Title
Introns and alternative splicing in choanoflagellates

Author
WESTBROOK, MARJORIE WRIGHT

Publication Date
2011

Peer reviewed|Thesis/dissertation

Introns and alternative splicing in choanoflagellates

By

Marjorie Wright Westbrook

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicole King
Professor Donald C. Rio
Professor G. Steven Martin
Professor Steven E. Brenner

Fall 2011

Introns and alternative splicing in choanoflagellates

by

Marjorie Wright Westbrook

Abstract

Introns and alternative splicing in choanoflagellates

By

Marjorie Wright Westbrook

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Nicole King, Chair

The first organisms to evolve were unicellular, and the vast majority of life has remained so for billions of years. Complex forms of multicellularity, requiring increased levels of cell adhesion, cell signaling and gene regulation, have evolved in only a few eukaryotic lineages [1, 2]. The comparison of genomes from choanoflagellates, the closest relatives of metazoans, with genomes from metazoans may reveal genomic changes underlying metazoan origins. I used this approach to investigate the evolution of introns during the origin of metazoans.

By analyzing the genome of the first choanoflagellate to be sequenced, *Monosiga brevicollis*, I found that its intron density rivals that of genes in intron-rich metazoans [3]. Many intron positions are conserved between choanoflagellates and metazoans, implying that their shared unicellular ancestor was also intron-rich. In my analysis of the *M. brevicollis* genome, I made the unexpected discovery that, unlike most choanoflagellate genes, the longest genes contain relatively few introns. Indeed, one *M. brevicollis* gene contains the longest stretch of intron-free coding sequence known to date. I also found a similar trend in the genome of a basal metazoan, the sponge *A. queenslandica*. However, most long genes in other metazoans are not depleted of introns, revealing a difference in gene structure between eumetazoans and their closest relatives that may have implications for how these genes are regulated.

The results of these analyses led me to investigate the evolution of alternative splicing during the emergence of metazoans. Intron-rich metazoan genes undergo complex patterns of developmentally regulated alternative splicing. My analysis of intron evolution revealed that the unicellular ancestor of metazoans was also intron-rich, raising the possibility that alternative splicing was common before the transition to multicellularity. To test this, I used transcriptome sequencing to detect alternative splicing in choanoflagellates and the early branching metazoan, *Hydra magnipapillata*. I found that alternative splicing, especially the skipping of entire exons, occurs less frequently in choanoflagellates than in *H. magnipapillata*.

Increased alternative splicing of already intron-rich genes may thus represent an augmentation of gene regulation that evolved during the origin of metazoans.

My analyses suggest that metazoans evolved from an intron-rich unicellular ancestor, setting the stage for complex patterns of alternative splicing to evolve during the transition to multicellularity. The connection between gene structure and alternative splicing provides an example of how non-coding features of eukaryotic genomes can impact the evolution of regulatory and morphological complexity.

Acknowledgements

Throughout my graduate career I had excellent mentors and colleagues. The work presented in this dissertation would not have been possible without their advice and collaboration.

My advisor, Nicole King, provided guidance and encouragement from the conception of this project, and was continually involved and supportive even when it strayed from her area of expertise. I am thankful for her mentorship, as well as the great contribution she has made to the scientific community by bringing choanoflagellates into the arena of molecular biology.

My committee contributed valuable feedback on all aspects of this project. I am particularly grateful to Donald Rio and Steven Brenner for sharing their expertise on alternative splicing. Steven Brenner was kind enough to let me attend his lab's group meetings and many members of the Brenner lab group, particularly Angela Brooks and Liana Lareau, were generous with their time and analytical tools. Without them the analysis of alternative splicing in choanoflagellates would not have come to fruition.

I had many wonderful collaborators at Berkeley and abroad. Bernard Degnan and Claire Larroux at the University of Queensland shared their genomic data from sponge, which made my comparative analyses considerably more interesting. Similarly, Bridgette Gaillot and Yvan Wegner at the University of Geneva shared their transcriptome data from cnidarians, and Yvan also provided valuable advice on analyses. At Berkeley, Uffe Hellenstein and Jason Stajich shared their expertise on the comparative genomics of introns. Leath Tompkins at the QB3 Vincent J. Coates sequencing facility was immensely helpful in the generation of RNA-seq libraries. Justin Choi at the Functional Genomics lab provided technical advice on RNA purification and quantification. Sean Ruddy in the Statistics department developed the statistical methods used for quantifying intron retention and differential isoform abundance.

The King lab was a fantastic place to work, and I am thankful to its members for making it an intellectually stimulating and fun place to be a graduate student. I am particularly indebted to Susan Young, Stephen Fairclough and Daniel Richter who helped me overcome several technical hurdles. I also feel lucky to have been a part of the MCB entering class of 2005. My classmates were both great colleagues and friends.

Finally, I am grateful to all my friends and family for supporting me in so many ways over the past six years. I especially thank my parents, Robin and Reeves Westbrook, who have always encouraged my interest in science.

Table of Contents

Tables and Figures

Chapter 4

Chapter 1: The evolution of spliceosomal introns

SUMMARY

Spliceosomal introns, a distinguishing feature of eukaryotic genes, consist of stretches of non-coding sequences within genes that are first transcribed but then removed by the ribozymal spliceosome before translation [4]. Spliceosomal introns evolved early within the eukaryotic lineage and increased in number and size in some groups while all but disappearing from others [5]. What evolutionary forces were responsible for this dynamic evolutionary history? One hypothesis proposes that introns are adaptive and play a key role in gene regulation by enabling alternative splicing, and have thereby contributed to the evolution of morphologically complex eukaryotes [6, 7]. A contrasting argument is that introns are deleterious and evolved by non-adaptive means in certain lineages [8, 9]. Recent genomic data from diverse unicellular eukaryotes present the opportunity to test these hypotheses, and may illuminate the interplay between adaptive and non-adaptive forces in the evolution of eukaryotic genomes.

INTRODUCTION

Spliceosomal introns are a ubiquitous feature of eukaryotic genes; at least one intron has been found in every eukaryotic lineage studied to date [10]. The first inklings of their existence came in the 1970s, when researchers studying adenovirus transcription found that RNA transcripts hybridized to non-contiguous stretches of genomic DNA [11-13]. Soon afterward, electron microscopic studies of RNA:DNA hybrids revealed non-hybridized intronic sequences looping out of RNA:DNA pairs [14]. These findings came as a great surprise; why transcribe large stretches of DNA only to remove them before translation? Biologists quickly began to consider how and why introns evolved. Forty years and many hypotheses later, the evolutionary history of introns is still a matter of debate [5, 15, 16].

The origins of genes in pieces

The emergence of spliceosomal introns

One of the most basic aspects of intron evolution is the timing of their initial origin, for which two alternative scenarios have been proposed. The “introns-early” hypothesis suggested that introns were present in the last universal common ancestor of all life (LUCA). According to this hypothesis, introns were then lost in archaebacteria and eubacteria, perhaps in response to pressure for faster replication times, but maintained in the eukaryotic lineage [17-20]. The opposing “introns-late” hypothesis proposed that introns were gained in eukaryotes after the split from archaebacteria and eubacteria [21-23]. Since introns are only found in modern-day eukaryotes these hypotheses have been inherently difficult to test.

This debate was sparked by Walter Gilbert's prescient essay "Why genes in pieces?" in which he proposed functional and evolutionary roles for introns [24]. Among these was the idea that introns allowed exons to be re-arranged into new combinations, a process termed exon shuffling. To test the introns-early hypothesis, various studies searched for signs of exon shuffling in ancient genes. If shuffling had occurred, exon boundaries should coincide with boundaries of functional protein domains and introns should be positioned in between codons rather than interrupting them [17]. These predictions were born out; in eukaryotes some protein domains in ancient genes are contained in single exons and introns are biased to occur in between codons [25, 26]. Exon shuffling has now been established as an important mechanism of evolution in eukaryotic gene families [27-30].

However, several studies suggested this is not reflective of a deeply ancestral condition [31, 32]. An important study looked at intron positions in genes that duplicated before the three major domains of life diverged [33]. If introns were present before these genes duplicated, and thus in the LUCA, their positions should be conserved between the paralogs. In a set of 10 such genes intron positions were not conserved. Similar studies using different sets of paralogs have also failed to find ancient conservation of intron positions, leaving little support for the idea that introns were present before the divergence of eukaryotes, eubacteria and archaebacteria [34]. The "introns-early" has now been largely abandoned.

Introns and the origins of the eukaryotes

If introns are indeed a eukaryotic invention, how and when did they first appear? Modern day spliceosomal introns may have evolved from the rare self-splicing, or "group II", introns found in some eubacteria and eukaryotic organelles [35, 36]. Group II introns are ribozymes that, often in combination with self-encoded proteins, catalyze their own splicing [37, 38]. Like transposable elements, group II introns are mobile and can even move between different species of eubacteria [39, 40]. Further, the mechanism of group II self-splicing bears striking similarities to the removal of introns by the spliceosome in modern eukaryotes [41]. A plausible scenario for the origin of spliceosomal introns is that group II introns invaded the genome of an early eukaryote and then lost the ability to self-splice after the evolution of a separate spliceosome [35, 36, 38]. A potential source of these invading group II introns is the alpha-proteobacteria endosymbiont that evolved into the mitochondria [42].

Once present, spliceosomal introns would have drastically changed the process of eukaryotic gene expression, and it has been hypothesized that they even sparked the evolution of the nucleus [43]. In comparison to translation and transcription, splicing is a relatively slow process [44, 45]. To prevent immature, unspliced mRNAs from being translated, splicing had to be sequestered into its own compartment – the nucleus [43]. This idea has been difficult to test but remains a tantalizing possibility for the origin of a defining feature of the eukaryotic cell.

Population dynamics and intron evolution
Initial origins aside, intron abundance and distribution evolved rapidly during the diversification of eukaryotes [46, 47]. Certain eukaryotes, such as vertebrates, have many introns in their genomes while others, such as the yeast *Saccharomyces cerevisiae*, have hardly any introns. This early observation, based on the human and yeast genomes led to the prediction that intron abundance was a genomic signature of "higher" eukaryotes and might confer a selective advantage that contributed to the evolution of complexity in multicellular eukaryotes [7, 48].

An alternative explanation for the genome-wide differences in intron abundance comes from the perspective of population genetics. In response to adaptive hypotheses, Lynch and Connery have proposed that introns accumulated in multicellular eukaryotes independently of an adaptive function [8]. In contrast with adaptive hypotheses, Lynch and Connery assumed that introns are selectively disadvantageous because their splice sites increase the mutational load of a gene. In other words, the possibility that a splice site could be mutated and that an unspliced intron would render the gene nonfunctional makes introns a hazard, though a relatively slight one [8].

Under this scenario, Lynch and Connery constructed an explanation for the seeming connection between morphological complexity and intron abundance. Selectively disadvantageous variants can spread throughout a small population by chance, i.e. genetic drift, but are much less likely to do so in large populations [8, 49]. Therefore, introns would be expected to accumulate in groups of eukaryotes with small population sizes but not in those with large population sizes. A fairly reliable negative correlate of population size is organism size, and multicellular organisms generally have smaller population sizes than unicellular ones [49]. Therefore, Lynch and Connery argue that introns are common in multicellular eukaryotes because they increased in abundance when population sizes became smaller [8, 49].

Notably, this line of reasoning extends beyond intron abundance to other non-coding features of the genome, such as mobile genetic elements, intergenic spacing and gene duplications, all of which increase in size or frequency in multicellular eukaryotes [9]. This model thereby provided a potential unifying explanation for many aspects of genome architecture without invoking any adaptive arguments [50]. From this viewpoint, dramatic genome-scale differences could be purely the result of genetic drift – some junk DNA may really just be junk after all. At least in the case of introns, though, neither the Lynch and Connery nor adaptive scenarios are entirely consistent with recent genomic data from unicellular eukaryotes.

Insights from comparative genomics

A survey of eukaryotic gene structure: yeast misleads the way
S. cerevisiae was the first eukaryote to have its genome sequenced [51], and the scarcity of introns in its genome led to the generalization that "simple" eukaryotes are intron-poor while multicellular eukaryotes are intron-rich. Although the set of

currently available genomes is more phylogenetically diverse, the connection between gene structure and multicellularity has still been subject to biases. To assess generalizations about gene structure more comprehensively, I analyzed annotations of 54 eukaryotic genomes from all of the major groups of eukaryotes (Table 1.1, Table S1.1). To allow for a more even-handed survey, only a subset of representative genomes from the heavily sequenced fungi and metazoans were analyzed. While transcript length has remained relatively constant across eukaryotes, exon length, intron length and intron abundance vary by orders of magnitude, reflecting dramatic changes in the exon-intron structure of eukaryotic genes (Table 1.1).

I then investigated if intron abundance correlates with multicellularity (Figure 1.1, panel A). Multicellularity has arisen several times throughout evolution [1], and this analysis included five groups that evolved multicellularity independently (basidiomycetes, brown algae, volvocine algae, land plants, and metazoans), making this a paraphyletic grouping. Although unicellular eukaryotes are generally less intron-rich than multicellular eukaryotes, there are many exceptions. Notably, a handful of unicellular eukaryotes have intron densities rivaling those of the most intron-rich multicellular taxa. These include the green algae *Chlamydomonas reinhardtii* and the choanoflagellates *Salpingoeca rosetta* and *Monosiga brevicollis*, each of which have an average of approximately seven introns per gene.

There are enough exceptions to the generalization that unicellular eukaryotes are intron-poor that its utility is questionable. Aside from Excavata, there are intron-rich unicellular eukaryotes in every major group (Table 1.1). Perhaps a more appropriate generalization is that while there is a great deal of variation of intron number in unicellular eukaryotes, multicellular eukaryotes are more consistently intron-rich.

To investigate whether intron length might correlate with multicellularity, I compared the mean and median intron lengths in unicellular and multicellular lineages (Figure 1.1, panel B). While median intron length is relatively constant, some metazoans have considerably longer mean intron lengths than other species, indicating that metazoans have a class of extremely long introns not present in other groups. These observations are limited to bilaterians; the basal metazoans *Nematostella vectensis*, *Trichoplax adhaerens*, and *Amphimedon queenslandica* have mean intron lengths similar to non-metazoans.

The extremely long introns found exclusively in bilaterians are notable as there are several reasons why longer introns may be selectively disadvantageous. Under the Lynch and Connery hypothesis, there are more potential sites for mutations that could alter the splicing of the intron [49]. Other studies have found that highly expressed genes tend to have smaller introns, indicating that long introns may impede transcription [52]. Extremely long introns could be attributed to genetic drift, but they are absent from plants whose population sizes are similar to invertebrates and some vertebrates [53]. The presence of long introns in bilaterian

metazoans but no other multicellular group therefore poses a challenge for both adaptive and non-adaptive hypotheses.

To summarize, my survey of eukaryotic gene structure yielded two observations inconsistent with current hypotheses about intron evolution: high intron densities in several unicellular eukaryotes and the restriction of extremely long introns to bilaterians. As the number and taxonomic breadth of sequenced eukaryotic genomes continues to increase it will be interesting to see if other anomalies arise. The finding of an intron-poor multicellular species or a unicellular species with very long introns would be particularly significant.

Reconstructing the evolutionary history of introns

The explosion of eukaryotic genome sequences has also enabled analyses of intron evolution and conservation by comparing intron positions in orthologous genes. If an intron occurs at the same position in two orthologs, it is inferred to have been in the last common ancestor of the compared species [54, 55]. Numerous methods, ranging from parsimony to more sophisticated maximum-likelihood approaches, have been developed to reconstruct ancestral intron states based on positional conservation among modern species [56-58].

These methods have consistently revealed extensive intron conservation between species, indicating that ancestral eukaryotes were intron-rich [59]. Early studies using a limited number of genomes found many introns conserved between the apicomplexan *Plasmodium falciparum* and crown group eukaryotes [55, 56]. A more recent study including more genomes and a maximum-likelihood method indicated a density of approximately three introns per kb in the ancestral eukaryote, which is comparable to intron densities in land plants and invertebrates [58]. Maximum likelihood methods can also be used to infer the rate of intron gain and loss in various lineages. Intriguingly, the observed rates of intron gain within eukaryotes are not high enough to account for the inferred intron-richness of their last common ancestor [60]. Additionally, little conservation was found among introns in paralogous genes that duplicated in the eukaryotic stem lineage [61]. Together, these results suggest an intron-free period early in the history of eukaryotes followed by massive intron gain at a rate greater than is observed in more recent evolution [61, 62].

More recent gene-structure evolution has rather been dominated by intron loss [3, 15, 55, 58]. Studies of fungi and alveolates have revealed that the last common ancestors of those groups were more intron-rich than any of the extant species [63, 64]. In contrast, one of the rare examples of net intron gain is the lineage leading from the ancestral opisthokont to metazoans. More focused studies on intron evolution within metazoans have found that this gain occurred very early in metazoan evolution; a striking 81% of human intron positions are conserved in the basal metazoan *N. vectensis* [65]. An analysis of intron positions in choanoflagellates, close unicellular relatives of animals, further pinpoints intron gain to the period during which metazoans evolved from their unicellular ancestors (Chapter 2 and

[3]). Though intron gain is associated with the transition to multicellularity in metazoans, intron evolution in the lineage leading to multicellular plants is dominated by loss [58], which is inconsistent with both the adaptive and non-adaptive hypotheses of intron evolution described above.

One concern about these analyses is the potential for the convergent gain of introns at the same position in multiple species. This concern is heightened by the hypothesis that introns only insert into genes at certain positions, termed proto-splice sites, thereby increasing the likelihood that observed "conservation" is due to parallel gain [66]. Although the mechanism behind intron gain is still unknown, simulations assuming the protosplice-site hypothesis is correct find that convergence can only account for a small fraction (5-10%) of conserved intron positions [67]. Nonetheless, the possibility of convergence remains an issue, especially in light of a recent population-level study in the metazoan *Daphnia pulex* that found several instances of parallel intron gain [68].

Caveats aside, the reconstruction of intron evolution provided by these studies has tested hypotheses about the evolutionary forces that influence intron gain and loss. Intron gain concomitant with the transition to multicellularity in metazoans agrees with non-adaptive intron gain in small populations but is also consistent with adaptive hypotheses. However, the lack of further gain within metazoans despite decreased population sizes in vertebrates and increased morphological complexity poses a challenge to both the adaptive and non-adaptive models [65]. Additionally, the lack of intron gain in the lineage leading to land plants is notable. A more detailed understanding of ancestral population dynamics and the functional roles of introns in these lineages will show if these observations can be incorporated into either of the prevailing hypotheses.

Introns: Free-loading or functional?

Functional elements in intronic sequence

Underlying the assertion that introns are slightly deleterious is the assumption that intronic sequence is nonfunctional and without potential to be adaptive [49]. However, molecular biologists have uncovered several functional roles for intronic sequence. Many enhancers, *cis*-elements important for regulating gene expression, occur within the introns of the genes they regulate [69-71]. In one case an intronic enhancer was found to regulate the developmental expression of the mouse *Gli3* transcription factor, which plays a role in body patterning through its regulation of Sonic Hedgehog signaling [72]. In addition to enhancers, ultraconserved elements occur in introns. An ultra-conserved element is a region of 200 bps or greater that is 100% conserved between human, mouse and rat [73]. One hundred of the 481 ultra-conserved elements in the human genome occur in introns [73]. These elements play roles in gene regulation and splicing, and their perturbation has been associated with, though not shown to be causative of, disease phenotypes [74].

In addition to cis-regulatory elements in intronic DNA, some transcribed intronic RNA has regulatory functions. MicroRNAs are small, approximately 21 bp, non-coding RNAs that down-regulate genes post-transcriptionally [75]. In mammals, 40% of canonical microRNAs occur within introns of protein coding genes [76]. Additionally, a new class of intronic microRNAs, termed mirtrons, was recently discovered in *Drosophila melanogaster* [77, 78]. Canonical microRNAs are processed from longer RNAs in a step-wise process by two enzymes, Drosha in the nucleus and subsequently Dicer in the cytoplasm. Mirtrons are contained in short introns and their removal by the spliceosome bypasses the need for Drosha, after which they are processed like canonical microRNAs [75]. To have the correct structure for Dicer recognition, mirtrons can only occur in short introns, presenting a link between intron length and function. Though short introns are relatively uncommon in mammals, some such introns contain mirtons [79]. These discoveries show that intronic sequence is not always useless and leave open the possibility that introns harbor additional, yet-to-be discovered functional elements.

The alternate genome

An indirect function of introns is alternative splicing. Also predicted by Walter Gilbert in 1978 [24], alternative splicing has since proven a widespread aspect of eukaryotic transcription [80]. Alternative splicing occurs when transcripts of a single gene are differentially spliced, allowing one gene to encode multiple proteins [81, 82]. There are several types of alternative splicing (Figure 1.2). In alternative donor/5' and acceptor/3' splice site usage, intronic sequence is added to an exon. Other types of alternative splicing include intron-retention, when introns are not spliced out, and exon skipping, when exons are spliced out along with their flanking introns. More complex patterns such as mutually exclusive or alternate first or last exons also occur [82]. Mechanistically, alternative splicing is regulated by interactions between the spliceosome, RNA binding proteins and sequences found in the involved exons and introns [83-85].

Transcriptome studies have revealed that alternative splicing is common in model metazoans and plants [82, 86]. In humans, a study that sequenced the transcriptomes of 15 tissues and cell types found that over 90% of human genes were alternatively spliced, and that in most cases the alternatively spliced isoform accounted for at least 15% of total transcripts from that gene [87]. Genome-wide studies have also found that different tissues have unique sets of isoforms, suggesting that alternative splicing plays a role in cell differentiation [87, 88].

However, there is much debate over how much of observed alternative splicing is functional and how much is the result of "messy" splicing [89, 90]. A common strategy to test whether alternative splicing is functional has been to look for conservation between species [91, 92]. The percentage of alternative splicing events conserved between human and mouse has been the focus of many studies [93-97]. Estimates vary greatly, from 27% to 67%, when all types of alternative splicing are included [93, 94]. There are fewer comparative studies from other lineages, but a comparison of the plants *Arabidopsis thaliana* and *Oryza sativa* (rice) found that

only 9% of isoforms were conserved [98]. Incompleteness of EST (Expressed Sequence Tag) libraries probably accounts for some of the differences in results, and estimates may change with increasing transcriptome coverage from RNA-seq studies. Nonetheless, it seems that a significant portion of alternative splicing events are non-functional or lineage specific.

Though the genomic view remains unclear, there are many examples of alternative splicing regulating organismal biology. In metazoans, alternative splicing regulates cell-cycle progression [99], apoptosis [100] and nematode and insect sex-differentiation [101, 102]. Alternative splicing also functions in the development of the vertebrate nervous system, where neuronal-specific RNA binding proteins regulate the splicing of many transcripts [103-105]. One such RNA binding protein, NOVA1/2/Pasilla, was shown to regulate alternative splicing in vertebrates and *D. melanogaster* by binding the same RNA sequence motif [106]. However, the positions of the motif itself were not conserved, and the set of alternative splice isoforms generated were non-overlapping [106]. This suggested that NOVA-dependent alternative splicing is ancient even in the absence of conserved splice isoforms, and could have exerted selective pressures on intron evolution throughout the diversification of bilaterians.

Differences in alternative splicing among eukaryotes

It has been suggested that alternative splicing evolved independently in plants and metazoans, facilitating increased organismal complexity [107]. The origin of alternative splicing has ramifications for its influence on intron evolution – a more ancient origin may have posed a long-standing selective pressure against losing introns. Several lines of investigation imply that alternative splicing did indeed evolve early in the history of eukaryotes. In a study of ESTs from 12 diverse species, alternative splicing was more common in ancient genes than in new ones, and the frequency of alternative splicing was the same in genes involved in basic cellular processes and those involved in processes unique to multicellular organisms [80]. The best predictor of alternative splicing was intron-richness. Therefore, since the ancestral eukaryote was intron-rich, alternative splicing may have originated very early in eukaryotic evolution [80].

If alternate splicing was present in the ancestral eukaryote, was it as pervasive as in modern day multicellular organisms? In comparisons of the levels and types of alternative splicing in diverse eukaryotes, unicellular eukaryotes generally have lower levels of alternative splicing. EST studies revealed alternative splicing in only 4.2% of *Cryptococcus neoformans* genes and 3% of *C. reinhardtii* genes [108, 109]. RNA-seq studies with deeper sequence coverage found alternative splicing in 8.6% of *Aspergillus oryzae* genes and 4.5% of *P. falciparum* genes [110, 111]. Contrastingly, estimates of RNA-seq based estimates of alternative splicing level in metazoans range from 25% in *C. elegans* [112] to 92% in humans [87] and in plants from 29% in *A. thaliana* [113] to 48% in *O. sativa* [114]. These results are consistent with an increase in alternative splicing in multicellular lineages from lower levels in ancestral eukaryotes. However, detection of alternative splicing

depends on sequence coverage, which is greater in well-studied multicellular organisms, and these estimates may change with additional transcriptome sequencing of unicellular species.

Another indication that alternative splicing is less common in unicellular eukaryotes comes from sequence features of their splice sites [107]. The core components of the spliceosome recognize specific sequences at splice sites (Figure 1.3), and mutational analysis has shown that the 4th, 5th and 6th base pairs of an intron (part of the 5' splice site recognition sequence) are particularly important for splice site selection [115]. Less conserved, or weak, splice sites have been associated with alternative splicing in vertebrates [116, 117]. Unicellular species tend to have highly conserved 5' splice sites, implying a low frequency of alternative splicing [107].

While most studies point to an increase in alternative splicing in multicellular lineages concordant with changes in morphological complexity, an argument to the contrary has been made based on splice site conservation [118]. A study that included many eukaryotes found that 5' splice site conservation decreased as intron number increased, and the authors argued that the intron-richness of the ancestral eukaryote implies it had a weak 5' splice site and high levels of alternative splicing [118]. To further test this idea, I analyzed splice site conservation in select intron-rich unicellular eukaryotes. I found that the critical 5th base pair downstream of the 5' splice site was more conserved in these species than in similarly intron-rich multicellular species (Figure 1.3). While only based on three unicellular species, this observation indicates that intron-rich unicellular eukaryotes may be exceptions to the negative correlation between intron-richness and splice site conservation. Since the ancestral eukaryote was unicellular the inference that it had weak splice sites prone to frequent alternative splicing is questionable.

Though there are obstacles to comparing levels of alternative splicing among eukaryotes, there are clear differences in the types of alternative splicing favored in various lineages. Numerous studies have found that exon skipping is the most common type of alternative splicing in metazoans, but not in any other groups (Figure 1.4) [119-121]. The high frequency of exon skipping in metazoans is likely due to a difference in the mechanism of splice site recognition [120]. The spliceosome recognizes pairs of splice sites, either across an exon, termed exon definition, or across an intron, termed intron definition. In exon definition, if the spliceosome fails to recognize a splice site the exon is skipped [122]. Contrastingly, in intron definition failure to recognize a splice site results in intron-retention [123]. Intron and exon size influence which mechanism is used. A study using *in vitro* splicing reactions found that intron definition only occurs across introns shorter than 250 bps, and in the case of longer introns splicing proceeds by exon definition [124]. Additionally, a survey of *D. melanogaster* and human ESTs found that skipped exons are associated with longer flanking introns [124]. Thus exon skipping may have evolved as the result of increasing intron lengths in metazoans, making this important regulatory mechanism the by-product of initially non-adaptive changes in gene-structure.

Towards a more complete understanding of intron evolution

Spliceosomal introns have proved an evolutionarily dynamic feature of eukaryotic genomes [59]. The adaptive and non-adaptive hypotheses described above provide alternative explanations for why intron abundance has increased in some lineages but decreased in others. However, the recent explosion of genomic data has yielded observations that challenge both of these hypotheses. Problematic for the non-adaptive theory are the many functional roles of introns, from harboring non-coding RNAs such as microRNAs to expanding the proteome through alternative splicing. [6, 79, 125]. The lack of an increase in intron abundance in land plants, despite decreased population sizes, also disagrees with the predictions of the non-adaptive hypothesis [58]. Additionally, the discovery of intron-rich unicellular species poses a challenge for both schools of thought (Figure 1.1). Thus, neither the adaptive nor non-adaptive hypothesis alone can currently explain all aspects of introns' evolution.

A better understanding of historical population dynamics and the functional roles of introns in diverse eukaryotes will show if unexpected observations can be reconciled with either hypothesis. However, a synthesis of the adaptive and non-adaptive scenarios may provide a more complete picture. The increase of intron length in metazoans is a striking example. This change has a direct mechanistic link to exon skipping, which has been elaborated and exploited in metazoan development [126]. Perhaps intron length initially increased by genetic drift, but then the utility of exon skipping in development presented a selective pressure to maintain increased intron lengths. This and other aspects of intron evolution are potentially cases of useful and novel forms of gene regulation evolving from initially disadvantageous changes, which could prove a repeating theme in the evolution of eukaryotic genomes.

Choanoflagellates and the study of intron evolution

Both the adaptive and non-adaptive hypotheses of intron evolution make specific predictions about changes in gene structure and regulation during transitions to multicellularity. As the closest outgroup to the well-studied Metazoa, choanoflagellates are uniquely positioned to test these hypotheses [3]. The non-adaptive hypothesis predicts that intron abundance should increase during the transition to multicellularity due to decreasing population sizes [49]. The adaptive hypothesis also predicts that intron abundance should increase during this transition but for different reasons, namely selection for functional roles of introns in gene regulation necessary for increased morphological complexity [7].

Comparative analyses of choanoflagellate genomes can test if intron number increased before or after the transition to multicellularity. Our studies using a single choanoflagellate genome have found that the ancestor of choanoflagellates and animals was already intron-rich, though intron number increased further during

early metazoan evolution (Chapter 2 and [3]). This finding is at least partially inconsistent with both hypotheses. The non-adaptive hypothesis must posit unexpected ancestral population dynamics, while the adaptive hypotheses is salvageable if introns had important functional roles prior to the evolution of metazoans. Studying levels and types of alternative splicing in choanoflagellates can provide insight into if this important functional role of introns was present before the transition to multicellularity (Chapter 4). In addition, the study of alternative splicing in choanoflagellates will reveal if aspects of alternative splicing such as widespread exon skipping are truly unique to metazoans (Chapter 4). Choanoflagellates thus have the potential to contribute to our understanding of intron evolution as well as the genomic regulatory landscape from which metazoans evolved.

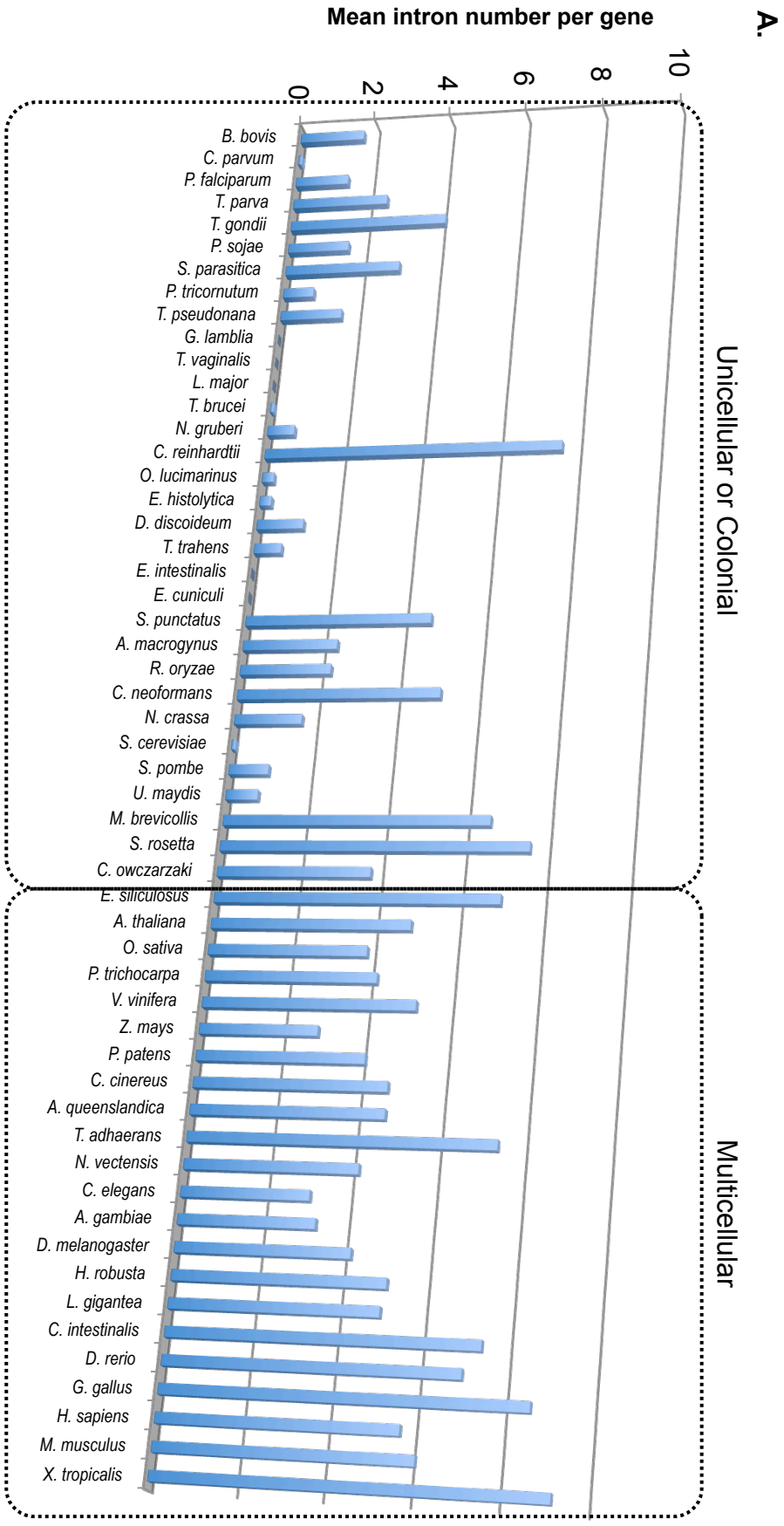
TABLES AND FIGURES

Table 1.1. Intron-exon structure in eukaryotic genomes

Classification	Intron Number per Gene		Intron Length (bp)		Exon Length (bp)		Transcript Length (bp)	
Species	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Chromalveolata								
Alveolata								
Apicomplexa								
Babesia bovis	1.7	1	395.4	205	555.8	211	1525.5	1176
Cryptosporidium parvum	0.1	0	373.3	143	1658.2	1190	1759.7	1277
Plasmodium falciparum	1.4	0	624.0	259	841.1	157	2006.0	1154
Theileria parva	2.5	2	364.8	196	399.7	165	1405.7	1094
Toxoplasma gondii	4.1	3	869.6	669	446.8	186	2274.4	1654
Stramenopila								
Oomycota								
Phytophthora sojae	1.6	1	503.6	260	535.8	272	1409.6	1085
Saprolegnia parasitica	3.0	2	323.8	158	336.1	181	1343.2	1058
Diatomophyceae								
Phaeodactylum tricornutum	0.8	0	462.0	89	840.3	629	1512.6	1261
Thalassiosira pseudonana	1.6	1	10388.0	90	609.2	387	1566.5	1238
Phaeophyceae								
Ectocarpus siliculosus	7.0	5	702.7	531	240.6	142	1920.4	1528
Excavata								
Diplomonadida								
Giardia lamblia	0.001	0	105.8	95	1054.7	434	1055.7	434
Parabasalia								
Trichomonas vaginalis	0.001	0	980.6	807	862.8	650	863.4	650
Kinetoplastidida								
Leishmania major	0.009	0	842.6	166	1590.5	1151	1604.4	1166
Trypanosoma brucei	0.1	0	692.3	314	1394.1	1064	1511.1	1184
Vahlkampfiid amoebae								
Naegleria gruberi	0.7	0	214.5	59	893.9	630	1505.1	1172
Viridiplantae								
Angiospermae								
Arabidopsis thaliana	4.9	3	375.7	270	260.8	146	1528.8	1378
Oryza sativa	3.9	2	655.8	415	312.6	157	1533.7	1388
Populus trichocarpa	4.2	3	577.4	407	276.5	152	1435.3	1253
Vitis vinifera	5.2	3	1177.5	493	240.0	142	1476.2	1249
Zea mays	2.9	1	850.3	403	287.5	152	1110.2	905
Bryophyta								
Physcomitrella patens	4.1	2	505.4	372	274.2	154	1384.3	1243
Chlorophyta								
Chlamydomonas reinhardtii	7.6	6	578.3	389	361.3	155	3108.7	2471
Ostreococcus lucimarinus	0.3	0	166.5	103	956.3	743	1255.5	1006
Volvox carteri	7.1	5	708.7	538	231.0	140	1859.9	1338
Amoebozoa								
Archamoebae								
Entamoeba histolytica	0.3	0	632.6	257	964.2	678	1260.1	971
Dictyostellida								
Dictyostelium discoideum	1.2	1	707.7	278	724.6	335	1626.0	1211
Apusozoa								
Thecamonas trahens	0.7	1	1119.2	449	1093.1	773	1855.8	1338
Fungi								
Microsporidia								
Encephalitozoon intestinalis	0.008	0	233.0	190	1040.8	824	1049.0	829
Encephalitozoon cuniculi	0.008	0	213.7	84	1048.3	830	1057.0	833
Chytridiomycota								
Spizellomyces punctatus	4.7	3	305.8	141	275.8	149	1563.4	1216
Blastocladiomycota								
Allomyces macrogynus	2.4	2	479.5	217	450.1	248	1539.2	1247

Zygomycota								
Rhizopus oryzae	2.3	2	350.3	180	308.9	165	1025.7	769
Dikarya								
Coprinus cinereus	4.7	4	298.9	153	251.2	147	1422.4	1172
Cryptococcus neoformans	5.1	4	321.7	164	263.1	156	1596.6	1355
Neurospora crassa	1.7	1	664.6	266	680.0	398	1827.0	1594
Saccharomyces cerevisiae	0.1	0	833.8	449	1388.1	1151	1459.1	1211
Schizosaccharomyces pombe	1.0	0	611.2	217	1024.1	508	2082.6	1815
Ustilago maydis	0.8	0	629.4	264	1049.7	567	1838.3	1505
Choanoflagellata								
Monosiga brevicollis	6.6	5	188.6	116	242.8	125	1829.3	1305
Salpingoeca rosetta	7.6	5	471.9	328	263.5	116	2252.2	1684
Filasterea								
Capsaspora owczarzaki	3.8	3	511.8	252	429.0	171	2072.7	1648
Metazoa								
Porifera								
Amphimedon queenslandica	4.7	2	421.2	255	217.0	122	1239.7	887
Placazoa								
Trichoplax adhaerans	7.4	5	283.6	134	162.3	102	1366.7	1042
Cnidaria								
Nematostella vectensis	4.2	2	799.2	436	206.5	122	1086.3	803
Nematoda								
Caenorhabditis elegans	3.1	2	572.4	353	207.4	146	854.8	511
Arthropoda								
Anopheles gambiae	3.3	2	1857.5	540	410.2	228	1755.6	1354
Drosophila melanogaster	4.2	3	1829.3	584	454.2	245	2359.7	1866
Annelida								
Helobdella robusta	5.1	3	524.1	255	201.6	129	1232.9	875
Mollusca								
Lottia gigantea	5.0	3	785.4	368	212.5	133	1280.5	957
Chordata								
Ciona intestinalis	7.4	5	717.6	498	170.4	136	1435.0	1138
Danio rerio	7.0	5	3038.6	1173	213.9	127	1703.0	1231
Gallus gallus	8.6	6	2797.7	980	175.2	124	1687.8	1268
Homo sapiens	5.7	3	6481.3	1741	237.7	128	1584.6	890
Mus musculus	6.1	3	5463.9	1545	258.1	130	1844.1	1119
Xenopus tropicalis	9.2	7	2329.0	993	179.2	119	1833.5	1473

Figure 1.1. Differences in intron frequency and length between unicellular and multicellular eukaryotes. (A) The mean number of introns per gene is shown for diverse unicellular and multicellular eukaryotes. While many unicellular eukaryotes are intron-poor, some are as or more intron-rich than multicellular eukaryotes. (B) The median and mean intron length is shown for the same species as in panel A. The vast majority of eukaryotes have similarly sized introns, the only exception being metazoans, which have significantly longer introns. Sources for the annotations of the various genomes are given in Supplementary Table 1.


A.
Mean intron number per gene
0
2
4
6
8
10
B. bovis
C. parvum
P. falciparum
T. parva
T. gondii
P. sojae
S. parasitica
P. tricornutum
T. pseudonana
G. lamblia
T. vaginalis
L. major
T. brucei
N. gruberi
C. reinhardtii
O. lucimarinus
E. histolytica
D. discoideum
T. trahens
E. intestinalis
E. cuniculi
S. punctatus
A. macrogynus
R. oryzae
C. neoformans
N. crassa
S. cerevisiae
S. pombe
U. maydis
M. brevicollis
S. rosetta
C. owczarzaki
E. siliculosus
A. thaliana
O. sativa
P. trichocarpa
V. vinifera
Z. mays
P. patens
C. cinereus
A. queenslandica
T. adhaerans
N. vectensis
C. elegans
A. gambiae
D. melanogaster
H. robusta
L. gigantea
C. intestinalis
D. rerio
G. gallus
H. sapiens
M. musculus
X. tropicalis
Unicellular or Colonial
Multicellular

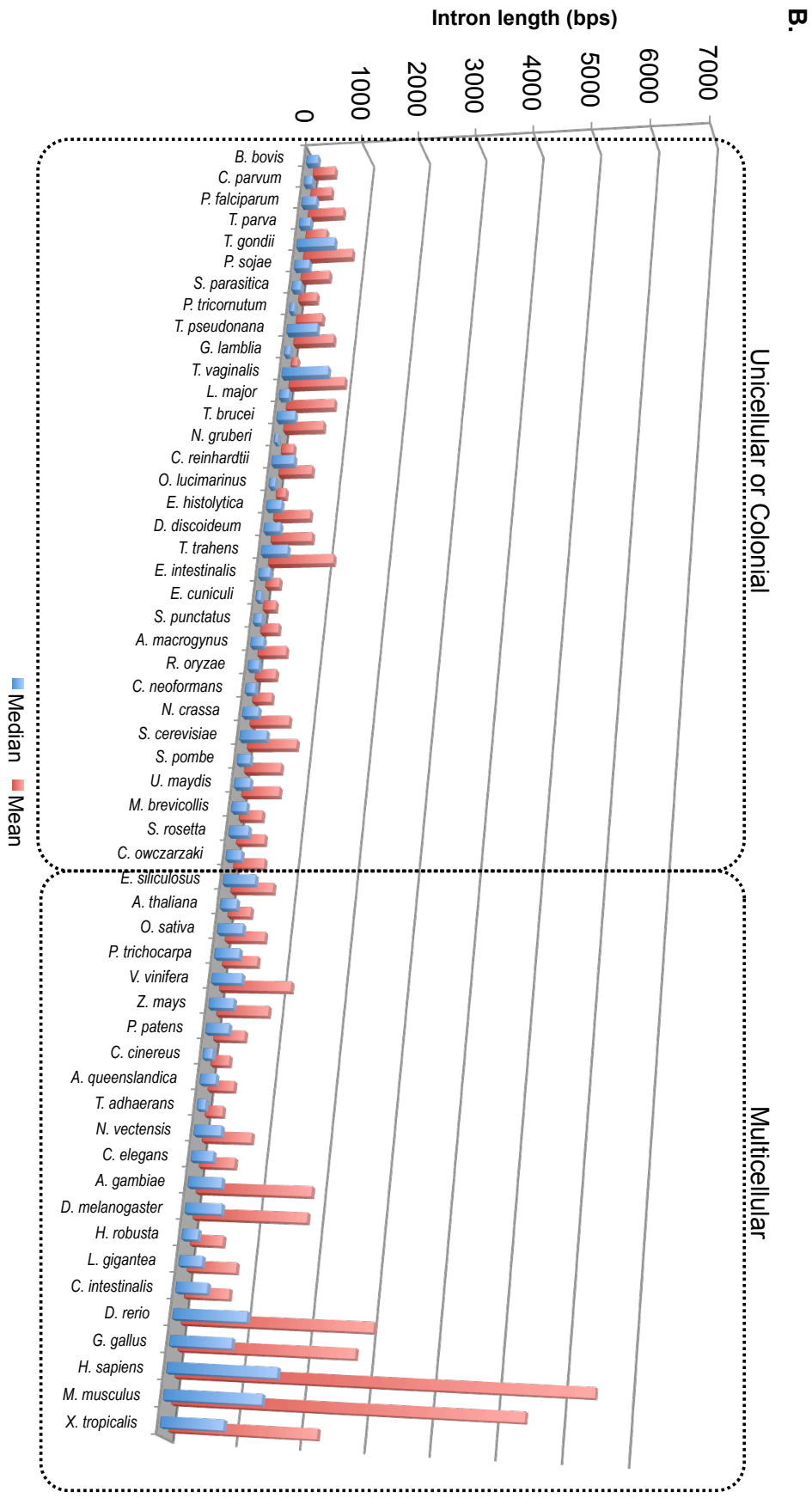

B.
Intron length (bps)
0
1000
2000
3000
4000
5000
6000
7000
Median
Mean
Unicellular or Colonial
Multicellular
B. bovis
C. parvum
P. falciparum
T. parva
T. gondii
P. sojae
S. parasitica
P. tricornutum
T. pseudonana
G. lamblia
T. vaginalis
L. major
T. brucei
N. gruberi
C. reinhardtii
O. lucimarinus
E. histolytica
D. discoideum
T. trahens
E. intestinalis
E. cuniculi
S. punctatus
A. macrogynus
R. oryzae
C. neoformans
N. crassa
S. cerevisiae
S. pombe
U. maydis
M. brevicollis
S. rosetta
C. owczarzaki
E. siliculosus
A. thaliana
O. sativa
P. trichocarpa
V. vinifera
Z. mays
P. patens
C. cinereus
A. queenslandica
T. adhaerans
N. vectensis
C. elegans
A. gambiae
D. melanogaster
H. robusta
L. gigantea
C. intestinalis
D. rerio
G. gallus
H. sapiens
M. musculus
X. tropicalis

Figure 1.2. Depiction of major subtypes of alternative splicing. Four basic types of alternative splicing events result from differential splice site usage: intron retention, where an entire intron is not spliced and remains in the mature mRNA; exon skipping, where an exon is spliced out along with the flanking introns and thus excluded from the mature transcript; alternative donor/5` splice site usage, where the spliceosome recognizes an alternate 5' splice site resulting in two different end sites for one exon; and alterative acceptor/3' splice site usage, where the splicesome recognizes an alternate 3' splice site resulting in two different stop sites for one exon.

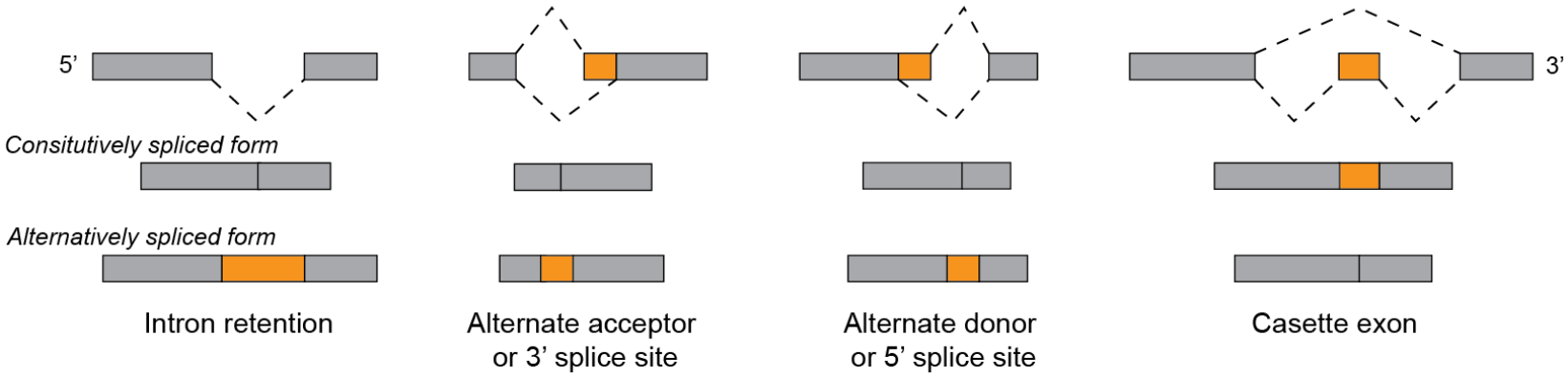

5’
3’
Consitutively spliced form
Alternatively spliced form
Intron retention
Alternate acceptor
or 3’ splice site
Alternate donor
or 5’ splice site
Casette exon

Figure 1.3. Higher levels of 5' splice site conservation in select unicellular Viridiplantae, Fungi, Choanoflagellata than in multicellular Viridplantae and Metazoa. Sequence conservation logos are shown for all predicted splice sites in the *Chlamydomonas reinhardtii* (Viridplantae), *Arabidopsis thaliana* (Viridiplantae), *Cryptococcus neoformans* (Fungi), *Monosiga brevicollis* (Choanoflagellata) and *Nematostella vectensis* (Metazoa) genomes. The overall height of the stacked letters indicates the level of conservation while the relative height of each letter indicates the frequency at which each base occurs at that position. Eight basepairs upstream and downstream of the splice site were included. Genome annotation sources are given in Supplementary Table 1.1.

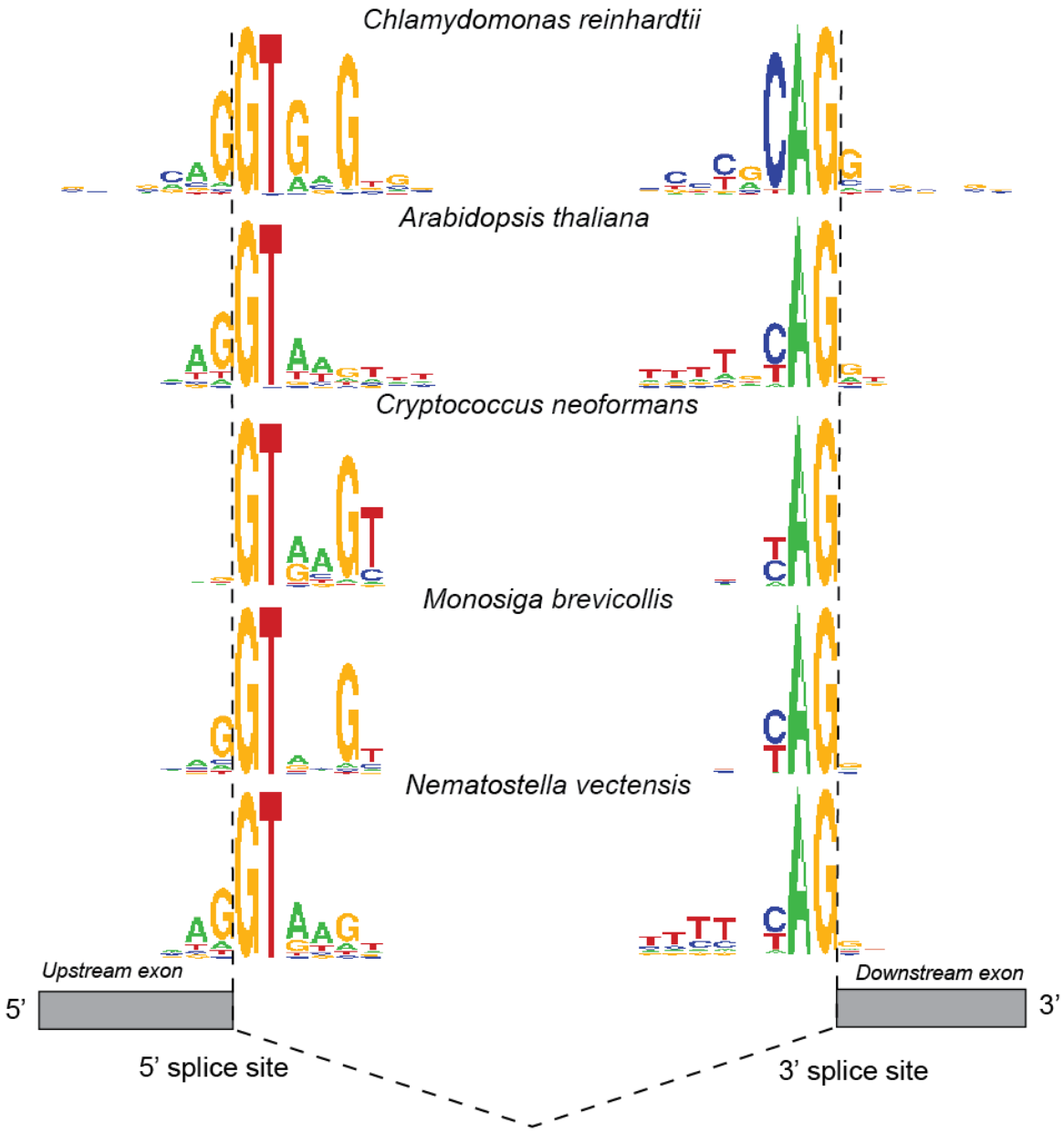

Chlamydomonas reinhardtii
Arabidopsis thaliana
Cryptococcus neoformans
Monosiga brevicollis
Nematostella vectensis
Upstream exon
Downstream exon
5'
3'
5' splice site
3' splice site

Figure 1.4. Metazoans show increased levels of exon skipping relative to other alternative splicing subtypes. For each species shown, the relative frequency of exon skipping was found by dividing the number of exon skipping events by the total number of alternative splicing events (including alternative 5', alternative 3' and intron retention in addition to exon skipping). Alternative splice isoforms were detected in ESTs. Data for human from *Sugnet CW, Kent WJ, Ares M Jr, Haussler D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. Pacific Symposium on Biocomputing, 9:66-77*. Data for all other species from *McGuire AM, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biology, 9:R50*.

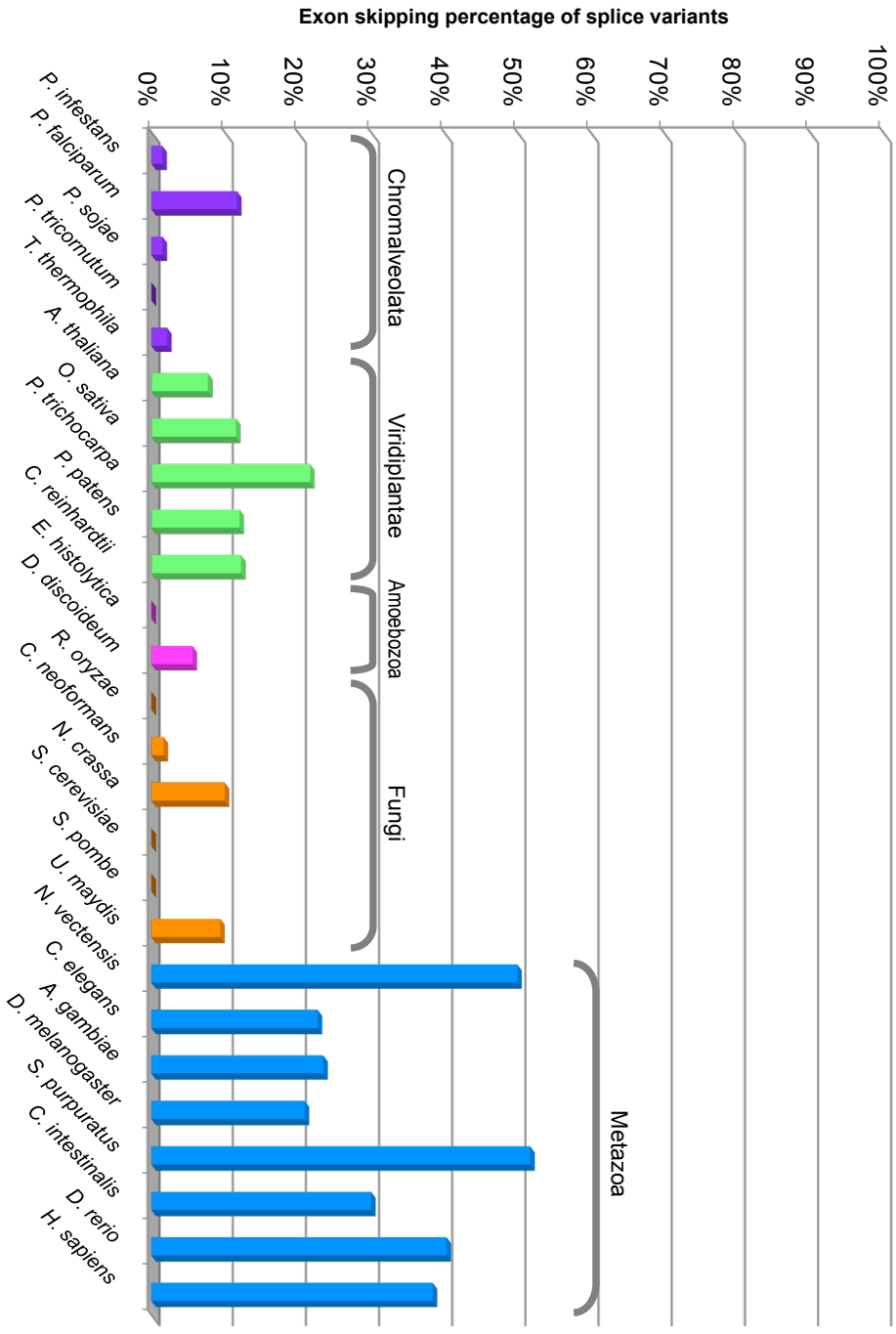

Exon skipping percentage of splice variants
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
P. infestans
P. falciparum
P. sojae
P. tricornutum
T. thermophila
A. thaliana
O. sativa
P. trichocarpa
P. patens
C. reinhardtii
E. histolytica
D. discoideum
R. oryzae
C. neoformans
N. crassa
S. cerevisiae
S. pombe
U. maydis
N. vectensis
C. elegans
A. gambiae
D. melanogaster
S. purpuratus
C. intestinalis
D. rerio
H. sapiens
Chromalveolata
Viridiplantae
Amoebozoa
Fungi
Metazoa

Table S1.1. Genome annotation sources

Species	Annotation source	GFF download reference
Babesia bovis	Piroplasmadb.org	http://beta.piroplasmadb.org/common/downloads/release-1.0/BbovisT2Bo/gff/
Cryptosporidium parvum	Cryptodb.org	http://cryptodb.org/common/downloads/release-4.4/Cparvum/gff/
Plasmodium falciparum	Plasmodb.org	http://plasmodb.org/common/downloads/release-7.2/Pfalciparum/gff/data/
Theileria parva	Piroplasmadb.org	http://beta.piroplasmadb.org/common/downloads/release-1.0/TparvaMuguga/gff/
Toxoplasma gondii	Toxodb.org	http://toxodb.org/common/downloads/release-6.3/Tgondii/gff
Phytophthora sojae	Broad Institute	http://www.broadinstitute.org/annotation/genome/Saprolegnia_parasitica/
Saprolegnia parasitica	Broad Institute	http://www.broadinstitute.org/annotation/genome/Saprolegnia_parasitica/
Phaeodactylum tricornutum	Joint Genome Institute	http://genome.jgi-psf.org/Phatr2/Phatr2.download/
Thalassiosira pseudonana	Joint Genome Institute	http://genome.jgi-psf.org/Thaps3/Thaps3.download/
Ectocarpus siliculosus	VIB/University of Gent	http://bioinformatics.psb.ugent.be/gdb/ectocarpus/Ectsi_gff3_LATEST.tar.gz
Giardia lamblia	Giardiadb.org	http://giardiadb.org/common/downloads/release-2.3/GintestinalisAssemblageA/gff/
Trichomonas vaginalis	Trichdb.org	http://trichdb.org/common/downloads/release-1.2/Tvaginalis/gff/
Leishmania major	TriTrypdb.org	http://tritrypdb.org/common/downloads/release-3.1/Lmajor/gff/
Trypanosoma brucei	TriTrypdb.org	http://tritrypdb.org/common/downloads/release-3.1/Tbrucei/gff/
Naegleria gruberi	Joint Genome Institute	http://genome.jgi-psf.org/Naegr1/Naegr1.download/
Arabidopsis thaliana	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Athaliana/annotation/
Oryza sativa	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Osativa/annotation/
Populus trichocarpa	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Ptrichocarpa/annotation/
Vitis vinifera	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Vvinifera/annotation/
Zea mays	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Zmays/annotation/
Physcomitrella patens	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Ppatens/annotation/
Chlamydomonas reinhardtii	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Creinhardtii/annotation/
Ostreococcus lucimarinus	Joint Genome Institute	http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.download/
Volvox carteri	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Vcarteri/annotation/
Entamoeba histolytica	Amoebadb.org	http://amoebadb.org/common/downloads/release-1.4/Ehistolytica/gff/
Dictyostelium discoideum	Dictybase.org	http://dictybase.org/Downloads/
Thecamonas trahens	Broad Institute	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
Encephalitozoon intestinalis	Microsporidiadb.org	http://microsporidiadb.org/common/downloads/release-1.4/Ecuniculi/gff/
Encephalitozoon cuniculi	Microsporidiadb.org	http://microsporidiadb.org/common/downloads/release-1.4/Eintestinalis/gff/
Spizellomyces punctatus	Broad Institute	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
Allomyces macrogynus	Broad Institute	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
Rhizopus oryzae	Broad Institute	http://www.broadinstitute.org/annotation/genome/rhizopus_oryzae/MultiDownloads.html

Coprinus cinereus	Broad Institute	http://www.broadinstitute.org/annotation/genome/coprinus_cinereus/MultiDownloads.html
Cryptococcus neoformans	Broad Institute	http://www.broadinstitute.org/annotation/genome/cryptococcus_neoformans/MultiDownloads.html
Neurospora crassa	Broad Institute	http://www.broadinstitute.org/annotation/genome/neurospora/MultiDownloads.html
Saccharomyces cerevisiae	Broad Institute	http://www.broadinstitute.org/annotation/genome/saccharomyces_cerevisiae.3/MultiDownloads.html
Schizosaccharomyces pombe	Broad Institute	http://www.broadinstitute.org/annotation/genome/schizosaccharomyces_group/MultiDownloads.html
Ustilago maydis	Broad Institute	http://www.broadinstitute.org/annotation/genome/ustilago_maydis.2/MultiDownloads.html
Monosiga brevicollis	Broad Institute	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
Salpingoeca rosetta	Broad Institute	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
Capsaspora owczarzaki	Broad Institute	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
Amphimedon queenslandica	Joint Genome Institute	ftp://ftp.jgi-psf.org/pub/JGI_data/ Amphimedon_queenslandica/ annotation/
Trichoplax adhaerans	Joint Genome Institute	http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html
Nematostella vectensis	Joint Genome Institute	http://genome.jgi-psf.org/Nemve1/ Nemve1.download.ftp. html
Caenorhabditis elegans	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/ caenorhabditis_elegans/
Anopheles gambiae	Vectorbase.org	http://www.vectorbase.org/GetData/Downloads/
Drosophila melanogaster	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/ drosophila_melanogaster/
Helobdella robusta	Joint Genome Institute	http://genome.jgi-psf.org/Helro1/Helro1.download.ftp.html
Lottia gigantea	Joint Genome Institute	http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html
Ciona intestinalis	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/ciona_intestinalis/
Danio rerio	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/danio_rerio /
Gallus gallus	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/gallus_gallus/
Homo sapiens	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/homo_sapeins/
Mus musculus	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/mus_musculus/
Xenopus tropicalis	Ensembl.org	ftp://ftp.ensembl.org/pub/release-62/gtf/xenopus_tropicalis/

Chapter 2: The genome of the choanoflagellate *Monosiga brevicollis*

SUMMARY

Choanoflagellates are the closest known relatives of metazoans. To reconstruct the genomic changes that accompanied the origin of metazoans, a research consortium of which I was a part sequenced and analyzed the genome of the unicellular choanoflagellate *Monosiga brevicollis*. The genome is small relative to metazoan genomes; at 42 Megabase pairs it encodes approximately 9,200 genes. Though compact, the genome is surprisingly intron-rich with a mean of 6.6 introns per gene. A comparative analysis of intron positions revealed that the last common ancestor of choanoflagellates and metazoans had similarly intron-rich genes. While *M. brevicollis* genes are as intron-rich as their metazoan orthologs, the mean intron length is much shorter. Many of these genes encode cell adhesion and signaling protein domains that are otherwise restricted to metazoans. These domains are often present in combinations that are not found in metazoans, suggesting that domain shuffling followed the divergence of the choanoflagellate and metazoan lineages. These results illuminate potential molecular mechanisms underlying the evolution of metazoan multicellularity and lay the foundations for future molecular studies of choanoflagellates.

Many of the results presented here were published as part of the following paper:

King, N., Westbrook, M.J., Young, S.L. et al. 2008. *The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans*. Nature. 451(1780): 783-8.

I performed and interpreted all of the analyses presented in this chapter.

INTRODUCTION

Choanoflagellates have fascinated evolutionary biologists for over a century because of their striking similarity to the "feeding cells" (choanocytes) of sponges, which raised the possibility that they might represent the closest living relatives of metazoans [127, 128]. Evidence supporting this relationship has accumulated from phylogenetic analyses of nuclear and mitochondrial genes [129-132], comparative genomics between the mitochondrial genomes of choanoflagellates, sponges, and other metazoans [133, 134], and the finding that choanoflagellates express homologs of metazoan signaling and adhesion genes [135-138]. Furthermore, species-rich phylogenetic analyses demonstrate that choanoflagellates are not derived from metazoans, but instead represent a distinct lineage that evolved before the origin and diversification of metazoans [134, 139]. By virtue of their phylogenetic position, studies of choanoflagellates provide an unparalleled window into the nature of the unicellular and colonial progenitors of metazoans [1].

Choanoflagellates are abundant and globally distributed microbial eukaryotes found in marine and freshwater environments [140, 141]. Like sponge choanocytes, each cell bears an apical flagellum surrounded by a distinctive collar of actin-filled microvilli, with which choanoflagellates trap bacteria and detritus, which they then ingest by phagocytosis. Using this highly effective means of prey capture, choanoflagellates link bacteria to higher trophic levels and thus play critical roles in oceanic carbon cycling and the microbial food web [142, 143]. Over 125 choanoflagellate species have been identified and all species have a unicellular life history stage. Some can also form simple colonies of equipotent cells through incomplete cytokinesis, although these differ substantially from the obligate associations of differentiated cells in metazoans [144, 145]. *Monosiga brevicollis* was selected as the first choanoflagellate to have its genome sequence because it readily grows at high concentrations in laboratory conditions. This sequenced culture originates from a single individual that was isolated from a marine environment (Church's Cave, Bermuda, 1986), although *M. brevicollis* can survive in a wide range of salinities (*personal observations*). *M. brevicollis* is seemingly strictly unicellular; in the laboratory it has never been observed to form colonies.

Based on molecular clock analyses, the lineages leading to choanoflagellates and metazoans diverged between 761 and 957 million years ago (MYA) [146]. The first metazoan fossils appeared from 635 to 575 MYA [147]. These numbers bracket the transition to multicellularity to between approximately 761 and 635 MYA, during which time a major snowball earth event occurred and atmospheric oxygen levels increased [148]. Comparative analyses of choanoflagellates and metazoans can provide insight into this major evolutionary transition. Studies of basal metazoans indicate that the ancestral metazoan was multicellular and had differentiated cell types, an epithelium, a body plan, and regulated development including gastrulation. In contrast, the last common ancestor (LCA) of choanoflagellates and metazoans was unicellular or possibly capable of forming simple colonies. Based on the morphological and functional similarities between choanoflagellates and the feeding choanocytes of sponges, this ancestor was likely similar to modern day choanoflagellates in terms of its morphology and lifestyle. The dramatic difference between the LCA of choanoflagellates and metazoans and the ancestral metazoan indicates that a great deal of biological innovation accompanied early metazoan evolution.

Despite their evolutionary and ecological importance, little is known about the genetics and cell biology of choanoflagellates. To gain insight into the biology of choanoflagellates and reconstruct the genomic changes that occurred during the early evolution of metazoans, I participated in a consortium that sequenced the *M. brevicollis* genome. My contributions focused on comparative analyses of gene structure, intron evolution, and protein domain content. I found that *M. brevicollis* genes were surprisingly intron-rich, and that the LCA of choanoflagellates and metazoans had similarly intron-rich genes. In addition, cell-signaling protein domains previously known only from metazoans are present in the *M. brevicollis* genome. However, these domains often occur in unique combinations, indicating

that domain shuffling took place during the early evolution of metazoan signaling pathways.

MATERIALS AND METHODS

Genome sequencing, assembly, and gene prediction
M. brevicollis genomic DNA was isolated and used to construct replicate libraries containing inserts of 2-3 kb, 6-8 kb, and 35-40 kb, each of which was used for Sanger paired end shotgun sequencing. The 41.6 Mb draft sequence of the *M. brevicollis* genome was generated from 8.5-fold redundant paired-end whole genome shotgun sequence coverage. Sequence data derived from six whole-genome shotgun (WGS) libraries was assembled using release 2.9.2 of the WGS assembler Jazz [149]. Out of 29,246 Expressed Sequence Tags (ESTs), 98.5% mapped to this assembly; this indicated that it is a nearly complete representation of the *M. brevicollis* genome.

The Joint Genome Institutes predicted 9,196 non-redundant genes using a variety of methods. The majority of these genes (87%) were predicted by the *ab initio* method FGENESH using a parameterization based on *M. brevicollis* full-length mRNAs and EST cluster consensus sequences that appeared to contain a full open reading frame. Only 13% of gene structure models were predicted using homology-based methods, specifically FGENESH+ and GeneWise. 90% of these gene predictions are complete models in the sense of having start and stop codons, 83% of the gene catalog aligns with proteins in the GenBank non-redundant database (e-value < 0.1) and 56% of the predicted genes encode Pfam domains. Furthermore, 46% of the gene catalog is supported by ESTs.

Intron evolution
To study intron loss and gain in orthologous genes, *M. brevicollis* genes were aligned to human (ENSEMBL models release 26.35.1), *Drosophila melanogaster* (BDGP4 ENSEMBL model release 41), *Nematostella vectensis* (JGI v1.0), *Phanerochaete chrysosporium* (JGI v2.0), *Cryptococcus neoformans A* (Broad Institute v3.0), *Arabidopsis thaliana* (TIGR release 5), *Chlamydomonas reinhardtii* (JGI v3.0), and *Tetrahymena thermophila* (TIGR, 2005) genes. In 473 cases, a human gene was found to have a mutual best hit to a gene from each of the other nine species, forming a tentative cluster of orthologous genes to be studied further.

Gene models are often incomplete at the 5' end, which may also have poorly determined splice sites, so the analysis was restricted to regions of highly conserved peptides in the orthologs of all five species. The independent identification of such regions in multiple species provides strong evidence for the accuracy of the gene models in these regions. Multiple alignments of the orthologous clusters were built using ClustalW and identified gap-free blocks flanked by fully conserved amino acids. Annotated splice sites within these regions were identified, with the additional requirements that 1) none of the peptides have a gap in the alignment closer than 3 amino acids from the splice site and 2) no two different peptides have splice sites at different positions closer than 4 amino acids. Empirically, these

requirements are necessary to avoid spurious detection of intron gains and losses due to ambiguities in either the multiple alignment or the gene models' splice sites. A final requirement was that at least 5 amino acids out of 10 in the flanking regions of the splice sites be either fully conserved or have strong functional similarity among all species. 1,989 intron splice sites at 1,054 highly reliable positions were identified by these requirements. Presence or absence of introns at these positions across was used to build a binary character matrix.

I used three different methods to reconstruct the evolutionary history of introns: Dollo parsimony, Roy-Gilbert maximum likelihood, and Csuros maximum likelihood. Dollo parsimony assumes that introns appearing at the same positions in orthologous genes were gained only once and then subsequently lost in as many lineages necessary to fit the observed phylogenetic pattern [150]. The ancestral state in all cases is a gene without introns. Intron gain and loss events were mapped onto an established species tree using PAUP 4.0b10 [151].

The Roy-Gilbert maximum likelihood method calculates intron loss rates and incorporates them into the estimation of ancestral intron contents [152]. This method was applied to the current data set using a PERL implementation written and made available by Jason Stajich and Scott Roy [63]. This method requires an out-group to infer ancestral intron states, so no inference is made for the most basal node.

The Csuros maximum likelihood method is a probabilistic model that estimates ancestral intron states and intron gain and loss rates for each branch [153]. This method was applied to the current data set using the Java application intronRates.jar made publicly available by the author (http://www.iro.umontreal.ca/~csuros/introns/). This model can also infer a number of "all zero" columns, or introns that were present in an ancestral state but lost in all extant taxa. The results shown here assume that there were no such "all zero" columns, but including "all zero" columns in the model does not dramatically change the results for this data set.

Protein domain content of M. brevicollis

The protein domain content of the *M. brevicollis* genome was annotated using Pfam v20 [154, 155] and SMART v5.1[156] with standard cutoff values. The initial analysis of the phylogenetic distribution of protein domains found in *M. brevicollis* included the species listed in Table S2.3. To identify domains found exclusively in choanoflagellates and other phylogenetic groups, lists were generated using the Pfam and SMART annotations of these genomes. The lists of Pfam and SMART domains were combined using Interpro ID numbers to eliminate overlap. The phylogenetic distribution of each domain thought to be unique to *M. brevicollis* and a given phylogenetic group was then checked by hand using the SMART and Pfam databases online in order to include additional species distribution information. The functions of domains identified as unique to *M. brevicollis* and metazoans were hand-curated.

Over- and under-represented protein domains in *M. brevicollis* as compared to humans and *Schizosaccharomyces pombe* were also identified. This analysis was done using SMART's genomic mode, to avoid over-counting domains due to redundant protein sets. Domains predicted by both SMART and Pfam were included and combined using Interpro ID numbers. The number of times each domain occurred in *M. brevicollis* was compared to its occurrence in *S. pombe* and humans. Significantly different numbers of domains were identified by the Chi-square test and ranked by their p-value. Two sets of comparisons were made, the first of which counted each domain only once per protein and the second of which counted all occurrences of each domain. The top ten over represented domains in *M. brevicollis*, when each domain is counted once per protein, as compared to humans and *S. pombe* are shown in Figure 2.4.

RESULTS

Gene structure and intron evolution

The ~41.6 million base pair (Mb) *M. brevicollis* genome contains approximately 9,196 genes, which were identified using a combination of homology-based and *ab initio* methods supplemented by nearly thirty thousand ESTs. Choanoflagellate genes have several distinguishing structural features (Table 2.1). Only 33% of the genome is intergenic, making the choanoflagellate genome compact relative to metazoan genomes. Though the genome is compact in terms of intergenic spacing, choanoflagellate genes have many introns. Choanoflagellates genes have on average 6.6 introns, making them almost as intron-rich as human genes, which have an average of 7.7 introns per gene. EST clusters, which typically do not cover the entire length of the gene, contained an average of 3.8 introns, placing a lower limit on intron density.

While the high intron density in choanoflagellates is similar to metazoan genomes, the size of these introns, at an average of 174 base pairs, is shorter than average metazoan intron lengths and more similar to that of fungi, amoebae and other unicellular eukaryotes (Table 2.1). The distribution of *M. brevicollis* intron lengths shows that most are very close to the average, and only a few are extremely long (Figure 2.2.A). To determine how this difference manifests itself in orthologous introns in *M. brevicollis* and metazoans, we identified 419 introns in *M. brevicollis* and humans that occurred at the same positions in well conserved genes (Figure 2.2. B). The average length of these introns in *M. brevicollis* is 132 base pairs as compared to 3,438 base pairs in humans, and the length distributions are significantly different between the two species (Kolmogorov-Smirnov comparison test, $D = 0.815$, $p < 0.01$). This suggests that orthologous introns have increased in length during the evolution of metazoans from their unicellular ancestors.

The observation that choanoflagellate genes are intron rich raises the possibility that the unicellular ancestor of metazoans had similarly complex genes. Recent studies on intron evolution suggest that genes in the ancestral eukaryote were

intron-rich, and that introns were subsequently lost in many lineages [56, 58, 63, 153]. Despite widespread loss throughout eukaryotes, these studies consistently show that intron gain outpaced loss in the lineage connecting the opisthokont ancestor to the bilaterian ancestor. The timing of this gain in relationship to the transition to multicellurity has important ramifications for hypotheses about why introns evolved [53]. To infer the evolutionary dynamics of *M. brevicollis* introns, and determine if the previously observed intron gain occurred before or after the transition to multicellularity in the metazoan lineage, we analyzed gains and losses of introns in these lineages.

Several methods have been developed to reconstruct the evolutionary history of introns in orthologous genes. To gain a comprehensive view of the possible alternative scenarios for intron evolution in *M. brevicollis* and early metazoans, I used a set of orthologous introns in well-conserved genes from nine species and three different phylogeny based methods; Csuros maximum likelihood (Figure 2.2), Roy-Gilbert maximum likelihood (Figure 2.2. A) and Dollo parsimony (Figure 2.2. B). Notably, the average number of introns per gene in this set of well-conserved orthologs was different from the average numbers of introns per gene for the entire genomes (12.4 vs. 7.7 introns/gene in humans, 11.7 vs. 5.8 in *N. vectensis*, 8.8 vs. 6.6 introns/gene in *M. brevicollis*, 6.5 vs. 5.3 in *C. neoformans*, and 8.8 vs. 4.4 in *A. thaliana*), which is consistent with the previously reported observation that introns tend to accumulate in highly conserved genes [58].

All models of intron evolution predicted a substantial gain of introns between the ancestral opisthokont and the LCA of choanoflagellates and metazoans, although this ancestor was at least if not more intron-rich than *M. brevicollis*. Therefore, introns were subsequently gained and lost in the choanoflagellate lineage, though the rate of loss was greater than the rate of gain (Table S2.1.). A net intron gain also occurred between the LCA of choanoflagellates and metazoans and the ancestral eumetazoan (Figure 2.3 and Figure S2.1). This observation is consistent with a proliferation of introns early in metazoan evolution. In all later metazoan evolution, the pattern of intron evolution is biased towards intron loss (Table S2.1), indicating that intron density increased early in metazoan evolution and was subsequently maintained or decreased.

Premetazoan history of protein domains and genes associated with metazoan multicellularity and development

Examination of the protein domain content of a genome has the potential to reveal the evolutionary history of protein families and previously unappreciated features of a non-model organism's biology [157-159]. Of the 1,798 Pfam protein domains identified in *M. brevicollis*, I found that the overwhelming majority 1,730 (96%) were shared with metazoans, 1,519 (85%) with fungi, and 1,410 (78%) with the amoebozoan *D. discoideum* [155]. 1,290 (72%) domains were shared among all these groups and are likely involved in basic cellular processes. I also identified protein domains that were found exclusively in *M. brevicollis* and other phylogenetic groupings (Table S2.2.). I found 78 protein domains were shared between *M.*

brevicollis and metazoans to the exclusion of other sequenced genomes, only two of which had been reported in previous EST studies of choanoflagellates [136]. In contrast, I found only two domains that were uniquely shared between *M. brevicollis* and Fungi, and none that were exclusive to *M. brevicollis* and *D. discoideum*. To ensure that the apparent enrichment for domains uniquely shared with metazoans was not due to an overrepresentation of metazoan specific domains in the Pfam database, I compared the percentage of metazoan specific domains found in *M. brevicollis* to the percentage of fungal specific domains. Approximately eight percent of domains previously thought to be unique to metazoans were found in *M. brevicollis*, in contrast with the less than one percent of fungal specific domains.

The 78 domains shared exclusively by *M. brevicollis* and metazoans are of special interest due to their potential contributions to metazoan origins (Table 2.2). Because genomic features shared by *M. brevicollis* and metazoans were likely present in their last common ancestor, this study extends the evolutionary history of this set of protein domains to the pre-metazoan era. Many of these domains are central to cell signaling and adhesion in metazoans (Table 2.2), suggestive of a role in the origin of multicellularity. One example is the Bruton's tyrosine kinase motif [160], which is involved in the regulation of cell proliferation through tyrosine kinase signaling in metazoans. Additional domains involved in tyrosine kinase signaling in metazoans were identified, including the phosphotyrosine binding domain (PTB/PID) and the SH3 domain binding protein 5 domain, indicating that a full set of phospho-tyrosine signaling machinery was in present in the LCA of choanoflagellates and metazoans.

The *M. brevicollis* genome also contains immunoglobulin (Ig) domains, which have both immune and adhesive functions in metazoans and have never previously been detected outside Metazoa (Table 2.2). The *M. brevicollis* genome encodes a total of 5 Ig domains that show affinity for either the I-set, V-set or C2-set subfamilies. Interestingly, the C1-set domain, which is not present in the *M. brevicollis* genome, is exclusively found in immune-related proteins, while the I-set, V-set and C2-set domains are found in both immune and adhesion proteins [161-163]. The absence of C1-set domains in the *M. brevicollis* genome suggests that the adhesive function of Ig domain-containing proteins may have evolved before their immune functions. In contrast to *M. brevicollis*, metazoan genomes possess between approximately 150 and 1,500 Ig domains (Table 2.3), consistent with an expansion of the Ig superfamily after the divergence of choanoflagellates and metazoans.

To gain insight into the biology of *M. brevicollis*, I next identified domains that were overrepresented in the *M. brevicollis* genome relative to other eukaryotic genomes, an approach that has proven fruitful in other genome studies [159]. Domains that are over-represented in *M. brevicollis* compared to humans include the FG-GAP domain and the hyaline repeat domain (Figure 2.4 A). The FG-GAP domain, a domain that is found in the extracellular portion of transmembrane proteins (e.g. α-integrins) and that mediates interactions with the ECM [164], occurs in 35 proteins

in the *M. brevicollis* genome and only 24 proteins in the human genome. The hyaline repeat occurs in 13 proteins in *M. brevicollis* as compared to only three proteins in humans. This predominantly extracellular domain is found in the human glycoprotein hyaline and the sea urchin protein hyalin, which forms an extracellular scaffold around the developing sea urchin embryo [165]. Notably, the five most significantly over-represented domains in *M. brevicollis* relative to *S. pombe* – the ankyrin, SH2, tyrosine protein kinase, PDZ and EGF-like domains – are important in metazoan cell-signaling pathways (Figure 2.4 B). EGF domains are particularly prominent in metazoan transmembrane proteins involved in inter-cellular signaling [29]. While the functions of these domains in choanoflagellates are unknown they may potentially mediate interactions with the external environment such as substrate attachment or quorum sensing.

Domain shuffling in the evolution of metazoan intercellular signaling networks
Metazoan multicellularity and development rests upon a set of signaling pathways that transduce extracellular cues to each cell's nucleus and cytoskeleton. Although all cellular organisms engage in cell signaling, the pathways required for metazoan development are distinct from those found in other multicellular lineages (e.g., fungi and plants). Traditionally, seven intercellular signaling pathways are considered unique metazoan development: receptor tyrosine kinase (RTK), nuclear hormone receptor (NHR), WNT, transforming growth factor-β (TGF-β), Janus kinase (Jak) / signal transducers and activators of transcription (STAT), Notch/Delta and hedgehog [166-168]. Analyses of sponge EST surveys have shown that at least six of these signaling networks (all but the NHR system) evolved prior to the radiation of extant eumetazoans, raising the possibility of even earlier origins [169-171]. Indeed, tyrosine kinase signaling has already been demonstrated in *M. brevicollis* [135, 136].

However, with the exception of tyrosine kinase signaling, no additional metazoan-specific signaling pathways were found in their entirety in the *M. brevicollis* genome. The nuclear hormone receptor, WNT and transforming growth factor-β (TGF-β) pathways were completely absent. In the case of Jak/STAT, Notch/Delta and hedgehog pathways, there are *M. brevicollis* genes that share conserved domains without aligning across the full span of what are often complex multidomain proteins. In these cases, we inferred that the re-arrangement of pre-existing protein domains occurred after the divergence of the choanoflagellate and metazoan lineages.

The Notch/Delta pathway in particular showed an extensive pre-metazoan evolutionary history of protein domain acquisition and re-arrangement (Figure 2.5). This pathway is a receptor-ligand system in which both components are transmembrane proteins and signaling occurs through cell-cell contact. Notch signaling is critical for cell fate specification during the development of *C. elegans*, *D. melanogaster* and other model bilaterians [172]. Notch and its ligands Delta and Jagged have a stereotypical protein domain arrangement that is conserved throughout metazoans (Figure 2.5) [173]. Notch contains one domain that is not

encoded by any other genes throughout metazoans, the NL (Notch-like) domain. It also contains three domains that occur in other signaling proteins – the EGF (Epidermal Growth Factor), NOD (Nucleotide-binding Oligomerization Domain), and Ankyrin domains.

In *M. brevicollis*, I found two distinct genes encoding Notch-like domains (Figure 2.5. A.). Though both genes contain transmembrane domains, neither encodes all of the other domains characteristic of Notch. However, one does have a single EGF domain and ankyrin repeats. Notch may have evolved through duplications of EGF domains and acquisition of the NOD domain in the metazoan lineage, or loss of these domains may have occurred in choanoflagellates.

The Notch homolog in *M. brevicollis* indicates that Notch was present in some form before the evolution of metazoans. However, no homologs of the Delta and Jagged ligands were found in the *M. brevicollis* genome, which raises questions about the ancestral function of this receptor. Surprisingly, another recently sequenced unicellular eukaryote, *Capsaspora owczarzaki*, was found to contain a protein domain characteristic of Delta and Jagged in metazoans, specifically the Delta Serrate Ligand (DSL) domain (Figure 2.5. B). *C. owczarzaki* is an independent lineage sister to choanoflagellates and metazoans. I found three distinct DSL domain containing proteins in the *C. owczarzaki* genome, one of which was present in 12 copies, a greater number than is found in any metazoan. However, these genes also encode other domains such as the Complement Control Protein (CCP) and Tyrosine kinase (YK) signaling domain, which are not found in metazoan Delta and Jagged, indicating that extensive domain acquisition and rearrangement occurred in DSL domain-containing proteins after the divergence of *C. owczarzaki* and the lineage eventually leading to metazoans.

Together, these results suggest that the metazoan Notch/Delta signaling system evolved through domain duplication and re-arrangement early in metazoan evolution. The presence of the Delta domain in *C. owczarzaki* and not *M. brevicollis* indicates that Delta was lost in choanoflagellates, and that there was a pre-metazoan ancestor that contained both of these domains. Since Notch and Delta interact through direct cell-cell contact in development, the presence of these domains together in a presumably unicellular ancestor suggests that their role in cell-fate specification may have been co-opted from yet-to-be discovered ancestral functions.

DISCUSSION

The evolution of metazoans from their unicellular common ancestor with choanoflagellates was a pivotal event in life's history, but little is known about the nature of this major evolutionary transition. Given the absence of a fossil record from this period, comparisons of modern genomes provide a valuable window into the origins of metazoans. The comparisons of the genomes of the choanoflagellate *M. brevicollis* and diverse metazoans described here have yielded several insights into their last common ancestor and the ensuing transition to multicellularity.

Gene structure is one aspect of genome evolution that may have been linked to transitions to multicellularity [9]. Previous genome sequencing projects have found that the genomes of multicellular land plants and metazoans tend to be much larger than those of unicellular eukaryotes, and that this difference is due mainly to increased intergenic spaces and intronic sequence rather than additional genes [9]. Using the genome of *M. brevicollis*, I was able to investigate whether these changes took place before or after the transition to multicellularity in the metazoan lineage. The *M. brevicollis* genome is small relative to metazoan genomes with intergenic sequences that are comparable in size to those of Fungi and other unicellular eukaryotes. The most parsimonious scenario is that the LCA of choanoflagellates and metazoans had a similarly compact genome and that an increase in genome size took place before the evolution of eumetazoans.

The genes of *M. brevicollis* contain numbers of introns similar to the most intron-rich metazoan lineages. By comparing intron positions in orthologous genes, I inferred that genes in the LCA of choanoflagellates and metazoans were at least as intron-rich as those in *M. brevicollis*, and that significant intron gain occurred in the lineage leading from the ancestral opisthokont to this ancestor. Intron gain also occurred after the divergence of choanoflagellates and metazoans on the lineage leading to the ancestral eumetazoan. This observation poses a challenge to current hypotheses about intron-gain that predict that intron-gain occurred after the transition to multicellularity ([49], Chapter 1). Intron-rich metazoan genes sometimes undergo complex patterns of alternative splicing during development and cell differentiation [85], and future studies in choanoflagellates have the potential to reveal how introns impact gene regulation in a unicellular context.

The *M. brevicollis* genome has also increased our understanding of how cell adhesion and signaling pathways unique to metazoans first evolved. My comparative study of the protein domain content of *M. brevicollis* revealed that many domains involved in cell adhesion and signaling in metazoans were already present in their unicellular ancestor. Furthermore, I found that pan-eukaryotic cell-signaling domains such as EGF and ankyrin that were known to have increased in abundance in metazoans relative to other groups also occur in large numbers in the *M. brevicollis* genome. The presence and abundance of these signaling domains in the unicellular *M. brevicollis* raises the possibility that their roles in intercellular communication and cell-differentiation were co-opted from unique, yet-to-be discovered ancestral functions.

The mechanism of invention of new genes on the metazoan stem, and their integration to create the cell signaling and transcriptional networks fundamental to metazoan biology, remains mysterious. Domain shuffling, which has frequently been proposed as an important mechanism for the evolution of metazoan multidomain proteins [29, 174], is implicated by the presence of essential metazoan signaling domains in *M. brevicollis* in unique combinations relative to metazoans. In the specific case of Notch, a nascent version is present in *M. brevicollis*, but lacks the

NOD domain and the extensive stretch of EGF repeats found in metazoan Notch orthologs, indicating that domain acquisition and/or duplication occurred in the metazoan lineage, or that domain loss occurred in choanoflagellates. Though the DSL domain characteristic of metazoan Notch ligands is not found in the *M. brevicollis* genome, it was identified in the unicellular sister group to choanoflagellates *C. owczarzaki*, and the unique combinations of DSL and other domains found in this organism implicate domain shuffling in the evolution of Notch ligands.

In addition to its utility in evolutionary studies of metazoan protein families, the *M. brevicollis* sequence opens the door to genome-enabled molecular and ecological studies of choanoflagellates, a diverse group of microbial eukaryotes that are important in their own right as bacterial predators in aquatic ecosystems. While *M. brevicollis* is strictly unicellular, other choanoflagellates facultatively form colonies, and the modulation of these associations by cell signaling, adhesion, transcriptional regulation, and environmental influences is poorly understood [144, 175]. An integrative approach that unites studies of choanoflagellate genomes, cell biology, and ecology with the biogeochemistry of the Proterozoic has the potential to reveal the intrinsic and extrinsic factors that influenced the origin of metazoans.

TABLES AND FIGURES

Table 2.1. *M. brevicollis* genome properties in a phylogenetic context

	Metazoa				Choanoflagellates	Fungi		Dictyostelium	Plants
	Hsap	*Cint*	*Dmel*	*Nvec*	***Mbre***	*Ccin*	*Ncra*	*Ddis*	*Atha*
Genome size (Mbp)	2,900	160	180	357	**42**	38	39	34	125
Total number of genes	23,224	14,182	14,601	18,000	**9,196**	13,544	9,826	13,607	27,273
Mean gene size (bp)	27,000	4,585	5,247	6,264	**3,004**	1,679	1,528	1,756	2,287
Mean intron density (introns/gene)	7.7	6.8	4.9	5.8	**6.6**	4.4	1.8	1.9	4.4
Mean intron length (bp)	3,365	477	1,192	903	**174**	75	136	146	164
Gene density (kb/gene)	127.9	11.9	13.2	19.8	**4.5**	2.7	4.0	2.5	4.5

Table 2.2. Functional classification of domains unique to choanoflagellates and metazoans

Cell Adhesion and Extracellular Matrix	
Cadherin*	Laminin G*
CUB	Laminin N-terminal
Ependymin	Reeler
Fibrillar collagen C-terminal	Somatomedin B
HYR*	Von Willebrand D*
Kunitz/bovine pancreatic trypsin inhibitor*	
Signal Transduction	
Antistasin family	Nine cysteines of family 3 GPCR
BTK motif	Pacifastin inhibitor (LCMII)
C1q*	Phosphotyrosine binding (IRS-1 type)
CBL proto-oncogene N-term, domain 1	Phosphotyrosine interaction (PTB/PID)
CBL proto-oncogene N-term, EF hand-like	PI3-kinase family, p85-binding
CBL proto-oncogene N-term, SH2-like	Plexin
ECSIT	Raf-like ras-binding
Flotilin family	Renin receptor-like protein
GoLoco motif	S-100/ICaBP type calcium binding
Heme NO binding associated	Seven transmembrane receptor, secretin family
Hormone receptor	SH3 domain-binding protein 5 (SH3BP5)
L27	Spin/Ssty family
Low-density lipoprotein receptor class A	TNF (Tumor Necrosis Factor)
Cell Adhesion and Signal Transduction	
Leucine rich repeat N-terminal	Immunoglobulin I-set*
Immunoglobulin	Immunoglobulin V-set*
Immunoglobulin c-2*	
Transcriptional Control	
Mbt repeat	STAT protein, DNA binding
p53 DNA-binding**	Zinc finger, C2HC type
PET	
Cytoskeletal Associated	
Nebulin repeat	Repeat in HS1/cortactin
Filament	Sarcoglycan complex subunit protein
Transporters/Channels	
Dihydropyridine sensitive L-type calcium channel	Organic anion transporter polypeptide (OATP)
Inward rectifier potassium channel	Progressive ankylosis protein (ANKH)
Enzymes	
Aspartyl/asparaginyl beta-hydroxylase	Galactosyl transferase
DNaseIc*	Glycosyl hydrolase family 59*
Cu_2 monooxygenase	Heparan sulfate 2-0-sulfotransferase*
Fzo-like conserved region	N-acetylglucosaminyltransferase-IV conserved reg.
Galactose-3-O-sulfotransferase	Phosphomevalonate kinase
Unknown	
Assoc. with transcription factors and helicases	PHR
Domain of unknown function (DUF758)	Protein of unknown function (DUF1241)
Domain of unknown function (DUF837)	Selenoprotein S (SelS)
Fukutin-related	Translocon-associated protein, δ subunit precursor

Hormone-sensitive lipase (HSL) N-terminus	Tropomyosin
MOFRL family*	Uncharacterized protein family (UPF0121)
N-terminal domain in C. elegans NRF-6	

*Present in bacteria **Partial domain present in *Zea mays* (Qi, 2003)

Table 2.3. Immunoglobulin domains in choanoflagellates and metazoans

	Metazoa			Choanoflagellates	Fungi		Dictyostelia	Plants
	Hsap	*Cint*	*Dmel*	*Mbre*	*Ccin*	*Ncra*	*Ddis*	*Atha*
Immunoglobulin*	1502	144	503	5	0	0	0	0

Figure 2.1. Phylogenetic placement of the choanoflagellate *M. brevicollis*. The close phylogenetic affinity between choanoflagellates and metazoans highlights the value of the *M. brevicollis* genome for investigations into metazoan origins, the biology of the last common ancestor of metazoans (filled circle) and the biology of the last common ancestor of choanoflagellates and metazoans (open circle). Genomes from the species shown were used as a starting point for the comparative analyses in this study.

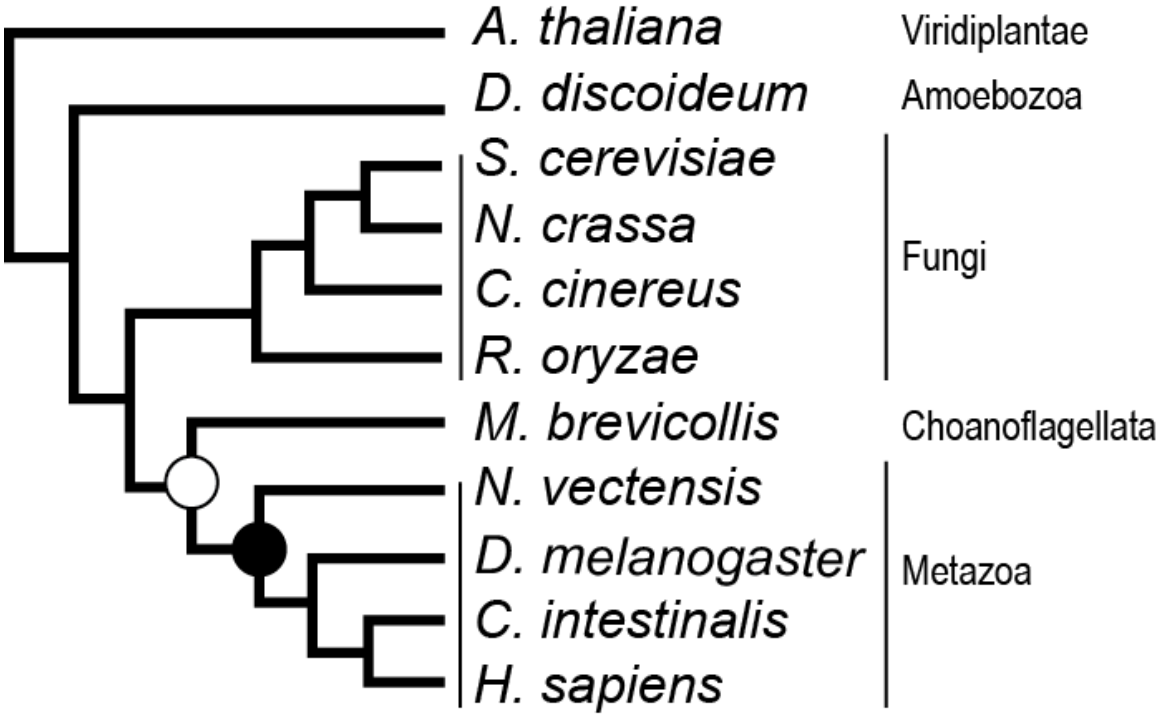
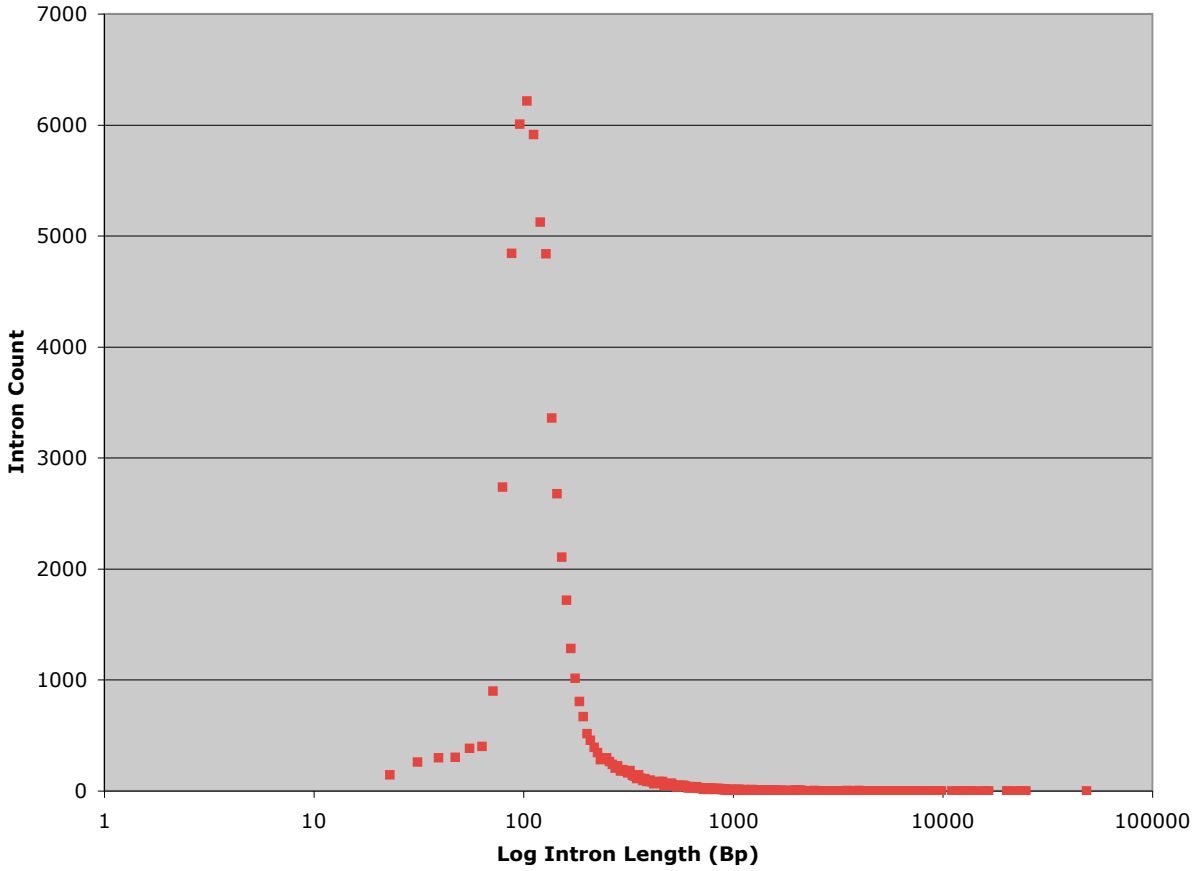


Figure 2.2. Distribution of *M. brevicollis* intron lengths. (A) Distribution of the lengths of the 60,636 introns from the *M. brevicollis* filtered gene models. (B) Distribution of the lengths of 419 introns that occur at the same positions in orthologous genes in *M. brevicollis* (red dots) and humans (blue dots).

A.



B.

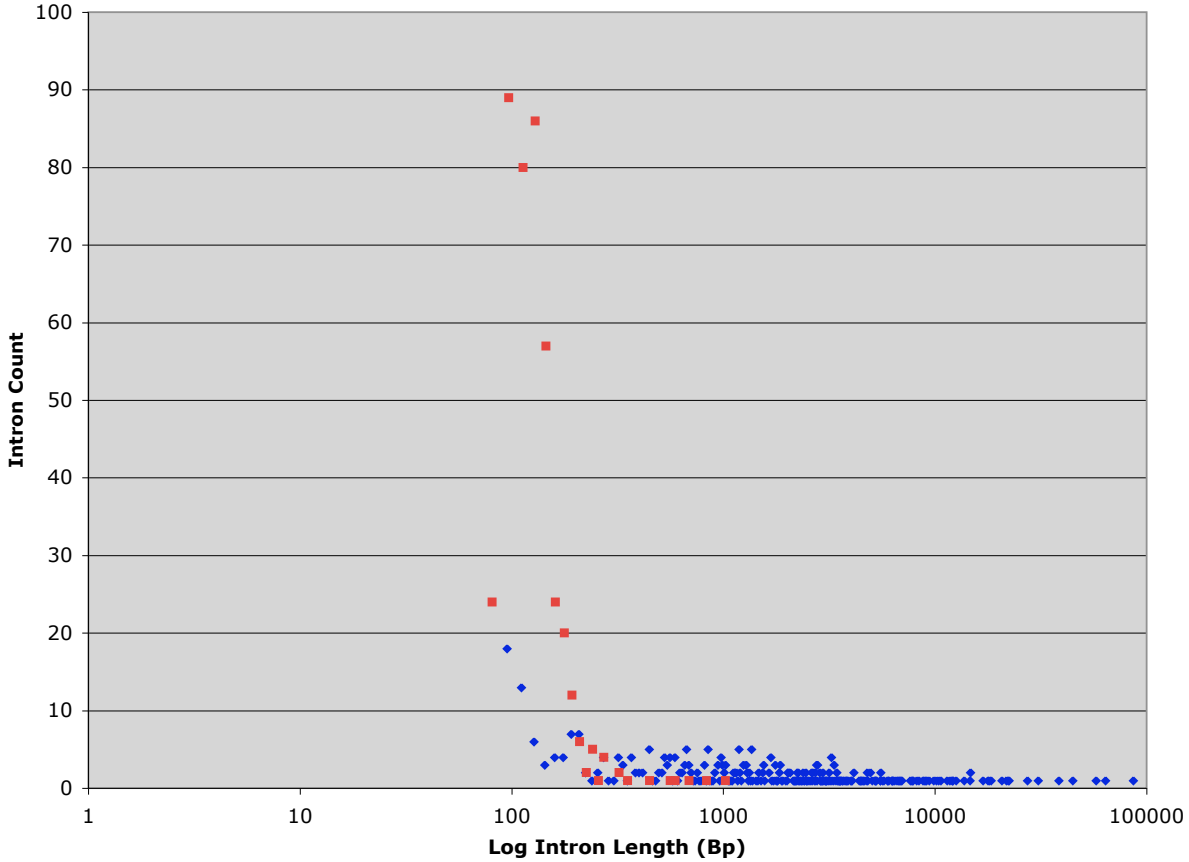

100
90
80
70
60
50
40
30
20
10
0
Intron Count
1
10
100
1000
10000
100000
Log Intron Length (Bp)

Figure 2.3. Intron gain preceded the origin and diversification of Metazoans. Ancestral intron content, intron gains and intron losses were inferred by the Csuros maximum likelihood method [153] from a sample of 1,054 intron positions in 473 highly conserved genes in representative metazoans (humans, *Drosophila melanogaster*, and *Nematostella vectensis*), *Monosiga brevicollis*, intron-rich fungi (*Cryptococcus neoformans A* and *Phanerochaete chrysosporium*), plants and green algae (*Arabadopsis thaliana* and *Chlamydomnas reinhardtii*), and a ciliate (*Tetrahymena thermophila*). Branches with more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. The inferred or observed number of introns present in ancestral and extant species are indicated by proportionally sized circles. As in Fig. 1, the last common ancestor of metazoans and the last common ancestor of choanoflagellates and metazoans are represented by a filled circle and an open circle, respectively.

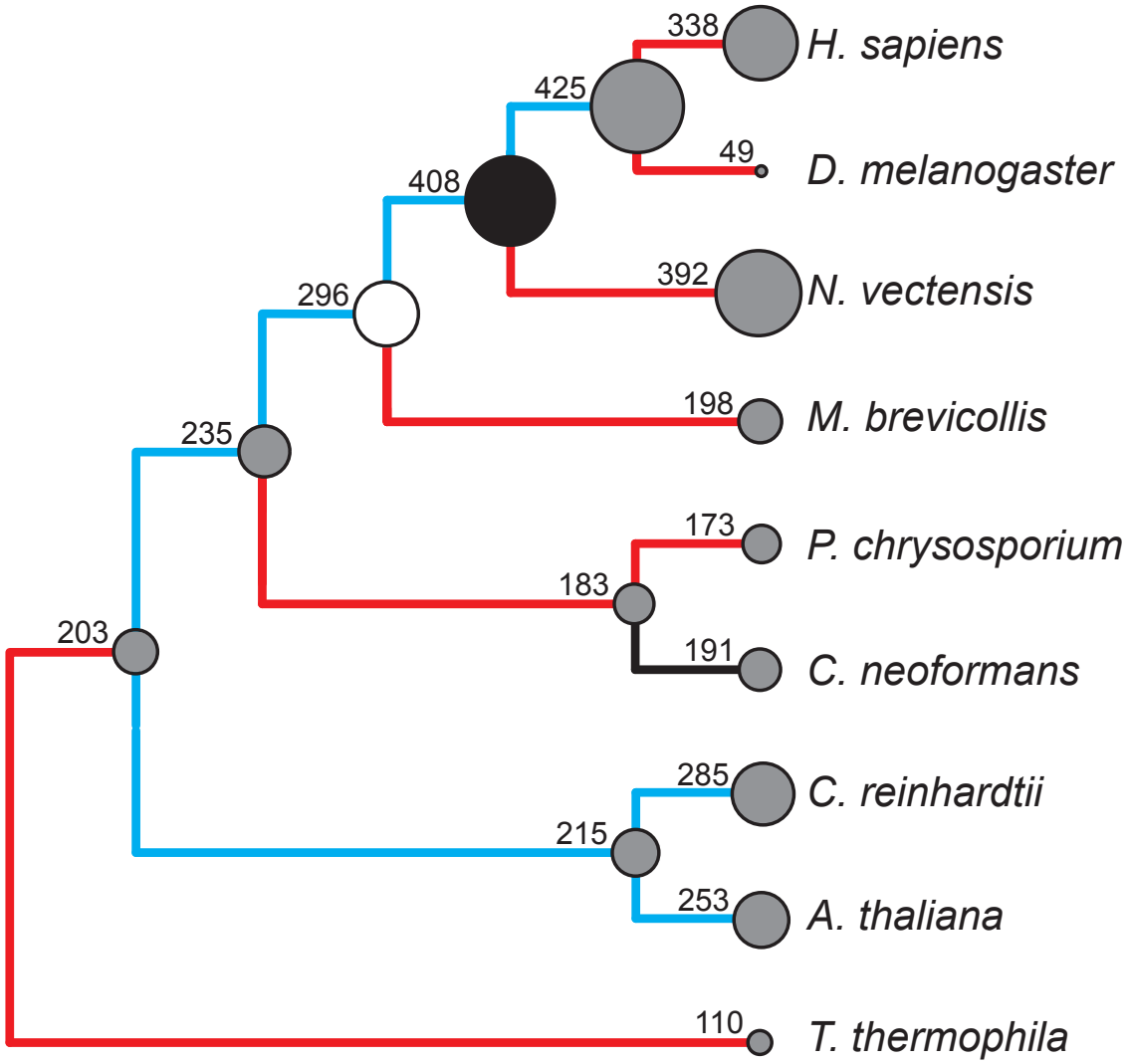
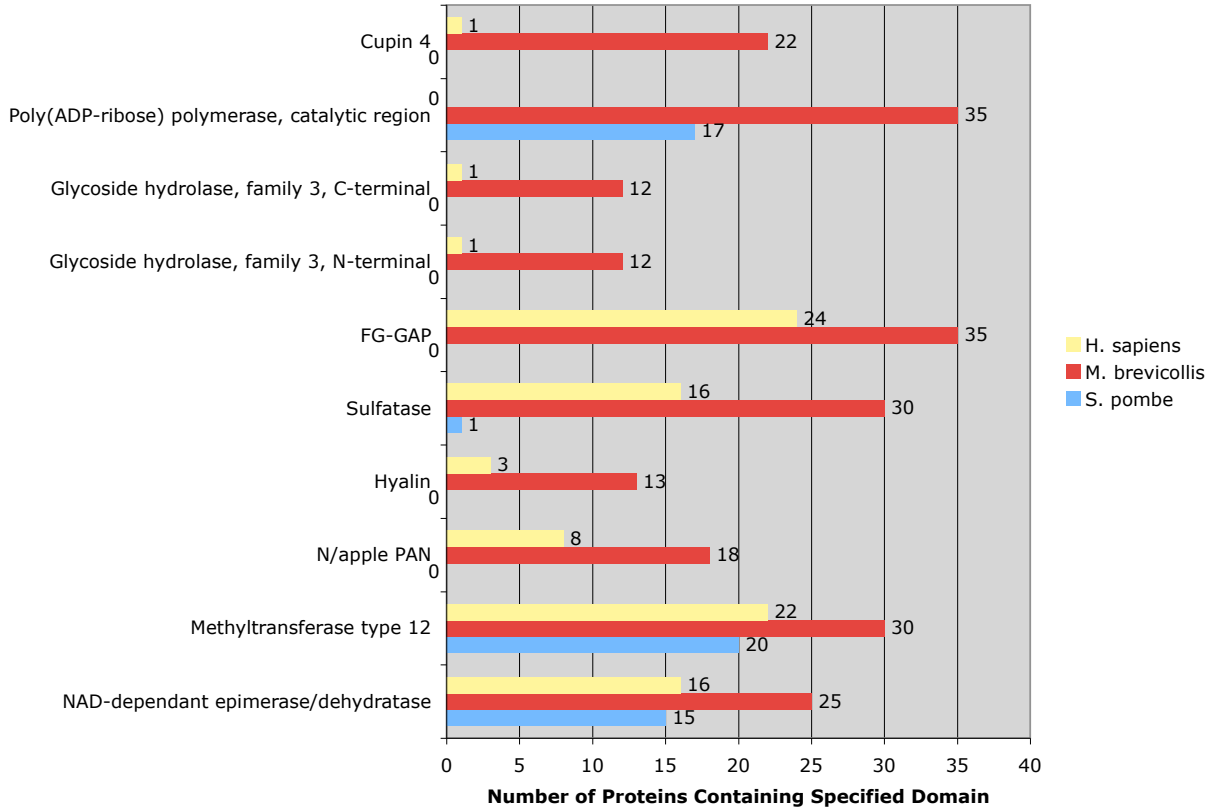


Figure 2.4. Domains significantly over-represented in choanoflagellates. Significantly over-represented domains in the choanoflagellate genome were identified by comparing the occurrence of PFAM domains excluding repeats (one hit per protein) in *M. brevicollis* to the human (A) and *S. pombe* (B) genomes. The ten most significantly over represented domains from each comparison as determined by a Chi-squared test are shown, with the most significantly over-represented domain shown at the top of the graphs. The number of proteins containing each domain is indicated.

A.



B.

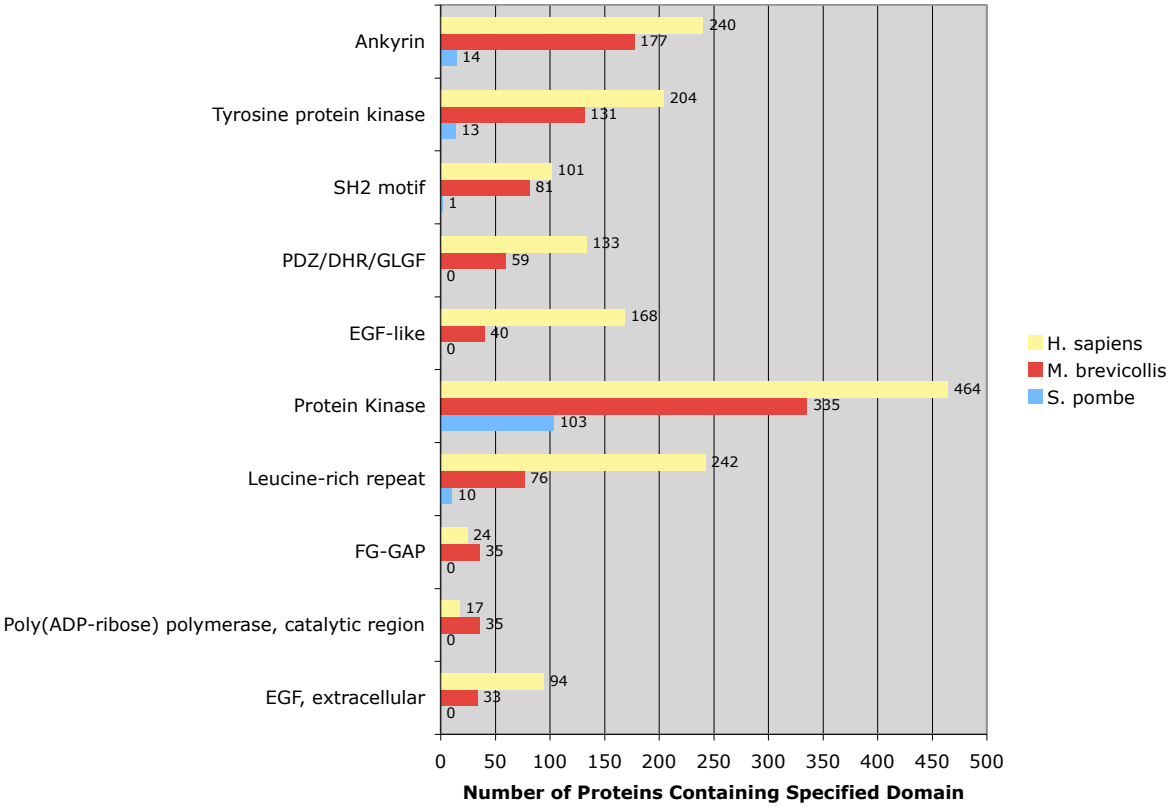

Ankyrin
240
177
14
Tyrosine protein kinase
204
131
13
SH2 motif
101
81
1
PDZ/DHR/GLGF
133
59
0
EGF-like
168
40
0
Protein Kinase
464
335
103
Leucine-rich repeat
242
76
10
FG-GAP
24
35
0
Poly(ADP-ribose) polymerase, catalytic region
17
35
0
EGF, extracellular
94
33
0
0 50 100 150 200 250 300 350 400 450 500
Number of Proteins Containing Specified Domain
H. sapiens
M. brevicollis
S. pombe

Figure 2.5. Domain shuffling and the pre-metazaon evolution of Notch and Delta. (A) Analysis of the draft gene set found that *M. brevicollis* possessed NL protein domain characteristic of metazoan Notch in two separate proteins, one of which contained additional domains found in metazoan Notch and represents a non-metazoan Notch homolog. (B) Analysis of the *Capsaspora owczarzaki*, the unicellular outgroup to choanoflagellates and metazoans, gene set identified genes that encode the DSL domain, which is characteristic of the Notch ligands Delta and Jagged, in unique arrangements not seen in metazoans. Protein domain abbreviations: NL – Notch-Like; DSL – Delta Serrate Ligand; TM – Transmembrane; EGF – Epidermal Growth Factor; NOD – Nucleotide-binding Oligomerization Domain; VWC – Von Wildebrand C domain; MNLL – N-terminal Notch-Like Ligand; CCP - Complement Control Protein; YK – Tyrosine Kinase; STK – Serine Threonine Kinase.

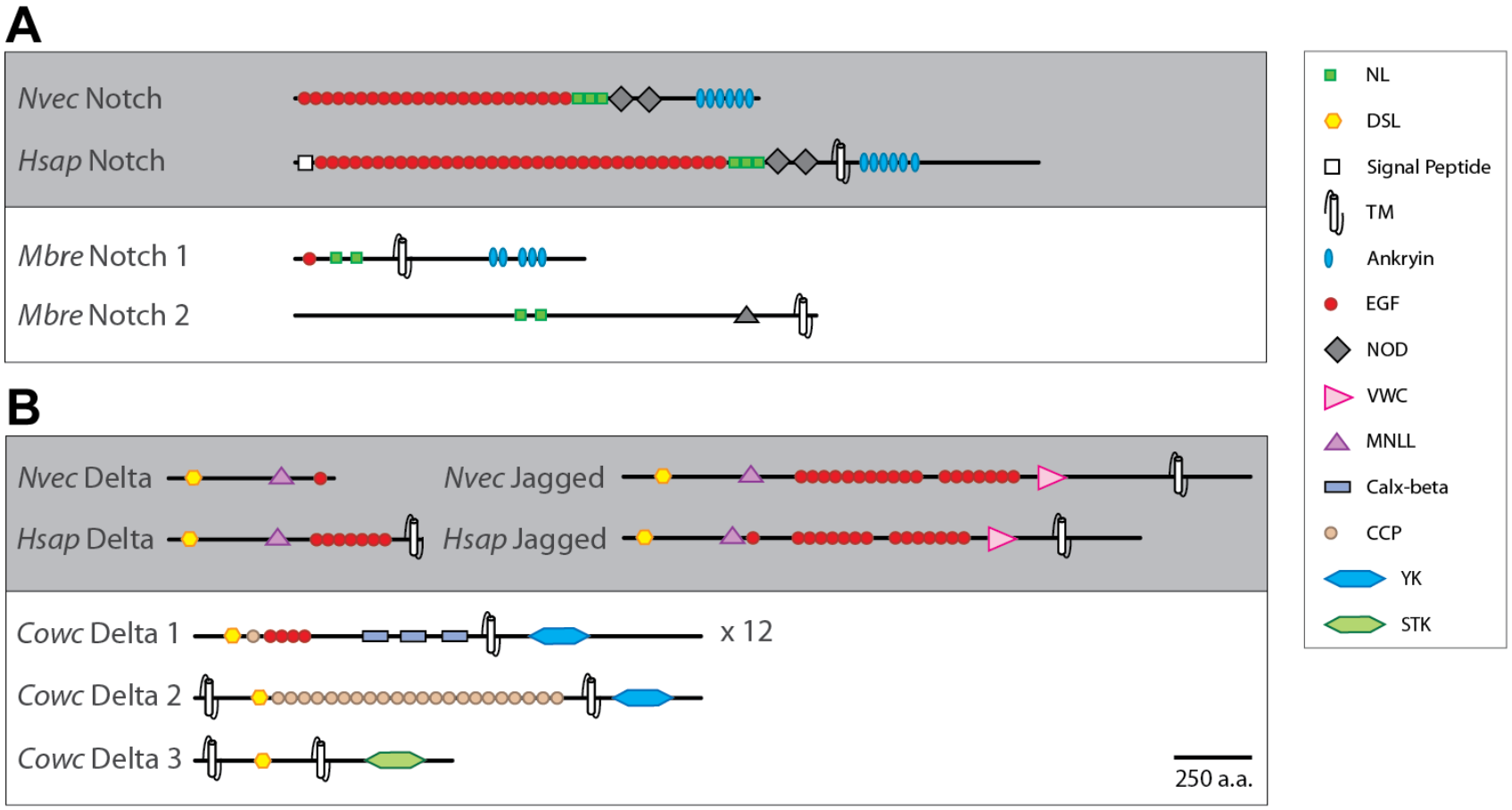

A
Nvec Notch
Hsap Notch
Mbre Notch 1
Mbre Notch 2
B
Nvec Delta
Nvec Jagged
Hsap Delta
Hsap Jagged
Cowc Delta 1
x 12
Cowc Delta 2
Cowc Delta 3
250 a.a.
NL
DSL
Signal Peptide
TM
Ankryin
EGF
NOD
VWC
MNLL
Calx-beta
CCP
YK
STK

Table S2.1. Intron gain and loss as calculated by Csuros maximum likelihood

Branch	Introns Gained	Introns Lost
Eukaryotic → *T. the*	64	157
Eukaryotic → Green plants ancestor	65	52
Green plants ancestor → *A. tha*	73	36
Green plants ancestor → *C. rei*	177	108
Eukaryotic → Opisthokont ancestor	56	23
Opisthokont → Basidomycete ancestor	75	126
Basidiomycete ancestor → *C. neo*	87	80
Basidiomycete ancestor → *P. chr*	32	42
Opisthokont → Choanoflagellate/Metazoan ancestor	61	0
Choanoflagellate/Metazoan ancestor → *M. bre*	69	167
Choanoflagellate/Metazoan → Eumetazoan ancestor	135	23
Eumetazoan ancestor → *N. vec*	12	29
Eumetazoan → Bilaterian ancestor	30	13
Bilaterian ancestor→ *D. mel*	21	397
Bilaterian ancestor → *H. sap*	1	89

Table S2.2. Protein domains unique to choanoflagellates and other groups

Domain Name	*Interpro ID*
Metazoa, Choanoflagellates, Fungi, and Dictyostelium	
Growth-Arrest-Specific Protein 2 Domain	IPR003108
Protein of unknown function (DUF1183)	IPR009567
Protein of unknown function (DUF1613)	IPR011671
Mss4 protein	IPR007515
UcrQ family	IPR004205
Diaphanous FH3 Domain	IPR010472
WSC domain	IPR002889
TAP C-terminal domain*	IPR005637
RasGAP C-terminus	IPR000593
GGL domain	IPR001770
Ras association (RalGDS/AF-6) domain	IPR000159
I/LWEQ domain	IPR002558
BTG family	IPR002087
Cysteine dioxygenase type I*	IPR010300
Fic protein family*	IPR003812
Fes/CIP4 homology domain (FCH)	IPR001060
GTPase-activator protein for Ras-like GTPase (Ras GAP)	IPR008936
RasGEF	IPR001895
RasGEF, N-terminal motif	IPR000651
Wiskott Aldrich syndrom homology region 2*	IPR003124
Alpha adaptin AP2, C-terminal domain	IPR003164
G-protein gamma like domain (GGL)	IPR001770
BTG domain	IPR002087
Metazoa, Choanoflagellates, and Fungi	
Arfaptin	IPR010504
ATP synthase D chain, mitochondrial (ATP5H)	IPR008689
Cation-dependent mannose-6-phosphate receptor	IPR000296
CP2 transcription factor family	IPR007604
CybS	IPR007992
Cytochrome c oxidase subunit Va	IPR003204
D-ala D-ala ligase C-terminus	IPR011095
Disintegrin	IPR001762
Dolichyl-phosphate-mannose-protein mannosyltransferase	IPR003342
Epoxide hydrolase N terminus	IPR010497
Forkhead domain	IPR001766
FRG1-like family	IPR010414
GDP/GTP exchange factor Sec2p	IPR009449
Golgi phosphoprotein 3 (GPP34)	IPR008628
HRDC (Helicase and RNase D C-terminal) domain	IPR002121
Inhibitor of Apoptosis domain	IPR001370
Microtubule associated	IPR012943
Peptidase C1-like family	IPR004134
Protein of unknown function (DUF1349)	IPR009784
Putative phosphatase regulatory subunit	IPR005036
Receptor L domain	IPR000494
RFX DNA-binding domain	IPR003150
SURF4 family	IPR002995
TEA/ATTS domain family	IPR000818
XPA protein C-terminus	IPR000465
XPA protein N-terminal	IPR000465

Metazoa, Choanoflagellates, and Dictyostelium

Tryptophan 2,3-dioxygenase*	IPR004981
DUF1632	IPR012435
Beta catenin interacting protein (ICAT)	IPR009428
DUF1394	IPR009828
RUN domain	IPR004012
Doublecortin	IPR003533
Translocon assoc. protein, gamma subunit	IPR009779
Hyaluronidase 2*	IPR013618
DUF1736	IPR013618
Fascin*	IPR010431
IRSp53/MIM homology domain (IMD)	IPR013606
Survival motor neuron protein (SMN)	IPR010304
Spectrin	IPR002017
Translocon-assoc protein, gamma subunit (TRAP-gamma)	IPR009779
Follistatin-N-terminal domain-like (FOLN)*	IPR003645

Metazoa and Choanoflagellates

Antistasin family	IPR004094
Aspartyl/asparaginyl beta-hydroxylase	IPR007803
Associated with TFs and helicases	IPR006576
BTK motif	IPR001562
C1q*	IPR001073
Cadherin*	IPR002126
CBL proto-oncogene N-term, domain 1	IPR003153
CBL proto-oncogene N-term, EF hand-like	IPR003153
CBL proto-oncogene N-term, SH2-like	IPR003153
Collagen triple helix	IPR000087
Cu_2 monooxygenase	IPR003153
CUB	IPR000859
Dihydropyridine sensitive L-type calcium channel	IPR000584
DNaseIc*	IPR008185
Domain of unknown function (DUF758)	IPR008477
Domain of unknown function (DUF837)	IPR008555
ECSIT	IPR010418
Ependymin	IPR001299
Fibrillar collagen C-terminal	IPR000885
Filament	IPR001664
Flotillin*	IPR004851
Fukutin-related	IPR009644
Fzo-like conserved region	IPR006884
Galactose-3-O-sulfotransferase	IPR009729
Galactosyl transferase	IPR002659
Glycosyl hydrolase family 59*	IPR001286
GoLoco motif	IPR003109
Heme NO binding associated	IPR011645
Heparan sulfate 2-0-sulfotransferase*	IPR007734
Hormone receptor	IPR000536
Hormone-sensitive lipase (HSL) N-terminus	IPR010468
HYR*	IPR003410
Immunoglobulin	IPR013151
Immunoglobulin c-2*	IPR003598
Immunoglobulin I-set*	IPR013098
Immunoglobulin V-set*	IPR013106
Integrin alpha	IPR013519
Inward rectifier potassium channel	IPR013521

Kunitz/bovine pancreatic trypsin inhibitor*	IPR002223
L27	IPR004172
Laminin G*	IPR001791
Laminin N-terminal	IPR008211
Leucine rich repeat N-terminal	IPR000372
Low-density lipoprotein receptor class A	IPR002172
Mbt repeat	IPR004092
MOFRL family*	IPR007835
N-AcetylglucosaminyltransferaseIV(GnT-IV) conserved region	IPR006759
Nebulin repeat	IPR013998
Nine cysteines of family 3 GPCR	IPR011500
NRF (N-ternminal domain in C. elegans NRF-6)	IPR006621
Organic anion transporter polypeptide (OATP)	IPR004156
p53 DNA-binding	IPR011615
Pacifastin inhibitor (LCMII)	IPR008037
PET	IPR010442
Phosphomevalonate kinase	IPR005919
Phosphotyrosine binding (IRS-1 type)	IPR013625
Phosphotyrosine interaction (PTB/PID)	IPR006020
PHR	IPR012983
PI3-kinase family, p85-binding	IPR003113
Plexin	IPR013548
Progressive ankylosis protein (ANKH)	IPR009887
Protein of unknown function (DUF1241)	IPR009652
Raf-like ras-binding	IPR003116
Reeler	IPR002861
Renin receptor-like protein	IPR012493
Repeat in HS1/cortactin	IPR003134
S-100/ICaBP type calcium binding	IPR013787
Sarcoglycan complex subunit protein	IPR006875
Selenoprotein S (SelS)	IPR009703
Seven transmembrane receptor, secretin family	IPR000832
SH3 domain-binding protein 5 (SH3BP5)	IPR007940
Somatomedin B	IPR001212
Spin/Ssty family	IPR003671
STAT protein, DNA binding	IPR013801
TNF (Tumor Necrosis Factor)	IPR006052
Translocon-associated protein, delta subunit precursor	IPR008855
Tropomyosin	IPR000533
Uncharacterized protein family (UPF0121)	IPR005344
Von willebrand D*	IPR001846
Zinc finger, C2HC type	IPR002515

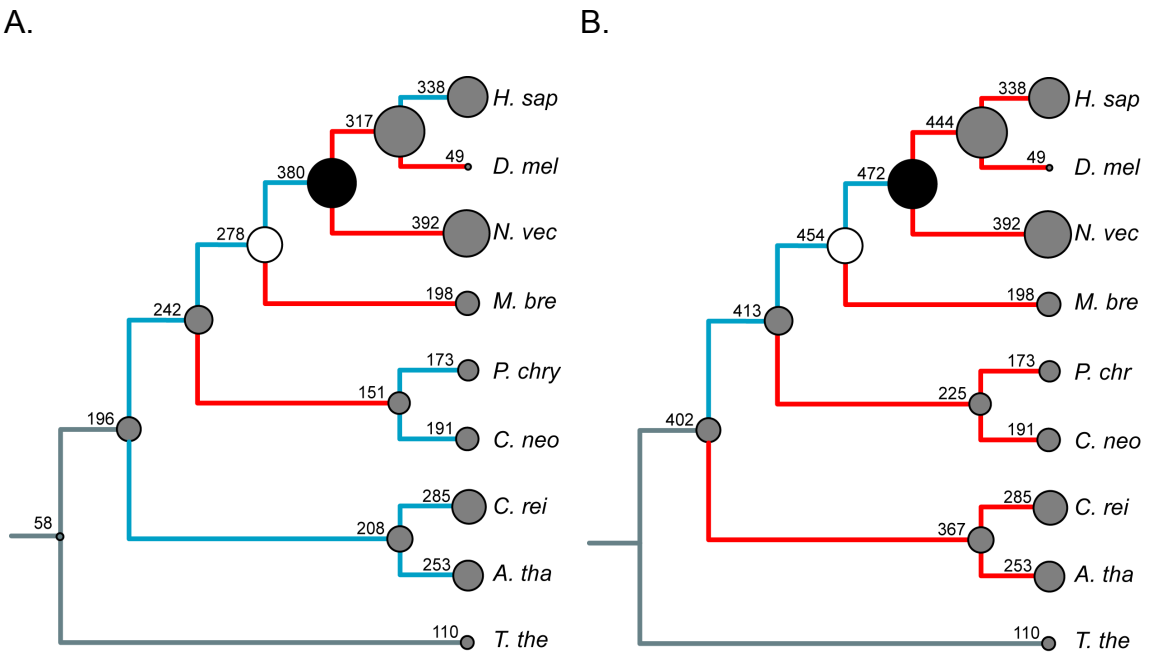
Fungi and Choanoflagellates

Anp1	IPR005545
YCII-related domain*	IPR005545

Table S2.3. Species included in comparative protein domain analysis

Dictyostelium	
Dictyostelium discoideum	*Dictyostelium discoideum AX4*
Fungi	
Aspergillus fumigatus	*Candida glabrata*
Cryptococcus neoformans	*Encephalitozoon cuniculi*
Eremothecium gossypii	*Kluyveromyces lactis*
Saccharomyces cerevisiae	*Schizosaccharomyces pombe*
Yarrowia lipolytica	
Metazoa	
Anopheles gambiae	*Apis mellifera*
Bos Taurus	*Caenorhabditis elegans*
Canis familiaris	*Ciona intestinalis*
Danio rerio	*Drosophila melanogaster*
Gallus gallus	*Homo sapiens*
Macaca mulatta	*Monodelphis domestica*
Mus musculus	*Pan troglodytes*
Rattus norvegicus	*Takifugu rubripes*
Tetraodon nigroviridis	*Xenopus tropicalis*
Unicellular eukaryotes	
Cryptosporidium hominis	*Cyanidioschyzon merolae*
Debaryomyces hansenii	*Giardia lamblia*
Monosiga brevicollis	*Plasmodium falciparum*
Thalassiosira pseudonana	

Figure S2.1. Intron evolution in Opisthokonta as assessed using Dollo parsimony and Roy-Gilbert maximum likelihood. Ancestral intron content and intron gains and losses were inferred using two additional methods: (A) Roy-Gilbert maximum likelihood and (B) Dollo parsimony methods. A sample of 1,054 intron positions in highly conserved sequences from 473 orthologs was used. Branches with at least 10% more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. Outgroup branches, for which intron loss could not be calculated, are grey. The inferred or observed number of introns present in ancestors and extant taxa are indicated next to proportionally sized circles. Species included are *Tetrahymena thermophila* (*T. the*), *Chlamydomonas reinhardtii* (*C. rei*), *Arabadopsis thaliana* (*A. tha*), *Cryptococcus neoformans* A (*C. neo*), *Phanerochaete chrysosporium* (*P. chr*), *Monosiga brevicollis* (*M. bre*), *Nematostella vectensis* (*N. vec*), *Drosophila melanogaster* (*D. mel*) and humans (*H. sap*).



Chapter 3: Exceptionally long exons in choanoflagellates, sponges and eumetazoans

SUMMARY

The intron-exon structure of eukaryotic genes varies greatly, both among species and among genes within the genomes of individual species. Complete genome sequences can be leveraged to study the extent of this variation more thoroughly. In the intron-rich choanoflagellates *M. brevicollis* and *S. rosetta*, two of the closest living relatives of animals, I found that the longest transcripts are encoded by genes consisting of exceptionally long exons (>10,000 bp), with remarkably few intervening introns. One *M. brevicollis* gene, which I named *gargantua*, contains the longest exon (59,595 bps) known in eukaryotes. The exon-intron structures of this and other *M. brevicollis* genes containing exceptionally long exons were experimentally validated. Genes with exceptionally long exons were also found in 19 of an additional 22 eukaryotic genomes analyzed, revealing that exceptionally long exons are a widespread, yet uncharacterized feature of eukaryotic genomes. Although widespread, the relative frequency of exceptionally long exons in the longest genes differs between choanoflagellates, sponges, and eumetazoans. In the choanoflagellates *M. brevicollis* and *S. rosetta*, and to a lesser extent in the sponge *Amphimedon queenslandica*, there is a genome-wide trend towards low intron-densities in the longest genes, and exceptionally long exons are accordingly common. In contrast, the longest eumetazoan genes are primarily intron-rich and exceptionally long exons are less frequent. The abundance of introns in the longest eumetazoan genes, and their striking absence from long choanoflagellate genes, highlights a difference in gene structure of potential importance to metazoan genome evolution.

INTRODUCTION

Unlike bacterial and archaeal genes, the protein coding sequences of eukaryotes are frequently interrupted by non-coding intronic sequence. Introns are an integral part of eukaryotic gene expression; they are transcribed and then removed from immature mRNAs by the spliceosome, a large complex of proteins and small RNAs [176]. The intron-exon structure of orthologous genes can vary markedly between species, and these differences are reflective of larger genome-wide trends [5, 47]. For example, the vertebrate *Xenopus tropicalis* has an average of 9.2 introns per gene (and over 250,000 introns in total) while there are only four introns in the entire genome of the unicellular *Giardia intestinalis* (Chapter 1). The evolutionary origin of these genome-wide differences in intron density has been the subject of many recent comparative genomics studies.

However, these genome-wide averages do not capture the full diversity of eukaryotic gene structure. Studies of genes that are extreme in some respect can reveal connections between gene structure, regulation and evolution. For example,

the extremely intron-rich *Drosophila Dscam* gene has illuminated how complex patterns of alternative splicing are generated [177, 178]. Additionally, a study of extremely long genes in eubacteria has uncovered a length-specific difference in tetranucleotide usage [179]. Genomes of less well-studied organisms may also contain unanticipated examples of extreme gene structures.

With the advent of next generation sequencing technologies, the number of sequenced eukaryotic genomes is rapidly increasing [180-182]. Among these are several unicellular lineages branching between metazoans and fungi whose genomes were sequenced in order to study the genomic underpinnings of multicellularity [183]. Within this group, the closest relatives of metazoans are the choanoflagellates [184], whose genome-enabled representatives are the species *Monosiga brevicollis* and *Salpingoeca rosetta*, and the parasitic *Capsaspora owczarzaki*, which is sister to choanoflagellates and metazoans [185]. Prior to the sequencing of these organisms, the lineage leading to metazoans had been associated with changes in gene structure and genome architecture, including widespread intron-gain [9]. Studies of gene structure in these unicellular outgroups may provide a finer resolution understanding of how such changes occurred, and also reveal previously unappreciated aspects of genome remodeling during metazoan origins.

Because choanoflagellates are phylogentically distant from well-studied model organisms, annotation of their genomes relies on a combination of *ab initio* gene prediction, homology searches with known genes from other genomes, and reference to sequenced cDNAs. Deep sequencing of cDNAs in particular provides experimental validation of gene predictions on a genome-wide scale [186] and sequenced cDNAs that span the junctions between two exons provide experimental support intron predictions. However, mapping short, spliced reads generated by the Illumina sequencing platform to a reference genome is technically challenging, and several methods attempting to produce accurate sequence maps have been developed [187-189]. These methods predict introns *de novo*, but some have proven to be less sensitive to short introns [190] like those found in choanoflagellates. An alternative approach for detecting introns is to computationally identify all possible introns based on the splice site signal sequence, and then determine which predictions have transcriptional support [106]. Using a combination of these methods to analyze deep transcriptome sequence data has the potential to provide accurate information on gene structure for a wide diversity of non-model eukaryotes.

In this study, I describe a novel observation about gene structure and utilize transcriptome data to experimentally validate the genes concerned. I have found that in the intron-rich choanoflagellates *M. brevicollis* and *S. rosetta*, the very longest genes have remarkably few introns relative to their overall length. They instead contain extremely long exons up to 60,000 base pairs in length. Transcriptome data from *M. brevicollis* indicated that these exons are transcribed in their entirety. A survey of diverse eukaryotic genomes showed that this type of gene structure is

widespread throughout eukaryotes but particularly prominent in choanoflagellates and the sponge, *Amphimedon queenslandica*. Contrastingly, the longest genes in eumetazoans are not generally depleted of introns, revealing a previously unappreciated difference in gene structure between eumetazoans and their closest relatives. The low density of introns in the longest genes of otherwise intron-rich species raises several questions about the evolution and functional impact of the gene structure of these already exceptional genes.

MATERIALS AND METHODS

ORF predictions
Genome sequences were downloaded from the following URLs: genome.jgi-psf.org (*Monosiga brevicollis*, *Nematostella vectensis*, *Chlamydomnas reinhardtii*, *Thalassiosira pseudonana*, *Ciona intestinalis, Trichoplax adheraens, Amphimedon queenslandica*), www.broadinstitute.org/science (*Neurospora crassa*, *Coprinus cinereus*, *Rhizopus oryzae*, *Aspergillus nidulans*), www.wormbase.org (*Caenorhabditis elegans*), flybase.org (*Drosophila melanogaster*), dictybase.org (*Dictyostelium discoideum*), www.yeastgenome.org (*Saccharomyces cerevisiae*), rgd.mcw.edu (*Rattus norvegicus*), www.plantgdb.org/AtGDB (*Arabidopsis thaliana*), www.ensembl.org/info/data/ftp/index.html (*Homo sapiens*), and www.hgsc.bcm.tmc.edu/projects/seaurchin (*Strongylocentrotus purpuratus*). ORFs were predicted using ORFfun.pl (© Jarrod Chapman, 2005) with no requirement for a start codon and a minimal length cutoff of 300 bps.

Validation of predicted gene structures by transcriptome sequencing and RT-PCR
Total RNA was isolated from *M. brevicollis* in log phase growth using the RNeasy isolation kit with on-column DNase treatment (Qiagen). Total RNA was isolated using the same method from *A. queenslandica* larvae, which were collected as previously described [191]. For RT-PCR, transcripts were reverse transcribed using gene-specific primers. Primers designed to flank exon-intron boundaries were used to confirm the splicing of specific introns.

For Illumina-based deep sequencing of the *M. brevicollis* transcriptome, total RNA was isolated as described above. The poly-A fraction was then purified using Dynal oligo(dT) beads (Invitrogen). The mRNA was fragmented to approximately 200-500 bps using a zinc acetate fragmentation buffer (Ambion). Double stranded cDNA was generated using random hexamer primers. I designed adaptors and multiplexing sequencing primers according to instructions provided by illumina. The adaptors were ligated to the cDNA as described in the Illumina RNA-seq library preparation protocol, with the exception of the adaptor mix concentration, which I diluted by a factor a ten to reduce the occurrence of adaptor concatemers. The libraries were size selected by using gel electrophoresis and excising a band slightly larger than 300 bps. The multiplexing sequencing primers were then used to amplify the library with 18 cycles of PCR. The quality and quantity of the library was assessed by biolanalyzer and qPCR-based assays.

101 bps paired ends reads were sequenced in two lanes of an Illumina GA2 sequencer. The total amount of sequence obtained was 10.3 Gbps. The spliced read alignment program Tophat was used to map the sequence to the genome [187]. 55.0% of this sequence was mapped uniquely to the *M. brevicollis* reference genome. To search for un-annotated introns in pre-existing genes, the package juncBASE was used to create a database of all possible introns based on the canonical dinucleotide splice site sequences between transcriptional start and stop sites [106]. The reads were then mapped to the resulting hypothetical exon-exon junction sequences using the short read alignment program Bowtie [192].

Protein domain predictions
Protein domains were predicted in all genes by Pfam release 23.0 [193]. Significantly overrepresented domains in long ORFs were determined by a chi-square test of independence comparing the occurrence of a given domain in long ORFs to its occurrence in the genome as a whole. Domain combinations found uniquely in *M. brevicollis* elORF genes were determined by using the online Pfam domain architecture tool.

Gene models and intron density analysis
Predicted gene models for the *M. brevicollis*, *S. rosetta*, *A. queenslandica*, *N. vectensis, T. adhaerans, H. robusta, L. gigantea, D. melanogaster, C. intestinalis and X. tropicalis* genomes were downloaded from the Joint Genome Institute (JGI) genome portal database (http://genome.jgi-psf.org/). Gene models for the *D. melanogaster, C. intestinalis and X. tropicalis* genomes were downloaded from Ensembl (relases 62) and the *H. sapiens* gene models from NCBI (release 36.53). The number of exons per gene was compared to the length of that gene's spliced transcript (not including introns) using PERL. The set of *M. brevicollis* gene models used in this analysis was automatically generated by the JGI and did not include several elORF-containing gene models that were experimentally validated by transcriptome data, and likely provides a conservative description of intron sparseness at the upper end of the length spectrum.

RESULTS

Large genes with exceptionally long exons
To investigate the early evolution of animal gene structures, I characterized the distributions of exon size and intron density in all predicted genes from the filasterean *C. owczarzaki*, the choanoflagellates *M. brevicollis* and *S. rosetta* and the sponge *A. queenslandica*. Typical genes from *C. owczarzaki, M. brevicollis*, *S. rosetta* and *A. queenslandica* are relatively intron-rich (averaging 3.8, 6.6, 7.6 and 4.0 introns/gene respectively) and have exons averaging 429, 247, 264 and 211 base pairs in length respectively. In contrast to the short exons typical of most of their genes, each of these species contains a set of atypical genes with unusually long exons, many of which exceed 10,000 bps in length. To identify the full complement of extremely long exons in each of these genomes while avoiding potential biases

from automated gene predictions, I used genomic open reading frames as a proxy for exons, reasoning that extremely long open reading frames represent coding sequences in which there has been selection against stop codons. By searching for genomic open reading frames, I found that the *M. brevicollis* and *S. rosetta* genomes contain 46 and 51 ORFs, respectively, that are longer than 10,000 bps (hereafter elORFs, for extremely long Open Reading Frames). Likewise, both the *C. owczarzaki* and *A. queenslandica* genomes contain 27 elORFs (Figure 1). In support of the assumption that these elORFs represent protein coding exons, no elORFs were detected in an artificial randomized genome of the same size and nucleotide content as the *M. brevicollis* genome, and the longest open reading frame was only 1,538 bps (Figure 1).

The abundance of elORFs in choanoflagellates and sponges raised the possibility that genes with comparably long exons are present in other genomes. To assess if elORFs are a common feature of eukaryotic genomes, I surveyed the genomes of representative animals, fungi, plants, and diverse single-celled eukaryotes. In 19 of the 22 eukaryotic genomes analyzed I identified multiple elORFs (Figure 1, Table S1), revealing that elORFs are phylogenetically widespread.

To gain insight into the evolutionary history and functional significance of elORFs, I tested whether relative elORF abundance in different species correlates either with distinguishing genomic characteristics or phylogeny. The relative abundance of elORFs in related lineages shows no strong phylogenetic pattern, implying that they are evolutionarily labile (Figure 1). In addition, there is no correlation between elORF abundance and genome size, contrary to what would be expected if elORFs occur with greater probability either in larger or more streamlined genomes. Likewise, there is no correlation between median ORF length and elORF abundance, suggesting that elORFs are not a secondary consequence of a genome-wide trend toward longer exons. Finally, there is no obvious connection between elORF abundance and the number of cell types or developmental complexity [194]. While the organisms with the most elORFs, *M. brevicollis* and *S. rosetta*, have relatively small genomes and little cell differentiation, elORFs are also found in complex multicellular organisms such as *D. melanogaster* and humans.

Validation of elORF-containing gene predictions

Because elORF-containing genes in choanoflagellates were predicted *ab initio* from draft genome sequences [195], experimental validation of these predictions was necessary. The size of elORF-containing genes makes traditional cloning and sequencing of cDNAs impossible. However, I was able validate the *M. brevicollis* elORF gene predictions using high throughput transcriptome sequencing, which yielded 10,250 Mbps of sequence (244 and 371-fold coverage of the *M. brevicollis* genome and transcriptome respectively), of which approximately 55% of the sequence reads aligned uniquely to the genome. In addition to showing that an elORF is transcribed across its entire length, this data can also be used to validate predicted introns and search for unpredicted introns within the elORFs.

These data revealed that the longest elORF from *M. brevicollis*, which spans 59,595 bps and was the longest ORF detected in all of the genomes analyzed, is transcribed in its entirety. The 59,595 bp elORF is the 3'-most exon of a seven-exon gene, which I named *gargantua* (Figure 2A). The predicted length of the entire *gargantua* transcript is 83,448 bps. The sequence coverage was higher at the 3' end of *gargantua*, presumably due to strand breakage during poly-A selection of mRNA. Nonetheless, sequence reads were aligned to the entire length of the transcript (Figure 3.2b).

Introns may be predicted independently of a pre-exsisting gene annotation based on the split alignment of sequence reads. I compared these intron predictions to the pre-existing elORF gene annotations. The *gargantua* gene annotation contains six predicted introns, and while four of these were predicted based on split read alignments, no additional introns were predicted based on split alignments within the extremely long exon (Figure 3.2A). To validate this result, I performed reverse transcriptase PCR (RT-PCR) across the six predicted introns. RT-PCR showed that all six predicted introns were spliced out of mature transcripts (Figure S3.1), indicating that Tophat failed to detect two introns. These introns were at the 5' end of the gene and were likely missed due to low sequence coverage.

I extended my analysis of the *M. brevicollis* transcriptome data to include all 44 elORF genes (two of which contain two elORFs). Within the entire set of *M. brevicollis* elORFs, there is no evidence of introns based on split read alignments, supporting their transcription as single exons. As a set, the genes containing elORFs have 227 introns, of which 52 are supported by split read alignments (Table S3.2). For the genome as a whole, approximately two thirds of the introns in the gene predictions have similar transcriptional support. As with *gargantua*, the sequence coverage in the other elORFs is biased to their 3' ends (Figure S3.2) and, additionally, some of these genes are not transcribed under standard growth conditions. Therefore, where sequence coverage was high (i.e. at the 3' ends of transcribed elORF genes), it was possible to rule out the existence of unpredicted introns, whereas the low coverage at the 5' ends of elORFs reduced my confidence about whether their might be unpredicted and undetected introns. However, the fact that there are no introns detected in the many regions of high sequence coverage within the elORFs, in combination with the absence of stop codons within these extremely long stretches of sequence, strongly supports their transcription as single exons uninterrupted by introns.

To provide a second test for the presence of unpredicted introns within the elORFs, I used the alternative splicing analysis package juncBASE [106] to create a database of all potential exon-to-exon junction sequences within elORFs based on the canonical AT|AG splice site sequence. 3,521,286 hypothetical exon-exon junctions were identified using this method. I then aligned the reads to these sequences, requiring that potentially novel introns have at least four reads at four different offsets spanning the corresponding exon-exon junction with a greater than 5 bp overlap. This prevented the misidentification of novel introns due to mismappings

of repetitive sequence or short pieces of sequence with chance similarity to the canonical splice site. Of the approximately 3.5 million candidate junction sequences within the elORFs, only two in distinct elORFs had reads alignments that met these criteria, and both were borderline cases. These potential novel introns occur within two distinct elORFs. Transcriptional data thus finds little evidence for unpredicted introns within elORFs, and instead indicates that these sequences are transcribed as single exons.

Protein domain composition and conservation of M. brevicollis *elORF genes*
To gain insight into the functions of proteins encoded by elORF genes, I investigated the protein domain composition. The *gargantua* gene encodes a predicted transmembrane protein containing multiple extracellular domains including EGF, von Willebrand factor D, and C8 domains, each of which function in diverse animal cell adhesion and extracellular matrix proteins (Figure 2A). Like *gargantua*, many other elORFs from *M. brevicollis* encode protein domains with extracellular functions. Of the 46 elORFs, 27 (58%) encode extracellular protein domains. In contrast, a significantly lower percentage of all genes in the *M. brevicollis* genome encode predicted extracellular domains (33%, Figure 3A). Three specific extracellular domains involved in signaling and adhesion in animals -- cadherin, EGF, and TNFR c6 domains -- are enriched in elORF-encoded proteins relative to the proteome as a whole (Figure 3A). Transmembrane domains are similarly overrepresented in elORF genes. The high frequency of extracellular and transmembrane domains suggests that many elORF-containing genes in *M. brevicollis* encode signaling or adhesion receptors.

Of the 44 elORF genes in *M. brevicollis* (two of which contain two elORFs), 43 have best reciprocal blast hits to an *S. rosetta* gene (which does not necessarily contain an elORF), indicating that the majority of these genes are conserved within choanoflagellates. In contrast, only four of the *M. brevicollis* elORF genes have clear orthologs in species outside of choanoflagellates. All four genes belong to the dynein heavy chain family, which consists of large motor proteins involved in vesicle trafficking and ciliary movement. Additional similarity was detected by BLAST between *M. brevicollis* and *A. queenslandica* elORF genes, but closer examination revealed that this was due shared sets of protein domains (Table S3). These domains typically occurred in different combinations in the two species and therefore did not suggest direct orthology between most elORF genes in choanoflagellates and sponges.

Intron sparseness and loss in long genes
Because of the abundance of elORFs in choanoflagellates, I hypothesized that choanoflagellates have a tendency to simplify the intron-exon structure of very long genes. If true, one would expect a general deficit of introns in all long transcripts. In contrast, if intron density is unrelated to transcript length, then the number of exons per gene should, on average, increase linearly with increasing transcript length. To test these alternative predictions, I examined the relationship between exon number and transcript length in all *M. brevicollis* and *S. rosetta* genes. Unlike shorter

transcripts (i.e. those under 10,000 bps in length), almost all (specifically 77 of 80 in *M. brevicollis* and 127 of 134 in *S. rosetta*) transcripts over 10,000 bps were encoded by genes with fewer introns than predicted based on the genome-wide averages (Figure 4, panels A and B).

I also examined the relationship between exon number and transcript length in metazoan genomes, including the basal animals *A. queenslandica*, *T. adhaerans*, and *N. vectensis,* and the bilaterians *L. gigantea*, *H. robusta*, *D. melanogaster*, *C. intestinalis*, *X. tropicalis* and *H. sapiens* (Figure 4C-F and S2). Genes in the early branching animals *A. queenslandica* and *T. adhaerens* also trended toward intron depletion as transcript length increased (Figure 4D), but not to the same extent as was observed in choanoflagellates. In contrast, the longest genes from *N. vectensis* and bilaterians contained high densities of introns and there was no genome-wide intron depletion in long genes (Figure 4, panels E and F, and Figure S2), making the intron-depletion of the longest genes in choanoflagellates more notable.

One possible explanation for the relative intron sparseness of long transcripts in choanoflagellates is that introns were lost from progenitors of these genes. In some species that have undergone genome-wide intron loss, the remaining introns show a positional bias towards the 5' ends of genes [196]. This bias may be due to a reverse transcriptase (RT) mediated mechanism of intron loss, whereby cDNAs reverse transcribed from the ends of messenger RNAs recombine homologously with the 3' ends of genes [197, 198]. In the *M. brevicollis gargantua* gene, the few introns are concentrated at the 5' end of the gene (Figure 1), consistent with the hypothesis of RT-mediated intron loss. Intron positions in all *M. brevicollis* elORF genes are slightly biased toward the 5' end, with 61% of introns occurring in the first half of the gene. In contrast, there is no such bias in intron position in the genome as a whole (51% of all introns occur in the first half of the gene).

DISCUSSION

elORF genes and the evolution of metazoan genomes

Although genes of average length in choanoflagellates are intron-rich, the longest genes are not. Instead, they frequently contain elORFs, exceptionally long stretches of coding sequence uninterrupted by introns. Indeed, these gene features were found in nearly all eukaryotic genomes analyzed.

Genes containing elORFs are abundant in the choanoflagellates *M. brevicollis* and *S. rosetta* and the early branching metazoans *A. queenslandica* and *T. adhaerens*, suggesting that this unusual gene structure was prominent in the last common ancestor of metazoans or that the biology of these organisms favors the evolution of elORFs. In contrast, exceptionally long genes in *N. vectensis* and bilaterians generally contain as many or more exons than expected based on the genome-wide average intron-density, suggesting that the exon-intron structure of long genes has changed during metazoan evolution.

Intron abundance in long genes has consequences for gene function and genome evolution in metazoans. In eumetazoans, alternative splicing of long, multi-exon genes has important roles in development and cell differentiation [88, 199]. For example, *Titin*, the longest gene in the human genome (and the only eukaryotic gene longer than *gargantua* as of Genbank release 171.0 [200]), contains 363 exons that are spliced into diverse isoforms that perform a wide array of functions in smooth and striated muscle cells [201]. In flies, the gene *Dscam* is broken up into 115 exons and can be alternatively spliced into as many as 38,016 isoforms, whose diversity regulates axon guidance in the developing nervous system [202]. This degree of alternative splicing would not be possible in an intron-depleted gene like *gargantua*, potentially limiting its functional coding capacity.

Gene structure also has important consequences for the evolution of new genes in metazoans. Multi-domain proteins often evolve through exon-shuffling, which is hypothesized to occur by non-homologous recombination in introns [203]. It has been proposed that exon-shuffling is a particularly important mode of gene evolution in metazoans [174]. In contrast to the intron-rich long genes of eumetazoans, the intron-poor elORF genes in *M. brevicollis* would leave little opportunity for evolution via this mechanism.

The evolution of exon size in long genes

The relative paucity of introns in long genes in choanoflagellates is a genome-wide trend; there are hardly any genes greater than 10,000 bps in length with an intron density similar to the genome-wide average. One explanation for this observation is that it might be selectively disadvantageous for the longest genes to contain many introns. Introns might be problematic for the transcription of long genes in organisms with rapid generation times, such as *M. brevicollis*. If a 10,000 bp stretch of coding sequence in *M. brevicollis* contained the expected 35 introns of average length, RNA polymerase II would have to transcribe approximately 6000 additional base pairs. The rate of transcription of large human genes has been measured at approximately 4 kbp per minute [44], making this amount of additional sequence non-trivial. Furthermore, the kinetics of splicing an intron, which occurs on the order of minutes, [204], which could further constrain the kinetics of expressing long, intron-rich genes.

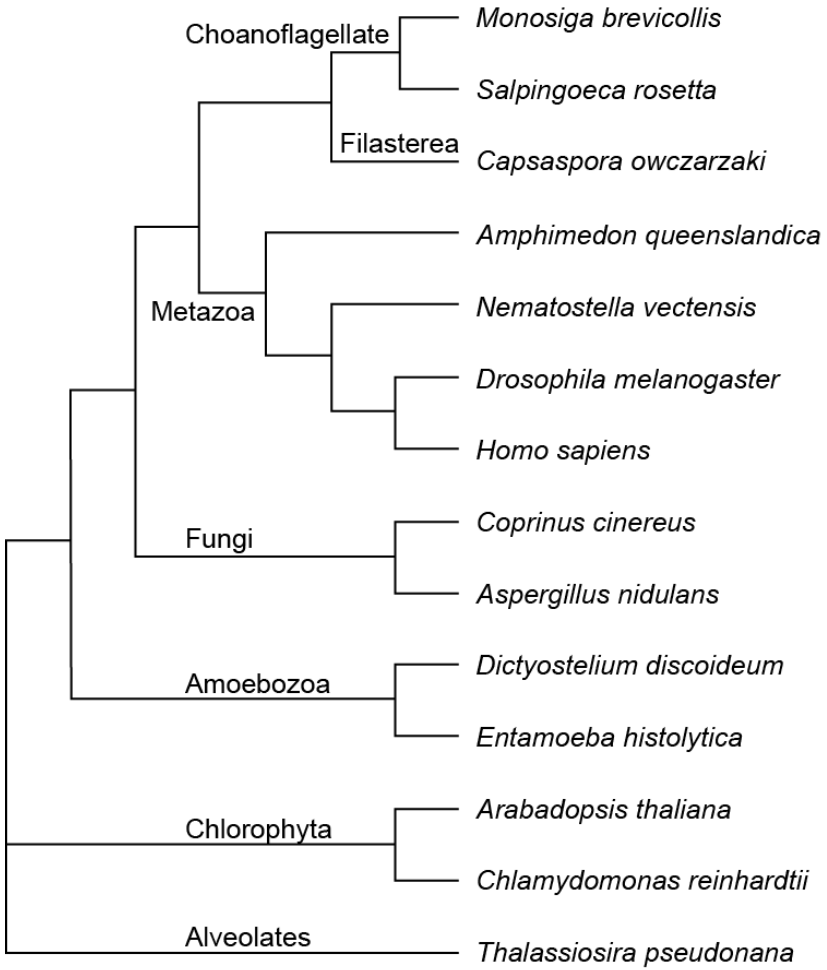
The finding that a specific class of genes has a distinct exon-intron structure is not without precedence. In humans and *C. elegans*, highly expressed genes have shorter-than-average intron lengths [52] and, in a wide variety of genomes, genes whose transcription is rapidly modulated are intron poor [205]. In addition, in eubacteria it has been shown that most of the cost of gene expression can be attributed to the processes of transcription and translation, not amino acid usage in the final protein product [206]. These studies, in combination with our findings, raise the possibility that low intron-densities are advantageous for certain classes of genes due to the kinetics and energetic costs of transcription and splicing.

Though the lack of widespread conservation of elORF genes makes it difficult to reconstruct their evolutionary history, several lines of evidence suggest that their unusual gene structure is the result of intron-loss from a more intron-rich ancestral state. Many *M. brevicollis* elORF genes contain conserved protein domains, but in entirely unique combinations, suggesting that they evolved by re-arrangements of pre-existing domains. Because multidomain proteins are hypothesized to evolve by exon shuffling [29], these unique combinations of domains within a single exon may reflect ancient exon shuffling events followed by intron loss. In addition, introns in elORF genes are slightly biased to occur towards the 5' ends of genes, consistent with RT-mediated intron loss [198]. RT-mediated intron loss also provides a mechanism by which many introns can be lost at once [197], and could lead to the formation of an elORF in one event rather than many sequential intron losses. Another potential mechanism for the evolution of elORFs is tandem segmental duplications [207]. Many elORFs contain stretches of repeated protein domains, suggesting that internal duplications, in addition to intron loss, contributed to elORF evolution.

In contrast with choanoflagellates, the longest genes in eumetazoans have not undergone a similar "streamlining". Why would intron loss in the longest genes occur in certain groups but not others? Lineage-specific variations in genome-wide rates of intron-loss have been reported in many studies [55, 58, 60, 63]. One proposed explanation is that introns are slightly disadvantageous because they increase mutational load, and therefore only persist in species with small population sizes wherein they can drift to fixation [9]. In contrast, in species with large population sizes introns are efficiently removed by negative selection. As discussed above, introns in long genes may be inherently more disadvantageous than ones in shorter genes due to the kinetics of transcription and splicing. Perhaps in choanoflagellates, the balance of selection and genetic drift is such that only the more detrimental introns in long genes are efficiently selected against. Another possibility is that introns in long genes in groups such as metazoans have become advantageous through their role in enabling complicated patterns of alternative splicing, and are maintained as a result of this function. Further studies on the link between intron conservation and the alternative splicing of long genes in eumetazoans may shed light on the forces influencing the evolution of this particular gene structure.

Evolutionary origins aside, the relative abundances of elORF genes represent a previously unappreciated difference in gene structure between choanoflagellates, sponges and eumetazoans, groups that have followed distinct evolutionary trajectories. This difference could have important consequences for the function and evolution of large, multidomain proteins in metazoans.

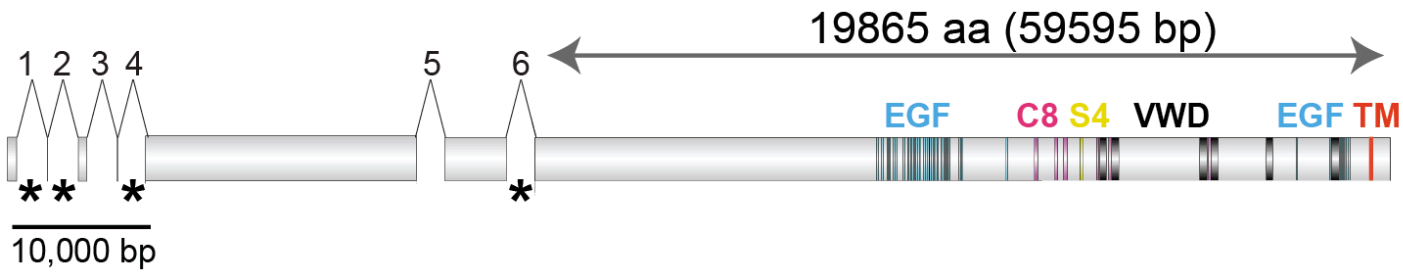
Figure 3.1. Unusually long ORFs and genome characteristics in diverse eukaryotes. *M. brevicollis* has an unusual abundance of genomic open reading frames (ORFs) greater than 10,000 base pairs. A randomized genome the same size and GC content as *M. brevicollis* had no ORFs greater than 10,000 bps, indicating that ORFs of this size do not occur by chance. ORFs were used as proxies for exons in this analysis to avoid potential biases of gene predictions. The lower length limit on ORF predictions was 300 bp, hence the average ORF length is greater than the average exon length for each species, respectively. Similar data for nine additional species is included in the Table S3.1.



	> 10,000 bp ORFs (#)	Longest ORF (bp)	Median ORF Length (bp)	Genome size (Mbp)
Monosiga brevicollis	46	59,597	395	41
Salpingoeca rosetta	51	27,443	428	55
Capsaspora owczarzaki	27	25,118	416	30
Amphimedon queenslandica	27	47,840	383	170
Nematostella vectensis	8	16,808	377	340
Drosophila melanogaster	10	27,803	371	180
Homo sapiens	15	21,851	359	2,910
Coprinus cinereus	0	8,468	392	38
Aspergillus nidulans	8	17,057	392	32
Dictyostelium discoideum	29	29,370	461	34
Entamoeba histolytica	2	15,215	470	24
Arabadopsis thaliana	0	7,524	380	125
Chlamydomonas reinhardtii	6	12,629	422	120
Thalassiosira pseudonana	6	17,513	407	34
Randomized *M. brevicollis* genome	0	1,538	350	41

Figure 3.2. Intron-exon structure of *M. brevicollis gargantua*, a gene containing the longest known eukaryotic exon. (A) *Gargantua* is a seven-exon gene that encodes a 27,816 amino acid protein. The last exon spans 59,595 base pairs and was identified as the longest ORF in our analysis of diverse eukaryotic genomes. This exon encodes multiple extracellular protein domains, including epidermal growth factor (EGF), C8, S4, von Willebrand type D (VWD), and transmembrane (TM) domains. Asterisks indicate introns that were predicted using RNA-seq data aligned by the spliced read-mapper tophat. In addition, all possible combinations of the dinucleotide acceptor and donor splice sites were used to generate a database of potential spliced sequences and the reads were then remapped to these sequences, but no potential introns were identified using this method. (B) RNA-seq data supporting the transcription of *gargantua*. Bars indicate the average number of reads that align to each 10 basepair segment of the transcript. There is a 3' coverage bias (presumably due to strand breakage during poly-A selection). Two different scales are shown to provide appropriate resolution for each end of the transcript.

A


19865 aa (59595 bp)
1 2 3 4
5
6
EGF
C8 S4
VWD
EGF TM
10,000 bp

B

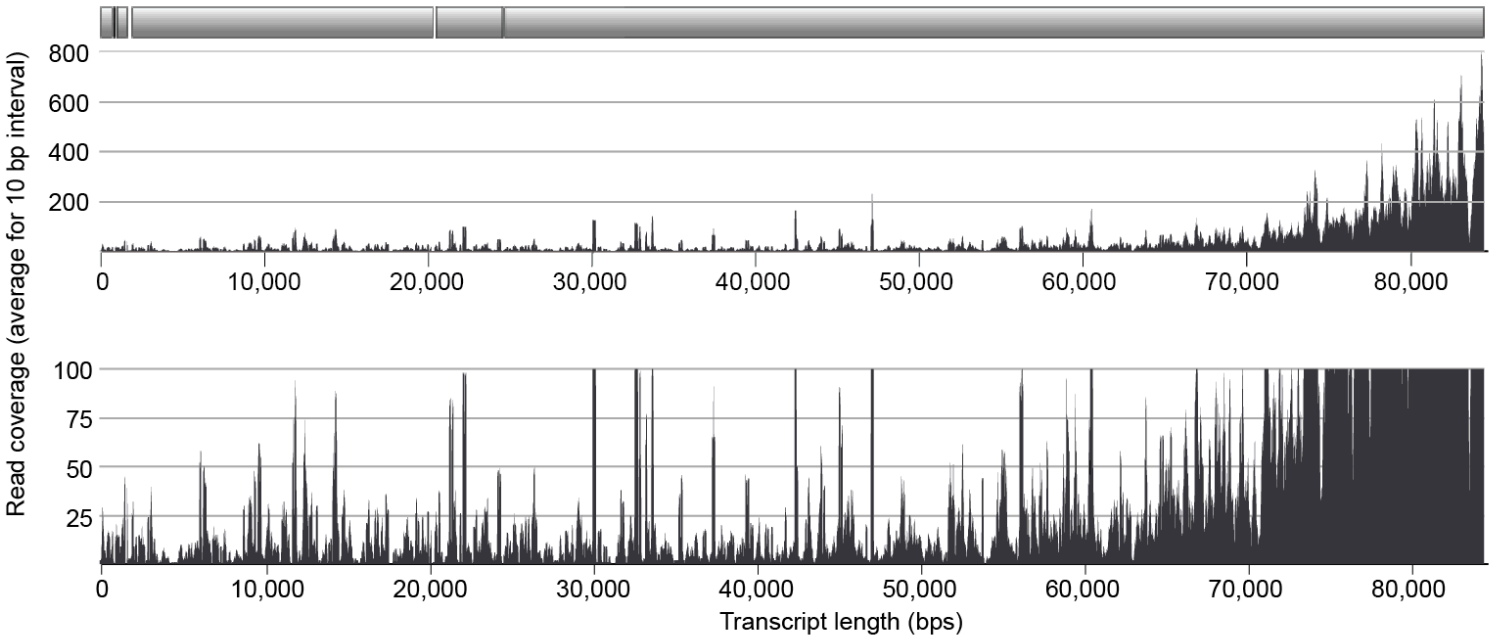
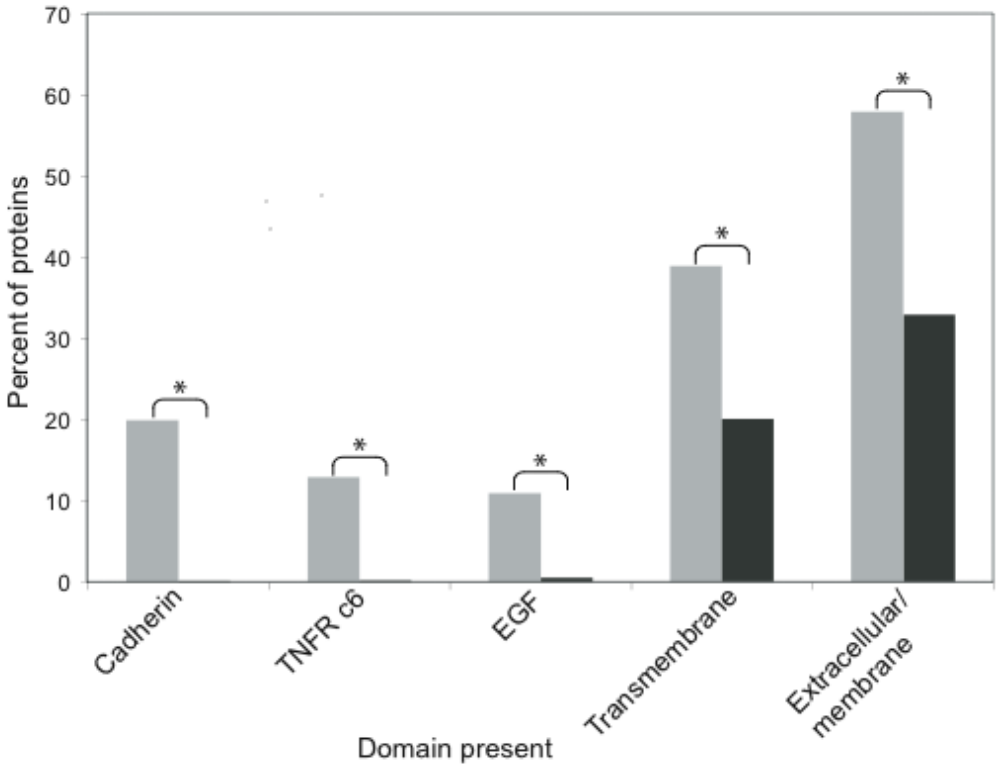

Read coverage (average for 10 bp interval)
800
600
400
200
0
10,000
20,000
30,000
40,000
50,000
60,000
70,000
80,000
100
75
50
25
Transcript length (bps)

Figure 3.3. Abundance and novel organization of extracellular protein domains in ulORF-containing genes. (A) The occurrence of Pfam-predicted protein domains [193] was compared between *M. brevicollis* ulORF genes (grey bars) and all genes (black bars). Cadherin, EGF (epidermal growth factor), TNFR_c6 (tumor necrosis factor c6), transmembrane and extracellular domains, as defined by GO annotations [208] of Pfam domains, are significantly more common in ulORF genes than in the *M. brevicollis* genome as a whole. Asterisks (*) indicate a chi-square p-value of less than 0.01. (B) Novel domain combinations were identified in *M. brevicollis* ulORFs by using the online Pfam domain architecture tool, which includes protein domain architectures from all Uniprot and Genpept sequences. Ten ulORFs contained domain combinations not found in any archived protein sequence. The length of the ulORF and the novel domain combination is shown. Extracellular domains are shaded.

A.


70
60
50
40
30
20
10
0
Percent of proteins
*
*
*
*
*
Cadherin
TNFR c6
EGF
Transmembrane
Extracellular/ membrane
Domain present

B.

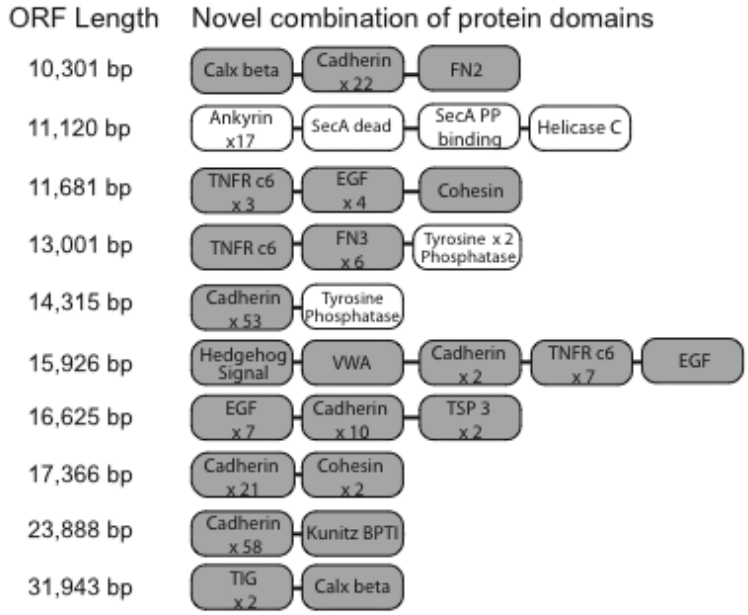

ORF Length
Novel combination of protein domains
10,301 bp
Calx beta
Cadherin x 22
FN2
11,120 bp
Ankyrin x17
SecA dead
SecA PP binding
Helicase C
11,681 bp
TNFR c6 x 3
EGF x 4
Cohesin
13,001 bp
TNFR c6
FN3 x 6
Tyrosine x 2 Phosphatase
14,315 bp
Cadherin x 53
Tyrosine Phosphatase
15,926 bp
Hedgehog Signal
VWA
Cadherin x 2
TNFR c6 x 7
EGF
16,625 bp
EGF x 7
Cadherin x 10
TSP 3 x 2
17,366 bp
Cadherin x 21
Cohesin x 2
23,888 bp
Cadherin x 58
Kunitz BPTI
31,943 bp
TIG x 2
Calx beta

Figure 3.4. Genome-wide intron density in *M. brevicollis* and animal genes. In *M. brevicollis* (A) and *S. rosetta* (B), almost all genes encoding transcripts greater than 10,000 bps (the unshaded portion of the graph) have fewer introns/kb than the genome-wide average. The number of exons per gene was plotted versus the length of the spliced transcript. For comparison, the number of exons predicted based on the average introns/kb of transcript is shown by a solid line. Regression analysis (dashed line) shows a negative trend in the number of exons for transcripts greater than 10,000 bps. The negative trend between length and exon number in the longest transcripts is also observed, though to a lesser extent, in *A. queenslandica* (C) and *T. adhaerans* (D). Contrastingly, in the eumetazoans *N. vectensis* (E) and *H. sapiens* (F) the longest transcripts are spliced from many exons and there is a positive trend in exon number for transcripts greater than 10,000 bps.

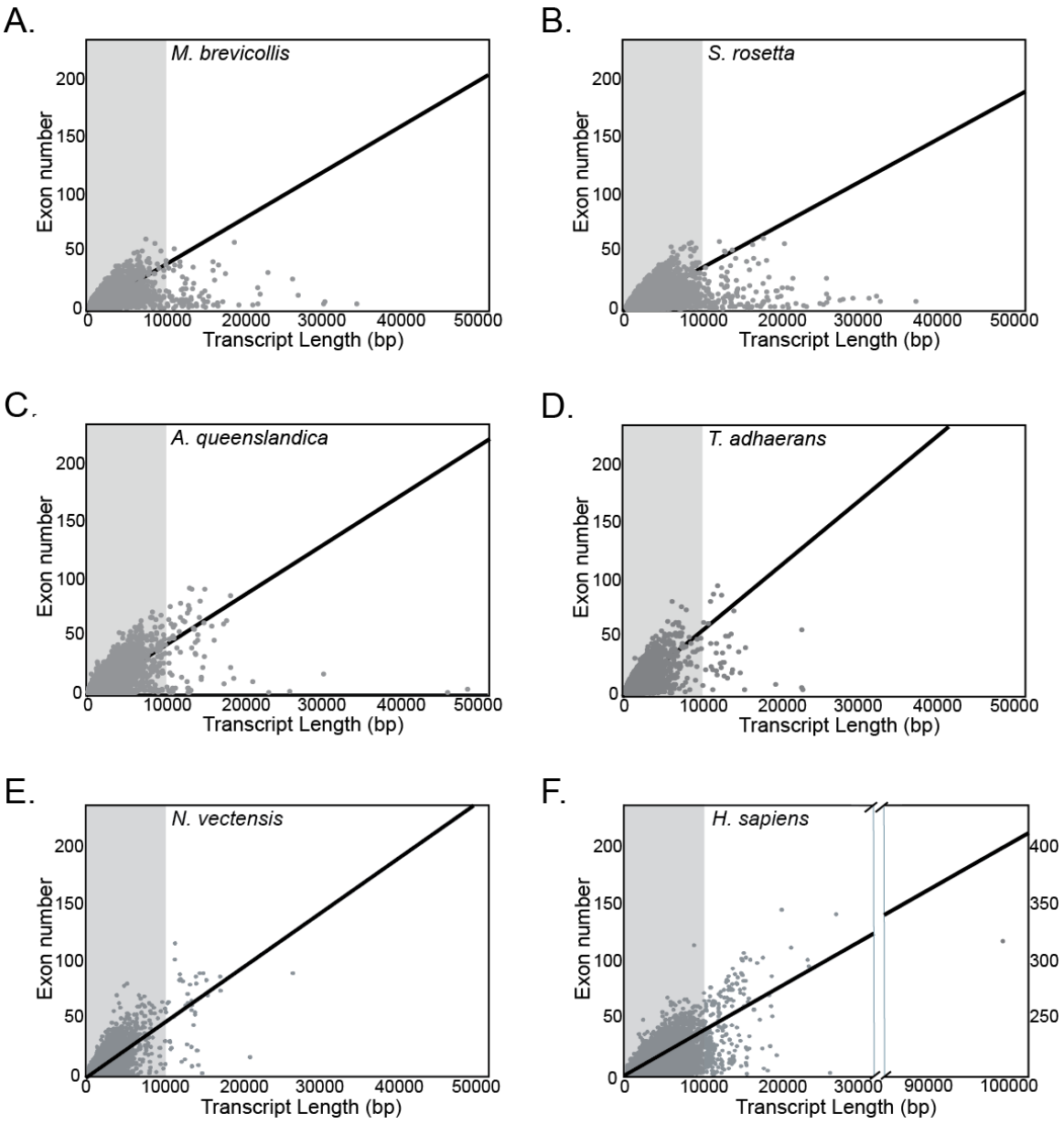

A.
M. brevicollis
B.
S. rosetta
C.
A. queenslandica
D.
T. adhaerans
E.
N. vectensis
F.
H. sapiens
Exon number
Transcript Length (bp)
0
50
100
150
200
10000
20000
30000
40000
50000
90000
100000
250
300
350
400

Table S3.1. Comparison of unusually long ORFs in phylogenetically diverse species

Classification[1]	Species	# ulORFs	Longest ORF (bp)	Median ORF Length (bp)	Genome Size (Mbp)
Choanoflagellates	*Monosiga brevicollis*	47	59,597	395	41
Choanoflagellates	*Salpingoeca rosetta*	51	27,443	428	55
Excavates	*Giardia lamblia*	38	24,216	422	12
Amoebazoa	*Dictyostelium discoideum*	29	29,370	461	34
Metazoa	*Amphimedon queenslandica*	27	47,840	383	170
Filasterea	*Capsaspora owczarzaki*	27	25,118	416	30
Metazoa	*Homo sapiens*	15	21,851	359	2,910
Metazoa	*Strongylocentrotus purpuratus*	14	22,331	356	800
Metazoa	*Drosophila melanogaster*	10	27,803	371	180
Metazoa	*Rattus norvegicus*	10	19,880	356	2750
Fungi	*Aspergillus nidulans*	8	17,057	392	32
Metazoa	*Nematostella vectensis*	8	16,808	377	340
Plants	*Chlamydomonas reinhardtii*	6	12,629	422	120
Heterokonts	*Thalassiosira pseudonana*	6	17,513	407	34
Metazoa	*Ciona intestinalis*	5	18,621	353	160
Metazoa	*Caenorhabditis elegans*	5	15,002	368	97
Fungi	*Neurospora crassa*	5	15,068	392	40
Amoebozoa	*Entamoeba histolytica*	2	15,215	470	24
Fungi	*Sacharomyces cerevisiae*	1	14,451	494	13
Plants	*Arabidopsis thaliana*	0	7,524	380	125
Fungi	*Coprinus cinereus*	0	8,468	392	38
Fungi	*Rhizopus oryzae*	0	9,833	407	35
	Randomized genome	0	1,538	350	41

Table S3.2. RNA-seq based validation of ulORF gene predictions

Tophat spliced read alignments			Denovo exon junction alignments	
Number of tophat predicted junctions within ulORFs	Number of introns in ulORF containing gene models	Number of introns in ulORF genes with supporting reads	Total number of denovo exon junctions within ulORFs	Number of denovo exon junctions in elORFs with transcriptional support
0	227	52	3,521,286	2

Table S3.3. Protein domains identified in *M. brevicollis* and *A. queenslandica* ulORFs

Extracellular (E), Intracellular (I), Membrane (M)	PFAM Protein domain	*M. brevicollis* ulORFs containing domain	*M. brevicollis* Total number in ulORFs	*A. queenslandica* ulORFs containing domain	*A. queenslandica* Total number in ulORFs
E	Cadherin	10	205	7	216
E	TNFR c6	6	24	0	0
E	EGF	5	26	2	3
E	IPT/TIG	4	110	1	3
E	EGF 2	2	11	0	0
E	Von Willebrand D	1	7	0	0
E	C8	1	6	0	0
E	Thrombospondin type 3 repeat	1	2	0	0
E	Laminin G2	1	1	1	1
E	Fibronectin 2	1	1	0	0
E	WAP	1	1	0	0
E	Laminin EGF	1	1	0	0
E	Fibronectin 3	1	1	0	0
E	Von Willebrand A	1	1	0	0
E	Hedgehog signal	1	1	0	0
E	Kunitz BPTI	1	1	0	0
E	Plethodontid receptivity factor	0	0	1	1
E	Immunoglobulin V-set	0	0	1	1
M	GPCR proteolytic site	1	1	0	0
I	AAA	5	6	0	0
I	Dynein heavy chain N-terminal 2	4	4	0	0
I	Dynein heavy chain	4	4	0	0
I	HEPN	0	0	4	4
I	Calx-beta	3	5	0	0
I	Cohesin	3	4	0	0
I	Tyrosine phosphatase	3	3	0	0
I	Dynein heavy chain N-terminal 1	2	2	0	0
I	Zinc finger C3HC4	2	2	0	0
I	Ankyrin	1	17	1	87
I	Photosystem I reaction center	1	1	0	0
I	Ubiquitin conjugating enzyme	1	1	0	0
I	SecA DEAD	1	1	0	0
I	SecA preprotein crosslinking	1	1	0	0
I	Helicase_C	1	1	0	0
I	Cu/Zn superoxide dismutase	1	1	0	0
I	Proprotein convertase P	1	1	1	1
I	Bacterial neuraminidase repeat	0	0	1	3

Figure S3.1. RT-PCR validation of Gargantua exon-intron structure. In addition to RNA-seq, Reverse-Transcriptase (RT) PCR across predicted intron junctions was used to confirm the exon/intron structure of *Gargantua*. The genomic DNA lanes show the size of unspliced product while the cDNA lanes show the predicted shorter product (indicated by a star) that results when the intronic sequence is spliced out. For intron 1, a band the size of the unspliced product is seen in addition to the spliced product, indicating that a portion of the transcripts retain the first intron. No amplification was observed in the no-reverse transcriptase controls (data not shown), confirming that the observed amplification reflects RNA transcripts and not genomic DNA contamination.

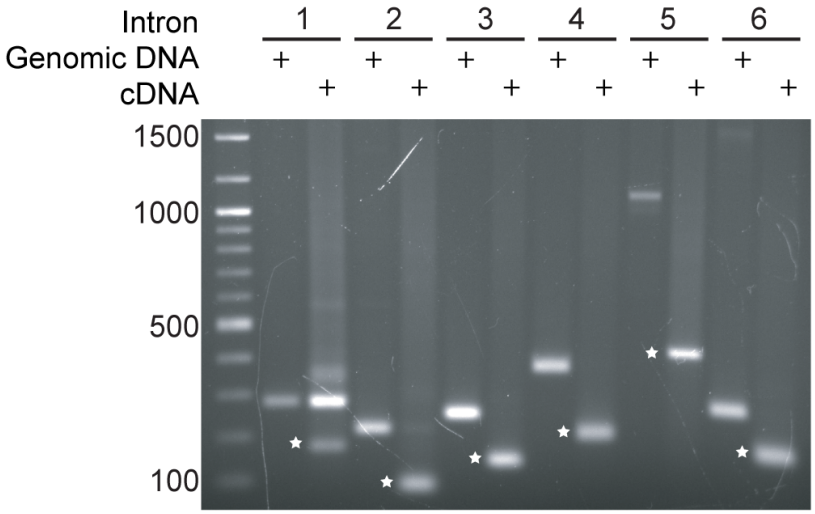
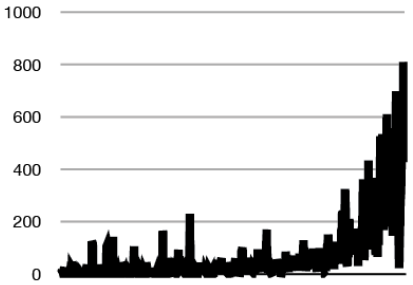
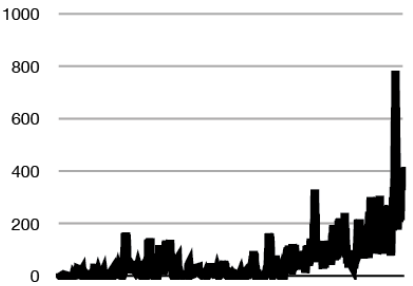


Figure S3.2. Transcriptional support of *M. brevicollis* elORFs. RNA-seq coverage of all predicted genomic reading frames greater than 10,000 bps (elORF) is shown. Each elORF was divided into 10 bp segments and the average number of reads mapping to each basepair within those segments was calculated and plotted versus the relative position of the segment within the elORF (normalizing the length of elORF for visualization purposes and taking the strand of the elORF into account, moving from the 5' end to the 3' end). Of the elORFs that were transcribed, there was a strong bias towards higher coverage at the 3' end, possibly due to strand breakage during the poly-A selection step of mRNA purification.

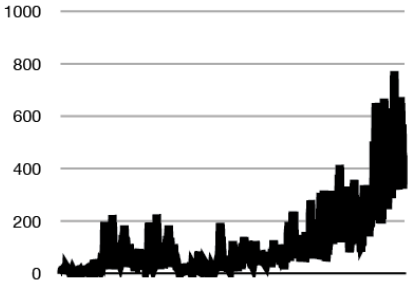
elORF #1
59597 bps



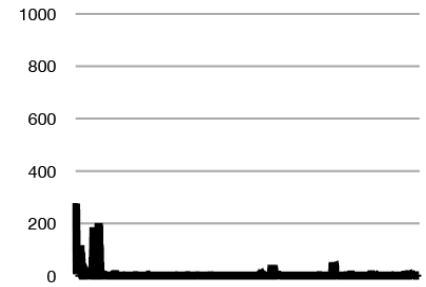
elORF #2
32153 bps



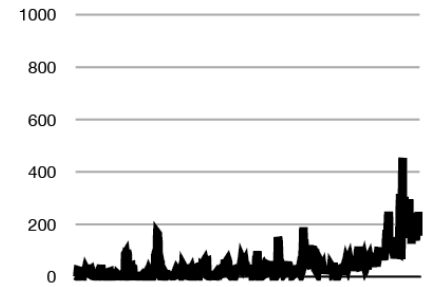
elORF #3
31943 bps



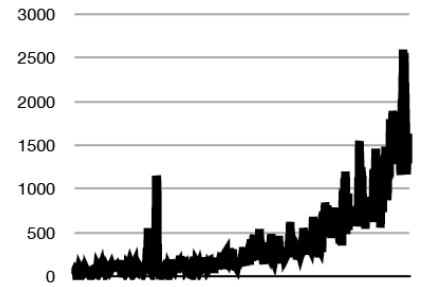
elORF #4
28883 bps



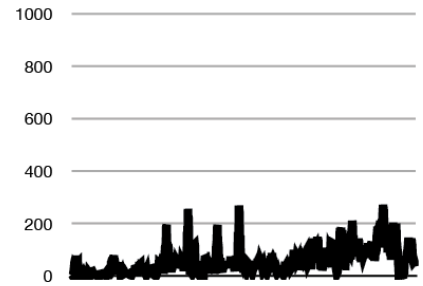
elORF #5
23888 bps



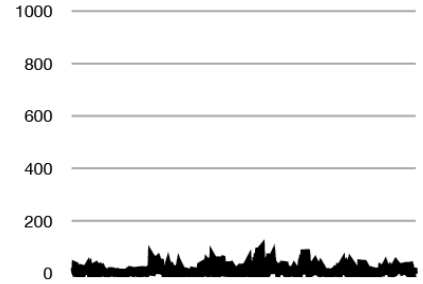
elORF #6
21440 bps



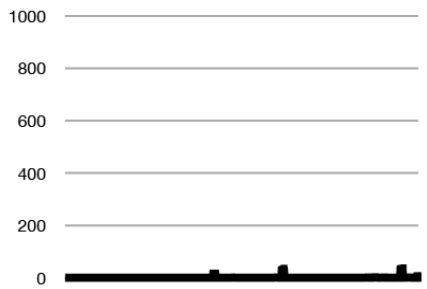
elORF #7
18767 bps



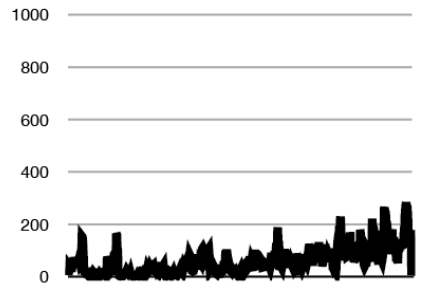
elORF #8
18551 bps



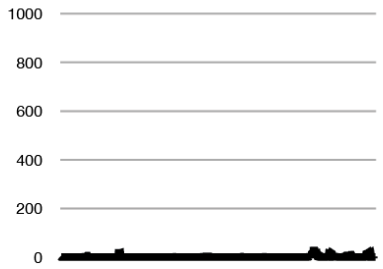
elORF #9
17801 bps



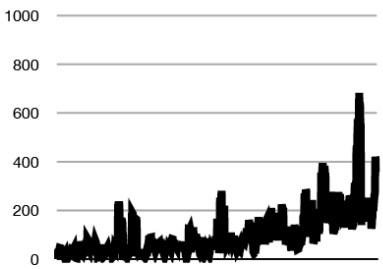
elORF #10
17366 bps



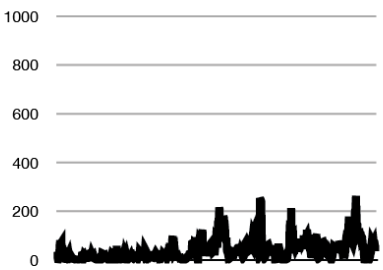
elORF #11
16625 bps



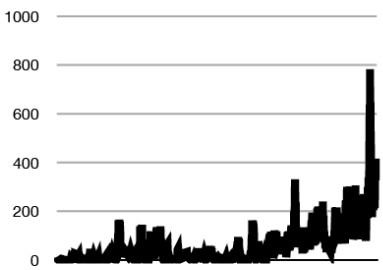
elORF #16
15242 bps



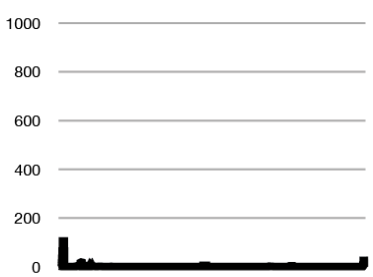
elORF #12
16274 bps



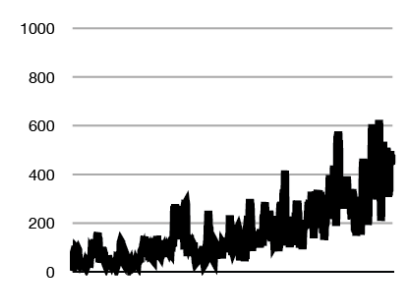
elORF #17
14582 bps



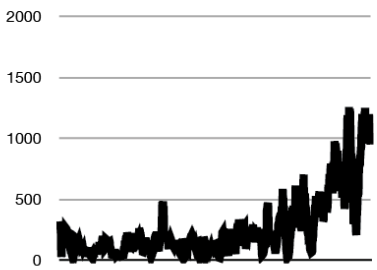
elORF #13
16202 bps



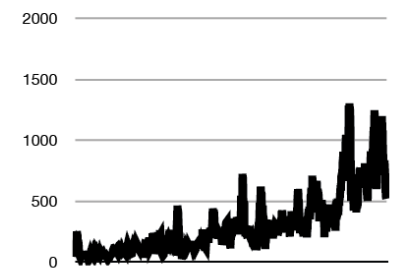
elORF #18
14315 bps



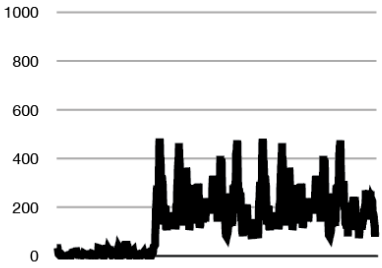
elORF #14
15926 bps



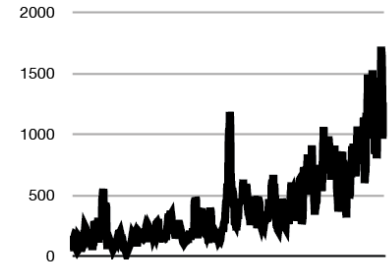
elORF #19
13706 bps



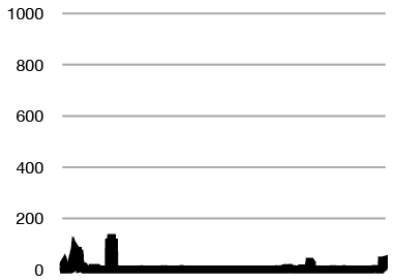
elORF #15
15380 bps



elORF #20
13682 bps

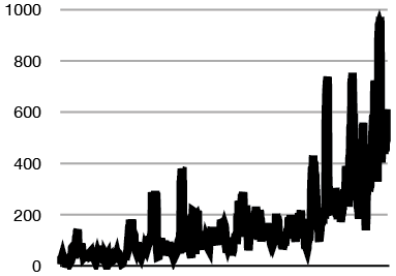


elORF #21
13358 bps

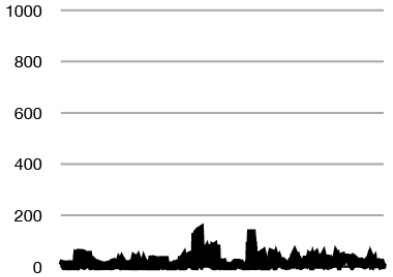


elORF #22
13019 bps

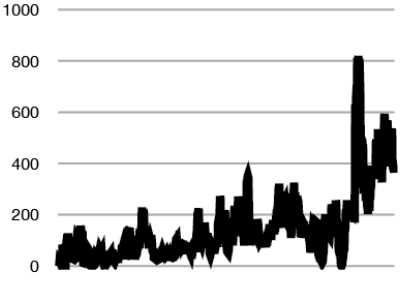
elORF #23
16202 bps



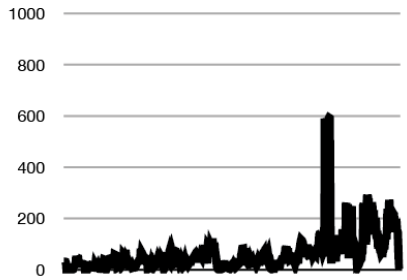
elORF #24
13001 bps



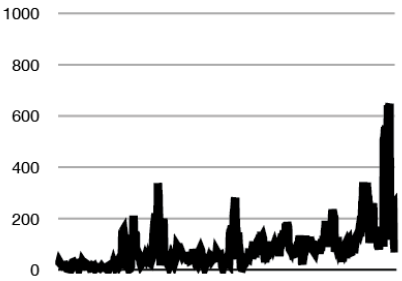
elORF #25
12740 bps



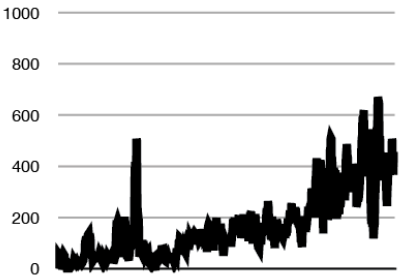
elORF #26
12599 bps



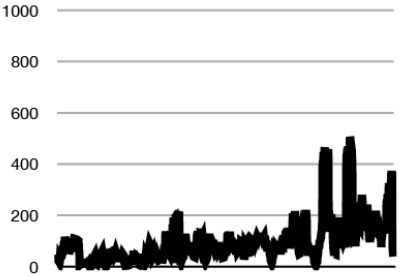
elORF #27
12587 bps



elORF #28
12266 bps

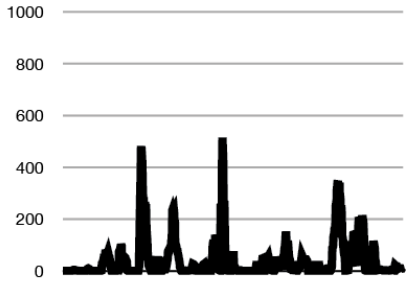


elORF #29
12260 bps



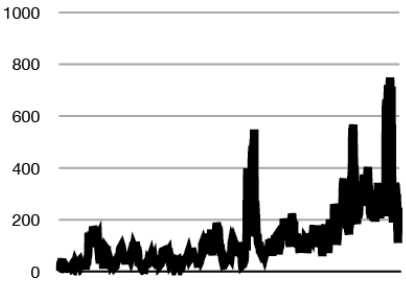
elORF #30
12125 bps

elORF #31
11969 bps

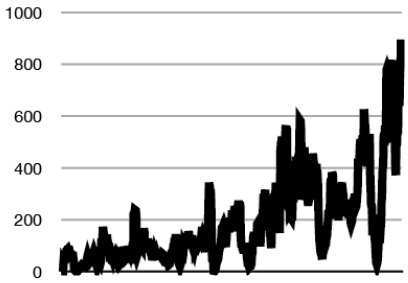


elORF #36
11255 bps

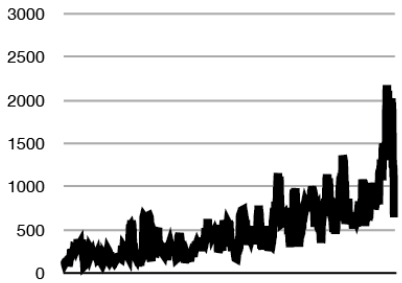
elORF #32
11924 bps



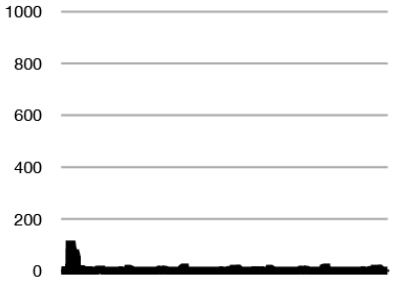
elORF #37
11120 bps



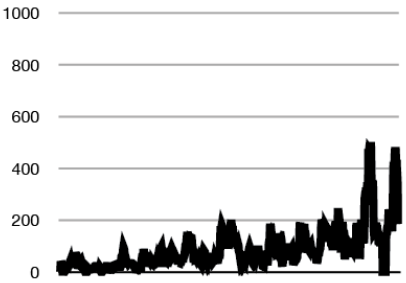
elORF #33
11849 bps



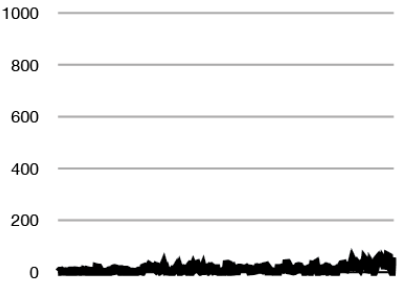
elORF #38
10991 bps



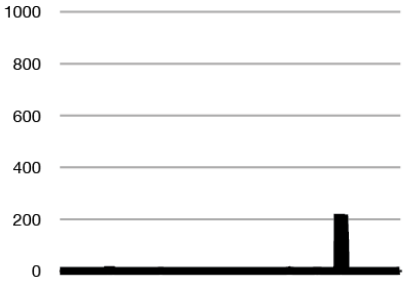
elORF #34
11681 bps



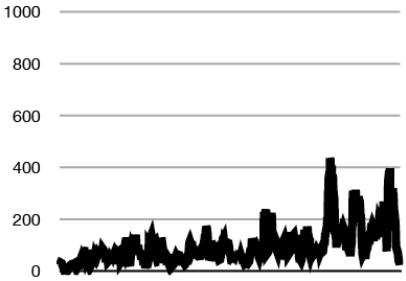
elORF #39
10700 bps



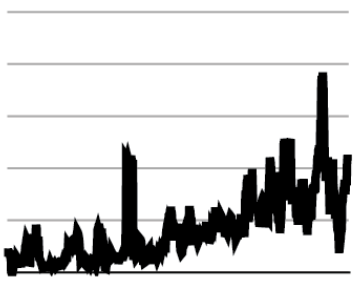
elORF #35
11468 bps



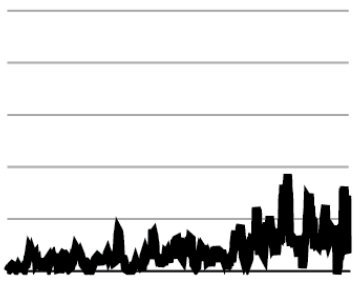
elORF #40
10658 bps



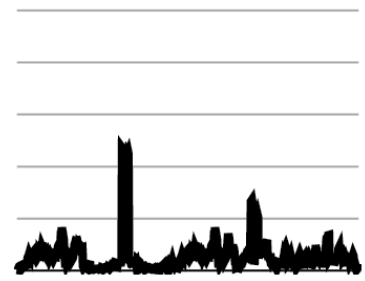
elORF #41
10475 bps



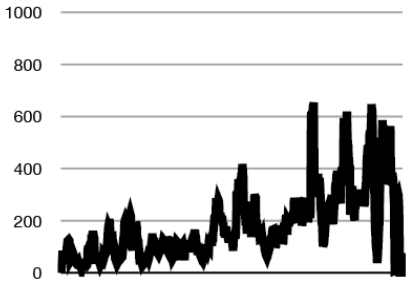
elORF #42
10451 bps



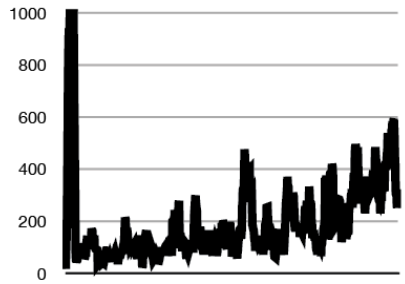
elORF #43
10385 bps



elORF #44
10301 bps



elORF #45
10082 bps



elORF #46
10025 bps

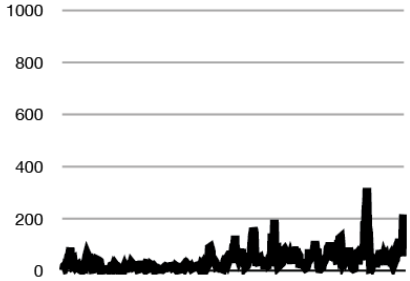
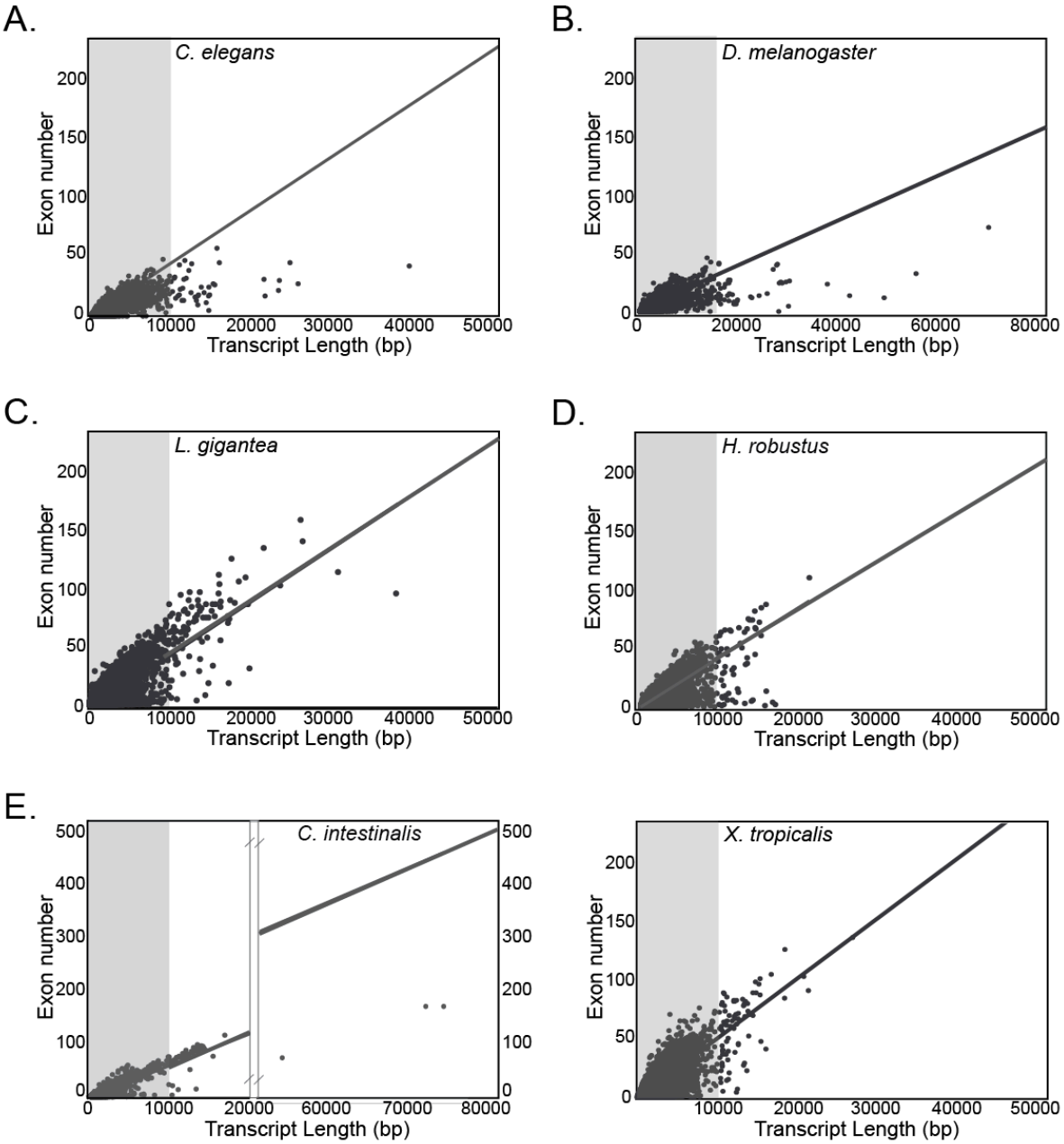


Figure S3.3. Intron density versus transcript length in bilaterians. Intron number versus transcript length was plotted for six additional bilaterian metazoan genomes (A – D). Solid lines indicate the number of expected introns based on the genome-wide intron density. In these genomes there were many instances of intron-rich genes with transcripts greater than 10,000 bps, and none of these genomes show as strong a trend towards intron depletion in long genes as is seen in choanoflagellates.



Chapter 4: Contrasting modes of alternative splicing in choanoflagellates and metazoans

SUMMARY

Alternative splicing, in which multi-exon genes are processed into a variety of isoforms, is a prominent type of gene regulation in metazoans, facilitating cell differentiation and intercellular signaling [199, 209]. However, little is known either about alternative splicing in unicellular eukaryotes or the role that it played in the origin of metazoans. To gain insight into the evolution of alternative splicing in metazoans, I studied this process in their closest unicellular relatives, choanoflagellates. I used transcriptome sequences from multiple life history stages in the choanoflagellate *S. rosetta* and environmental stress conditions in the choanoflagellate *M. brevicollis* to assay alternative splicing. Alternative splicing was detected in 2.4 and 8.3% of *M. brevicollis* and *S. rosetta* genes, respectively. In both species, alternative splicing was associated with the use of non-canonical splice sites. In *S. rosetta*, which has multiple life history stages, there were examples of cell type-specific splice isoforms. The most common form of alternative splicing in both choanoflagellates was intron retention, while exon skipping was the rarest. As a point of comparison I analyzed transcriptome data from an early branching metazoan, the cnidarian *H. magnipapillata*. In contrast to choanoflagellates, intron retention in *H. magnipapillata* was rare while exon skipping was common, as has been observed in other metazoans [120, 126]. These results imply that while alternative splicing may have played functional roles in their unicellular ancestors, a shift in the preferred type of alternative splicing occurred early in the evolution of developmentally complex metazoans.

INTRODUCTION

A key component of multicellularity is the differential regulation of gene expression by cell type and developmental stage [194, 210, 211]. Much attention has been given to the evolution of transcriptional regulatory networks in multicellular organisms [211, 212]. However, alternative splicing, which enables one gene to encode multiple proteins with potentially different functions by varying which parts of the coding sequence are included in the final transcript [82], is emerging as another form of regulatory novelty that may have been important in the evolution of multicellular organisms [107].

Alternatively spliced variants are categorized based on how the exon-intron structure of the transcript is altered. In alternate 5' or 3' splice site usage, a shift in the location of splicing alters how much of an intron is removed from the transcript. Entire exons can be omitted from the final transcript by exon skipping, which occurs when the spliceosome removes an exon along with its two flanking introns. Finally, failure of the spliceosome to remove an intron leads to intron retention in the final transcript, which will add coding sequence or introduce a premature stop codon.

Specific patterns of alternative splicing can be regulated by the interaction of trans-acting RNA binding proteins and sequence elements near the involved splice sites, affecting their recognition by the spliceosome [112, 213].

Recent transcriptome sequencing studies have revealed that alternative splicing is common in metazoans and plants. In metazoans, the percentage of genes for which alternatively spliced isoforms was detected is high: 25% in *C. elegans* [112]; 61% in *D. melanogaster* [214]; and 95% in humans [87, 215]. Similarly, in plants, alternative splicing has been observed in 29% and 48% of *A. thaliana* and *O. sativa* genes, respectively [113, 114]. Contrastingly, little support for alternative splicing has been found in extant unicellular eukaryotes. In the intron-rich unicellular eukaryotes *Chlamydomonas reinhardtii* (Viridiplantae) and *Cryptococcus neoformans* (Fungi), traditional Sanger sequencing of ESTs revealed alternative splicing in 3% and 4.2% of genes, respectively [108, 109]. While deep next generation sequencing has not been widely used to analyze alternative splicing in unicellular species, one such study in *Plasmodium falciparum* (Chromalveolata), which at 1.4 introns per gene has relatively little opportunity for alternative splicing, found alternative splicing in 4.5% of genes [111]. A similar study in the fungus *Aspergillus oryzae*, which has a mean of 1.9 introns per gene [216], found that 8.6% of genes had at least one alternatively spliced isoform [111]. Also indicative of a low frequency of alternative splicing in unicellular eukaryotes is the high level of sequence conservation flanking the canonical dinucleotide splice site; variability in these sequences has been associated alternative splicing [107, 117].

In multicellular organisms alternative splicing is regulated by development and cell signaling [82], but the functional relevance of alternative splicing in unicellular eukaryotes has been ascertained in only a few cases. In *S. pombe*, intron retention in a cyclin (*rem1*) controls the switch between the mitotic and meiotic cell cycles [217]. Alternative splicing is also regulated during different life-cycle stages in *P. falciparum* [218]. In other unicellular eukaryotes alternative splicing is involved in responding to changes in the extracellular environment, e.g. light-dependent regulation of photosynthesis in *C. reinhardtii* [219] and response to heat stress in the diatom *Chaetoceros compressum* [220]. The life cycles and ecological ranges of many unicellular eukaryotes are not completely characterized, and estimates of the frequency of alternative splicing may increase with the inclusion of additional life history stages and environmental conditions, as well as higher sequence coverage.

Although comparisons of the amounts of alternative splicing in different species are somewhat fraught due to differing levels of sequence coverage and uneven sampling of different life history stages or conditions, the relative prominence of the various types of alternative splicing shows clear lineage-specific differences [119, 126]. In unicellular eukaryotes as diverse as the fungus *C. neoformans* and the diatom *P. tricornutum*, intron retention is the most frequently detected form while exon skipping is relatively rare, typically accounting for less than 10% of alternative splicing events [120]. Intron retention is also the predominant type of alternative splicing observed in plants [98]. In contrast, in metazoans exon skipping occurs

more frequently than intron retention in all lineages surveyed to date, accounting for as much as 50% of all alternative splicing events in some species [94, 120]. Inclusion or omission of an exon from a transcript can alter the activity of the protein product by adding or removing a functional domain [82]. Exon skipping is functionally important in several uniquely metazoan processes. For example, a network of exon skipping regulated by trans-acting RNA binding proteins is critical in the development of the vertebrate nervous system [104], and exon skipping in specific transcripts is involved in learning and memory [221]. Exon skipping also plays critical roles in apoptosis [100] and sex-determination [102], processes that are central to the evolution of multicellularity.

However, almost all data on the frequency and function of alternative splicing in metazoans comes from bilaterians. A more complete picture of the apparent association between alternative splicing and metazoan evolution must include early branching metazoans and appropriate unicellular outgroups. The closest unicellular relatives of metazoans are the choanoflagellates [184], whose genome-enabled representatives are the species *M. brevicollis* and *S. rosetta* ([3] and S.R. Fairclough, manuscript in preparation). These species have intron densities comparable to most intron-rich metazoan taxa, with averages of 6.6 and 7.6 introns per gene in *M. brevicollis* and *S. rosetta*, respectively (Chapter 1). Many intron positions in orthologous genes from *M. brevicollis* and metazoans are conserved, indicating that these introns evolved prior to the divergence of the choanoflagellate and metazoan lineages [3]. Thus the complex patterns of alternative splicing associated with intron-rich genes in bilaterians could also have a pre-metazoan origin.

In this study, I analyzed transcriptome data from *M. brevicollis* and *S. rosetta* to elucidate the relationship between the origin of metazoans and the evolution of exon skipping as the predominant form of alternative splicing. To capture regulated alternative splicing events, I incorporated transcriptome data from environmental stress conditions in *M. brevicollis* and different life history stages in *S. rosetta*. I found that intron retention is the most common type of alternative splicing in choanoflagellates while exon skipping is the least common. As a point of comparison, I also analyzed transcriptome data from the basal animal *Hydra magnipapillata* and, in contrast to choanoflagellates, exon skipping was the most common type of alternative splicing detected. These observations support a shift in the preferred type of alternative splicing early in metazoan evolution, allowing abundant exon skipping to be a source of increased regulatory complexity in metazoans.

MATERIALS AND METHODS

M. brevicollis culture conditions and UV treatment
M. brevicollis was grown at 25°C in natural seawater infused with Ward's cereal grass (5 g/L, Scholar Chemistry #9448606) in polystyrene culture dishes (Falcon) [222]. The cells were co-cultured with the bacterium *E. aerogenes* as a food source.

UV treatments were performed when cells were in log-phase growth (between 2 and 4 million cells per mL). To prevent unwanted UV absorption, the cereal grass media was removed and the dishes were rinsed two times with natural seawater, after which only a thin layer of liquid remained on the choanoflagellates attached to the bottom of the dish. The dishes were exposed to a UV-B light source for the amount of time necessary to deliver 500 J/m2 of UV irradiation, as was determined by a dosimeter (approximately three minutes). To assess viability, cells were stained with propidium iodide (10 ug/ml) for 15 minutes and visualized with fluorescence microscopy. Cell number was measured using a Brightline hemacytometer (Hausser Scientific).

M. brevicollis transcriptome sequencing
Sequencing libraries were prepared from two replicates of UV-treated and control samples of *M. brevicollis*. Replicate treatments were performed on different days. Cells were harvested 12 hours post-treatment. Total RNA was collected using the RNeasy isolation kit with on-column DNase treatment (Qiagen) and mRNA was purified using Dynal oligo(dT) beads (Invitrogen). Sequencing libraries were then prepared as described in chapter three. Multiplexing sequencing primers were used to introduce a unique six base pair barcode into each of the sample libraries, after which the libraries were amplified with 18 cycles of PCR. The concentrations of the libraries were determined using a quantitative PCR-based assay. Equal amounts of each library were pooled and 101 bps paired end reads were sequenced using two lanes of an Illumina GAIIx sequencer at the Vincent J. Coates Genomic Sequencing Laboratory (QB3, University of California, Berkeley).

Detection of alternative splicing in M. brevicollis, S. rosetta, and H. magnipapillata
In addition to the data that I collected as described above, I also analyzed transcriptome sequence data from two additional species, the choanoflagellate *S. rosetta* and the cnidarian *H. magnipapillata*, that were collected as part of other studies (*S. rosetta:* S.R. Fairclough, manuscript in preparation; *H. magnipapillata*: Y. Wegner, manuscript in preparation). For all three species, reads were mapped to the genome using the spliced read alignment program Tophat [187]. This program predicts introns without reference to gene annotations, but does require pre-assumed minimum and maximum intron lengths. I specified the minimum intron length as five base pairs and the maximum intron length as 50,000 bps. To increase the probability of detecting extremely short exons, I aligned the reads with the microexon-search option enabled. For *M. brevicollis* and *S. rosetta*, there was sequence data from a variety of conditions. For each species, all the sequence was aligned to the genome simultaneously and then mapped reads were separated by sample based on their multiplexing barcode.

Intron positions can be predicted independently of pre-existing gene annotations based on the split alignment of reads to a reference genome. However, such approaches may incorrectly predict introns due to misalignment of short or repetitive stretches of sequence. To obtain a high confidence set of intron predictions, I filtered the split read alignments using an entropy-based method that

is part of a pre-existing package, JuncBASE, designed to analyze alternative splicing in RNA-seq data [106]. The settings I used required each intron to have at least four split reads spanning it at four different offsets, which eliminates intron predictions that may be due to alignment errors. This step eliminated 40.5, 37.1, and 72.3 percent of intron predictions from *M. brevicollis*, *S. rosetta* and *H. magnipapillata* respectively (the higher percentage in *H. magnipapillata* was largely because the transcriptome was sequenced to lower coverage and many splice sites had fewer than four supporting reads). To analyze the conservation of splice site sequences, I extracted the five base pairs of sequence in the reference genome that flanked each of the remaining intron predictions using the Bioperl seqIO module [223]. I then used the WebLogo program to generate sequence logos, which graphically display the levels of conservation within a consensus site sequence [224].

I used the high confidence set of intron predictions to identify alternative splicing without reference to any pre-existing gene annotations. I wrote an algorithm that identifies introns whose positions conflict in such a way that they could not occur in the same transcript, indicating that there are multiple splice isoforms at that locus. The way in which the splice sites conflict with one another is used to classify alternative splicing events into four types: alternative 5' splice site, alternative 3' splice site, alternative 5' and 3' splice sites, and exon skipping. This approach does not attempt to assemble entire transcripts from the sequenced reads, but rather identifies alternative splicing events on a splice site by splice site basis.

While these types of alternative splicing may be detected by conflicts in splice site positions, intron retention is more difficult to identify. This is because some low level of sequence coverage is expected in intronic regions due to cDNAs reverse transcribed from mRNA transcripts that were not fully processed. To circumvent this issue, I quantified the level and uniformity of sequence coverage within the predicted introns. For an intron to qualify as retained, I first required that it have a substantial level of sequence coverage relative to its flanking exonic sequence. Specifically, the number of reads aligned to the intron normalized by its length had to be had to be at least 20% of the number of reads that were split across the intron. Second, I required that the level of sequence coverage be uniform across the length of the intron. To quantify uniformity of coverage, each intron was divided into ten bins and the read count for each bin was compared to its neighbors. I required that the difference between adjacent bins be less than 80% of the total number of reads mapped to the two bins. This effectively eliminates introns with large spikes or sudden drop-offs in sequence coverage. In addition, if any bin has no reads mapped to it, the difference between it and the adjacent bin will be 100%, and this intron will thus not be counted as retained. The combination of these two statistics identified introns with substantial and uniform levels of sequence coverage across their entire length. I manually examined intron retention events identified by this method and determined that these parameters were conservative (Figure S4.2), thus the estimates presented in this study are likely to represent lower bounds on levels of intron retention.

To identify alternative splicing events that are differentially regulated by environmental conditions (*M. brevicollis*) or life history stages (*S. rosetta*), I analyzed the number of reads mapping to alternative splice sites in different samples. To normalize for variations in overall transcript expression or sequence coverage between samples, I determined the rate of alternative splice site usage relative to its associated constitutive sites, an approach that is modeled after methods developed for the analysis of splice junction microarrays [225]. The rate of alternative 5' or 3' splice site usage was calculated as the number of reads supporting the alternative splice site divided by the total number of reads supporting both the alternative and constitutive splice sites. Similarly, the rate of exon skipping was calculated as the number of reads supporting the splice site that skips the alternate exon divided by the total number of reads mapping to that site and the two splice sites that result in inclusion of the exon.

RESULTS

Past comparative studies of alternative splicing have relied on mining EST data [119, 120]. However, the depth of sequence coverage achieved by traditional EST studies, which has been relatively low for lesser-studied organisms, limits the amount of alternative splicing detected by this approach. High-throughput sequencing of transcriptomes (RNA-seq) can efficiently provide increased levels of transcriptome coverage, even for organisms in which little or no previous transcriptome data exists. To study alternative splicing in choanoflagellates, a key lineage for understanding the evolution of gene regulation in metazoans, I analyzed high-throughput transcriptome data from two species, *M. brevicollis* and *S. rosetta*. To capture potentially condition-specific alternative splicing, RNA was collected from *M. brevicollis* in standard growth conditions and after UV-induced DNA damage. Unlike *M. brevicollis*, *S. rosetta* has multiple cell types as part of its life history, and differentiation of the various cell types can be controlled in the laboratory [145, 175]. To capture cell type splice isoforms, I analyzed transcriptome data from cultures with attached, solitary, colonial or mixed *S. rosetta* cell types.

The total amounts of transcriptome data obtained were 10.3 and 20.9 Gigabases of sequence for *M. brevicollis* and *S. rosetta* respectively (Table 4.1). For both species, slightly over 50% of the paired-end sequence reads were uniquely aligned to the reference genome (Table 4.1), resulting in 328-fold coverage of the coding sequence in *M. brevicollis* and 427-fold coverage of the coding sequence in *S. rosetta*. Because choanoflagellates are co-cultured with a bacterial food source, a substantial fraction of the reads were prokaryotic, which at least partially accounts for the high percentage of un-aligned reads.

This level of sequence coverage was sufficient to detect many examples of alternative 5' and 3' splice site usage, as well as exon skipping, in both species of choanoflagellates (Figure 4.S3). In *M. brevicollis*, a total of 509 events involving these three types of alternative splicing were observed. These events were spread across 221 genes, or 2.4% of the total set of *M. brevicollis* genes. In *S. rosetta*, substantially

more alternative splice site usage was observed. A combined total of 2,522 alternate 5' splice site, alternate 3' splice site and exon-skipping events were detected. These events were spread across 974 genes, corresponding to 8.4% of all *S. rosetta* genes. Though the level of sequence coverage in *S. rosetta* was higher than in *M. brevicollis* (427 vs. 328-fold coverage), it was only greater by a factor of 1.3. From these results, alternative splicing appears to be more common in *S. rosetta* genes than in those of *M. brevicollis*. The increased level of alternative splicing detected in *S. rosetta* may reflect intrinsic biological differences between the species, or the treatment of *M. brevicollis* with UV may not stimulate alternative splicing at levels comparable to what is observed during *S. rosetta* cell differentiation.

Although the level of alternative splicing observed is contingent upon the depth of sequence coverage, the relative proportions of the different types of alternative splicing are more robust to differences in the amount of sequence data. It has been previously reported that exon skipping is the most common type of alternative splicing in bilaterians, but the least common type in plants and unicellular eukaryotes [98, 119, 120]. To test if the prominence of exon skipping observed in bilaterians is unique to metazoans, I determined the relative frequencies of exon-skipping, alternate 5' and 3' splice site usage, and intron retention in choanoflagellates. Like other unicellular eukaryotes and plants, exon skipping was rare in choanoflagellates, accounting for just 2.3% and 3.9% of alternative splicing events in *M. brevicollis* and *S. rosetta*, respectively (Figure 4.1A – B). Also similar to other non-metazoan lineages was the high frequency of intron retention observed in choanoflagellates, which accounted for 86.3% and 67.5% of alternative splicing events in *M. brevicollis* and *S. rosetta*, respectively (Figure 4.1A – B). The relative usage of different types of alternative splicing in choanoflagellates is thus more similar to what has been observed in other unicellular eukaryotes and plants than in their sister lineage, the metazoans.

The relative abundance of intron retention in comparison to other types of alternative splicing in choanoflagellates, in combination with the converse observation in bilaterians, suggests that a shift in the preferred type of alternative splicing took place after the divergence of these two lineages. To further resolve when in metazoan evolution this shift took place, I analyzed high throughput transcriptome data from a basal metazoan, the cnidarian *H. magnipapillata*. The depth of sequence coverage for *H. magnipapillata* was lower than that obtained for choanoflagellates; with 3.1 Gigabases of sequence aligned to the reference genome yielding 131.1-fold coverage of the coding sequence.

Despite the lower amount of sequence coverage, the levels of alternative 5' and 3' splice site usage, as well as exon skipping, in *H. magnipapillata* were greater than those observed in choanoflagellates (Figure S4.3). In contrast, the amount of intron-retention is approximately 100 times lower in *H. magnipapillata* than in choanoflagellates (Figure S4.3). This difference is also reflected in the relative proportion of intron retention, which accounts for only 2.1% of alternative splicing events in *H. magnipapillata* (Figure 4.1C). Exon-skipping, while comprising a much

higher percentage of alternative events than in choanoflagellates, was still not the most common type of alternative splicing, in contrast to what has been reported for many bilaterians [119, 120]. Alternate 5' and 3' splice site usage each accounted for a slightly higher percentage of alternative splicing events than exon skipping (Figure 4.1C). Although the relative frequency of exon skipping in *H. magnipapillata* may be lower than in some bilaterian lineages, the ratio of exon skipping to intron retention is still high (Figure 4.1D). The opposite is true for choanoflagellates, in which intron-retention outnumbers exon skipping by more that 10:1 (Figure 4.1D). The stark difference in these ratios suggests that the shift from intron retention to exon skipping as the predominant mode of alternative splicing occurred early in metazoan evolution.

Non-canonical dinucleotide usage in alternative splice sites
In both constitutively and alternatively spliced introns, the first and last two nucleotides of intronic sequence are highly conserved throughout eukaryotes [4]. For introns spliced by the major (U2) spliceosome, the canonincal dinucleotide sequences found at the 5' and 3' splice sites are GT and AG (GT…AG). However, a small minority of U2 introns contain non-canonical dinucleotide sequences, specifically AT…AC and GC…AG [226, 227]. Although the preexisting gene annotations of *M. brevicollis* and *S. rosetta* contained only canonical splice site sequences, I identified many examples of non-canonical splice sites based on transcriptome sequence data (Table 4.2). In *M. brevicollis*, I found 571 examples of non-canonical splice sites, or 1.15% of the total number of sites. Non-canonical splice sites were even more common in *S. rosetta*, in which there were 5,147 non-canonical sites, accounting for over 5% of all observed splice sites.

Although the intronic dinucleotide sequence is the most highly conserved part of the splice site, sequence conservation does extend further into the flanking exon and intron [227]. The level of extended conservation is variable between species; splice sites in unicellular eukaryotes tend to have higher levels of conservation than splice sites in multicellular species (Chapter 1, [107, 118]). As would be expected from this trend, I found that canonical splice sites in choanoflagellates had significant sequence conservation outside of the dinucleotide sequence, particularly at the fifth intronic position (Figure 4.2A – B). In contrast, the non-canonical splice sites had little conservation outside of the dinucleotide sequences, with the exception of the 5' GC sequence in *M. brevicollis*, which was highly conserved (Figure 4.2A - B). Notably, AT…AC is also the dinucleotide sequence used by the minor (U12) spliceosome. However, U12 introns also have a specific pattern of extended conservation at the 5' end of the intron [228], which was not seen in choanoflagellates. Additionally, choanoflagellates have lost the components of the U12 spliceosome [229], so AT…AC introns in choanoflagellates are most likely non-canonical U2 introns rather than U12 introns.

Because low levels of splice site sequence conservation, or "weakened" splice sites, have been linked with alternative splicing elsewhere [230], I investigated if non-canonical splice sites in choanoflagellates were associated with alternative splicing.

The rates of exon-skipping and alternative 5' and 3' splice site usage in non-canonical splice sites were similar to what was seen for canonical sites. However, I found that non-canonical splice sites were preferentially associated with another type of alternative splicing, tandem short-distance splice site usage or wobble splicing (Figure 4.2C – D). In wobble splicing, both the 5' and 3' splice sites are shifted a small number of nucleotides in the same direction [231-233]. This type of alternative splice site usage is strikingly common in *S. rosetta* non-canonical splice sites; 30% of GC…AG and 49% of AT…AC splice sites are involved in wobble splicing. The case was similar for *M. brevicollis* AT…AC sites, where wobble splicing occurred in 23% of sites. In contrast, wobble splicing was less frequent among the highly conserved *M. brevicollis* GC…AG sites, where it was seen in only 3% of sites.

Notably, the majority of these wobble-splicing events consisted of a pair of one canonical and one non-canonical splice site. In 90% of cases, the 5' and 3' sides of the splice sites were separated by the same number of nucleotides, which preserves the reading frame of the transcript. However, the distance between the two splice sites was short (mean = 3.72 bps) and in almost 50% of cases in *S. rosetta* the splice site is shifted by two base pairs, which would change only one amino acid of the protein sequence. Thus, although widespread in *M. brevicollis* and *S. rosetta*, the functional significance of non-canonical splice sites and their association with wobble splicing in choanoflagellates is unclear.

Functional relevance of alternative splicing in choanoflagellates

Although the impact of splicing variations such as wobble-splicing and intron-retention remains a matter of debate [86, 234], the affect of exon skipping on protein structure and function is often more straightforward to predict, such as when it results in the exclusion of an entire functional protein domain [235]. In this study, I detected 84 and 304 exon skipping events in *M. brevicollis* and *S. rosetta*, respectively (Figure 4.S3). As multiple exon skipping events often occurred in a single gene, a total of 44 *M. brevicollis* and 144 *S. rosetta* genes were associated with exon skipping. To gain insight into the functional implications of alternative splicing in choanoflagellates, I investigated exon skipping events on a gene-by-gene basis (Table S4.1). Although most choanoflagellate genes containing exon skipping events did not have clear orthologs in other species, several of those that did undergo functionally relevant exon skipping in metazoans, specifically *src*, *myc*, *rab14* and *annexin A7* [236-239]. In many cases where direct orthology could not be assigned, conserved protein domains were indicative of the gene's function (Table S4.2). Many of the protein domains found in choanoflagellate genes associated with exon skipping are known to be involved in cell signaling in metazoans, including RhoGEF, TNFR, PDZ, PTB, and protein kinase domains (Table S4.2). In *S. rosetta*, exon skipping was especially prominent in one group of serine/threonine kinases, the TKL family. Three different TKL protein kinases in *S. rosetta* (gene IDs: 2727, 8621 and 7945) contained multiple exon skipping events. One of these kinases (08165) displayed a particularly complex pattern of alternative splicing (Figure 4.3A).

As exon skipping is often regulated by cell-type in metazoans, I attempted to identify statistically significant differences in rates of alternative exon inclusion between the various *S. rosetta* cell-types. To do this, I analyzed the number of reads supporting each exon skipping event, normalized by the total number of reads aligned to that region to account for any changes in overall transcript level. However, this data did not meet the distributional requirements of methods commonly used to analyze differential expression in transcriptome sequence data with small sample sizes [240, 241]. Although a genome-wide statistical analysis of differential regulation of exon skipping was not possible, I did examine the rates of exon skipping in the TKL protein kinase described above. I found that the tenth exon was always skipped in one of the colonial cell-types (Figure 4.3B). In contrast, the level of this exon-skipping event was low in solitary attached cells. Alternative splicing may be more widely regulated by cell-type in *S. rosetta*, although the answer to this question awaits better statistical methods and additional transcriptome data.

DISCUSSION

A major shift in alternative splicing accompanied the origin of metazoans

Exon skipping is uniquely prominent in metazoans; in all other lineages alternative splicing is dominated by intron retention [119, 120]. In this study I report on the first analysis of alternative splicing in RNA-seq data from the metazoan sister group, the choanoflagellates, and the basal metazoan *H. magnipapillata*. I found that, similar to other non-metazoan lineages, intron retention is by far the most common type of alternative splicing in choanoflagellates. Contrastingly, in *H. magnipapillata* intron retention was the least common type of alternative splicing while exon skipping was more frequent, as has been seen in other metazoans. These results pinpoint the shift from intron retention to exon skipping to the early evolution of metazoans, specifically the period after their divergence from choanoflagellates but before the eumetazoan radiation.

One potential cause of the shift from intron retention to exon skipping is a difference in how the spliceosome accurately identifies pairs of splice sites in metazoans versus other lineages. In metazoans, where exons are located amid long stretches of intronic sequences containing many potential cryptic splice sites, identification of the correct intron-exon boundaries is aided by the interaction of spliceosomal components with other proteins that span across the exon, a process that is termed exon definition [122, 123, 242, 243]. In contrast, for unicellular eukaryotes with small intron sizes, such as *S. pombe*, accurate splice site recognition seems relies on interactions that bridge the intervening intronic sequence, or intron definition [243, 244]. The recognition of splice sites in pairs puts a limit on the amount of sequence that can intervene, and both exon definition and intron definition become prone to errors in splicing when the exon or intron size, respectively, becomes greater than approximately 500 bps [245].

Exon and intron definition are mechanistically linked to exon skipping and intron retention respectively – when a 5’ splice site that is usually identified via exon

definition is mutated, the upstream exon will be skipped [242]. However, if the 5' splice site is usually recognized through intron definition the downstream intron will be retained [244]. The intron sizes observed in this study suggest that *H. magnipapillata*, with a median intron size of 518 bps, would typically employ exon definition while choanoflagellates, with median intron sizes of 157 bps in *M. brevicollis* and 228 bps in *S. rosetta*, would favor intron definition. Widespread intron definition in choanoflagellates and exon definition in *H. magnipapillata*, predisposing choanoflagellates to intron retention and *H. magnipapillata* to exon skipping, may at least partially explain the dramatic differences observed in the relative frequencies these types of alternative splicing.

Splice site sequence conservation and alternative splicing in choanoflagellates
Another difference in gene structure between metazoans and unicellular eukaryotes is the level of conservation at the splice site sequences that mark the beginning and end of each intron. In metazoans, there is little sequence conservation outside of the first and last two intronic positions, while in unicellular eukaryotes conservation extends farther into the intron [107]. It has been suggested that in metazoans, the binding of SR (serine/arginine-rich) proteins to exonic sequences in pre-mRNA facilitates precise splicing, while in unicellular eukaryotes accurate splicing relies primarily on highly conserved intronic splice site sequences [246]. Indeed, the SR protein family underwent a significant expansion in metazoans, while being lost entirely in some unicellular lineages [247]. The demarcation of exons by SR proteins may have relaxed the selective constraints on splice site sequences in metazoans, leading to their decreased conservation. Support for this hypothesis comes from *S. cerevisiae*, an organism that contains no SR proteins, in which removal of an intron with mutationally weakened splice sites was restored by the ectopic expression of a mammalian SR protein [248]. The exon definition strategy in metazoans provides a route to alternative splicing not present in unicellular eukaryotes, as the trans-acting SR proteins can be regulated at the transcriptional or post-transcriptional level during cell differentiation or signaling.

Like other unicellular eukaryotes, I found that canonical splice sites in choanoflagellates have conservation that extends beyond the dinucleotide sequences. However, in the choanoflagellate *S. rosetta*, and to a lesser extent *M. brevicollis*, I also found a substantial number of splice sites with non-canonical dinucleotide sequences. In general, these sites had lower levels of extended conservation than the canonical sites, with the notable exception of the highly conserved GC…AG splice site in *M. brevicollis*. Apart from the *M. brevicollis* GC…AG site, the other less conserved non-canonical splice sites often occurred in tandem with canonical ones, producing a pattern that has been termed wobble splicing [231, 232]. Wobble splicing in these pairs of tandem sites allows for alternative splice site usage without compromising the strength of the constitutive site. However, the functional relevance this type of alternative splicing in metazoans is unclear; global surveys have not detected regulation by cell type and suggest that wobble splicing may be a reflection of noise in splice site recognition [234]. Nevertheless, there are

isolated examples of wobble splicing producing functionally significant isoforms [249, 250].

In *S. rosetta*, the distance between the tandem splice sites is typically short and identical between the 5' and 3' sites. This type of offset maintains the reading frame of the transcript and alters the protein sequence by only a few amino acids, and in the majority of cases usage of the non-canonical site is not likely to affect protein function. However, there may be rare instances in which it does, thus providing a path to alternative splicing that is more accessible to organisms such as choanoflagellates that maintain highly conserved splice sites. Notably, the non-canonical splice sites in *S. rosetta* and *M. brevicollis* were not identified in pre-existing gene annotations, potentially because of their close proximity to canonical sites, and may also have been missed in other organisms. Future transcriptome studies will reveal if this type of alternative splicing is common in additional unicellular lineages.

The impact of gene structure on the evolution of regulatory complexity
The differences in alternative splicing among metazoans and choanoflagellates highlight the connection between exon-intron gene structure and regulatory complexity. Although choanoflagellates and metazoans have similar numbers of introns (Chapter 1), exon skipping is much more common in metazoans, where it plays important roles in cell signaling and development [82]. As suggested above, this change may have initially been triggered by increased intron sizes, and then subsequently evolved into an important mechanism of gene. This underlying change in gene structure, greater intron lengths, was part of a more extensive increase in non-coding sequence that occurred during early metazoan evolution, and which may have been a non-adaptive by-product of the transition to multicellularity [50]. Indeed, while there are conserved and potentially functional regions within large metazoan introns [251], much of the increase in size can be accounted for by insertions of selfish or repetitive DNA elements [252]. The increased frequency of exon skipping in metazoan genomes may thus be the result of non-adaptive changes in genome architecture, an aspect of the transition to multicellularity that could have had widespread affects on the evolution of regulatory complexity.

TABLES AND FIGURES

Table 4.1. Summary of transcriptome sequencing data from *M. brevicollis* and *S. rosetta*

	M. brevicollis	*S. rosetta*
Number of conditions	2 (environmental)	8 (life cycle stages)
Read length (bps)	101 bps	68 bps
Total number of reads	101.5 million	322.4 million
Total sequence amount	10.3 Gigabases	20.9 Gigabases
Percentage of uniquely aligned paired-end reads	54.0%	53.5%
Number of introns predicted by alignments	53,669	78,138

Figure 4.1. Intron retention is the predominant form of alternative splicing in choanoflagellates. The relative frequencies of exon skipping, alternate 5' splice site usage, alternate 3' splice site usage, and intron retention in *M. brevicollis* (A), *S. rosetta* (B), and *H. magnipapillata* (C) are shown. Intron retention is the most common type of alternative splicing in choanoflagellates, while it is the least common in *H. magnipapillata*. (D) The ratio of exon skipping to intron retention seen in *M. brevicollis*, *S. rosetta*, and *H. magnipapillata* shows a shift in the preferred type of alternative splicing from intron retention in choanoflagellates to exon skipping in *H. magnipapillata*.

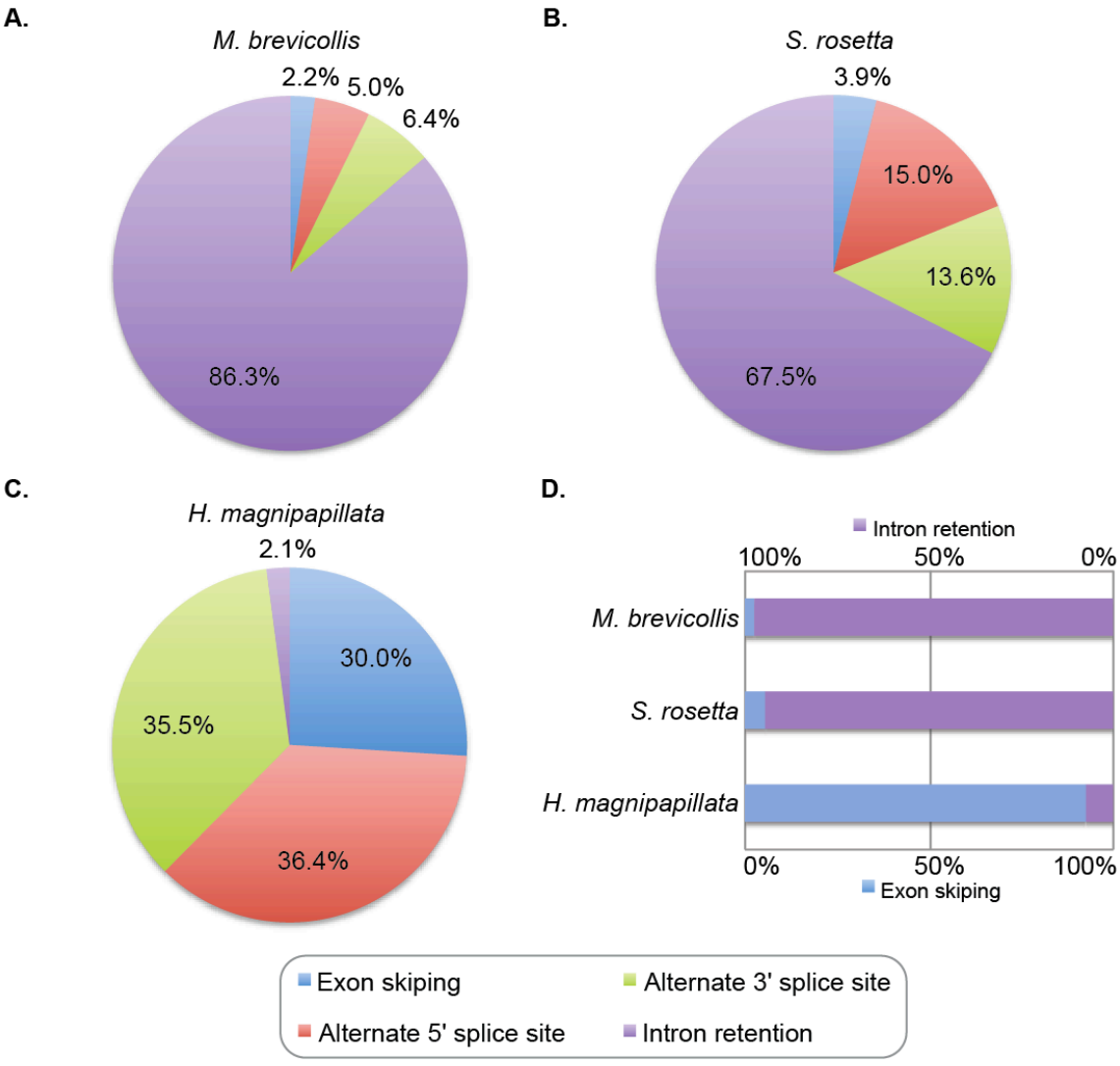


Table 4.2. Canonical and non-canonical dinucleotide splice site usage in choanoflagellates

	M. brevicollis	*S. rosetta*
GT . . . . . AG	98.85 %	93.41%
GC . . . . . AG	1.00 %	3.70 %
AT . . . . . AC	0.15 %	2.89 %

Figure 4.2. Non-canonical splice sites are associated with alternative splicing. (A – B) Splice site conservation in *M. brevicollis* (A) and *S. rosetta* (B). The splice site sequences from a high confidence set of introns predicted based on RNA-seq evidence were separated according to the dinucleotide sequence immediately following and proceeding the 5' and 3' splice sites, respectively. The overall height of the letters indicates the level of conservation at that position, while the height of the individual letters represents their relative proportions. The level of conservation of the surrounding sequence was similar for both the canonical (GT-AG) and non-canonical (GC-AG and AT-AC) dinucleotide sequences. The 5' splice site sequence in the AT-AC splice sites does not show the conservation typical of U12 introns, indicating that these introns are not processed by the minor (U12) spliceosome [228]. (C – D) Splice sites were grouped by dinucleotide sequence and the percent of splice sites with and without associated alternative sites was calculated for *M. brevicollis* (C) and *S. rosetta* (B). This analysis included exon-skipping, alternate 5' and 3' splice site usage, and wobble splicing, cases in both the 5' and 3 splice sites are shifted in the same direction. In both species, the percentage of non-canonical splice sites involved in at least one of these types of alternative splicing is higher than what is seen for canonical splice sites. In all panels, non-canonical splice sites are highlighted in bold.

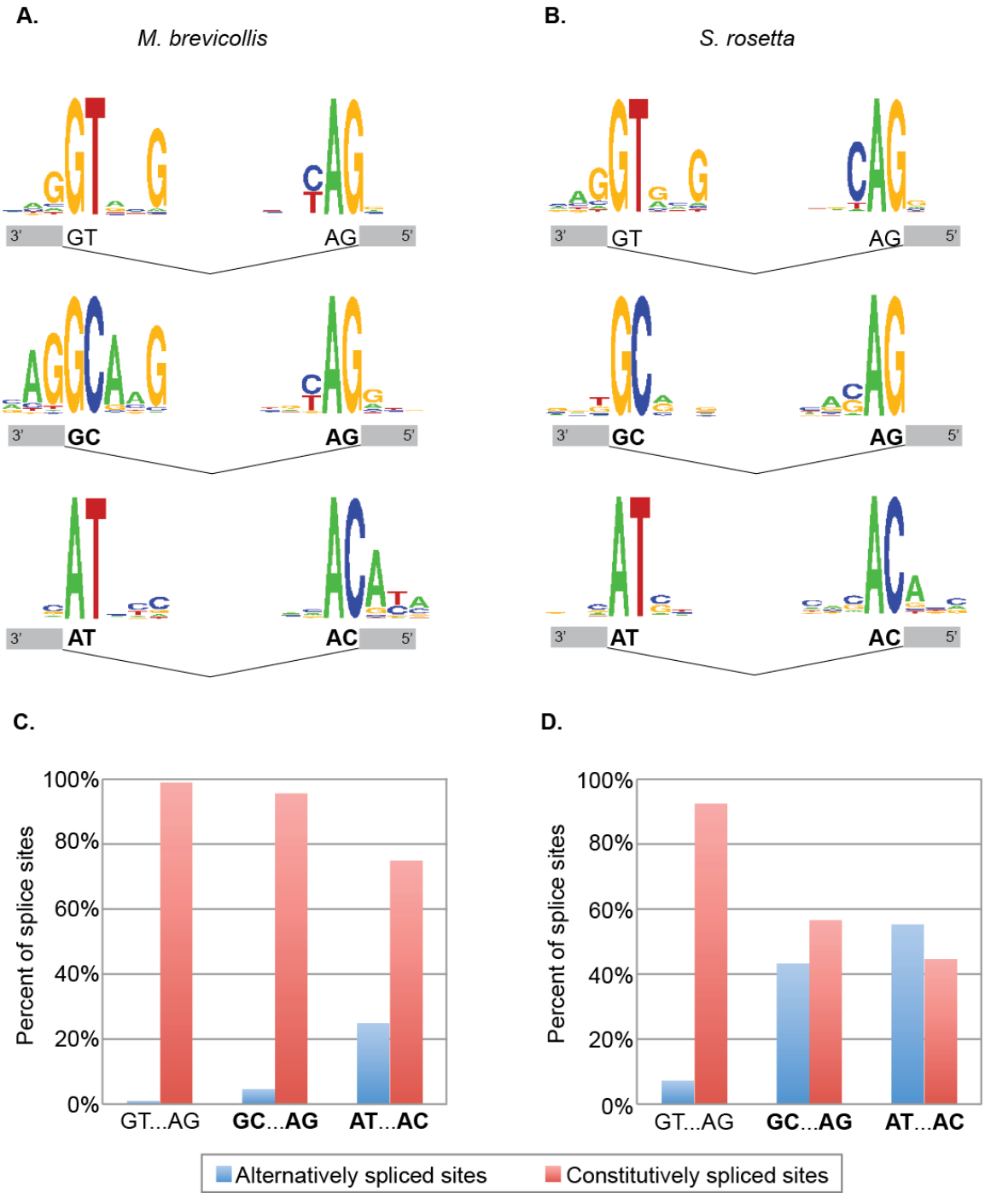

A.
M. brevicollis
B.
S. rosetta
3'
GT
AG
5'
GC
AG
AT
AC
C.
D.
Percent of splice sites
100%
80%
60%
40%
20%
0%
GT...AG
GC...AG
AT...AC
Alternatively spliced sites
Constitutively spliced sites

Figure 4.3. A TKL protein kinase in *S. rosetta* has cell-type specific splice isoforms. (A) Numerous exon skipping events are seen in an *S. rosetta* TKL family serine-threonine protein kinase (gene ID: 08165). In rosette colonies, inclusion of exon 10 was never detected. The exon skipping event concerned is highlighted in red. (B) The "rate of exon skipping" can be quantified by dividing the number of reads supporting the alternative splice site by the total number of reads supporting the alternative site and the two constitutive sites. The rate of exon skipping for exon 10 observed in various life-cycle stages of *S. rosetta* is shown. Data for attached cells isolated from three different culture conditions is shown. Exon 10 was skipped at a higher rate in colonial and swimming cell types than in attached cells. As exon 10 is always skipped in rosette colonies, the rate of exon skipping for this cell type is one.

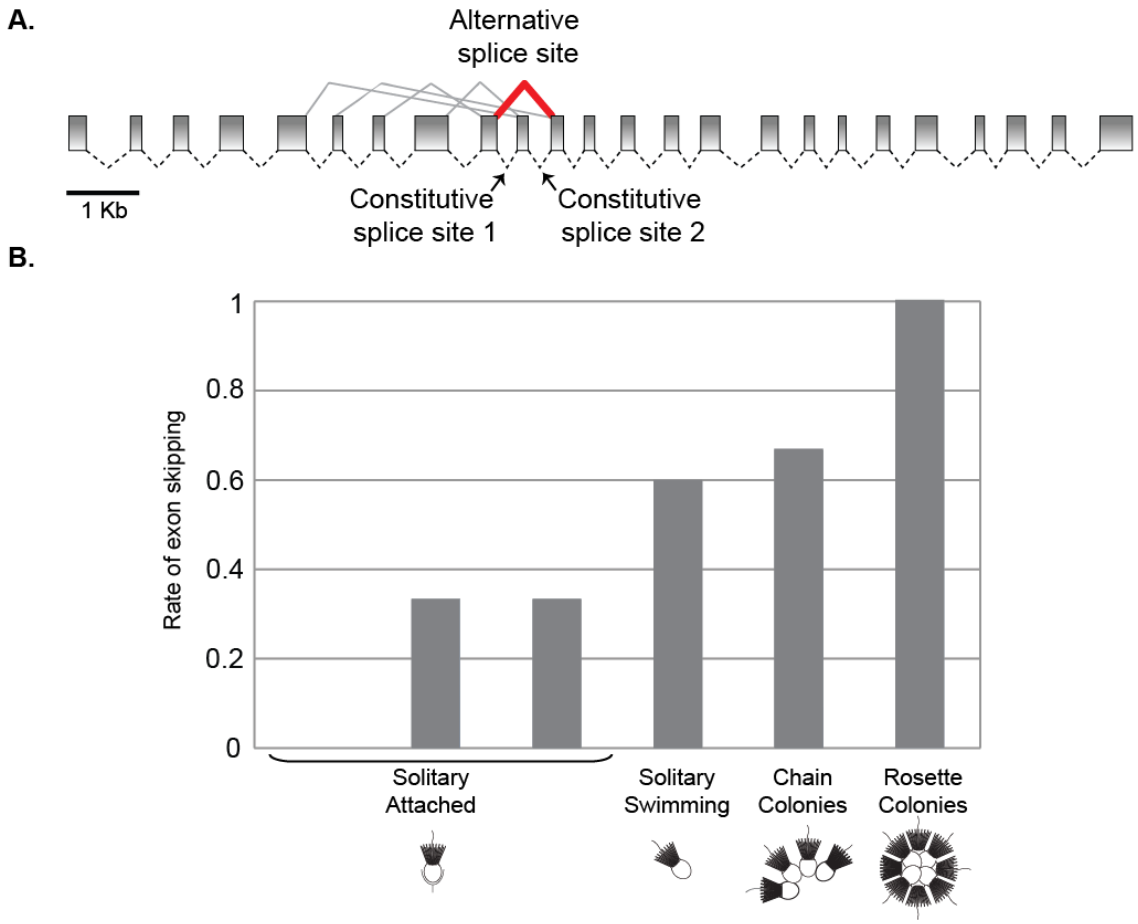
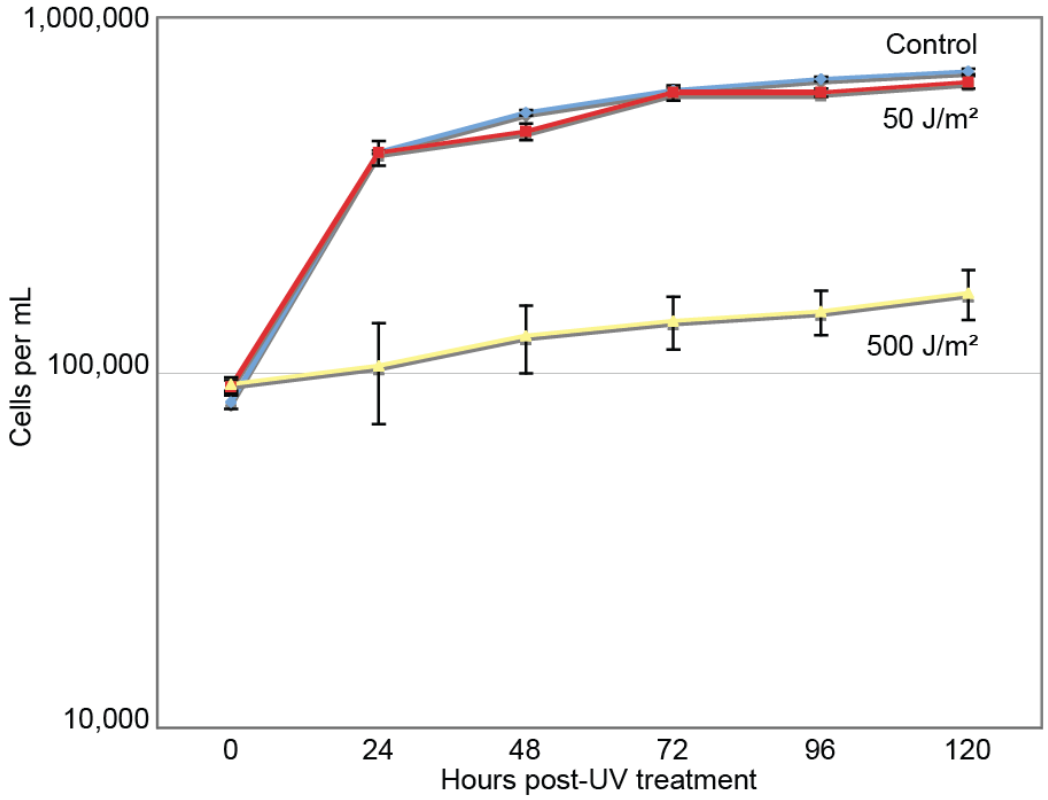


Figure S4.1. The affect of UV-irradiation on *M. brevicollis* viability and growth. (A) Cells were treated with 50 J/m^2 (red line) or 500 J/m^2 (yellow line) of UV-B irradiation, or left untreated (blue line). After the 500 J/m^2 treatment cell growth was inhibited for the entire period of monitoring (120 hours), while the 50 J/m^2 treatment had no affect on cell growth as compared to the control. (B) Cell viability 12 hours after UV treatment was measured using propidium iodide to label non-viable cells. While high doses of UV irradiation inhibited cell growth, their affect on cell viability was similar to the relatively mild affect seen in the lower dosage treatments, suggesting that cells persisted in a growth-arrested state after the 500 J/m^2 treatment.

A.


1,000,000
100,000
10,000
Cells per mL
Control
50 J/m²
500 J/m²
0
24
48
72
96
120
Hours post-UV treatment

B.

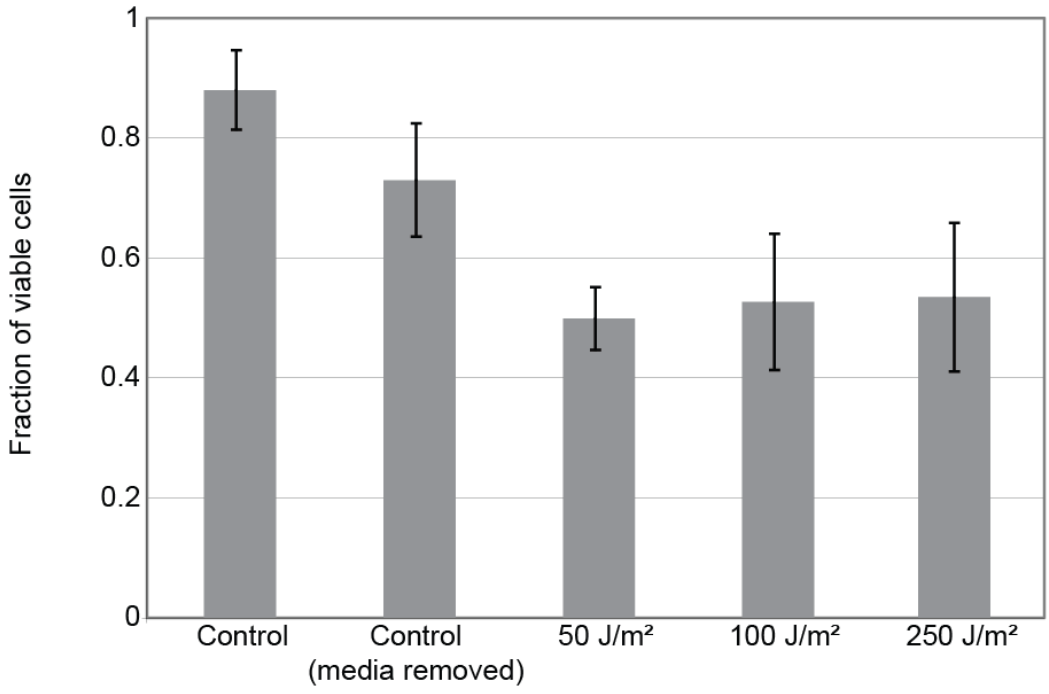

1
0.8
0.6
0.4
0.2
0
Fraction of viable cells
Control
Control
(media removed)
50 J/m²
100 J/m²
250 J/m²

Figure S4.2. Examples of sequence coverage in retained introns. Graphical representations of sequence coverage of retained introns and the flanking exonic sequence in *M. brevicollis* is shown. The intronic sequence is represented in black and the flanking exonic sequence in red. Intron retention was identified using two parameters; the level of coverage relative to the flanking exons and the uniformity of coverage across the length of the intron. Retained introns were required to have at least 20% as much sequence coverage (normalized to length) as the immediately flanking exonic sequence. Uniformity of coverage was assessed by comparing neighboring bins as described in the methods. The largest relative difference between neighboring bins could not exceed 80% in retained introns. These statistics, coverage (cvg.) level and uniformity, are displayed above each intron. Typical retained introns are shown in the left column (panels A, C, and E), while borderline retained introns (ones for which these statistics were near the cutoff values) are shown in the right column (panels B, D, and F).

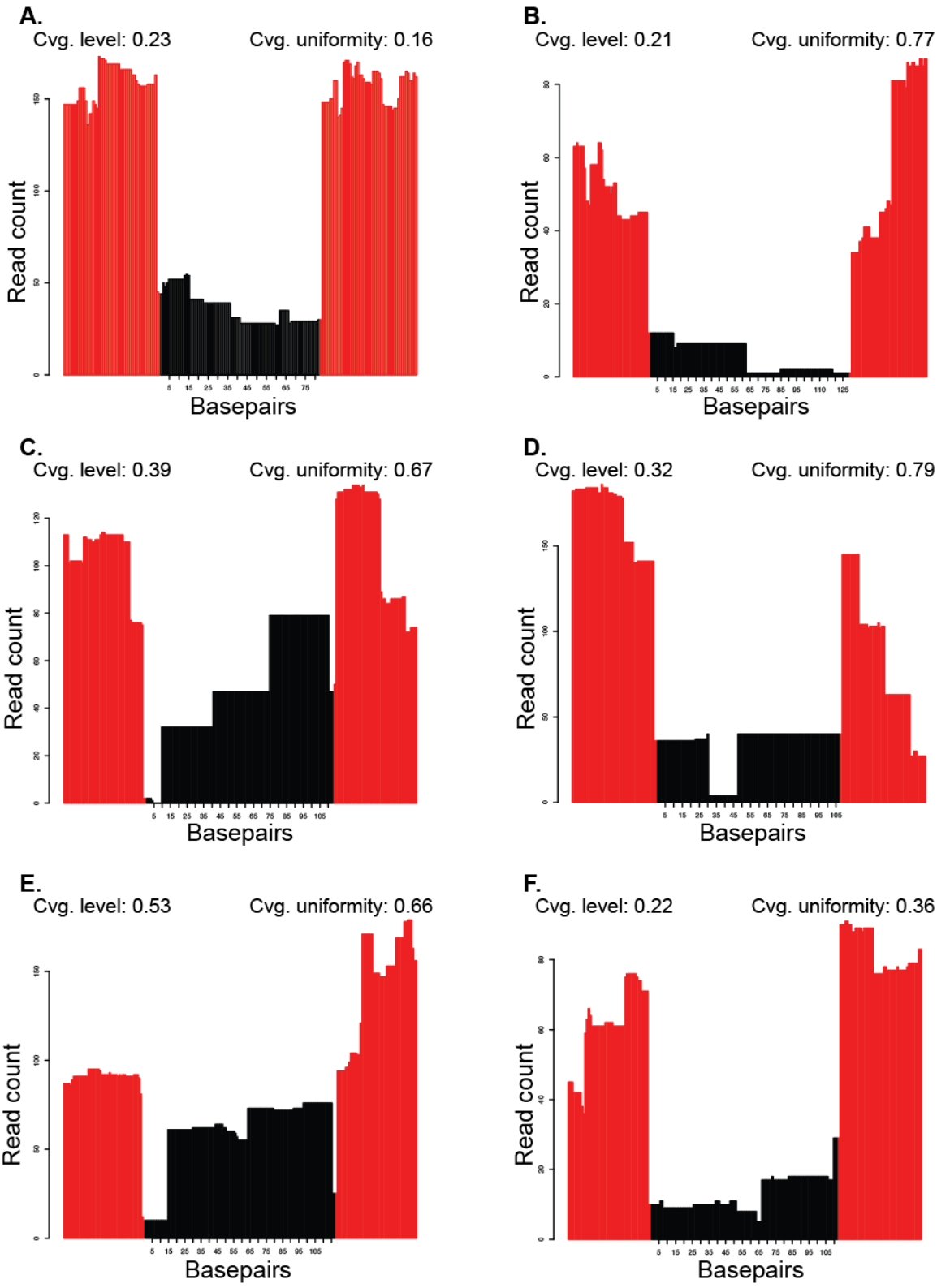

A.
Cvg. level: 0.23
Cvg. uniformity: 0.16
Read count
Basepairs
B.
Cvg. level: 0.21
Cvg. uniformity: 0.77
Read count
Basepairs
C.
Cvg. level: 0.39
Cvg. uniformity: 0.67
Read count
Basepairs
D.
Cvg. level: 0.32
Cvg. uniformity: 0.79
Read count
Basepairs
E.
Cvg. level: 0.53
Cvg. uniformity: 0.66
Read count
Basepairs
F.
Cvg. level: 0.22
Cvg. uniformity: 0.36
Read count
Basepairs

Figure S4.3. Levels of alternative splicing in choanoflagellates and the cnidarian *H. magnipapillata*. The number of alternative splicing events observed in the choanoflagellates *M. brevicollis* and *S. rosetta*, and the cnidarian *H. magnipapillata*, was classified according to the type of event (exon skipping, alternate 5' splice site usage, alternate 3' splice site usage or intron retention). The number of intron retention events observed in choanoflagellates was approximately 100 times greater than the number observed in *H. magnipapillata*.

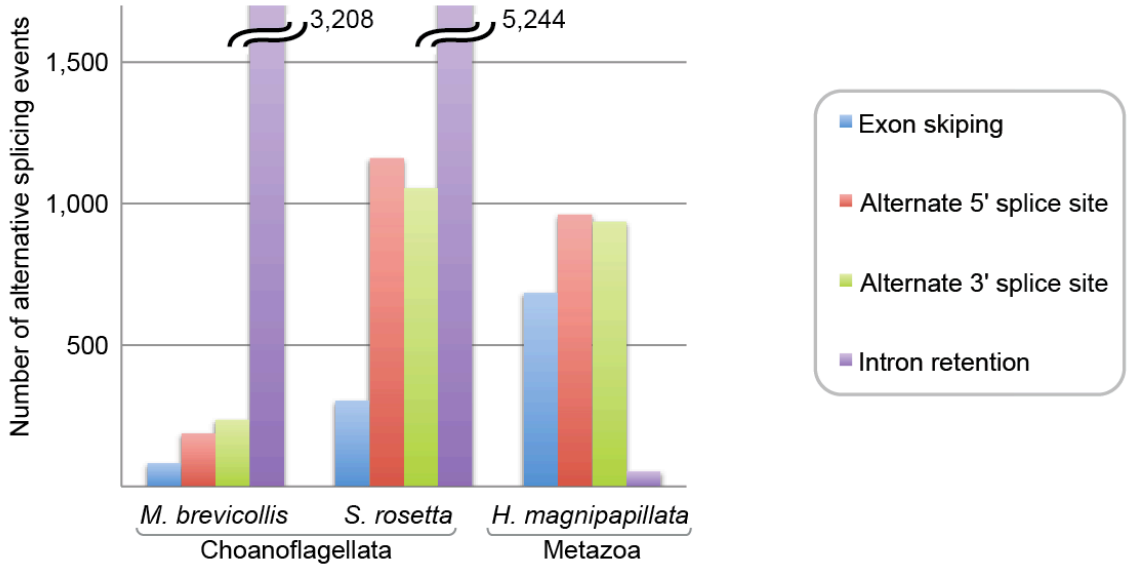


Table S4.1. Conserved genes with exon skipping splice isoforms

M. brevicollis	*S. rosetta*
Myc transcription factor	Acetyl coA Carboxylase
Src tyrosine kinase	Annexin A7
	Calponin
	Cathepsin Z
	DNA cross-link repair 1A protein
	Dynein heavy chain
	Inosine Triphosphate Pyrophosphatase
	NAD+ kinase
	NAK protein kinase
	PKC-interacting protein PICOT
	Presequence peptidase
	Proprotein convertase substillin/kexin 5
	Protein disulfide-isomerase
	Rab14

Table S4.2. Protein domains encoded by genes with exon skipping splice isoforms

M. brevicollis	*S. rosetta*
Ankyrin repeat	Ankyrin repeat
DNA helicase	Annexin
PDZ	AAA (ATPase associated with various cellular activities)
Protein kinase	Peptidase M17
Protein tyrosine kinase	Chaperone DnaJ
RRM (RNA recognition motif)	EF hand
Sushi	Leucine rich repeat
TNFR (Tumor Necrosis Factor Receptor)	MULE transposase
WD40 (WD domain, G-beta repeat)	Myb-like DNA binding
	PBI (PDZ Binding Interface)
	Protein kinase
	PTB (Phospho-Tyrosine Binding)
	RhoGEF (Guanine nucleotide Exchange Factor)
	RRM (RNA recognition motif)
	WWP
	Zinc finger C2H2 DNA binding

Bibliography

1. King, N., *The unicellular ancestry of animal development.* Developmental Cell, 2004. **7**(3): p. 313-25.
2. Knoll, A.H., *The multiple origins of complex multicellularity.* Annual Review of Earth and Planetary Sciences, 2011. **39**: p. 217-239.
3. King, N., et al., *The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans.* Nature, 2008. **451**(7180): p. 783-8.
4. Valadkhan, S., *The spliceosome: a ribozyme at heart?* Biological Chemistry, 2007. **388**(7): p. 693-7.
5. Rodríguez-Trelles, F., R. Tarrío, and F.J. Ayala, *Origins and evolution of spliceosomal introns.* Annual Reviews in Genetics, 2006. **40**: p. 47-76.
6. Tarrío, R., F.J. Ayala, and F. Rodríguez-Trelles, *Alternative splicing: a missing piece in the puzzle of intron gain.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(20): p. 7223-8.
7. Mattick, J.S. and M.J. Gagen, *The evolution of multi-tasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms.* Molecular Biology and Evolution, 2001. **18**(9): p. 1611-1630.
8. Lynch, M. and A.O. Richardson, *The evolution of spliceosomal introns.* Current Opinion in Genetics and Development, 2002. **12**(6): p. 701-10.
9. Lynch, M., *The origins of genome architecture.* 2007, Sunderland, MA: Sinauer Associates.
10. Nixon, J.E., et al., *A spliceosomal intron in Giardia lamblia.* Proceedings of the National Academies of Science, 2002. **99**(6): p. 3701-5.
11. Berget, S.M., C. Moore, and P.A. Sharp, *Spliced segments at the 5' terminus of adenovirus 2 late mRNA.* Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(8): p. 3171-5.
12. Chow, L.T., et al., *An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA.* Cell, 1977. **12**(1): p. 1-8.
13. Sambrook, J., *Adenovirus amazes at Cold Spring Harbor.* Nature, 1977. **268**(5616): p. 101-104.
14. Breathnach, R., J.L. Mandel, and P. Chambon, *Ovalbumin gene is split in chicken DNA.* Nature, 1977. **270**(5635): p. 314-9.
15. Koonin, E.V., *Intron-dominated genomes of early ancestors of eukaryotes.* Journal of Heredity, 2009. **100**(5): p. 618-23.
16. Fedorova, L. and A. Fedorov, *Introns in gene evolution.* Genetica, 2003. **118**: p. 123-31.
17. Gilbert, W., *Genes-in-pieces revisited.* Science, 1985. **228**(4701): p. 823-4.
18. Gilbert, W. and M. Glynias, *On the ancient nature of introns.* Gene, 1993. **135**(1-2): p. 137-44.
19. Patthy, L., *Exons--original building blocks of proteins?* BioEssays : news and reviews in molecular, cellular and developmental biology, 1991. **13**(4): p. 187-92.
20. Doolittle, F.W., *Genes in pieces: were they ever together?* Nature, 1978. **272**: p. 581-582.

21. Palmer, J.D. and J.M. Logsdon, *The recent origins of introns.* Current Opinion in Genetics and Development, 1991. **1**(4): p. 470-7.
22. Logsdon, J.M. and J.D. Palmer, *Origin of introns--early or late?* Nature, 1994. **369**(6481): p. 526.
23. Logsdon, J.M., *The recent origins of spliceosomal introns revisited.* Current opinion in Genetics and Development, 1998. **8**(6): p. 637-48.
24. Gilbert, W., *Why genes in pieces?* Nature, 1978. **271**(5645): p. 501.
25. Long, M., C. Rosenberg, and W. Gilbert, *Intron phase correlation and the evolution of the intron/exon structure of genes.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**: p. 12495-12499.
26. Liu, M. and A. Grigoriev, *Protein domains correlate strongly with exons in multiple eukaryotic genomes - evidence of exon shuffling?* Trends in Genetics, 2004. **20**: p. 339-403.
27. Kaessmann, H., et al., *Signatures of domain shuffling in the human genome.* Genome Research, 2002. **12**: p. 1642-1650.
28. Vogel, C., et al., *Structure, function and evolution of multidomain proteins.* Current Opinion in Structural Biology, 2004. **14**: p. 208-216.
29. Tordai, H., et al., *Modules, multidomain proteins and organismic complexity.* FEBS Journal, 2005. **272**(19): p. 5064-78.
30. Basu, M.K., et al., *Evolution of protein domain promiscuity in eukaryotes.* Genome Research, 2008. **18**(3): p. 449-461.
31. Nguyen, H.D., M. Yoshihama, and N. Kenmochi *Phase distribution of spliceosomal introns: implications for intron origin*. BMC Evolutionary Biology, 2006. **6**, DOI: 10.1186/1471-2148-6-69.
32. Qiu, W.G., N. Schisler, and A. Stoltzfus, *The evolutionary gain of spliceosomal introns: sequence and phase preferences.* Molecular Biology and Evolution, 2004. **21**: p. 1252-1263.
33. Cho, G. and W.F. Doolittle, *Intron distribution in ancient paralogs supports random insertion and not random loss.* Journal of Molecular Evolution, 1997. **44**: p. 573-584.
34. Rzhetsky, A., et al., *Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(13): p. 6820-5.
35. Cech, T.R., *The generality of self-splicing RNA: relationship to nuclear mRNA splicing.* Cell, 1986. **44**(2): p. 207-10.
36. Lambowitz, A.M. and S. Zimmerly, *Mobile group II introns.* Annual Review of Genetics, 2004. **38**: p. 1-35.
37. Toor, N., G. Hausner, and S. Zimmerly, *Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases.* RNA, 2001. **7**(8): p. 1142-52.
38. Bonen, L. and J. Vogel, *The ins and outs of group II introns.* Trends in Genetics, 2001. **17**(6): p. 322-31.
39. Zimmerly, S., G. Hausner, and X. Wu, *Phylogenetic relationships among group II intron ORFs.* Nucleic Acids Research, 2001. **29**(5): p. 1238-50.

40. Cousineau, B., et al., *Retrotransposition of a bacterial group II intron.* Nature, 2000. **404**(6781): p. 1018-21.
41. Sontheimer, E.J., P.M. Gordon, and J.A. Piccirilli, *Metal ion catalysis during group II intron self-splicing: parallels with the spliceosome.* Genes and Development, 1999. **13**(13): p. 1729-41.
42. Cavalier-Smith, T., *Intron phylogeny: a new hypothesis.* Trends in Genetics, 1991. **7**(5): p. 145-8.
43. Martin, W. and E.V. Koonin, *Introns and the origin of nucleus-cytosol compartmentalization.* Nature, 2006. **440**(7080): p. 41-5.
44. Singh, J. and R.A. Padgett, *Rates of in situ transcription and splicing in large human genes.* Nature Structural and Molecular Biology, 2009. **16**(11): p. 1128-33.
45. Vassilenko, K.S., et al., *Unidirestional constant rate motion of the ribosomal scanning particle during eukaryotic translation initiation.* Nucleic Acids Research, 2011. **39**(13): p. 5555-5567.
46. Rogozin, I.B., et al., *Analysis of evolution of exon-intron structure of eukaryotic genes.* Briefings in Bioninformatics, 2005. **6**(2): p. 118-134.
47. Jeffares, D.C., T. Mourier, and D. Penny, *The biology of intron gain and loss.* Trends in Genetics, 2006. **22**(1): p. 16-22.
48. Mattick, J.S., *Challenging the dogma: hidden layers of non-protein coding RNAs in the development of complex organisms.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2003. **25**: p. 930-939.
49. Lynch, M., *Intron evolution as a population-genetic process.* Proceedings of the National Academies of Science, 2002. **99**(9): p. 6118-23.
50. Lynch, M., *The frailty of adaptive hypotheses for the origins of organismal complexity.* Proceedings of the National Academies of Science, 2007. **104**(Suppl 1): p. 8597-604.
51. Goffeau, A., et al., *Life with 6000 Genes.* Science, 1996. **274**(5287): p. 546-567.
52. Castillo-Davis, C.I., et al., *Selection for short introns in highly expressed genes.* Nature Genetics, 2002. **31**(4): p. 415-418.
53. Lynch, M. and J.S. Conery, *The origins of genome complexity.* Science, 2003. **302**(5649): p. 1401-4.
54. Fedorov, A., A.F. Merican, and W. Gilbert, *Large-scale comparison of intron positions among animal, plant, and fungal genes.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**: p. 16128–1613.
55. Rogozin, I.B., et al., *Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.* Current Biology, 2003. **13**(17): p. 1512-7.
56. Roy, S.W., *Intron-rich ancestors.* Trends in Genetics, 2006. **22**(9): p. 468-71.
57. Csuros, M., *Malin: maximum likelihood analysis of intron evolution in eukaryotes.* Bioinformatics, 2008. **24**(13): p. 1538-9.
58. Carmel, L., et al., *Three distinct modes of intron dynamics in the evolution of eukaryotes.* Genome Research, 2007. **17**(7): p. 1034-44.
59. Roy, S.W. and W. Gilbert, *The evolution of spliceosomal introns: patterns, puzzles and progress.* Nature Reviews in Genetics, 2006. **7**(3): p. 211-221.

60. Roy, S.W. and W. Gilbert, *Rates of intron loss and gain: implications for early eukaryotic evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(16): p. 5773-8.
61. Sverdlov, A.V., et al., *A glimpse of a putative pre-intron phase of eukaryotic evolution.* Trends in Genetics, 2007. **23**(3): p. 105-8.
62. Roy, S.W. and D. Penny, *Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution.* Genome Research, 2006. **16**(10): p. 1270-5.
63. Stajich, J.E., F.S. Dietrich, and S.W. Roy, *Comparative genomic analysis of fungal genomes reveals intron-rich ancestors.* Genome Biology, 2007. **8**(10): p. R223.
64. Csurös, M., I.B. Rogozin, and E.V. Koonin, *Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach.* Molecular Biology and Evolution, 2008. **25**(5): p. 903-11.
65. Putnam, N.H., et al., *Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization.* Science, 2007. **317**(5834): p. 86-94.
66. Dibb, N.J., *Proto-splice site model of intron origin.* Journal of Theoretical Biology, 1991. **151**(3): p. 405-16.
67. Sverdlov, A.V., et al., *Conservation versus parallel gains in intron evolution.* Nucleic Acids Research, 2005. **33**(6): p. 1741-8.
68. Li, W., et al., *Extensive, recent intron gains in Daphnia populations.* Science, 2009. **326**(5957): p. 1260-2.
69. Jones, E.A., et al., *Regulation of the PMP22 gene through an intronic enhancer.* The Journal of Neuroscience, 2011. **31**(11): p. 4242-50.
70. Smith, A.N., et al., *A regulatory element in intron 1 of the cystic fibrosis transmembrane conductance regulator gene.* The Journal of Biological Chemistry, 1996. **271**(17): p. 9947-54.
71. Muller, F., et al., *Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord.* Development, 1999. **126**(10): p. 2103-16.
72. Abbasi, A.A., et al. *Human intronic enhancers control distinct sub-domains of Gli3 expression during mouse CNS and limb development*. BMC Developmental Biology, 2010. **10**.
73. Bejerano, G., et al., *Ultraconserved elements in the human genome.* Science, 2004. **304**(5675): p. 1321-5.
74. Baira, E., et al., *Ultraconserved elements: genomics, function and disease.* RNA biology, 2008. **5**(3): p. 132-4.
75. Krol, J., I. Loedige, and W. Filipowicz, *The widespread regulation of microRNA biogenesis, function and decay.* Nature Reviews in Genetics, 2010. **11**(9): p. 597-610.
76. Rodriguez, A., et al., *Identification of mammalian microRNA host genes and transcription units.* Genome Research, 2004. **14**(10A): p. 1902-10.
77. Okamura, K., et al., *The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila.* Cell, 2007. **130**(1): p. 89-100.
78. Ruby, J.G., C.H. Jan, and D.P. Bartel, *Intronic microRNA precursors that bypass Drosha processing.* Nature, 2007. **448**(7149): p. 83-6.

79. Berezikov, E., et al., *Mammalian mirtron genes.* Molecular Cell, 2007. **28**(2): p. 328-36.
80. Irimia, M., et al. *Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing*. BMC Evolutionary Biology, 2007. **7**.
81. Modrek, B. and C. Lee, *A genomic view of alternative Splicing.* Nature Genetics, 2002. **30**: p. 13-19.
82. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
83. Graveley, B.R., *Sorting out the complexity of SR protein functions.* RNA, 2000. **6**(9): p. 1197-211.
84. Barash, Y., et al., *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-9.
85. Wang, Z. and C.B. Burge, *Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.* RNA, 2008. **14**(5): p. 802-13.
86. Barbazuk, W.B., Y. Fu, and K.M. McGinnis, *Genome-wide analyses of alternative splicing in plants: opportunities and challenges.* Genome Research, 2008. **18**(9): p. 1381-92.
87. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
88. Yeo, G., et al. *Variation in alternative splicing across human tissues*. Genome Biology, 2004. **5**.
89. Pickrell, J.K., et al. *Noisy splicing drives mRNA isoform diversity in human cells*. PLoS Genetics, 2010. **6**.
90. Severing, E.I., et al. *Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome*. BMC Genomics, 2009. **10**.
91. Lareau, L.F., et al., *The evolving roles of alternative splicing.* Current Opinion in Structural Biology, 2004. **14**(3): p. 273-82.
92. Irimia, M., et al., *Quantitative regulation of alternative splicing in evolution and development.* Bioessays, 2009. **31**(1): p. 40-50.
93. Thanaraj, T.A., F. Clark, and J. Muilu, *Conservation of human alternative splice events in mouse.* Nucleic Acids Research, 2003. **31**(10): p. 2544-52.
94. Sugnet, C.W., et al., *Transcriptome and genome conservation of alternative splicing events in humans and mice.* Pacific Symposium on Biocomputing, 2004: p. 66-77.
95. Nurtdinov, R.N., et al., *Low conservation of alternative splicing patterns in the human and mouse genomes.* Human Molecular Genetics, 2003. **12**(11): p. 1313-20.
96. Kan, Z., D. States, and W. Gish, *Selecting for functional alternative splices in ESTs.* Genome Research, 2002. **12**(12): p. 1837-45.
97. Resch, A., et al., *Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation.* Nucleic Acids Research, 2004. **32**(4): p. 1261-9.

98. Wang, B.B. and V. Brendel, *Genomewide comparative analysis of alternative splicing in plants.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(18): p. 7175-80.
99. Shin, C. and J.L. Manley, *The SR protein SRp38 represses splicing in M phase cells.* Cell, 2002. **111**(3): p. 407-17.
100. Moore, M.J., et al., *An alternative splicing network links cell-cycle control to apoptosis.* Cell, 2010. **142**(4): p. 625-36.
101. Demir, E. and B.J. Dickson, *fruitless splicing specifies male courtship behavior in Drosophila.* Cell, 2005. **121**(5): p. 785-94.
102. Stothard, P. and D. Pilgrim, *Sex-determination gene and pathway evolution in nematodes.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2003. **25**(3): p. 221-31.
103. Boutz, P.L., et al., *A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons.* Genes and Development, 2007. **21**(13): p. 1636-52.
104. Calarco, J.A., et al., *Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein.* Cell, 2009. **138**(5): p. 898-910.
105. Ule, J., et al., *Nova regulates brain-specific splicing to shape the synapse.* Nature Genetics, 2005. **37**(8): p. 844-52.
106. Brooks, A.N., et al., *Conservation of an RNA regulatory map between Drosophila and mammals.* Genome Research, 2011. **21**(2): p. 193-202.
107. Ast, G., *How Did Alternative Splicing Evolve.* Nature Reviews in Genetics, 2004. **5**(10): p. 773-82.
108. Loftus, B.J., et al., *The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans.* Science, 2005. **307**(5713): p. 1321-4.
109. Labadorf, A., et al. *Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii*. BMC Genomics, 2010. **11**.
110. Wang, B., et al., *Survey of the transcriptome of Aspergillus oryzae via massively parallel mRNA sequencing.* Nucleic Acids Research, 2010. **38**(15): p. 5075-87.
111. Sorber, K., M.T. Dimon, and J.L. DeRisi, *RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts.* Nucleic Acids Research, 2011. **39**(9): p. 3820-35.
112. Ramani, A.K., et al., *Genome-wide analysis of alternative splicing in Caenorhabditis elegans.* Genome Research, 2011. **21**(2): p. 342-8.
113. Filichkin, S.A., et al., *Genome-wide mapping of alternative splicing in Arabidopsis thaliana.* Genome Research, 2010. **20**(1): p. 45-58.
114. Lu, T., et al., *Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq.* Genome Research, 2010. **20**(9): p. 1238-49.
115. Roca, X., R. Sachidanandam, and A.R. Krainer, *Determinants of the inherent strength of human 5′ splice sites.* RNA, 2005. **11**(5): p. 683-98.
116. Sorek, R., et al., *Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons.* Molecular Cell, 2004. **14**(2): p. 221-31.
117. Koren, E., G. Lev-Maor, and G. Ast *The emergence of alternative 3′ and 5′ splice site exons from constitutive exons*. PLoS Computational Biology, 2007. **3**.

118. Irimia, M., D. Penny, and S.W. Roy, *Coevolution of genomic intron number and splice sites.* Trends in Genetics, 2007. **23**(7): p. 321-5.
119. Kim, E., A. Magen, and G. Ast, *Different levels of alternative splicing among eukaryotes.* Nucleic Acids Research, 2007. **35**(1): p. 125-31.
120. McGuire, A.M., et al. *Cross-kingdom patterns of alternative splicing and splice recognition*. Genome Biology, 2008. **9**.
121. Krutovsky, V. and T.E. Koralewski *Evolution of exon-intron structure and alternative splicing*. PloS One, 2011. **6**.
122. Robberson, B.L., G.J. Cote, and S.M. Berget, *Exon definition may facilitate splice site selection in RNAs with multiple exons.* Molecular and Cellular Biology, 1990. **10**(1): p. 84-94.
123. Berget, S.M., *Exon recognition in vertebrate splicing.* The Journal of Biological Chemistry, 1995. **270**(6): p. 2411-4.
124. Fox-Walsh, K.L., et al., *The architecture of pre-mRNAs affects mechanisms of splice-site pairing.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(45): p. 16176-81.
125. Xing, Y. and C. Lee, *Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes.* Nature Reviews in Genetics, 2006. **7**(7): p. 499-509.
126. Keren, H., G. Lev-Maor, and G. Ast, *Alternative splicing and evolution: diversification, exon definition and function.* Nature Reviews in Genetics, 2010. **11**(5): p. 345-55.
127. James-Clark, H., *On the spongiae ciliatae as infusoria flagellata; or observations on the structure, animality, and relationship of Leucosolenia botryoides.* Annals and Magazine of Natural History, 1868. **1**: p. 133-142; 188-215; 250-264.
128. Saville Kent, W., *A Manual of the Infusoria.* 1880-1882, London: David Bogue.
129. Steenkamp, E.T., J. Wright, and S.L. Baldauf, *The protistan origins of animals and fungi.* Molecular Biology and Evolution, 2006. **23**(1): p. 93-106.
130. Medina, M., et al., *Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa.* International Journal of Astrobiology, 2003. **2**(3): p. 203-211.
131. Philippe, H., et al., *Phylogenomics of eukaryotes: impact of missing data on large alignments.* Molecular Biology and Evolution, 2004. **21**(9): p. 1740-52.
132. Lang, B.F., et al., *The closest unicellular relatives of animals.* Current Biology, 2002. **12**(20): p. 1773-8.
133. Burger, G., et al., *Unique mitochondrial genome architecture in unicellular relatives of animals.* Proceedings of the National Academies of Science, 2003. **100**(3): p. 892-7.
134. Lavrov, D.V., et al., *Mitochondrial genomes of two demosponges provide insights into an early stage of animal evolution.* Molecular Biology and Evolution, 2005. **22**(5): p. 1231-9.
135. King, N. and S.B. Carroll, *A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution.* Proceedings of the National Academies of Science, 2001. **98**(26): p. 15032-7.

136. King, N., C.T. Hittinger, and S.B. Carroll, *Evolution of key cell signaling and adhesion protein families predates animal origins.* Science, 2003. **301**(5631): p. 361-3.
137. Segawa, Y., et al., *Functional development of Src tyrosine kinases during evolution from a unicellular ancestor to multicellular animals.* Proceedings of the National Academies of Science, 2006. **103**(32): p. 12021-6.
138. Snell, E.A., et al., *An unusual choanoflagellate protein released by Hedgehog autocatalytic processing.* Proceedings of the National Academies of Science, 2006. **273**(1585): p. 401-7.
139. Rokas, A., D. Kruger, and S.B. Carroll, *Animal evolution and the molecular signature of radiations compressed in time.* Science, 2005. **310**(5756): p. 1933-8.
140. Buck, K.R. and D.L. Garrison, *Distribution and Abundance of Choanoflagellates (Acanthoecidae) across the Ice-Edge Zone in the Weddell Sea, Antarctica.* Marine Biology, 1988. **98**(2): p. 263-269.
141. Thomsen, H.A. and J. Larsen, *Loricate Choanoflagellates of the Southern-Ocean with New Observations on Cell-Division in Bicosta-Spinifera (Throndsen, 1970) from Antarctica and Saroeca-Attenuata Thomsen, 1979, from the Baltic Sea.* Polar Biology, 1992. **12**(1): p. 53-63.
142. Hartmut, A., et al., *Functional diversity of heterotrophic flagellates in aquatic ecosystems*, in *The Flagellates*, B. Leadbeater and J.C. Green, Editors. 2000, Taylor & Francis: London. p. 240-268.
143. Boenigk, J. and H. Arndt, *Bacterivory by heterotrophic flagellates: community structure and feeding strategies.* Antonie van Leeuwenhoek, 2002. **81**(1-4): p. 465-80.
144. Leadbeater, B.S.C., *Life-history and ultrastructure of a new marine species of Proterospongia (Choanoflagellida).* Journal of the Marine Biological Association U.K., 1983. **63**: p. 135-160.
145. Fairclough, S.R., M.J. Dayel, and N. King, *Multicellular development in a choanoflagellate.* Current Biology, 2010. **20**(20): p. R875-6.
146. Douzery, E.J., et al., *The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils?* Proceedings of the National Academy of Sciences, 2004. **101**(43): p. 15386-91.
147. Xiao, S. and M. Laflamme, *On the eve of animal radiation: phylogeny, ecology and evolution of the Ediacara biota.* Trends in Ecology & Evolution, 2009. **24**(1): p. 31-40.
148. Shen, Y., T. Zhang, and P.F. Hoffman, *On the coevolution of Ediacaran oceans and animals.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(21): p. 7376-81.
149. Chapman, J.A., *Whole genome shotgun assembly in theory and practice*, in *Physics.* 2004, University of California, Berkeley: Berkeley.
150. Kondrashov, F.A. and E.V. Koonin, *Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences.* Trends in Genetics, 2003. **19**(3): p. 115-9.
151. Swofford, D.L., *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, 2002, Sinauer: Sunderland, MA.

152. Roy, S.W., Gilbert, W, *Complex early genes.* Proceedings of the National Academy of Sciences, 2005. **102**(6): p. 1986-1991.
153. Csuros, M. *Likely scenarios of intron evolution*. in *Proceedings of the Comparative Genomics: RECOMB 2005 International Workshop*. 2005. Dublin, Ireland: Springer-Verlag.
154. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Research, 2006. **34**(Database issue): p. D247-51.
155. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Research, 2004. **32**(Database issue): p. D138-41.
156. Letunic, I., et al., *SMART 5: domains in the context of genomes and networks.* Nucleic Acids Res, 2006. **34**(Database issue): p. D257-60.
157. Yang, S., R.F. Doolittle, and P.E. Bourne, *Phylogeny determined by protein domain content.* Proceedings of the National Academies of Science, 2005. **102**(2): p. 373-8.
158. Eichinger, L., et al., *The genome of the social amoeba Dictyostelium discoideum.* Nature, 2005. **435**(7038): p. 43-57.
159. Consortium, T.H.G.S., *Insights into social insects from the genome of the honeybee Apis mellifera.* Nature, 2006. **443**(7114): p. 931-49.
160. Lindvall, J.M., et al., *Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling.* Immunology Reviews, 2005. **203**: p. 200-15.
161. Chothia, C. and E.Y. Jones, *The molecular structure of cell adhesion molecules.* Annual Reviews in Biochemistry, 1997. **66**: p. 823-62.
162. Harpaz, Y. and C. Chothia, *Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains.* Journal of Molecular Biology, 1994. **238**(4): p. 528-39.
163. Williams, A.F. and A.N. Barclay, *The immunoglobulin superfamily--domains for cell surface recognition.* Annual Reviews in Immuniology, 1988. **6**: p. 381-405.
164. Baneres, J.L., et al., *A minimized human integrin alpha(5)beta(1) that retains ligand recognition.* Journal of Biological Chemistry, 2000. **275**(8): p. 5888-903.
165. Wessel, G.M., et al., *A molecular analysis of hyalin--a substrate for cell adhesion in the hyaline layer of the sea urchin embryo.* Developmental Biology, 1998. **193**(2): p. 115-26.
166. Pires-daSilva, A. and R.J. Sommer, *The evolution of signalling pathways in animal development.* Nature Reviews in Genetics, 2003. **4**(1): p. 39-49.
167. Barolo, S. and J.W. Posakony, *Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling.* Genes and Development, 2002. **16**(10): p. 1167-81.
168. Gerhart, J., *1998 Warkany lecture: signaling pathways in development.* Teratology, 1999. **60**(4): p. 226-39.
169. Nichols, S.A., et al., *Early evolution of animal cell signaling and adhesion genes.* Proceedings of the National Academies of Science, 2006. **103**(33): p. 12451-6.

170. Adell, T., I. Nefkens, and W.E. Muller, *Polarity factor 'Frizzled' in the demosponge Suberites domuncula: identification, expression and localization of the receptor in the epithelium/pinacoderm(1).* FEBS Letters, 2003. **554**(3): p. 363-8.
171. Adell, T., A.N. Thakur, and W.E. Muller, *Isolation and characterization of Wnt pathway-related genes from Porifera.* Cell Biology International, 2007. **31**(9): p. 939-49.
172. Artavanis-Tsakonas, S., M.D. Rand, and R.J. Lake, *Notch signaling: cell fate control and signal integration in development.* Science, 1999. **284**(5415): p. 770-6.
173. Gazave, E., et al., *Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes.* BMC Evolutionary Biology, 2009. **9**: p. 249.
174. Ekman, D., A.K. Bjorklund, and A. Elofsson, *Quantification of the elevated rate of domain rearrangements in metazoa.* Journal of Molecular Biology, 2007. **372**(5): p. 1337-48.
175. Dayel, M.J., et al., *Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta.* Developmental Biology, 2011. **357**: p. 73-82.
176. Bentley, D., *The mRNA assembly line: transcription and processing machines in the same factory.* Current Opinion in Cell Biology, 2002. **14**(3): p. 336-42.
177. Lee, C., et al., *Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution.* RNA, 2010. **16**(1): p. 91-105.
178. Graveley, B.R., *Mutually exclusive splicing of the insect Dscam pre-mRNA is directed by competing intronic RNA secondary structures.* Cell, 2005. **123**: p. 65-73.
179. Reva, O. and B. Tummler, *Think big--giant genes in bacteria.* Environmental Microbiology, 2008. **10**(3): p. 768-777.
180. Cantu, D., et al., *Next Generation Sequencing Provides Rapid Access to the Genome of Puccinia striiformis f. sp. tritici, the Causal Agent of Wheat Stripe Rust.* PloS One, 2011. **6**(8): p. e24230.
181. Dalloul, R.A., et al., *Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis.* PLoS Biology, 2010. **8**(9).
182. Metzker, M.L., *Sequencing technologies - the next generation.* Nature Reviews in Genetics, 2010. **11**(1): p. 31-46.
183. Ruiz-Trillo, I., et al., *The origins of multicellularity: a multi-taxon genome initiative.* Trends in Genetics, 2007. **23**(3): p. 113-8.
184. Shalchian-Tabrizi, K., et al., *Multigene phylogeny of choanozoa and the origin of animals.* PloS One, 2008. **3**(5): p. e2098.
185. Ruiz-Trillo, I., et al., *Capsaspora owczarzaki is an independent opisthokont lineage.* Current Biology, 2004. **14**(22): p. R946-7.
186. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature Biotechnology, 2010. **28**(5): p. 511-5.

187. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
188. Au, K.F., et al., *Detection of splice junctions from paired-end RNA-seq data by SpliceMap.* Nucleic Acids Research, 2010. **38**(14): p. 4570-8.
189. Jean, G., et al., *RNA-Seq read alignments with PALMapper.* Current Protocols in Bioinformatics, 2010. **Chapter 11**: p. 116.
190. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature Biotechnology, 2011. **29**(7): p. 644-52.
191. Leys, S.P. and B.M. Degnan, *Cytological basis of photoresponsive behavior in a sponge larva.* Biological Bulletin, 2001. **201**: p. 323-338.
192. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
193. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Research, 2008. **Database Issue 36**: p. D281-D288.
194. Valentine, J.W., *On the Origin of Phyla,* 2004, Chicago: University of Chicago Press.
195. King, N., et al., *The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans.* Nature, 2008. **451**(7180): p. 783-8.
196. Jeffares, D.C., T. Mourier, and D. Penny, *The biology of intron gain and loss.* Trends in Genetics, 2005. **22**(1): p. 16-22.
197. Stajich, J.E. and F.S. Dietrich, *Evidence of mRNA-mediated intron loss in the human-pathogenic fungus Cryptococcus neoformans.* Eukaryotic Cell, 2006. **5**(5): p. 789-793.
198. Sverdlov, A.V., et al., *Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion.* Gene, 2004. **338**: p. 85-91.
199. Kim, E., A. Goren, and G. Ast, *Alternative splicing: current perspectives.* Bioessays, 2008. **30**(1): p. 38-47.
200. Benson, D.A., et al., *GenBank.* Nucleic Acids Research, 2005. **33**(Database Issue): p. D34-D38.
201. Labeit, S., et al., *Expression of distinct classes of Titin isoforms in striated and smooth muscles by alternative splicing, and their conserved interaction with filamins.* Journal of Molecular Biology, 2006. **362**(4): p. 664-81.
202. Hattori, D., et al., *Dscam diversity is essential for neuronal wiring and self-recognition.* Nature, 2007. **449**(7159): p. 223-7.
203. Vogel, C., S.A. Teichmann, and J. Pereira-Leal, *The relationship between domain duplication and recombination.* Journal of Molecular Biology, 2005. **346**(1): p. 355-365.
204. Audibert, A., D. Weil, and F. Dautry, *In vivo kinetics of mRNA splicing and transport in mammalian cells.* Molecular Cell Biology, 2002. **22**(19): p. 6706-6718.
205. Jeffares, D.C., C.J. Penkett, and J. Bahler, *Rapidly regulated genes are intron poor.* Trends in Genetics, 2008. **24**(8): p. 375-8.
206. Stoebel, D.M., A.M. Dean, and D.E. Dykhuizen, *The cost of expression of Escheria coli lac operon proteins is in the process, not in the products.* Genetics, 2008. **178**: p. 1653-1660.

207. Bjorklund, A.K., D. Ekman, and A. Elofsson, *Expansion of protein domain repeats.* PLoS Computational Biology, 2006. **2**(8): p. e144.
208. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
209. Kalsotra, A. and T.A. Cooper, *Functional consequences of developmentally regulated alternative splicing.* Nature Reviews in Genetics, 2011. **12**(10): p. 715-29.
210. Buss, L., *The Evolution of Individuality,* 1988, Princeton: Princeton University Press.
211. Carroll, S.B., *Endless forms: the evolution of gene regulation and morphological diversity.* Cell, 2000. **101**(6): p. 577-80.
212. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-51.
213. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.* Nature Reviews in Molecular and Cell Biology, 2009. **10**(11): p. 741-54.
214. Graveley, B.R., et al., *The developmental transcriptome of Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 473-9.
215. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nature Genetics, 2008. **40**(12): p. 1413-5.
216. Wang, K., D.W. Ussery, and S. Brunak, *Analysis and prediction of gene splice sites in four Aspergillus genomes.* Fungal Genetics and Biology, 2009. **46**: p. S14-S18.
217. Moldón, A., et al., *Promoter-driven splicing regulation in fission yeast.* Nature, 2008. **455**(7215): p. 997-1000.
218. Singh, N., et al., *Conservation and developmental control of alternative splicing in maebl among malaria parasites.* Journal of Molecular Biology, 2004. **343**(3): p. 589-99.
219. Falciatore, A., et al., *The FLP proteins act as regulators of chlorophyll synthesis in response to light and plastid signals in Chlamydomonas.* Genes and Development, 2005. **19**(1): p. 176-87.
220. Kinoshita, S., et al., *A novel heat stress-responsive gene in the marine diatom Chaetoceros compressum encoding two types of transcripts, a trypsin-like protease and its related protein, by alternative RNA splicing.* European Journal of Biochemistry / FEBS, 2001. **268**(17): p. 4599-609.
221. Valluet, A., et al., *B-raf alternative splicing is dispensable for development but required for learning and memory associated with the hippocampus in the adult mouse.* PloS One, 2010. **5**(12): p. e15272.
222. King, N., et al., *Starting and maintaining Monosiga brevicollis cultures.* Cold Spring Harbor Protocols, 2009. **2**.
223. Stajich, J.E., et al., *The Bioperl Toolkit: Perl modules for the life sciences.* Genome Research, 2002. **12**(10): p. 1611-8.
224. Crooks, G.E., et al., *WebLogo: A sequence logo generator.* Genome Research, 2004. **14**: p. 1188-1190.

225. Blanchette, M., et al., *Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila.* Genes and Development, 2005. **19**(11): p. 1306-14.
226. Burset, M., A. Seledtsov, and V.V. Solovyev, *Analysis of canonical and non-canonical splice sites in mammalian genomes.* Nucleic Acids Research, 2000. **28**(21): p. 4364-4375.
227. Sheth, N., et al., *Comprehensive splice site-analysis using comparative genomics.* Nucleic Acids Research, 2006. **34**(14): p. 3955-3967.
228. Alioto, T.S., *U12DB: a database of orthologous U12-type spliceosomal introns.* Nucleic Acids Research, 2006. **35**(1): p. 110-115.
229. Bartschat, S. and T. Samuelsson, *U12 type introns were lost at multiple occasions during evolution.* BMC Genomics, 2010. **11**: p. 106.
230. Chen, L. and S. Zheng, *Identify alternative splicing events based on position-specific evolutionary conservation.* PLoS One, 2008. **3**(7): p. e2806.
231. Lv, J., et al., *Molecular determinants and evolutionary dynamics of wobble splicing.* Molecular Biology and Evolution, 2009. **26**(5): p. 1081-92.
232. Hiller, M. and M. Platzer, *Widespread and subtle: alternative splicing at short-distance tandem sites.* Trends in Genetics, 2008. **24**(5): p. 246-55.
233. Hiller, M., et al., *Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.* Nature Genetics, 2004. **36**(12): p. 1255-7.
234. Tsai, K.W. and W.C. Lin, *Quantitative analysis of wobble splicing indicates that it is not tissue specific.* Genomics, 2006. **88**(6): p. 855-64.
235. Tress, M.L., et al., *The implications of alternative splicing in the ENCODE protein complement.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(13): p. 5495-500.
236. Bodescot, M. and O. Brison, *Characterization of new human c-myc mRNA species produced by alternative splicing.* Gene, 1996. **174**(1): p. 115-20.
237. Rick, M., et al., *Nuclear localization of Annexin A7 during murine brain development.* BMC Neuroscience, 2005. **6**(25).
238. Brignatz, C., et al., *Alternative splicing modulates autoinhibition and SH3 accessibility in the Src kinase Fyn.* Molecular Cell Biology, 2009. **29**(24): p. 6438-48.
239. Gan, Q., et al., *Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq.* Cell Research, 2010. **20**(7): p. 763-83.
240. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biology, 2010. **11**: p. R106.
241. Robinson, M.D. and G.K. Smyth, *Moderated statistical tests for assessing differences in tag abundance.* Bioinformatics, 2007. **23**(21): p. 2881-87.
242. Talerico, M. and S.M. Berget, *Effect of 5' splice site mutations on splicing of the preceding intron.* Molecular and Cellular Biology, 1990. **10**(12): p. 6299-305.
243. Ellis, J.D., et al., *Spatial mapping of splicing factor complexes involved in exon and intron definition.* Journal of Cell Biology, 2008. **181**(6): p. 921-34.

244. Romfo, C.M., et al., *Evidence for splice site pairing via intron definition in Schizosaccharomyces pombe.* Molecular and Cellular Biology, 2000. **20**(21): p. 7955-70.
245. Sterner, D.A., T. Carlo, and S.M. Berget, *Architectural limits on split genes.* Proceedings of the National Academies of Science, 1996. **93**: p. 15081-85.
246. Ram, O. and G. Ast, *SR proteins: a foot on the exon before the transition from intron to exon definition.* Trends in Genetics, 2007. **23**(1): p. 5-7.
247. Barbosa-Morais, N.L., M. Carmo-Fonesca, and S. Aparicio, *Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion.* Genome Research, 2006. **16**: p. 66-77.
248. Shen, H. and M.R. Green, *RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans.* Genes and Development, 2006. **20**: p. 1755-65.
249. Tsai, K.W., H.C. Tseng, and W.C. Lin, *Two wobble-splicing events affect ING4 protein subnuclear localization and degradation.* Experimental Cell Research, 2008. **314**(17): p. 3130-41.
250. Raho, G., et al., *Detection of novel mRNA splice variants of human tumor suppressor gene ING4 alternative splicing.* Oncogene, 2007. **26**: p. 5247-57.
251. Knightley, P.D. and D.J. Gaffney, *Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents.* Proceedings of the National Academies of Science, 2003. **100**(23): p. 13402-6.
252. Vinogradov, A.E., *Growth and decline of introns.* Trends in Genetics, 2002. **18**(5): p. 232-6.