Origin of introns — early or late?

SIR — Kersanach *et al.*[1] assert that the existence of five spliceosomal introns in identical positions in nuclear genes encoding glyceraldehyde-3-phosphate dehydrogenases (GAPDHs) of eubacterial (chloroplast *GapA/B*) and uncertain (cytosolic *GapC*) ancestry provides "strong evidence in favour of the 'introns-early' hypothesis" for the origin of spliceosomal introns, and thus "lends strong support to the exon theory of genes". Two of these five identical intron positions were already known in 1988 (refs 2, 3) and led us to comment on their evolutionary significance in 1991 (ref. 4). Despite the new data, we still consider that all of these shared positions represent "parallel insertion of different introns"[4]. Our rationale, part of which we elaborate below, is fourfold: (1) these five GAPDH introns have a very limited phylogenetic distribution, as do most of the numerous introns in this gene; (2) there are likely to be preferred sites for intron insertion (the proto-splice site[5], which is evident at four of the shared GAPDH intron positions), and thus the likelihood of parallel insertion is probably much higher than Kersanach *et al.*[1] calculate; (3) many of the putatively ancient GAPDH exons (averaging only 7.2 codons in a hypothetical ancestral GAPDH gene containing all 47 known introns) as well as some present-day GAPDH exons (as small as five nucleotides!) are in our view too tiny to be useful in ancient gene assembly as posited by the exon theory of genes[6]; (4) common ancestry would require the relatively recent, massive and coincidental loss of introns from many eubacterial lineages.

Kersanach *et al.*[1] in their Fig. 1 fail to include numerous GAPDH genes that do not contain introns, including all of the many sequenced protistan and eubacterial genes. The protistan lineages are critical to our understanding of eukaryotic diversity; in particular, *Giardia* and *Trypanosoma* represent deeply branching lineages for which no spliceosomal introns have been reported (ref. 4, and our unpublished data). When all examined GAPDH genes are considered (see figure), it becomes apparent that each of the five 'shared' introns is present only in two highly restricted and distantly related lineages. Thus, the possibility must be considered that these introns arose independently and recently in each specific lineage, as opposed to being homologous and ancient, and lost pervasively throughout eukaryotes and prokaryotes. In fact, virtually all the 47 known GAPDH introns are very limited in their phylogenetic distribution (see Fig. 1 of ref. 1).

Acceptance of these five 'shared' introns as homologous and ancient forces one to postulate their retention for much or all of eubacterial evolution before the endosymbiotic origin of plastids, as well as their subsequent independent loss from many eubacterial lineages in parallel[4]. By extension, this 'introns-early' view requires that the estimated thousands of other introns in nuclear genes of eubacterial ancestry survived for most of eubacterial history only to be extinguished recently and completely from all eubacteria. We believe a more plausible explanation is that nuclear genes of plastid and mitochondrial ancestry acquired all their introns subsequent to their transfer from the organelles to the nucleus. Consistent with this proposal, indeed indicative of a post-endosymbiotic origin of all spliceosomal introns, we find lineage-specific covariance between intron numbers in nuclear genes of eukaryotic and organellar/eubacterial ancestry. In conclusion, the data in ref. 1 are best explained by the late, lineage-specific acquisition of introns rather than primordial intron retention. The introns in question are not, in our view, homologous, but instead are the result of recent, independent parallel insertion into conserved target sites.

John M. Logsdon Jr
Jeffrey D. Palmer
Department of Biology,
Indiana University,
Bloomington,
Indiana 47405, USA

	25	28	29	30	46
Zea GapC1	+	–	+	–	–
Zea GapC4	+	–	+	–	+
Pisum GapC1	+	–	+	–	+
Arabidopsis GapC	+	–	+	–	–
Chlamydomonas GapC	–	–	+	–	–
Chondrus GapC	–	–	–	–	–
Gracilaria GapC	–	–	–	–	–
Homo	–	+	–	–	–
Gallus	–	+	–	–	–
Drosophila m Gap1,2	–	–	–	–	–
Drosophila h	–	–	–	–	–
Schistosoma	–	–	–	–	–
Caenorhabditis e gpd1,4	–	–	–	+	–
Caenorhabditis e gpd2,3	–	–	–	–	–
Caenorhabditis b gpd2,3	–	–	–	–	–
Ustilago	–	–	–	–	–
Agaricus GPD1,2	–	–	–	–	–
Phanerochaete	–	–	–	–	–
Schizophyllum	–	–	–	–	–
Podospora	–	–	–	–	–
Claviceps	–	–	–	–	–
Cryphonectria	–	–	–	–	–
Glomerella	–	–	–	–	–
Curvularia	–	–	–	–	–
Cochliobolus	–	–	–	–	–
Aspergillus	–	–	–	–	–
Kluyveromyces	–	–	–	–	–
Zygosaccharomyces	–	–	–	–	–
Saccharomyces Gap1,2,3	–	–	–	–	–
Phytophthora	–	–	–	–	–
Dictyostelium	–	–	–	–	?
Entamoeba	–	–	–	–	?
Trypanosoma b gapG	–	–	–	–	–
Trypanosoma c gapG	–	–	–	–	–
Leishmania gapG	–	–	–	–	–
Trypanosoma b gapC	–	–	–	–	–
Leishmania gapC	–	–	–	–	–
Trichomonas	–	–	–	–	–
Giardia	–	–	–	–	?
Thermotoga	–	–	–	–	?
Thermus	–	–	–	–	–
Bacillus c	–	–	–	–	–
Bacillus m	–	–	–	–	–
Bacillus s	–	–	–	–	–
Corynebacterium	–	–	–	–	–
Clostridium	–	–	–	–	–
Streptococcus	–	–	–	–	–
Rhodobacter	–	–	–	–	–
Zymomonas	–	–	–	–	–
Pseudomonas	–	–	–	–	–
Escherichia c gapA	–	–	–	–	–
Escherichia c gapA (13)	–	–	–	–	?
Escherichia b gapA (3)	–	–	–	–	?
Escherichia h gapA (2)	–	–	–	–	?
Escherichia v gapA (3)	–	–	–	–	?
Escherichia f gapA (3)	–	–	–	–	?
Salmonella t gapA	–	–	–	–	?
Salmonella sp. *gapA* (15)	–	–	–	–	?
Citrobacter gapA	–	–	–	–	?
Enterobacter gapA	–	–	–	–	?
Klebsiella gapA	–	–	–	–	?
Serratia m gapA	–	–	–	–	?
Serratia o gapA	–	–	–	–	?
Escherichia c gapB	–	–	–	–	–
Escherichia c gapC	–	–	–	–	?
Anabaena gap1	–	–	–	–	–
Anabaena gap2	–	–	–	–	–
Anabaena gap3	–	–	–	–	–
Chondrus GapA/B	–	–	–	–	–
Gracilaria GapA/B	–	–	–	–	–
Chlamydomonas GapA/B	+	–	+	–	–
Zea GapA	–	–	–	+	–
Arabidopsis GapA	–	–	–	+	–
Pisum GapA	–	–	–	+	–
Arabidopsis GapB	–	+	–	+	+
Pisum GapB	–	+	–	+	+

Distribution of the five shared introns among the 116 GAPDH genes of known exon/intron structure. Plus sign, presence of an intron; minus, absence of intron; ?, missing data. Intron positions are numbered according to ref. 1. Genes are organized from top to bottom by major phylogenetic groups: plants/algae, animals, fungi, protists, eubacteria and nuclear genes of plastid/eubacterial ancestry. All sequences are available from GenBank except for *Dictyostelium* (from M. Smith and R. Doolittle) and *Gracilaria* (from Y.-H. Zhou and M. Ragan). Multiple GAPDH sequences from the same eukaryotic species are grouped on the same line if they have the same genomic structure, except for the anciently diverged *gapC* (cytoplasmic) and *gapG* (glycosomal) of trypanosomatids. Multiple GAPDH genes from the same eubacterial species are shown on separate lines because they represent ancient duplications. Numbers in parantheses indicate numbers of isolates within a eubacterial species whose orthologous (*gapA*) genes have been sequenced. Multiple species of the same genus are distinguished by the first letter of the species name. The *Chlamydomonas*, *Chondrus* and *Gracilaria* nuclear-encoded plastid GAPDH genes are denoted *GapA/B* because their relationships to land plant *GapA* and *GapB* genes are unclear.

SIR — "Lightning never strikes the same place twice", declares a common, but incorrect, aphorism. If lightning struck with equal probability at every possible location, multiple events would indeed be vanishingly rare. However, chances vary from site to site, so that some sites are struck repeatedly, and others not at all.

Kersanach *et al.*[1] have found five matching intron positions between the ten positions known among genes of the *gapAB* family (eukaryotic nuclear genes