0.11)	2.2 (0.38)	157,084 (0.06)
– Mammalia	33	3,100.6 (0.16)	20,003 (0.09)	35,762 (0.23)	23.2 (0.23)	269.0 (0.12)	1.0 {1.2}(0.18)	6.3 (0.47)	190,608 (0.11)

Note: Genome contents provided in % or Mbs are calculated from assembled genome sizes. Unique or non-repetitive exon contents are denoted with {}. Summary statistics for additional clades and for genome contents based on estimated genome sizes are provided in Supplementary Tables S14-S15.

bioRxiv preprint doi: https://doi.org/10.1101/283549; this version posted March 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 6. Contribution (%) of 7 intron features, 15 genomic traits and cellular complexity to the first seven principal components as estimated with the phylogenetic Principal Component Analyses (phyloPCA).

Intron features only	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Cumulative % of var	**68.09**	**19.07**	**7.04**	3.22	2.2	0.24	0.15
Intron number	**16.935**	10.744	0.666	4.684	18.121	41.191	7.659
CDS with introns (%)	11.568	**19.174**	**20.104**	25.391	23.682	0.026	0.055
Intron density	11.763	6.245	**67.774**	0.539	13.618	0.059	0.001
Intron weighted-mean size	4.864	**54.072**	0.803	14.782	3.767	16.261	5.451
Intron content	**19.986**	2.249	0.062	0.543	5.538	0.042	71.58
Unique intron genome	**19.665**	0.775	0.218	4.937	20.867	39.104	14.434
Intron repeats	**15.219**	6.741	10.373	49.123	14.406	3.317	0.82
15 traits	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Cumulative % of var	**50.18**	**14.39**	**12.28**	6.2	5.54	4.27	3.2
Cellularity type	1.250	0	0.020	95.603	0.368	0.600	1.848
Number CDS	6.842	**12.412**	6.212	0.001	0.001	1.717	2.969
CDS with introns (%)	3.316	**15.560**	7.007	0.281	5.366	12.077	30.978
Genome size	**9.195**	0.980	9.305	0.030	0.472	8.715	6.303
Repeat genome content	7.262	1.852	10.440	0.138	12.259	5.600	0.684
Unique nc-genome content	**9.476**	0.009	6.467	0.012	2.063	9.464	4.721
Intron content	**11.637**	2.943	0.007	0.438	0.902	6.417	0.004
Unique intron genome	**10.966**	4.039	0.143	0.346	3.140	4.793	0.046
Intron repeat content	**10.253**	0.312	1.742	0.387	8.792	7.684	0.439
Intron density	4.597	**15.291**	5.384	0.290	1.476	5.013	26.913
Intron weighted-mean size	3.757	0.075	**27.306**	1.682	7.321	10.892	9.427
Exon average size	4.796	**14.971**	4.033	0.067	2.436	24.303	1.409
Exon content	6.528	**12.239**	9.952	0.381	3.547	0.294	0.356
Unique exon content	6.155	8.347	10.293	0.156	15.378	0.197	0.058
Exon repeat content	3.971	**10.971**	1.689	0.188	36.480	2.236	13.845

Note: Since all genome-feature values were log10- transformed prior to analysis, phyloPCAs were performed with 457 species (see Methods). Only results from the phyloPCAs performed with the reference tree topology (Figure 1a) and genome features based on assembled genome sizes are shown here. Results for the phyloPCAs performed with other tree topologies and estimated genome sizes, as well as from the comparative PCA, are provided in Supplementary Tables S13-S14. The noticeable contribution of some variables for the first three PCA components is highlighted with bold font.