Extremely Intron-Rich Genes in the Alveolate Ancestors Inferred with a Flexible Maximum-Likelihood Approach

Miklós Csúrós,*,† Igor B. Rogozin,‡ and Eugene V. Koonin‡

*Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec, Canada; †Collegium Budapest Institute for Advanced Study, Budapest, Hungary; and ‡National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

Chromalveolates are a large, diverse supergroup of unicellular eukaryotes that includes Apicomplexa, dinoflagellates, ciliates (three lineages that form the alveolate branch), heterokonts, haptophytes, and cryptomonads (three lineages comprising the chromist branch). All sequenced genomes of chromalveolates have relatively low intron density in protein-coding genes, and few intron positions are shared between chromalveolate lineages. In contrast, genes of different chromalveolates share many intron positions with orthologous genes from other eukaryotic supergroups, in particular, the intron-rich orthologs from animals and plants. Reconstruction of the history of intron gain and loss during the evolution of chromalveolates using a general and flexible maximum-likelihood approach indicates that genes of the ancestors of chromalveolates and, particularly, alveolates had unexpectedly high intron densities. It is estimated that the chromalveolate ancestor had, approximately, two-thirds of the human intron density, whereas the intron density in the genes of the alveolate ancestor is estimated to be slightly greater than the human intron density. Accordingly, it is inferred that the evolution of chromalveolates was dominated by intron loss. The conclusion that ancestral chromalveolate forms had high intron densities is unexpected because all extant unicellular eukaryotes have relatively few introns and are thought to be unable to maintain numerous introns due to intense purifying selection in their, typically, large populations. It is suggested that, at early stages of evolution, chromalveolates went through major population bottlenecks that were accompanied by intron invasion.

Introduction

Spliceosomal introns that interrupt most of the protein-coding genes and the concurrent splicing machinery that mediates intron excision and exon splicing are among the defining features of eukaryotes (Doolittle 1978; Gilbert 1978; Mattick 1994; Deutsch and Long 1999). To date, all eukaryotes with sequenced genomes, including parasitic protists with compact genomes, previously suspected to be intronless, 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) full complement of spliceosomal proteins (Collins and Penny 2005). Different species dramatically vary in their intron density, ranging from a few introns per genome to over 8 per gene (Logsdon 1998; Mourier and Jeffares 2003; Jeffares et al. 2006). Despite the ubiquity of introns in eukaryotic genomes, their natural history is poorly understood. To what extent introns are to be regarded as “junk DNA” as opposed to functional parts of the genome remains an open question. There are many reports on the contribution of introns to the regulation of gene expression (Bourdon et al. 2001; Le Hir et al. 2003; Rose 2004; Ying and Lin 2005), but it is unclear how general such functional roles of introns might be.

Much uncertainty also remains with regard to the origin and subsequent evolution of introns. For the last 30 years, the study of intron evolution had been coached, primarily, as a debate between the so-called introns-early and introns-late concepts. The introns-early view (more recently revived in the form of “introns-first”) holds that introns were part of the very first protein-coding genes and contributed to the emergence of proteins via recombination between RNA molecules that encoded short peptide

Key words: eukaryotic genome evolution, spliceosomal introns, chromalveolates, maximum likelihood, intron-rich ancestors.

E-mail: csuros@iro.umontreal.ca; koonin@ncbi.nlm.nih.gov.

doi:10.1093/molbev/msn039
Advance Access publication February 21, 2008

© 2008 The Authors.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/ uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
phylogenies that include, mostly, uncontested evolutionary formation of ancestral introns than intron-poor genomes). Intron-rich species (which are expected to convey more information about ancestral introns than intron-poor genomes). Among the available fungal and animal genomes, we selected a diverse set of land plant genomes and two fungal and animal genomes. We collected gene structure data from all publicly available, complete, annotated chromalveolate genomes in which spliceosomal introns are not uncommon. As outgroups, we used all available land plant genomes and two green algal genomes, as well as a comparable number of genomes from animals and fungi. Among the available fungal and animal genomes, we selected a diverse set of intron-rich species (which are expected to convey more information about ancestral introns than intron-poor genomes). Throughout the evolutionary analyses, a fixed organismal phylogeny that includes, mostly, uncontested evolutionary relationships was used; the tree includes 3 relevant eukaryotic supergroups, Chromalveolata, Plantae, andunikonts, with the relationship between them remaining unresolved and represented as a trifurcation (Adl et al. 2005; Keeling et al. 2005). The genome sequences were extracted from GenBank, the NCBI RefSeq database, or the Joint Genome Institute database; the details on the sources of protein sequences and exon-intron structure are given in supplementary table S1 (Supplementary Material online). The following eukaryotic species were included in the analysis: Plasmodium berghei (Pber), Plasmodium chabaudi chabaudi (Pcha), Plasmodium falciparum (Pf), Plasmodium yoelii yoelii (Pyoe), Theileria annulata (Tann), Theileria parva (Tpar), Paramaecium tetraurelia (Ptet), Tetrahymena thermophila (Tthe), Phaeodactylum tricornutum (Ftri), Phytophthora ramorum (Pram), Phytophthora sojae (Psoj), Arabidopsis thaliana (Ath), Oryza sativa ssp. japonica (Osat), Populus trichocarpa (Ptri), Chlamydomonas reinhardtii (Crei), Ostreococcus tauri (Otau), Apis mellifera (Amel), Homo sapiens (Hsap), Tribolium castaneum (Tcas), Coprinus cinereus (Ccim), Phycycomyces blakesleeanus (Pbla), Phanerochaete chrysosporium (Pchr), and Rhizopus oryzae (Rory).

Paralogous Gene Sets

Sets of paralogous genes were constructed by updating and extending the database of eukaryotic clusters of orthologous genes (KOGs) as follows. First, the KOG database covering 7 eukaryotic genomes (Tatusov et al. 2003) was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/COG/KOG/. Subsequently, each KOG was used as query to search clade-specific databases of protein sequences using the PSI-Blast program (Altschul et al. 1997; Schaffer et al. 2001). The searches were performed using command-line tools of the NCBI software development kit (Version 6.1, obtained from ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/ncbi.tar.gz). Using a Blast database for each of fungi, chromalveolates, and insects, the searches were performed using three iterations (switch –j 3) of PSI-Blast (blastpgp executable); for human and plant sequences, no iterations were used. For each KOG query, sequences with an E-value <10^{-9} were retained if they had a Blast hit score within 50% of the best hit for the species. In a further filtering step, reversed position-specific Blast search (Marchler-Bauer et al. 2007) was used to query each retained protein sequence against the Conserved Domain Database (CDD) of KOGs (ps blast executable with default parameters). Only those sequences passed this filter for which the highest scoring KOG hit was the same as the KOG used in the initial PSI-Blast search and the second highest scoring KOG had less than 90% of the highest score. Sequences from the same genome that were thus assigned to the same KOG comprised paralogous sets.

Materials and Methods

Data

We collected gene structure data from all publicly available, complete, annotated chromalveolate genomes in which spliceosomal introns are not uncommon. As outgroups, we used all available land plant genomes and two green algal genomes, as well as a comparable number of genomes from animals and fungi. Among the available fungal and animal genomes, we selected a diverse set of intron-rich species (which are expected to convey more information about ancestral introns than intron-poor genomes). Throughout the evolutionary analyses, a fixed organismal phylogeny that includes, mostly, uncontested evolutionary relationships was used; the tree includes 3 relevant eukaryotic supergroups, Chromalveolata, Plantae, and unikonts, with the relationship between them remaining unresolved and represented as a trifurcation (Adl et al. 2005; Keeling et al. 2005). The genome sequences were extracted from GenBank, the NCBI RefSeq database, or the Joint Genome Institute database; the details on the sources of protein sequences and exon-intron structure are given in supplementary table S1 (Supplementary Material online). The following eukaryotic species were included in the analysis: Plasmodium berghei (Pber), Plasmodium chabaudi chabaudi (Pcha), Plasmodium falciparum (Pf), Plasmodium yoelii yoelii (Pyoe), Theileria annulata (Tann), Theileria parva (Tpar), Paramaecium tetraurelia (Ptet), Tetrahymena thermophila (Tthe), Phaeodactylum tricornutum (Ftri), Phytophthora ramorum (Pram), Phytophthora sojae (Psoj), Arabidopsis thaliana (Ath), Oryza sativa ssp. japonica (Osat), Populus trichocarpa (Ptri), Chlamydomonas reinhardtii (Crei), Ostreococcus tauri (Otau), Apis mellifera (Amel), Homo sapiens (Hsap), Tribolium castaneum (Tcas), Coprinus cinereus (Ccim), Phycycomyces blakesleeanus (Pbla), Phanerochaete chrysosporium (Pchr), and Rhizopus oryzae (Rory).

Paralogous Gene Sets

Sets of paralogous genes were constructed by updating and extending the database of eukaryotic clusters of orthologous genes (KOGs) as follows. First, the KOG database covering 7 eukaryotic genomes (Tatusov et al. 2003) was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/COG/KOG/. Subsequently, each KOG was used as query to search clade-specific databases of protein sequences using the PSI-Blast program (Altschul et al. 1997; Schaffer et al. 2001). The searches were performed using command-line tools of the NCBI software development kit (Version 6.1, obtained from ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/ncbi.tar.gz). Using a Blast database for each of fungi, chromalveolates, and insects, the searches were performed using three iterations (switch –j 3) of PSI-Blast (blastpgp executable); for human and plant sequences, no iterations were used. For each KOG query, sequences with an E-value <10^{-9} were retained if they had a Blast hit score within 50% of the best hit for the species. In a further filtering step, reversed position-specific Blast search (Marchler-Bauer et al. 2007) was used to query each retained protein sequence against the Conserved Domain Database (CDD) of KOGs (psblast executable with default parameters). Only those sequences passed this filter for which the highest scoring KOG hit was the same as the KOG used in the initial PSI-Blast search and the second highest scoring KOG had less than 90% of the highest score. Sequences from the same genome that were thus assigned to the same KOG comprised paralogous sets.
"weaving" method to select a plausible orthologous set, which was then validated using a likelihood-based phylogeny comparison. The weaving method (Supplementary Material online) constructs a phylogeny of molecular sequences within a fixed species tree. The key technique consists of building a rooted evolutionary tree from sequences associated with two organismal lineages resulting from a speciation event, in the following manner. First, pairwise distances are computed from a multiple alignment of the sequences (the alignment is computed on the fly for each application of this technique.) Second, the tree is built by applying the Neighbor-Joining algorithm (Saitou and Nei 1987; Studier and Keppler 1988) to the distances. Inner nodes of the tree are subsequently classified as speciation or duplication nodes. Duplication nodes that have descendant speciation nodes are split so that paralogous gene lineages are identified for which duplication predates the speciation event. Only one sequence is kept as a representative from each gene lineage. This technique is applied to each bifurcation of the species tree by proceeding from the terminal taxa towards the root. The result of this weaving procedure is a set of putative orthologous lineages, which are presumably the result of gene duplication predating the root of the species tree.

The largest set of putative orthologs was elected from each set of paralogs, and a phylogeny was constructed using Neighbor-Joining. The resulting distance-based phylogeny was compared with the species tree. For this comparison, the PAML package (Yang 2007) was used to compute likelihood scores for protein sequence evolution along both phylogenies (with the software option of the WAG+Γ amino acid replacement model). A set of putative orthologs was considered valid if it contained representatives of at least 18 species and the log-likelihood score with the distance-based phylogeny was greater than the log-likelihood score with the species tree by at most 0.4 (this threshold was established by surveying the distribution of these scores across all KOGs).

Throughout the ortholog identification phase, sequences were aligned using MUSCLE (Edgar 2004) and distances were computed using the heuristic of Sonnhammer and Hollich (Sonnhammer and Hollich 2005) in conjunction with the VTML240 amino acid scoring matrix (Muller et al. 2002).

Orthologous Intron Sites

For each set of orthologous proteins, a multiple alignment was constructed using MUSCLE (Edgar 2004), the corresponding coding sequences were aligned using the protein alignment as the guide, and the intron sites were projected onto the alignment as described previously (Rogozin et al. 2003). Aligned intron-containing sites with identical phases were considered orthologous. Sites were propagated to further analysis by computationally inspecting sequence conservation around them. For each intron site within each sequence, the number of nongap amino acid positions had to be at least 4 on both the left- and right-hand sites to be categorized solid. If the number of solid positions at a site was at least 18, then it was included in the intron data set. In solid positions, 0 and 1 were used to encode absence and presence of the intron, respectively, whereas in nonsolid positions, and for missing sequences, an ambiguity character was used. The intron data set was compiled by concatenating the intron site information from all orthologous sets.

Likelihood-Based Analysis of Intron Evolution

The intron data set was analyzed in a likelihood framework described previously (Csuros 2005; Csuros et al. 2007). Briefly, the procedure is as follows. It is assumed that intron sites evolve independently under a Markov model (Steel 1994). The intron state (encoded by 0 and 1 for absence and presence) changes on each branch of the phylogeny according to the probabilities

\[
p_0 \to 0 (e) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \quad p_0 \to 1 (e) = \frac{\lambda}{\lambda + \mu} - \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}
\]

\[
p_1 \to 0 (e) = \frac{\mu}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \quad p_1 \to 1 (e) = \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}
\]

where \( \lambda \) denotes branch-specific intron gain rate, \( \mu \) denotes branch-specific loss rate, and \( t \) stands for branch length. These latter parameters were set by numerical optimization of the likelihood function, while taking into account a correction for missing intron sites (Felsenstein 1992). The intron density at an ancestral node was computed as an expected value conditioned on the observed data, by summing posterior probabilities (Csuros et al. 2007). The extent of intron gains and losses along individual branches are estimated analogously, using conditional expectations.

We experimented with rate variation models in which intron sites belong to discrete loss and gain rate categories. In rate variation models, each site category is defined by a pair of gain and loss rate-modifying factors \((\alpha, \beta)\) that apply to all branches of the tree such that loss and gain rates \( \alpha_1 \) and \( \beta_1 \) are plugged into the state transition probabilities of equation (1). We used the Bayesian Information Criterion (Schwarz 1978) to select the best rate variation model, which had two loss rates’ classes.

Confidence Intervals

Confidence interval (CIs) for estimates of ancestral intron density were obtained by using 1 000 bootstrap replicates. In each iteration, a new data set was generated by randomly selecting the same number of intron sites (independently and uniformly, with replacement). The likelihood of the new data set was maximized numerically to set gain and loss rates \( \alpha_1 \) and \( \beta_1 \), and the likelihood score \( L(\alpha_1, \beta_1) \) was plugged into the state transition probabilities of equation (1). The CIs were obtained by discarding the 25 largest and the 25 smallest values from the bootstrap estimates.

Results

Shared and Unique Intron Positions in Orthologous Genes of Chromalveolates and Other Eukaryotes

The data set analyzed here consisted of 394 orthologous gene sets from 23 eukaryotes, including 11 chromalveolates, where each set was represented in at least 18
species. The species were selected to combine the chromalveolates with complete annotated genome sequences available with a maximum representation of intron-rich outgroups. The data set contained 7030 intron-bearing sites in conserved, unambiguously aligned regions of the orthologous protein sequences (Materials and Methods).

A crucial observation is that introns are rarely found in the same position between distant chromalveolate species, with the exception of introns in *Plasmodium*, which often share positions with introns in *Theileria*, as reported previously (Roy and Penny 2006). Previous analyses have shown similar patterns of intron sharing at slightly lower levels, due to sparser taxonomic sampling (Rogozin et al. 2003; Nguyen et al. 2007; Roy and Penny 2007a). Surprisingly, in many cases, chromalveolate introns are more likely to share position with introns in orthologous genes of animals, fungi, or plants than with other chromalveolates (table 1). Thus, almost half of *Phytophthora* intron positions coincide with those in orthologous genes of animals, fungi, or plants. This pattern of intron sharing suggests that differential lineage-specific intron loss was a substantial, if not the primary, contributor to the observed differences in the exon–intron structure of orthologous genes among the chromalveolates.

Table 1

<table>
<thead>
<tr>
<th>Introns Shared between Taxa*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theileria</td>
</tr>
<tr>
<td>Theileria (692)</td>
</tr>
<tr>
<td>Plasmodium (195)</td>
</tr>
<tr>
<td>Ciliates (824)</td>
</tr>
<tr>
<td>Phytophthora (288)</td>
</tr>
</tbody>
</table>

* Numbers denote percentages, computed as the fraction of all introns from taxa in the row’s clade that coincide with at least one member of the column’s set. “AFV” column refers to all animals, fungi, plants, and green algae in the data set. The numbers of intron-bearing sites are shown in parentheses in the row headers.

Intron Gain and Loss Dynamics in Chromalveolate Lineages

The gain and loss of introns in chromalveolates were reconstructed using a likelihood framework that incorporated branch-specific intron loss and gain rates, as well as rate variation across sites embodied by two loss-rate categories where about one-fifth of modern intron sites lose introns at a 60% lower rate than the rest of the sites. The parameters of the rate categories were set by numerical optimization along with other model parameters. The number of rate categories was picked using a correction for model complexity to the likelihood score (Materials and Methods). The model imposes no constraints on the sequence of events occurring in the same site, that is, introns in a given position of an orthologous gene set can be lost and regained or gained independently in different lineages.

The reconstruction revealed a remarkable variation in intron loss and gain dynamics among chromalveolate lineages (fig. 1; the reconstructions for each of the individual set of orthologs are available at http://www.iro.umontreal.ca/~csuros/introns/Chroma23/). The exon–intron structure of orthologous genes has not changed much within the Apicomplexan genera, that is, individual species of *Theileria* and *Plasmodium* maintained the same intron density, with balanced gains and losses affecting 3–4% of their introns. This is in agreement with the recent findings of Roy and Hartl who demonstrated the stasis of gene structures within the *Plasmodium* genus (Roy and Hartl 2006). The branch leading to the *Theileria* ancestor, where ~20% of modern *Theileria* introns were gained, is characterized by a slight net loss, with losses outnumbering gains, approximately, 2-fold. Intron abundance was reduced more drastically in other alveolate branches, where losses outnumber gains 3- to 6-fold (the ciliate branch and the Alveolata–Apicomplexa branch), or even more than 20-fold (the Apicomplexa-*Plasmodium* branch), in agreement with the previous conclusions on the high prevalence of intron loss in Apicomplexa (Roy and Penny 2007a). The present reconstruction indicates that evolution of gene structure in heterokont lineages was also dominated by massive loss of introns (fig. 1).

The extensive intron loss is often accompanied by modest but nonnegligible intron gain. Among the chromalveolates, in the extreme case of the diatom *P. tricornutum*, these recent gains account for ~90% of the few introns present in the genes of this organism (fig. 2). A similar pattern has been detected and thoroughly discussed by Roy and Penny for the diatom *Thalassiosira pseudonana* (Roy and Penny 2007b). *Phytophthora* is estimated to have gained a comparable number of introns in the same time interval but underwent a less extreme reduction such that about 50% of the introns predate the chromalveolate ancestor (fig. 1). In alveolates, recent lineage-specific gains (on branches below the apicomplexan and ciliate ancestors) account for 20–30% of the extant introns in *Theileria, Plasmodium,* and *Tetrahymena* and ~46% of the extant introns in *Paramecium* (fig. 2).

High Intron Density in Deep Ancestors of Chromalveolates

We considered four alveolate lineages: *Plasmodium*, *Theileria*, *Paramecium*, and *Tetrahymena*. The key aspects of chromalveolate intron evolution are apparent in the pattern of intron sharing between these lineages (table 2). First, introns that are shared between alveolate and nonalveolate organisms most often appear in only one alveolate lineage (specifically, in 72% of the cases). Considering the relatively low level of parallel intron gain in the same position (estimated at <20% even for the most distant eukaryotes; Sverdlov et al. 2005; Carmel, Rogozin, et al. 2007), these shared introns were, most likely, present in the alveolate
ancestor, and so their presence in only a subset of the chromalveolate lineages attests to extensive, lineage-specific intron loss. Second, introns that are unique to chromalveolates exhibit an even more skewed distribution among lineages than introns that are conserved outside the supergroup. Indeed, introns that are shared with other eukaryotes are significantly more likely to appear in multiple chromalveolate lineages than supergroup-specific introns. Introns that appear in, at least, two chromalveolate lineages and are shared with nonchromalveolates are significantly more likely to appear in three or more lineages than chromalveolate-specific introns ($P < 6.2 \times 10^{-3}$, one-tailed Fisher’s exact test). This difference is likely to stem from a combination of the substantial between-sites variation of the intron loss rate (Carmel, Wolf, et al. 2007) and the relatively recent origin of some chromalveolate-specific introns (Roy and Penny 2007b).

The inferred intron densities in the ancestors of alveolates and chromalveolates are remarkably high (fig. 3). Specifically, the alveolate ancestor is estimated to have had a slightly greater intron density than humans, whereas the ancestor of the chromalveolate supergroup would have ~65% of that density. Strikingly, the estimated intron density of the alveolate ancestor is somewhat greater than the intron density in the plant (~91% of the human density) and opisthokont (~78% of the human density) ancestors estimated with the same method (fig. 3). The latter estimates were only slightly higher than those obtained previously with more constrained ML models (Nguyen et al. 2005; Roy and Gilbert 2005b; Carmel, Wolf, et al. 2007; Csuros et al. 2007).

Although modern alveolates have an intron density that is at least 60% lower than the current estimate for the ancestral form, ~72% of the inferred ancestral alveolate intron positions are shared by at least one extant, nonalveolate eukaryote. The uncertainty of the ancestral alveolate density estimate is relatively high (~19% coefficient of variation in bootstrap experiments), but even conservative estimates exceed two-thirds of the modern human intron density.
Table 2
Distribution of Shared Introns in Alveolate Lineages (Plasmodium, Theileria, Paramecium, and Tetrahymena)*

<table>
<thead>
<tr>
<th>Lineages</th>
<th>Shared</th>
<th>Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>299</td>
<td>999</td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>125</td>
</tr>
<tr>
<td>3 or 4</td>
<td>22</td>
<td>10</td>
</tr>
</tbody>
</table>

* The “Shared” column shows the distribution of the number of alveolate lineages in which a site is occupied by introns that are also shared with at least one nonalveolate lineage. The “Unique” column shows the same distribution for intron-bearing sites that are unique to alveolates.

Density ($P < 0.05$ in bootstrap experiments). Similar high estimates were obtained with the three possible branching orders for the supergroups and when different conservation criteria were applied for identification of homologous intron sites (Supplementary Material online). Furthermore, the possibility of a numerical optimization artifact was ruled out by examining the surface of the likelihood function (Supplementary Material online).

Discussion

The results of this study appear counterintuitive in that a very high intron density is confidently predicted for the ancestors of eukaryotic groups that (at least, so far) do not contain a single intron-rich species. This prediction became possible because a general and flexible ML method was applied to a diverse set of species. Adequate taxon sampling is particularly crucial for the reconstruction of evolution characterized by dramatic reduction of intron frequency in multiple lineages within a eukaryotic group such as the chromalveolates. In a case like this, evidence of a high ancestral intron density in the examined group can be obtained only through analysis of a diverse set of species because the genomes within the group share very few intron positions with each other but, collectively, retain many ancestral intron positions shared with some intron-rich genomes outside the group. The estimates of the rates of parallel intron gain obtained here are, generally, compatible with the previous estimates (Sverdlov et al. 2005; Carmel, Rogozin, et al. 2007) and indicate that the shared introns are predominately ancestral rather than those acquired in different lineages independently. The high level of intron conservation between chromalveolates and representatives of other eukaryotic supergroups, such as plants and animals, suggests that the intriguing possibility that, at least, some of the conserved introns retain ancestral functions throughout eukaryotic evolution. Indeed, introns often affect the expression of genes at several levels including mRNA export, stability, and translation efficiency (Le Hir et al. 2003). However, the loss of most of the ancestral introns in some of the chromalveolate lineages indicates that if such ancestral functions of introns exist, they are not unconditionally essential.

The trend toward an upward revision of inferred ancestral intron densities is seen in recent reconstructions (Carmel, Wolf, et al. 2007; Csuros et al. 2007) compared with previous analyses, even those performed with methods that might be prone to statistical bias (Roy and Gilbert 2005b). Conceivably, given that the current collection of eukaryotic genomes (improved as it is) hardly can be considered representative of each supergroup, even the estimates in this work are conservative.

Another important factor is the number of orthologous genes included in the data set as this determines the number of intron sites. Given the large loss and gain rate variation between lineages, several thousand sites are necessary to produce accurate rate estimates. In addition, with too few intron sites, models with rate variation cannot be used because there is insufficient information to partition the sites into rate categories. For instance, when only half of the intron sites contained in the present data set are analyzed, a constant-rate model has almost as much statistical support as a two-loss-rates’ model, and a model complexity penalty (e.g., the Bayesian Information Criterion) will
out of heterokonts, was apparently dominated by intron loss, presumably, following independent increases in the effective population size in each lineage. To a large extent, this extensive elimination of introns might have been mediated by retrotransposon activity as suggested by Roy and Penny (Roy and Penny 2007a). In future genome analyses, it would be of interest to investigate other correlates of intensive purifying selection in these lineages of chromalveolates, such as the extent of gene loss.

Conclusions

The results of this work indicate that ancestral forms in a eukaryotic supergroup that consists exclusively of unicellular and relatively intron-poor organisms were, in all likelihood, extremely intron rich — possibly, more so than modern multicellular eukaryotes with the most complex genomes. Given the extensive lineage-specific intron loss that apparently dominated the evolution of chromalveolates, this conclusion could be reached only by analyzing a large set of orthologous genes from a representative set of species. As shown here and elsewhere, the ancestors of plantae and unikonts are also estimated to have been intron rich although, paradoxically, somewhat less so than the chromalveolate and alveolate ancestors. For the remaining two eukaryotic supergroups that include only protists, Rhizaria and Excavates, there are currently no sufficiently intron-rich genomes to perform similar reconstructions. When such genomes become available, it will become possible to obtain a reasonably complete scenario of early evolution of eukaryotic gene structure.

Supplementary Material

Supplementary table S1 is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was supported by the Intramural Research Program of the NIH (National Library of Medicine, National Center for Biotechnology Information) and by a research grant from the National Sciences and Engineering Research Council of Canada.

Literature Cited


Aoife McLysaght, Associate Editor

Accepted January 30, 2008