BMC Evolutionary Biology

Research article

Patterns of intron gain and conservation in eukaryotic genes

Liran Carmel, Igor B Rogozin, Yuri I Wolf and Eugene V Koonin*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

Email: Liran Carmel - carmel@ncbi.nlm.nih.gov; Igor B Rogozin - rogozin@ncbi.nlm.nih.gov; Yuri I Wolf - wolf@ncbi.nlm.nih.gov; Eugene V Koonin* - koonin@ncbi.nlm.nih.gov

* Corresponding author

Published: 12 October 2007

BMC Evolutionary Biology 2007, **7**:192 doi:10.1186/1471-2148-7-192

Received: 18 June 2007
Accepted: 12 October 2007

This article is available from: http://www.biomedcentral.com/1471-2148/7/192

Abstract

Background: The presence of introns in protein-coding genes is a universal feature of eukaryotic genome organization, and the genes of multicellular eukaryotes, typically, contain multiple introns, a substantial fraction of which share position in distant taxa, such as plants and animals. Depending on the methods and data sets used, researchers have reached opposite conclusions on the causes of the high fraction of shared introns in orthologous genes from distant eukaryotes. Some studies conclude that shared intron positions reflect, almost entirely, a remarkable evolutionary conservation, whereas others attribute it to parallel gain of introns. To resolve these contradictions, it is crucial to analyze the evolution of introns by using a model that minimally relies on arbitrary assumptions.

Results: We developed a probabilistic model of evolution that allows for variability of intron gain and loss rates over branches of the phylogenetic tree, individual genes, and individual sites. Applying this model to an extended set of conserved eukaryotic genes, we find that parallel gain, on average, accounts for only ~8% of the shared intron positions. However, the distribution of parallel gains over the phylogenetic tree of eukaryotes is highly non-uniform. There are, practically, no parallel gains in closely related lineages, whereas for distant lineages, such as animals and plants, parallel gains appear to contribute up to 20% of the shared intron positions. In accord with these findings, we estimated that ancestral introns have a high probability to be retained in extant genomes, and conversely, that a substantial fraction of extant introns have retained their positions since the early stages of eukaryotic evolution. In addition, the density of sites that are available for intron insertion is estimated to be, approximately, one in seven basepairs.

Conclusion: We obtained robust estimates of the contribution of parallel gain to the observed sharing of intron positions between eukaryotic species separated by different evolutionary distances. The results indicate that, although the contribution of parallel gains varies across the phylogenetic tree, the high level of intron position sharing is due, primarily, to evolutionary conservation. Accordingly, numerous introns appear to persist in the same position over hundreds of millions of years of evolution. This is compatible with recent observations of a negative correlation between the rate of intron gain and coding sequence evolution rate of a gene, suggesting that at least some of the introns are functionally relevant.

Background

The presence of spliceosomal introns and the concurrent splicing machinery are one of the principal distinctive features of eukaryotic genomes [1-3]. Indeed, even early branching eukaryotes that were once suspected of being intronless have been shown in recent years to posses at least a few introns [4-7]. This suggests that the evolution of introns is tightly linked to the central aspects of the evolution of eukaryotes, and it even has been proposed that introns were the driving force behind the emergence of the eukaryotic nucleus and other features of the eukaryotic cell [8,9].

Thanks to their ubiquity and potential major role in the evolution of eukaryotes, intron evolution has drawn considerable attention [1-3,10]. It is generally accepted that introns are units of evolution such that their presence/absence pattern is a result of stochastic processes of loss and gain. The details of these processes, however, remain elusive. Only in recent years, with the accumulation of genomic data, evolution of introns has become amenable to a systematic, genome-wide analysis. However, different attempts to accomplish such a study led to incongruent conclusions regarding the prevalence, rates, and timing of intron loss and gain during the evolution of eukaryotes [11-18]. Apparently, these discrepancies are due, primarily, to incomplete underlying evolutionary models and biased estimation techniques [10].

Recently, we have obtained more conclusive results by developing a comprehensive probabilistic model of intron evolution, and by compiling a data set that is considerably larger than any one previously used [19,20]. The probabilistic models used so far to study intron evolution can be classified into two groups: branch-specific and gene-specific ones. The branch-specific models assume that the processes of intron gain and loss along a branch are determined only by the properties of the branch, regardless of the particular gene in question [17,18,21]. Conversely, gene-specific models assume that these processes are determined only by the particular gene, independent on the branch in question [12]. Obviously, in reality, the characteristics of intron gain and loss processes vary considerably both across genes and across branches. Thus, each of the models employed thus far seem to describe only one facet of a more complex reality. By contrast, our model allows for the variability of intron gain and loss characteristics with respect to both genes and branches such that any of the previously suggested models can be shown to be a special case of this comprehensive model [20]. Moreover, this model is even more realistic in that it also includes rate variability between sites, with respect to both intron gain and intron loss. In order to estimate the model parameters, we devised an expectation-maximization (EM) algorithm that can also be used to reconstruct ancestral states [22]. Combining this algorithm with the profile likelihood technique allows one, in addition to all of the above, to compute confidence intervals of the model parameters.

Here, we describe, in detail, the developed model of intron evolution and the derivation of an improved version of the EM algorithm used to estimate its parameters. Previously, we have applied this comprehensive model to a set of 391 conserved genes from 19 eukaryotic species, to investigate evolutionary trends in gene structure both at the lineage level [20] and at the gene level [19]. The results of this analysis suggest that introns invaded eukaryotic genomes at early stages of eukaryogenesis in a nearly neutral process. At early times, during periods of major transitions in the eukaryotic evolution that led to population bottlenecks, these introns seem to have vastly proliferated in the ancient genomes. Gradually, a considerable fraction of the introns appear to become involved in various cellular functions, mostly, regulation of gene expression [19].

In the present work, we focus on the process of intron gain, and address the causes of the high level of intron position conservation between eukaryotic taxa. It had been already noticed that many intron positions are shared between distant eukaryotic taxa [11,13]. For example, plants and animals share up to 25% of the intron positions [13]. However, these findings can be explained by either remarkable conservation of ancient introns or by parallel, independent, intron gains in the same positions, or (perhaps, most likely) by a combination of both these factors. The previous analyses that have attempted to quantify the relative contributions of evolutionary conservation and parallel gain in intron position sharing have differed widely, with estimates of the extent of parallel gain ranging from nearly 0% to nearly 100% [11-13,18,23]. Using our probabilistic model, we developed a rigorous measure for assessing the amount of parallel gain of introns. We found that, overall, parallel gain is responsible for ~8% of the shared intron positions, with the rest due to shared ancestry. However, we also demonstrate substantial heterogeneity in the extent of parallel gain, with almost none in closely related lineages, but up to ~20% in distant ones, such as plants and unikonts. On the whole, these results support the notion that intron positions are highly conserved during evolution.

Results and discussion

Notation

The primary input component in the study of intron evolution consists of G sets of aligned protein-coding sequences of orthologous genes from S species. Each nucleotide in these alignments is substituted by 0 or 1 depending on whether or not an intron is present following the respective position. We allow for missing data by using a third symbol (*) to indicate lack of knowledge about the presence or absence of an intron. Consequently, every site in the alignments, called *pattern*, is a vector of length S over the alphabet (0,1,*) and is denoted by ω. Let Ω be the total number of unique patterns in the entire set of G alignments, and let n_{gp} be the count of the number of times pattern ω_p (p = 1, ..., Ω) is found in the multiple alignment of gene g. Assuming that the sites evolve independently, the set $M_g = (n_{g1}, ..., n_{g\Omega})$ fully characterizes the multiple alignment of the g th gene.

Let T be a rooted, bifurcating phylogenetic tree with S leaves (terminal nodes), describing the evolutionary relationships between the S species above. The total number of nodes in T is $N = 2S - 1$, and we index them by t = 0,1, ..., N - 1, with the convention that zero is the root node. The state of node t is described by the variable q_t, which can take the values 0 and 1 (and * in leaves). We use V_t for the set of all leaves such that node t is among their ancestors. The entire collection of leaves is, obviously, V_0. The parent node of t is denoted P(t). We use the notations q_t^P and V_t^P for $q_{\mathrm{P}(t)}$ and $V_{\mathrm{P}(t)}$, respectively. Analogously, the two direct descendents of the node t are denoted L(t) and R(t), and we use the notations q_t^L, q_t^R, V_t^L, and V_t^R for $q_{\mathrm{L}(t)}$, $q_{\mathrm{R}(t)}$, $V_{\mathrm{L}(t)}$, and $V_{\mathrm{R}(t)}$, respectively. The branches are indexed by the node into which they are leading, and Δ_t denotes the length (in time units) of the t th branch. Hereinafter we assume that the tree topology, as well as all the branch lengths Δ_1, ..., Δ_{N-1}, are known.

The probabilistic model

A bifurcating phylogenetic tree can be viewed as a graphical model depicting the probabilistic model

$$\Pr(q_0)\prod_{t=1}^{N-1}\Pr(q_t \mid q_t^P). \qquad (1)$$

We use the notation $\pi_i = \Pr(q_0 = i)$ to describe the prior probability of the root, and $T_{ij}(g, t) = \Pr(q_t = j \mid q_t^P = i, g)$ to describe the transition probability for gene g along branch t. In our model, we assume that this transition probability depends on both the gene and the branch, and that it takes the explicit form

$$T(g,t) = \begin{pmatrix} 1-\xi_t(1-e^{-\eta_g\Delta_t}) & \xi_t(1-e^{-\eta_g\Delta_t}) \\ 1-(1-\phi_t)e^{-\theta_g\Delta_t} & (1-\phi_t)e^{-\theta_g\Delta_t} \end{pmatrix}. \qquad (2)$$

Here, η_g and θ_g are non-negative parameters which determine, respectively, the intron gain and loss rates per site for gene g. That is, along branch t the gene's contribution to intron gain and retention probabilities per site is $1-e^{-\eta_g\Delta_t}$ and $e^{-\theta_g\Delta_t}$, respectively. We assume that each branch is characterized by an intrinsic *branch-specific intron gain coefficient*, ξ_t, as well as an intrinsic *branch-specific intron loss coefficient*, ϕ_t, both of which are bounded, $0 \le \xi_t, \phi_t \le 1$.

In other fields of molecular evolution, it had been long realized that the precision of the analysis significantly improves if one allows for rate variability across sites [24-26]. Typically, such rate variability is modeled by introducing a *rate variable*, r, which scales, for each site, the time units of the phylogenetic tree, $\Delta_t \leftarrow r \cdot \Delta_t$. This rate variable is a random variable that is distributed according to a distribution function with non-negative domain and unit mean, typically, the unit-mean gamma distribution. The rate variability captures rate variations among sites of the same gene. Specifically, there are fast-evolving sites ($r >> 1$), as well as slow-evolving ones ($r << 1$). In our model of intron evolution, we extend this idea by assuming that the gain and loss processes are subject to rate variability, independently of each other. Hence, a site can have any combination of gain and loss rates, for example, it can be fast to gain introns but slow to lose them. To implement this approach, we use two independent rate variables, r^η and r^θ, that are used to scale, for each site, the gene-specific gain rate, $\eta_g \leftarrow r^\eta \cdot \eta_g$, and the gene-specific loss rate, $\theta_g \leftarrow r^\theta \cdot \theta_g$, respectively. We further assume that the distributions of these rate variables are independent of the genes, and are explicitly given by

$$\begin{aligned} r^\eta &\sim \nu\delta(\eta) + (1-\nu)\Gamma(\eta;\lambda_\eta) \\ r^\theta &\sim \Gamma(\theta;\lambda_\theta). \end{aligned} \qquad (3)$$

Here, $\Gamma(x; \lambda)$ is the unit-mean gamma distribution of the variable x with the shape parameter λ, $\delta(x)$ is the Dirac delta-function, and ν is the fraction of sites that are assumed to have zero gain rate. The existence of these *zero sites* reflects the notion that introns cannot be gained at any location within genes, but rather are preferentially inserted at specific locations, contingent on particular sequence motifs known as proto-splice sites [27-29], the density of other introns in the neighborhood, the chromatin exposure, and more. According to this interpretation, 1 - ν measures the density of potential intron insertion sites. Importantly, using the same value of ν for the entire tree does not mean that the proto-splice sites are constant throughout the evolution, or are identical for different lineages. It only means that, on the average, the fraction of potential insertion sites, whatever is their concrete nature, is similar across the lineages throughout the course of eukaryotic evolution. The incorporation of such invariant sites in a rate variability model appears natural for intron evolution and has proved beneficial also in other fields of molecular evolution [30-32]. By contrast, intron loss does not have an invariant counterpart because the assumption is that, once an intron is gained, it can always be lost. Therefore, the loss rate variable is assumed to be distributed according to a gamma distribution, which is by far the most popular distribution for describing rate variability [24,33].

In practice, the rate distributions in (3) are rendered discrete [34]. We assume that the gain rate variable can take K_η discrete values $r_1^\eta = 0, r_2^\eta, \ldots, r_{K_\eta}^\eta$ with probabilities $f_1^\eta = \nu, f_2^\eta \ldots, f_{K_\eta}^\eta$ such that $\sum_{k=1}^{K_\eta} f_k^\eta = 1$. Analogously, we assume that the loss rate variable can take K_θ discrete values $r_1^\theta, \ldots, r_{K_\theta}^\theta$ with probabilities $f_1^\theta, \ldots, f_{K_\theta}^\theta$ such that $\sum_{k=1}^{K_\theta} f_k^\theta = 1$. For a particular gain rate value r_k^η, we denote the actual gain rate $r_k^\eta \cdot \eta_g$ by η_{kg}. Similarly, for a particular loss rate value r_k^θ, we denote the actual loss rate $r_k^\theta \cdot \theta_g$ by θ_{kg}.

For notational clarity, we aggregate the model's parameters into a small number of sets. To this end, let $\Xi_t = \{\xi_t, \phi_t\}$ be the set of parameters that are specific to branch t, and let $\Xi = (\Xi_1, ..., \Xi_{N-1})$ be the set of all branch-specific parameters. Similarly, let $\Psi_g = (\eta_g, \theta_g)$ be the set of parameters that are specific for gene g, and let $\Psi = (\Psi_1, ..., \Psi_G)$ be the set of all gene-specific parameters. Additionally, we denote by $\Lambda = (\pi_0, \nu, \lambda_\eta, \lambda_\theta)$ the "global" parameters that determine the rate variability and the prior probability of an intron absent in the root. When the distinction between the different sets of parameters is irrelevant, we shall use $\Theta = (\Xi, \Psi, \Lambda)$ as the set of all the model's parameters. We achieve further succinctness in notations by denoting the actual gene-specific rate values for particular values r_k^η and $r_{k'}^\theta$ of the rate variables as $\Psi_{kk'g} = (\eta_{kg}, \theta_{k'g})$.

The Expectation-Maximization algorithm

We estimate the parameters of the model using maximum likelihood. As the probability model includes observed random variables (state of the tree leaves) as well as hidden random variables (state of internal nodes in the tree and value of actual rate variables), the expectation-maximization algorithm is a natural tool to use [35]. This is a hill climbing iterative algorithm that requires two steps in each iteration – the expectation (E) step followed by the maximization (M) step. The details of the algorithm are given under Methods.

The total number of parameters in the model is $2G + 4S$, where G is the number of genes and S is the number of species. For data sets in the hundreds of genes, this sums up to >1000 parameters. For infinite data, maximum likelihood estimators are known to be nonbiased and efficient. However, as each gene typically goes through a small number of intron-related events during its lifetime, the information content of our data is limited, and cannot straightforwardly support the estimation of such a large number of parameter. To overcome this, we adopt a two-phase approach in the data analysis. In the first phase, that we denoted *homogeneous evolution*, it is assumed that all the genes have identical intron gain and loss rates, formally, that $\theta_g = \theta_0$ and $\eta_g = \eta_0$ for any gene g. In the second phase, denoted *heterogeneous evolution*, all the global and branch-specific parameters are fixed, which allows for estimation of the gene-specific parameters θ_g and η_g that can now take different values for different genes. Except for the rate variability within genes, the model of evolution under the homogeneous phase resembles the branch-specific models, and consequently the EM algorithm used in this part has similar structure to the EM algorithm developed by Nguyen *et al.* [18,36].

Using simulations, we showed that this approach yields highly accurate evolutionary reconstructions: a relative error of 1%, 3%, and 11% in estimating the number of introns in internal nodes, the number of loss events along each branch, and the number of gain events along each branch, respectively [20]. The current analysis is pattern-centered rather than gene-centered and thus we have used the results of the homogeneous phase only. While slightly less accurate than the heterogeneous counterparts, the reconstructions after the homogeneous phase are still highly accurate: a relative error of 2%, 4%, and 12% in estimating the number of introns in internal nodes, the number of losses along each branch, and the number of gains along each branch, respectively [20]. Also notable is that, apart from the fraction of zero sites ν, the rate variability within genes can be ignored without having any significant effect on the results (Ref. [20] and Additional file 1), indicating that those sites that are capable of gaining introns do so at comparable rates.

The EM algorithm was applied to the set of 391 orthologous genes from 19 eukaryotic species ([20] and see Methods), under the homogeneous evolution assumption, and the profile likelihood technique was used to estimate the 95% confidence interval for each parameter (Additional file 1). In computing the confidence intervals for a particular parameter, we allow all other parameters to vary. Sometimes, different combinations of parameters yield similar likelihood values, an observation that has already been made formal for simpler models [18]. This occasionally leads to wide single-parameter confidence intervals, especially for the intron gain and loss rates of deep nodes. Importantly, this does not reflect similarly large errors in the ensuing inference, as was demonstrated by an exhaustive simulation study [19,20].

One in 7 nucleotides is a potential intron insertion site

The maximum-likelihood estimate for the fraction of zero sites is $\nu = 0.862$, suggesting a potential intron insertion site every 7 nucleotides. Taking into account the 95% confidence interval for ν, [0.631, 0.928], the density of potential insertion sites is estimated to be 1 site per 3-14 nucleotides. Based on a smaller data set, Nguyen *et al.* [18] computed a 95% confidence interval of 1 potential insertion site in 9–14 nucleotides. Although this estimate falls within our confidence interval, our results suggest, generally, a denser population of potential insertion sites. The present estimate is also compatible with the high density of intron insertion sites (one site per 9.7 nucleotides) observed for some large protein families, e.g., small G proteins [37]. The problem of estimating the number of zero sites is of fundamental importance in maximum likelihood techniques, so both Csuros [17] and Nguyen *et al.* [18] developed different heuristic procedure to handle it. We took the alternative and, arguably, more natural approach of integrating this estimate into the model as part of the gain rate variability distribution (see above) such that additional, *ad hoc* computations are not required.

One should be cautious about identifying potential intron insertion sites with proto-splice sites. More realistically, the density of potential insertion sites reflects the overall average (across species) of the impact of multiple factors such as differential tendencies to be inserted into different proto-splice sites, local densities of pre-existing introns, and degree of chromatin exposure. Furthermore, this density estimation strongly depends on the specific data set as the parameter ν "absorbs" the information on zero sites. Thus, if intronless genes are added to the data, only this parameter will be heavily affected.

With the total number of sites $N_0 = 289{,}902$ our calculations yield 39,962 sites that can gain introns (95% confidence interval 20,891 to 106,901). Accordingly, the 5,755 sites that are actually occupied by introns in the genes analyzed here comprise 14.4% of all sites that can potentially gain introns (95% confidence interval 5.4% to 27.5%). Thus, even when data from 19 species are combined, the density of the sites actually occupied by introns is still far below the density of sites capable of hosting introns, which emphasizes the inadequacy of any analysis that considers only those sites in which introns are, actually, observed.

The number of shared intron positions is much greater than expected by chance

Numerous intron positions are shared even between taxa that had diverged during the early stages of the eukaryotic evolution [11,13]. To quantify this conservation more precisely and to estimate how surprising it is, compared to the random expectation, we propose the following measure: let N_i and N_j be the number of intron positions in species i and j, respectively; let S_{ij} be the number of intron positions shared by the two species; and let N be the total number of sites capable of gaining introns. Then, approximately, we would expect that N_iN_j/N positions will be shared between the two species by chance alone. We define the *conservation level*

$$c_{ij} = \log\left(\frac{NS_{ij}}{N_iN_j}\right)$$

as the log-ratio between the observed number and the number expected by chance. Positive values designate that S_{ij} exceeds random expectation, whereas negative values indicate that S_{ij} is below expectation. Even if we take for N its lower 95% confidence interval bound of 20,891 (see the preceding section), almost all pairs show a positive value, namely, a greater than random number of shared positions (Fig. 1). The only exception is the pair *S. cerevisiae – P. falciparum* (the two most intron-poor species in our set) that do not have even a single shared position.

The number of positions shared by chance between two species is a random variable distributed, approximately, according to a binomial distribution with the probability of success $p = N_iN_j/N^2$, and N experiments. Therefore, it is easy to associate a p-value with any observed number of shared positions, S_{ij} measuring how improbable it is to obtain by chance this value or a greater one. An overall significance level of 0.05 is equivalent to a Bonferoni-corrected significance level of 0.0003 (overly conservative as we assume that all pairs are independent). The calculations indicate that only 20 species pairs out of 171, all involving the intron-poor *S. cerevisiae* or one of the apicomplexans, had a number of shared positions that was indistinguishable from the random expectation; all other pairs had a significant excess of shared intron positions (Additional file 2).

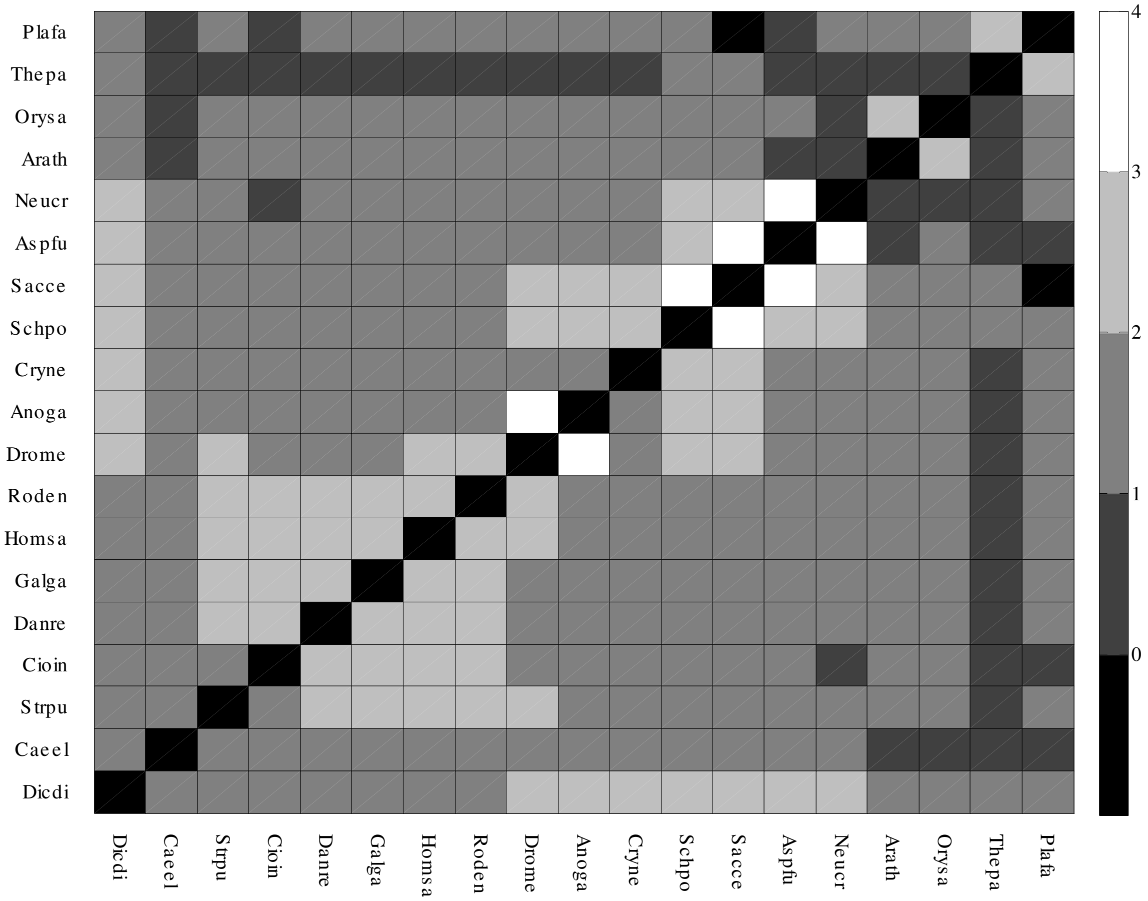

Figure 1

Conservation of intron positions between eukaryotic species. The scale to the right shows the pairwise conservation level of intron positions, measured as the log-ratio of the observed number of shared positions to the number expected by chance (see text). The expected value of this ratio is 0, so the positive values indicate an excess of shared intron positions, and the negative values indicate an unexpected deficit of such positions. Species and lineage abbreviations: Anoga (*Anopheles gambiae*), Arath (*Arabidopsis thaliana*), Aspfu (*Aspergillus fumigatus*), Caeel (*Caenorhabditis elegans*), Cioin (*Ciona intestinalis*), Cryne (*Cryptococcus neoformans*), Danre (*Danio rerio*), Dicdi (*Dictyostelium discoideum*), Drome (*Drosophila melanogaster*), Galga (*Gallus gallus*), Homsa (*Homo sapiens*), Neucr (*Neurospora crassa*), Orysa (*Oryza sativa*), Plafa (*Plasmodium falciparum*), Sacce (*Saccharomyces cerevisiae*), Schpo (*Schizosaccharomyces pombe*), Strpu (*Strongylocentrotus purpuratus*), Thepa (Theileria parva), roden (*Mus musculus* and *Rattus norvegicus* combined).

Twelve out of thirteen shared intron positions reflect common ancestry

The significant excess of shared intron positions over the random expectation can be explained by one of two factors or a combination thereof. The first explanation is that this observation reflects genuine evolutionary conservation; the implication is that, once an intron is gained, it is hard to lose, perhaps, due to functional importance of introns or because the deletion event itself is destructive. The second explanation is that different lineages gain introns at the same position, independently. The chance of such parallel gain is not necessarily as low as it seems at first sight because introns are, apparently, preferentially inserted into proto-splice sites (see above). Accordingly, regions of high sequence conservation might gain introns at exactly the same position.

The two explanations have opposing impacts on our understanding of the evolution of eukaryotic genes. If introns are persistent, many must have been gained early in the eukaryotic evolution, and the later evolution involved multiple losses [16,38]. By contrast, if parallel gain is the dominant mechanism, many of the extant introns are evolutionarily young, and the eukaryotic evolution involves, primarily, multiple introns gains [12].

In order to estimate the extent of parallel gain, we examined all the sites that host introns in at least two species. Assuming that these sites are not invariant, the probability was estimated that the corresponding pattern had arisen from parallel gain, by computing the probability of the last common ancestor of the intron-bearing species to lack an intron. This approach assumes that, if the last common ancestor had an intron in a particular position, no parallel gain occurred, i.e., highly unlikely scenarios involving at least two gains and one loss at the same site are neglected. The probabilities of parallel gain for all the patterns observed in our data set are summarized in Additional file 3. Overall, the data set harbors 3176 sites with shared intron positions (741 unique patterns) out of which 317 positions (10%) are expected to result from parallel intron gain. As inferences involving the root of the tree are prone to significantly elevated standard errors [20], the calculation was repeated using only patterns that do not have the root as the last common ancestor of all intron-bearing species. This calculation yielded 2913 sites with shared intron positions (568 unique patterns) of which 229 (~7.9%) are expected to be due to parallel intron gain. Each of these calculations provides a rough estimate of the level of errors introduced into some of the recent studies that explicitly excluded the possibility of parallel gain [13,16,21].

Importantly, however, the contributions of parallel gains to the emergence of different patterns of intron sharing are widely different; in particular, some rare patterns are explained (almost) entirely by parallel gain and do not reflect evolutionary conservation (Additional file 3). For example, for the single site that harbors introns only in humans and *N. crassa*, the probability is >0.99 that it results from parallel gain. Considering somewhat more frequent patterns, 11 sites harbor introns in *C. intestinalis*, *A. thaliana* and *O. saliva*, with the probability of parallel gain ~0.875, and another 12 sites harbor introns only in *C. elegans* and *C. intestinalis*, with the probability of parallel gain ~0.8.

Generally, the distribution of parallel gains in comparisons of specific clades is, obviously, more informative than overall counting (Table 1). To obtain this information, for each internal node *t* (excluding the root of the tree), all patterns that have 1s in the two sub-clades stemming from *t*, V_t^L and V_t^R were tallied, and the probability that *t* is in state zero was computed. By this analysis, in which 728 unique patterns were included, we found that nearly 20% of the shared intron positions between plants and unikonts, thought to have diverged more than a billion and a half years ago [39], are due to parallel gain. In fungi and metazoa, diverged ~1.4 billion years ago, >10% of the shared positions are estimated to derive from parallel gains. In contrast, many recently diverged clades show almost no parallel gain (e.g., humans versus rodents, birds versus mammals, *Aspergillus* versus *Neurospora*, and flies versus mosquitoes). Table 2 lists all patterns for which the estimated contribution to parallel gain was greater than two sites. The estimated total number of sites with parallel gain is 248. Out of the 728 unique patterns,

Table 1: The estimated number of parallel gains on the branches stemming out of each of the internal nodes in the phylogenetic tree (excluding the root; see Additional file 4)

internal node	Subclade_1	subclade_2	total number of shared sites	total number of parallel gains [95% confidence inrterval]	% parallel gains [95% confidence inrterval]
AME	Unikonts	Magnoliophyta	630	122.8 [38.5 – 229.3]	19.5 [6.1 – 36.4]
Unikonts	Dicdi	Opisthokonts	212	4.9 [1.6 – 15.3]	2.3 [0.7 – 7.2]
Opisthokonts	Metazoa	Fungi	606	70.7 [23.0 – 123.1]	11.7 [3.8 – 20.3]
Metazoa	Caeel	Coelomata	374	24.0 [7.8 – 38.7]	6.4 [2.1 – 10.4]
Coelomata	Deuterostomia	Diptera	350	7.0 [2.4 – 11.5]	2.0 [0.7 – 3.3]
Deuterostomia	Strpu	Chordata	1395	4.7 [1.6 – 8.2]	0.3 [0.1 – 0.6]
Diptera	Drome	Anoga	192	0.0 [0.0 – 0.1]	0.0 [0.0-0.0]
Fungi	Cryne	Ascomycota	223	7.0 [2.4 – 13.7]	3.1 [1.1 – 6.1]
Ascomycota	Schpo	ScAfNc	82	0.3 [0.0 – 0.5]	0.3 [0.0 – 0.7]
ScAfNc	Sacce	Pezizomycotina	5	0.0 [0.0 – 0.1]	0.5 [0.1 – 1.2]
Magnoliophyta	Arath	Orysa	1337	0.2 [0.1 – 0.5]	0.0 [0.0-0.0]
Chordata	Cioin	Vertebrata	822	2.5 [0.9 – 4.2]	0.3 [0.1 – 0.5]
Vertebrata	Danre	Amniota	1701	0.3 [0.1 – 0.5]	0.0 [0.0-0.0]
Apicomplexa	Thepa	Plafa	113	3.9 [0.4 – 8.0]	3.4 [0.3 – 7.1]
Pezizomycotina	Aspfu	Neucr	221	0.1 [0.0 – 0.3]	0.1 [0.0 – 0.1]
Amniota	Galga	Mammals	1659	0.1 [0.0 – 0.1]	0.0 [0.0-0.0]
Mammals	Homsa	Roden	1448	0.0 [0.0-0.0]	0.0 [0.0-0.0]

Species abbreviations are as in Fig. 1. AME stands for the last common Ancestor of Multicellular Eukaryotes

Table 2: Patterns that are estimated to contribute more than two sites to the total count of parallel gains

Pattern	total number of patterns	Estimated number of parallel gains
Caeel, Arath, Orysa	20	14.9
Cryne, Arath, Orysa	28	12.6
Cioin, Arath, Orysa	11	9.6
Caeel, Cioin	12	9.6
Strpu, Danre, Galga, Homsa, Arath, Orysa, Roden	39	7.5
Strpu, Cioin, Danre, Galga, Homsa, Arath, Orysa, Roden	32	6.1
Strpu, Cryne	16	5.8
Strpu, Arath, Orysa	13	5.7
Danre, Galga, Homsa, Arath, Orysa, Roden	13	5.3
Cioin, Cryne	5	4.5
Caeel, Cryne	6	4.3
Strpu, Danre, Galga, Homsa, Cryne, Roden	32	4.1
Thepa, Plafa	65	3.3
Caeel, Thepa	23	3.0
Caeel, Strpu, Cioin, Danre, Galga, Homsa, Arath, Orysa, Roden	4	2.9
Aspfu, Arath	3	2.6
Dicdi, Strpu, Danre, Galga, Homsa, Aspfu, Neucr, Roden	4	2.6
Caeel, Strpu, Danre, Galga, Homsa, Drome, Anoga, Arath, Orysa, Roden	7	2.3
Caeel, Strpu, Danre, Galga, Homsa, Schpo	14	2.3
Arath, Orysa, Thepa, Plafa	3	2.3
Cioin, Danre, Galga, Homsa, Arath, Orysa, Roden	7	2.2
Danre, Galga, Homsa, Drome, Anoga	3	2.0

Species and lineages abbreviations are as in Figure 1.

the 22 in this list (3%) account for 116 parallel gains, i.e., ~47% of the total estimated number.

Given that most of the other studies report overall estimates on a smaller (8 species) data set [13], a direct comparison of their results with the present ones is not feasible due to the apparent non-uniformity in the extent of parallel gain in different parts of the tree. Nevertheless, an overall parallel gain of ~10% has been computed also for the smaller data set. Rogozin *et al.* [13] used simulations to assess the extent of parallel gain and showed that, under various assumptions regarding the density of proto-splice sites, parallel gains could be responsible to ~2–40% of the shared intron positions. Accounting more accurately for the likely density of proto-splice sites, Sverdlov et al. estimated the contribution of parallel gains to the observed sharing of intron positions to be in the range of 5–10% [23]. An even higher estimate, 18.5% parallel gains, was obtained by Nguyen et al. using a branch-specific maximum-likelihood model [18].

Intron retention

We conclude, therefore, that a substantial majority of the shared intron positions are due to evolutionary conservation, hence intron positions tend to be retained for long times. To further validate this conclusion, we explicitly computed the probability of an intron to survive along given paths in the phylogenetic tree. To this end, let *B* denote the set of branches that comprise a path in the tree. Then, intron retention probability along this path is

$$P_B = \sum_{k'=1}^{K_\theta} f_{k'}^{\theta} \prod_{t \in B} (1-\phi_t) e^{-\theta_{k'} g \Delta_t},$$

where $(1-\phi_t)e^{-\theta_{k'} g \Delta_t}$ is the retention probability along branch *t*, and $f_{k'}^{\theta}$ is the probability of being at loss-rate category *k*'.

In accord with the above conclusions on the high level of evolutionary conservation, this probability is, typically, high, even for very long paths in the phylogenetic tree (Table 3). For example, an intron present in the last common ancestor of the metazoa has a probability of 0.83 to be retained in humans whereas an intron present in the last common ancestor of multicellular life (AME) has a probability of 0.57 to be retained in extant plants.

A complementary approach involves the computation of the probability of extant introns to be of ancient origin. To calculate this probability, we assumed that a site is known to host an intron in one species, and that no information is available on this site in other species (that is, their state is *). Then, we used our model to compute the probability that this intron was present in any of the ancestors of that species (Table 4 and Fig. 2). Clearly, these probabilities decay quite slowly with evolutionary time. For instance,

Table 3: Probabilities of an intron to be retained along selected paths in the phylogenetic tree

If an intron was present in	chances are	that it would be present in	95% confidence interval
AME	0.63	Homsa	[0.57 – 0.68]
Metazoa	0.83	Homsa	[0.79 – 0.84]
Deuterostomia	0.86	Homsa	[0.85 – 0.88]
Vertebrata	0.95	Homsa	[0.94 – 0.96]
Mammals	0.96	Homsa	[0.95 – 0.97]
Fungi	0.01	Sacce	[0.00 – 0.01]
Fungi	0.26	Aspfu	[0.24 – 0.27]
Apicomplexa	0.26	Plafa	[0.20 – 0.33]
Apicomplexa	0.69	Thepa	[0.57 – 0.80]
AME	0.57	Arath	[0.50 – 0.64]
AME	0.57	Orysa	[0.50 – 0.65]

Species and lineage abbreviations are as in Fig. 1.

an intron in *C. elegans*, *H. sapiens* and *D. melanogaster* has a probability of 0.44, 0.69, and 0.68, respectively, to have been present in the last common ancestor of metazoa.

Conclusion

Despite the extensive attention given to the evolution of eukaryotic gene structure over the last three decades, the fundamental characteristics of this process remain controversial. In particular, depending on the methods and data sets used, different researchers have reached opposite conclusions on the causes of the remarkably high fraction of shared introns in orthologous genes from distant eukaryotic species. Some attribute it (almost) entirely to a remarkable evolutionary conservation of intron positions and others, largely, to parallel gain of introns. To resolve these contradictions, it is important to analyze the evolution of introns by using a probabilistic model that minimally relies on arbitrary assumptions. To this end, we developed a model that allows for variability of intron gain and loss rates over branches of the phylogenetic tree, individual genes, and individual sites. Applying this model to an extended set of conserved eukaryotic genes, we found that parallel gain, on average, accounts for only ~8% of the shared intron positions. However, the distribution of parallel gains over the phylogenetic tree of eukaryotes is highly non-uniform such that there are, practically, no parallel gains in closely related lineages, whereas for distant lineages, such as animals and plants, parallel gains might have contributed up to 20% of the shared intron positions. Given the distinctly non-uniform distribution of the inferred gain events over the phylogenetic tree of eukaryotes [20], most of the recently diverged

Table 4: Probability of an intron in extant species to be inherited from an ancestral node

intron is present in	Probability	that it is also present in	95% confidence interval
Dicdi	0.71	AME	[0.46 – 0.92]
Caeel	0.19	AME	[0.10 – 0.74]
Strpu	0.31	AME	[0.17 – 0.79]
Cioin	0.22	AME	[0.12 – 0.75]
Danre	0.29	AME	[0.16 – 0.78]
Galga	0.29	AME	[0.16 – 0.78]
Homsa	0.30	AME	[0.17 – 0.78]
Roden	0.30	AME	[0.17 – 0.78]
Drome	0.30	AME	[0.17 – 0.78]
Anoga	0.28	AME	[0.16 – 0.78]
Cryne	0.29	AME	[0.17 – 0.78]
Schpo	0.39	AME	[0.24 – 0.82]
Sacce	0.27	AME	[0.16 – 0.77]
Aspfu	0.27	AME	[0.15 – 0.77]
Neucr	0.21	AME	[0.11 – 0.75]
Arath	0.31	AME	[0.18 – 0.79]
Orysa	0.30	AME	[0.17 – 0.78]
Thepa	0.30	Apicomplexa	[0.10 – 0.77]
Plafa	0.46	Apicomplexa	[0.21 – 0.81]

Confidence intervals are wide due to the large amount of uncertainty associated with the input patterns (each pattern has 18 out of 19 sites with the presence/absence of the intron marked as unknown). Species and lineages abbreviations are as in Table 1.

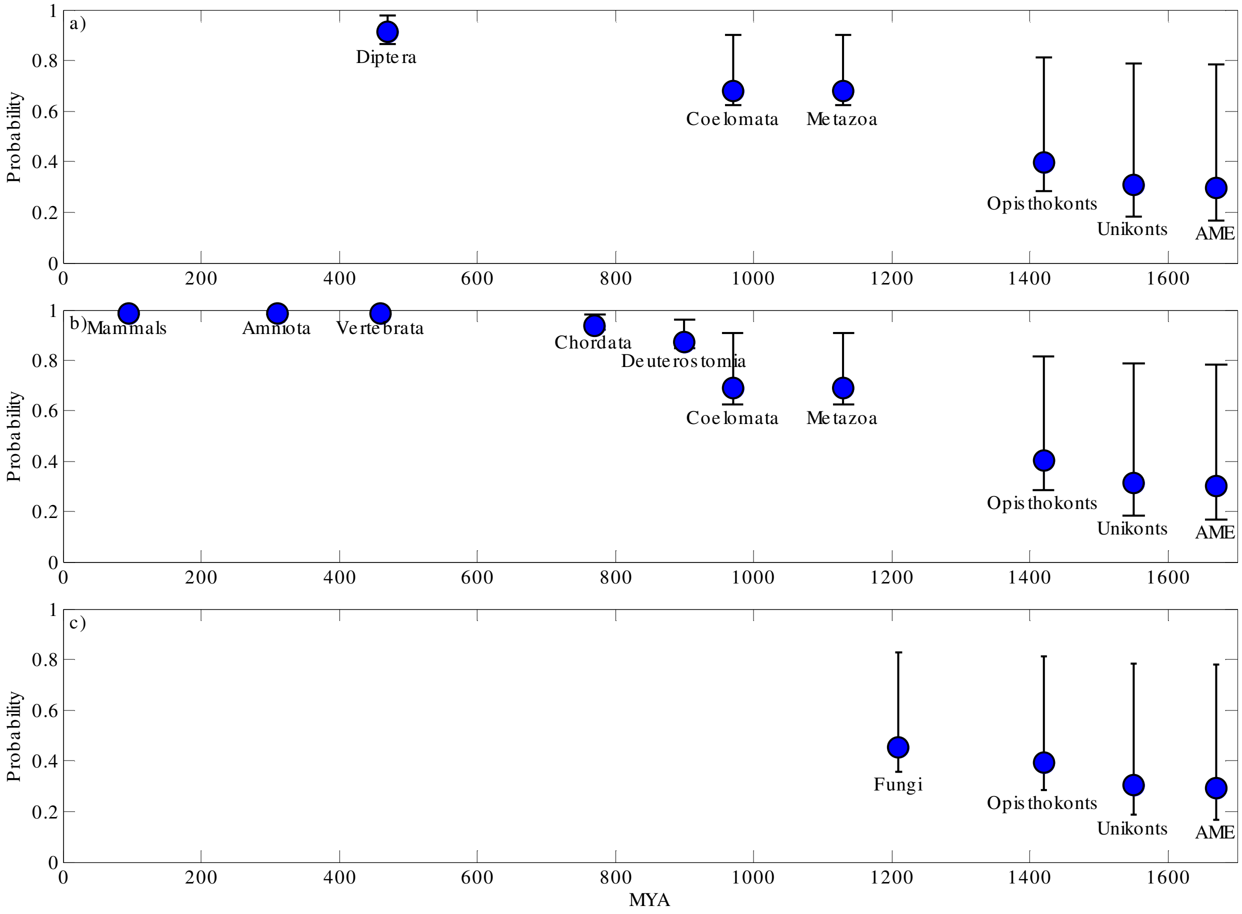

Figure 2
The probability of an intron in extant species to be present in ancient ancestors. **a)** an intron in *D. melanogaster*, **b)** an intron in *H. sapiens*, **c)** an intron in *C. neoformans*. AME stands for the last common ancestor of multicellular eukaryotes.

lineages have amassed very few gains during their separate evolution; by contrast, deeply diverged lineages, such as animals and plants appear to have gone through independent stages of extensive intron gain, which would explain the greater share of parallel gains.

We estimated that ancestral introns have a high probability to be retained in extant genomes, and conversely, many of the extant introns are ancient. The reasons for this remarkable endurance of a substantial fraction of the introns are not clear. One possibility is mechanistic, i.e., removing existing introns might be imprecise and hence potentially deleterious. Another possibility is functional, i.e., introns might have acquired many functional roles since entering eukaryotic genomes. The latter possibility is compatible with the recent observation of a negative correlation between the rate of intron gain and coding sequence evolution rate of a gene [19] which suggests that at least some of the introns are functionally relevant.

Methods

The data set

The methods and criteria used in compiling the data set have been described previously [19,20]. Briefly, the analyzed data set consisted of the reliable alignments of 391 genes from 19 eukaryotic species (a total of 289,902 sites). These include 9 animals (*Caenorhabditis elegans*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, rodents (*Mus musculus* and *Rattus norvegicus* combined), *Drosophila melanogaster*, *Anopheles gambiae*); 5 fungi (*Cryptococcus neoformans*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Aspergillus fumigatus*, *Neurospora crassa*); two plants (*Arabidopsis thaliana*, *Oryza sativa*); two apicomplexans (*Theileria parva*, *Plasmodium falciparum*); and the amoebozoan *Dictyostelium discoideum*.

Phylogenetic tree topology

Throughout this paper we assumed the traditional "crown-group" tree topology (Additional file 4). Specifically, the root position is between the Apicomplexa and the common ancestor of multicellular eukaryotes (plants and animals) [40], as opposed to the alternative Unikont-Bikont division [41]. We furthermore assume the Coelomata topology (Deuterostomia and insects are grouped together to the exclusion of nematodes) [42,43] as opposed to the Ecdysozoa topology (insects and nematodes are grouped together to the exclusion of Deuterostomia) [44,45]. The results, however, are not sensitive to the exact tree topology, as explicitly shown elsewhere [19,20].

The EM Algorithm

For each site, the S leaves form a set of observed random variables, their states being described by the corresponding pattern ω_p. The states of all the internal nodes, denoted σ, form a set of hidden random variables, that is, random variables whose states are not observed. In order to account for rate variability across sites, we associate with each pattern two hidden random variables, ρ_p^η and ρ_p^θ, that determine the value of the rate variables in that site. To sum up, the observed random variables are ω_p, and the hidden random variables are (σ, ρ_p^η, ρ_p^θ).

We assume that sites within a gene, as well as the genes themselves, evolve independently. Therefore, the total likelihood can be decomposed as

$$L(M_1,\ldots,M_G \mid \Theta) = \prod_{g=1}^{G} L(M_g \mid \Xi,\Psi_g,\Lambda) = \prod_{g=1}^{G}\prod_{p=1}^{\Omega} L(\omega_p \mid \Xi,\Psi_g,\Lambda)^{n_{gp}}.$$

and so

$$\log L(M_1,\ldots,M_G \mid \Theta) = \sum_{g=1}^{G}\sum_{p=1}^{\Omega} n_{gp} \log L(\omega_p \mid \Xi,\Psi_g,\Lambda). \tag{4}$$

According to the well-known EM paradigm [35], log $L(M_1, \ldots, M_G|\Theta)$ is guaranteed to increase as long as we maximize the auxiliary function

$$Q(\Theta,\Theta^0) = \sum_{g=1}^{G}\sum_{p=1}^{\Omega} n_{gp} Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0), \tag{5}$$

where

$$\begin{aligned} &Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0) = \\ &\quad \sum_{\sigma,\rho_p^\eta,\rho_p^\theta} \Pr(\sigma,\rho_p^\eta,\rho_p^\theta \mid \omega_p,\Xi^0,\Psi_g^0,\Lambda^0) \log \Pr(\omega_p,\sigma,\rho_p^\eta,\rho_p^\theta \mid \Xi,\Psi_g,\Lambda). \end{aligned} \tag{6}$$

If we replace the formal summing over all states of ρ_p^η and ρ_p^θ in (6) by a direct sum, we get

$$\begin{aligned} &Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0) = \\ &\quad \sum_{k=1}^{K_\eta}\sum_{k'=1}^{K_\theta}\sum_{\sigma} \Pr(\sigma,\rho_p^\eta = k,\rho_p^\theta = k' \mid \omega_p,\Xi^0,\Psi_g^0,\Lambda^0) \log \Pr(\omega_p,\sigma,\rho_p^\eta = k,\rho_p^\theta = k' \mid \Xi,\Psi_g,\Lambda). \end{aligned} \tag{7}$$

Using our notational conventions, we can write the first term in (7) as

$$\begin{aligned} &\Pr(\sigma,\rho_p^\eta = k,\rho_p^\theta = k' \mid \omega_p,\Xi^0,\Psi_g^0,\Lambda^0) = \\ &\quad \Pr(\rho_p^\eta = k,\rho_p^\theta = k' \mid \omega_p,\Xi^0,\Psi_g^0,\Lambda^0) \cdot \Pr(\sigma \mid \omega_p,\Xi^0,\Psi_{gkk'}^0), \end{aligned} \tag{8}$$

and the second term as

$$\begin{aligned} &\log \Pr(\omega_p,\sigma,\rho_p^\eta = k,\rho_p^\theta = k' \mid \Xi,\Psi_g,\Lambda) = \log \Pr(\rho_p^\eta = k \mid \Xi,\Psi_g,\Lambda) + \\ &\quad + \log \Pr(\rho_p^\theta = k' \mid \Xi,\Psi_g,\Lambda) + \log \Pr(\omega_p,\sigma \mid \Xi,\Psi_{gkk'}) = \\ &\quad \log f_k^\eta + \log f_{k'}^\theta + \log \Pr(\omega_p,\sigma \mid \Xi,\Psi_{gkk'}). \end{aligned} \tag{9}$$

Substituting (8) and (9) back into (7) gives

$$\begin{aligned} &Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0) = \\ &\quad \sum_{k=1}^{K_\eta}\sum_{k'=1}^{K_\theta} \left[\Pr(\rho_p^\eta = k,\rho_p^\theta = k' \mid \omega_p,\Xi^0,\Psi_g^0,\Lambda^0) \right] \cdot \\ &\qquad \cdot \left[\sum_{\sigma} \Pr(\sigma \mid \omega_p,\Xi^0,\Psi_{gkk'}^0) \cdot \left\{ \log f_k^\eta + \log f_{k'}^\theta + \log \Pr(\omega_p,\sigma \mid \Xi,\Psi_{gkk'}) \right\} \right] \end{aligned}$$

Denoting by $w_{gpkk'}$ and $Q_{gpkk'}$ the first and second square brackets, respectively, this expression becomes

$$Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0) = \sum_{k=1}^{K_\eta}\sum_{k'=1}^{K_\theta} w_{gpkk'} Q_{gpkk'}, \tag{10}$$

And, consequently.

$$Q(\Theta,\Theta^0) = \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K_\eta}\sum_{k'=1}^{K_\theta} n_{gp} w_{gpkk'} Q_{gpkk'} \tag{11}$$

The E-step

In this step, the function Q(Θ, Θ^0) or, equivalently, the set of coefficients $w_{gpkk'}$ and $Q_{gpkk'}$ is computed by using inward-outward recursion on the tree.

The inward (γ) recursion

Here we suggest a variation on the well-known Felsenstein's pruning algorithm [46]. Let us associate with each node t (except for the root) a vector $\gamma_i^{gpkk'}(t) = \Pr(V_t \mid q_t^P = i, \Xi^0, \Psi^0_{gkk'})$. This is the probability of observing the nodes V_t (which are a subset of the pattern ω_p) for a gene g, when the gain and loss rate variables are r_k^η and $r_{k'}^\theta$, respectively, and the parent node of t is known to be in state i. By definition, this function is initialized at all leaves ($t \in V_0$) by

$$\gamma(t \in V_0) = \begin{cases} \begin{pmatrix} 1 - \xi_t(1 - e^{-\eta_{gk}\Delta_t}) \\ 1 - (1 - \phi_t)e^{-\theta_{gk'}\Delta_t} \end{pmatrix} & q_t = 0 \\ \begin{pmatrix} \xi_t(1 - e^{-\eta_{gk}\Delta_t}) \\ (1 - \phi_t)e^{-\theta_{gk'}\Delta_t} \end{pmatrix} & q_t = 1. \end{cases} \quad (12)$$

Here and in the derivations to follow, we omit the superscript from γ. For all internal nodes (except for the root), γ is computed using the recursion

$$\gamma_i(t) = \sum_{j=0}^{1} T_{ij}(g,t)\tilde{\gamma}_j(t), \quad (13)$$

where $\tilde{\gamma}_j(t)$ is defined as $\gamma_j(\mathrm{L}(t))\gamma_j(\mathrm{R}(t))$. This is easy to see, as

$$\begin{aligned} \gamma_i(t) &= \Pr(V_t \mid q_t^P = i) = \Pr(V_t^L, V_t^R \mid q_t^P = i) = \sum_{j=0}^{1} \Pr(V_t^L, V_t^R, q_t = j \mid q_t^P = i) = \\ &= \sum_{j=0}^{1} \Pr(q_t = j \mid q_t^P = i) \cdot \Pr(V_t^L \mid q_t = j, q_t^P = i) \cdot \Pr(V_t^R \mid V_t^L, q_t = j, q_t^P = i). \end{aligned} \quad (14)$$

The first term is, simply, the definition of T_{ij} (g, t). Given q_t, V_t^L is independent of q_t^P, thus the second term is just Pr(V_t^L | $q_t = j$) = $\gamma_j(t^L)$. By similar argument, the third term is simply Pr(V_t^R | $q_t = j$) = $\gamma_j(t^R)$. Substituting these results in (14), we recover the recursion formula (13).

The γ-recursion allows for computing the likelihood of any observed pattern ω_p, given the values of the rate variables:

$$\begin{aligned} \Pr(\omega_p \mid \Xi^0, \Psi^0_{gkk'}) &= \Pr(V_0 \mid \Xi^0, \Psi^0_{gkk'}) = \Pr(V_0^L, V_0^R \mid \Xi^0, \Psi^0_{gkk'}) = \\ &= \sum_{i=0}^{1} \Pr(V_0^L, V_0^R, q_0 = i \mid \Xi^0, \Psi^0_{gkk'}) = \\ &= \sum_{i=0}^{1} \Pr(q_0 = i \mid \Xi^0, \Psi^0_{gkk'}) \cdot \Pr(V_0^L \mid q_0 = i, \Xi^0, \Psi^0_{gkk'}) \cdot \Pr(V_0^R \mid V_0^L, q_0 = i, \Xi^0, \Psi^0_{gkk'}). \end{aligned}$$

Given q_0, V_0^R is independent of V_0^L, and so

$$\Pr(V_0^R \mid V_0^L, q_0 = i, \Xi^0, \Psi^0_{gkk'}) = \Pr(V_0^R \mid q_0 = i, \Xi^0, \Psi^0_{gkk'}),$$

and

$$\Pr(\omega_p \mid \Xi^0, \Psi^0_{gkk'}) = \sum_{i=0}^{1} \pi_i \tilde{\gamma}_i(0). \quad (15)$$

One of the useful features of this recursion is that is allows to treat missing data fairly easily. Only a single option has to be added to the initialization phase (12),

$$\gamma(t \in V_0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad q_t = *. \quad (16)$$

The outward (α-β) recursion

Once the γ-recursion is computed, we can use it to compute a second, complementary, recursion. To this end, let us associate with each node t (except for the root node) a matrix $\alpha_{ij}^{gpkk'}(t) = \Pr(q_t = j, q_t^P = i \mid \omega_p, \Xi^0, \Psi^0_{gkk'})$. It is convenient to define for each node t (except for the root node) a vector $\beta_j^{gpkk'}(t) = \sum_{i=0}^{1} \alpha_{ij}^{gpkk'}(t) = \Pr(q_t = j \mid \omega_p, \Xi^0, \Psi^0_{gkk'})$.

Upon the computation of α, β is readily computed, too. Again, omitting the superscripts, α can be initialized from its definition on the two direct descendants of the root,

$$\alpha(\mathrm{D}(0)) = \frac{1}{\Pr(\omega_p|\Xi^0, \Psi^0_{gkk'})} \begin{cases} \begin{pmatrix} \pi_0\gamma_0(\bar{\mathrm{D}}(0))T_{00}(g,\mathrm{D}(0)) & 0 \\ \pi_1\gamma_1(\bar{\mathrm{D}}(0))T_{10}(g,\mathrm{D}(0)) & 0 \end{pmatrix} & \mathrm{D}(0) \in V_0, q_0^D = 0 \\ \begin{pmatrix} 0 & \pi_0\gamma_0(\bar{\mathrm{D}}(0))T_{01}(g,\mathrm{D}(0)) \\ 0 & \pi_1\gamma_1(\bar{\mathrm{D}}(0))T_{11}(g,\mathrm{D}(0)) \end{pmatrix} & \mathrm{D}(0) \in V_0, q_0^D = 1 \\ \begin{pmatrix} \pi_0\gamma_0(\bar{\mathrm{D}}(0))\tilde{\gamma}_0(\mathrm{D}(0))T_{00}(g,\mathrm{D}(0)) & \pi_0\gamma_0(\bar{\mathrm{D}}(0))\tilde{\gamma}_1(\mathrm{D}(0))T_{01}(\mathrm{D}(0)) \\ \pi_1\gamma_1(\bar{\mathrm{D}}(0))\tilde{\gamma}_0(\mathrm{D}(0))T_{10}(\mathrm{D}(0)) & \pi_1\gamma_1(\bar{\mathrm{D}}(0))\tilde{\gamma}_1(\mathrm{D}(0))T_{11}(\mathrm{D}(0)) \end{pmatrix} & \mathrm{D}(0) \notin V_0. \end{cases} \quad (17)$$

Here, D(0) stands for any one of the direct descendants of the root, and $\bar{\mathrm{D}}(0)$ is its sibling. For any other internal node, α is computed using the outward-recursion

$$\alpha(t) = \begin{pmatrix} \beta_0(\mathrm{P}(t))\tilde{\gamma}_0(t)T_{00}(g,t)/\gamma_0(t) & \beta_0(\mathrm{P}(t))\tilde{\gamma}_1(t)T_{01}(g,t)/\gamma_0(t) \\ \beta_1(\mathrm{P}(t))\tilde{\gamma}_0(t)T_{10}(g,t)/\gamma_1(t) & \beta_1(\mathrm{P}(t))\tilde{\gamma}_1(t)T_{11}(g,t)/\gamma_1(t) \end{pmatrix}. \quad (18)$$

To prove this recursion, let us start with the definition of α,

$$\alpha_{ij}(t) = \Pr(q_t = j, q_t^P = i \mid \omega_p) = \Pr(q_t = j, q_t^P = i \mid V_0) = \\ \Pr(q_t^P = i \mid V_0) \cdot \Pr(q_t = j \mid q_t^P = i, V_0) = \beta_i(\mathrm{P}(t)) \cdot \Pr(q_t = j \mid q_t^P = i, V_0), \tag{19}$$

and make the decomposition $V_0 = V_t + \bar{V}_t$, with $\bar{V}_t$ being the set of all leaves such that node t is not among their ancestors. But, given q_t^P, the state of node t is independent on $\bar{V}_t$, and therefore (19) becomes

$$\alpha_{ij}(t) = \beta_i(\mathrm{P}(t)) \cdot \Pr(q_t = j \mid q_t^P = i, V_t) \tag{20}$$

From Bayes formula,

$$\Pr(q_t = j \mid q_t^P = i, V_t) = \frac{\Pr(q_t = j, V_t \mid q_t^P = i)}{\Pr(V_t \mid q_t^P = i)} = \\ \frac{\Pr(q_t = j \mid q_t^P = i) \cdot \Pr(V_t \mid q_t = j, q_t^P = i)}{\gamma_i(t)} = \frac{T_{ij}(g,t)}{\gamma_i(t)} \cdot \Pr(V_t \mid q_t = j, q_t^P = i). \tag{21}$$

But, given q_t, V_t is independent of P(t) and therefore

$$\Pr(V_t \mid q_t = j, q_t^P = i) = \Pr(V_t \mid q_t = j) = \tilde{\gamma}_j(t). \tag{22}$$

Combining (22) and (21) in (20), we get

$$\alpha_{ij}(t) = \frac{\tilde{\gamma}_j(t)\beta_i(\mathrm{P}(t))}{\gamma_i(t)} T_{ij}(g,p),$$

which is just another form of (18). Finally, for each leaf that is not a descendant of the root

$$\alpha(t) = \begin{cases} \begin{pmatrix} \beta_0(\mathrm{P}(t)) & 0 \\ \beta_1(\mathrm{P}(t)) & 0 \end{pmatrix} & q_t = 0 \\ \begin{pmatrix} 0 & \beta_0(\mathrm{P}(t)) \\ 0 & \beta_1(\mathrm{P}(t)) \end{pmatrix} & q_t = 1. \end{cases} \quad t \in V_0, \mathrm{P}(t) \neq 0 \tag{23}$$

When missing data are present, two simple modifications are required. First, we have to add to the initialization phase (17) an option

$$\alpha(\mathrm{D}(0)) = \frac{1}{\Pr(\omega_p | \Xi^0, \Psi^0_{gkk'})} \begin{pmatrix} \pi_0\gamma_0(\bar{\mathrm{D}}(0))T_{00}(g,\mathrm{D}(0)) & \pi_0\gamma_0(\bar{\mathrm{D}}(0))T_{01}(\mathrm{D}(0)) \\ \pi_1\gamma_1(\bar{\mathrm{D}}(0))T_{10}(\mathrm{D}(0)) & \pi_1\gamma_1(\bar{\mathrm{D}}(0))T_{11}(\mathrm{D}(0)) \end{pmatrix} \quad \mathrm{D}(0) \in V_0, q_0^D = *$$

Second, we have to add to the finalization phase (23) an option

$$\alpha(t) = \begin{pmatrix} \beta_0(\mathrm{P}(t))T_{00}(g,t) & \beta_0(\mathrm{P}(t))T_{01}(g,t) \\ \beta_1(\mathrm{P}(t))T_{10}(g,t) & \beta_1(\mathrm{P}(t))T_{11}(g,t) \end{pmatrix} \quad q_t = *.$$

These inward-outward recursions are the phylogenetic equivalent of the backward-forward recursions known from hidden Markov models, and other versions of this method have been developed previously [47,48]. The version developed here can be shown to be the realization of the junction tree algorithm [49] on rooted bifurcating trees. The junction tree algorithm is a scheme to compute marginal probabilities of maximal cliques on graphs by means of belief propagation on a modified junction tree. Indeed, the matrix α computes marginal probabilities of pairs (t, P(t)), but such pairs are nothing but maximal cliques on rooted bifurcating trees.

Computing the coefficients w$_{\mathrm{gpkk'}}$

Here we show that the γ-recursion is sufficient to compute the coefficients $w_{gpkk'}$. From the definition, $w_{gpkk'} = \Pr(\rho_p^\eta = k, \rho_p^\theta = k' \mid \omega_p, \Xi^0, \Psi_g^0, \Lambda^0)$. Using the Bayes formula Pr(x, y | z) = Pr(x, y, z)/$\Sigma_{x,\, y}$ Pr(x, y, z), we can rewrite it as

$$w_{gpkk'} = \frac{\Pr(\rho_p^\eta = k, \rho_p^\theta = k', \omega_p \mid \Xi^0, \Psi_g^0, \Lambda^0)}{\sum_{h,h'} \Pr(\rho_p^\eta = h, \rho_p^\theta = h', \omega_p \mid \Xi^0, \Psi_g^0, \Lambda^0)} = \\ = \frac{\Pr(\rho_p^\eta = k \mid \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\rho_p^\theta = k' \mid \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\omega_p \mid \Xi^0, \Psi^0_{gkk'})}{\sum_{h,h'} \Pr(\rho_p^\eta = h \mid \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\rho_p^\theta = h' \mid \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\omega_p \mid \Xi^0, \Psi^0_{ghh'})}.$$

But $\Pr(\rho_p^\eta = k \mid \Xi^0, \Psi_g^0, \Lambda^0)$ is the current estimate of the probability of the gain rate variable to have the value r_k^η, namely $(f_k^\eta)^0$. Similarly, $\Pr(\rho_p^\theta = k' \mid \Xi^0, \Psi_g^0, \Lambda^0)$ is just $(f_{k'}^\theta)^0$. Therefore, the expression for the coefficients $w_{gpkk'}$ reduces to

$$w_{gpkk'} = \frac{(f_k^\eta)^0 (f_{k'}^\theta)^0 \Pr(\omega_p \mid \Xi^0, \Psi^0_{gkk'})}{\sum_{h,h'} (f_h^\eta)^0 (f_{h'}^\theta)^0 \Pr(\omega_p \mid \Xi^0, \Psi^0_{ghh'})}. \tag{24}$$

The function $\Pr(\omega_p \mid \Xi^0, \Psi^0_{gkk'})$ is the likelihood of observing pattern ω_p for gain and loss rate variables r_k^η and $r_{k'}^\theta$, respectively. This is readily computed upon completion of the γ-recursion, using (15).

Computing the coefficients $Q_{gpkk'}$

Here we show that those coefficients require the *α*, *β*-recursion. By definition,

$$Q_{gpkk'} = \sum_{\sigma} \Pr(\sigma \mid \omega_p, \Xi^0, \Psi^0_{gkk'}) \cdot [\log f^{\eta}_{k} + \log f^{\theta}_{k'} + \log \Pr(\omega_p, \sigma \mid \Xi, \Psi_{gkk'})].$$

The probability $\Pr(\omega_p, \sigma \mid \Xi, \Psi_{gkk'})$ is the likelihood of a particular realization of the tree, thus from (1)

$$\log \Pr(\omega_p, \sigma \mid \Xi, \Psi_{gkk'}) = \sum_{i=0}^{1} \delta(q_0, i) \cdot \log \pi_i + \sum_{i,j=0}^{1} \sum_{t=1}^{N-1} \delta(q_t, j) \delta(q^P_t, i) \cdot \log T_{ij}(g,t). \tag{25}$$

Here, $\delta(a, b)$ is the Kronecker delta function, which is 1 for $a = b$ and 0 otherwise. Denote the expectation over $\Pr(\sigma \mid \omega_p, \Xi^0, \Psi^0_{gkk'})$ by E_σ. Applying it to (25), we get

$$E_{\sigma}[\log \Pr(\omega_p, \sigma \mid \Xi, \Psi_{gkk'})] = \sum_{i=0}^{1} \log \pi_i \cdot E_{\sigma}[\delta(q_0, i)] + \sum_{i,j=0}^{1} \sum_{t=1}^{N-1} \log T_{ij}(g,t) \cdot E_{\sigma}[\delta(q_t, j) \delta(q^P_t, i)].$$

But $E_{\sigma}[\delta(q_0, i)] = \Pr(q_0 = i \mid \omega_p, \Xi^0, \Psi^0_{gkk'}) = \beta_i(0)$, and similarly $E_{\sigma}[\delta(q_t, j) \delta(q^P_t, i)] = \alpha_{ij}(t)$.

Hence, $Q_{gpkk'}$ is given by

$$\begin{aligned} Q_{gpkk'} &= \sum_{\sigma} \Pr(\sigma \mid \omega_p, \Xi^0, \Psi^0_{gkk'}) [\log f^{\eta}_{k} + \log f^{\theta}_{k'} + \log \Pr(\omega_p, \sigma \mid \Xi, \Psi_{gkk'})] = \\ &= \log f^{\eta}_{k} + \log f^{\theta}_{k'} + \sum_{i=0}^{1} \beta_i(0) \log \pi_i + \sum_{i,j=0}^{1} \sum_{t=1}^{N-1} \alpha_{ij}(t) \log T_{ij}(g,t). \end{aligned} \tag{26}$$

The M-step

Substituting (26) in (11), we obtain an explicit form of the function whose maximization guarantees stepping up-hill in the likelihood landscape,

$$\begin{aligned} Q = &\sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} n_{gp} w_{gpkk'} (\log f^{\eta}_{k} + \log f^{\theta}_{k'}) + \\ &+ \sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} n_{gp} w_{gpkk'} [\beta^{gpkk'}_{0}(0) \log \pi_0 + \beta^{gpkk'}_{1}(0) \log \pi_1] + \\ &+ \sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{gp} w_{gpkk'} \alpha^{gpkk'}_{00}(t) \log[1 - \xi_t (1 - e^{-\eta_{gk} \Delta_t})] + \\ &+ \sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{gp} w_{gpkk'} \alpha^{gpkk'}_{01}(t) [\log \xi_t + \log(1 - e^{-\eta_{gk} \Delta_t})] + \\ &+ \sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{gp} w_{gpkk'} \alpha^{gpkk'}_{10}(t) \log[1 - (1 - \phi_t) e^{-\theta_{gk'} \Delta_t}] + \\ &+ \sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{gp} w_{gpkk'} \alpha^{gpkk'}_{11}(t) [\log(1 - \phi_t) - \theta_{gk'} \Delta_t]. \end{aligned} \tag{27}$$

Actually, any increase in Q is sufficient to guarantee an increase in the likelihood, suggesting that the precise maximization of Q is not particularly important. Therefore, we speed up the computations by performing low-tolerance maximization with respect to each of the parameters individually. Except for the parameters λ_η and λ_θ, it is easy to differentiate Q twice with respect to any parameter. This lends itself to using simple zero-finding algorithms of which we chose the Newton-Raphson algorithm [50].

Maximizing Q with respect to the shape parameters λ_η and λ_θ is more involved, as Q depends on these parameters only through the discrete approximation of the rate variability distributions (3). In our implementation, we used Yang's quantile method [34] to compute the discrete levels of the gamma distributions such that each level has equal probability. Formally, $f^{\eta}_{1} = \nu, f^{\eta}_{k} = (1 - \nu)/(K_\eta - 1)$ for $k = 2, ..., K_\eta$ and $f^{\theta}_{k} = 1/K_\theta$ for $k = 1, ..., K_\theta$. To perform the maximization in this case, we used Brent's maximization algorithm that does not require derivatives [50].

Reconstruction of Ancestral States and Events

Given the *α*, *β*, *γ*-recursions on the tree with the final model parameters, it is straightforward to reconstruct the history of intron evolution, and to assign gains and losses to specific branches. The number of introns in an internal node t for a gene g, assuming gain rate variable r^{η}_{k} and loss rate variable $r^{\theta}_{k'}$, given that the observed pattern is ω_p, is $n_{gp} \omega_{gpkk'} \beta^{gpkk'}_{1}(t)$. Similarly, the number of loss events along the branch t is $n_{gp}\omega_{gpkk'}\alpha_{10}^{gpkk'}(t)$, and the number of gain events along this branch is $n_{gp}\omega_{gpkk'}\alpha_{01}^{gpkk'}(t)$.

Confidence Intervals

In order to obtain confidence intervals on the model parameters, we used the profile likelihood technique. In brief, if ϑ is one parameter in the model, and $\bar{\Theta}$ is the set of remaining parameters, then the profile likelihood of ϑ is defined as

$$L(\vartheta) = \max_{\bar{\Theta}} \Pr(\vartheta, \bar{\Theta}).$$

That is, we compute the maximum likelihood under the constraint that the value of ϑ is given. If we denote the overall maximum likelihood by $L(\Theta) = \max_{\Theta} \Pr(\Theta)$, then the likelihood ratio

$$\lambda = -2\ln\frac{L(\vartheta_0)}{L(\Theta)}, \qquad (28)$$

under the hypothesis that $\vartheta = \vartheta_0$ is distributed according to the χ^2 distribution with one degree of freedom. In order to find the 95% confidence interval of the parameter ϑ around its optimal value ϑ_0, we find the value of ϑ for which the likelihood ratio (28) exceeds the value 3.84 (95%-percentile of the $\chi^2(1)$ distribution). This value is found numerically using Ridder's method [50].

Authors' contributions

LC contributed to the development of the probabilistic model, developed the EM algorithm, and drafted the manuscript; IBR collected the data and contributed to the development of the probabilistic model; YIW contributed to the development of the probabilistic model; EVK conceived of the study, provided the biological interpretation of the results, and finalized the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Maximum likelihood estimation of the model parameters. Each parameter is associated with three values – the lower 95% confidence bound, the optimal value, and the upper 95% confidence bound. The confidence intervals were computed using the profile likelihood technique.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-192-S1.doc]

Additional file 2

Comparison of the number of shared intron positions with the number expected by chance. All pairs of species (20), where the number of shared intron positions is not significantly greater than the number expected by chance alone.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-192-S2.doc]

Additional file 3

Pattern-by-pattern analysis of parallel gains. The order of species in each pattern is Dicdi, Caeel, Strpu, Cioin, Danre, Galga, Homsa, Roden, Drome, Anoga, Cryne, Schpo, Sacce, Aspfu, Neucr, Arath, Orysa, Thepa, and Plafa. The frequency of a pattern is the number of times it was observed in our data.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-192-S3.doc]

Additional file 4

*The phylogenetic tree of eukaryotes used in the present study. Species and lineage abbreviations: Caeel (*Caenorhabditis elegans*), Strpu (*Strongylocentrotus purpuratus*), Cioin (*Ciona intestinalis*), Danre (*Danio rerio*), Galga (*Gallus gallus*), Homsa (*Homo sapiens*), roden (*Mus musculus *and* Rattus norvegicus *combined), Drome (*Drosophila melanogaster*), Anoga (*Anopheles gambiae*), cryne (*Cryptococcus neoformans*), Schpo (*Schizosaccharomyces pombe*), Sacce (*Saccharomyces cerevisiae*), Aspfu (*Aspergillus fumigatus*), Neucr (*Neurospora crassa*), Arath (*Arabidopsis thaliana*), Orysa (*Oryza sativa*), Thepa (*Theileria parva*), Plafa (*Plasmodium falciparum*), Dicdi (*Dictyostelium discoideum*), AME (Ancestor of Multicellular Eukaryotes).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-192-S4.doc]

Acknowledgements

We thank Jacob Goldberger for useful discussions regarding aspects of the EM algorithm. This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

1. Lynch M, Richardson AO: **The evolution of spliceosomal introns.** *Curr Opin Genet Dev* 2002, **12(6):**701-710.
2. Roy SW, Gilbert W: **The evolution of spliceosomal introns: patterns, puzzles and progress.** *Nat Rev Genet* 2006, **7(3):**211-221.
3. Rodríguez-Trelles F, Tarrío R, Ayala FJ: **Origins and evolution of spliceosomal introns.** *Annu Rev Genet* 2006, **40:**47-76.
4. Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J: **A spliceosomal intron in Giardia lamblia.** *Proc Natl Acad Sci U S A* 2002, **99(6):**3701-3705.
5. Simpson AG, MacQuarrie EK, Roger AJ: **Eukaryotic evolution: early origin of canonical introns.** *Nature* 2002, **419(6904):**270.
6. Vanacova S, Yan W, Carlton JM, Johnson PJ: **Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis.** *Proc Natl Acad Sci U S A* 2005, **102(12):**4430-4435.
7. Russell AG, Shutt TE, Watkins RF, Gray MW: **An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia.** *BMC Evol Biol* 2005, **5:**45.
8. Koonin EV: **The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate?** *Biol Direct* 2006, **1(1):**22.

9. Martin W, Koonin EV: **Introns and the origin of nucleus-cytosol compartmentalization.** *Nature* 2006, **440(7080):**41-45.
10. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: **Analysis of evolution of exon-intron structure of eukaryotic genes.** *Brief Bioinform* 2005, **6(2):**118-134.
11. Fedorov A, Merican AF, Gilbert W: **Large-scale comparison of intron positions among animal, plant, and fungal genes.** *Proc Natl Acad Sci U S A* 2002, **99(25):**16128-16133.
12. Qiu WG, Schisler N, Stoltzfus A: **The evolutionary gain of spliceosomal introns: sequence and phase preferences.** *Mol Biol Evol* 2004, **21(7):**1252-1263.
13. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.** *Curr Biol* 2003, **13(17):**1512-1517.
14. Roy SW, Fedorov A, Gilbert W: **Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain.** *Proc Natl Acad Sci U S A* 2003, **100(12):**7158-7162.
15. Roy SW, Gilbert W: **The pattern of intron loss.** *Proc Natl Acad Sci U S A* 2005, **102(3):**713-718.
16. Roy SW, Gilbert W: **Complex early genes.** *Proc Natl Acad Sci U S A* 2005, **102(6):**1986-1991.
17. Csuros M: **Likely scenarios of intron evolution.** *Comparative Genomics Lecture Notes in Computer Science* 2005, **3678:**47-60.
18. Nguyen HD, Yoshihama M, Kenmochi N: **New maximum likelihood estimators for eukaryotic intron evolution.** *PLoS Comput Biol* 2005, **1(7):**e79.
19. Carmel L, Rogozin IB, Wolf YI, Koonin EV: **Evolutionarily conserved genes preferentially accumulate introns.** *Genome Res* 2007, **17:**1045-1050.
20. Carmel L, Wolf YI, Rogozin IB, Koonin EV: **Three distinct modes of intron dynamics in the evolution of eukaryotes.** *Genome Res* 2007, **17:**1034-1044.
21. Roy SW, Gilbert W: **Rates of intron loss and gain: implications for early eukaryotic evolution.** *Proc Natl Acad Sci U S A* 2005, **102(16):**5773-5778.
22. Carmel L, Rogozin IB, Wolf YI, Koonin EV: **An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes.** *Comparative Genomics Lecture Notes in Computer Science* 2005, **3678:**35-46.
23. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: **Conservation versus parallel gains in intron evolution.** *Nucleic Acids Res* 2005, **33(6):**1741-1748.
24. Nei M, Chakraborty R, Fuerst PA: **Infinite allele model with varying mutation rate.** *Proc Natl Acad Sci U S A* 1976, **73(11):**4164-4168.
25. Uzzell T, Corbin KW: **Fitting discrete probability distributions to evolutionary events.** *Science* 1971, **172(988):**1089-1096.
26. Felsenstein J: **Inferring Phylogenies.** Sunderland, MA , Sinauer Associates; 2004.
27. Dibb NJ: **Proto-splice site model of intron origin.** *J Theor Biol* 1991, **151(3):**405-416.
28. Dibb NJ, Newman AJ: **Evidence that introns arose at proto-splice sites.** *Embo J* 1989, **8(7):**2015-2021.
29. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: **Reconstruction of ancestral protosplice sites.** *Curr Biol* 2004, **14(16):**1505-1508.
30. Gu X, Fu YX, Li WH: **Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites.** *Mol Biol Evol* 1995, **12(4):**546-557.
31. Hasegawa M, Kishino H, Yano T: **Man's place in Hominoidea as inferred from molecular clocks of DNA.** *J Mol Evol* 1987, **26(1-2):**132-147.
32. Mayrose I, Friedman N, Pupko T: **A Gamma mixture model better accounts for among site rate heterogeneity.** *Bioinformatics* 2005, **21 Suppl 2:**ii151-ii158.
33. Jin L, Nei M: **Limitations of the evolutionary parsimony method of phylogenetic analysis.** *Mol Biol Evol* 1990, **7(1):**82-102.
34. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39(3):**306-314.
35. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society, Series B* 1977, **39:**138.
36. Nguyen HD, Yoshihama M, Kenmochi N: **Authors' reply.** *PLoS Comput Biol* 2006, **2:**e83.
37. Stoltzfus A, Logsdon JM Jr., Palmer JD, Doolittle WF: **Intron "sliding" and the diversity of intron positions.** *Proc Natl Acad Sci U S A* 1997, **94(20):**10739-10744.
38. Roy SW: **Intron-rich ancestors.** *Trends Genet* 2006, **22(9):**468-471.
39. Hedges SB, Kumar S: **Genomic clocks and evolutionary timescales.** *Trends Genet* 2003, **19:**200-206.
40. Hedges SB: **The origin and evolution of model organisms.** *Nat Rev Genet* 2002, **3(11):**838-849.
41. Stechmann A, Cavalier-Smith T: **The root of the eukaryote tree pinpointed.** *Curr Biol* 2003, **13(17):**R665-6.
42. Wolf YI, Rogozin IB, Koonin EV: **Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis.** *Genome Res* 2004, **14:**29-36.
43. Rogozin IB, Wolf YI, Carmel L, Koonin EV: **Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements.** *Mol Biol Evol* 2007, **24(4):**1080-1090.
44. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA: **Evidence for a clade of nematodes, arthropods and other moulting animals.** *Nature* 1997, **387(6632):**489-493.
45. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia.** *Mol Biol Evol* 2005, **22(5):**1246-1253.
46. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17(6):**368-376.
47. Friedman N, Ninio M, Pe'er I, Pupko T: **A structural EM algorithm for phylogenetic inference.** *J Comput Biol* 2002, **9(2):**331-353.
48. Siepel A, Haussler D: **Phylogenetic estimation of context-dependent substitution rates by maximum likelihood.** *Mol Biol Evol* 2004, **21(3):**468-488.
49. Castillo E, Gutierrez JM, Hadi AS: **Expert systems and probabilistic network models (Monographs in Computer Science).** New York , Springer; 1996.
50. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: **Numerical recipes in C: The art of scientific computing.** 2nd edition. New York , Cambridge University Press; 1992.