

Research article

Open Access

Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing

Manuel Irimia*^{1,2}, Jakob Lewin Rukov³, David Penny¹ and Scott William Roy*¹

Address: ¹Allan Wilson Centre for Molecular Evolution and Ecology, Massey University, Palmerston North, New Zealand, ²Departament de Genètica, Universitat de Barcelona, Barcelona, Spain and ³Department of Molecular Biology, University of Copenhagen, Copenhagen, Denmark

Email: Manuel Irimia* - mirimia@gmail.com; Jakob Lewin Rukov - JLRukov@bi.ku.dk; David Penny - D.Penny@massey.ac.nz; Scott William Roy* - scottwroy@gmail.com

* Corresponding authors

Published: 4 October 2007

Received: 20 December 2006

BMC Evolutionary Biology 2007, **7**:188 doi:10.1186/1471-2148-7-188

Accepted: 4 October 2007

This article is available from: <http://www.biomedcentral.com/1471-2148/7/188>

© 2007 Irimia et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Alternative splicing has been reported in various eukaryotic groups including plants, apicomplexans, diatoms, amoebae, animals and fungi. However, whether widespread alternative splicing has evolved independently in the different eukaryotic groups or was inherited from their last common ancestor, and may therefore predate multicellularity, is still unknown. To better understand the origin and evolution of alternative splicing and its usage in diverse organisms, we studied alternative splicing in 12 eukaryotic species, comparing rates of alternative splicing across genes of different functional classes, cellular locations, intron/exon structures and evolutionary origins.

Results: For each species, we find that genes from most functional categories are alternatively spliced. Ancient genes (shared between animals, fungi and plants) show high levels of alternative splicing. Genes with products expressed in the nucleus or plasma membrane are generally more alternatively spliced while those expressed in extracellular location show less alternative splicing. We find a clear correspondence between incidence of alternative splicing and intron number per gene both within and between genomes. In general, we find several similarities in patterns of alternative splicing across these diverse eukaryotes.

Conclusion: Along with previous studies indicating intron-rich genes with weak intron boundary consensus and complex spliceosomes in ancestral organisms, our results suggest that at least a simple form of alternative splicing may already have been present in the unicellular ancestor of plants, fungi and animals. A role for alternative splicing in the evolution of multicellularity then would largely have arisen by co-opting the preexisting process.

Background

Alternative splicing (AS) of transcripts is common in diverse eukaryotic lineages. By this mechanism, a variety of transcripts and proteins are produced from a single

gene, contributing to increased transcriptome and proteome diversity. AS has been reported in a wide range of eukaryotic groups including plants, apicomplexans, diatoms, amoebae, animals and fungi [1-5]. However, it is

unclear and hard to assess whether this process has arisen independently in the different lineages (as suggested by some authors, e.g. [6]) or whether it was already present in their last common ancestor. The spliceosome, the machinery responsible for the splicing of introns in eukaryotic genes, is ancestral to all extant eukaryotic groups with the last common ancestor possessing a complex machinery, similar to that found in most modern organisms [7]. In addition, we recently argued that eukaryotic ancestors had weak 5' splice site boundary consensus sequences [8], a characteristic that is linked to the presence of AS in modern organisms [6]. These ancestral traits thus allow for the possibility that AS arose early in eukaryotic evolution.

How can we begin to address this issue? If AS arose independently in different lineages, we might expect that different classes of genes would show varying levels of AS in separate lineages, reflecting differential evolutionary histories. In particular, genes with regulatory functions, such as transcription factors [9,10] or signal transducers [11], exhibit high levels of alternative splicing in mammals, consistent with a central role for AS in generating the complexity of mammalian ontology [9], while basic enzymatic functions show less splice variation [9]. By the same reasoning, if AS arose along with the rise of organismal complexity in different multicellular lineages, genes central to this complexity would likely have high AS frequencies, while conserved ancient eukaryotic gene functions might have lower AS frequencies.

Notably, the finding of significant AS in the intron-rich pathogenic unicellular fungus *Cryptococcus neoformans* [3] demonstrates that widespread AS is not restricted to multicellular or highly-differentiated organisms. Indeed, from an evolutionary viewpoint, it is not likely that AS would evolve "in order for" multicellularity to develop; rather, it is possible that AS already existed (in at least a simple form) and was then later co-opted for multicellular development.

Intron/exon structure may be an important determinant for evolution of AS. Genes with more introns have more opportunities for AS. This would be consistent with vertebrates' higher intron numbers and AS frequencies [12]. However, recent results have shown that vertebrate intron number is not particularly high by historical metazoan standards [13-17], and that early eukaryotic ancestors likely harbored relatively high intron numbers [15,18-22].

We studied patterns of AS in 12 well-annotated genomes from plants, fungi and animals. We compared frequencies of AS of genes of different classes according to their gene structure, evolutionary origins, phylogenetic distribution

and functionality. Our major findings include: (i) ancient genes (conserved in both plants and animals/fungi) are equally likely to have known AS as 'newer' genes; (ii) ancient functions are carried by genes that show relatively high levels of AS; (iii) genes found across all lineages (suggesting that they are essential for eukaryotic life) are no more likely to show AS than are genes that have been lost in one or more lineages; and (iv) there is a strong relationship between intron number and the existence of known AS across genes. We interpret our results to support the notion that a potentially widespread AS may have been present at least as early as the unicellular ancestor of animals, fungi and plants.

Results

Intron/exon structure and AS frequency

We found a clear positive relationship between average intron number per gene and occurrence of AS across 12 animal, fungus, and plant species (Figure 1A). This relationship is also seen across genes within each of the eight species with significant frequencies of AS: higher intron number is associated with higher AS levels within each genome (Figure 1B). In particular, there is a steady increase in incidence of AS among genes with up to 6-10 introns. Given estimates of high intron densities in the plant animal ancestor (at least as high as modern *Caenorhabditis* species [15,18,23]), this finding is consistent with frequent ancestral AS.

Age of alternatively spliced genes

The KOG database [24] includes groups of orthologous genes for seven animal, fungus, and plant species. For each of the four species with significant AS levels in the KOG database (*H. sapiens*, *C. elegans*, *D. melanogaster* and *A. thaliana*) we divided gene families into four groups: 1) common (and thus presumably ancestral) to plants, animals, and fungi (PAF); 2) common to fungi and animals (AF); 3) specific to animals (A); 4) specific to a single lineage (LSE) (Figure 2A).

Figure 2B shows the percentage of genes in each group with known alternative splicing for each species. AS was found in all groups. LSE genes showed the lowest frequency of AS in each species (significantly lower than the whole set of genes in fly, worm and *Arabidopsis*, $p < 0.0001$ by Fisher exact tests), and AF and A genes showing the highest frequency. Interestingly, the most ancient group (PAF) showed relatively high levels of AS (significantly higher than the whole set of genes in worms and *Arabidopsis*, $p < 0.0001$ by Fisher exact test, and not significantly different to this set in humans and flies), indicating no constraints against evolution of AS in ancient eukaryotic genes. In particular, we identified 36 KOGs whose genes are highly alternatively spliced in all four species, which could reflect that these gene functions have been alterna-

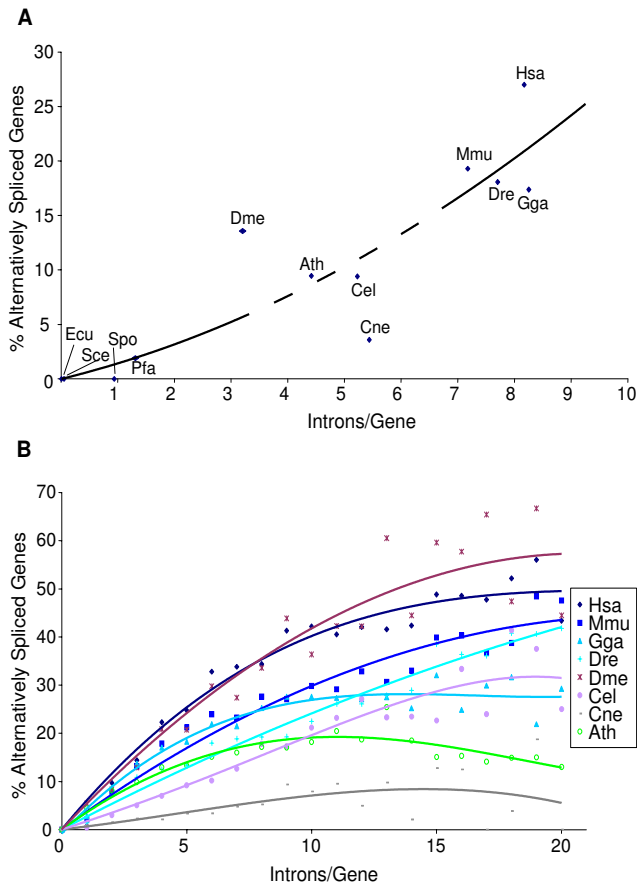


Figure 1
Intron/exon numbers and AS frequency. **A:** Percentage of alternatively spliced genes in different eukaryote genomes vs. the average number of introns per gene. Discontinuous line is an estimated interval for intron density of the ancestor of animals and plants (from 3.5 [18] to 7.0 [15]). **B:** Frequency of AS versus intron numbers per gene for the 8 species showing relatively high values of AS. Abbreviations: Hsa (*Homo sapiens*), Mmu (*Mus musculus*), Gga (*Gallus gallus*), Dre (*Danio rerio*), Cel (*Caenorhabditis elegans*), Dme (*Drosophila melanogaster*), Ath (*Arabidopsis thaliana*), Sce (*Saccharomyces cerevisiae*), Spo (*Schizosaccharomyces pombe*), Ecu (*Encephalitozoon cuniculi*), Pfa (*Plasmodium falciparum*), Cne (*Cryptococcus neoformans*).

tively spliced in the plant/animal ancestor (see Additional file 1).

Gene dispensability and alternative splicing
 Among KOG's shared between *A. thaliana* and animals and/or fungi, we determined AS in 'indispensable' genes (those shared across all seven species in the KOG database) and 'dispensable' genes (absent from one or more opisthokonts); both classes of genes showed high AS levels (Figure 3). Furthermore, no correlation was found between a KOG's PGL (Propensity for Gene Loss, a meas-

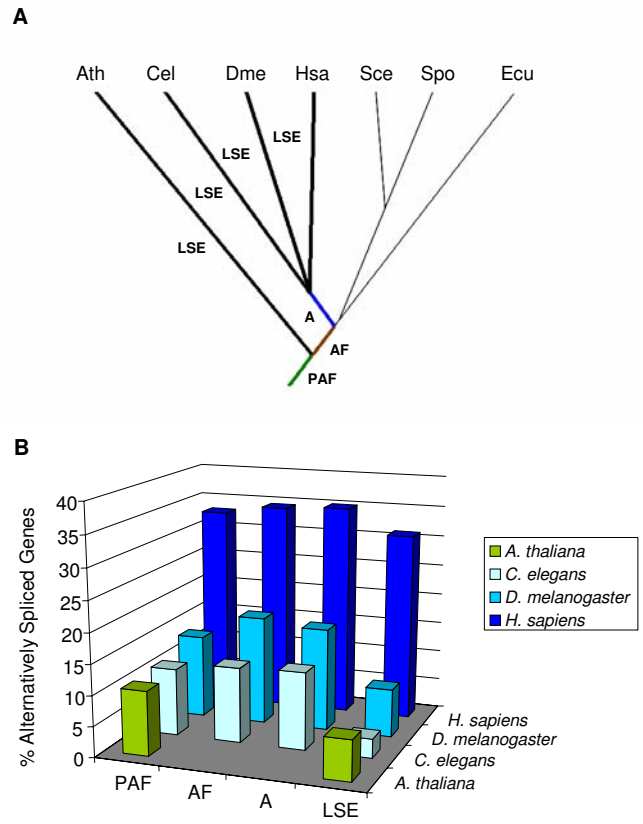


Figure 2
Evolutionary origin of alternatively spliced genes. **A:** Phylogenetic tree showing the relations between the seven species included in the KOG database and used in this study. PAF (green line) corresponds to the group of KOG's likely appeared before the split of animals, plants and fungi; AF (brown line), KOG's likely appeared in the fungamal ancestor; A (blue line), clusters of orthologous likely appeared in the ancestors of nematodes, insects and vertebrates; LSE's (four wide black lines) correspond to lineage specific expansions of plants, nematodes, insects and vertebrates. **B:** Percentage of AS for genes according to phylogenetic origin. PAF: ancestral to plants, animals and fungi. AF: ancestor of animals and fungi. A: animals. LSE: lineage specific expansions. Hsa (*Homo sapiens*), Cel (*Caenorhabditis elegans*), Dme (*Drosophila melanogaster*), Ath (*Arabidopsis thaliana*), Sce (*Saccharomyces cerevisiae*), Spo (*Schizosaccharomyces pombe*), Ecu (*Encephalitozoon cuniculi*). Note that in *A. thaliana* genes can only group into PAF or LSE.

ure of a gene's likelihood to be lost in evolution [25]) and AS in any species (data not shown).

Thus, genes encoding basic and highly conserved cellular functions are no less likely to be alternatively spliced than are other genes (for instance those involved in multicellularity or other complex functions).

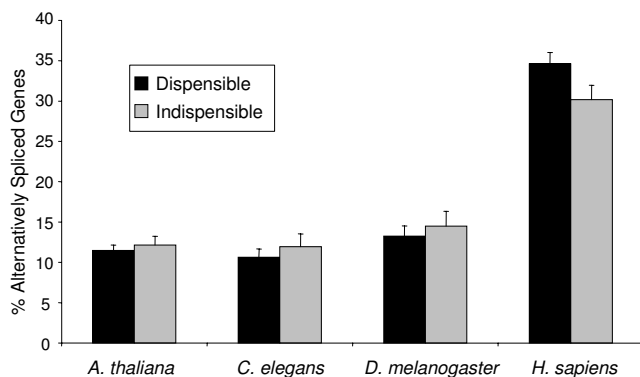


Figure 3
Gene dispensability and alternative splicing. Percentage of alternatively spliced genes according to gene dispensability in evolution. All the gene functions were present in the common ancestor of animals, plants and fungi. Dispensable genes (black): the KOG's to which they belong was lost in at least one of the animal or fungal species included in KOG database. Indispensable genes (grey): KOG's present in the seven studied species.

Cellular location of alternatively spliced genes

The level of AS by cellular location for 6 eukaryotes is shown in Figure 4 and in detail in Additional file 2. Genes for proteins in most cellular locations showed AS. In particular, we were interested in the level of AS of genes encoding extracellular proteins, since many of these genes are likely to be important in the intercellular structures and communication vital to multicellularity (consistent with this notion, genes encoding extracellular products are much less frequent in the unicellular fungus *C. neoformans* (4/4578, 0.09%) than in the multicellular species (ranging from 0.32–3.19% across species)). Such extracellular-associated genes did not show elevated AS rates. Instead, genes encoding proteins located in the nucleus and plasma membrane generally have higher proportions of AS.

Functional profile of alternatively spliced genes

AS levels across species for molecular function (F) and biological process (P) GO categories are shown in Figures 5 and 6, respectively, and in detail in Additional files 3 and 4, respectively. Again, gene categories generally associated with multicellularity (development, sensory-related functions) did not show elevated AS rates. Among molecular functions, protein kinase activity, RNA binding and calcium ion binding generally had high AS frequencies while monooxygenase activity, receptor activity, transporter activity and heme binding had much lower AS across all species. Biological processes showed greater variation across species

Functional profiling of alternatively spliced genes thus shows that most genes encoding most cellular functions exhibit AS. Some functions seem to be especially amenable to AS, perhaps due to these gene functions being particularly improved by the production of multiple products. AS is particularly prevalent in genes associated with regulation and signaling, consistent with previous observations [9,11]. Interestingly, the overrepresentation of AS in genes with regulatory functions previously observed in mammals is observed across lineages.

Discussion

Patterns of genome-wide AS usage are similar in different eukaryotic lineages

We studied patterns of genome-wide AS in 12 eukaryotic genomes. Our major findings include:

- 1) a correspondence between intron number and frequency of known AS, both within and between species;
- 2) evidence that ancient genes show relatively high levels of known AS.
- 3) no evidence for elevated AS in recently evolved genes – in fact, the most recently evolved genes are less likely to have known AS;
- 4) no clear evidence for elevated AS in classes of genes thought to be important in the rise of multicellularity;
- 5) a variety of similarities in the patterns of AS across diverse species, which could reflect patterns inherited from a putative AS-rich plant-animal ancestor.

Alternatively spliced genes in animals, plants and fungi have several common features. In all studied species, AS is widely found across genes with different functions, although those associated with regulation (i.e. protein kinase activity or RNA binding) show consistently higher AS levels. Also, the species studied showed similar patterns of AS usage in genes of different evolutionary age. Finally, intron number per gene was related with AS frequency in a similar manner in all species.

The simplest hypothesis to explain these similarities is that AS is homologous in these groups, inherited from their common ancestor, and that AS patterns in the common ancestor might have been similar. However, caution is necessary in interpreting this result. Given the high rates of gains and losses of AS events, the alternative hypothesis of convergent evolution of AS patterns in the different lineages, although less parsimonious, cannot be excluded.

Cellular Component		<i>Hsa</i> (34,16%-35,53%)				<i>Mmu</i> (25,44%-26,48%)				<i>Dme</i> (18,61%-20,66%)			
GO Entry	Name	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign
GO:0005634	nucleus	3255	35,82	0,8277		2902	30,94	0,0000	*(+)	763	22,02	0,3450	
GO:0005886	plasma membrane	426	36,15	1,0000		221	30,32	0,6911		268	36,19	0,0000	*(+)
GO:0005737	cytoplasm	847	41,44	0,0004	*(+)	717	33,47	0,0000	*(+)	298	23,83	0,3085	
GO:0005856	cytoskeleton	276	44,20	0,0060	*(+)	221	38,01	0,0005	*(+)	18	50	0,0294	*(+)
GO:0005783	endoplasmic reticulum	372	32,26	1,0000		357	21,57	0,2206		41	29,27	0,7265	
GO:0005840	ribosome	194	20,62	0,0001	*(-)	413	16,95	0,0000	*(-)	46	21,74	1,0000	
GO:0005739	mitochondrion	506	32,02	0,7580		560	16,61	0,0000	*(-)	152	15,79	1,0000	
GO:0005576	extracellular region	379	31,13	0,5496		139	23,02	1,0000		180	9,44	0,0010	*(-)

Cellular Component		<i>Cel</i> (8,26%-9,58%)				<i>Cne</i> (3,35%-4,47%)				<i>Ath</i> (8,14%-8,81%)			
GO Entry	Name	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign
GO:0005634	nucleus	1204	10,30	0,3119		663	3,62	1,0000		1875	11,89	0,0000	*(+)
GO:0005886	plasma membrane					177	7,91	0,0704	*(+)	118	13,56	0,3266	
GO:0005737	cytoplasm	307	11,40	0,6102		848	3,07	0,7445		390	16,41	0,0000	*(+)
GO:0005856	cytoskeleton	139	7,19	1,0000									
GO:0005783	endoplasmic reticulum	68	8,82	1,0000		124	4,84	1,0000		77	18,18	0,0393	*(+)
GO:0005840	ribosome	124	14,52	0,2087		29	3,45	1,0000		290	9,66	1,0000	
GO:0005739	mitochondrion	121	6,61	1,0000		217	3,28	1,0000		2861	8,60	0,0125	*(+)
GO:0005576	extracellular region	108	6,48	1,0000		4				85	4,71	1,0000	

Figure 4

AS frequency for GO categories for cellular locations. For each category, green/red colored AS frequency indicates that the frequency is higher/lower than the average, with (*) denoting statistical significance. In the "Total" column, the total number of genes of each category is shown (categories represented by less than 35 genes are shown in blue). In parenthesis, for each species, 95% confidence interval for the average of alternatively spliced genes in all Cellular location categories. *p*-values are given after multiple testing correction. Abbreviations: *Hsa* (*H. sapiens*), *Mmu* (*M. musculus*), *Dme* (*D. melanogaster*), *Cel* (*C. elegans*), *Cne* (*C. neoformans*), *Ath* (*A. thaliana*).

Ancient genes and functions show relatively high levels of AS

Also consistent with a relatively early origin of AS is our finding that recently evolved gene families (LSE) showed the lowest frequency of AS in all species (Figure 2B). Many of these genes are likely associated with newly evolved and complex lineage-specific traits in plants and animals. On the other hand, ancient gene families (PAF), which were already present in the ancestor, are highly alternatively spliced in modern organisms (Figure 2B) and a wide range of fundamental ancient functions are currently performed in eukaryotic cells by alternatively spliced genes (Figures 5 and 6), indicating no constraints for ancient genes to be alternatively spliced. Finally, gene functions that show consistently higher (e.g. RNA binding and protein kinase activity) or lower (e.g. mono-oxygenase activity) levels of AS across eukaryotes could have also had similar relative levels in the common ancestor.

Alternatively splicing and proteomic networks

We found no clear relationship between incidence of AS and gene dispensability (Figure 3). Indispensable genes tend to have large numbers of interaction partners, occu-

pying central positions in proteomic networks, while dispensable genes usually occupy external positions in the interacting networks [25]. These results thus suggest that AS may be integrated across all levels of eukaryotic proteomic networks.

Intron-rich gene structures is the main requirement for AS

We show that intron/exon number correlates strongly with the frequency of AS within and between genomes (Figure 1). In accordance, most intron-reduced genomes, such as those of most microsporidia and ascomycetes, show no AS [6] and other relatively reduced genomes, such as amoebas [26] or apicomplexans [27] do not exhibit high frequencies of AS. Interestingly, LSE genes, found to have significantly lower levels of AS, show lower average intron numbers than the other groups.

Our and others' previous work has shown that the plant-animal ancestor was at least moderately intron-rich (with at least as many introns as modern *Caenorhabditis* species), and that lower modern densities in some lineages reflect widespread intron loss [15,19,21,23,28-30]. Taken with present results, there are two important potential

Molecular Function		<i>Hsa</i> (36,91%-37,90%)				<i>Mmu</i> (28,59%-29,56%)				<i>Dme</i> (17,46%-18,44%)			
GO Entry	Name	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign
GO:0003723	RNA binding	473	41,01	1,0000		363	36,36	0,0384	*(+)	238	26,47	0,0212	*(+)
GO:0004672	protein kinase activity	31	61,29	0,1478		37	45,95	0,5439		113	27,43	0,2368	
GO:0005509	calcium ion binding	835	41,20	0,3115		712	31,60	1,0000		89	17,98	1,0000	
GO:0005524	ATP binding	1328	46,31	0,0000	*(+)	1297	38,09	0,0000	*(+)	12	25,00	1,0000	
GO:0003676	nucleic acid binding	750	43,47	0,0085	*(+)	941	39,43	0,0000	*(+)	348	14,94	1,0000	
GO:0008270	zinc ion binding	1600	40,44	0,1448		1533	34,83	0,0000	*(+)	13	23,08	1,0000	
GO:0004674	protein ser/thr kinase activity	499	48,30	0,0000	*(+)	409	37,65	0,0026	*(+)	134	32,84	0,0007	*(+)
GO:0003779	actin binding	211	44,08	0,6819		167	38,92	0,0973		118	38,98	0,0000	*(+)
GO:0016887	ATPase activity	106	60,38	0,0000	*(+)	105	33,33	1,0000		32	34,38	0,5296	
GO:0004713	protein-tyr kinase activity	444	48,20	0,0000	*(+)	353	37,11	0,0163	*(+)	39	20,51	1,0000	
GO:0016740	transferase activity	929	38,00	1,0000		877	30,33	1,0000		58	15,52	1,0000	
GO:0016787	hydrolase activity	714	38,66	1,0000		653	26,49	1,0000		63	9,52	1,0000	
GO:0005515	protein binding	3218	36,73	1,0000		2061	32,07	0,0293	*(+)	133	23,31	1,0000	
GO:0003677	DNA binding	986	37,73	1,0000		896	31,25	1,0000		203	27,59	0,0138	*(+)
GO:0003700	transcription factor activity	894	37,25	1,0000		703	26,60	1,0000		301	18,94	1,0000	
GO:0004871	signal transducer activity	307	43,00	0,6237		119	26,05	1,0000		24	12,50	1,0000	
GO:0008233	peptidase activity	152	41,45	1,0000		199	22,11	0,4102		48	8,33	1,0000	
GO:0003735	struct const of ribosome	270	21,85	0,0000	*(-)	506	17,79	0,0000	*(-)	157	14,01	1,0000	
GO:0016491	oxidoreductase activity	458	36,46	1,0000		437	20,59	0,0008	*(-)	200	8,00	0,0008	*(-)
GO:0030528	transc regulator activity	99	36,36	1,0000		103	27,18	1,0000		227	16,74	1,0000	
GO:0005215	transporter activity	342	35,67	1,0000		300	25,67	1,0000		122	16,39	1,0000	
GO:0004497	monooxygenase activity	69	37,68	1,0000		67	16,42	0,3141					
GO:0020037	heme binding	100	35,00	1,0000		124	16,94	0,0321	*(-)				
GO:0004872	receptor activity	1419	28,26	0,0000	*(-)	1811	13,86	0,0000	*(-)	108	15,74	1,0000	
GO:0004984	olfactory receptor activity	421	0,71	0,0000	*(-)	800	0,50	0,0000	*(-)	60	0,00	0,0001	*(-)

Molecular Function		<i>Cel</i> (9,45%-10,25%)				<i>Cne</i> (3,42%-4,77%)				<i>Ath</i> (8,57%-9,16%)			
GO Entry	Name	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign
GO:0003723	RNA binding	261	15,33	0,0811		52	7,69	1,0000		395	19,49	0,0000	*(+)
GO:0004672	protein kinase activity	402	13,43	0,2936		37	5,70	1,0000		662	9,37	1,0000	
GO:0005509	calcium ion binding	182	16,48	0,0858						208	10,10	1,0000	
GO:0005524	ATP binding	832	14,66	0,0001	*(+)					1448	8,70	1,0000	
GO:0003676	nucleic acid binding	524	11,45	1,0000						708	13,56	0,0005	*(+)
GO:0008270	zinc ion binding	841	7,49	0,2341						849	9,78	1,0000	
GO:0004674	protein ser/thr kinase activity	203	21,67	0,0000	*(+)	27	7,41	1,0000		707	8,06	1,0000	
GO:0003779	actin binding	33	27,27	0,0926						48	6,25	1,0000	
GO:0016887	ATPase activity	76	13,16	1,0000		62	4,84	1,0000		132	5,30	1,0000	
GO:0004713	protein-tyr kinase activity	101	8,91	1,0000						572	8,04	1,0000	
GO:0016740	transferase activity	607	11,04	1,0000						94	5,32	1,0000	
GO:0016787	hydrolase activity	623	9,15	1,0000						234	9,83	1,0000	
GO:0005515	protein binding	1408	9,09	1,0000		78	3,85	1,0000		1028	10,41	1,0000	
GO:0003677	DNA binding	916	7,75	0,3705		37	2,70	1,0000		1287	10,10	1,0000	
GO:0003700	transcription factor activity	467	9,64	1,0000		41	0,00	1,0000		1666	10,80	0,0765	
GO:0004871	signal transducer activity	170	15,29	0,3939		37	2,70	1,0000		99	6,06	1,0000	
GO:0008233	peptidase activity	184	8,15	1,0000		9	0,00	1,0000		106	10,38	1,0000	
GO:0003735	struct const of ribosome	135	13,33	1,0000		121	0,00	0,1445		350	10,29	1,0000	
GO:0016491	oxidoreductase activity	408	6,62	0,3318		19	0,00	1,0000		352	11,65	1,0000	
GO:0030528	transc regulator activity	48	10,42	1,0000						101	7,92	1,0000	
GO:0005215	transporter activity	210	6,67	1,0000		47	4,26	1,0000		300	9,00	1,0000	
GO:0004497	monooxygenase activity	92	0,00	0,0018	*(-)					259	3,47	0,0131	*(-)
GO:0020037	heme binding	126	0,79	0,0008	*(-)					218	3,21	0,0201	*(-)
GO:0004872	receptor activity	1032	6,01	0,0001	*(-)	6	0,00	1,0000		31	0,00	1,0000	
GO:0004984	olfactory receptor activity												

Figure 5
AS frequency for GO categories for molecular function. For each category, green/red colored AS frequency indicates that the frequency is higher/lower than the average, with (*) denoting statistical significance. In the "Total" column, the total number of genes of each category is shown (categories represented by less than 35 genes are shown in blue). In parenthesis, for each species, 95% confidence interval for the average of alternatively spliced genes in all Molecular Function categories. p-values are given after multiple testing correction. Abbreviations: Hsa (*H. sapiens*), Mmu (*M. musculus*), Dme (*D. melanogaster*), Cel (*C. elegans*), Cne (*C. neoformans*), Ath (*A. thaliana*).

Biological Process		<i>Hsa</i> (34,38%-35,48%)				<i>Mmu</i> (25,64%-26,74%)				<i>Dme</i> (21,27%-22,40%)			
GO Entry	Name	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign
GO:0006397	mRNA processing	58	39,66	1,0000		49	26,53	1,0000		30	46,67	0,0592	
GO:0045449	regulation of transcription	116	43,97	0,7309		98	35,71	0,6335		141	24,11	1,0000	
GO:0007155	cell adhesion	407	43,49	0,0053	*(+)	302	34,77	0,0151	*(+)	172	23,84	1,0000	
GO:0006470	protein amino acid dephosph.	156	57,05	0,0000	*(+)	110	33,64	1,0000		80	23,75	1,0000	
GO:0006468	protein amino acid phosph.	565	47,08	0,0000	*(+)	469	36,67	0,0000	*(+)	242	26,86	0,9797	
GO:0007242	intracellular signaling cascade	381	46,46	0,0001	*(+)	330	30,61	1,0000		145	22,07	1,0000	
GO:0006350	transcription	940	35,53	1,0000		714	32,21	0,0043	*(+)	30	26,67	1,0000	
GO:0006355	reg. of transc. DNA-dep	1622	37,18	0,7395		1499	32,42	0,0000	*(+)	79	21,52	1,0000	
GO:0006512	ubiquitin cycle	246	36,59	1,0000		209	24,88	1,0000		23	0,00	0,0465	*(-)
GO:0006508	proteolysis	486	39,92	0,3208		549	26,41	1,0000		499	9,22	0,0000	*(-)
GO:0006915	apoptosis	242	38,43	1,0000		174	27,59	1,0000		88	20,45	1,0000	
GO:0016567	protein ubiquitination	341	44,57	0,0036	*(+)	243	37,45	0,0019	*(+)				
GO:0006810	transport	530	39,06	0,6845		479	26,51	1,0000		124	14,52	0,7217	
GO:0006281	DNA repair	137	38,69	1,0000		96	30,21	1,0000		73	9,59	0,1367	
GO:0008283	cell proliferation	226	32,74	1,0000		85	21,18	1,0000		242	22,73	1,0000	
GO:0005975	carbohydrate metabolism	222	33,78	1,0000		167	25,15	1,0000		136	10,29	0,0090	*(-)
GO:0006457	protein folding	226	26,11	0,0740		214	14,95	0,0014	*(-)	110	17,27	1,0000	
GO:0006629	lipid metabolism	192	35,94	1,0000		122	27,87	1,0000		177	16,38	1,0000	
GO:0006412	protein biosynthesis	344	22,38	0,0000	*(-)	552	18,48	0,0003	*(-)	172	14,53	0,2765	
GO:0006811	ion transport	284	31,34	1,0000		222	26,58	1,0000		42	16,67	1,0000	
GO:0007186	G-prot coupled rec. prot. sign.	815	11,29	0,0000	*(-)	1313	4,49	0,0000	*(-)	218	19,27	1,0000	
GO:0008152	metabolism	419	41,53	0,0751		347	24,21	1,0000					
GO:0007165	signal transduction	1438	28,72	0,0000	*(-)	669	21,23	1,0000	*(-)	366	19,95	1,0000	
GO:0007275	development	395	30,38	0,8292		287	22,65	1,0000		80	11,25	0,3018	
GO:0006118	electron transport	379	35,36	1,0000		360	20,28	0,1390		40	15,00	1,0000	
GO:0007600	sensory perception	432	11,11	0,0000	*(-)	153	14,38	0,0083	*(-)	16	12,50	1,0000	
GO:0007608	sensory perception of smell	254	1,18	0,0000	*(-)	47	2,13	0,0003	*(-)	69	4,35	0,0015	*(-)

Biological Process		<i>Cel</i> (10,26%-11,51%)				<i>Cne</i> (3,19%-4,18%)				<i>Ath</i> (8,46%-9,13%)			
GO Entry	Name	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign	Total	%AS	corr. p	Sign
GO:0006397	mRNA processing	52	19,23	1,0000						40	10,00	1,0000	
GO:0045449	regulation of transcription	167	11,38	1,0000						577	9,53	1,0000	
GO:0007155	cell adhesion	52	25,00	0,0836						30	6,67	1,0000	
GO:0006470	protein amino acid dephosph.	129	3,88	0,0963		19	5,26	1,0000		50	16,00	1,0000	
GO:0006468	protein amino acid phosph.	408	13,73	1,0000		87	4,60	1,0000		962	7,59	1,0000	
GO:0007242	intracellular signaling cascade	165	12,73	1,0000						68	4,41	1,0000	
GO:0006350	transcription	372	8,60	1,0000						80	11,25	1,0000	
GO:0006355	reg. of transc. DNA-dep	630	11,27	1,0000		29	3,45	1,0000		939	9,27	1,0000	
GO:0006512	ubiquitin cycle	73	13,70	1,0000						77	14,29	1,0000	
GO:0006508	proteolysis	311	9,65	1,0000		18	5,56	1,0000		430	9,30	1,0000	
GO:0006915	apoptosis	20	10,00	1,0000						151	5,30	1,0000	
GO:0016567	protein ubiquitination	154	2,60	0,0030	*(-)					490	8,78	1,0000	
GO:0006810	transport	700	8,43	0,4139		62	6,45	1,0000		360	11,67	0,9895	
GO:0006281	DNA repair	46	6,52	1,0000		43	0,00	1,0000		63	12,70	1,0000	
GO:0008283	cell proliferation												
GO:0005975	carbohydrate metabolism	105	8,57	1,0000						359	10,03	1,0000	
GO:0006457	protein folding	100	11,00	1,0000		38	2,63	1,0000		201	10,95	1,0000	
GO:0006629	lipid metabolism	89	12,36	1,0000						180	7,22	1,0000	
GO:0006412	protein biosynthesis	185	15,14	1,0000		146	0,68	0,7907		403	10,92	1,0000	
GO:0006811	ion transport	248	12,90	1,0000						22	9,09	1,0000	
GO:0007186	G-prot coupled rec. prot. sign.	438	5,94	0,0049	*(-)					24	12,50	1,0000	
GO:0008152	metabolism	374	9,63	1,0000		19	0,00	1,0000		605	8,43	1,0000	
GO:0007165	signal transduction	239	12,97	1,0000		31	3,23	1,0000		223	6,73	1,0000	
GO:0007275	development	354	11,02	1,0000						137	6,57	1,0000	
GO:0006118	electron transport	290	5,86	0,0531						636	5,03	0,0052	*(-)
GO:0007600	sensory perception												
GO:0007608	sensory perception of smell												

Figure 6

AS frequency for GO categories for biological process. For each category, green/red colored AS frequency indicates that the frequency is higher/lower than the average, with (*) denoting statistical significance. In the "Total" column, the total number of genes of each category is shown (categories represented by less than 35 genes are shown in blue). In parenthesis, for each species, 95% confidence interval for the average of alternatively spliced genes in all Biological process categories. p-values are given after multiple testing correction. Abbreviations: Hsa (*H. sapiens*), Mmu (*M. musculus*), Dme (*D. melanogaster*), Cel (*C. elegans*), Cne (*C. neoformans*), Ath (*A. thaliana*).

implications. First, retention of ancestral intron densities was likely an important condition for modern AS. Thus, in fact AS played an important role in the emergence of organismal complexity [31,32], differential retention of ancestral introns would have profound consequences for morphological evolution across lineages.

Second, intron-rich ancestors are likely to have had significant AS. All thoroughly studied intron-rich genomes show relatively high frequencies of AS, suggesting both that a complex gene structure favors AS and that AS could have an important role in most non-reduced genomes, with high numbers of introns per gene. As we mention above, this is especially interesting in light of accumulating evidence that the last common ancestor of plants and opisthokonts was at least moderately intron rich [15,18-22] (with an estimated intron density between ~ 3.5 [18] and ~ 7.0 [15] introns per gene) and that it had weak consensus 5' splice site boundaries [8]. Among modern eukaryotes, both high intron number and weak 5'ss are characteristic of diverse species with widespread alternative splicing [8]. Therefore, these studies together strongly suggest the presence of AS in plant-animal ancestor

Intron numbers, spliceosomal errors, functionality and origin of AS

It is important to note that our present results do not address the functionality of alternative splice variants (and thus of AS) either in early eukaryotes or in modern organisms, a topic currently under debate [33,34]. Alternative transcripts produced from the same gene might: (i) encode different functions, (ii) reflect nonfunctional (but common) variants or (iii) represent rare spliceosomal errors, which will all appear in EST databases and thus in EST-based AS annotations. It seems likely that all three cases contribute to modern transcriptome variability. If in fact our argument is correct (that early eukaryotes already utilized extensive AS), it would be interesting to know how levels of AS functionality have changed through time. Increased requirements on proteome and regulatory flexibility could have driven an increase in functional AS. In tandem, refinements in the spliceosomal machinery could have increased splicing fidelity through eukaryotic evolution, disproportionately decreasing nonfunctional AS variants.

Interestingly, the evolutionary origin of functional AS is likely related to mis-splicing (splicing errors). AS might have evolved from mis-splicing as the early eukaryotic cells evolved to use and benefit from multiple splicing outputs. Therefore, the widespread production of multiple splice forms could be a main requirement for the origin of functional AS. Thus, our results along with the likely existence of weak splice sites in early eukaryotes [8] do not prove that early eukaryotes had functional AS, but

they strongly suggest that the last plant-animal ancestor had at least such additional splice variants available for potential participation and recruitment in biological processes.

AS in unicellular organisms and the origin of multicellularity

A striking potential implication of our results is that AS already existed in the plant-animal ancestor, a rather ancient and "primitive" unicellular eukaryote. As seen in some modern unicellular organisms (e.g. *Cryptococcus*), AS could have played an extensive role in the biology and evolution of these ancestral unicellular eukaryotes.

In this case, AS would have predated multicellularity and could perhaps have been recruited to allow the rise of multicellular complexity. This would resemble other biological processes, like apoptosis, whose origin precedes the rise of multicellular organisms, although the co-option of this ability was crucial for the advent of multicellularity [35].

Reliability of AS databases to answer evolutionary questions

Though we restricted our analysis to well-annotated genomes from long- and deeply-studied species with wide cDNA/EST coverage [36], it is likely that many alternative transcripts are not represented in current annotations, introducing the possibility of sampling biases. For instance, some gene types have been more thoroughly studied, and therefore may show higher proportions of annotated AS. However, such differences are unlikely to explain our central conclusions, since they are largely based on shared similarities, not differences (similar incidence of alternative splicing in old and new genes, dispensable and indispensable genes, genes of different functional classes), and clear associations across large numbers of genes. Indeed, EST coverage in *C. elegans* does not correlate with fractions of predicted alternatively spliced genes across different GO categories or gene ages, suggesting that the observed patterns are not due to EST sampling (Additional file 5).

Another potential problem associated with EST based annotation is the source of these data, especially in the case of humans. Many human EST libraries derive from cancerous or abnormal cell lineages, thought to contain aberrant, disease related alternative splice variants [37]. If these variants are more frequent among some groups of genes, this could introduce a bias in our results. However, our results for functional ontologies are based on data from a variety of species and the patterns presented here are consistent among all the species despite this potential source of noise in human databases.

We used genome annotation databases for this analysis because they are constructed using very similar approaches and so they might be more suitable for comparing these species. Supporting the quality of the studies databases, we found in the thorough analysis of the *C. elegans* Wormbase dataset that the vast majority of alternatively spliced isoforms included are well supported by experimental evidence, and only very few cases represent annotation mistakes.

Finally, it should be noted that the current results concern only presence/absence of AS, rather than number of alternative transcripts. Since alternatively spliced genes may produce from 2 to hundreds of isoforms, the effects on the transcriptome output will be quite different across genes, and further studies should address this important issue. Instead, we have concentrated on known AS presence/absence in a gene, which is likely to be less sensitive to differences in EST sampling. Similarly, different positions of the AS events in each gene may produce very different outputs with different effects on the organism's fitness. Our analysis does not take differences in function between splice variants into account. However, these considerations are unlikely to affect our conclusions about AS in early eukaryotes.

To verify our hypotheses on the emergence of AS, further studies of conservation of AS mechanisms (i.e. use of splicing regulators), splicing boundaries, and expression patterns will be necessary. Characterization of levels and patterns of AS in diverse additional eukaryotes, particularly unicellular intron-rich species, will also be important. Species of apicomplexans [4] and diatoms [5] have already been shown to have AS. EST and genome sequencing projects will provide data to assess whether AS was an ancestral feature of eukaryotic organisms, playing another important role in the complex RNA processing of early eukaryotes [7,38].

Conclusion

We find similar patterns of genome-wide AS usage in different eukaryotic lineages. We show that ancient genes and functions (present in the common ancestor of plants and opisthokonts) have high levels of AS in modern organisms indicating no bias against AS of these genes. These genes were also likely intron-rich in the common ancestor [15,18-22], which we find to be the main requirement for AS. Since the spliceosomal machinery is widely conserved throughout eukaryotes [7,38], our results favor the hypothesis that some form of AS appeared relatively early in eukaryotic evolution, at least in the unicellular common ancestor of plants, animals and fungi (around 1300 million years ago [39,40], quite early in the evolution of extant eukaryotes [41]). This implies AS appeared before the rise of multicellular organ-

isms, and could therefore have an important role in the biology of ancient unicellular organisms.

Methods

Datasets and resources

GenBank genome annotations were downloaded from NCBI webpage [42] or Ensembl database [43] for six metazoa: human (*Homo sapiens* (NCBI 36 Ensembl 38.36)), mouse (*Mus musculus* (NCBI m35 Ensembl 38.35)), chicken (*Gallus gallus* (WASHUC1 Ensembl 38.1n)), zebra fish (*Danio rerio* (Zv5 Ensembl 38.35e)), fruitfly (*Drosophila melanogaster* (FlyBase release 4.1)), worm (*Caenorhabditis elegans* (WS150 Wormbase 38.150a)); four fungi: *Cryptococcus neoformans* B3501-A (NC 006670, NC 006679–NC 006687, NC006691–NC006694), *Schizosaccharomyces pombe* 972h (AL672256-8.1), *Saccharomyces cerevisiae* YJM789 (AAF000000000.1), and *Encephalitozoon cuniculi* GB-M1 (AL391737.1, AL590442-50.1); one plant: *Arabidopsis thaliana* (NC 003070.5, NC 003071.3, NC 003074.4, NC 003075.3, NC 003076.4, based on TAIR genome annotations); and one apicomplexan: *Plasmodium falciparum* HB3 (AANS000000000.1).

Quality of the databases

For many species, there are currently various genomic databases available having information on AS. In each such case, we used the richest and most up-to-date database, containing the largest number of described alternatively spliced isoforms. Each of the databases was constructed by automatic predictions of gene structures, generally combining different software, and then confirmed by mapping ESTs and cDNAs onto genomic sequences and usually manually curated. Described alternatively spliced isoforms are based on alignments of ESTs and cDNAs onto these gene models. For each genome, some subsets of genes are manually annotated and thoroughly studied. Detailed explanations of the methods using in deriving these databases are available from the primary references [3,44-46] and from the Ensembl, TAIR and NCBI web pages.

To better understand these genome annotations we further explored one of them, the Wormbase annotation of *C. elegans*. We studied each gene that was annotated to be alternatively spliced. We found that 97.8% of the isoforms had one or more kinds of experimental support (RNA, ORF sequence tags (OSTs) and/or ESTs), described as "confirmed by cDNA(s)" or "partially confirmed by cDNA(s)", thus only ~2.2% of isoforms are predictions. In addition, for each case we aligned the different isoforms against the genomic sequence. In only 2.4% of cases, we found that slight errors, generally one or two base indels, were responsible for the annotation of alternative splicing.

To test the effects of sampling biases we analyzed the coverage of ESTs per Kb for each gene and for each category of gene. For each gene we counted the number of matching ESTs per gene available in Wormbase and divided it by the length of the longest transcript. Importantly, no correlation was found between EST coverage and percentage of genes that were alternatively spliced for any of the GO classifications (cellular location, molecular function, or biological process), or for age of gene (Additional file 4).

Evolutionary analyses

For evolutionary analyses we used the Eukaryotic Clusters of Orthologous Groups (KOGs), which includes putative ortholog sets for seven species: *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi*. This database is suitable for the study of protein functions from an evolutionary perspective, addressing issues such as origin of gene functions or their dispensability during eukaryotic evolution [47]. Data were downloaded from the corresponding NCBI webpage [48] and linked to current genome annotation databases. Data was carefully filtered for repetitions resulting from database linking and from the updating of annotations of the genes included in the KOG database.

Gene Ontology analyses

Gene Ontology annotations for cellular location (C), molecular function (F) and biological process (P) for genes from *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, *C. neoformans* and *A. thaliana* were obtained from Gene Ontology Consortium website [49]. This database was linked to Ensembl or NCBI gene ID's, using UniProt ID's [50], if necessary. Data was carefully filtered to avoid redundancies due to database linking. We analyzed in *H. sapiens* a total of 18589 entries in 458 C-GO categories, 36653 entries in 2063 F-GO categories and 29162 in 1867 P-GO categories; in *M. musculus*, 19002 entries in 450 C-GO categories, 33461 entries in 1898 F-GO categories and 24793 in 2119 P-GO categories; in *D. melanogaster*, 5811 entries in 411 C-GO categories, 11916 in 1571 F-GO categories and 20307 in 1657 P-GO categories; in *C. elegans*, 7151 entries in 174 C-GO categories, 21627 in 837 F-GO categories and 9579 in 562 P-GO categories; in *C. neoformans* 4578 entries in 279 C-GO categories, 3298 in 969 F-GO categories and 5564 in 884 P-GO categories; in *A. thaliana*, 26592 entries in 281 C-GO categories, 34959 in 1255 F-GO categories and 27845 in 1210 P-GO categories.

Statistical analysis

Percentages of alternatively spliced genes were calculated for each category under study. The correspondent 95% confidence interval (CI) was calculated for each percent-

age using the standard formula: $a \pm 1.96 \sqrt{\frac{a(1-a)}{N}}$, where a is the fraction of alternatively spliced genes in a given group and N the total number of genes in that group.

To assess statistical under/overrepresentation of AS in each studied category, we used Fisher exact tests (assuming the one-sided probability for similarity of samples). For each of the three blocks of GO terms presented in Figures 4, 5 and 6 (cellular location, molecular function and biological process, respectively), we corrected for multiple testing using full Bonferroni correction.

In the Additional files 2, 3 and 4, we excluded groups of genes that contained less than 35 genes.

Analysis of alternative isoforms

Gene, intron and exon information was extracted from their annotation using a PERL script "Intron_finder.pl" as previously described [36]. For each gene, custom PERL scripts assessed intron number and alternative splicing (i.e. the position or length of any alignable intron or internal exon is different in at least two different isoforms). AS events in *P. falciparum* were extracted from [51].

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

MI carried out the data collection, genomic and statistical analyses, designed and conceived the study and drafted the manuscript. SWR and JLR helped to draft the manuscript and participated in the interpretation and analyses of the data. DP participated in the design of the study, coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Ancient alternatively spliced KOG's. List of 36 ancient KOG (appeared before the split of animals, fungi and plants) that show high AS incidence in A. thaliana, D. melanogaster, C. elegans and H. sapiens.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-7-188-S1.xls>]

Additional file 2

Cellular locations and alternative splicing. List of different cellular locations and their AS frequency for A. thaliana, D. melanogaster, C. elegans, C. neoformans, M. musculus and H. sapiens.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-7-188-S2.xls>]

Additional file 3

Molecular functions and alternative splicing. List of different molecular functions and their AS frequency for A. thaliana, D. melanogaster, C. elegans, C. neoformans, M. musculus and H. sapiens.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-7-188-S3.xls>]

Additional file 4

Biological Processes and alternative splicing. List of different biological processes and their AS frequency for A. thaliana, D. melanogaster, C. elegans, C. neoformans, M. musculus and H. sapiens.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-7-188-S4.xls>]

Additional file 5

EST/cDNAs Sampling Bias Control. Control for EST/cDNAs sampling bias. It has been performed in C. elegans. The document contains 4 figures, corresponding to: cellular locations (C), molecular functions (F), biological process (P) and species groups.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2148-7-188-S5.pdf>]

Acknowledgements

MI was supported by Fundación Caixa Galicia, SWR by the Allan Wilson Centre of Molecular Ecology and Evolution and JLR by a Carlsberg Foundation Grant (21-00-0680). We thank Irene Sicilia and Ignacio Maeso for helpful comments and discussions during the preparation of this manuscript, Klaus Schliep for helping with the statistical analyses and Tim White and Michael Woodhams for their help in programming.

References

1. Yatzkan E, Yarden O: **The B regulatory subunit of protein phosphatase 2A is required for completion of macroconidiation and other developmental processes in Neurospora crassa.** *Mol Micro* 1999, **31**:197-209.
2. Ye D, Lee CH, Queener SF: **Differential splicing of Pneumocystis carinii f. sp. carinii inosine 5'-monophosphate dehydrogenase pre-mRNA.** *Gene* 2001, **263**:151-158.
3. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, Allen JE, Bosdet IE, Brent MR, Chiu R, Doering TL, Donlin MJ, D'Souza CA, Fox DS, Grinberg V, Fu J, Fukushima M, Haas BJ, Huang JC, Janbon G, Jones SJM, Koo HL, Krzywinski MI, Kwon-Chung JK, Lengeler KB, Maiti R, Marra MA, Marra RE, Mathewson CA, Mitchell TG, Pertea M, Riggs FR, Salzberg SL, Schein JE, Shvartsbeyn A, Shin H, Shumway M, Specht CA, Suh BB, Tenney A, Utterback TR, Wickes BL, Wortman JR, Wye NH, Kronstad JW, Lodge JK, Heitman J, Davis RW, Fraser CM, Hyman RW: **The genome of the Basidiomycetous yeast and human pathogen Cryptococcus neoformans.** *Science* 2005, **307**:1321-1324.
4. Li L, Brunk BP, Kissinger JC, Pape D, Tang K, Cole RH, Martin J, Wylie T, Dante M, Fogarty SJ, Howe DK, Liberator P, Diaz C, Anderson J, White M, Jerome ME, Johnson EA, Radke JA, Stoeckert CJ Jr., Waterston RH, Clifton SW, Roos DS, Sibley LD: **Gene discovery in the Apicomplexa as revealed by EST sequencing and assembly of a comparative gene database.** *Genome Res* 2003, **13**:443-454.
5. Kinoshita S, Kaneko G, Lee JH, Kikuchi K, Yamada H, Hara T, Itoh Y, Watabe S: **A novel heat stress-responsive gene in the marine diatom Chaetoceros compressum encoding two types of transcripts, a trypsin-like protease and its related protein, by alternative RNA splicing.** *Eur J Biochem* 2001, **268**:4599-4609.
6. Ast G: **How did alternative splicing evolve?** *Nat Rev Genet* 2004, **5**:773-782.
7. Collins L, Penny D: **Complex spliceosomal organization ancestral to extant eukaryotes.** *Mol Biol Evol* 2005, **22**:1053-1066.
8. Irimia M, Penny D, Roy SW: **Coevolution of genomic intron number and splice sites.** *Trends Genet* 2007, **23**:321-325.
9. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, RIKEN GER Group, GSL Members, Hayashizaki Y, Gaasterland T: **Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome.** *Genome Res* 2003, **13**:1290-1300.
10. Taneri B, Snyder B, Novoradovsky A, Gaasterland T: **Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific.** *Genome Biol* 2004, **5**:R75.
11. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucl Acids Res* 2001, **29**:2850-2859.
12. Kim H, Klein R, Majewski J, Ott J: **Estimating rates of alternative splicing in mammals and invertebrates.** *Nat Genet* 2004, **36**:915-916.
13. Guiliano DB, Hall N, Jones SJ, Clark LN, Corton CH, Barrell BG, Blaxter ML: **Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes.** *Genome Biol* 2002, **3**:R57.
14. Banyai L, Patthy L: **Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues.** *FEBS Lett* 2004, **565**:127-132.
15. Roy SW, Gilbert W: **Complex early genes.** *Proc Natl Acad Sci USA* 2005, **102**:1986-1991.
16. Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, Ferrier D, Benes V, de Jong P, Weissenbach J, Bork P, Arendt D: **Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii.** *Science* 2005, **310**:1325-1326.
17. Roy SW: **Intron-rich ancestors.** *Trends Genet* 2006, **22**:468-471.
18. Csurós M: **Likely scenarios of intron evolution.** Springer LNCS 3678; 2005:47-60.
19. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: **Analysis of evolution of exon-intron structure of eukaryotic genes.** *Brief Bioinform* 2005, **6**:118-134.
20. Nguyen HD, Yoshihama M, Kenmochi N: **New maximum likelihood estimators for eukaryotic intron evolution.** *PLoS Comput Biol* 2005, **1**:e79.
21. Yoshihama M, Nakao A, Nguyen HD, Kenmochi N: **Analysis of ribosomal protein gene structures: implications for intron evolution.** *PLoS Genet* 2006, **2**:e25.
22. Roy SW, Gilbert W: **Rates of intron loss and gain: implications for early eukaryotic evolution.** *Proc Natl Acad Sci USA* 2005, **102**:5773-5778.
23. 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.
24. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, Rao BS, Smirnov S, Sverdlov A, Vasudevan S, Wolf Y, Yin J, Natale D: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
25. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.** *Genome Res* 2003, **13**:2229-2235.
26. Davis CA, Brown MPS, Singh U: **Functional Characterization of Spliceosomal Introns and Identification of U2, U4, and U5 snRNAs in the Deep-Branching Eukaryote Entamoeba histolytica.** *Eukaryotic Cell* 2007, **6**:940-948.
27. Singh N, Preiser P, Renia L, Balu B, Barnwell J, Blair P, Jarra W, Voza T, Landau I, Adams JH: **Conservation and developmental control of alternative splicing in maebl among malaria parasites.** *J Mol Biol* 2004, **343**:589-599.
28. Slamovits CH, Keeling PJ: **A high density of ancient spliceosomal introns in oxymonad excavates.** *BMC Evol Biol* 2006, **6**:34.
29. Roy SW, Gilbert W: **The evolution of spliceosomal introns: patterns, puzzles and progress.** *Nat Rev Genet* 2006, **7**:211-221.
30. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: **Conservation versus parallel gains in intron evolution.** *Nucl Acids Res* 2005, **33**:1741-1748.

31. Modrek B, Lee C: **A genomic view of alternative splicing.** *Nat Genet* 2002, **30**:13-19.
32. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes.** *Nucl Acids Res* 2007, **35**:125-131.
33. Sorek R, Shamir R, Ast G: **How prevalent is functional alternative splicing in the human genome?** *Trends Genet* 2004, **20**:68-71.
34. Rukov JL, Irimia M, Mork S, Lund VK, Vinther J, Arctander P: **High Qualitative and Quantitative Conservation of Alternative Splicing in *Caenorhabditis elegans* and *Caenorhabditis briggsae*.** *Mol Biol Evol* 2007, **24**:909-917.
35. Bidle KD, Falkowski PG: **Cell death in planktonic, photosynthetic microorganisms.** *Nat Rev Micro* 2004, **2**:643-655.
36. Collins L, Penny D: **Investigating the intron recognition mechanism in eukaryotes.** *Mol Biol Evol* 2006, **23**:901-910.
37. Baranova AV, Lobashev AV, Ivanov DV, Krukovskaya LL, Yankovsky NK, Kozlov AP: **In silico screening for tumour-specific expressed sequences in human genome.** *FEBS Lett* 2001, **508**:143-148.
38. Kurland CG, Collins LJ, Penny D: **Genomics and the irreducible nature of eukaryote cells.** *Science* 2006, **312**:1011-1014.
39. Douzery EJP, Snell EA, Baptiste E, Delsuc F, Philippe H: **The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?** *Proc Natl Acad Sci USA* 2004, **101**:15386-15391.
40. Hedges SB, Blair JE, Venturi ML, Shoe JL: **A molecular timescale of eukaryote evolution and the rise of complex multicellular life.** *BMC Evol Biol* 2004, **4**:2.
41. Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW: **The tree of eukaryotes.** *Trends Ecol Evol* 2005, **20**:670-676.
42. **NCBI** [<http://www.ncbi.nlm.nih.gov>]
43. **Ensembl** [<http://www.ensembl.org>]
44. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl Automatic Gene Annotation System.** *Genome Res* 2004, **14**:942-950.
45. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucl Acids Res* 2003, **31**:5654-5666.
46. Hild M, Beckmann B, Haas SA, Koch B, Solovyev V, Busold C, Fellenberg K, Boutros M, Vingron M, Sauer F, Hoheisel JD, Paro R: **An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome.** *Genome Biol* 2003, **5**:R3.
47. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin YJ, DA N: **A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.** *Genome Biol* 2004, **5**:R7.
48. **KOG database** [<http://ftp.ncbi.nih.gov/pub/COG/KOG/>]
49. **Gene Ontology Consortium** [<http://www.geneontology.org>]
50. **UniProt** [<http://www.ebi.uniprot.org>]
51. **Scrpps Genome Centre** [<http://www.sgc.ucsd.edu/autodb/browse.php?db=PfalSDB2>]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

