Intron Length Coevolution across Mammalian Genomes

Peter A. Keane[1] and Cathal Seoighe*,[1]

[1]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland

***Corresponding author:** E-mail: Cathal.Seoighe@nuigalway.ie

Associate editor: John Parsch

Abstract

Although they do not contribute directly to the proteome, introns frequently contain regulatory elements and can extend the protein coding potential of the genome through alternative splicing. For some genes, the contribution of introns to the time required for transcription can also be functionally significant. We have previously shown that intron length in genes associated with developmental patterning is often highly conserved. In general, sets of genes that require precise coordination in the timing of their expression may be sensitive to changes in transcript length. A prediction of this hypothesis is that evolutionary changes in intron length, when they occur, may be correlated between sets of coordinately expressed genes. To test this hypothesis, we analyzed intron length coevolution in alignments from nine eutherian mammals. Overall, genes that belong to the same protein complex or that are coexpressed were significantly more likely to show evidence of intron length coevolution than matched, randomly sampled genes. Individually, protein complexes involved in the cell cycle showed the strongest evidence of coevolution of intron lengths and clusters of coexpressed genes enriched for cell cycle genes also showed significant evidence of intron length coevolution. Our results reveal a novel aspect of gene coevolution and provide a means to identify genes, protein complexes and biological processes that may be particularly sensitive to changes in transcriptional dynamics.

***Key words:* coevolution, introns, transcription, coexpression.**

Introduction

The origin, prevalence, and function of introns in eukaryotic genomes have long been a subject of great interest in molecular evolution. Although typically they do not contribute directly to the protein coding sequences of genes, introns have been shown to play several important functional roles. Introns frequently contain regulatory elements and allow for the process of alternative splicing, expanding the protein coding capacity of the genome, and giving rise to additional modes of gene regulation (Chorev and Carmel 2012). In mammals, introns comprise up to 95% of transcribed primary protein-coding sequences (Mattick and Gagen 2001), imposing a significant cost in terms of the energy required for transcription and in terms of the potential for erroneous transcripts resulting from mis-splicing, particularly for genes with longer introns (Fox-Walsh et al. 2005). The time taken to transcribe introns can be significant. Estimates of the rate of transcription in eukaryotes range from 1 to 4 Kb per minute (Ardehali and Lis 2009) with a median rate of transcription estimated from over 2,000 human genes at 1.5 Kb per minute (Veloso et al. 2014). As an extreme example of the time taken to transcribe long introns, the human dystrophin gene, which is 2.3 Mb in length, 99% of which is intronic, can take up to 16 h to transcribe (Tennyson et al. 1995), corresponding to a rate of 2.4 Kb per minute. In contrast, genes that are highly expressed tend to have short introns (Castillo-Davis et al. 2002), and genes that require rapid induction in response to stress are intron poor (Jeffares et al. 2008).

The evolution and functional significance of large introns in mammals has been much debated. Lynch (2002, 2006) proposed that intron expansion is a largely a semi-neutral process, resulting from insufficient purifying selection to remove insertions in organisms with small effective population sizes. The genome design model asserts that intron length reflects gene function (Vinogradov 2006). This model proposes that housekeeping genes tend to have short introns because they require very little regulation (Eisenberg and Levanon 2003). Conversely, genes that require more complex modes of gene regulation, such as tissue-specific genes are longer due, in part, to the presence of intronic regulatory elements (Castillo-Davis et al. 2002; Vinogradov 2006). In opposition to this view, a number of authors have argued that the short introns in highly and ubiquitously expressed genes, such as housekeeping genes, are the result of selection for efficiency, resulting in reduced energetic costs of transcription (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Seoighe et al. 2005).

For some genes, the transcriptional delay resulting from the time taken to transcribe introns is functionally significant. This intron delay is particularly important in embryonic development, where precise temporal control of gene expression is required to ensure proper development of the embryo (Swinburne and Silver 2008). During embryonic development, cells undergo rapid divisions creating a temporal constraint on transcription (Shermoen and O'Farrell 1991). In Drosophila, introns in genes expressed by the embryo

doi:10.1093/molbev/msw151 Advance Access publication August 22, 2016

during this period are short, allowing for rapid transcription between cell cycles (Artieri and Fraser 2014). Intron length can also contribute to more complex forms of gene regulation. Delayed expression, resulting in part from the transcription of introns, combined with negative feedback loops, can create oscillating patterns of gene expression. The *Hes7* gene, which is involved in somite segmentation during development, represents an example of this. Removal of introns from *Hes7* in mouse resulted in *Hes7* being expressed 19 min early and the abolition of expression oscillations, resulting in severe developmental defects (Takashima et al. 2011). The intron lengths, though in general not the intron sequences, of *Hes7* as well as several other genes involved in developmental patterning are highly conserved, suggesting that these genes are under purifying selection to maintain intron length (Seoighe and Korir 2011). Intron length also appears to be relevant in protein complex formation as genes that occur within the same complex usually have very similar length (Chen et al. 2009).

Here, we investigated the extent of intron length coevolution across several mammalian genomes using whole genome alignments together with reconstructed ancestral sequences. We hypothesized that sets of genes that require precise coordination in the timing of their expression may be sensitive to changes in intron length and that natural selection can act to maintain the relative lengths of such genes, resulting in correlated changes in intron content. We found that, overall, coexpressed genes and genes belonging to the same protein complex were significantly more likely to have coevolving intron lengths than randomly sampled genes, matched for total intron content. We also found a number of gene modules that showed particularly strongly correlated intron length evolution that may reflect selection acting on relative transcript lengths. We propose that transcript length coevolution can reveal functional constraints on intron lengths and that intron length coevolution may suggest biological processes that require precise temporal regulation of gene expression.

Results

We hypothesized that sets of genes that require precise coordination in the timing of their expression may be particularly sensitive to evolutionary changes in intron content, and that changes in intron content due to insertions and deletions would be correlated among such sets of genes. To identify sets of genes where this may be the case, we compared changes in intron content across the species phylogeny for 9,396 genes from nine mammalian genomes. The phylogenetic tree representing the nine species used in this analysis is shown in figure 1.

Protein Complexes Show Significant Evidence of Intron Length Coevolution

Multi-protein complexes are often precisely coregulated in order to ensure proper assembly and function (Papp et al. 2003) and thus we predicted that the intron content of members of some protein complexes may display evidence of coevolution. Subunits of protein complexes tend to be coexpressed (Jansen et al. 2002), a pattern which is conserved across a wide range of species (Webb and Westhead 2009). Studies in yeast have shown that gene expression levels among sets of interacting proteins coevolve (Fraser et al. 2004), suggesting that the mechanisms that control this coregulation may also coevolve. It has also been suggested that gene length contributes to the coregulation of protein complexes, as genes encoding subunits of multi-protein complexes tend to have similar lengths (Chen et al. 2009). Furthermore, protein–protein interactions are known to be largely conserved across mammals (Pérez-Bercoff et al. 2013).

We sought to determine whether intron length coevolution could be detected among subunits of known protein complexes, potentially representing a novel aspect of the coevolutionary dynamics of members of multi-protein complexes. To do this, we downloaded all human protein complexes from the core set of the CORUM database of mammalian protein complexes (Ruepp et al. 2009). We then carried out a randomization test to determine if members of the same protein complex show significantly greater intron length correlation compared to sets of randomly sampled genes with similar intron content (see “Materials and Methods” section). For this test, we only considered complexes for which at least three genes were present in our original 9,396 gene dataset, resulting in 440 complexes that could be tested.

First, we asked if intron length coevolution could be detected by considering all interacting proteins from across all protein complexes at once. This was done using a randomization test described in the “Materials and Methods” section. We found evidence of intron length coevolution of genes belonging to the same protein complex using this approach ($P < 0.0001$; fig. 2*A*). We next asked which protein complexes showed the greatest evidence of coevolution. We repeated the randomization test, this time considering each protein complex individually. Of the 440 complexes tested, the members of 44 clusters showed evidence of intron length coevolution (uncorrected $P < 0.05$), the top 10 of which are shown in table 1. Given that there was overall evidence of coevolution of members of the same protein complex and an excess of nominally significant complexes over the null expectation we applied the *q*-value method (Storey and Tibshirani 2003) to investigate properties of the collection of statistical tests applied to protein complexes. The proportion, π_0, of tests estimated to conform to the null hypothesis was 0.9, indicating that 10% of tests were not consistent with the null hypothesis (fig. 3*A*). Just one complex, annotated in CORUM as the NCOR1 complex, had a *q*-value below 0.05 (*q*-value = 0.04). The full list of nominally significant complexes and their q-values is provided in supplementary table S1, Supplementary Material online.

Intron Length Coevolution in Coexpressed Genes

We next asked if intron length coevolution is a general feature of genes that are coregulated. Sets of genes that are coregulated include coexpressed genes. Indeed, interacting proteins

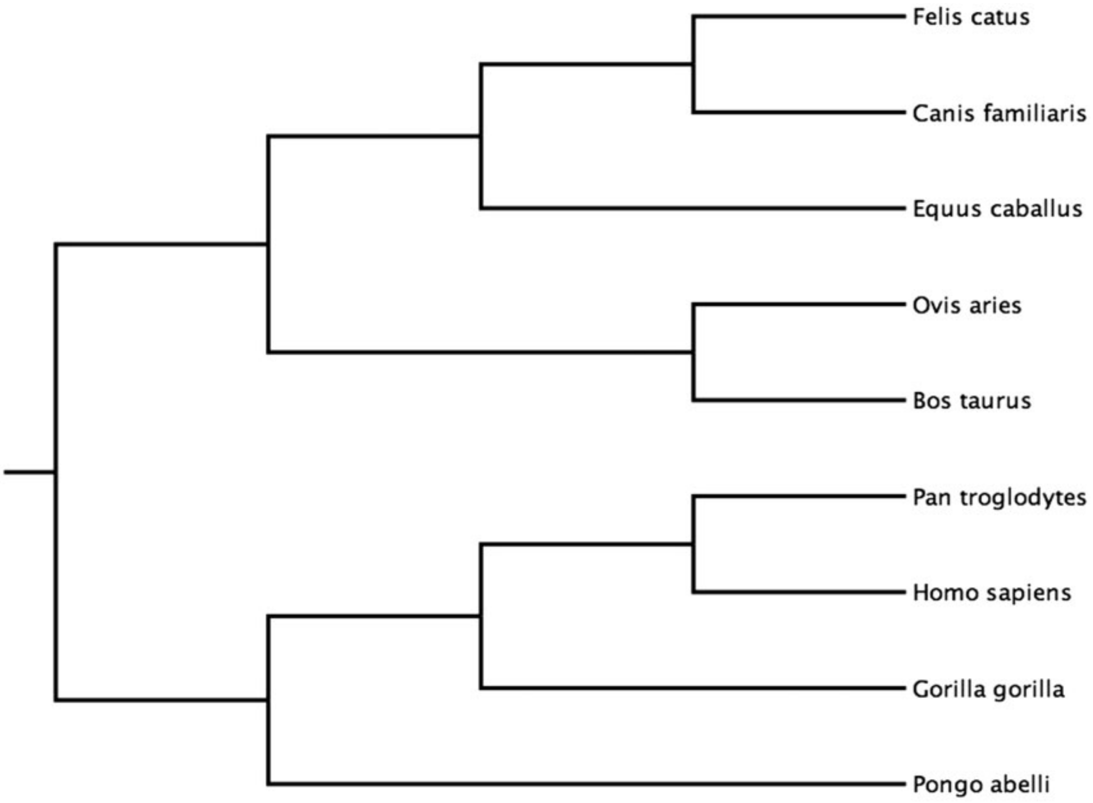

Fig. 1. Phylogenetic tree of the nine mammals included in this study.

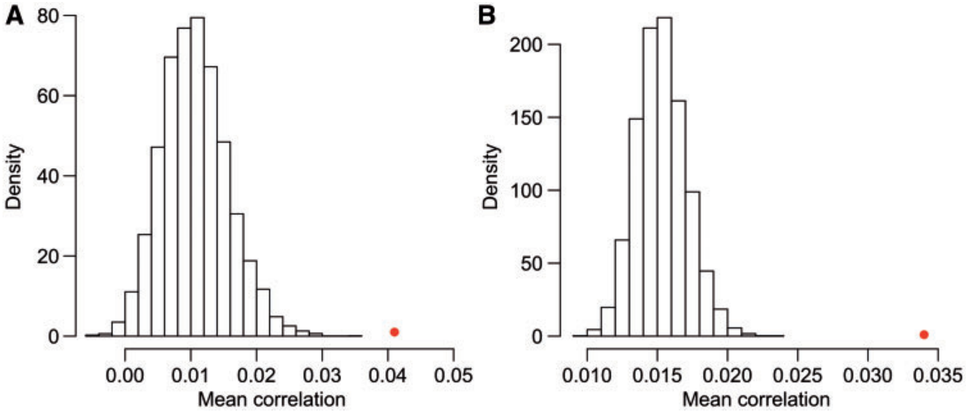

Fig. 2. Distribution of the mean partial Pearson correlation values from 10,000 randomizations of (A) CORUM protein complexes and (B) GTEx coexpression clusters. The mean value of the real data is shown as a red dot.

are often coexpressed (Jansen et al. 2002). We constructed a gene coexpression network containing 60 clusters using publicly available gene expression data from the Genotype-Tissue Expression (GTEx) consortium (see supplementary table S2, Supplementary Material online, for a list of genes included in each coexpression cluster) (Lonsdale et al. 2013). Although we allowed for overlapping clusters in our method (see “Materials and Methods” section), we found that the clusters identified were largely nonoverlapping and independent (supplementary fig. S1, Supplementary Material online). For this analysis, we only considered the 50 clusters for which at least three genes were present in our original 9,396 gene dataset. To first test if intron length coevolution could be detected across the gene coexpression network as a whole, we applied the randomization test used for protein complexes to the 50 clusters. Similarly to protein complexes, we found that coexpressed genes were significantly more likely to be coevolving compared to sets of random genes ($P < 0.0001$; fig. 2B). We next applied this test to each cluster individually to determine which clusters were most enriched for intron length coevolution. We found 17 clusters with significant evidence of coevolution ($P < 0.05$; uncorrected P-values; table 2). π_0 was estimated to be 0.3, suggesting that 70% of tests were not consistent with the null hypothesis (fig. 3B). All 17 tests remained significant after correction for multiple testing using the q-value method (q-value < 0.05; table 2). These clusters

Table 1. Top 10 Significantly Coevolving CORUM Protein Complexes.

CORUM complex ID	Complex description	No. of genes analyzed	Mean Pearson coefficient	*P*-value	*q*-value
1413	NCOR1 complex	8	0.43	<0.0001	0.04
2446	ITGA9-ITGB1-FIGF complex	3	0.77	0.001	0.12
5589	LINC complex, S-phase	6	0.42	0.001	0.16
799	DMAP1-associated complex	8	0.32	0.002	0.16
1179	CENP-A NAC-CAD complex	6	0.42	0.002	0.16
657	Retromer complex (SNX1, SNX2, VPS35, VPS29, VPS26A)	3	0.66	0.004	0.23
570	p300-CBP-p270-SWI/SNF complex	6	0.36	0.005	0.26
781	URI complex (Unconventional prefoldin RPB5 Interactor)	6	0.35	0.005	0.26
351	Spliceosome	69	0.04	0.007	0.28
1332	Large Drosha complex	10	0.19	0.007	0.28

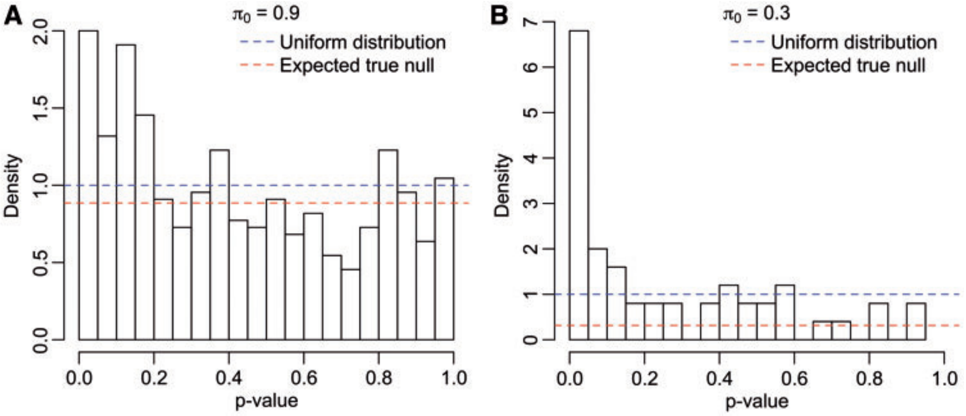

Fig. 3. Distribution of p-values obtained from 10,000 randomizations of (A) CORUM protein complexes and (B) GTEx coexpression clusters. π_0, the expected proportion of tests consistent with the null hypothesis is shown as a red dotted line. The uniform distribution, expected under the null hypothesis is shown in blue.

Table 2. Significantly Coevolving GTEx Coexpression Clusters.

ID	Most significantly enriched GO Biological Process	No. of genes	Mean correlation	*P*-value	*q*-value
2	Synaptic transmission	429	0.06	<0.0001	<0.0001
9	Nucleic acid metabolic process	83	0.09	<0.0001	<0.0001
5	Cell cycle	171	0.04	0.0003	0.002
4	Synaptic transmission	213	0.05	0.004	0.004
8	ncRNA metabolic process	122	0.03	0.01	0.031
15	Ribosome biogenesis	85	0.04	0.015	0.037
30	Cardiac muscle tissue development	30	0.1	0.02	0.037
14	Extracellular muscle organization	102	0.05	0.021	0.037
21	Extracellular muscle organization	49	0.08	0.021	0.037
19	Cellular respiration	64	0.04	0.024	0.038
10	Muscle system process	110	0.03	0.03	0.038
13	Carboxylic acid metabolic process	63	0.04	0.033	0.038
11	Translation elongation	34	0.05	0.033	0.038
34	Peptide hormone processing	19	0.09	0.038	0.038
46	Carboxylic acid transport	7	0.2	0.038	0.038
1	Spermatogenesis	511	0.02	0.039	0.038
44	Surfactant homeostasis	7	0.19	0.049	0.045

are highlighted in the gene coexpression network in supplementary figure S1, Supplementary Material online.

It is possible that the intron content coevolution we observed for coexpressed genes or members of protein complexes reflects similar patterns of intron length change in different expression classes (e.g., highly or broadly expressed genes). To explore this possibility, we repeated the randomization tests, matching on median, or maximum gene expression across GTEx tissues in addition to intron length (see "Materials and Methods" section). The randomization test remained significant for both the protein complexes ($P = 0.0016$ and $P = 0.0001$, matching on median or

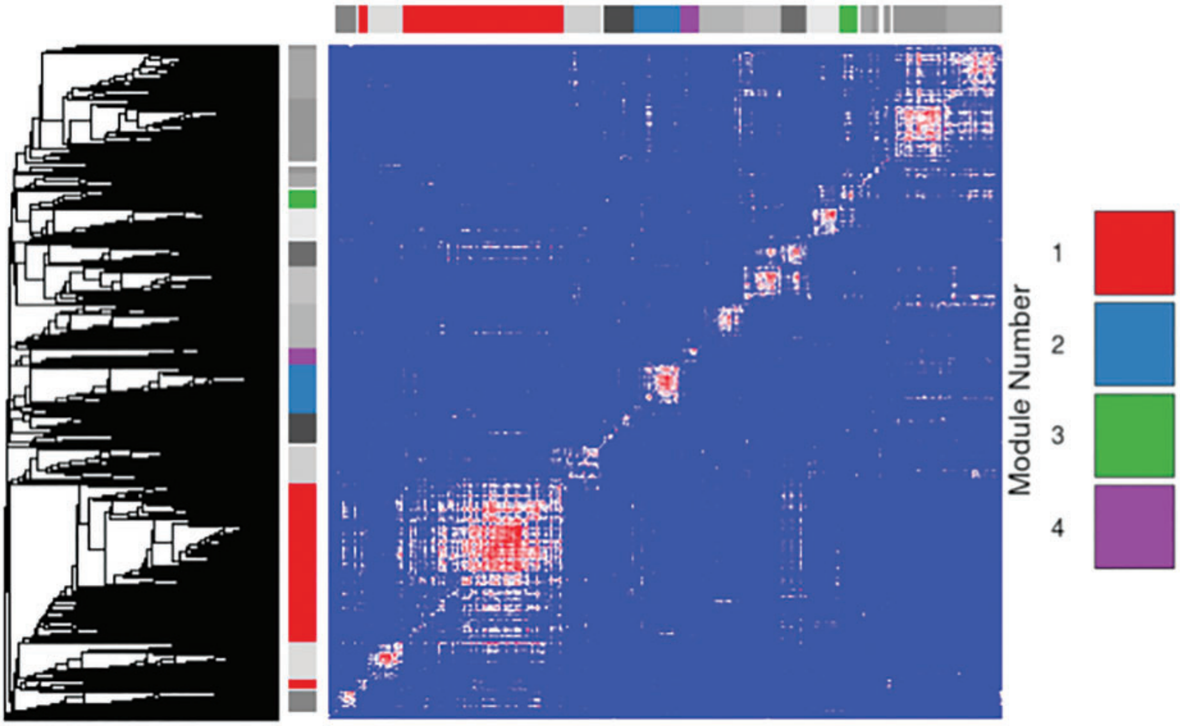

Fig. 4. Heatmap showing the results of hierarchical clustering of the Simpson coefficient. Gene modules identified using the dynamicTreeCut method are indicated by colored bars along the *x*- and *y*-axes. Four modules that were significantly enriched for specific biological processes are colored red, blue, green, and purple. Modules that were tested, but not significantly enriched are shown in grey.

Table 3. GO Biological Processes Most Significantly Enriched in Gene Coevolution Modules.

Module number	GO term	Description	Size of module	Expected count	Count	Size of term	Log odds ratio	FDR
1	GO:0048513	Organ development	2355	324.8	417	2601	0.39	3.3×10^{-5}
2	GO:0048856	Anatomical structure development	661	147.4	205	4374	0.48	0.001
3	GO:0006396	RNA processing	264	9.9	27	762	1.09	0.035
4	GO:0032879	Regulation of localization	251	20.7	46	1721	0.95	0.003

maximum expression, respectively) and gene coexpression clusters ($P = 0.0002$ and $P < 0.0001$). Thus, we did not find evidence that expression class was the driver of intron content coevolution of members of protein complexes or coexpression clusters.

We also considered the possibility that correlated change in intron length of coexpressed genes and members of protein complexes results from genomic colocation. It has previously been shown that genes in close proximity to each other tend to have similar patterns of gene expression and similarly sized introns (Urrutia and Hurst 2003; Batada et al. 2007). To investigate the possible impact of this on our analysis, we repeated our randomization test but this time excluding all gene pairs within a genomic distance of 1 Mb. We found that this did not significantly alter our results, as both protein complexes and coexpressed genes were still found to be significant after the exclusion of these gene pairs ($P < 0.0001$ for both the protein complex and coexpressed genes). Therefore, we conclude that chromosomal location or linkage does not significantly contribute to the patterns of intron length coevolution we observed.

Functional Enrichment in Coevolving Gene Modules

Considering all genes, we constructed a network of genes with edges corresponding to gene pairs that showed highly correlated change in intron content across the phylogeny (partial Pearson correlation > 0.7; see "Materials and Methods" section for details), and then calculated the Simpson coefficient for each pair of genes in the resulting network. Hierarchical clustering of this matrix identified 14 distinct modules (fig. 4). To assess the sensitivity of this method to the choice of correlation threshold, we repeated this analysis at various thresholds ranging from 0.1 to 0.9. We found that this did not significantly alter the structure or content of these gene modules (supplementary fig. S2, Supplementary Material online), indicating that this analysis is robust to choice of threshold.

Four of these modules were functionally enriched for biological processes (FDR < 0.05; see supplementary table S3, Supplementary Material online, for a list of genes in each of these coevolution modules). The most significantly enriched biological process for each of these modules is shown in table 3 (for a complete list of enriched processes, see supplementary table S4, Supplementary Material online). Module 1 was most significantly enriched for organ development. Interestingly, both modules 1 and 2 were enriched for biological processes related to development, in particular brain development (supplementary table S4, Supplementary Material online), suggesting that coevolution of intron length may be particularly important in this process.

Next, we examined the patterns of change in intron content across the phylogeny for each of the four coevolving modules that were enriched for biological processes (supplementary fig. S3, Supplementary Material online; see "Materials and Methods" section for details). We were interested to determine whether the coordinated change in intron length observed in these enriched modules reflected a trend towards longer or shorter intron lengths in these gene sets along specific lineages. In general, increasing intron length was observed more frequently than decreases in intron length, consistent with what has been reported previously (Gelfman et al. 2012). Modules 1 and 3 showed some similarity in the rate of intron change across branches in the primate clade (supplementary fig. S3*A* and *C*, Supplementary Material online). In modules 2 and 4, much of the signal appeared to arise from rapid increase in intron length in the enriched gene sets on individual branches (supplementary fig. S3*B* and *D*, Supplementary Material online).

In the case of Pearson correlation, coordinated change in intron content along a single branch among members of a gene set may be sufficient to generate a signal of coevolution. In contrast, the rank-based Spearman correlation would require coordinated changes on multiple branches. To identify signals of coevolution likely to result from coordinated changes in multiple branches, we repeated the analysis using Spearman correlation (see "Materials and Methods" section for details). With a lower threshold in this case ($\rho > 0.5$), we were able to construct 16 modules from this data, four of which were functionally enriched (FDR < 0.05; supplementary fig. S4, Supplementary Material online; see supplementary table S5, Supplementary Material online, for a list of genes in each enriched module). These modules were enriched for developmental processes, including central nervous system development (modules S2 and S3; supplementary table S6, Supplementary Material online), consistent with our results from analysis of the partial Pearson correlations. Interestingly, we also found that these modules were enriched for several metabolic processes, as well as several processes related to the cell cycle (module S1; supplementary table S6, Supplementary Material online). There was also a significant overlap between these modules and modules identified using Pearson correlation (supplementary table S7, Supplementary Material online). Although some of the Pearson modules overlapped significantly with more than one Spearman module, the same themes emerged. In both cases, the enriched coevolving modules were enriched for developmental and cell cycle processes. Clear modules were only visible at lower thresholds than for the Pearson correlation, although traces of these modules persisted at higher thresholds (supplementary fig. S5, Supplementary Material online).

Effects of Species Choice

The nine species in figure 1 were selected from the 15 species for which alignments were available as a tradeoff between maximizing the number of species and maximizing the number of genes for which an ortholog was found in all species. We also considered a second tree (supplementary fig. S6, Supplementary Material online), consisting of ten species and including a slightly reduced number (8719) of genes. Results obtained using this set of species was similar to those obtained on the original data set. Both protein complexes and coexpressed genes were significantly more likely to show evidence of intron length coevolution compared to sets of random genes with similarly sized introns ($P = 0.0001$ and $P < 0.0001$ for the CORUM protein complexes and GTEx coexpressed genes, respectively), consistent with our previous findings from the nine species tree. We were able to identify 13 coevolving modules (supplementary fig. S7, Supplementary Material online), one of which was significantly enriched for developmental processes, though not, specifically, for brain development (supplementary tables S8 and S9, Supplementary Material online). We also tested the effects of the human-centric approach we adopted to defining gene models (see "Materials and Methods" section) and instead defined introns according to gene models annotated in cow. Evidence for intron length coevolution remained significant for both the protein complexes and GTEx coexpression clusters ($P = 0.005$ and $P < 0.0001$, respectively).

Coevolution of Intergenic Lengths

Although we hypothesized that genes that require precise coordination in the timing of their expression may show evidence of coevolution in intron content, this does not imply that selection acting on transcription time is the only possible cause of coevolution in intron content. For example, similarities in the genomic context of functionally related genes may result in similar patterns of insertion and deletion events across lineages that could result in correlated evolution. Furthermore, introns, particularly first introns, contain regulatory elements, and functionally related genes may have similar regulatory complexity. Changes in the selection pressures acting on functionally related sets of genes along distinct lineages could allow them to become more or less permissive to loss or gains of regulatory sites through deletions or insertions, potentially giving rise to correlated change in intron content across the phylogeny. As a test of these alternative and other similar potential sources of intron content coevolution, we redid the analysis but this time treating the length of the intergenic region as the property of interest. Genes belonging to the same protein complex did not show evidence of coevolution of the length of intergenic regions ($P = 0.72$). Coexpressed genes showed marginally significant evidence ($P = 0.07$) of coevolution of the lengths of intergenic regions, but the extent of coevolution for intergenic lengths was significantly less than for intron lengths (e.g., partial Pearson correlation coefficients of genes belonging to the same GTEx coexpression clusters were significantly higher in the case of introns than intergenic regions; $P = 7.4 \times 10^{-8}$). However, coevolution modules similar to the modules shown in figure 4 were also observed for intergenic lengths and these were also significantly enriched for genes associated with developmental processes, but not for cell cycle processes (supplementary fig. S8 and tables S10 and S11, Supplementary Material online). Therefore, we conclude that there may be alternative explanations for coevolution of intron lengths of members of some of these modules; however, the coevolution of intron content only, in the case of members of protein complexes and the far stronger coevolution of intron content than of intergenic content for coexpressed gene clusters is consistent with a role for transcription time in intron length coevolution.

Discussion

In mammals, introns comprise up to 95% of the length of primary protein coding transcripts (Mattick and Gagen 2001), raising questions as to the evolution and functional consequences of such large introns. While some theories have argued that the expansion of introns in eukaryotes is largely a semi neutral process resulting from the failure to prevent insertions that have only a slightly deleterious effect on fitness in organisms with small effective population sizes (Lynch 2002, 2006), others have focused more on the functional roles of introns, particularly in gene regulation. The genome design theory suggests that intron length reflects gene function (Vinogradov 2006). Others have argued that the shorter introns found in highly expressed genes are due to selection for transcriptional efficiency (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Seoighe et al. 2005)

In this study, we carried out an investigation into intron length coevolution in mammals. We have shown that evolutionary changes in intron content are positively correlated among several sets of coregulated genes, including multiprotein complexes and coexpressed genes. Transcriptional coregulation of protein complex subunits is conserved over evolutionary time (Webb and Westhead 2009). Our results show that members of the same protein complex are significantly more likely to show coevolution of intron length compared to sets of randomly sampled genes with similar intron content. Coordinated expression of the components of protein complexes is important for the proper formation of the complex (Papp et al. 2003). The time that it takes to transcribe genes that encode members of protein complexes places a constraint on the evolution of the intron content of coregulated genes, particularly for genes such as those involved in development or formation of certain protein complexes that are sensitive to small changes in expression timing.

Coevolution in intron content was more apparent for some protein complexes than for others. For example, the NCOR1 complex (table 2), a complex that is involved in chromatin modification (Underhill et al. 2000). Interestingly, four of the top five complexes, including NCOR1, with the strongest evidence of intron length coevolution are annotated in CORUM (Ruepp et al. 2009) as being involved in the cell cycle, a time sensitive process. These are NCOR1, CENP-A NCOR/CAD complex, DMAP1 associated complex, and the LINC complex. Strikingly, our analysis of gene coexpression clusters is in agreement with this result, as coexpression clusters that were enriched for cell cycle genes were also found to be significantly coevolving (clusters 1 and 5; supplementary table S12, Supplementary Material online). The cell cycle is a dynamic process that requires precise temporal regulation of protein complex components at various time points (Lichtenberg et al. 2005). The presence of cell cycle processes in the set of genes associated with high levels of intron length coevolution is consistent with our initial hypothesis that genes that require precise coordination in the timing of their expression may show signs of intron length coevolution. We propose that for a substantial proportion of such genes large changes in intron length may have deleterious consequences.

Delays in gene expression that result from the transcription of introns are thought to be important in embryonic development (Swinburne and Silver 2008; Takashima et al. 2011). We have previously found that genes with the greatest conservation in intron length are highly enriched for processes relating to developmental patterning (Seoighe and Korir 2011). Although, we found little evidence for conserved sequences in the introns of these genes, we could not exclude that the conservation of intron lengths was the result of conserved functional elements within these introns that limited the scope for changes in length through insertion and deletion. The present study takes our original observation further by showing that, not only is the intron content of genes involved in developmental processes conserved, but when changes are observed in the intron content they tend to be shared among coexpressed sets of genes involved in these processes.

Our results suggest that constraints acting on transcription time are not likely to be the only source of intron content coevolution. Our initial hypothesis was that genes that are particularly sensitive to changes in expression timing (e.g., genes involved in developmental patterning) may be particularly sensitive to changes in intron content and that sets of genes with coordinated expression timing (e.g., members of tightly regulated protein complexes) may show evidence of coevolution of intron content. However, this does not preclude that other factors may also cause coevolution of intron content. For example, functionally related genes may sometimes occur in similar genomic contexts that may experience similar mutational patterns across different lineages, setting up correlation in intron length. As a means of exploring these alternatives, we compared the coevolution of intron content to the coevolution of the lengths of intergenic sequences. In the case of members of protein complexes and coexpressed gene pairs, the correlation in intron content was either not found or found to a much lesser extent when we considered intergenic length instead of intronic content, reducing the plausibility of shared genomic context or selective pressures acting to preserve gene regulation as the cause of correlated intron content of these genes. Some of the functionally enriched clusters with coevolving intron content could be reproduced with intergenic lengths rather than intron content, suggesting that these alternative explanations may apply in these cases. This included clusters that were enriched for brain development and consequently we cannot conclude that the coevolution of the intron content of genes involved in brain development reflects the effects of constraints acting on transcription timing.

In summary, we hypothesized that the primary transcript length of sets of genes that require precise coordination in expression timing coevolves. In support of this hypothesis, we found that, in general, genes belonging to the same protein complex and pairs of coexpressed genes are significantly more likely to have coevolving intron length than randomly sampled gene pairs, matched for intron content. We found that some biological processes that are particularly time-sensitive, such as the cell cycle, are particularly enriched for genes with coevolving intron content. Our results suggest that for some sets of genes, intron length is functionally relevant and evolves under natural selection, representing a novel aspect of intron evolution and of the coevolutionary dynamics of coregulated genes.

Materials and Methods

Data

Gene models for human protein coding genes were downloaded from Ensembl release 74 via BioMart (Cunningham et al. 2015). To identify insertion and deletion events (indels) in the introns of these genes, we downloaded whole genome alignments of 15 eutherian mammals from Ensembl Compara release 74. This data included ancestral sequences, reconstructed using the Enredo-Pecan-Ortheus (EPO) pipeline, at each branch of the phylogenetic tree, allowing indels to be placed on the branch on which they are inferred to have occurred (Hubbard et al. 2009). We selected nine species for which complete ortholog sets were available for a large number of genes. This selection was made on the basis of a trade-off between maximizing the number of species included in the tree and maximizing the number of genes for which orthologs were available for each species in the tree. The latter consideration was important for our analyses, particularly in order to have good coverage of protein complexes. These species were *Felis catus, Canis familiaris, Equus caballus, Ovis aries, Bos taurus, Pan troglodytes, Homo sapiens, Gorilla gorilla*, and *Pongo abelii*. Only alignments containing all nine species were used for analysis, resulting in a dataset consisting of 9396 genes. The phylogenetic tree of these species is shown in figure 1.

Partial Correlation of Intron Change

A prediction of our hypothesis is that changes in intron length due to insertions and deletions should be correlated among sets of coevolving genes. To test this, we first calculated the intron content of each gene, defined as the sum of the lengths of all introns in the canonical transcript, using the positions of the intron/exon boundaries of the human gene as a reference. Canonical transcripts were obtained directly from Ensembl. For protein coding genes, these are defined as the transcript with the longest coding sequence without stop codons. Using the indels inferred on the branches of the phylogeny, we also calculated the intron content at each ancestral node. This was done by first calculating the intron content of the last common ancestor of all nine species by subtracting the length of all indels along the human lineage from the intron content of the human gene, and then applying the set of all indels leading to each node. We then calculated the proportional change in the intron content along each branch of the phylogeny. Partial Pearson correlation of this proportional change was calculated for each pair of genes, controlling for the median (over genes) change in intron content along each branch. This allowed us to control for differences in the overall rate of evolution along different branches.

Identification of Coevolving Gene Modules

From the partial correlations calculated above, we constructed a network with edges corresponding to pairs of highly correlated genes (partial Pearson coefficient > 0.7). This network comprised 9,286 genes with 1.2×10^6 edges. To extract modules of coevolving genes, we calculated the Simpson coefficient for each pair of genes *a* and *b* using the formula

$$\frac{|A \cap B|}{\min(|A|, |B|)}$$

where *A* and *B* are the sets of genes that share an edge with genes *a* and *b* respectively. We then converted this similarity measure to a distance (1 – Simpson coefficient), and carried out average linkage hierarchical clustering of the resulting matrix in R (version 3.1.1). Modules were extracted from the dendrogram using the dynamicTreeCut method (version 1.62) (Langfelder and Zhang 2014). To investigate the functions associated with these modules, we carried out Gene Ontology (GO) enrichment analysis using GOseq (version 1.16.2) (Young et al. 2010). We included intron content as a covariate, to exclude the possibility that the observed functional enrichment was a consequence of shared intron content of genes in certain functional classes (with genes with larger or smaller intron content having had similar proportional changes in intron content over the phylogeny). Each module was tested for enriched biological processes. Correction for multiple testing of the *P*-values produced by GOseq was done using the Benjamini–Hochberg method in R. To rank these processes in order of enrichment, we calculated the log odds ratio, and the 95% confidence interval of the log odds ratio for each enriched GO term using logistic regression with intron length included as a covariate.

Visualization of Intron Length Change

To show the changes in intron length across the phylogeny of the gene coevolution modules, we carried out a linear regression of the proportional change in intron content against the median change along each branch, for each gene in the module. We then calculated the residuals for these regressions and plotted the mean value obtained on the phylogenetic tree of the nine species. This was carried out for all functionally enriched gene modules and is represented in supplementary figure S3, Supplementary Material online, using colors to denote the value of the residuals on each branch.

Test for Significantly Coevolving Gene Sets

We developed a randomization-based test that can be used to determine if a given set of genes is significantly more likely to be correlated than expected, compared to random sets of genes with similarly sized introns. First, to account for the possibility that genes with similar intron content may evolve in a similar way, we divided the genes into five intron size classes ($<$5 Kb, 5–20 Kb, 20–50 Kb, 50–100 Kb and $>$100 Kb). For each gene within a set of interest, we randomly selected another gene at random from among all genes of the same intron size class. We then calculated the mean partial Pearson correlation of the random pairings. This was repeated 10,000 times and produced a *P*-value that represents the proportion of times the mean correlation of the random gene set was greater than or equal to that of the real data. A gene set was deemed to be significant if this *P*-value was less than 0.05. To control for the false discovery rate from these tests, we calculated π_0, the expected proportion of tests that are consistent with the null hypothesis that coevolution in the gene set is no different to randomly selected genes (Storey and Tibshirani 2003). This was done using the *q*-value package in R (version 2.0.0) (Storey 2015).

We applied this test to a set of human protein complexes obtained from the core set of the CORUM database of mammalian protein complexes (Ruepp et al. 2009), and to gene coexpression networks constructed using publicly available gene expression data obtained from the GTEx consortium (Lonsdale et al. 2013).

To investigate a possible influence of gene expression level on intron length coevolution, we modified the above test to also control for gene expression level in addition to intron size class. To do this, we first calculated the median and maximum expression value for each gene across all tissues of the GTEx gene expression dataset (Lonsdale et al. 2013). Genes were then placed into one of five categories based on their median or maximum expression, and during the randomization test, genes were randomly resampled from both their relevant intron size class and gene expression class.

Construction of Gene Coexpression Networks

We downloaded gene expression data from the GTEx consortium (data release V4) via the GTEx web portal (www.gtexportal.org). This data consists of Reads Per Kilobase per Million mapped reads (RPKM) expression values calculated from RNA sequencing applied to 42 human tissues/cell-types, sampled from 190 individuals (Lonsdale et al. 2013). To reduce noise, we only considered genes with RPKM $>$ 0.5 in at least 10% of all samples (22,718 genes). Pearson correlations were then calculated for each gene pair. We constructed a gene coexpression network with edges corresponding to correlated gene pairs ($|r| > 0.7$). Gene clusters were then identified using the ClusterONE algorithm (version 1.0), which can identify densely connected and overlapping gene clusters within large biological networks (Nepusz et al. 2012). ClusterONE was run with a minimum cluster size of 10 and maximum matching coefficient between clusters of 0.5. Clusters above this threshold were merged to form single clusters. All other parameters were set at default values. For unweighted networks, ClusterONE determines the statistical significance of the clusters using a one-sided Mann–Whitney U test conducted on the number of inward versus outward edges. A low *P*-value is obtained when the number of inward edges is significantly greater than the number of outward edges (Nepusz et al. 2012). A cluster was deemed significant if this *P*-value was less than 0.05. To relate these clusters to gene function, we annotated each cluster using the most significantly enriched biological process identified using the GOstats package (version 2.30.0) in R (Falcon and Gentleman 2007). GO terms with an FDR $<$ 0.05 were deemed to be significantly enriched in the coexpression clusters. To visualize the coexpression network, we carried out hierarchical clustering of the adjacency matrix for the network in R. The results of this are shown as a heatmap in supplementary figure S1, Supplementary Material online.

Availability of Software

The analysis presented here was conducted primarily using custom Perl scripts. In order to ensure reproducibility of our results, we have made these scripts available on GitHub at https://github.com/petebio/Intron_coevolution together with detailed usage instructions.

Supplementary Material

Supplementary figures S1–S8, tables S1–S12, and references are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Araxi Urrutia and an anonymous reviewer for insightful comments that significantly improved the manuscript. This work was supported by Science Foundation Ireland (Grant number 07/SK/M1211b). P.K. was funded under the Simulation Science (SimSci) PhD program. SimSci is funded under the Program for Research in Third Level Institutions and cofunded under the European Regional Development Fund.

References

Ardehali MB, Lis JT. 2009. Tracking rates of transcription and splicing in vivo. *Nat Struct Mol Biol* 16:1123–1124.

Artieri CG, Fraser HB. 2014. Transcript length mediates developmental timing of gene expression across Drosophila. *Mol Biol Evol* 31(11): 2879–2889.

Batada NN, Urrutia AO, Hurst LD. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet* 23:480–484.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet* 31:415–418.

Chen X, Shi S, He X. 2009. Evidence for gene length as a determinant of gene coexpression in protein complexes. *Genetics* 183:751–754. 1SI–5SI.

Chorev M, Carmel L. 2012. The function of introns. *Front Genet* 3:55.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res* 43:D662–D669.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet* 19:362–365.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinform Oxf Engl* 23:257–258.

Fox-Walsh KL, Dou Y, Lam BJ, Hung S, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci USA* 102:16176–16181.

Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA* 101:9033–9038.

Gelfman S, Burstein D, Penn O, Savchenko A, Amit M, Schwartz S, Pupko T, Ast G. 2012. Changes in exon–intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res* 22:35–50.

Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al. 2009. Ensembl 2009. *Nucleic Acids Res* 37:D690–D697.

Jansen R, Greenbaum D, Gerstein M. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res* 12:37–46.

Jeffares DC, Penkett CJ, Bähler J. 2008. Rapidly regulated genes are intron poor. *Trends Genet* 24:375–378.

Langfelder P, Zhang B. 2014. dynamicTreeCut: Methods for detection of clusters in hierarchical clustering dendrograms. Available from: https://cran.r-project.org/web/packages/dynamicTreeCut/index.html.

Lichtenberg U. d, Jensen LJ, Brunak S, Bork P. 2005. Dynamic complex formation during the yeast cell cycle. *Science* 307:724–727.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The genotype-tissue expression (GTEx) project. *Nat Genet* 45:580–585.

Lynch M. 2002. Intron evolution as a population-genetic process. *Proc Natl Acad Sci USA* 99:6118–6123.

Lynch M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol* 23:450–468.

Mattick JS, Gagen MJ. 2001. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol* 18:1611–1630.

Nepusz T, Yu H, Paccanaro A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 9:471–472.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Pérez-Bercoff Å, Hudson CM, Conant GC. 2013. A conserved mammalian protein interaction network. *PLoS One* 8:e52581.

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W. 2009. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 36(Database issue):D646–D650.

Seoighe C, Gehring C, Hurst LD. 2005. Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet* 1:e13.

Seoighe C, Korir PK. 2011. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. *BMC Bioinform* 12:S16.

Shermoen AW, O'Farrell PH. 1991. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell* 67:303–310.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100:9440–9445.

Storey JD. 2015. qvalue: Q-value estimation for false discovery rate control. Rpackage version 2.0.0. Available from: http://qvalue.princeton.edu/, http://github.com/jdstorey/qvalue.

Swinburne IA, Silver PA. 2008. Intron delays and transcriptional timing during development. *Dev Cell* 14:324–330.

Takashima Y, Ohtsuka T, González A, Miyachi H, Kageyama R. 2011. Intronic delay is essential for oscillatory expression in the segmentation clock. *Proc Natl Acad Sci USA* 108:3300–3305.

Tennyson CN, Klamut HJ, Worton RG. 1995. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet* 9:184–190.

Underhill C, Qutob MS, Yee S-P, Torchia J. 2000. A novel nuclear receptor corepressor complex, N-CoR, contains components of the mammalian SWI/SNF complex and the corepressor KAP-1. *J Biol Chem* 275:40463–40470.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res* 13:2260–2264.

Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, Ljungman M. 2014. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res* 24:896–905.

Vinogradov AE. 2006. "Genome design" model: Evidence from conserved intronic sequence in human–mouse comparison. *Genome Res* 16:347–354.

Webb EC, Westhead DR. 2009. The transcriptional regulation of protein complexes; a cross-species perspective. *Genomics* 94:369–376.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 11:R14.