Intron-Dominated Genomes of Early Ancestors of Eukaryotes

EUGENE V. KOONIN

From the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

Address correspondence to Eugene V. Koonin at the address above, e-mail: koonin@ncbi.nlm.nih.gov.

Abstract

Evolutionary reconstructions using maximum likelihood methods point to unexpectedly high densities of introns in protein-coding genes of ancestral eukaryotic forms including the last common ancestor of all extant eukaryotes. Combined with the evidence of the origin of spliceosomal introns from invading Group II self-splicing introns, these results suggest that early ancestral eukaryotic genomes consisted of up to 80% sequences derived from Group II introns, a much greater contribution of introns than that seen in any extant genome. An organism with such an unusual genome architecture could survive only under conditions of a severe population bottleneck.

Key words: effective population size, endosymbiosis, group II self-splicing introns, origin of eukaryotes, spliceosomal introns

Spliceosomal introns that interrupt the protein-coding genes and the spliceosome, a highly complex RNA–protein machine that mediates intron excision and exon splicing is among the hallmarks of eukaryotes (Doolittle 1978; Gilbert 1978; Mattick 1994; Deutsch and Long 1999). All eukaryotes with sequenced genomes, including unicellular organisms with compact genomes, previously suspected to be intron-less, have been shown to possess at least a few introns (Nixon et al. 2002; Simpson et al. 2002; Vanacova et al. 2005) and a (nearly) complete spliceosome (Collins and Penny 2005). The intron density in eukaryotes ranges from a few introns per genome to more than 8 per gene (Logsdon 1998; Mourier and Jeffares 2003; Jeffares et al. 2006). The intron content of a genome is generally thought to be determined by the efficiency of purifying selection that itself depends on the effective population size of the corresponding organism (Lynch 2007). The population-genetic theory has it that, because of their large effective population size leading to highly efficient purifying selection, most of the unicellular eukaryotes cannot sustain more than a small number of short introns (Lynch 2002, 2006; Lynch and Richardson 2002; Lynch and Conery 2003). In contrast, proliferation of introns is possible in genomes of multicellular eukaryotes (plants and animals) that have small populations subject to relatively inefficient purifying selection pressure but high levels of random drift.

For the first 25 years after the discovery of splicing, the study of intron evolution revolved, primarily, around the more or less speculative debate between the introns-early and introns-late concepts. The introns-early hypothesis (more recently recast in the form of “introns-first”) posits that very first protein-coding genes already contained introns that contributed to the emergence of proteins via recombination between RNA molecules that encoded short peptides (Doolittle 1978; Gilbert 1978; Gilbert and Glyniás 1993; Gilbert et al. 1997; Jeffares et al. 2006). The introns-late concept counters that the primordial genes were intronless, and prokaryotic genes have retained that primitive state, whereas eukaryotic genes were invaded by introns only after (or during) the onset of the eukaryotic lineage (Stoltzfus et al. 1994; Logsdon et al. 1995; Logsdon 1998). Given the absence of the spliceosome and spliceosomal introns in prokaryotes, the failure of key predictions of the introns-early hypothesis, such as those on differences in intron phase distributions among ancient and more recent introns (Rogozin et al. 2003) and conservation of intron positions between ancient paralogs (Cho and Doolittle 1997; Sverdlov et al. 2007), and the uncertainty surrounding other types of evidence, such as intron–domain correspondence (Roy and Gilbert 2006), the original introns-early concept does not seem to be a viable contender anymore (Koonin 2006).

However, as far as evolution of eukaryotes themselves is concerned, introns do seem to be a very early acquisition, perhaps, contemporary with the origin of the eukaryotic cell, as indicated by the presence of at least a few introns and, most importantly, a (nearly) full-fledged spliceosome in all eukaryotes with sequenced genomes (Collins and Penny 2005; Koonin 2006; Roy 2006; Roy and Gilbert 2006). An intriguing question is, what were the dynamics of intron
content over the course of eukaryotic evolution? Given that the last eukaryotic common ancestor (LECA) was, in all likelihood, unicellular, and most of the extant unicellular eukaryotes have intron-poor genes, it might seem logical to infer that LECA had a low intron density, and subsequently, some of the eukaryotic lineages have accrued many more introns. However, the latest comparative genomic analyses increasingly suggest that this scenario is unlikely to be correct.

Reconstruction of Intron Gain–Loss Events during the Evolution of Eukaryotes

Numerous intron positions in orthologous genes are conserved at great evolutionary depths: for instance, orthologous genes from human and Arabidopsis share up to 25% of intron positions (Fedorov et al. 2002; Rogozin et al. 2003). Sequencing of multiple genomes of diverse eukaryotes created the possibility for reconstruction of intron content in ancestral forms for which purpose maximum parsimony and maximum likelihood (ML) methods have been applied (Rogozin et al. 2005). The maximum parsimony approaches are straightforward but are intrinsically overrelent on the conservation of the analyzed characters (intron positions, in this case) and might severely underestimate the intron content in ancestral genomes. More sophisticated ML methods that model the rates of intron gain and loss along branches of a phylogenetic tree have the potential to yield higher and, conceivably, more realistic estimates. The results of such reconstructions strongly suggest that protein-coding genes of ancient eukaryotic ancestors including LECA already possessed intron density comparable to that found in modern, moderately intron-rich genomes (Csuros 2005; Nguyen et al. 2005; Roy and Gilbert 2005a, 2005b; Carmel et al. 2007). The estimate for LECA, although characterized by a relatively high uncertainty due to methodological reasons, conservatively indicates the presence of at least 2 introns per kilobase of protein-coding DNA (Figure 1). Moreover, a recent reconstruction of intron content in the ancestors of alveolates and chromalveolates that employed an advanced ML technique yielded an unexpected result: It was inferred that the ancestors of these intron-poor unicellular eukaryotes had high intron densities, comparable to those in vertebrates, and also that LECA contained at least 3 introns per kilobase of coding DNA (Csuros et al. 2008). Thus, the ancestors of at least 3 of the 5 eukaryotic supergroups (Keeling et al. 2005), Chromalveolata, Plantae, and Unikonts (animals, fungi, and amoebzoa), most likely had intron-rich genomes (the genomic data for the remaining 2 supergroups are currently insufficient for specific inferences). Accordingly, the history of eukaryotic genes appears to have been, to a large extent, a story of extensive loss of ancestral introns, perhaps, punctuated with a few episodes of major gain (Roy 2006; Carmel et al. 2007). These conclusions, which suggest that the above estimate of at least 2 introns per kilobase of protein-coding DNA in LECA is highly conservative, have interesting implications for the genome architecture and population dynamics of early eukaryotic ancestors.

The First Eukaryotes: Did They Have Intron-Dominated Genomes?

Although the ultimate origin of eukaryotic spliceosomal introns is not known with certainty, the current dominant hypothesis has it that they are derivatives of prokaryotic Group II self-splicing introns that gave rise both to introns and to the active RNA moieties of the spliceosome (Lambowitz and Zimmerly 2004; Robart and Zimmerly 2005; Martin and Koonin 2006). Recently, this hypothesis received a strong boost from the resolved structure of a Group II intron that showed extensive similarities to the structures of spliceosomal snRNAs and the ends of introns themselves (Toor et al. 2008).

It has been further proposed that the invasion of the ancestral eukaryotic genome by Group II introns was triggered by the mitochondrial endosymbiosis (Martin and Koonin 2006). Under this hypothesis, the α-proteobacterial ancestor of the mitochondria contained multiple Group II introns (this is compatible with the relatively high abundance of these elements in some α-proteobacteria [Robart and Zimmerly 2005]) that became unleashed, in part, because of recurrent release of the symbiont DNA into the host cell. This scenario of intron invasion does not depend on the nature of the organism that hosted the
mitochondrial endosymbiont, that is, whether it was a typical archaeon as posited by the symbiotic hypotheses of eukaryogenesis (Embley and Martin 2006; Martin and Koonin 2006) or a distinct protoeukaryotic form as suggested by the archezoan hypothesis (Kurland et al. 2006; Poole and Penny 2007).

As suggested by the above estimates, the invasion of Group II introns led to a fairly high density of introns over the supposedly brief time interval that separated the acquisition of the mitochondrial endosymbiont from the advent of LECA. Let us assume that the intron invasion was instantaneous on the evolutionary scale, that is, rapid enough to disregard intron loss and deterioration. This assumption appears credible if the invasion was the direct consequence of endosymbiosis (Martin and Koonin 2006). The implications for the genome architecture of the immediate predecessors of LECA seem to be striking. Group II introns are complex elements that encode a large protein containing a reverse transcriptase domain and several accessory domains; accordingly, these elements have a (nearly) uniform size of approximately 2.5 kb (Lambowitz and Zimmerly 2004).

Thus, under the (near) instantaneous invasion scenario, that is, assuming that intron invasion occurred faster than substantial loss of intronic sequences, the median size of the introns in the (pre)LECA genome would be considerably greater than in any extant genomes, and the mean size would be greater than that in any modern forms, with the exception of mammals and some other vertebrates (Figure 2). In modern bacteria, Group II introns reside, mostly, in intergenic regions (and therefore should be more properly regarded as retroelements rather than bona fide introns), presumably, because, if an intron invades a functionally important protein-coding sequence, it is rapidly weeded out by the highly efficient purifying selection that affects large prokaryotic populations (Robart and Zimmerly 2005). By contrast, protein-coding genes of endosymbiotic organelles, namely, fungal and plant mitochondria and plant and algal chloroplasts, often carry bona fide, self-splicing introns (Toro et al. 2007). The latter situation could be the model for the events that transpired during the mitochondrial endosymbiosis except that the original intron invasion of protein-coding regions that quickly reached the inferred high intron density must have been a much more dramatic event, a virtual genome catastrophe that could realize only under the conditions of a major population bottleneck (see Implications for Eukaryogenesis: Intron Invasion Was Accompanied by a Major Population Bottleneck That Enabled Key Eukaryotic Innovations). All known genomes of nonparasitic prokaryotes possess short intergenic regions and a uniform mean gene size of about 1000 nucleotides (Koonin and Wolf 2008). Assuming that the prokaryotic host of the mitochondrial endosymbiont and, accordingly, the emerging eukaryote at the earliest stages of eukaryogenesis possessed the typical prokaryotic genome architecture and that Group II introns inserted randomly into the coding and noncoding sequences, one can estimate the fraction of its genome allotted to introns. The estimated intron density of 2 introns per kilobase of the coding sequence and the mean intron length of 2.5 kb yield: $\frac{5}{6} \times 0.85 = 0.71$, that is, more than 70% of the genome of the pre-LECA eukaryotes would be occupied by introns, by far the greatest intron content compared with modern eukaryotic genomes (Figure 3). Thus, the ancient eukaryotic forms might have had literally intron-dominated genomes so
that the rest of eukaryotic evolution, with some notable exceptions like the expansion of introns in mammals and, possibly, some short episodes of substantial intron gain (Carmel et al. 2007), could be a story of intron loss and shrinkage.

It is worth noting that the above inference of the extremely high fraction of intron sequences in the genomes of primordial eukaryotes is predicated on the scenario under which the primary invasion of introns antedated the emergence of the spliceosome (Martin and Koonin 2006). In principle, the reverse order of events is conceivable whereby the spliceosome that, clearly, preceded LECA (Collins and Penny 2005) evolved prior to intron invasion, perhaps, performing a different function, and was available to catalyze the excision of the invading Group II introns. However, considering the apparent homology between the catalytic snRNAs of the spliceosome and the ribozyme part of Group II introns (Toor et al. 2008), the invasion-first scenario appears distinctly more plausible.

**Implications for Eukaryogenesis: Intron Invasion Was Accompanied by a Major Population Bottleneck That Enabled Key Eukaryotic Innovations**

Nonfunctional, in particular, intronic DNA is subject to purifying selection that purges nonfunctional sequences from genomes in large populations where the selection pressure is strong. Nonfunctional sequences can be fixed and survive for extended time intervals only in small populations with weak selective pressure. According to population-genetic theory, the condition for the fixation of introns is \( Ng/mu < 1 \) where \( Ng \) is the effective number of genes per locus (related to the more traditional effective population size), \( n \) is the number of nucleotides that are required for intron splicing, and \( u \) is mutation rate per nucleotide per generation (Lynch 2007). Conservatively, it can be assumed that splicing requires, approximately, 25 nucleotides/intron (Lynch 2002). We first consider an extreme, “worst case” scenario under which all the introns are acquired by the pre-LECA eukaryote, literally, in one coup. Assuming that the organism in question possessed, approximately, 5000 genes (a reasonable size after the putative fusion of 2 prokaryotic genomes) with the intron density of \( \sim 2 \) introns/gene (see above), we get \( \sim 10^4 \) introns per genome and \( u \approx 2.5 \times 10^{-9} \). Using a conservative estimate of \( u \approx 5 \times 10^{-9} \) (Lynch 2007), the condition for intron fixation after rapid invasion is \( Ng < 1/mu = 1/2.5 \times 10^5 \times 5 \times 10^{-9} \) or \( Ng < 10^{6} \). According to this estimate, the stage of eukaryogenesis evolution between the mitochondrial endosymbiosis and the advent of LECA was characterized by an extremely severe population bottleneck (Figure 4). Of course, the literally simultaneous invasion of all introns assumed in this scenario is unlikely, so this should be regarded as an absolute low bound of \( Ng \) during eukaryogenesis. The most conservative, “best case” scenario would have the introns invading one by one, each new intron arriving only after the fixation of the preceding one. Then \( Ng < 1/mu = 1/2.5 \times 5 \times 10^{-9} \) or \( Ng < 10^{7} \). Taken literally, this scenario, with the successive invasion of all \( \sim 10^4 \) introns, is unlikely as well, so the actual characteristic \( Ng \) values of pre-LECA eukaryotes, in all likelihood, lie in the range between these estimates, presumably, below \( Ng \sim 10^{6} \), a characteristic value in unicellular eukaryotes that prevents massive fixation of introns (Lynch 2007).

This bottleneck that enabled the fixation of numerous, large introns also would have been the key condition for the other dramatic innovations associated with eukaryogenesis including extensive duplication of many key genes (Makarova et al. 2005) and the emergence of a variety of eukaryote-specific cellular structures. Within the framework of the symbiotic scenario of eukaryogenesis, it appears likely that the symbiosis directly caused the bottleneck (Martin and Koonin 2006). Conceivably, then, the course of early history of eukaryotes, in terms of population dynamics, can be represented as transition from a typical prokaryotic population, with \( Ng \sim 10^{9} \) to the bottleneck with, perhaps, \( Ng \sim 10^{6} \) to a partial rebound reaching \( Ng \sim 10^{6} \), a typical value for unicellular eukaryotes (Figure 4).

**Conclusion**

Although the introns-early theory in its original form does not seem to be viable, self-splicing introns, in all likelihood, existed even at the earliest stages of the evolution of life and invaded the emerging eukaryotic genomes, giving rise to splicosomal introns, shortly after the mitochondrial endosymbiosis (Koonin 2006; Koonin et al. 2006). As a result of the invasion of self-splicing introns, the early eukaryotic ancestor might have an unusual genome that consisted of \( \sim 70\% \) of intronic DNA. Such a heavy burden of (at least, originally) nonfunctional DNA could be sustained only under...
conditions of a population bottleneck that would also precipitate other innovations associated with eukaryogenesis.

Funding
Intramural Research Program of the DHHS/NIH (National Library of Medicine, National Center for Biotechnology Information).

Acknowledgments
I thank Igor Rogozin for help with data collection for Figures 2 and 3 and Michael Lynch for most helpful discussions and critical reading of the manuscript.

References


Corresponding Editor: Michael Lynch