A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes

Supporting Information

Miklós Csűrös  Igor B. Rogozin  Eugene V. Koonin

May 23, 2011
## Contents

1 Supporting Methods 3
   1.1 Species and annotated genomes 3
   1.2 Homolog identification 4
   1.3 Ortholog set selection 5
   1.4 Input table of intron presence and absence 5

2 Convergence 6
   2.1 Priors and proposals 6
   2.2 Sampling 7

3 Ambiguity at Amoeboza and Chromalveolates 11

4 Ancestral phase distributions 16

5 Model fit 19

6 Ancestral inference 21

7 Simulations 24

8 Nægleria gruberi 26
1 Supporting Methods

1.1 Species and annotated genomes

We included the following species in the study: *Aureococcus anophagefferens* (Aano), *Aedes aegypti* (Aaeg), *Agaricus bisporus* (Abis), *Anopheles gambiae* (Agam), *Allomyces macrognus* ATCC 38327 (Amac), *Apis mellifera* (Amel), *Aspergillus nidulans* FGSC A4 (Anid), *Acyrthosiphon pisum* (Apis), *Arabidopsis thaliana* (Ath), *Babesia bovis* (Bbov), *Batrachochytrium dendrobatidis* (Bden), *Branchiostoma floridae* (Bflo), *Botryotinia fuckeliana B05.10* (Bfuc), *Brugia malayi* (Bmal), *Bombyx mori* (Bmor), *Coccomyxa sp. C-169* (C169), *Chlorella sp. NC64a* (C64a), *Caenorhabditis briggsae* (Cbri), *Caenorhabditis elegans* (Cele), *Coprinopsis cinerea okayama7#130* (Ccint), *Cochliobolus heterostrophus* C5 (Chet), *Coccidioides immitis RS* (Cimm), *Ciona intestinalis* (Cint), *Cryptococcus neoformans var. neoformans* (Cneo), *Chlamydomonas reinhardtii* (Crei), *Capitella teleta* (Ctel), *Capsaspora owczarzaki ATCC 30864* (Cowc), *Dictyostelium discoideum* (Ddis), *Dictyostelium purpureum* (Dpur), *Drosophila melanogaster* (Dmel), *Drosophila mojavensis* (Dmoj), *Daphnia pulex* (Dpul), *Danio rerio* (Drer), *Entamoeba dispar* (Dis), *Entamoeba histolytica* (Ehis), *Emiliania huxleyi* (Ehux), *Fragilariopsis cylindrus* (Fcyl), *Phanerochaete chrysosporium* (Fchr), *Phaeodactylum tricornutum* (Ftri), *Gallus gallus* (Ggal), *Gibberella zeae PH-1* (Gzea), *Hydra magnipapillata* (Hmag), *Helobdella robusta* (Hrob), *Homo sapiens* (Hsap), *Ixodes scapularis* (Isca), *Laccaria bicolor* (Lbic), *Lottia gigantea* (Lgig), *Micromonas sp. RCC299* (M299), *Monosiga brevicollis* (Mbre), *Mucor circinelloides* (Mcir), *Mycosphaerella fijiensis* (Mfij), *Mycosphaerella graminicola* (Mgra), *Magnaporthes grisea* 70-15 (Mgri), *Melampsora laricis-populina* (Mlar), *Micromonas pusilla CCMP1545* (Mpus), *Neurospora crassa OR74A* (Ncra), *Nägleria gruberi* (Ngru), *Nematostella vectensis* (Nvec), *Nasonia vitripennis* (Nvit), *Ostreococcus sp. RCC809* (O809), *Ostreococcus lucimarinus* (Oluc), *Oryza sativa japonica* (Osat), *Ostreococcus taurii* (Otou), *Phytophthora capsici* (Pcap), *Plasmidium falciparum* (Pfal), *Puccinia graminis* (Pgr), *Pediculus humanus* (Phum), *Phaeosphaeria nodorum SN15* (Pnod), *Physcomitrella patens subsp. patens* (Ppat), *Phytophthora ramorum* (Pram), *Pyrenophora tritici-repentis Pt-1C-BFP* (Prep), *Proterospongia sp. ATCC 50818* (Prsp), *Phytophthora sojae* (Psjo), *Paramecium tetraurelia* (Ptet), *Plasmodium vivax* (Pviv), *Plasmodium yoelii yoelii* (Pyoe), *Rhizopus oryzae* (Rory), *Sorghum bicolor* (Sbic), *Saccharomyces cerevisiae* (Scer), *Schizosaccharomyces japonicus yFS275* (Sjap), *Schistosoma mansoni* (Sman), *Selaginella moellendorffii* (Smoe), *Schizosaccharomyces pombe* (Spom), *Spizellomyces punctatus* DAOM BR117.
(Spun), Strongylocentrotus purpuratus (Spur), Sporobolomyces roseus (Sros), Sclerotinia sclerotiorum 1980 UF-70 (Sscl), Trichoplax adhaerens (Tadh), Theileria annulata (Tann), Tribolium castaneum (Tcas), Toxoplasma gondii (Tgon), Taenopygia guttata (Tgut), Theileria parva (Tpar), Thalassiosira pseudonana (Tpse), Tetrahymena thermophila (Tthe), Ustilago maydis 521 (Umay), Uncinocarpus reesii 1704 (Uree), Volvox carteri (Vcar), Vitis vinifera (Vvin).

The following annotated genomes were downloaded from the Refseq database (version 40) of the National Center for Biotechnology Information [ftp://ftp.ncbi.nih.gov/refseq/release/]: Aaeg, Agam, Amel, Anid, Apis, Atha, Bbov, Bflo, Bfuc, Bmal, Bcbr, Ccin, Cele, Cimm, Cint, Cneo, Crei, Ddis, Dmel, Dmoj, Drer, Edis, Ehis, Ggal, Gzea, Hmag, Hsap, Isca, M299, Mgrl, Ncra, Nvec, Nvit, Oluc, Osat, Pfal, Phum, Pnod, Ppat, Prep, Ptet, Pvin, Pyoe, Sbic, Scer, Sjap, Sman, Spom, Spur, Sscl, Tadh, Tann, Tcas, Tgon, Tgut, Tpar, Tthe, Umay, Uree, Vvin.

The following genome sequence data were produced by the US Department of Energy Joint Genome Institute [http://www.jgi.doe.gov/] in collaboration with the user community: Aano, Abis, C169, C64a, Chet, Ctel, Dpul, Dpur, Ehux, Fchr, Fclyl, Ftri, Hrob, Lbic, Lgig, Mbre, Mcir, Mfij, Mgra, Mlar, Mpus, Ngru, O809, Otav, Pbla, Pcap, Pgra, Pram, Psoj, Smoe, Sros, Tpse, Vcar.

The following annotated genomes were downloaded from the MIT Broad Institute [http://www.broadinstitute.org/]: Amac, Bden, Prsp, Rory, Spun. The annotated Bbom genome was downloaded from the SilkDB (Institute of Sericulture and Systems Biology, Southwest University, China) [http://silkworm.genomics.org.cn/].

1.2 Homolog identification

Orthologs were identified by slightly modifying our previously described procedure (Csûrös et al., 2008) We used the eukaryotic ortholog groups (so-called euKOGs and euNOGs) from the eggNog database (Muller et al., 2010). The ortholog groups were not used directly, but only for “seeding” a homolog identification step based on bidirectional best hits between each genomes protein sequences, and the seed groups. Specifically, we used 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) to build BLAST databases on individual genomes and the seed groups. First, the seed groups were used to query each genome. Hits between a seed sequence and the query were retained if (1) had an E-value at most $10^{-9}$, (2) had an amino acid identity at
least 50%, and (3) had a score within 90% of the best BLAST hit with the same query. Subsequently, the retained sequences were used to query the seed groups to filter out ambiguous group assignments. Sequences passed the filtering step if they had the highest scoring hit to the same seed group used in the initial search, and hit other groups with at most 98% of the best score.

1.3 Ortholog set selection

In a second phase, the collected paralogs were submitted to our ortholog “weaving” protocol described before (Csúrsos et al., 2008) which, by reconciling the gene phylogeny with the given species phylogeny, selected at most one sequence for each species that are highly likely to be orthologs. Finally, the resulting candidate ortholog sets were further filtered by verifying their phylogenetic relationships. In particular, we required a non-negative log-likelihood ratio between the neighbor-joining tree and the known species phylogeny, computed by using PhyML (Guindon and Gascuel, 2003).

1.4 Input table of intron presence and absence

Following a well-established analysis pipeline (Rogozin et al., 2003, 2005; Csúrsos et al., 2008) we selected homologous intron-bearing positions by projecting the exon-intron structure onto aligned orthologous coding sequences. The resulting intron dataset is a table of intron absence and presence, in which each column corresponds to a splice site projected onto an unambiguous alignment column (retaining splice phase information), and each row corresponds to one of the 99 terminal taxa. Table entries may be 1 (splice site is present), 0 (no splice site), or “ambiguous,” denoting a missing ortholog or an uncertain alignment. The table contains all columns with at most 24 ambiguous entries, and was produced using the Malin software package (Csűrsös, 2008) with its default settings for ambiguity. The alignments were done using Muscle (Edgar, 2004) with its default settings.
2 Convergence

2.1 Priors and proposals

Recall that the model is defined by the parameters

$$\theta = (\alpha_{\text{gain}}, \alpha_{\text{loss}}, \pi, (t_{uv}, q_{uv}) : uv \in T).$$

The prior distribution was uniform for every parameter: over the range $[0, 10]$ for $\alpha$ and $t_{uv}$, and over the range $[0, 1]$ for $\pi$ and $q_{uv}$.

In a typical MCMC proposal, a subset of model parameters was chosen, and then multiplied by a random value. Random multipliers were generated by an exponential distribution restricted to some range $[e^{-\sigma}, e^{\sigma}]$. More precisely, the pseudorandom multipliers were drawn as $M_\sigma = \exp(U_{-\sigma,\sigma})$, using a pseudorandom uniform $U_{-\sigma,\sigma}$ over the interval $(-\sigma, \sigma)$. In an Edge move, for instance, the proposed parameter vector $\theta'$ was produced by setting the length of a randomly picked to $t'_{uv} = t_{uv} \cdot M_\sigma$, and keeping all the other parameter values unchanged.

The following table shows the proposals used during MCMC sampling.

| Move     | Parameter          | Max. multiplier ($e^\sigma$) | Frequency |
|----------|--------------------|------------------------------|-----------|
| Root     | $\pi$              | 1.44                         | 2/m       |
| Var      | $\alpha_{\text{gain}}, \alpha_{\text{loss}}$ | 1.2                          | 1/m       |
| Edge     | $t_{uv}$           | 1.44                         | 20/m      |
| Rates    | $q_{uv}$           | 1.44                         | 20/m      |
| Resize   | all $t_{uv}$ in a subtree | 1.44                          | 20/m      |
| StarRates| all $q_{uv}$ incident to a node | 1.44                          | 20/m      |

For Edge and Rates moves a random edge was selected uniformly. For a Resize move, a random non-terminal node was selected uniformly, and all edge lengths were multiplied with the same random $M_\sigma$. For a StarRates move, a random non-terminal node was selected uniformly, and the ratios on each incident edge (children, and the parent — except for the root) were multiplied with a separately drawn random $M_\sigma$.

After the burn-in period, the multiplier range was reduced by setting $\sigma \leftarrow \sigma/2$ for each move type. The Var move was applied only to rate variations with more than 1 discrete categories. The frequency denominator $m = 82 + v$ depends on the types of permitted Var moves: $m = 82$ for no rate variation across sites, $m = 83$ if loss or gain, but not both vary across sites, and $m = 84$ if both vary across sites.
2.2 Sampling

Sampling data was collected after a burn-in period of 500 thousand steps, subsampling every 1000th subsequent position, across 100 independently launched chains. Figure 1 illustrates the convergence of the log-likelihoods to the same value across all the chains. Figure 2 illustrates the mixing of the sampled reconstructions at LECA. In Figure 3, a mixing efficiency is calculated, determined by across- and within-chain variations. Sampled reconstructions at a given step are independent across chains, and therefore can be used to estimate the standard deviation for the estimated posterior distribution. We define the mixing coefficient as the ratio between the estimated standard deviation within a chain, and the standard deviation of the estimated quantity. With sufficiently sparse subsampling, the coefficient approaches 1. Figure 3 illustrates the superior mixing at most ancestral nodes.

Figure 1

MCMC sampling traces: log-likelihood. The Y axis shows the log-likelihood; 12 of the 100 chains are grouped by fours on the plots. The vertical line on each plot shows the end of the burn-in period, and a horizontal dashed line shows the harmonic average of the likelihoods across all chains after burn-in.

Figure 2

MCMC sampling traces after burnin: LECA. The Y axis shows the number of introns at LECA at different sampled reconstructions across 12 of the 100 chains, grouped by fours. The horizontal dashed line shows the average across all chains after burn-in.

Figure 3

Mixing at ancestral estimates. The standard deviation (SD) was estimated within chains, and across chains. For within-chain variation, the SD was calculated from the 500 post-burnin sampled ancestral intron counts, and then averaged across the chains. For across-chain variation, the SD was calculated from the 100 chains at each MCMC step, and then averaged across the steps.
Mixing efficiency for ancestral reconstruction

- Heterokonts: 68%
- xF4: 79%
- Amoebozoa: 59%
- Chromista: 81%
- Ascomycota: 83%
- Alveolates: 90%
- Chromalveolates: 90%
- LECA: 92%
- Entamoeba: 43%
3 Ambiguity at Amoeboza and Chromalveolates

Figure 4 illustrates that the ambiguity at the amobozoan ancestor is clearly caused by the impossibility of inferring the timing of intron losses. Figure floatnbr5 illustrates the more complicated dependencies between chromalveolate ancestors. The tree below facilitates the interpretation by showing the assumed phylogenetic relationships between protists, and the naming of ancestral nodes. Table 6 shows the lack of signal in chromalveolate algae, and the strong intron conservation seen in human introns.
Sampled ancestral reconstructions at the ancestor of Amoebozoa in an MCMC trace. The plots show introns lost on the branches ("::l") to the inferred number of introns in the Amoebozoan ancestor ("Amoebozoa::p"). The coloring of the dots indicate when the state was visited.

Sampled ancestral reconstructions at the ancestors of chromalveolate algae an MCMC trace. The plots show introns lost on the branches ("::l") or inferred to be present ("::p") at various positions. The coloring of the dots indicate when the state was visited.

Introns in shared positions between some nodes in the data set. The first data row in each subtable gives the reference taxon and the number of sites in which the site has an intron in the reference group’s members. The subsequent rows list the intron-sharing patterns between the reference and other taxa $t$. “Shared” sites have introns in both taxa; “reference-specific” sites have introns in the reference taxon, opposing absent and ambiguous entries in $t$. “Taxon-specific” sites have introns in $t$ but not in the reference taxon. Notice that the counts include only unambiguously aligned sites, and, thus, they may not sum up to as expected. For instance, only 654 of 780 (84%) human introns can be unambiguously projected onto homologous amphioxus sequence, and 537 of those amphioxus positions correspond to a splice site. For higher-level taxa, “sharing” was established by at least one terminal descendant taxon: e.g., 568 human introns are projected to an annotated splice site in at least one arthropod genome. Percentages in parentheses are calculated with respect to all introns for the taxa in question: either the percentage of introns in the reference (“15% of all human introns do not align with amphioxus introns), or those in $t$ (“20% of amphioxus introns do not align with human introns).
### Human introns in shared positions with other taxa

| Taxon            | introns | shared | human-specific | taxon-specific |
|------------------|---------|--------|----------------|----------------|
| Hsap [human]     | 780     |        |                |                |
| Bflo [amphioxus]| 711     | 537    | 117 (15%)      | 140 (20%)      |
| Arthropods       | 1424    | 568    | 200 (26%)      | 701 (49%)      |
| Tadh [Trichoplax]| 893     | 564    | 145 (19%)      | 244 (27%)      |
| Fungi            | 2565    | 318    | 462 (59%)      | 2036 (79%)     |
| Dictyostelium    | 212     | 82     | 672 (86%)      | 110 (52%)      |
| Embryophyta [land plants] | 853 | 146 | 618 (79%) | 614 (72%) |
| Apicomplexa      | 611     | 79     | 622 (80%)      | 469 (77%)      |
| Phytophthora     | 165     | 45     | 692 (89%)      | 104 (63%)      |

### Phytophthora introns in shared positions with other taxa

| Taxon | introns | shared | Phytophthora-spec. | taxon-spec. |
|-------|---------|--------|--------------------|-------------|
| Phytophthora | 165 | | | |
| diatoms | 208 | 7 | 145 (88%) | 197 (95%) |
| Ehux [E. huxleyi] | 353 | 3 | 69 (60%) | 302 (86%) |
| Apicomplexa | 611 | 27 | 128 (78%) | 556 (91%) |
| Dictyostelium | 212 | 27 | 132 (80%) | 175 (82%) |
| Chlorophyta [green algae] | 1781 | 42 | 120 (72%) | 1658 (93%) |
| Embryophyta [land plants] | 853 | 50 | 111 (67%) | 768 (90%) |

### T. pseudonana introns in shared positions with other taxa

| Taxon            | introns | shared | Tpse-specific | taxon-specific |
|------------------|---------|--------|---------------|----------------|
| Tpse [T. pseudonana] | 115 | | | |
| Bacillariophyceae [pennate diatoms] | 121 | 28 | 78 (68%) | 81 (67%) |
| Phytophthora      | 165     | 4      | 109 (95%)     | 128 (78%)      |
| Ehux [E. huxleyi] | 353     | 3      | 69 (60%)      | 302 (86%)      |
| Apicomplexa       | 611     | 4      | 105 (91%)     | 539 (88%)      |
| Dictyostelium     | 212     | 4      | 106 (92%)     | 181 (85%)      |
| Chlorophyta [green algae] | 1781 | 8 | 105 (91%) | 1505 (85%) |
| Embryophyta [land plants] | 853 | 5 | 108 (94%) | 726 (85%) |

### E. huxleyi introns in shared positions with other taxa

| Taxon            | introns | shared | Ehux-specific | taxon-specific |
|------------------|---------|--------|---------------|----------------|
| Ehux [E. huxleyi] | 353     | | 302 (86%) | 69 (60%)      |
| Phytophthora      | 165     | 3      | 302 (86%)     | 69 (60%)       |
| Chlorophyta [green algae] | 1781 | 44 | 297 (84%) | 1163 (65%) |
| Bilateria        | 2742    | 47     | 306 (87%)     | 1883 (43%)     |
4 Ancestral phase distributions

Individual splice site histories were reconstructed using the maximum-likelihood model found during sampling. Figure 7 and Table 8 show the reconstructed phases at sites represented in terminal taxa.

**Figure 7**
Intron phase distribution at different nodes. Branch widths are proportional to (inferred) intron density. Pie charts show the relative frequencies of phase-1, -2 and -0 splice sites. Notice that the reconstruction covers only the splice sites in the data set (where phase can be established), and not the extrapolated all-absent profiles.

**Table 8**
Intron phase distribution, gains and losses at different nodes. Pie charts show the relative frequencies of phase-1, -2 and -0 splice sites.
| Species   | Loss Phase 1 | Loss Phase 2 | Loss Phase 0 |
|-----------|--------------|--------------|--------------|
| Micromonas |              |              |              |
| xG2       |              |              |              |
| Volvocales |              |              |              |
| Opisthokonts |            |              |              |
| Holozoa   |              |              |              |
| Vertebrates |             |              |              |
| Aves      |              |              |              |
| Culicidae |              |              |              |
| xF10      |              |              |              |
| xF12      |              |              |              |
| xF9       |              |              |              |
| Eurotiomycetes |        |              |              |
| Sordariomycetes |          |              |              |
| Leotiomycetes |           |              |              |
| xH2       |              |              |              |
| xVvin     |              |              |              |
| xSmoe     |              |              |              |
| M299      |              |              |              |
| Edis      |              |              |              |
| Ddis      |              |              |              |
| Tgut      |              |              |              |
| Cowc      |              |              |              |
| Pbla      |              |              |              |
| Rory      |              |              |              |
| Chet      |              |              |              |
| Mfij      |              |              |              |
| Sscl      |              |              |              |
| Scer      |              |              |              |
| Tann      |              |              |              |
| Pyoe      |              |              |              |
| Tgon      |              |              |              |
| Ehux      |              |              |              |
| Node       |              |              |              |
| Introns   |              |              |              |
| Phase 1   |              |              |              |
| Phase 2   |              |              |              |
| Phase 0   |              |              |              |
| 776       | 12%          | 23%          | 63%          |
| 410       | 18%          | 18%          | 63%          |
| 130       | 23%          | 15%          | 60%          |
| 1110      | 22%          | 23%          | 53%          |
| 1043      | 23%          | 23%          | 53%          |
| 708       | 19%          | 21%          | 58%          |
| 904       | 24%          | 23%          | 51%          |
| 511       | 28%          | 26%          | 45%          |
| 349       | 27%          | 27%          | 44%          |
| 283       | 28%          | 27%          | 44%          |
| 272       | 28%          | 27%          | 44%          |
| 493       | 28%          | 23%          | 47%          |
| 499       | 28%          | 23%          | 47%          |
| 403       | 25%          | 24%          | 50%          |
| 126       | 26%          | 26%          | 47%          |
| 55        | 29%          | 28%          | 42%          |
| 19        | 25%          | 20%          | 54%          |
| 883       | 23%          | 24%          | 51%          |
| 217       | 22%          | 30%          | 46%          |
| 195       | 30%          | 23%          | 46%          |
| 321       | 26%          | 26%          | 46%          |
| 748       | 19%          | 20%          | 60%          |
| 836       | 24%          | 23%          | 51%          |
| 533       | 20%          | 24%          | 55%          |
| 595       | 26%          | 29%          | 43%          |
| 604       | 28%          | 29%          | 42%          |
| 599       | 28%          | 30%          | 41%          |
| 246       | 31%          | 23%          | 45%          |
| 172       | 26%          | 31%          | 42%          |
| 301       | 28%          | 28%          | 43%          |
| 272       | 29%          | 25%          | 44%          |
| 120       | 28%          | 29%          | 42%          |
| 495       | 29%          | 25%          | 44%          |
| 548       | 30%          | 22%          | 47%          |
| 148       | 21%          | 26%          | 52%          |
| 9 34%      | 16%          | 34%          |
| 0 41%      | 26%          | 31%          |
| 2 24%      | 28%          | 47%          |
| 111 28%   | 27%          | 43%          |
| 51 23%    | 32%          | 43%          |
| 17 25%    | 31%          | 43%          |
| 3 73%     | 18%          | 8%           |
| 56 35%    | 24%          | 40%          |
| 12 19%    | 40%          | 40%          |
| 5 25%     | 30%          | 44%          |
| 17 13%    | 22%          | 64%          |
| 36 21%    | 25%          | 53%          |
| 54 25%    | 27%          | 47%          |
| 27 26%    | 30%          | 42%          |
| 0 24%     | 30%          | 44%          |
| 50 20%    | 21%          | 57%          |
| 12 30%    | 14%          | 54%          |
| 79 42%    | 33%          | 24%          |
| 93 29%    | 32%          | 38%          |
| 5 33%     | 18%          | 47%          |
| 30 31%    | 30%          | 37%          |
| 112 27%   | 27%          | 44%          |
| 52 29%    | 23%          | 46%          |
| 14 21%    | 25%          | 53%          |
| 84 17%    | 15%          | 66%          |
| 9 40%     | 20%          | 39%          |
| 26 29%    | 25%          | 45%          |
| 24 30%    | 35%          | 34%          |
| 1 30%     | 7%           | 61%          |
| 141 33%   | 19%          | 47%          |
| 18 32%    | 20%          | 42%          |
5 Model fit

Model fit was assessed by computing the harmonic average of the likelihood in a trace, after the burn-in period, denoted as $H$. In the limit, the statistic $H$ is proportional to the so-called marginal likelihood, and the ratio of $H$ between two models may serve as a proxy for the Bayes factor (Newton and Raftery, 1994). The distribution of the statistic was computed from the 100 MCMC traces, in order to interpret the significance of differences seen between $H$ of different models. In one set of experiments, we selected the rate variation model with a trifurcating LECA. As Figure 9 shows, multiple gain categories do not improve the models dramatically. Multiple loss categories give significant improvements, but the returns are diminishing beyond 4 categories. The most likely root resolution ($\text{CU,P}$) is the bifurcation between plants and the other two supergroups, but the average likelihoods do not differ more than between the trifurcating traces.

**Figure 9** Model fit for different models and phylogeny rootings. The X axis plots the natural logarithm of $H$ for different models, computed from 500 subsampled MCMC steps. Rate variation models are denoted by $L_xG_y$ for $x$ loss and $y$ gain categories. The shaded area shows the empirical cumulative distribution function (cdf) of $H$ in 100 traces with the $L4G1$ model chosen for the ancestral inference. Alternative model scores are projected onto the cdf for establishing significance levels: for instance, $L5G1$ has a score $H$ surpassing 64 of 100 $L4G1$ traces. Different LECA rooting models (tri is the rooting chosen for the inference) are denoted as shown below.
6 Ancestral inference

Ancestral density and confidence intervals. The plots show the posterior distribution of the ancestral intron density inferred from the sampling chains. On each plot, the horizontal red line shows the median (by the dot) and the 95% (±47.5%) credible interval around it estimated from 50000 subsampled MCMC steps. The Dollo parsimony and maximum-likelihood (ML) reconstructions are also shown.

Number of ancestral introns in different reconstructions. The plots compare the number of ancestral introns inferred by MCMC sampling (median in 50000 subsampled steps) with the inference by Dollo parsimony, and by the maximum-likelihood (ML) model found during sampling.

Lineage-specific loss and gain rates ($\bar{\mu}_{uv}$ and $\bar{\lambda}_{uv}$). On both plots, the X axis shows the median rate seen in the 50000 MCMC-sampled steps. The Y axis plots the rate in the maximum-likelihood model. The two models do not differ much in their parameters.
Ancestral presence by Dollo and MCMC

\[ y = 1.0131x \]

\[ R^2 = 0.79368 \]

Ancestral presence by ML and MCMC

\[ y = 1.0138x \]

\[ R^2 = 0.96474 \]
7 Simulations

The consistency of different inference methods was assessed by generating 100 random simulated data sets using the median model inferred from the MCMC sampling (same sampling settings as before) on the main data set (“Same model”). In particular, we generated 7200 random intron absence-presence profiles, and performed ancestral reconstruction using Dollo parsimony, 100 independently run MCMC samples, and the maximum likelihood model found using sampling. The reconstruction was compared to the known ancestral intron counts.

In order to assess the methods’ robustness, we performed 100 additional simulations using a “Mixed model,” in which model parameters were multiplied by randomly drawn random variables. At each intron site, the root presence probability ($\pi$), as well as the site-specific rate multipliers ($\gamma_j, \nu_j$) were multiplied by random lognormal variables $\ln N(0, \sigma^2)$ with $\sigma = 0.2$ for $\pi$, $\sigma = 0.1$ for the gain rate multiplier $\nu_j$, and $\sigma = 0.3$ for the loss rate multiplier $\gamma_j$. Concomitantly, the edge length parameters were also multiplied independently (among themselves, and across sites) by exponentially distributed random variables with mean 1.

In a final set of experiments, we assessed inference robustness with respect to missing orthologs. Using 100 simulated sets, we replaced randomly chosen entries with ambiguity. Namely, we split the sites into blocks of 34 (average number of sites per gene in the main data set), and erased presence-absence information independently within each block by 13% probability.

We chose the number of simulated intron-bearing sites in each set of experiments to yield similar number of informative entries as the main data set (containing 14% ambiguous entries), generating fewer sites without ambiguity (7200 sites for “Same model” and “Mixed model”), and to 9000 sites with “Missing orthologs” (since some sites without unambiguous intron presence are filtered out in the inference).

Figure 13 Estimation error in simulated data at some key nodes. Each point on the plots shows the actual number of ancestral introns along the X axis, and the estimation error (actual number subtracted from inferred number) along the Y axis. The 25% error level is shown by the slanted dotted lines. The MCMC and ML procedures have very small empirical bias, but Dollo parsimony is typically completely off target.
8 Nægleria gruberi

We identified 68 homolog groups represented in the *Nægleria gruberi* genome (Fritz-Laylin *et al.*, 2010), using bi-directional best BLAST hits in the same way as for the main data set (§1.2). For groups with multiple Nægleria paralogs, we selected the one with the highest BLAST score. The 68 selected Nægleria genes were aligned to the multiple alignment profile constructed for their respective homolog groups (without recomputing the alignment between the 99 eukaryotes of the original data set). The alignments scored aligned intron positions alongside protein sequences, as described before (Csürös *et al.*, 2007). The intron presence-intron table was computed from the alignments in the same way as for the main data set (§1.4). The resulting table comprises 4812 intron sites with 141 ambiguous Nægleria entries.

The data set contains 97 aligned Nægleria introns, from which 57 fall into a homologous position with at least one other eukaryote. Twenty-three of those are inferred to have been absent at the trifurcating LECA (average posterior presence probability 9%), and 34 are inferred to have been present (average probability 95%). At the chromalveolate and unikont ancestors, the numbers are the same. The shared Nægleria intron counts do not differ significantly across deep-branching unikonts (Capsaspora, Dictyostelium, Allomyces), either. Two introns are inferred to have been inserted in parallel between Ngru and the Metazoan ancestor. At the common ancestor of plants and green algæ, only 26 Nægleria introns are inferred to have been present.

Nægleria introns projected onto the data set. The table lists the number of introns shared between Naegleria and other eukaryotic taxa. The table rows list the intron-sharing patterns between the Ngru and other taxa *t* (“Taxon” column). “Shared” sites have introns in both taxa, “Ngru-specific” sites have introns in Nægleria, opposing absent and ambiguous entries in *t*. “taxon-specific” sites have introns in *t* but not in Ngru. Notice that the counts include only unambiguously aligned sites, and, thus, they may not sum up to as expected. For instance, only 88 of the 97 Ngru introns can be unambiguously projected onto homologous Dictyostelium sequences, and 25 of those correspond to a splice site in at least one Dictyostelium genome, the remaining 63 are “Ngru-specific.”

Table 14

| Taxon | Shared | Ngru-specific | t-specific |
|-------|--------|--------------|-----------|
| Capsaspora | 57     | 34           | 0         |
| Dictyostelium | 57     | 34           | 0         |
| Allomyces | 57     | 34           | 0         |

26
| Taxon | introns | shared | Ngru-specific | taxon-specific |
|-------|---------|--------|---------------|---------------|
| Ngru *Naegleria gruberi* | 97 | | | |
| 99 eukaryotes | 4772 | 57 | 40 (41%) | 4574 (96%) |
| Dictyostelium | 127 | 25 | 63 (65%) | 97 (76%) |
| Metazoa | 1767 | 40 | 57 (59%) | 1676 (95%) |
| Hsap [*Homo sapiens*] | 530 | 32 | 64 (66%) | 485 (92%) |
| Cowc [*Capsaspora owczarzaki*] | 217 | 25 | 64 (66%) | 186 (86%) |
| Choanoflagellates Mbre+Prsp | 567 | 25 | 68 (70%) | 528 (93%) |
| Amac [*Allomyces macrogynus*] | 132 | 24 | 65 (67%) | 104 (79%) |
| Chytridiomycota Bden+Spun | 378 | 28 | 66 (68%) | 343 (91%) |
| Fungi | 1454 | 41 | 56 (58%) | 1375 (95%) |
| Unikonts | 3160 | 51 | 46 (47%) | 3013 (95%) |
| Alveolates | 555 | 24 | 69 (71%) | 517 (93%) |
| Chromalveolates | 997 | 31 | 66 (68%) | 935 (94%) |
| Atha [*Arabidopsis thaliana*] | 460 | 25 | 59 (61%) | 426 (93%) |
| Embryophyta [land plants] | 543 | 27 | 69 (71%) | 505 (93%) |
| Viridiplantae | 1483 | 30 | 67 (69%) | 1420 (96%) |
 References

Csürös, M., Rogozin, I. B., and Koonin, E. V. (2008). Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum likelihood approach. *Molecular Biology and Evolution, 25*(5), 903–911.

Csürös, M. (2008). Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics, 24*(13), 1538–1539.

Csürös, M., Holey, J. A., and Rogozin, I. B. (2007). In search of lost introns. *Bioinformatics, 23*(13), i87–i96.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792–1797.

Fritz-Laylin, L. K. *et al.* (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell, 140*, 631–642.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology, 52*(5), 696–704.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L. J., and Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research, 38*, D190–195.

Newton, M. E. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B, 56*(1), 3–48.

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., and Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology, 13*, 1512–1517.

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., and Koonin, E. V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Briefings in Bioinformatics, 6*(2), 118–134.