The relationship between non-protein-coding DNA and eukaryotic complexity

Ryan J. Taft, Michael Pheasant, and John S. Mattick*

Summary

There are two intriguing paradoxes in molecular biology—the inconsistent relationship between organismal complexity and (1) cellular DNA content and (2) the number of protein-coding genes—referred to as the C-value and G-value paradoxes, respectively. The C-value paradox may be largely explained by varying ploidy. The G-value paradox is more problematic, as the extent of protein coding sequence remains relatively static over a wide range of developmental complexity. We show by analysis of sequenced genomes that the relative amount of non-protein-coding sequence increases consistently with complexity. We also show that the distribution of introns in complex organisms is non-random. Genes composed of large amounts of intronic sequence are significantly overrepresented amongst genes that are highly expressed in the nervous system, and amongst genes downregulated in embryonic stem cells and cancers. We suggest that the informational paradox in complex organisms may be explained by the expansion of *cis*-acting regulatory elements and genes specifying *trans*-acting non-protein-coding RNAs. *BioEssays* 29:288–299, 2007. © 2007 Wiley Periodicals, Inc.

Introduction

The relationship between genomic information, DNA content and biological complexity has been a subject of considerable interest since the dawn of the genetic age.[1,2] The definition of biological complexity is itself a matter of debate, but may be considered a combination of metabolic and developmental complexity, the latter broadly defined as the number[3] and different types of cells, and the degree of cellular organization.

ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Australia.

Funding agency: We thank the Australian Research Council for financial support. RJT is supported by a United States National Science Foundation Graduate Research Fellowship.

*Correspondence to: John S. Mattick, ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072, Australia. E-mail: j.mattick@imb.uq.edu.au

DOI 10.1002/bies.20544
Published online in Wiley InterScience (www.interscience.wiley.com).

There is generally a close relationship between genome size and genetic capacity in the prokaryotes, whose genomes are haploid and primarily composed of closely packed protein-coding sequences. Bacterial genome sizes vary from obligate endosymbionts such as *Carsonella ruddii* (0.16 Mb, ~182 protein-coding genes)[4] which rely heavily on host functions, to free-living bacteria with considerable metabolic and catabolic capacity, such as *Burkholderia xenovorans* (9.7 Mb, 8,602 protein-coding genes).[5] With the exception of a few species whose genomes appear to be in the process of degradation, prokaryotes contain relatively low amounts of non-protein-coding sequences, on average ~12%, which mainly function as 5′ and 3′ regulatory elements that control gene expression at the transcriptional and translational levels, and also encode a limited number of infrastructural or regulatory RNAs.[6]

In contrast to the prokaryotes, there are major incongruities between both cellular DNA content and the number of protein-coding genes in relation to developmental complexity in the eukaryotes, whose genomes contain larger amounts of intragenic (i.e. intronic) and intergenic non-protein-coding sequences, totaling almost 98% in humans. These sequences have been considered, in the main, to be genetically inert. Coupled with the presence of large amounts of 'repetitive', usually transposon-derived, sequences in animal and plant genomes, this has led to the conclusion that complex eukaryotes can happily tolerate, and have either retained or accumulated, variable amounts of "junk" or "selfish" DNA,[7–9] in contrast to rapidly dividing, free-living cells that have been under sustained pressure to streamline their genomes.[10] The possibility that these sequences may themselves have function has, with some exceptions,[11] rarely been considered, due to the underlying assumption that most genes are synonymous with proteins, and the accompanying expectation that the number of (protein-coding) genes embedded in these otherwise junk-laden genomes would emerge as the underlying dictate of biological complexity.[12] The latter has now been contradicted by genome sequencing, which raises the question of where in genomes the information that underpins biological function and developmental complexity lies, and how this information scales with increasing complexity.

This article reviews the relationship of cellular DNA content and gene number to organismal complexity. We show that incongruities in the former are due primarily to varying ploidy in different lineages. We suggest that the lack of correlation in the latter is explained by the evolution of an increasingly sophisticated architecture to control gene expression and gene product diversity during multicellular ontogeny, involving both the expansion of *cis*-acting regulatory elements acting at multiple levels (e.g. chromatin architecture, transcription, splicing, RNA modification and editing, mRNA translation and RNA stability), and the expansion of non-protein-coding genes specifying regulatory RNAs that fulfill a wide range of functions in cell and developmental biology.

The C-Value paradox

The C-value paradox or enigma refers to the historical observation that the amount of cellular DNA in different organisms does not correlate with their relative biological complexity, or at least that there appear to be significant exceptions to such correlations. Iconic examples include amphibians and amoebae, which have much more DNA per cell than mammals.[13]

The term C-value was first introduced in 1950 by Swift,[14] and the C-value paradox nearly a quarter of a century later by Thomas.[15] There has been considerable uncertainty regarding the meaning of 'C-value' which has had a rather loose definition ranging from the complete complement of DNA per nucleus to the haploid genome size, which has in turn led to some confusion.[16] It is important to note, however, that until recently the C-value was biochemically ascertained and measured in picograms, and as such was generally a crude estimate that did not give any insight into the sequence composition or nature of the genetic information in the genome concerned. It is also now clear that there are at least four variables affecting estimations of the C-value—the ambiguity of the term, polyploidy, repetitive sequences and experimental errors—of which polyploidy may be the most significant.[2]

Polyploidy is rare in mammals,[17] but relatively common in amphibians and fish,[18] the vertebrate classes with the highest known C-values. However, catalogued C-values are not usually corrected for ploidy.[13] For example, the C-value for the salamander genus *Ambystoma*[19] has been taken to indicate a genome size 10–25 times larger than other vertebrates[20] although polyploidy is known in *Ambystomatidae*,[21] and a recent genetic map suggests that the salamander genome may not be greatly dissimilar in size to other vertebrate genomes.[20] The lungfish also has a high C-value, superficially suggesting that its genome is over an order of magnitude larger than primates, but is again known to be polyploid.[22] Other groups of organisms that exhibit a wide range of C-values, such as crustaceans and insects, are also frequently polyploid.[23]

There may also be significant measurement errors stemming from different experimental methodologies, interfering compounds and physiological states.[24–26] For example, there are widely differing estimates of the DNA content of the lungfish *Proteopterus aethiopicus*, with measurements ranging from 40 to 130 pg.[24]

In addition to lungfish and amphibians, amoebae are often cited as the most-dramatic example of the lack of correlation between genome size and biological complexity.[27] There are many problems with this conclusion, including a likely variation in ploidy, since some amoeba have hundreds of chromosomes,[28] which may be generally related to cell size, not just in protists but in many different organisms,[29] as well as the presence of significant amounts of contaminating DNA from their prey.[24] The amoeba genome is probably smaller than 20 pg, far less than the 700 pg commonly cited,[24] but the effective or haploid genome size of such protists may be in fact far lower in light of the recently sequenced *Paramecium tetraurelia* genome, which is only 72 Mb (i.e. approximately ~0.07 pg).[30] There is also wide variation in C-values and ploidy in plants,[23,31] which has led to the suggestion that C-value might be best defined as "mean basic" genome size, i.e. the size of a single unique copy of the genome.[31]

In this context, it is worth noting that a widely accepted measure of relative complexity is the minimum amount of information required to specify the ontogeny and operation of a system (referred to as Kolmogorov or program-size complexity).[32] Interestingly, the *minimum* genome sizes observed across metazoa show a consistent increase with complexity from simple nematode worms to insects to vertebrates. This holds true for both biochemical and genome sequence measurements, as well as for computationally compressed genome sizes (Table 1). Importantly there is no clade of organisms whose minimum genome size is not substantially larger than the minimum observed in a clade of lower complexity.

It is clear that gene, segmental or whole genome duplication is one of the two major sources of raw evolutionary material for genome expansion and the generation of new functions.[33] This has been well documented in yeast where there appears to have been at least one round of whole genome duplication, followed by large-scale but highly selective gene and genomic sequence losses, retaining genes that evolved useful alternative (non-redundant) paralogous functions and shedding many of those that were redundant.[34] Similar observations have been made in flowering plant *Arabidopsis thaliana*.[35] There is also evidence that the vertebrates have undergone at least one round of duplication compared to other metazoans, which may have provided the platform for their subsequent evolution.[36] However, the extent to which genome duplication via polyploidy has contributed to greater functional complexity, as opposed to allelic diversity, in extant organisms is difficult to assess,

Table 1. Minimum genome size per clade

	Biochemical genome measurements		Sequenced genomes		
Clade	**Mb**	**Common name**	**Mb**	**Compressed size (Mbits)**	**Common name**
Nematodes	29	Root-knot nematode	97	240	Nematode
Insects	88	Hessian fly	118	312	Fruitfly
Non-vertebrate chordates	68	Larvacean	159	344	Sea squirt
Fish	342	Green pufferfish	385	776	Pufferfish
Amphibians	929	Ornate burrowing frog	1,700	3,312	Frog
Birds	949	Common pheasant	1,047	2,544	Chicken
Reptiles	1,027	Skink			
Mammals	1,692	Bent-winged bat	2,445	5,952	Dog

Minimum genome size per clade as an indicator of the minimum information required to specify a representative of the clade. Biochemical assessments, taken from the Animal Genome Size Database,[13] sample a large number of species but are affected by a number of factors (see text). Figures for sequenced genomes, obtained through the UCSC genome browser (see Hinrichs AS *et al.* 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34:D590–598) are more accurate but far fewer species have been sampled. The compressed size in megabits is an approximation of the Kolmogorov complexity of the genome and is calculated as the size of the gzip compressed genome sequence file size.

especially, for example, when one compares different protists, amphibians with mammals, or wheat with other grasses.

The other major source of raw evolutionary material and of variation in genome size is transposon-derived sequences,[37–40] which comprise almost half of the human and mouse genomes[41,42] (Table 2). These sequences are often but somewhat pejoratively referred to as 'repetitive' sequences. Nonetheless, there is increasing evidence that at least some have acquired functions (for examples see Refs [43–46]), but it remains unclear what proportion have done so and therefore what contribution they make to genetic complexity in different lineages.[29,47]

However, it is also clear that deletion of sequences occurs at a significant frequency over evolutionary time.[29,48] Therefore, any extant genome sequence must reflect a balance between (i) sequence acquisition by duplication and transposition, a proportion of which will have been fixed by positive and subsequently negative selection following acquisition of function, and (ii) sequence loss, which will be largely restricted to those that have not yet acquired function or that are functionally redundant. Again, what proportion of these sequences fall into these categories is presently impossible to determine. One may safely predict, however, that the more ancient a sequence, the more likely it will have suffered one or the other fate (i.e. acquired function or have been deleted), and, therefore, the more likely that those that remain are functional. This may throw doubt on the use of ancient 'repeats' (transposon-derived sequences) as an index of the rate, and

Table 2. Genomic feature comparison across eight species

Species	**CDS Mb (%)**	**UTR Mb (%)**	**Intron Mb (%)**	**Intergenic Mb (%)**	**Total Genome size Mb**	**Repeats Mb (%)**
D. discoideum	21.1 (61.7)	NA	2.5 (7.4)	10.6 (30.9)	34.2	3.4 (10.0)
C. elegans	25.4 (25.3)	2.2 (2.2)	30.4 (30.4)	42.3 (42.1)	100.3	12.9 (12.9)
D. melanogaster	21.8 (16.6)	6.5 (4.9)	38.4 (29.1)	65.1 (49.4)	132.0	16.2 (12.3)
T. nigroviridis	30.5 (8.9)	NA	190.1 (55.5)	121.9 (35.6)	342.0	10.3 (3.0)
G. galus	25.0 (2.4)	4.9 (0.5)	345.0 (32.7)	679.0 (64.4)	1054.0	104.3 (9.9)
M. musculus	27.5 (1.1)	25.8 (1.0)	757.0 (29.3)	1773.0 (68.6)	2583.0	1092.6 (42.3)
H. sapiens	31.7 (1.1)	30.0 (1.1)	1009.0 (35.2)	1795.0 (62.6)	2866.0	1391.0 (48.5)

Sequence analysis was accomplished using the featureBits program of the UCSC genome browser for each species (*C. elegans* = ce2, *D. melanogaster* = dm2, *C. intesinalis* = ci1, *T. nigrovirdis* = tetNig2, *G. galus* = galGal1, *M. musculus* = mm7, and *H. sapiens* = hg17) and the corresponding gene annotation track (dictyBase, NCBI refGene, flyBaseGene, Genscan, Ensemble ensGene, UCSC knownGene, UCSC knownGene). In the case of *Dictyostelium*, we constructed a partial UCSC browser using genome sequence and gene annotations obtained from dictyBase (http://www.dictybase.org/); repeats were identified from the relevant literature. It should be noted that exon discovery will not significantly affect these data, other than altering proportions of sequence in either intronic or intergenic bins. NA, data Not Available.

variance of the rate, of neutral evolution in mammalian genomes.[42]

The G-value paradox

As noted earlier, the apparent inconsistencies between DNA content and organismal complexity, leaving aside the fact that some may be ascribed to measurement errors and polyploidy, were widely assumed to reflect the likelihood that the relevant genetic information was obscured by the presence of variable amounts of non-functional evolutionary debris in different species or lineages. Reciprocally, it was assumed that the relevant indicator, the underlying number of 'genes', would in fact scale appropriately with complexity.[12] This depends on the definition of a gene,[49] which has clearly evolved over recent decades,[40] but has in the main been taken to mean protein-coding sequences and associated regulatory elements, on the general expectation that most genetic information is expressed as and transacted by proteins, and that proteins carry out most cellular functions, including most regulatory functions.

The expectation that increased developmental complexity would be reflected in an increased number of protein-coding genes has not been borne out, and has been termed the G-value paradox.[49] Predictions of the estimated number of protein-coding genes in the human genome prior to genome sequencing ranged from as low as 50,000 to as high as 140,000,[50] whereas the latest estimates from genome analysis indicate that humans have approximately 20,000 protein-coding genes,[51] similar to other vertebrates such as the mouse,[42] chicken[52] and pufferfish.[53] Unexpectedly, the nematode worm *Caenorhabditis elegans,* which comprises only 1,000 cells, has ~50% more annotated protein-coding genes (~19,300 genes)[54] than the far more complex insects (~13,500 genes)[55] and nearly as many genes as currently estimated for vertebrates. Indeed, there is no more than 50% variation in the extent of annotated protein-coding sequences between slime molds, nematodes, insects and vertebrates (Table 2). Moreover, despite their considerable developmental and neurological complexity, mammalia do not appear to have any more, and at present are predicted to have less, protein-coding genes than plants such as *A. thaliana* (latest estimate ~26,000)[56] and rice (~37,000)[57] or protists such as *Paramecium tetraurelia* (~40,000)[30] and *Tetrahymena thermophila* (~27,000)[58] (Fig. 1B).

Thus, although these estimates may have significant errors, particularly in plants,[59] and are still being refined[57,60] (see below), whatever differentiates humans from fish, and vertebrates from worms and protists, does not presently appear to be reflected in the extent of protein-coding sequences or the number of protein-coding genes (Table 2, Fig. 1B). This post-genomic realization has been termed the gene number, or G-value, paradox.[49] Part of this paradox can be explained by an increased utilization of alternative splicing, which allows a greater range of protein isoforms to be expressed,[61] which clearly occurs in the complex organisms, although this in turn necessitates an increase in regulation.

Indeed, there is ample evidence that complex organisms utilize a wide range of gene regulatory processes in addition to alternative splicing, including chromatin architecture, promoter selection, RNA modification and editing, RNA localization, translation, and RNA stability, among others. Recent studies have suggested that regulatory information scales more than linearly with function in organisms and other integrated systems, and may ultimately come to dominate their information content.[62,63] Most, if not all, of this regulatory information would be expected to reside outside of protein coding sequences.

The expansion of non-protein-coding sequences in eukaryotic genomes

We have previously shown a strong correlation between the amount of non-protein-coding sequence and biological complexity by examining the ratio of non-protein-coding DNA to total genome size (nc/tg), which intrinsically corrects for varying ploidy or partial ploidy.[62,64] For the purposes of this study, we were primarily concerned with broad increases in complexity between different classes of organisms, such as prokaryotes and eukaryotes, single cell and simple multicellular organisms, and invertebrates and vertebrates. However, this correlation itself exhibited some incongruities, including the initial annotation of the mosquito (*Anopheles gambiae*) genome which indicated an unusually high proportion of non-coding sequences (93%), clustering it with vertebrates, rather than with other insects. The data for eukaryotes has since been reassessed and, due largely to improvements in the resolution and annotation of sequenced genomes, shows a more consistent scaling, in line with our earlier predictions (Fig. 1A).

Interestingly, the annotation history of completed genomes shows that genomes, such as human, which were predicted to have high gene counts were initially annotated correspondingly, followed by a gradual decrease.[60] Conversely, there is some indication that there has been an increase in protein-coding gene number in genomes where initial predictions, conditioned by corresponding expectations, were low.[55,65] In the case of *Anopheles*, recent EST sequencing and a re-evaluation of its genome size suggest that its genome is in fact 81–89% non-coding, congruent with other arthropods.[66,67] Likewise, the silkworm *Bombyx mori* [68] has an nc/tg value of 0.8. The increased annotation fidelity of vertebrate genomes reveals a clustering at the high end of the nc/tg spectrum, which includes the compact puffer-fish genomes.[53,69] We suspect that recently sequenced genomes from lower eukaryotes, which are currently predicted to have high

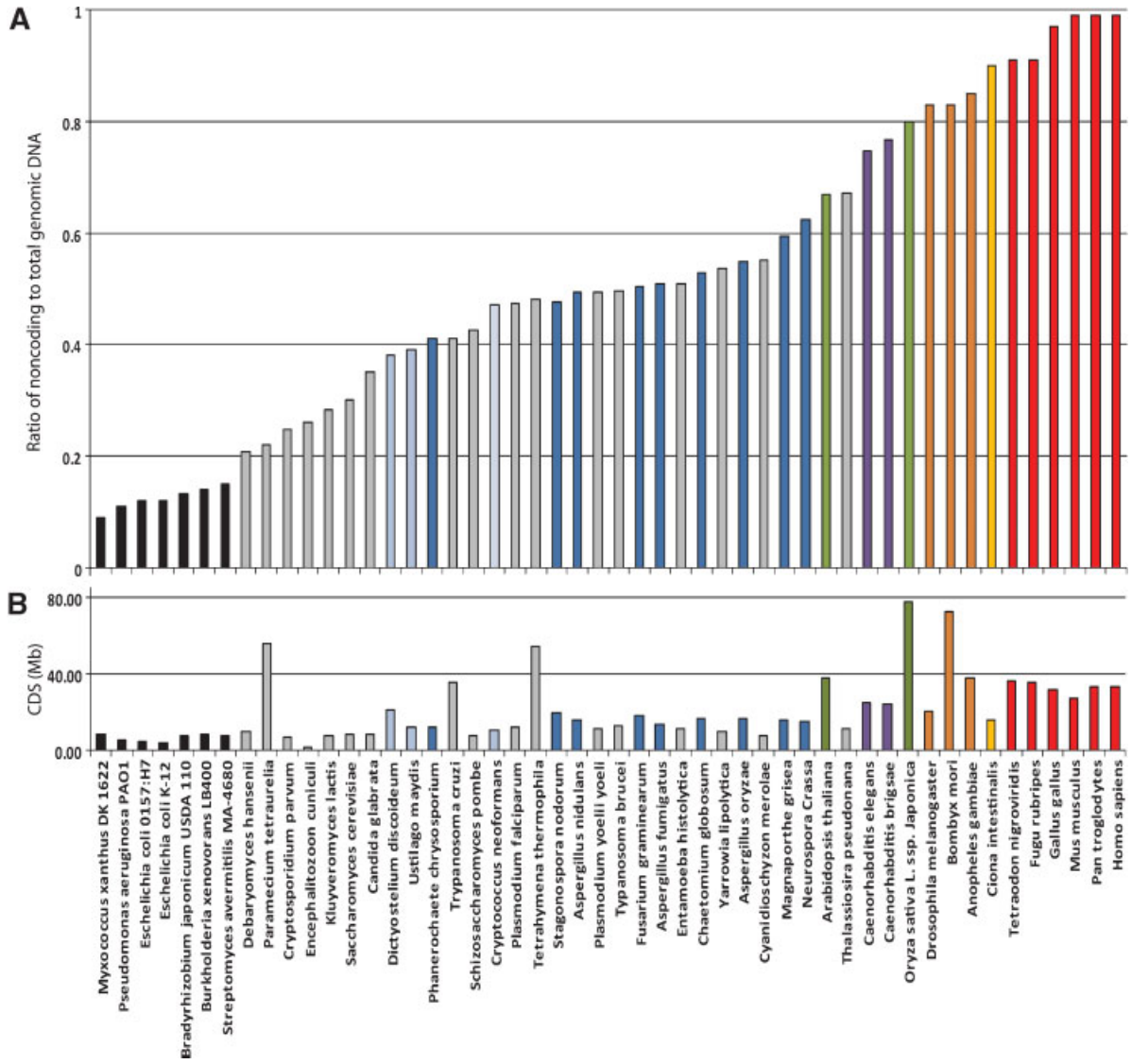

Figure 1. The fraction of non-protein-coding DNA and megabases of protein-coding sequence (CDS) per haploid genome across species. **A:** The ratio of the total bases of non-protein-coding to the total bases of genomic DNA per sequenced genome across phyla (i.e. the percent ncDNA). The four largest prokaryote genomes and two well-known bacterial species are depicted in black. Single-celled organisms are shown in gray, organisms known to be both single and multicellular depending on lifecycle are light blue, basal multicellular organisms are blue, plants are green, nematodes are purple, arthropods are orange, chordates are yellow, and vertebrates are red. Species names are listed below section B of the figure. **B:** The amount (in megabases) of CDS per genome for species ranked by fraction of non-protein-coding DNA. A description of the calculation methods and references for each genome can be obtained upon request.

proportions of non-coding sequence, for example the diatom *Thalassiosira pseudonan,*[70] are mis-annotations due insufficient EST coverage and to the fact that many gene prediction algorithms have been trained on higher eukaryotes or prokaryotes.[71] Moreover, two sequenced protist species, *Paramecium tetraurelia* and *Tetrahymena thermophila*, despite having an unexpectedly large amount of predicted protein-coding sequences (56 Mb and 55 Mb, respectively) and genomes nearly as large as *C. elegans*,[30,58,72] have nc/tg ratios that place them broadly in the same cluster as other unicellular eukaryotes (Fig. 1).

Intron size and distribution

Despite the fact that introns, consistent with the nc/tg trend discussed above, are much larger in developmentally complex organisms (Table 3), are transcribed, include significant

Table 3. Intron comparison across eight species

	Number of genes			Introns	Intron size		
Species	**total**	**with introns**	**%**	**per gene**	**median**	**mean**	**max**
D. discoideum	16362	11386	69.6	1.53	104	145	2842
C. elegans	19957	19413	97.3	5.46	67	315	103002
D. melanogaster	14034	11467	81.7	4.22	82	1179	132737
*C. intestinalis**	246	222	90.2	7.06	339	826	161573
T. nigroviridis	28060	21219	75.6	8.07	281	1110	150886
G. galus	24747	23476	94.9	6.93	733	2347	675663
M. musculus	18908	17031	90.1	9.84	1360	5413	996015
H. sapiens	20181	17753	88.0	9.10	1609	6486	1096453

The number of genes, introns with genes and introns per gene were calculated using the UCSC genome browser for each species and the corresponding gene annotation track (see Table 2 for more detail). For these analyses, we considered introns obligate CDS and UTR free regions within a known protein-coding transcriptional locus. Intron size analysis was accomplished using the UCSC genome browser and the corresponding mRNA track. Unique introns from cataloged mRNAs were analyzed, yielding a transcript-based (as opposed to genomic location based) description of intron distribution. It should be noted that these data are likely affected by the amount of EST and mRNA coverage (e.g., *Drosophila* has much higher coverage than *T. nigroviridis).* Descriptive statistics are also affected by the fact that introns sizes for large genomes are right-skewed. *The current data for *C. intestinalis* are limited, and are included here for the purposes proving an estimate of Ciona's intronic landscape.

amounts of conserved sequences, and house all known small nucleolar RNAs and a large fraction of microRNAs, they are still thought to be largely devoid of important genetic information. Correspondingly, their presence has been commonly rationalized as either the remnants of the early assembly of genes and have been subjected to minimal pressure for their removal in complex organisms compared to microorganisms—the ''introns-early'' hypothesis—and/or the result of the increased capacity of multicellular organisms to accumulate evolutionary debris from transposons and other sources—''introns late''.[10,73–75]

In either case, there has been little suggestion that the retention and/or expansion of introns in complex organisms is due to selection for functions encoded within them,[11,74] and that, with the exception of a limited amount of obviously conserved sequence (currently estimated at about 5%), they are presumed to be evolving neutrally.[76] This presumption makes a specific prediction, which the availability of multiple whole genome sequences now makes it possible to test. If introns are largely genetically inert, their lengths should be relatively random within different types of genes, resulting from either primordial events in gene assembly or the random accumulation of sequence from transposon insertions, etc. However, this is not what is observed. We examined the relationship between the total intronic sequence (TIS) within annotated protein-coding genes and their functions as defined by their Gene Ontology (GO) designations. (GO is a formalized consistent vocabulary, whose category terms are related to one another in a hierarchical manner, which is used to describe the biology of a gene product in three independent dimensions: (i) its molecular function, (ii) the biological process in which it participates, and (iii) the cellular component wherein it is located.)

We found significant GO category enrichment for genes with the highest and lowest TIS lengths in human, mouse, fly and worm (Fig. 2). To explore if this effect was driven by repeats, we examined the most-repeat-dense genome of our set, human. GO enrichment was not significantly affected by repeat removal: greater than 85% of GO terms are identical between repeat-included and repeat-removed enrichment sets (Table 4). GO enrichment trends are also unaffected by repeats in the mouse genome, or if genes are examined using the ratio of cumulative intron to coding sequence (CDS) size, which corrects for the size of protein-coding exons (data not shown). Human enrichment sets show that large TIS genes are strongly associated with neural functions and processes. Similar enrichment is observed in large introns gene sets in mouse and fly, but not worm (Table 5). We speculate that this is due to the relatively small amount of neural tissue and GO annotations ($\sim$6500) in *C. elegans* (compared to $>$15,000 in human).[77] Conversely, when examining across species for small TIS GO enrichment, we find terms for ribosomal processes and functions in human, mouse and worm, although not in fly (Table 5). This may be due to the depth of *Drosophila* GO annotation. If it is validated that ribosomal related genes generally have small TIS compared to their host gene sets, it would argue strongly against the hypothesis that intronic sequences can tolerate indiscriminate insertion and therefore generally increase in size over time.

GO annotation is a relatively blunt tool, limited by electronic annotation, curation bias and GO ID term ambiguities and relationships. In order to examine if there was systematic tissue bias based on TIS size, we obtained the publicly available Genomics Institute of the Novartis Research Foundation (GNF) Expression Atlas 2 and sets for human

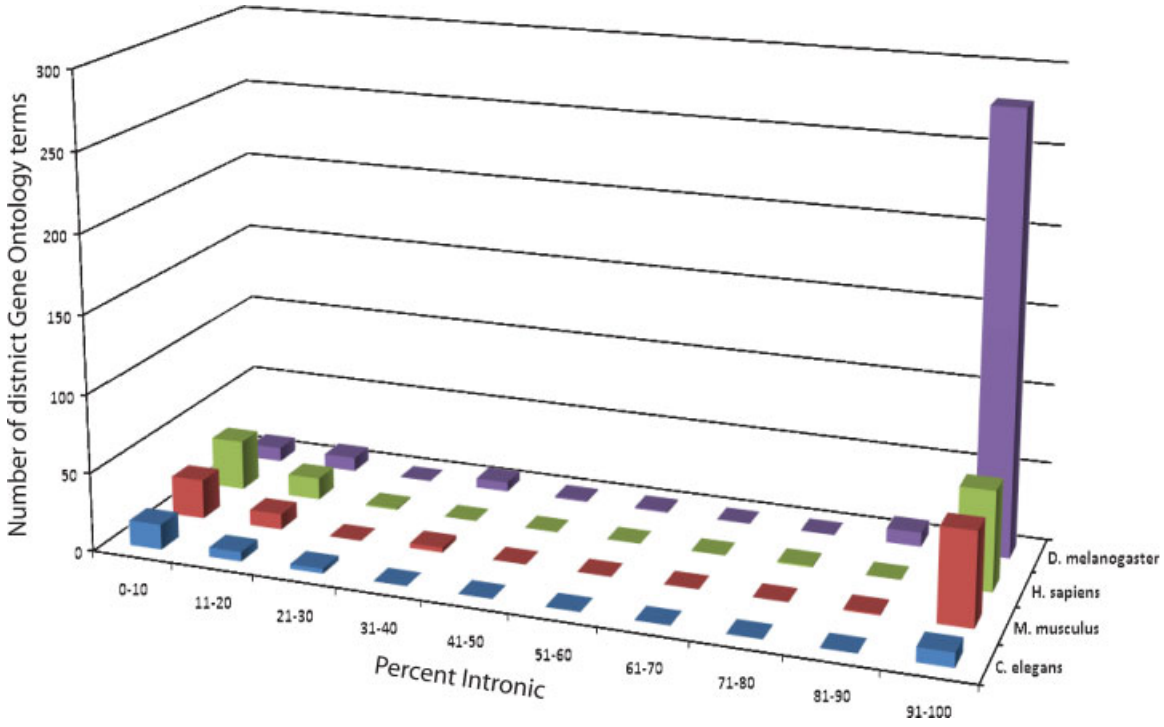

Figure 2. Protein-coding genes binned by percent total intronic sequence (TIS) show a non-random distribution in the number of statistically significant Gene Ontology terms per decile. An equal number of genes were evaluated for each decile based on TIS for human, mouse, fruit fly, and round worm (e.g. assuming ~20,000 genes in the human genome, ~2,000 were analyzed in each decile). Clusters of enriched terms are evident for genes with very little (i.e. 0–20%) and large amounts (i.e. 91–100%) of TIS. Gene Ontology enrichments were determined using a Fisher's Exact test, using threshold criteria of two fold enrichment and a *P* value of less than 10^{-6}.

Table 4. GO enrichment for largest TIS human genes

GO ID	GO term
GO:0007215	glutamate signaling pathway
GO:0007409	axonogenesis
GO:0030182	neuron differentiation
GO:0031175	neurite morphogenesis
GO:0048667	neuron morphogenesis during differentiation
GO:0007417	central nervous system development
GO:0000904	cellular morphogenesis during differentiation
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
GO:0048699	neurogenesis
GO:0007167	enzyme linked receptor protein signalling pathway
GO:0030029	actin filament-based process
GO:0030036	actin cytoskeleton organization and biogenesis
GO:0048731	system development
GO:0007399	nervous system development
GO:0006813	potassium ion transport

GO terms which are at least twofold enriched (and a Fisher's Exact *P* value $<10^{-6}$) for the 10% of human genes with the largest TIS measured with or without repeats.

(including the Gladstone array dataset) and mouse,[78] the Arbeitman et al. life cycle expression set for fruit fly,[79] and the Kim et al. life cycle set for *C. elegans*[80] through the UCSC Genome Browser. Using tissue type or life cycle stage as a surrogate for gene ontology type, we used a Fisher's Exact test to assess if the large or small TIS genes were over-represented among genes highly or lowly expressed in a tissue type.

Strikingly, both large and small TIS genes yield significant enrichment biases in the human GNF Atlas 2 data sets. In agreement with the GO analyses, large TIS genes are significantly over-represented among highly expressed genes in nervous system tissues, as well as in the uterus (Table 6). Interestingly, large TIS genes are over-represented amongst under-expressed genes in several cell types, including immunologic, embryonic stem and cancer cells, which are undifferentiated and pluripotent, or de-differentiated.[80] Small TIS genes, however, are over-represented among highly expressed genes in heart, bone marrow, lung, and pancreatic islets (data not shown). The enrichment trends for large TIS genes are consistent in mouse (data not shown), although no significant tissue or life cycle expression patterns could be gleaned from the *D. melanogaster* or *C. elegans* datasets. We

Table 5. GO enrichment for genes with respect to total intronic sequence across 4 species

GO Terms common to the largest 10% TIS genes

Organism	GO term category	GO identifier	GO term
D,H,M	biological process	GO:0000904	cellular morphogenesis during differentiation
D,H,M		GO:0006468	protein amino acid phosphorylation
C,D,H,M		GO:0007155	cell adhesion
D,H,M		GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
D,H,M		GO:0007268	synaptic transmission
D,H,M		GO:0007399	nervous system development
D,H,M		GO:0007409	axonogenesis
D,H,M		GO:0019226	transmission of nerve impulse
D,H,M		GO:0030182	neuron differentiation
D,H,M		GO:0031175	neurite morphogenesis
D,H,M		GO:0048667	neuron morphogenesis during differentiation
D,H,M		GO:0048699	neurogenesis
D,H,M		GO:0048731	system development
D,H,M	cellular component	GO:0005578	extracellular matrix (sensu Metazoa)
D,H,M		GO:0031012	extracellular matrix
D,H,M		GO:0045202	synapse
D,H,M	molecular function	GO:0003779	actin binding
D,H,M		GO:0004672	protein kinase activity
D,H,M		GO:0004674	protein serine/threonine kinase activity
D,H,M		GO:0004713	protein-tyrosine kinase activity
D,H,M		GO:0005083	small GTPase regulator activity
C,H,M		GO:0005085	guanyl-nucleotide exchange factor activity
C,D,H,M		GO:0005216	ion channel activity
D,H,M		GO:0005261	cation channel activity
D,H,M		GO:0008092	cytoskeletal protein binding
C,D,M		GO:0015268	alpha-type channel activity
D,H,M		GO:0030695	GTPase regulator activity

GO Terms common to the smallest 10% TIS genes

Organism	GO Category	GO ID	GO Term
C, H, M	molecular function	GO:0003735	structural constituent of ribosome
C, H, M	cellular component	GO:0005840	ribosome
C, H, M		GO:0030529	ribonucleoprotein complex

We used the latest builds of the human (*Homo sapiens*, hg17), mouse (*Mus musculus*, mm7), fruit fly (*Drosophila melanogaster*, dm2) and nematode (*Caenhorabditis elegans*, ce2) genomes in the UCSC genome database as our initial datasets. Protein-coding genes for each genome were obtained from the hg17.knownGene, mm7.knownGene, dm2.flyBaseGene and ce2.sangerGene SQL tables respectively. These genomes and gene annotations sets were chosen due to their well-curated protein-coding sequences and Gene Ontology annotations. Alternative isoforms were grouped to one canonical gene (most 5′ start and most 3′ stop of clustered transcripts) according to the provided annotations. Intronic regions were defined as bases within the genomic bounds of a canonical gene that were never annotated as CDS or UTR in any isoform, and aggregated to give a total. We also excluded regions annotated as CDS or UTR by the NCBI RefSeq gene annotation set in each genome.[126] In a small number of cases, genes were removed from the analysis if they mapped to more than one locus. Genes were then binned into deciles based on their total intronic sequence (TIS) size, and these bins were assessed for possible Gene Ontology (GO) enrichment.[77] Gene Ontology structure and vocabulary were obtained from the GO website (http://www.geneontology.org). Annotations for each species were taken from the Gene Ontology Annotations (GOA) database at the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/). Human and mouse Known Genes annotations were derived from the GOA gene_association Table. *C. elegans* and *D. melanogaster* annotations were derived from the GOA gene_product Table. A Perl script and SQL code were created to calculate enrichment of terms and Fisher's Exact P-values against a background of all GO annotated genes in the molecular function, biological process, and cellular compartment ontologies for each species. Any GO term with less than two-fold enrichment or a *P*-value greater than 10^{-6} was discarded. We chose these criteria to ensure that after correcting conservatively for multiple-hypothesis testing (bonferroni with at most 2000 tests) our P-values would still be significant at less than 10^{-3}. This relatively relaxed cut off was chosen to allow comparison between species with high (e.g. human) and low (e.g. *C.elegans*) GO annotations. In the organism column C = *C. elegans*, D = *D. melanogaster*, H = *H. sapiens*, and M = *M. musculus*. Genes with small intronic sequence show enrichment for ribosomal-related terms in *C. elegans*, human, and mouse. Threshold criteria are two-fold enrichment and a *P* value of less than 10^{-6}. Nervous system enrichment is common across human, mouse and *Drosophila* for genes with large intronic sequence.

suspect that this is due to the restrictions of the statistical methods that we employed and the limited data sets in these organisms. The Arbeitman et al. expression data[79] covers less than one third of known *Drosophila* transcripts (4,028), and the Kim et al. expression set[81] is specific to developmental stage rather than tissue.

These results suggest a highly nonrandom distribution of intronic sequences in relation to gene function and tissue

Table 6. Expression enrichment in *Homo sapiens* of genes with large total intronic sequence size

Gene set: ≥9-fold overexpressed		Gene set: ≥2-fold underexpressed	
***P* Value**	**Tissue Type**	***P* Value**	**Tissue Type**
E-25	prefrontal cortex	E-47	leukemia chronic myelogenous k562
E-21	fetal brain	E-36	thymus
E-19	occipital lobe	E-36	BM-CD71+ early erythroid
E-15	amygdala	E-36	leukemia lymphoblastic molt4-1
E-15	parietal lobe	E-34	PB-BDCA4+ dentritic cells
E-14	thalamus	E-33	colorectal adenocarcinoma
E-13	hypothalamus	E-31	PB-CD56+ NK cells
E-11	subthalamic nucleus	E-31	PB-CD8+ T-cells
E-11	medulla oblongata	E-30	PB-CD4+ Tcells
E-09	caudate nucleus	E-28	PB-CD19+ B-cells
E-09	globus pallidus	E-26	BM-CD34+
E-09	cingulated cortex	E-24	leukemia promyelocytic hl60
E-08	whole brain	E-24	bone marrow
E-08	cerebellum peduncles	E-24	tonsil
E-07	spinal cord	E-20	fetal lung
E-06	temporal lobe	E-16	lymphoma burkitts raji
Gladstone Dataset		E-16	lymph node
E-19	whole brain	E-15	cardiac myocytes
E-12	uterus	E-13	pancreatic islets
E-11	spinal cord	*Gladstone Dataset*	
		E-09	thymus
		E-06	whole blood
		E-06	embryonic stem cell

Human GNF expression tissue enrichment with respect to total intronic sequence size. Using the genome and gene database builds outlined in the Table 5 legend, we analyzed large TIS genes for enrichment among highly and lowly expressed genes in 72 tissue types. We used expression threshold criteria to obtain groups of overexpressed and underexpressed genes within the GNF Atalas data sets. Using tissue type as a surrogate for gene ontology type, we examined for enrichment using a Fisher's Exact test against a background of all genes within the probe set (e.g. genes ≥9-fold overexpressed). Threshold criteria for inclusion were 2-fold enrichment and a *P* value of 10^{-6} for overexpression gene sets and 1.5-fold enrichment and a *P* value of 10^{-6} for underexpression gene sets, due to differences in the total number of genes/probes in each set. As with our GO term enrichment analysis, we did not directly correct for multiple-hypothesis testing but in practice we performed less than 100 tests per sample.

expression. This is in agreement with a number of recent reports that examined intron size distributions but limited their analyses to conserved regions.[82–85] In this study, we ignored conservation, instead basing our work on the hypothesis that many undiscovered *cis*-acting non-protein-coding DNA and *trans*-acting non-protein-coding RNA elements reside in introns, and that these may be adaptively plastic and therefore undetected by conservation alone, as has been shown for a number of transcriptional regulatory elements (see below). However, we also acknowledge that there are other, not mutually exclusive, explanations for these observations. For example, recent experiments in yeast have shown that introns may improve transcriptional and translational yield.[86] Additionally, genes involved in house-keeping processes are generally highly expressed and may have been evolutionarily selected to be compact by deletion of intronic sequences,[82,84,85,87–89] perhaps permitted by their lower requirement for tissue-specific regulation, which potentially artificially inflates the tissue-specific associations seen with large TIS genes in this study. Genes involved in differentiation and development, nonetheless, are known to be significantly larger than average and to contain large amounts of regulatory information in their introns, consistent with a "genome design" model.[84,39]

Regulatory sequences and non-protein-coding RNAs in complex organisms

For years, the relationship between cellular DNA content and developmental complexity has been obscured by variations in ploidy and by the assumption that the large amounts of "non-genic" information (that is, DNA sequences that do not code for proteins) in introns and intergenic regions of the genome were mostly non-functional. This may be incorrect. Genome sequence comparisons have shown that multi-cellular organisms exhibit significant conservation of non-protein-coding DNA,[42,90,91] indicating that a sizeable fraction of these sequences have genetic function. Moreover, a recent comparative study of *Drosophila* genomes indicated that a large fraction of the non-protein-coding sequences are functionally important and subject to both purifying selection and adaptive evolution.[92] In addition, it seems clear that the extent of promoter-enhancer regions, especially around genes involved in differentiation and development, have greatly expanded in complex organisms.[93–95] Many of these

sequences appear to be evolving rapidly at the primary sequence level while still maintaining similar function, indicating that important regulatory information resides not only in conserved[95,96] but also in non-conserved regions,[97–99] whose extent is as yet unknown. *Cis*-regulatory sequences in vertebrates have also been shown to undergo shuffling, increasing the number of recognizably conserved elements.[100] The length of UTRs in mRNAs has also expanded in the complex organisms, particularly in mammals,[101] suggesting an increase in *cis*-acting regulatory sequences that control translation and mRNA half-life.

It has also become evident that the genomes of complex organisms express large numbers of non-protein-coding RNAs (ncRNAs), some, if not many, of which have regulatory functions. Both cDNA and genome tiling array transcriptome analyses have shown that the majority (at least 70%) of the mammalian genome is transcribed in extremely complicated patterns of interlaced and overlapping transcripts, many of which are not polyadenylated.[101–104] Most mammalian genes also have antisense transcripts that appear to play a role in regulation of their expression.[105] The majority of the *Drosophila* genome has also been shown to be transcribed.[106] Many of these non-coding transcripts show developmental regulation and common structures,[102,103,103,107,108] and, in those cases that have been examined in more detail, specific cellular locations and functions.[109–112] Interestingly, known functional ncRNAs show different extents of primary sequence conservation, indicating that they are evolving at different rates under different structure–function constraints and selection pressures,[113,114] including positive selection during adaptive radiation,[112] indicating that lack of conservation does not necessarily imply lack of function.[113]

These observations suggest that there may be a vast hidden layer of RNA regulatory information in complex organisms and that increasing amounts of genetic information in these organisms is expressed as and transacted by RNA.[115,116] This suggestion is supported by the finding that many genetic phenomena in the higher organisms, such as imprinting, co-suppression, RNA interference and chromatin modification, involve RNA signaling (for recent reviews see Refs [116,117]). It is also supported by the unfolding discovery of new classes of small regulatory RNAs, including increasing numbers of miRNAs that control a range of developmental processes in plants and animals, via control of mRNA translation and degradation,[114,118,119] and other RNAs such as snoRNAs that modify other RNAs[120,121] and testis-specific piRNAs whose function has yet to be determined but which are evolving rapidly.[122,123]

The evidence for a central role of RNA in eukaryotic evolution and gene regulation has been presented in detail elsewhere.[62,104,116] We do not claim that all transcribed sequences are necessarily functional. Indeed, there may be a reservoir of such transcripts that are themselves simply raw material for evolution.[124] Nonetheless, the observation that the genomes of multicellular organisms contain large amounts of non-protein-coding sequences that scale consistently with developmental complexity, indicates that, in addition to important innovations in proteins involved in developmental regulation and cell signaling, most of which were in place at the base of the metazoan radiation,[125] the expansion of the *cis*- and *trans*-acting regulatory architecture has been a crucial factor in the evolution of the more developmentally complex organisms. Accordingly, we suggest that the realization that non-protein-coding sequences in complex organisms contain large amounts of regulatory information, much of which is transacted by RNA, would finally resolve the informational enigma that has confounded genetics and genomics for so long.

References

1. Cavalier-Smith T. 1985. Introduction: the evolutionary significance of genome size. In: Cavalier-Smith T, editors. The evolution of genome size. New York: John Wiley & Sons.
2. Gregory TR. 2005. Genome size evolution in animals. In: Gregory TR, editors. The evolution of the genome. New York: Elsevier.
3. Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. PLoS Comput Biol 2:e48.
4. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, et al. 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314:267.
5. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, et al. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct Integr Genomics 6:165–185.
6. Gottesman S. 2005. Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet 21:399–404.
7. Ohno S. 1972. So much ''junk'' DNA in our genome. In: Smith HH, editors. Evolution of Genetic Systems. New York: Gordon and Breach.
8. Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. Nature 284: 604–607.
9. Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603.
10. Gilbert W, Marchionni M, McKnight G. 1986. On the antiquity of introns. Cell 46:151–154.
11. Mattick JS. 1994. Introns: evolution and function. Curr Opin Genet Dev 4:823–831.
12. Bird AP. 1995. Gene number, noise reduction and biological complexity. Trends Genet 11:94–100.
13. Gregory TR. 2005. Animal genome size database. http://www.genomesize.com/.
14. Swift H. 1950. The constancy of desoxyribose nucleic acid in plant nuclei. Proc Natl Acad Sci USA 36:643–654.
15. Thomas CA Jr. 1971. The genetic organization of chromosomes. Annu Rev Genet 5:237–256.
16. Greilhuber J, Dolezel J, Lysak MA, Bennett MD. 2005. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. Ann Bot (Lond) 95:255–260.
17. Svartman M, Stone G, Stanyon R. 2005. Molecular cytogenetics discards polyploidy in mammals. Genomics 85:425–430.
18. Becak ML, Kobashi LS. 2004. Evolution by polyploidy and gene regulation in Anura. Genet Mol Res 3:195–212.
19. Licht LE, Lowcock LA. 1991. Genome size and metabolic-rate in salamanders. Comp Biochem Physiol B: Biochem Mol Biol 100:83–92.
20. Smith JJ, Kump DK, Walker JA, Parichy DM, Voss SR. 2005. A comprehensive expressed sequence tag linkage map for tiger

salamander and Mexican axolotl: enabling gene mapping and comparative genomics in Ambystoma. Genetics 171:1161–1171.
21. Gregory TR, Mable BK. 2005. Polyploidy in animals. In: Gregory TR, editors. The evolution of the genome. New York: Elsevier.
22. Vervoort A. 1980. Tetraploidy in Protopterus (Dipnoi). Cellular and Molecular Life Sciences (CMLS) 36:294–296.
23. Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu Rev Genet 34:401–437.
24. Vinogradov AE. 2005. Genome size and chromatin condensation in vertebrates. Chromosoma 113:362–369.
25. Gregory TR. 2005. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann Bot (Lond) 95:133–146.
26. Bennett MD, Leitch IJ. 2005. Nuclear DNA amounts in angiosperms: progress, problems and prospects. Ann Bot (Lond) 95:45–90.
27. Gregory TR, Hebert PD. 1999. The modulation of DNA content: proximate causes and ultimate consequences. Genome Res 9:317–324.
28. Knight J. 2002. All genomes great and small. Nature 417:374–376.
29. Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot (Lond) 95:147–175.
30. Genoscope 2006. Paramecium Genome Browser. http://www.genoscope.cns.fr/externe/Francais/Projets/Projet_FN/.
31. Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. Biol J Linn Soc 82:651–663.
32. Li M, Vitanyi PMB. 1997. An introduction to Kolmogorov complexity and its applications. New York: Springer-Verlag.
33. Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169–187.
34. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341–345.
35. Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res 16:934–946.
36. Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314.
37. Brosius J. 1991. Retroposons--seeds of evolution. Science 251:753.
38. Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. Trends Genet 17:23–28.
39. Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63.
40. Brosius J. 2005. Disparity, adaptation, exaptation, bookkeeping, and contingency at the genome level. Paleobiology 31:1–16.
41. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
42. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
43. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, et al. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7:597–606.
44. Nishihara H, Smit AF, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res 16:864–874.
45. Xie X, Kamal M, Lander ES. 2006. A family of conserved noncoding elements derived from an ancient transposable element. Proc Natl Acad Sci USA 103:11659–11664.
46. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90.
47. Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc 80:227–250.
48. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100:11484–11489.
49. Hahn MW, Wray GA. 2002. The g-value paradox. Evol Dev 4:73–75.
50. Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, et al. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nature Genet 25:235–238.
51. Goodstadt L, Ponting CP. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol 2.
52. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.
53. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297:1301–1310.
54. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol 1:E45.
55. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol 3: RESEARCH0083.
56. Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK Jr, et al. 2005. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biol 3:7.
57. Project IRGS. 2005. The map-based sequence of the rice genome. Nature 436:793–800.
58. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. 2006. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol 4:e286.
59. Jabbari K, Cruveiller S, Clay O, Le Saux J, Bernardi G. 2004. The new genes of rice: a closer look. Trends Plant Sci 9:281–285.
60. Southan C. 2004. Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics 4:1712–1726.
61. Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O. 2005. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene 364:53–62.
62. Mattick JS. 2004. RNA regulation: a new genetics? Nat Rev Genet 5: 316–323.
63. Mattick JS, Gagen MJ. 2005. Accelerating networks. Science 307:856–858.
64. Taft RJ, Mattick JS. 2003. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biol Preprint Depository http://genomebiology.com/2003/5/1/ P1.
65. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, et al. 2005. The genome of the social amoeba Dictyostelium discoideum. Nature 435:43–57.
66. Mongin E, Louis C, Holt RA, Birney E, Collins FH. 2004. The Anopheles gambiae genome: an update. Trends Parasitol 20:49–52.
67. Kriventseva EV, Koutsos AC, Blass C, Kafatos FC, Christophides GK, et al. 2005. AnoEST: toward A. gambiae functional genomics. Genome Res 15:893–899.
68. Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, et al. 2004. The genome sequence of silkworm, Bombyx mori. DNA Res 11:27–35.
69. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.
70. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, et al. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306:79–86.
71. Brent MR. 2005. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res 15:1777–1786.
72. Zagulski M, Nowak JK, Le Mouel A, Nowacki M, Migdalski A, et al. 2004. High coding density on the largest Paramecium tetraurelia somatic chromosome. Curr Biol 14:1397–1404.
73. de Souza SJ. 2003. The emergence of a synthetic theory of intron evolution. Genetica 118:117–121.
74. Fedorova L, Fedorov A. 2003. Introns in gene evolution. Genetica 118: 123–131.
75. Roy SW. 2003. Recent evidence for the exon theory of genes. Genetica 118:251–266.
76. Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol 23:450–468.

77. Consortium GO. 2006. The Gene Ontology (GO) project in 2006. Nucleic Acids Res 34:D322–326.
78. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067.
79. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, et al. 2002. Gene expression during the life cycle of Drosophila melanogaster. Science 297:2270–2275.
80. Kufe DW, Pollock RE, Weichselbaum RR, Bast RC Jr, Gansler TS, et al. 2003. Cancer Medicine. Hamilton (Canada): BC Decker Inc.
81. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, et al. 2001. A gene expression map for Caenorhabditis elegans. Science 293:2087–2092.
82. Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N, et al. 2005. Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. Hum Mol Genet 14:2533–2546.
83. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050.
84. Vinogradov AE. 2006. ''Genome design'' model: Evidence from conserved intronic sequence in human-mouse comparison. Genome Res 16:347–354.
85. Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, et al. 2006. Intron size in mammals: complexity comes to terms with economy. Trends Genet, epub in advance of publication:doi10. 1016.
86. Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW. 2006. Introns regulate RNA and protein abundance in yeast. Genetics 174: 511–518.
87. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet 21:203–207.
88. Coulombe-Huntington J, Majewski J. 2006. Characterization of intron loss events in mammals. Genome Res, epub in advance of publication:gr.5703406:.
89. Vinogradov AE. 2006. 'Genome design' model and multicellular complexity: golden middle. Nucleic Acids Res, epub in advance of publication:doi10.1093/nar/gkl773.
90. Inada DC, Bashir A, Lee C, Thomas BC, Ko C, et al. 2003. Conserved noncoding sequences in the grasses. Genome Res 13:2030–2041.
91. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, et al. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302:1033-1035.
92. Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–1152.
93. Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76:8–32.
94. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, et al. 2005. Evolution and functional classification of vertebrate gene deserts. Genome Res 15:137–145.
95. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502.
96. Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. Science 314: 786.
97. Frith MC, Ponjavic J, Fredman D, Kai C, Kawai J, et al. 2006. Evolutionary turnover of mammalian transcription start sites. Genome Res 16:713–722.
98. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. 2006. Heterotachy in mammalian promoter evolution. PLoS Genet 2:e30.
99. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279.
100. Sanges R, Kalmar E, Claudiani P, D'Amato M, Muller F, et al. 2006. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. Genome Biol 7:R56.
101. Frith MC, Pheasant M, Mattick JS. 2005. The amazing complexity of the human transcriptome. Eur J Hum Genet 13:894–897.
102. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154.
103. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. 2005. The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
104. Mattick JS, Makunin IV. 2006. Non-coding RNA. Hum Mol Genet 15: R17–R29.
105. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, et al. 2005. Antisense transcription in the mammalian transcriptome. Science 309:1564–1566.
106. Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, et al. 2006. Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nat Genet 38:1151–1158.
107. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, et al. 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16:11–19.
108. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res 16:885–889.
109. Prasanth KV, Prasanth SG, Xuan Z, Hearn S, Freier SM, et al. 2005. Regulating gene expression through RNA nuclear retention. Cell 123: 249–263.
110. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, et al. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 309:1570–1573.
111. Ginger MR, Shore AN, Contreras A, Rijnkels M, Miller J, et al. 2006. A noncoding RNA is a potential marker of cell fate during mammary gland development. Proc Natl Acad Sci USA 103:5781–5786.
112. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172.
113. Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22:1–5.
114. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, et al. 2006. Diversity of microRNAs in human and chimpanzee brain. Nat Genet 38:1375–1377.
115. 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.
116. Mattick JS. 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays 25:930–939.
117. Bernstein E, Allis CD. 2005. RNA meets chromatin. Genes Dev 19:1635–1655.
118. Bartel DP, Chen CZ. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet 5:396–400.
119. Mattick JS, Makunin IV. 2005. Small regulatory RNAs in mammals. Hum Mol Genet 14:R121–132.
120. Bachellerie JP, Cavaille J, Huttenhofer A. 2002. The expanding snoRNA world. Biochimie 84:775–790.
121. Kishore S, Stamm S. 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science 311:230–232.
122. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442:199–202.
123. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442:203–207.
124. Brosius J. 2005. Waste not, want not--transcript excess in multicellular eukaryotes. Trends Genet 21:287–288.
125. Technau U, Rudd S, Maxwell P, Gordon PM, Saina M, et al. 2005. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet 21:633–639.
126. Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–504.