Published in final edited form as:

Trends Genet. 2008 October ; 24(10): 475–478. doi:10.1016/j.tig.2008.08.003.

Nested genes and increasing organizational complexity of metazoan genomes

Raquel Assis¹, Alexey S. Kondrashov¹, Eugene V. Koonin², and Fyodor A. Kondrashov³

¹Center for Computational Medicine and Biology and the Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA

²National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

³Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093, USA

Abstract

The most common form of protein-coding gene overlap in eukaryotes is a simple nested structure, whereby one gene is embedded in an intron of another. Analysis of nested protein-coding genes in vertebrates, fruit flies and nematodes revealed substantially higher rates of evolutionary gains than losses. The accumulation of nested gene structures could not be attributed to any obvious functional relationships between the genes involved and represents an increase of the organizational complexity of animal genomes via a neutral process.

Introduction

Eukaryotes are typically more complex than prokaryotes on the molecular, systems and phenotypic scales of biological organization. In particular, genomes of multicellular eukaryotes possess a complex architecture that involves substantial overlapping of their transcribed regions [1–4] and protein-coding genes [5–7], forming an interleaving mosaic of exon and intron sequences. Although it is clear that such complex genome organization is made possible by the presence of introns, the rates and mechanisms of evolutionary events leading to gains and losses of overlapping gene arrangements have not been studied previously.

Previous studies of the evolution of genome complexity have primarily relied on correlations between the abundances of various genomic elements (introns, transposons, gene size, etc.) and the product of the effective population size and mutation rate [8–13]. However, claims of causality based on such correlative analyses are always inconclusive, because other potentially important factors can never be excluded. In an attempt to circumvent some of the limitations of the correlative approach, we explored the evolution of genomic complexity in a more direct manner, by tracing the evolutionary dynamics of nested pairs of protein-coding genes in animals. This study covers only one, perhaps not even the most common, class of interleaved gene arrangements because we left out the numerous intron-contained small RNA genes [14]. Nevertheless, even this limited analysis clearly reveals the ongoing increase of the organizational complexity of animal genomes and suggests that this process occurs via a nonselective route.

Evolutionary dynamics of nested gene structures

The most common form of overlap between protein-coding genes in eukaryotes is a nested gene structure, and in a majority of such structures, the internal gene lies entirely within one intron of the external gene [6,7]. Thus, we investigated the evolution of this class of nested gene structures in vertebrates, *Drosophila* and *Caenorhabditis*. A search of NCBI annotation records yielded 428, 815, 440 and 608 nested gene pairs in *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. briggsae* genomes, respectively. After eliminating gene pairs that might have been misannotated (see Supplementary Material), we arrived at sets of 128, 792, 429 and 233 nested gene pairs, respectively. Only a small minority of the protein sequences encoded by internal genes from each of these three major taxa show significant sequence similarity to internal genes products in the other two taxa (data not shown), suggesting that either these structures emerged independently and relatively late during evolution or that they were extensively and repeatedly lost.

By examining gene annotations and constructing sequence alignments, we identified the closest species with a completely sequenced genome in which each nested gene structure was absent. Absence of the nested structure in an appropriate outgroup species indicates its emergence (gain) in the respective lineage, whereas presence of the nested structure in the outgroup indicates its loss (Figure 1). Gains were found in all three taxa, with the emergence of 55 internal genes in at least 40 independent events in vertebrates, 52 internal genes in at least 48 events in *Drosophila* and 22 internal genes in as many events in *Caenorhabditis*. The rate of these acquisitions was approximately uniform throughout the course of evolution (Figure 2). By contrast, losses of nested gene structures were much rarer, with none detected in vertebrates, 17 in *Drosophila* and 2 in *Caenorhabditis*.

Acquisition of nested gene structures

At least four scenarios are plausible for the formation of a nested gene structure: (i) an internal gene can evolve by insertion of a DNA sequence into an intron of a pre-existing gene, (ii) an internal gene can evolve *de novo* from an intronic sequence of a pre-existing gene, (iii) a gene can become internal after an adjacent gene acquires an additional exon(s) or (iv) a gene can become internal after fusion of two genes that flank it from the opposite sides (Figure 3).

By comparing the gene structures and encoded protein sequences of internal and external genes to complete gene sets from the respective species, we deduced the mechanisms of formation of nested gene structures in vertebrates (Table 1). Nearly all nested gene structures seem to have emerged by insertion of a DNA sequence, which arose by gene duplication or retrotransposition, into an intron of a pre-existing gene. The origin of an internal gene was classified as a retrotransposition when it was intronless in a given species, whereas its non-nested ortholog in a sister species contained introns. A duplication at the DNA level was inferred when both the internal gene and a non-nested ortholog in a sister species had introns. In cases where the internal gene and a non-nested ortholog were both intronless, retrotransposition and duplication at the DNA level could not be discriminated. Five internal genes in humans are candidates for *de novo* origin from intron sequences (see Supplementary Data), including one gene with no sequence similarity beyond apes [*placenta-specific 4 (PLAC4)*] and another with no similarity beyond old world monkeys [*saitohin (STH)*] (Table S1). Analysis of the 12 recently sequenced *Drosophila* genomes showed that the majority of *de novo* genes originate in introns [15]. Consistent with this observation, we found 11 internal genes in *D. melanogaster* with no sequence similarity to any genes in the genome of the closely related *D. yakuba*. We did not identify any nested gene structures that evolved via the remaining two scenarios.

No functional significance of nested gene structures

At least three hypotheses could explain the parallel accumulation of nested gene structures in different taxa. First, a nested structure might confer a selective advantage because of a functional or co-regulatory relationship between its members [16–20]. Second, according to the transcriptional collision model, members of a nested gene structure could interfere with each other's transcription [21–23], resulting in alternative expression of these genes in different tissues or during different times in development. Finally, acquisition of a nested gene structure could be a neutral process [8–13,17], driven by the presence of numerous long introns that provide niches for insertion of genes. Each of these hypotheses leads to a distinct prediction about the relationship between the expression of internal and external genes in a nested pair. The functional co-regulation hypothesis predicts a positive correlation between levels of their expression in similar tissues, the transcriptional collision hypothesis predicts a negative correlation and the neutral hypothesis predicts no correlation.

To discriminate between these three hypotheses, we analyzed gene expression data from human and *D. melanogaster* genomes (see Supplementary Material). We compared correlations of gene expression in 109 and 752 nested gene pairs in humans and *D. melanogaster*, respectively, to 1000 random sets of 109 and 752 adjacent gene pairs from corresponding genomes. There was no significant difference in mean correlation coefficients of gene expression levels between nested and adjacent genes in either human (0.33 ± 0.03 for nested and 0.33 ± 0.0008 for adjacent pairs) or *D. melanogaster* (0.041 ± 0.014 for nested and 0.030 ± 0.00046 for adjacent gene pairs), which is consistent with the neutral hypothesis. The observation that external genes have substantially more and longer introns than average in the respective species (Ref. [7] and Supplementary Material) is also compatible with the neutral hypothesis. Furthermore, examination of the available functional information for nested gene pairs (Table S1) did not reveal any obvious connections [7]. Fixation of originally neutral or even slightly deleterious sequence segments, such as introns and transposable elements, through genetic drift acting in relatively small populations is a common phenomenon in eukaryotic evolution that might be partially responsible for the evolution of complex phenotypes [8–13]. The increase in organizational complexity of intron-rich genomes via emergence of nested gene structures seems to be another facet of this process.

Predicting the course of genome structure evolution

The neutral hypothesis implies that the preferential evolutionary gain of nested gene structures is caused by metazoan genomes being far from neutral equilibrium with respect to birth and death of intron-contained genes [8]. We estimated the rate of acquisition of nested gene structures as ~0.4, 0.9 and 0.2 events per million years in the *H. sapiens*, *D. melanogaster* and *C. elegans* lineages, respectively (see Supplementary Material). Because animal genomes currently contain ~500–800 nested gene pairs, these rates indicate that nested gene structures began to emerge ~1 billion years ago, perhaps concurrent with the substantial intron gain that apparently occurred at the onset of metazoan evolution [24]. These results suggest that metazoan introns are still far from saturation by internal genes and that the organizational complexity of metazoan genomes will continue to increase for many millions of years via the emergence of new nested gene structures. By the time metazoan genomes reach organizational complexity equilibrium, the overlap of functional elements is expected to be much greater than what we observe in extant taxa and will probably include numerous Russian doll-like nested structures. This process has already begun in fruit flies, with the *D. melanogaster* genome containing six cases where a nested gene structure is nested in another gene.

Conclusions and perspective

We have shown that the evolution of metazoan genomes is accompanied by a steady rise in the prevalence of nested arrangements of protein-coding genes, leading to increasingly complex genome architectures. In addition to nested protein-coding genes, animal genomes contain numerous complex arrangements, including incomplete overlaps of protein coding regions and their untranslated regions and various RNA genes [2–6,25]. In particular, a substantial fraction of microRNA and small nucleolar RNA genes are either fully contained within introns of protein-coding genes or overlap with protein-coding exons [14]. It will be of major interest to determine whether the trend of increasingly complex genome organization reported here applies to RNA genes or incompletely overlapping gene structures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Mironov AA, et al. Frequent alternative splicing of human genes. *Genome Res.* 1999; 9:1288–1293. [PubMed: 10613851]
2. Willingham AT, et al. Transcriptional landscape of the human and fly genomes: nonlinear and multifunctional modular model of transcriptomes. *Cold Spring Harb Symp Quant Biol.* 2006; 71:101–110. [PubMed: 17480199]
3. Makalowska I, et al. Overlapping genes in vertebrate genomes. *Comput Biol Chem.* 2005; 29:1–12. [PubMed: 15680581]
4. Kapranov P, et al. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet.* 2007; 8:413–423. [PubMed: 17486121]
5. Veeramachaneni V, et al. Mammalian overlapping genes: the comparative perspective. *Genome Res.* 2004; 14:280–286. [PubMed: 14762064]
6. Misra S, et al. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* 2002; 3:research0083.1–0083.22. [PubMed: 12537572]
7. Yu P, et al. Nested genes in the human genome. *Genomics.* 2005; 86:414–422. [PubMed: 16084061]
8. Lynch M, Conery JS. The origins of genome complexity. *Science.* 2003; 302:1401–1404. [PubMed: 14631042]
9. Lynch M. The origins of eukaryotic gene structure. *Mol Biol Evol.* 2006; 23:450–468. [PubMed: 16280547]
10. Yi SV. Non-adaptive evolution of genome complexity. *Bioessays.* 2006; 28:979–982. [PubMed: 16998825]
11. Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A.* 2007; 104:8597–8604. [PubMed: 17494740]
12. Lynch M. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A.* 2002; 99:6118–6123. [PubMed: 11983904]
13. Lynch, M. *The Origins of Genome Architecture.* Sinauer Associates, Inc; 2007.
14. Mattick JS, Makunin IV. Small regulatory RNAs in mammals. *Hum Mol Genet.* 2005; 14:R121–R132. [PubMed: 15809264]
15. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature.* 2007; 450:203–218. [PubMed: 17994087]
16. Henikoff S, Eghtedarzadeh MK. Conserved arrangement of nested genes at the *Drosophila* Gart locus. *Genetics.* 1987; 117:711–725. [PubMed: 3123310]
17. Habib AA, et al. The OMgp gene, a second growth suppressor within the NF1 gene. *Oncogene.* 1998; 16:1525–1531. [PubMed: 9569019]

18. Jaworski DM, et al. Potential regulatory relationship between the nested gene DDC8 and its host gene tissue inhibitor of metalloproteinase-2. *Physiol Genomics*. 2007; 28:168–178. [PubMed: 16985004]
19. Furia M, et al. Dense cluster of genes is located at the ecdysone-regulated 3C puff of *Drosophila melanogaster*. *J Mol Biol*. 1993; 231:531–538. [PubMed: 8510164]
20. Davies W, et al. Expression patterns of the novel imprinted genes Nap115 and Peg13 and their non-imprinted host genes in the adult mouse brain. *Gene Expr Patterns*. 2004; 4:741–747. [PubMed: 15465498]
21. Crampton N, et al. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Res*. 2006; 34:5416–5425. [PubMed: 17012275]
22. Osato N, et al. Transcriptional interferences *in cis* natural antisense transcripts of humans and mice. *Genetics*. 2007; 2:1299–1306. [PubMed: 17409075]
23. Da Lage JL, et al. A nested alpha-amylase gene in *Drosophila ananassae*. *J Mol Evol*. 2003; 57:355–362. [PubMed: 14629045]
24. Carmel L, et al. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol*. 2007; 7:192. [PubMed: 17935625]
25. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007; 447:799–816. [PubMed: 17571346]

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi: 10.1016/j.tig. 2008.08.003.

Figure 1.

Phylogenetic analysis of gains and losses of nested gene structures. Gain or loss of a nested gene structure must have occurred if, within a pair of sister species, the structure is present in one but absent in the other. **(a)** Absence of the nested structure in the outgroup indicates its gain in sister 1. **(b)** Presence in the outgroup indicates its loss in sister 2.

Figure 2.

Dynamics of gain and loss of nested gene structures. Gains and losses of internal genes are labelled on the (a) vertebrate, (b) *Drosophila* and (c) nematode phylogenies in red and blue, respectively. Nested gene structures that have a different nested state in the most distant outgroup, and therefore cannot be resolved between gains or losses, are shown in green. Independent events, or those that occur in different introns, are shown in parentheses. Events that could not be timed with a high enough resolution are shown on the side of each phylogeny.

Figure 3.

Scenarios for the origin of a nested gene structure. **(a)** Evolution of an internal gene by insertion of a DNA sequence into an intron of a pre-existing gene. **(b)** *de novo* evolution of a gene from an intronic sequence of a pre-existing gene. **(c)** Internalization of a gene after exon(s) acquisition of an adjacent gene. **(d)** Internalization of a gene via fusion of two flanking genes. Color key: pink, internal gene; green, exons of the external gene; blue and yellow, flanking genes.

Table 1

Mechanisms of origin of human internal genes

	After human – mouse	After human – opossum	After human – chicken	Total
Duplication	2	2	4	8
Retrotransposition	3	11	6 (5)	21 (20) ^a
Duplication or retrotransposition	2	3 (2)	16 (2)	21 (6)
<i>De novo</i> candidates	2	1	2	5

Independent events are shown in parentheses.

^aOne retrotransposition event was not dated to the degree of accuracy as other cases.