Letters

Footprints of primordial introns on the eukaryotic genome

A recent paper in this journal[1] presented an analysis of PHASE DISTRIBUTION OF INTRONS (see Glossary) in *Caenorhabditis elegans* with respect to the EXON THEORY OF GENES. This theory, also known as introns-early, contains two specific postulations about the phase distribution. First, the observed excess of PHASE-ZERO INTRONS in ANCIENT EUKARYOTIC GENES is at least in part due to the presence of ancient phase zero introns. Second, the observed excess of SYMMETRIC EXONS in these genes is the legacy of ancient intron shuffling. Wolf *et al.*[1] argued that the ancient eukaryotic genes do not have higher proportions of phase-zero introns or of symmetric exons than do genes recently transferred from prokaryotes, and hence suggested that there was no detectable signal of ANCIENT INTRONS. However, an error in their computer programs led to their data being presented wrongly, and they have published a new data table as an erratum[2]. This new table shows that their earlier conclusion was incorrect and that the phase distribution of introns in 'recent' genes compared with 'ancient' genes does indeed vary in the ways predicted by the exon theory of genes: ancient genes do have higher fractions of phase-zero introns and symmetric exons than do recent ones. Here, we show this reanalysis and, furthermore, provide new evidence for the existence of ancient phase-zero introns: those intron positions identified as putatively ancient by virtue of a wide phylogenetic distribution lie preferentially in phase zero.

Several years ago, Long *et al.*[3] argued that the skewed phase distribution (more introns in phase zero than in phase one or two) in ancient conserved genes could be best explained by the hypothesis that the original genes were created by exon shuffling, primarily using introns lying in phase zero. They observed further that there was a significant excess of symmetric exons (which could be shuffled into an intron without destroying the reading frame) and argued that this too was a signal of exon shuffling in the last universal common ancestor (LUCA). Since then, work on such ancient conserved genes has found a correlation between phase-zero introns and compact elements of the protein structure (called modules)[4–6], supporting the notion that some fraction of the current phase-zero intron positions might be primordial, whereas the rest of the intron positions (including most phase-one and phase-two positions) might be more recent[5,6].

The question of the origin of the spliceosomal introns has often been seen in the extremes of an 'all introns early'[7,8] versus 'all introns late'[9–11] debate. However, recent work[5,6] stresses a mixed model, where many introns were created in the course of evolution, possibly by gene or exon duplication followed by a splicing event that converts the duplicated region into a pair of exons separated by a new intron[12]. Nevertheless, this mixed model does assert that some fraction of the phase-zero introns is primordial and that the original genes were assembled by exon shuffling.

The paper by Wolf *et al.*[1] challenged the notion of ancient introns by attempting to identify in the *C. elegans* genome a group

Glossary

Ancient genes: Eukaryotic genes shared across the prokaryotes, hence thought to be ancestral and to date back to the last universal common ancestor (LUCA).
Ancient introns: Introns dating back to the last universal common ancestor (the existence of which is being debated here). The exon theory of genes holds that these introns facilitated the modular assembly of complex genes from small exons, and that most or all lie in phase zero.
Exon theory of genes: The notion that there were (mostly phase-zero) ancient introns and that these introns were integral in assembling the first genes through recombination.
PAM and BLOSUM matrix scores: Entries in the substitution matrices used by pair-wise sequence comparison programs (notably BLAST).
Phase distribution of introns: the patterns of intron positions in a gene or set of genes with respect to their positions within codons (between codons or after the first or second nucleotide of a codon).
Phase-zero introns: Introns located between codons, thus ensuring the integrity of the open reading frame. Phases one and two begin after the first and second bases of codons, respectively.
Symmetric exons: Exons flanked by introns in the same phase that, therefore, could be shuffled into an intron without destroying the reading frame.

Table 1. Intron phase distributions and exon symmetry in different classes of *Caenorhabditis elegans* genes

	No. of genes	Intron data available	Introns/ gene	Average density	Phase 0 introns		Phase 1 + 2 introns		Total	Symmetric internal exons		Non-symmetric internal exons		Total internal exons
'Nematode-specific'	8901	6798	4.1	1.400	12621	45.1%	15391	54.9%	28012	8386	39.5%	12828	60.5%	21214
'Eukaryotic'	5714	4217	5.8	1.375	10546	43.5%	13715	56.5%	24261	7680	38.3%	12364	61.7%	20044
'Recent bacterial'	185	156	6.2	1.343	376	39.2%	584	60.8%	960	261	32.5%	543	67.5%	804
'Recent archael'	13	8	5.1	1.547	17	41.5%	24	58.5%	41	14	42.4%	19	57.6%	33
'Recent uncertain'	44	38	4.6	1.319	81	46.0%	95	54.0%	176	47	34.1%	91	65.9%	138
Total 'recent'		202	5.8		474	40.3%	703	59.7%	1177	322	33.0%	653	67.0%	975
'Ancient bacterial'	2395	1771	6.1	1.273	4840	44.5%	6047	55.5%	10887	3452	37.9%	5664	62.1%	9116
'Ancient archael'	430	328	4.9	1.230	690	42.6%	931	57.4%	1621	503	38.9%	790	61.1%	1293
'Ancient uncertain'	894	671	6.2	1.215	1968	47.4%	2187	52.6%	4155	1388	39.8%	2096	60.2%	3484
Total 'ancient'		999	5.8		2658	46.0%	3118	54.0%	5776	1891	39.6%	2886	60.4%	4777

The tabular material is from Ref. 2. The genes were taken from the WormPep18 dataset and the intron data from the Sanger center. The intron density is introns/100 codons. The totals of introns in the 'recent' and 'ancient' set is given and the relevant percentages are in boldface. For phase-zero introns, $\chi^2 = 13.0$, $P = 0.0003$; for symmetric internal exons; $\chi^2 = 14.7$, $P = 0.00012$.

  PII: S0168-9525(01)02375-7

Box 1. Controversy over the sisterhood of Eukarya and Archaea

The proposal of the sisterhood of eukarya and archaea, based largely on trees constructed from rRNA sequences[a], has been recently called into question by contradictory evidence [reviewed in Ref. b]. For instance, Koonin *et al.* found that 44% of *Methanococcus jannaschii* genes had their closest homologue in bacteria, whereas only 13% had their closest homologue in eukaryotes, and they suggested that the origin of archaea might lie in a fusion of eukarya and bacteria [Ref. c, reviewed in Ref. d]. Such a model of evolution suggests a different interpretation of the data studied here. In this model, the 'ancient uncertain' and 'ancient bacterial' groups are the most convincingly 'ancient' (meaning vertically inherited from the LUCA), because the presence of a gene in both bacteria and eukarya would suggest presence in the LUCA, whereas presence in only archaea and eukarya would not. In this case ('recent' versus 'ancient bacterial' and 'ancient uncertain'), the phase-zero bias is 45.3% versus 38.4% and the symmetric exon bias is 39.6% versus 33.0%. χ^2 is 11.0 and 11.1, respectively; both *P* values are less than 0.001. This evolutionary model could also explain the failure of Wolf *et al.* (this issue) to find a difference between the intron pattern in 'ancient bacterial' and 'ancient archaeal' genes.

Other models propose a clade containing bacteria and archaea[d,e] (in which the 'ancient uncertain' would be the most convincingly ancient and, again, one would not expect a difference between 'ancient bacterial' and 'ancient archaeal'). In this case, the phase-zero bias is 47.4% versus 40.3% and the symmetric exon bias is 39.8% versus 33.0%. χ^2 is 18.6 and 15.0, respectively; *P* values are 0.00002 and 0.0001.

References

a Woese, C.R. *et al.* (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–4579

b Forterre, P. and Philippe, H. (1999) Where is the root of the universal tree of life? *BioEssays* 21, 871–879

c Koonin, E.V. *et al.* (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin of the archaea. *Mol. Microbiol.* 25, 619–637

d Doolittle, W.F. and Logsdon, J.M., Jr. (1998) Do archaea have a mixed heritage? *Curr. Biol.* 8, R209–R211

e Gupta, R.S. (1998) What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Mol. Microbiol.* 29, 695–707

of prokaryotic genes that appear to have been transferred recently. They then compared intron phase distribution patterns for these recent genes to those of a set that they identified as ancient. They analyzed a total of 18 576 *C. elegans* genes, 13 987 of which had intron data, using BLASTP to find homologues. They identified 6798 of the intron-containing genes as 'nematode-specific' (i.e. no BLAST match to any other organism), 4217 as 'eukaryotic' (i.e. some eukaryotic matches but no match to prokaryotic organisms), 202 genes as 'recent transfers' from prokaryotes (i.e. having prokaryotic hits ten orders of magnitude more significant than the non-nematode eukaryote hits, or having prokaryotic but not eukaryotic hits), 1771 as 'ancient bacterial' (i.e. having eukaryotic hits as well as a bacterial hit that is ten orders of magnitude more significant than the best archaeal hit, or having no archaeal hit), 328 as 'ancient archaeal' (i.e. having eukaryotic hits as well as an archaeal hit that is ten orders of magnitude more significant than the best bacterial hit, or having no bacterial hit), and the rest, 671, as 'ancient uncertain'.

The ancient archaeal and ancient uncertain groups are interpreted by Wolf *et al.*[1] as vertically inherited through the 'standard model' of evolution; the ancient bacterial group as transfers from bacteria before the nematode radiation; and the recent groups as transfers after the nematode radiation. We consider first the two best-defined groups, those of recent transfers and ancient (i.e. ancient archaeal and ancient uncertain, leaving aside the problematic ancient bacterial for now).

The introns in the recent transfers appear at about the same density as in all the other genes. However, the details of the intron distributions are different. Table 1 shows the corrected data for these *C. elegans* genes[2]. First, the introns in the ancient genes show higher phase-zero bias than the recent ones: 46.0% versus 40.3%; second, the ancient genes have a greater excess of symmetrical exons than the recent ones: 39.6% versus 33.0%. These effects are not only in the predicted direction; they are highly significant with χ^2 values of 13.0 and 14.7, respectively (*P* values ~0.0001 that the recent patterns are different from the ancient). Thus, both these findings distinguish the intron pattern in the recently transferred genes from those found in all the other genes of *C. elegans* and support the suggestion that there are traces of ancient introns in modern eukaryotes.

What, then, of the ancient bacterial group? This group contains *C. elegans* genes with hits in eukaryotes and bacteria, in which the bacterial hit is ten orders of magnitude stronger than the best archaeal hit (or in which there is no archaeal hit). The authors assert that this is the pattern expected of ancient transfers from bacteria to eukaryotic ancestors of worms. Even leaving aside the fact that this accepts as fact a still-debated view of ancient evolution (the sisterhood of Archaea and Eukarya, see Box 1), there are problems with this interpretation. In this model, this is not

Table 2. Phase patterns of introns in *Caenorhabditis elegans* genes

	Phase zero	χ^2	Phase one	Phase two	Total introns
Exclusively animal *C. elegans* positions	663 50.8%		306 23.4%	337 25.8%	1306
Widely distributed *C. elegans* positions	145 62.2%	10.4	46 19.7%	42 18.0%	233

Widely distributed signifies that an intron is also found at that position in a non-animal copy of the gene. The introns are from the database of 6568 intron positions in the set of ancient genes that have three-dimensional structures for the corresponding proteins compiled by Fedorov and collaborators (unpublished). The χ^2 is calculated comparing phase zero with (phase one + phase two).

Table 3. Phase patterns of introns in different kingdoms[a]

	Phase zero	χ^2	Phase one	Phase two	Total introns
Animal					
Exclusively animal	1607 50.2%		860 26.9%	735 23.0%	3202
Animal and other kingdom(s)	329 59.8%	17.4	127 23.1%	94 17.1%	550
Exclusively vertebrate	504 48.2%		324 31.0%	217 20.8%	1045
Vertebrate and other kingdom(s)	194 61.8%	17.8	69 22.0%	51 16.2%	314
Exclusively invertebrate	906 51.6%		423 24.1%	428 24.4%	1757
Invertebrate and other kingdom(s)	246 58.9%	7.2	96 23.0%	76 18.2%	418
Plant					
Exclusively plant	1166 59.9%		398 20.5%	381 19.6%	1945
Plant and other kingdom(s)	323 66.1%	6.1	99 20.2%	67 13.7%	489
Fungi					
Exclusively fungi	303 42.7%		236 33.3%	170 24.0%	709
Fungi and other kingdom(s)	118 54.1%	8.7	56 25.7%	44 20.2%	218
Protist					
Exclusively protist	34 38.6%		28 31.8%	26 29.5%	88
Protists and other kingdom(s)	41 56.9%	5.3	16 22.2%	15 20.8%	72
All					
Only one kingdom	3110 52.3%		1522 25.6%	1312 22.1%	5944
Two or more kingdoms	371 60.4%	14.7	142 23.1%	101 16.4%	614
Three or more kingdoms	51 64.6%		13 16.5%	15 19.0%	79

The introns are from the database of 6568 intron positions in the set of ancient genes that have three-dimensional structures for the corresponding proteins compiled by Fedorov and collaborators (unpublished). The χ^2 is calculated comparing phase zero with (phase one + phase two).

only the pattern expected of true ancient transfers – it is also expected of cases of archaeal loss (e.g. the substantial non-orthologous gene displacement found in *Methanococcus jannaschii* by Koonin *et al.*[13]) or rapid evolution down the archaeal line. Even more strikingly, it is also the pattern expected of transfers *from* eukaryotes *to* bacteria. To try to address these concerns, the original authors identified 727 experimentally characterized mitochondrial proteins encoded by nuclear genes from different eukaryotes and compared them with the *C. elegans* genes. They found that about half of these 727 genes had their closest matches in the ancient bacterial group, supporting the claim that some of the members of that group are, in fact, the result of transfers from bacteria or organelles to eukaryote genomes.

However, this says nothing of the 1400 or so genes in the ancient bacterial group that are not matched by these mitochondrial genes. This result only supports the notion that eukaryotic transfers from mitochondria are found in the ancient bacterial group. (This is not surprising, as one intuitively expects these genes to fall in the ancient bacterial group.) It does not support the notion that all or even most of the members of this group are the results of ancient transfers. In fact, both the fraction of phase-zero introns and the fraction of symmetric exons are intermediate between the values found for the vertically inherited group (ancient archaeal and ancient uncertain) and those found for the recent group. This is exactly the pattern expected if the ancient bacterial group is a mosaic of ancient and transferred genes.

We should step back for a moment to consider whether the methods used by Wolf *et al.*[1] are sufficient for identifying any but the most obvious instances of lateral transfer to eukaryotes. In fact, the whole classification of modern *C. elegans* genes into nematode-specific, eukaryotic, ancient, and recent transfers is problematic. First, even if we concede that the category Wolf *et al.* labeled 'recent transfers' are such, their method cannot differentiate transfers *to* the worm lineage *from* prokaryotes versus transfers *from* the worm lineage *to* prokaryotes. Second, the classification in Wolf *et al.* misuses BLAST. BLAST evaluates the chance that two proteins are related, which is not the same as evolutionary distance. PAM and BLOSUM MATRIX SCORES are based on average substitution rates for a large sample of proteins, which are at best a rough approximation for a given amino acid in a given protein at a given position. Scores are scaled using the rate of occurrence of an amino acid in this sample, a quantity not necessarily relevant to notions of evolutionary distance. Although the strength of BLAST scores is expected to correlate with evolutionary distance, it is not a surrogate for evolutionary distance. Third, any sequence comparison program is limited by only being able to identify relationships if the residues have not changed too much in the course of evolution, and the programs have no way of dissociating nearness and distance notions from the problems of different rates of evolutionary change of homologous proteins down different branches of descent. In general, the assertion that a difference of ten orders of magnitude (40 bits in score) is sufficient evidence for lateral transfer might not be justifiable and awaits further analysis and simulations.

We think it extremely unlikely that only 2770 of the 13 987 nematode genes with introns in the database originated in the LUCA and that 6798 are truly nematode specific (implying that they have no homologues elsewhere). Biologically, it is far more probable that the past evolution of these genes proceeded through duplication followed by divergence, in which case there are full-length homologues elsewhere in the eukaryotes and prokaryotes, or they were created by exon shuffling, in which case there are homologous regions in other genes throughout Nature.

These arguments call into question the identification of any genes as 'recent transfers', unless the scoring were so extreme and the phylogenetic distribution so limited as to be unambiguous. However, the fact that the intron distribution in these recent transfers is actually different from the rest of the genes does support the notion that there is some truth in this distinction. A more restrictive definition of recent transfers might even lead to a greater deviation in intron pattern, still more closely approximating an equal insertion into the three phases.

Another indication of ancient phase-zero introns

Another expected legacy of ancient phase-zero introns is that putatively ancient

introns (those widely phylogenetically distributed) should have a stronger phase-zero bias. We tested this notion by examining widely distributed *C. elegans* intron positions: those present in two or more kingdoms. We used a database of ancient genes with known corresponding protein structures (culled from the Protein Data Bank) and the introns from their homologues, compiled by Fedorov and collaborators (unpublished). Of 6568 total intron positions, *C. elegans* introns are found at 1539 positions within 280 gene families. At 233 of these *C. elegans* positions there are also introns in the copies of the gene from non-animal organisms. Such coincident positions are candidates for the positions of ancient introns, based on this very wide phylogenetic distribution. (These coincident intron positions do, in fact, show a stronger correlation with module boundaries[6].) If these positions are indeed ancient, and if the phase-zero bias of introns is due to the presence of ancient introns, this ancient subset of *C. elegans* introns should have a higher phase-zero bias than should the set of animal-unique *C. elegans* introns.

Table 2 shows that this ancient subset does have a significantly stronger phase-zero bias (62.2% versus 50.8%), with a χ^2 of 10.4 ($P = 0.0013$). This is a striking verification of the expectation that part of the phase-zero bias is due to the presence of truly ancient introns. Table 3 shows further that this is a general phenomenon. There is a highly-significant difference between the phase bias of animal intron positions at which introns are found in two or more kingdoms and those positions that are exclusively animal. The same is true for invertebrates and vertebrates separately, for plants, and for fungi, and there is a significant difference for protists. This trend is seen even further in positions where introns are found in three or more kingdoms. This result suggests that there were in fact ancient intron positions and that the observed excess of phase-zero positions is, at least in part, due to their presence in modern eukaryotes.

These twin findings – first, that the intron phases in ancient genes, defined by BLAST scores, differ from those of recently transferred genes by having a stronger phase-zero bias, and second, that phylogenetically ancient intron positions also show a stronger phase-zero bias – support the theory that introns, mostly in phase zero, were present in the LUCA.

Furthermore, as phase-zero, widely phylogenetically-distributed intron positions are highly correlated with the boundaries of modules[6], the subset of introns possessing one characteristic expected of ancient intron positions – wide phylogenetic distribution – also possesses two other traits expected of ancient positions. This is a result easily explained on a mixed intron-origin model, but not easily explained on strictly insertional ones.

Scott W. Roy*
Benjamin Peter Lewis
Alexei Fedorov
Walter Gilbert
Dept of Molecular and Cellular Biology, Biological Laboratories, Harvard University, Cambridge, MA 02138, USA.
*scottroy@fas.harvard.edu

References

1 Wolf, Y.I. *et al.* (2000) No footprints of primordial introns in a eukaryotic genome. *Trends Genet.* 16, 333–334
2 Wolf, Y.I. *et al* (2001) No footprints of primordial introns in a eukaryotic genome [Author correction]. *Trends Genet*. 17, 146
3 Long, M. *et al.* (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. U. S. A.* 92, 12495–12499
4 de Souza, S.J. *et al.* (1996) Intron positions correlate with module boundaries in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 93, 14632–14636
5 de Souza, S.J. *et al.* (1998) Towards a resolution of the introns early/late debate: only phase-zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5094–5099
6 Roy, S.W. *et al.* (1999) Centripetal modules and ancient introns. *Gene* 238, 85–91
7 Doolittle, W.F. (1978) Genes in pieces: were they ever together? *Nature* 272, 581–582
8 Gilbert, W. (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol*. 52, 901–905
9 Logsdon, J.M., Jr *et al.* (1995) Seven newly-discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci. U. S. A.* 92, 8507–8511
10 Cho, G. and Doolittle, R.F. (1997) Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.* 44, 573–584
11 Logsdon, J.M. Jr. (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev*. 8, 637–648
12 Venkatesh, B. *et al.* (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10267–10271
13 Koonin, E.V. *et al.* (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin of the archaea. *Mol. Microbiol.* 25, 619–637

Footprints of primordial introns on the eukaryotic genome: still no clear traces

Response from Yuri I. Wolf, Fyodor A. Kondrashov and Eugene V. Koonin

The slightly lower fraction of PHASE-ZERO INTRONS (see Glossary) and the slightly lower excess of SYMMETRICAL EXONS over non-symmetrical exons in *Caenorhabditis elegans* genes that are thought to be recent transfers from prokaryotes compared with ANCIENT GENES has been proposed to be evidence for persistence of PRIMORDIAL INTRONS in ancient genes. We show here that there is no significant difference, by both of these criteria, between ancient genes and genes thought to be old horizontal acquisitions of bacterial genes, including those from the progenitors of mitochondria. If primordial introns were indeed detectable in analyses of modern genes, they should have shown up in such a comparison. We conclude therefore that no traces of primordial introns are detectable in the nematode genome.

In their letter (this issue), Roy and colleagues suggest that a re-analysis of our data on intron features in the nematode *Caenorhabditis elegans* genes that appear to have different evolutionary histories is compatible with the notion that some phase-zero introns

Glossary

Ancient genes: Eukaryotic genes that are thought to date back to the last universal common ancestor (LUCA). Typically, these are genes that are conserved in Archaea and Bacteria and appear to fit the standard model of evolution, with distinct archaeo-eukaryotic and bacterial branches.
Phase-zero introns: Introns located between codons. Phases one and two begin after the first and second bases of codons, respectively.
Primordial introns: Introns thought to be inherited from LUCA.
Polyphyletic: Occurring independently in different evolutionary lineages (e.g. polyphyletic gene loss).
Symmetric exons: Exons flanked by introns in the same phase.