Available online at www.sciencedirect.com

ScienceDirect

Genomics 90 (2007) 44–48

GENOMICS

www.elsevier.com/locate/ygeno

Patterns of gene duplication and intron loss in the ENCODE regions suggest a confounding factor

Sourav Chatterji [a,*], Lior Pachter [b,c]

[a] *Genome Center, University of California at Davis, Davis, CA 95616, USA*
[b] *Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720, USA*
[c] *Department of Computer Science, University of California at Berkeley, Berkeley, CA 94720, USA*

Received 29 November 2006; accepted 22 March 2007
Available online 11 May 2007

Abstract

The exon–intron structure of eukaryotic genes allows for phenomena such as alternative splicing, nonsense-mediated decay, and regulation through untranslated regions. However, the evolution of the exon structure of genes is not well elucidated because of limited and phylogenetically sparse data sets. In this study, we use the phylogenetically diverse sequencing of the ENCODE regions to study gene structure evolution in mammalian genomes. This first phylogenetically diverse study of gene structure changes offers insights into the mode and tempo of mammalian gene structure evolution. The genes undergoing structure changes appear to be moderately to highly expressed in germline cells and show levels of selection similar to those of other ENCODE genes. Patterns of gene duplication of the affected genes are more complex than expected. The number of sampled genomes is sufficiently dense to infer that certain gene duplications happened after intron loss. Thus, although gene duplication is highly correlated with intron loss, we conclude that structural changes in genes are not necessarily due to a loss of constraint following gene duplication as previously suggested.

Keywords: Gene evolution; Exon–intron structure; Intron loss; Gene duplication; Gene structure evolution; Mammalian genome evolution; ENCODE

The elucidation of the biology of protein-coding genes is one of most important problems in molecular biology. There has been a considerable amount of investigation of the evolution of genes at the nucleotide and amino acid levels. However, in addition to coding for proteins, eukaryotic genes have an exon–intron structure and introns are spliced out of precursor mRNAs. This structure allows for phenomena such as alternative splicing, nonsense-mediated decay, and regulation through untranslated regions. Because changes in gene structure are rare [1], it has been difficult to obtain large data sets for studying the evolution of gene structure. Consequently, there have been few systematic studies on the evolution of gene structure [2–4]. Most of these studies involved the comparison of gene structure in large sets of orthologous genes in widely separated eukaryotic genomes. Because of the very large evolutionary distances between the genomes, several issues about gene structure remain unsolved. These include a precise elucidation of the mechanisms underlying intron gain/loss and a complete elucidation of the relationship between gene duplication and change in gene structure.

The evolutionary history of an intron can be reconstructed by comparing the presence/absence of the intron in the phylogenetic tree relating the orthologous genes. There are two approaches to reconstructing this evolutionary history from the phylogenetic tree. The maximum parsimony approach [2,5] infers the evolutionary history that can explain the phylogenetic tree most parsimoniously with regard to intron gain and loss events. The parsimony approach assumes that intron gain and loss are comparatively rare. However, if species being studied are phylogenetically sparse, the parsimony approach may give incorrect or ambiguous answers because of parallel intron gain and loss. On the other hand, the maximum likelihood approach [6–8] infers the evolutionary history with the highest probability according to a particular model of intron evolution. The results of the likelihood

* Corresponding author.
E-mail address: schatterji@ucdavis.edu (S. Chatterji).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.03.008

approach depend on the assumptions in the underlying model and different likelihood models infer vastly disparate results for phylogenetically diverse data sets such as the one in the study by Rogozin et al. [2].

A denser phylogenetic sampling of genomes can help us make inferences about the evolutionary history of an intron with greater confidence and the ENCODE project [9] has provided such a resource. A key part of the project has been the sequencing of multiple species orthologous to 1% of the human genome. The September 2005 release of ENCODE contains 546 Mb of genomic sequence from 44 vertebrates. This includes about 500 Mb of sequences from 38 mammalian genomes. In addition, the human ENCODE sequences have been rigorously annotated as part of the GENCODE project [10]. Thus the ENCODE regions offer an unprecedented opportunity to study the evolution of gene structure in mammals.

The study of gene structure changes among related species requires accurate annotation of protein-coding genes in all sequences. As discussed above, the ENCODE sequences have a well-curated set of human annotations. However, ENCODE sequences in nonhuman species have little experimental evidence to support gene annotation. In a previous study [11], we developed GeneMapper, a reference-based annotation program for transferring annotations from a well-curated genome to other sequences. It was shown that GeneMapper is able to predict gene structure with very high fidelity (with an accuracy of ~92% at the gene level and ~97% at the exon level) and compares favorably to other reference-based programs. In particular, GeneMapper effectively annotates genes that have undergone structural changes. In this paper, we use GeneMapper annotations of the ENCODE sequences to study the evolution of gene structure in mammalian lineages.

Previous studies on intron gain and loss have found widely diverse rates of intron gain and loss in various eukaryotic lineages [1,2]. In mammals, it has been observed that intron gain is very rare (or nonexistent) and our results on ENCODE regions are consistent with this previous study [12]. Some other studies have suggested that intron gain/loss follows gene duplication as gene duplication removes the selective constraints on the gene [13,14]. Most of these studies have been in nonmammalian lineages and have been phylogenetically sparse [15]. In this study, we have used the phylogenetically dense sequence coverage in the ENCODE regions to obtain a more complete elucidation of the relationship between intron gain/loss and duplication events. We show that, although gene structure changes may follow duplication events in mammalian lineages, there is also evidence that structural changes can precede duplication. Our results point to an explanation for the correlation between duplication propensity and structural plasticity based on a confounding factor.

Results

Human GENCODE annotations were used as reference to annotate the nonhuman sequences. These annotations were used to search for changes in gene structure in the mammalian sequences. A phylogenetic analysis of gene structure changes in the mammalian lineages was used to identify 11 genes with instances of intron loss (Table 1 and Supplementary Data Set 1). No intron gains were observed. Some genes were found to have lost more than one intron, resulting in 15 distinct cases of intron loss. A single instance of intron loss was detected in the bat lineage whereas the rest of the instances of intron loss were in the rodent (mouse/rat) and shrew lineages. A particularly interesting example is the gene AC018512.8 (a microfibrillar-associated protein), where the second and third introns were lost in the mouse lineage and the fourth intron was lost in the rat lineage. In fact, in this gene, introns are lost in this gene in fugu and zebrafish also. This example suggests the presence of "hotspots" for structural changes.

Rates of intron gain/loss

It is apparent from the results above that intron loss occurs at a much higher rate than intron gain in mammalian lineages. In fact, to the best of our knowledge, no instance of recent intron gain has been detected in mammalian lineages. It also appears that some lineages (such as rodents and shrew) have a much higher rate of gene structure change than other lineages (such as primates). The difference in rates might be related to differences in generation times. These observations are consistent with results in a previous study comparing the structure of human and mouse genes [12].

Mechanisms of intron loss

The classical theory of intron loss states that introns are lost by recombination of reverse-transcribed mRNA transcript with the genome [16]. Since reverse transcriptase operates from the 3′ to the 5′ end and may terminate prematurely, this theory predicts that more introns should be lost from the 3′ end than from the 5′ end. Because of the involvement of reverse transcriptase, this theory also predicts that introns should be lost in tandem. While we did not find that the lost introns show bias toward the 3′ end of genes, all the cases of multiple intron loss did occur in tandem. An alternative theory of intron loss hypothesizes that introns are lost by genomic deletion [17,18]. This theory predicts that intron lost is inexact and a small number of codons are added or lost from the flanking coding

Table 1
Intron loss events in the ENCODE regions

Gene	Species	Introns lost	Function
Psmd4	Rat	1	Modulates intestinal fluid secretion
Ddx18	Mouse, Rat	1	Putative RNA-dependent helicase
Irf1	Rat	1	Regulates MHC class I genes
Eef1a1	Shrew	2	Protein biosynthesis in ribosomes
Flna	Shrew	1	Actin binding protein
TAZ	Bat	1	Cardiolipin metabolism
G6pdx	Shrew	1	Nucleic acid synthesis
KIAA0404	Mouse, Rat	1	Unknown function
Mfap1a	Rat	1	Creatine kinase
Ckmt1	Mouse, Rat	3	Component of microfibrils
Atp50	Shrew	2	Component of F-type ATPase

sequence during intron deletion. However, all the intron losses in our data set are exact.

Gene expression

For a gene structure change to be passed on to subsequent generations, it has to occur in the germline. Indeed, it has been previously observed that genes expressed in the germline are more susceptible to gene structure change [3]. We obtained gene expression levels in 79 human and 61 mouse tissues from the GNF Gene Expression Atlas 2 [19]. For each gene with structure change, the maximum expression level across all germline tissues was compared to the median value across all tissues. It was found that all the genes had moderate to high expression levels in at least one germline tissue (more than 0.9 above the median on the log scale). Of these genes, four were highly expressed (more than 2 above the median on the log scale). It should be pointed out that it is possible that the genes with moderate expression levels might be expressed at high levels. This is because not all the genes were covered by the mouse experiments and the coverage of some other genes was incomplete. It is also possible that these genes are expressed in tissues that were not sampled in the experiments. Furthermore, some of the gene structure changes occurred in the rat, shrew, and bat lineages and expression levels might have changed in these species. In any event, the evidence seems to indicate that genes that have undergone structural change are expressed in at least moderate levels in germline cells.

Selection

Introns are believed to play a selective role in evolution [20,21], but their exact role is not elucidated. We therefore looked at selective constraints on genes undergoing structure changes. GeneMapper was used to create multiple alignments of all human genes and their orthologs in other species. We used these alignments to measure ω, the ratio of synonymous and nonsynonymous substitution rates for all genes undergoing gene structure evolution (Supplementary data set 2). The value of ω is a measurement of the nature of selection undergone by a gene. If $\omega << 1$, a gene is likely to be under purifying selection. On the other hand, a value of $\omega >> 1$ suggests that a gene is under positive selection. As the biological functions of most genes are expected to be conserved during evolution, genes are expected to be under purifying selection. All 11 genes with intron loss were under strong to moderate purifying selection ($\omega < 0.20$). In addition, six genes were under very strong purifying selection ($\omega < 0.05$). The ω values are similar to that expected in a typical gene. Therefore, it appears that gene structure changes are not related to any changes in selective constraint.

Gene duplication

It has been suggested that intron gain/loss is accelerated in genes with duplications as a result of a reduction in selective pressure [13,14]. If this hypothesis is true in mammalian lineages, most cases of intron loss should follow gene duplication. We tested this hypothesis by studying duplication events in the six genes with intron loss in mouse and rat using the complete genome sequences available for those species. We searched for homologs of each gene in human, mouse, and rat genomes using BLAT [22]. Four genes had multiple copies in at least one of the three genomes. For each gene with a homolog, the homolog with the highest sequence identity was identified as the one formed by the most recent duplication event. The locations of the duplication event and intron loss were then identified on the phylogenetic tree relating the three species. This association of gene duplication with intron loss is depicted in Fig. 1. It is interesting to note that, in two genes (Ddx18 and Mfap1a), the most recent duplication event occurred after the intron loss. In two other genes (Ckmt1 and Psmd4), the most recent duplication occurred in the human lineage (which had no structure change). We also found that all the genes are under strong or moderate purifying selection. Therefore, all available evidence indicates that intron loss does not occur due to relaxation of selection pressure (caused by duplication). Thus, while it may superficially appear that duplicated genes undergo intron loss [13], many of the duplication events may occur after intron loss.

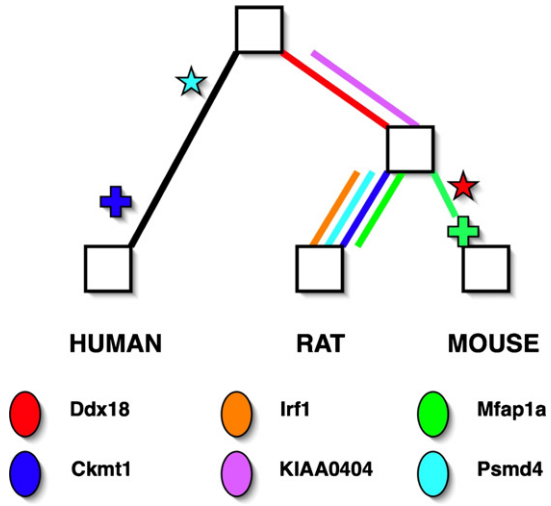

Fig. 1. Relationship between duplication events and intron losses for each gene. Each gene is assigned a separate color on the phylogenetic tree relating the species. Colored edges on the tree show when intron losses occurred. The colored stars show when retrotransposition events occurred for the associated gene, whereas colored plus signs indicate the occurrence of local duplication events for the gene. The locations of the symbols and edges indicate the relative order of the associated events. The gene Mfap1a (green) is interesting because intron loss occurred twice in separate introns (one in the mouse and the other in the rat). In the mouse lineage, both the loss of the intron and local duplication occurred after separation from the mouse/rat ancestor. Moreover, we were able to infer that the duplication event occurred after the intron loss.

Discussion

Our study provides direct evidence that gene structure changes are not caused by reduction of selection pressure due to duplication. In fact, we show that many duplication events occur after gene structure change. Our observationis also supported by the fact that all the genes with gene structure

change are under purifying selection ($\omega < 0.20$). This provides evidence for a common underlying cause for intron loss and gene duplication. We speculate that changes are induced by a mechanism mediated by reverse transcriptase. The fact that the genes that we identified are moderately to highly expressed in germline cells is also consistent with a reverse splicing mechanism.

Although the data set used in our study is seemingly small in that we find only 15 events, this is compensated for by the phylogenetically dense sampling of species with which to examine these events. For example, to illustrate the limitations inherent in using a phylogenetically sparse species set we mention the only previous gene structure evolution study in mammals which was done in human, mouse, and rat, with fugu as the out-group [12]. Analyzing the fifth intron of the gene Mfap1a where introns are lost in both fugu and rat, it is clear that, without more species, it is impossible to decide with confidence whether these events are due to parallel intron gains in human/mouse or intron losses in fugu and rat. A phylogenetic analysis of the gene structure in all the ENCODE species makes it clear that the scenario of two intron losses is the most parsimonious explanation. We were also able to make more subtle inferences about the mechanisms of intron loss than previously possible because of the quality of the ENCODE annotations and the associated expression measurements.

Our analysis has also resulted in a resource for the scale-up phase of the ENCODE project, namely the annotation of orthologs to ENCODE human genes in other species. These annotations are robust with respect to frameshift errors caused by poor sequence coverage and to structural changes of the type studied in this paper. We have also generated high-quality alignments of the GENCODE genes which should be a useful resource for other studies of gene function and structure. All these resources are publicly available in the supplementary website http://bio.math.berkeley.edu/genemapper/encode/.

Methods

Generation of orthology maps

An annotation pipeline was developed to generate GeneMapper annotations and gene alignments of the September 2005 version of ENCODE sequences. Mercator (Colin Dewey and Lior Pachter, unpublished) was used to create an orthology map relating the sequences in various species. The orthology map created by Mercator can be incomplete due to factors such as low sequence identity, incomplete anchor coverage, and microrearrangements. We therefore extended the Mercator orthology map by using extrapolation. For example, if an unmapped region had mapped regions both upstream and downstream, we looked for the orthologs of the unmapped region between the orthologs of its nearest mapped upstream and downstream region. In addition, we searched for orthologs in both strands which helped us detect inversions missed by Mercator.

Annotation and gene alignments

We downloaded GENCODE annotations of the human sequences from the UCSC browser [23]. The extended orthology map created in the previous step was used to determine the approximate location of the ortholog of each human GENCODE gene in the nonhuman species. GeneMapper was then used to annotate every nonhuman species by transferring the human GENCODE annotation. Gene alignments for each GENCODE gene and their orthologs were also created at this step. Details about the alignment algorithm are provided below.

Identification of gene structure changes

GeneMapper annotations were used to identify genes that underwent gene structure changes. All cases of putative gene structure change were manually verified for any discrepancies. For each instance of intron gain and loss, the presence/absence of the intron in each ENCODE species was used to label the leaves of the phylogenetic tree relating the species (obtained from Margulies et al., submitted to the ENCODE companion issue of Genome Research). A parsimony analysis was then used to locate intron gain and loss in the tree. In fact, gene structure changes are so rare that the location of gene structure change could be identified by manual inspection.

Gene alignments and sequence evolution

GeneMapper iteratively creates a gene profile of orthologous genes while transferring genes from the reference species to multiple target species. The gene profile is essentially an alignment of the reference gene (GENCODE gene in this case) and its orthologs in the other species. Therefore, the gene profile can be used to guide a gene alignment to study gene evolution. Unlike global alignment programs, GeneMapper carefully models the evolution of genes, taking into account the fact that they havea codon structure. The evolution of codons are modeled using 64*64 substitution matrices. Furthermore, GeneMapper uses exact dynamic programming while adding each ortholog to the gene profile. The profile-based approach was used to generate alignments for each GENCODE gene. GeneMapper alignments were then used to examine selection in genes undergoing gene structure change. For each such gene, we calculated ω, the ratio of synonymous substitutions to nonsynonymous substitution, between the reference human ENCODE gene and the ortholog in the species undergoing gene structure change. The substitution rates and ω values are available in Supplementary Data Set 3.

Acknowledgments

We thank Colin Dewey for providing Mercator maps of the ENCODE regions. We also thank the GENCODE and HAVANA teams for organizing the EGASP workshop during which we began work on this project. S.C. and L.P. were partially funded by NIH Grants R01:HG02632-1 and U01: HG003150-01.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.03.008.

References

[1] S.W. Roy, W. Gilbert, Rates of intron loss and gain: implications for early eukaryotic evolution. Proc. Natl. Acad. Sci. USA 102 (16) (2005) 5773–5778.
[2] I.B. Rogozin, Y.I. Wolf, A.V. Sorokin, B.G. Mirkin, E.V. Koonin, Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution, Curr. Biol. 13 (17) (2003) 1512–1517.
[3] A. Coghlan, K.H. Wolfe, Origins of recently gained introns in caenorhabditis, Proc. Natl. Acad. Sci. USA 101 (31) (2004) 11362–11367.
[4] S.W. Roy, W. Gilbert, The pattern of intron loss, Proc. Natl. Acad. Sci. USA 102 (3) (2005) 713–718.
[5] C.B. Nielsen, B. Friedman, B. Birren, C.B. Burge, J.E. Galagan, Patterns of intron gain and loss in fungi, PLoS Biol. 2 (12) (2004) e422.
[6] S.W. Roy, W. Gilbert, Complex early genes, Proc. Natl. Acad. Sci. USA 102 (6) (2005) 1986–1991.
[7] W.-G. Qiu, N. Schisler, A. Stoltzfus, The evolutionary gain of spliceosomal

introns: sequence and phase preferences, Mol. Biol. Evol. 21 (7) (2004) 1252–1263.

[8] H.D. Nguyen, M. Yoshihama, N. Kenmochi, New maximum likelihood estimators for eukaryotic intron evolution, PLoS Comput. Biol. 1 (7) (2005) e79.

[9] E.A. Feingold, P.J. Good, M.S. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, F.S. Collins, The encode (encyclopedia of dna elements) project, Science 306 (5696) (2004) 636–640.

[10] R. Guigo, P. Flicek, J.F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V.B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T.R. Gingeras, J. Harrow, T. Hubbard, S.E. Lewis, M.G. Reese, Egasp: the human encode genome annotation assessment project, Genome Biol. 7 (Suppl. 1) (2006) S2.1–S2.31 (1465-6914 (Electronic)).

[11] S. Chatterji, L. Pachter, Reference based annotation with genemapper, Genome Biol. 7 (4) (2006) R29.

[12] S.W. Roy, A. Fedorov, W. Gilbert, Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain, Proc. Natl. Acad. Sci. USA 100 (12) (2003) 7158–7162.

[13] C.I. Castillo-Davis, T.B.C. Bedford, D.L. Hartl, Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites, Mol. Biol. Evol. 21 (7) (2004) 1422–1427.

[14] H. Lin, W. Zhu, J. Silva, X. Gu, C. Buell, Intron gain and loss in segmentally duplicated genes in rice, Genome Biol. 7 (5) (2006) R41.

[15] M. Yandell, C.J. Mungall, C. Smith, S. Prochnik, J. Kaminker, G. Hartzell, S. Lewis, G.M. Rubin, Large-scale trends in the evolution of gene structures within 11 animal genomes, PLoS Comput. Biol. 2 (3) (2006) e15.

[16] L.B. Bernstein, S.M. Mount, A.M. Weiner, Pseudogenes for human small nuclear rna u3 appear to arise by integration of self-primed reverse transcripts of the rna into new chromosomal sites, Cell 32 (2) (1983) 461–472.

[17] W.J. Kent, A.M. Zahler, Conservation, regulation, synteny, and introns in a large-scale c. briggsae-c. elegans genomic alignment, Genome Res. 10 (8) (2000) 1115–1125.

[18] S. Cho, S.-W. Jin, A. Cohen, R.E. Ellis, A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution, Genome Res. 14 (7) (2004) 1207–1220.

[19] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, J.B. Hogenesch, Large-scale analysis of the human and mouse transcriptomes, Proc. Natl. Acad. Sci. USA 99 (7) (2002) 4465–4470.

[20] J.M. Comeron, M. Kreitman, The correlation between intron length and recombination in drosophila. dynamic equilibrium between mutational and selective forces, Genetics 156 (2000) 1175–1190 (0016-6731 (Print)).

[21] M. Lynch, The origins of eukaryotic gene structure, Mol. Biol. Evol. 23 (2) (2006) 450–468.

[22] W. Kent, Blat-the blast-like alignment tool, Genome Res. 12 (4) (2002) 656–664.

[23] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, W.J. Kent, The ucsc genome browser database, Nucleic Acids Res. 31 (3) (2003) 51–54.