Title

Computational Methods of Identification of Pseudogenes Based on Functionality: Entropy and GC Content

Authors

Balakirev, Evgeniy S
Chechetkin, Vladimir R
Lobzin, Vasily V
et al.

Publication Date

2014

DOI

10.1007/978-1-4939-0835-6_4

Peer reviewed

Chapter 4

Computational Methods of Identification of Pseudogenes Based on Functionality: Entropy and GC Content

Evgeniy S. Balakirev, Vladimir R. Chechetkin, Vasily V. Lobzin, and Francisco J. Ayala

Abstract

Spectral entropy and GC content analyses reveal comprehensive structural features of DNA sequences. To illustrate the significance of these features, we analyze the *β-esterase* gene cluster, including the *Est-6* gene and the ψ*Est-6* putative pseudogene, in seven species of the *Drosophila melanogaster* subgroup. The spectral entropies show distinctly lower structural ordering for ψ*Est-6* than for *Est-6* in all species studied. However, entropy accumulation is not a completely random process for either gene and it shows to be nucleotide dependent. Furthermore, GC content in synonymous positions is uniformly higher in *Est-6* than in ψ*Est-6*, in agreement with the reduced GC content generally observed in pseudogenes and nonfunctional sequences. The observed differences in entropy and GC content reflect an evolutionary shift associated with the process of pseudogenization and subsequent functional divergence of ψ*Est-6* and *Est-6* after the duplication event. The data obtained show the relevance and significance of entropy and GC content analyses for pseudogene identification and for the comparative study of gene–pseudogene evolution.

Key words Pseudogenes, Entropy, GC content, *Drosophila melanogaster* subgroup, β-*esterase* gene cluster, *Est-6*, ψ*Est-6*

1 Introduction

Identification of pseudogenes and distinguishing them from functional genes is an important but difficult problem. To resolve the issue, a number of computational approaches are available, all rooted in an initial homology-based search to identify putative pseudogene sequences [1–18]. These approaches have greatly enhanced our capability to identify pseudogenes. However, they have multiple limitations: they display inconsistency in a number of detected pseudogenes [19] and they underestimate the total pseudogene number [20]. Furthermore, they fail in differentiating paralogous genes from duplicated pseudogenes, as well as in identifying "cryptic" pseudogenes (pseudogenes that do not show any clear disruptive features, such as internal stop codons or not-in-frame indels [21]),

Laura Poliseno (ed.), *Pseudogenes: Functions and Protocols*, Methods in Molecular Biology, vol. 1167,
DOI 10.1007/978-1-4939-0835-6_4,

because they limit their analysis to sequence-level homology [22, 23]. There is also increasing evidence showing pseudogenes that are functional in spite of carrying disruptive mutations (*reviewed in* refs. 24–28), which further complicates the problem of distinguishing between genes and pseudogenes.

Here, we describe entropy and GC-content analyses and apply them to characterize the functional gene *Est-6* and its putative pseudogene ψ*Est-6* in seven species of the *Drosophila melanogaster* subgroup. We show that both features (entropy and GC content) are useful for pseudogene identification and for the comparative study of gene–pseudogene evolution.

1.1 Entropy

Tandem repeats and scattered DNA repeats play important roles in the structural organization of chromatin and in the regulatory mechanisms [29]. Repeats modified by quasi-random mutations and insertions/deletions also play important roles in protein-coding fragments attributed to a fraction of "unique DNA" in genomic sequences (*for a review see*, e.g., refs. 30, 31). Such periodicities emerge because the coding concordant with B-DNA double helix pitch, with quasi-repeated package of nucleosomes, with cooperative binding with regulatory proteins, etc., exhibit evolutionary preference. Another highly reproducible feature is related to periodicities of period $p=3$ in the protein-coding regions ([30–33], *and references therein*). The typical scenario may be as follows: (1) take a stretch of any tandem repeats, e.g., TTA|TTA|TTA …; (2) replace randomly some nucleotides in this stretch (generally, random insertions/deletions should also be taken into account). The resulting stretch can be defined as a sequence with hidden periodicities. The frequency of random replacements may be biased in the different sites of a repeat. For instance, if the random replacements occur only in the first and second positions, this would yield the patterns NNA|NNA|NNA…, where N is any nucleotide. The period $p=3$ has proved to be typical of both genes and pseudogenes [34, 35].

Spectral methods for analysis of DNA sequences are commonly used to reveal hidden periodicities, to study the correlations between different sequences, and to investigate long-range correlations (*for a general review of methods and further references*, *see* ref. 30). The use of the Fourier transform allows one to obtain statistical criteria that possess a self-averaging property beginning from relatively short sequences (≥100–200 bp). Within the framework of such a technique we can study both the separate elements of mosaic structure and the coupling between the elements. The strategy is to compare the observed characteristics of real sequences with those of random sequences with the same nucleotide composition. This comparison is important for revealing segmentation effects, as well as the effects of compositional variations arising in evolution and as a consequence of the differences in the environmental conditions

for different organisms. In this way one can obtain convenient and universal criteria for structural regularity. Such criteria do not depend on nucleotide composition, thereby allowing to compare the structural characteristics of DNA sequences with different nucleotide compositions. Thus, spectral analysis allows, in principle, to reveal all the structural features of the analyzed DNA sequence.

1.2 GC Content

GC content is (G+C)/(A+T+G+C), where G is the number of guanines, C is the number of cytosine, A is the number of adenines, and T is the number of thymines. GC content is a fundamental property of DNA sequences and there is a considerable amount of research related to mean genomic GC content demonstrating that this property is the result of many factors interacting in a highly complex manner [36]. It has been suggested to consider GC content as a relatively stable characteristic (or parameter) of human genomes, like fundamental mathematical, physical, or chemical constants [37]. GC content explains the intrinsic nucleosome-forming preferences and may have significant influence on chromatin structure in eukaryotic genomes [38]. In prokaryotes, the GC content remains fairly constant within species, but varies widely across microbial species, which may be viewed as a response to environmental adaptation [39–41]. The considerable variation in GC content cannot be explained by neutral processes, implying a role for natural selection [42, 43]. Increase of GC content creates CpG islands associated with gene regulation [44]. The gain and loss of CpG may be a fundamental characteristic of the human-specific genotype. The CpG dinucleotide is vital for regulation and it not only conveys genomic data but also enables epigenomic variation [45]. Thus, the comparative analysis of the GC-content may reveal much information about features of gene and pseudogene evolution.

1.3 The β-esterase Gene Cluster in Drosophila

The β-*esterase* gene cluster is located on the left arm of chromosome 3 of *D. melanogaster*, at 68F7-69A1 in the cytogenetic map. The cluster comprises two tandemly duplicated genes, first described as *Est-6* and *Est-P* [46], with coding regions separated by only 193 bp. The coding regions are 1,686 and 1,691 bp long, respectively, and consist of two exons (1,387 and 248 bp) and a small (51 bp in *Est-6* and 56 bp in *Est-P*) intron [47].

The *Est-6* gene is well characterized (*reviews in* refs. 48–50). The gene encodes the major β-carboxylesterase (EST-6) that is transferred by *D. melanogaster* males to females in the seminal fluid during copulation [51] and affects the female's consequent behavior and mating proclivity [52]. Less information is available for *Est-P*. Based on several lines of evidence (transcriptional activity, intact splicing sites, no premature termination codons, presence of initiation and termination codons) it was concluded that *Est-P* is a functional gene [46]. Some alleles of the *Est-P* produce a catalytically active esterase [53] corresponding to the previously identified EST-7 isozyme [54].

However, molecular population analysis detected premature stop codons within the *Est-P* coding region and some other indications suggesting that *Est-P* might be in fact a pseudogene, which was labeled ψ*Est-6* [21, 35, 55, 56]. The β-*esterase* gene cluster in other *Drosophila* species also includes two (or three, in *D. pseudoobscura* and related species) closely linked genes [49, 50, 57–60].

Previously, we have investigated nucleotide variability of the *Est-6* gene and ψ*Est-6* in a *D. melanogaster* sample from a natural population of California [21, 35, 61–63], and also in three populations of East Africa (Zimbabwe), Europe (Spain), and South America (Venezuela) [55, 56, 64]. We have detected different patterns of nucleotide variation in *Est-6* and ψ*Est-6* indicating different evolutionary trends in these two genes and suggesting that ψ*Est-6* could represent a putative pseudogene.

Here we describe the entropy and GC content analyses to characterize the functional gene *Est-6* and putative pseudogene ψ*Est-6* in seven species of the *D. melanogaster* subgroup. We show that both approaches are useful for pseudogene identification and for comparative studies of gene–pseudogene evolution.

2 Materials and Methods

2.1 Drosophila Strains and Species

The *D. melanogaster* strains have been previously described [55, 62, 64]. Strains of *Drosophila simulans*, *Drosophila sechellia*, *Drosophila mauritiana*, *Drosophila erecta*, *Drosophila teissieri*, and *Drosophila orena* were obtained from the Drosophila Species Stock Center (Bowling Green, Ohio).

2.2 DNA Extraction, Amplification, and Sequencing

The procedures for *Est-6* and ψ*Est-6* amplification, cloning, and sequencing were described earlier [35, 61, 62]. For each line, the sequences of both strands were determined, using 24 overlapping internal primers spaced, on average, by 350 nucleotides. At least two independent PCR amplifications were sequenced in both directions to prevent possible PCR or sequencing errors. The data for seven species of *D. melanogaster* subgroup are from Balakirev et al. [65]; see also GenBank accessions AY695919 (*D. simulans*), AY695920 (*D. sechellia*), AY695921 (*D. mauritiana*), AY695922 (*D. teissieri*), AY695923 (*D. erecta*), and AY695924 (*D. orena*). The population data for *D. melanogaster* are from Balakirev and Ayala [55, 56, 64]; see GenBank accessions AF147095–147102; AF150809–AF150815; AF217624–AF217645; AF526538–AF526559; AY247664–AY247713; AY247987–AY248036; AY368077–AY368109; AY369088–AY369115.

2.3 DNA Sequence Analysis

The esterase sequences were assembled using the program SeqMan (Lasergene, DNASTAR, Inc., 1994–1997). Multiple alignment was carried out manually and using the program CLUSTAL W [66].

Model-based phylogeny reconstructions were performed with the MEGA, version 5 [67] using the neighbor-joining (NJ) and maximum-likelihood (ML) methods. GC content was computed using the DnaSP program, version 5.10.01 [68] and PROSEQ, version 2.9 [69]. The Wilcoxon and Mann–Whitney tests were used to evaluate the significance of pairwise differences in GC content.

2.4 Spectral Structural Entropy

Spectral entropy characterizes the structural regularity of a nucleotide sequence. In our case it allows one to assess the comparative rates and positional distribution of mutations, as well as their influence on the regularity of the nucleotide sequences for *Est-6* and ψ*Est-6*. Therefore, it allows one to shed additional light on gene function. For the convenience of the reader, below we reproduce the main definitions (*for details and further references see* refs. 30, 70–72).

Fourier harmonics corresponding to nucleotides of type α (α is *A*, *C*, *G*, or *T*) in a sequence of length *M* are calculated as

$$\rho_\alpha(q_n) = M^{-1/2} \sum_{m=1}^{M} \rho_{m,\alpha} e^{-iq_n m}, \quad q_n = 2\pi n/M, \quad n = 0,1,\ldots,M-1. \tag{1}$$

Here $\rho_{m,\alpha}$ indicates the position occupied by the nucleotide of type α; $\rho_{m,\alpha} = 1$ if the nucleotide of type α occupies the *m*-th site, and 0 otherwise. The amplitudes of Fourier harmonics (or structure factors) are expressed as

$$F_{\alpha\alpha}(q_n) = \rho_\alpha(q_n)\rho_\alpha^*(q_n), \tag{2}$$

where the asterisk denotes complex conjugation. The zeroth harmonics, depending only on the nucleotide composition, do not contain structural information and will be discarded below. The structure factors will always be normalized with respect to the mean spectral values, which are determined by the exact sum rules,

$$f_{\alpha\alpha}(q_n) = F_{\alpha\alpha}(q_n) / \bar{F}_{\alpha\alpha}; \quad \bar{F}_{\alpha\alpha} = N_\alpha(M - N_\alpha)/M(M-1), \tag{3}$$

where N_α is the total number of nucleotides of type α in a sequence of length *M*. Correct judgment on the significance of hidden periodicities revealed by Fourier analysis needs application of proper statistical criteria. Random sequences of the same nucleotide composition serve as a reference for assessing the observed regularities in a given DNA sequence.

The spectral entropy provides the quantitative measure of order/disorder in DNA sequence and is defined by the sum

$$S_\alpha = -\sum_{n=1}^{M-1} f_{\alpha\alpha}(q_n) \ln f_{\alpha\alpha}(q_n) \tag{4}$$

over the spectrum. Its values are strictly negative. The value of spectral entropy for a counterpart random sequence having the same nucleotide composition is the highest and the corresponding mean characteristics averaged over the ensemble of various random realizations are given by [73, 74]

$$< S_\alpha >_{\text{random}} \cong -0.422785\ldots(M-1);\ < (\Delta S_\alpha)^2 >_{\text{random}} \cong 0.579736\ldots(M-1). \quad (5)$$

The order in a DNA sequence can be related to hidden periodic patterns. The lower (or more negative) values of spectral entropy indicate the higher ordering of DNA sequence or the more pronounced periodic patterns in comparison with random sequences of the same nucleotide composition. Otherwise, the higher (or closer to zero) values of spectral entropy indicate the higher frequency of point random mutations or the stronger variability in the corresponding stretches. Therefore, spectral entropy provides a useful tool for the quantitative comparison of structural ordering in stretches with different genetic functions [35, 65, 73, 75]. In what follows, the values of spectral structural entropy will always be normalized per harmonic, i.e., $S \equiv S/(M-1)$. We will use the sums $S_A + S_T$ and $S_G + S_C$ remaining invariant on the direct and complementary strands. The total entropy is determined by the sum over entropies for particular nucleotides,

$$S_{\text{total}} = S_A + S_T + S_G + S_C. \quad (6)$$

3 Results

We present here examples of the entropy characteristics and GC content analysis of the β-*esterase* gene cluster in seven species belonging to the *D. melanogaster* subgroup. These seven species belong to three complexes: (1) the *melanogaster* complex, composed by *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia*; (2) the *yakuba* complex, composed by *D. teissieri* (*D. yakuba* and *D. santomea* are the two other species in this complex); and (3) a complex represented by *D. erecta* and *D. orena* [76–79]. The phylogenetic relationships of the *Est-6* and ψ*Est-6* in the seven species are presented in Fig. 1. Alternative tree reconstruction methods implemented in the MEGA version 5 [67] yield identical topologies (data not shown). The tree in Fig. 1 is consistent with those derived from other genes as well as using entire genomes [80–84].

3.1 Entropy Analysis

Here we present the results of a comparative analysis of DNA sequences related to the functional gene *Est-6* and the putative pseudogene ψ*Est-6* in four populations of *Drosophila melanogaster* (Zimbabwe, Spain, California, and Venezuela). The total number of sequence pairs is 79. A similar analysis was also performed for six

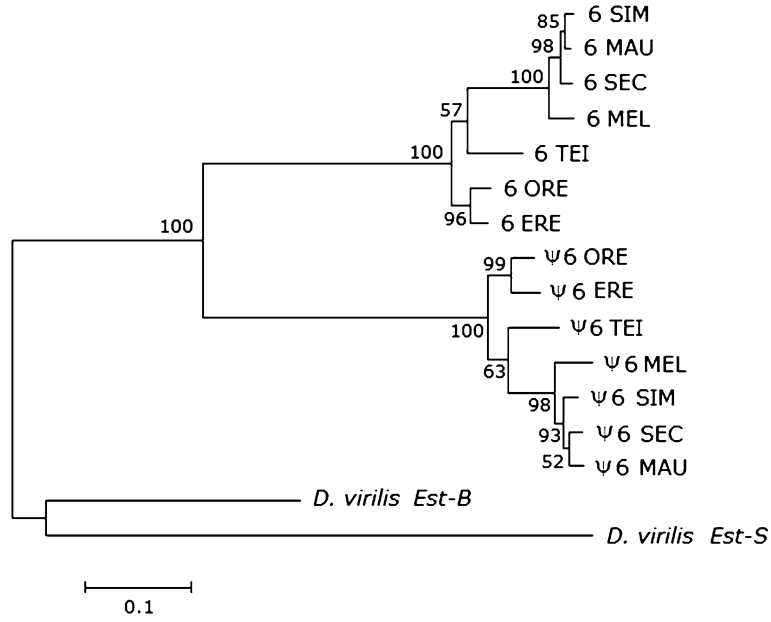

Fig. 1 Maximum likelihood tree of the β-*esterase* genes based on Kimura 2-parameter + gamma model of substitution. The tree is constructed using the coding (exon I + exon II) sequence for each gene. Numbers at the nodes are bootstrap percent support values based on 1,000 replications in Maximum likelihood analysis. The sequences of the *esterase-B* (AB679281) and *esterase-S* (XM002056666) of *D. virilis* are used as an outgroup. MEL = *D. melanogaster*, SIM = *D. simulans*, SEC = *D. sechellia*, MAU = *D. mauritiana*, TEI = *D. teissieri*, ORE = *D. orena*, and ERE = *D. erecta*. See the text for GenBank accession numbers and species abbreviations. *6* = *Est-6*; ψ*6* = ψ*Est-6*

other species of the *Drosophila melanogaster* subgroup: *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. erecta*, *D. teissieri*, and *D. orena* (one pair of sequences per each species). All sequences were aligned, introns and indels were removed, so that the length of the remaining sequences was $M = 1{,}608$ nt.

The comparative characteristics for *Est-6* and ψ*Est-6* in populations of *D. melanogaster* are presented in Fig. 2. The divergence between the two sets of data for gene and pseudogene proved to be statistically significant by Fisher criterion based on the ratio of intra- and intergroup scattering [85], $\Pr < 10^{-3}$. The highest divergence was observed for the sum of entropies $S_G + S_C$ and for GC content (Figs. 2b, d). The difference in the sum of entropies $S_G + S_C$ corresponding to *Est-6* and ψ*Est-6* correlates with the difference in GC content (Fig. 3). The corresponding Pearson correlation coefficient between these differences is $k = -0.324$ ($\Pr = 0.004$). The counterpart correlations between differences in $S_A + S_T$ and that in AT content appear to be insignificant.

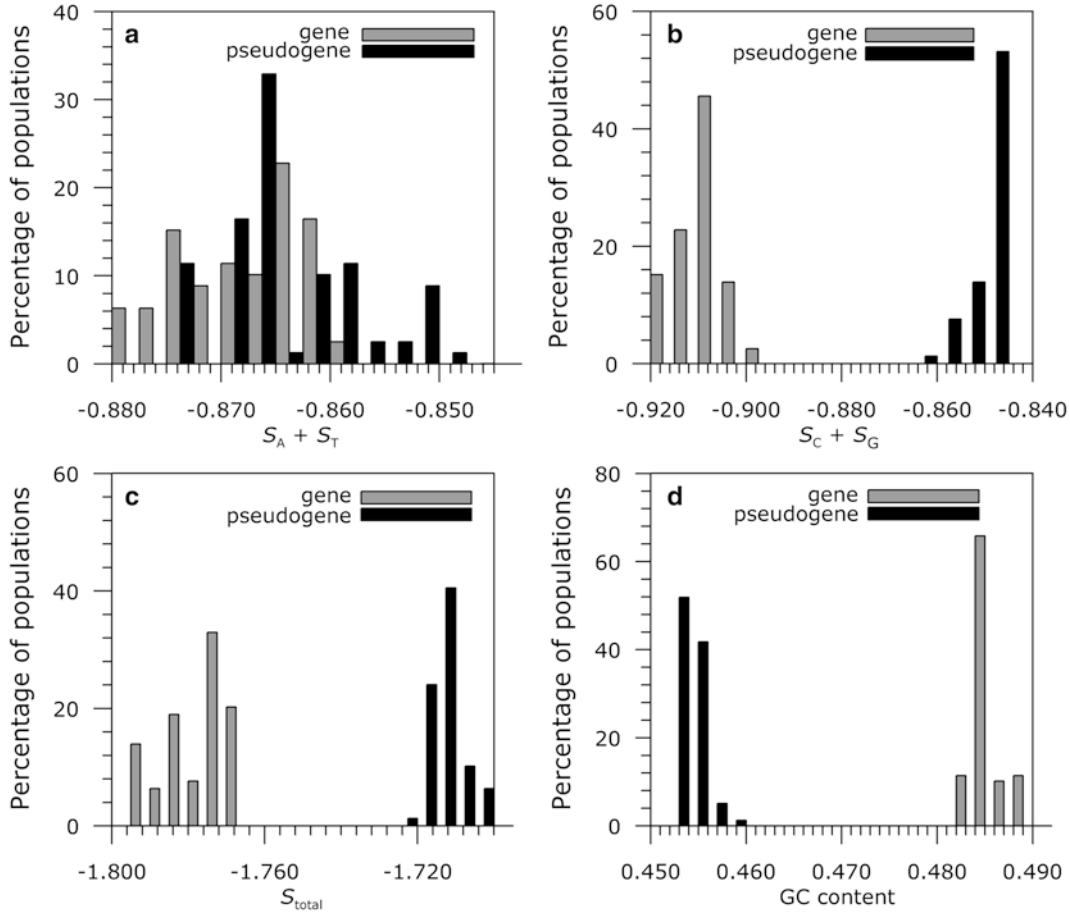

Fig. 2 Distribution of spectral entropies per harmonic and GC content for *Est-6* and ψ*Est-6* over four geographic populations of *D. melanogaster* (California, Zimbabwe, Spain and Venezuela). The spectral entropies are defined by Eqs. 4 and 6 and should be compared with random counterparts defined by Eq. 5

The higher (closer to zero) entropy for ψ*Est-6* relative to *Est-6* can be attributed to the higher accumulation of mutations in the pseudogene relative to the functional gene. In order to assess the relative mutation accumulation, we performed a simulation by introducing random point mutations into a sequence for *Est-6*. Two scenarios were considered: in the first one, multiple random mutations in a given site were permitted, whereas in the second scenario, multiple mutations in the same site were excluded (scenarios I and II, respectively). In both scenarios the mutations in *Est-6* were introduced up to equalizing total entropies for *Est-6* and ψ*Est-6*. The number of realizations in simulations for the each pair of sequences was 300. The corresponding results are shown in Fig. 4. The mean values of mutations averaged over populations were 313 and 256, with the mean scattering 231 and 149 in scenario I and II, respectively. The mean value of scattering, N_s, averaged over 300 realizations can be approximately related to the corresponding mean value of mutations, N_m, as $N_s \approx 0.5N_m + 4.4N_m^{1/2}$ (scenario I, Fig. 4a)

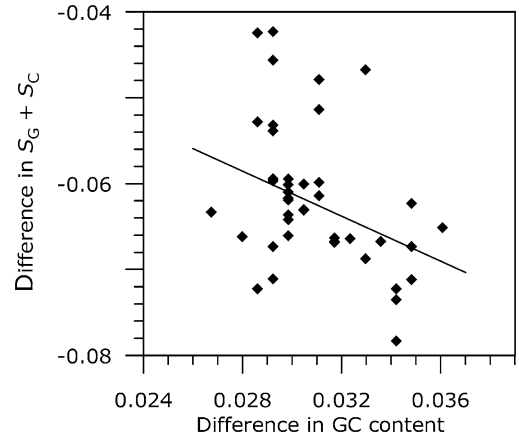

Fig. 3 Scatter diagram for the difference in the sum of entropies $S_G + S_C$ corresponding to the gene *Est-6* and the pseudogene ψ*Est*-6, in relation to GC content. The line corresponds to the best linear fit and denotes a decreasing trend

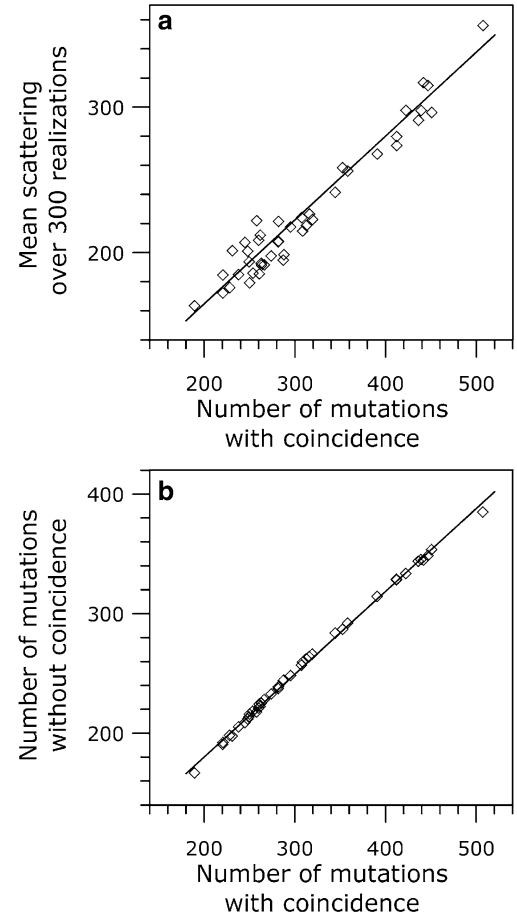

Fig. 4 (**a**) The relationship between mean number of mutations and mean scattering averaged over 300 realizations in scenario I with multiple mutations permitted at the same site. (**b**) The relationship between mean numbers of mutations averaged over 300 realizations in scenarios I and II with permitted and excluded multiple mutations at the same site. In both scenarios, the mutations in *Est-6* were introduced up to equalizing total entropies for *Est-6* and ψ*Est-6*. The lines correspond to the best linear fit and denote the trend

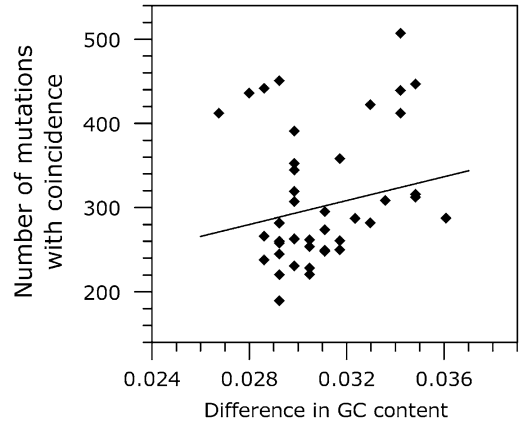

Fig. 5 Scatter diagram for the mean number of mutations averaged over 300 realizations in scenario I with permitted multiple mutations at the same site in reference to the difference in GC content between *Est-6* and ψ*Est-6*. The line corresponds to the best linear fit and denotes the trend

and $N_s \oplus 9.3N_m^{1/2}$ (scenario II, data not shown). The approximation of scattering was chosen by analogy with Poisson distribution. The mean values of mutations in the two scenarios are approximately related as $N_{mII} \approx 0.8 N_{mI}$ (Fig. 4b). The mean number of mutations in both scenarios proved to be correlated with the difference in GC content related to *Est-6* and ψ*Est-6* (Fig. 5). The corresponding Pearson correlation coefficient is $k = 0.230$ ($\Pr = 0.04$).

In Table 1 the values of the relative spectral structural entropy

$$S_{\alpha,\mathrm{rel}} = \frac{\left(< S_\alpha >_{\mathrm{random}} - S_\alpha\right)}{\left|< S_\alpha >_{\mathrm{random}}\right|}, \tag{7}$$

as well as the relative normalized deviations

$$r_\alpha = \frac{\left(< S_\alpha >_{\mathrm{random}} - S_\alpha\right)}{< \left(\Delta S_\alpha\right)^2 >_{\mathrm{random}}^{1/2}} \tag{8}$$

obeying approximately Gaussian statistics for random sequences are presented for seven species of the *D. melanogaster* subgroup. These data also prove the higher structural ordering in *Est-6* relative to ψ*Est*-6, indicating a higher incidence of mutations and a stronger variability in the sequences corresponding to ψ*Est-6*. Further details for similar analyses in seven species of the *D. melanogaster* subgroup may be found in Balakirev et al. [65].

3.2 GC Content Analysis

Table 2 and Fig. 6 show the distribution of GC content (GCc). The average GC:AT ratios for all and coding positions are not significantly different between *Est-6* and ψ*Est-6* ($P > 0.05$, Fisher's exact test), but the ratio is significantly different for the third codon

Table 1
Relative normalized deviations for structural entropy, $r_{\alpha,\,rel}$ and r_{rel}

	Nucleotides				
	A	**G**	**T**	**C**	**Total**
Est-6					
MEL	−0.21	2.66	0.36	1.37	2.09
SIM	−0.48	2.70	−0.12	0.83	1.46
SEC	−0.11	3.56	0.56	0.21	2.11
MAU	−0.23	2.42	−0.18	1.14	1.58
TEI	1.16	3.24	1.00	1.04	3.22
ORE	1.86	2.82	1.66	0.75	3.55
ERE	2.57	3.28	1.67	1.44	4.48
Average	0.65 ± 0.46	2.95 ± 0.16	0.71 ± 0.29	0.97 ± 0.16	2.64 ± 0.43
ψ*Est-6*					
MEL	−1.10	0.33	−0.13	−1.40	1.15
SIM	−0.08	0.94	0.43	−1.35	0.02
SEC	−0.38	0.78	0.44	−1.69	0.43
MAU	−0.00	−0.71	−0.11	−1.59	1.21
TEI	1.53	0.41	1.38	−2.03	0.64
ORE	1.23	1.54	−0.45	−0.07	1.13
ERE	1.21	2.27	0.16	−0.60	1.53
Average	0.34 ± 0.37	0.79 ± 0.36	0.25 ± 0.22	−1.25 ± 0.26	0.87 ± 0.20

The threshold normalized deviation $r_0 = 1.64$ corresponds to probability Pr = 0.05 for the random counterparts. Species abbreviations as in Fig. 1

position ($P = 0.0143$, Fisher's exact test). Total GCc is significantly lower in ψ*Est-6* than in *Est-6* (46.6 % vs. 49.7 %; Wilcoxon test $P = 0.0156$; Mann–Whitney test $P = 0.006$), mostly due to GC3, the third codon position (46.0 % vs. 55.1 %; Wilcoxon $P = 0.0156$; Mann–Whitney $P = 0.006$). For coding positions the difference in GCc between ψ*Est-6* and *Est-6* is not significant (47.0 % vs. 46.9 %; Wilcoxon $P = 0.6875$; Mann–Whitney $P = 0.9015$). Thus, the most pronounced difference in base composition between *Est-6* and ψ*Est-6* is GC content at the third codon position (GC3).

The dispersion of the GC values between exon I and exon II is lower for ψ*Est-6* than for *Est-6* (Table 2 and Fig. 6). GC content varies narrowly in the ψ*Est-6* exons (40.2–48.1 %) but more broadly in *Est-6* (36.1–58.5 %). Exon I has significantly higher GC content than exon II for both genes (47.7 and 51.9 % vs. 40.6 and

Table 2
GC content (%) and overall GC to AT ratio of the *Est-6* and ψ*Est-6* genes in seven species of the *D. melanogaster* subgroup

Gene and species	Exon I GC	Exon II GC	Total GC	Total GC:AT	Exon I GC3	Exon II GC3	Total GC3	Total GC3:AT3	Exon I GCc	Exon II GCc	Total GCc	Total GCc:ATc
6 MEL	50.4	37.8	48.5	0.942	55.2	35.4	52.1	1.088	48.0	39.0	46.6	0.873
6 SIM	51.9	37.8	49.8	0.992	59.2	36.6	55.7	1.257	48.3	38.4	46.8	0.880
6 SEC	51.9	37.0	49.7	0.988	58.5	32.9	54.6	1.203	48.7	39.0	47.2	0.894
6 MAU	51.9	37.0	49.7	0.988	59.6	34.1	55.7	1.257	48.1	38.4	46.6	0.873
6 TEI	52.6	40.2	50.7	1.028	61.1	41.5	58.1	1.387	48.3	39.6	47.0	0.887
6 ORE	52.0	34.1	49.3	0.972	57.2	34.1	53.6	1.155	49.4	34.1	47.1	0.890
6 ERE	52.4	37.0	50.0	1.000	58.9	37.8	55.7	1.257	49.1	36.6	47.2	0.894
Average	*51.9*	*37.3*	*49.7*	*0.988*	*58.5*	*36.1*	*55.1*	*1.227*	*48.6*	*37.9*	*46.9*	*0.883*
ψ*6* MEL	46.7	39.8	45.6	0.838	45.5	36.6	44.1	0.789	47.2	41.5	46.4	0.866
ψ*6* SIM	47.8	40.2	46.6	0.873	46.4	36.6	44.9	0.815	48.5	42.1	47.5	0.905
ψ*6* SEC	47.9	40.2	46.7	0.876	47.0	39.0	45.8	0.845	48.3	40.9	47.2	0.894
ψ*6* MAU	47.6	40.7	46.5	0.869	46.8	40.2	45.8	0.845	48.0	40.9	46.9	0.883
ψ*6* TEI	48.7	41.9	47.7	0.912	49.0	43.9	48.2	0.931	48.6	40.9	47.4	0.901
ψ*6* ORE	47.8	41.1	46.7	0.876	46.8	43.9	46.4	0.866	48.2	39.6	46.9	0.883
ψ*6* ERE	47.7	40.2	46.5	0.869	47.5	41.5	46.5	0.869	47.8	39.6	46.5	0.869
Average	*47.7*	*40.6*	*46.6*	*0.873*	*47.0*	*40.2*	*46.0*	*0.852*	*48.1*	*40.8*	*47.0*	*0.887*

GC3 represents GC content at the third codon position. GCc represents GC content at the coding positions. For other comments *see* Fig. 1

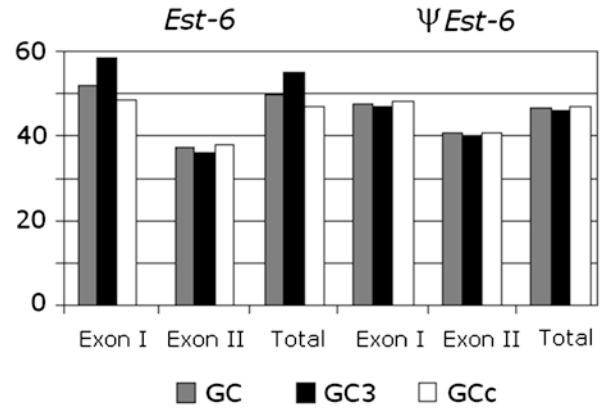

Fig. 6 Percent of GC content in *Est-6* and ψ*Est-6* in seven species of the *D. melanogaster* subgroup

37.3 %; Wilcoxon $P=0.0156$; Mann–Whitney $P=0.006$). For *Est-6* there is pronounced difference between GC3 (58.5 %) and GCc (48.6 %) in exon I (Wilcoxon $P=0.0156$; Mann–Whitney $P=0.006$); for ψ*Est-6* the difference is much less pronounced: 47.0 % vs. 48.1 %. For exon II there is no difference between GC3 and GCc for either *Est-6* (Wilcoxon $P=0.2187$; Mann–Whitney $P=0.1649$) or ψ*Est-6* (Wilcoxon $P=0.8125$; Mann–Whitney $P=0.9015$). Thus, GC content has different evolutionary patterns in exon I and II.

For the *Est-6* gene, GC3 does not correlate between exon I and exon II (Pearson correlation coefficient $r=0.5410$, $P=0.2099$; Spearman correlation coefficient $rho=0.4680$, $P=0.2512$). However, there is marginally significant correlation for ψ*Est-6* ($r=0.7515$, $P=0.0515$; $rho=0.725$, $P=0.0758$). The same pattern is detected for the total GC content: there is no correlation between exon I and II for the *Est-6* gene ($r=0.1055$, $P=0.8219$; $rho=0.0$, $P=1.0$) but there is marginally significant correlation for ψ*Est-6* ($r=0.8057$, $P=0.0287$; $rho=0.580$, $P=0.1556$). For *Est-6* there is no correlation between the sequence length and GC3 content ($r=0.1925$, $P=0.3396$) as well as AT content (Spearman nonparametric correlation $rho=0.2335$, $P=0.6142$). However, for ψ*Est-6* both interrelationships are significant ($r=0.7566$, $P=0.0245$; $rho=0.8367$, $P=0.0189$).

3.3 Codon Usage Bias

The scaled chi-square ($S\chi^2$, [86]) measures departure from equal use of synonymous codons, estimated by a χ^2 statistic scaled by dividing it by the number of codons analyzed; the higher the values, the higher the bias, and 0 indicates a perfectly uniform codon usage. The $S\chi^2$ values are 0.180 and 0.138 for *Est-6* and ψ*Est-6*, respectively, indicating that codon bias is higher for *Est-6* (Wilcoxon $P<0.0001$; Mann–Whitney $P<0.0001$). The $S\chi^2$ values are very similar for exon I (0.223) and exon II (0.234) of *Est-6*, but there is noticeable difference between exon I (0.138) and exon II (0.229) in ψ*Est-6*.

The effective number of codons (ENC, [87]) ranges from 20, which means that the bias is at a maximum so that only one codon is used from each synonymous codon group, to 61, which means no bias (all synonymous codons are equally used in each codon group). For the full gene, ENC is 53.12 for *Est-6*, but is 56.16 for ψ*Est-6* (Wilcoxon $P<0.0001$; Mann–Whitney $P<0.0001$). Codon bias is higher for exon II than for exon I in both genes. For exon I, ENC is 51.80 and 56.51 and for exon II ENC is 48.35 and 50.16, for *Est-6* and ψ*Est-6*, respectively.

The codon bias index (CBI, [88]) is a measure of deviation from uniform use that achieves values between 0 and 1 for random use and maximum bias, respectively. CBI is higher for *Est-6* (average 0.2765) than for ψ*Est-6* (average 0.2490) (Wilcoxon $P<0.0001$; Mann–Whitney $P<0.0001$), and higher for exon II than for exon I. The CBI values are 0.285 and 0.250 for the full length of *Est-6* and ψ*Est-6*, respectively; for *Est-6* the CBI values are 0.471 for exon II, but 0.322 for exon I; for ψ*Est-6*, CBI is 0.459 for exon II and 0.263 for exon I.

4 Concluding Remarks

Two different approaches, entropy and GC content analyses, reveal significantly different patterns of evolution in the functional gene *Est-6* and in the putative pseudogene ψ*Est-6* in seven species of the *D. melanogaster* subgroup. GC content was suggested previously as a useful parameter to characterize the human genome [37]. Here we have shown that entropy could be an additional useful and powerful parameter to distinguish functional genes from putative pseudogenes and to investigate gene–pseudogene evolution.

Significantly higher values of entropy for ψ*Est-6* than for *Est-6* are observed in all seven species of the *D. melanogaster* subgroup. The significantly lower structural ordering (regularity) of ψ*Est-6* in comparison with *Est-6* is compatible with the suggestion that ψ*Est*-6 might be a pseudogene.

An interesting feature of the entropy in *Est-6* and ψ*Est-6* is its being nucleotide-dependent: the entropy values for nucleotides A and T are similar for *Est-6* and ψ*Est-6*, but for G and C the entropy values are lower for *Est-6*, which indicates that entropy increase is not a purely random process. There are also significant correlations between intraspecific variability and interspecific divergence in entropy data for both genes. These observations are consistent with other ψ*Est-6* characteristics that combine features of functional and nonfunctional genes [24, 25, 35, 55, 56, 63, 65]. Furthermore, the entropy nucleotide-dependent characteristics of *Est-6* and ψ*Est-6* are related to their position within codons. In all seven species, the main difference is in the GC content of the third codon positions, which is significantly higher in *Est-6* than in ψ*Est*-6. There are few differences between *Est-6* and ψ*Est-6* in AT overall content and in GC content at the first and second codon positions. Interestingly, for *Est-6* there are noticeable differences in the average values of the GC content between the third position (GC3) and the overall coding (GCc) positions, while these differences do not exist for ψ*Est-6* (Table 2).

Besides the high GC content at the third codon position, codon bias indices are higher for *Est-6* than for ψ*Est-6*, and there is a clear difference in codon bias between exon I and exon II in both genes. Codon usage bias in *Drosophila* species is well established [86] and has been attributed both to mutational biases within selectively neutral sequences and to selection to improve translational efficiency [89–93]. The higher values of codon bias for *Est-6* than for ψ*Est-6* probably indicate higher level of selection to improve translational efficiency in *Est-6*; relaxed translational selection in ψ*Est*-6 would lead to increased AT content.

A similar trend in GC content has been observed in comparative investigations of pseudogenes and their functional homologs of *Drosophila* [86, 94–98]. A comprehensive survey of *Caenorhabditis elegans*, *Saccharomyces*, *D. melanogaster*, and human pseudogenes

shows that the nucleotide composition of pseudogenes is invariably intermediate between genes and intergenic regions [99], while *Drosophila* pseudogenes have nearly the same composition as intergenic DNA. In *D. melanogaster*, GC content is uniformly higher at silent sites in coding regions than in the putatively neutrally evolving introns [100].

There is evidence of a general positive correlation between GC3 content and functionality [101, 102]. GC-rich genes tend to be of a greater transcriptional and mitogenic significance than AT-rich genes [101]. Moreover, third-base GC retention also identifies critical amino acids within individual proteins, as indicated by nonrandom patterns of codon variation between homologous genes [101, 102]. Sequence analysis of human receptor tyrosine kinase genes confirms that functionally important transmembrane hydrophobic amino acids are specified by codons containing GC third bases significantly more often than in transmembrane neutral amino acids. Amino acids encoded by GC third bases thus appear more tightly linked to cell function and survival than are those encoded by AT third bases. The same pattern appears in tumor-associated genes undergoing either loss-of-function mutation or rearrangements. As in gene–pseudogene comparisons, genes undergoing loss-of-function mutation tend to be GC-poor, whereas those involved in rearrangements tend to be GC-rich. Moreover, actively transcribed genes use more often C and G at synonymous sites than low-expressed genes [86, 103]. The overall data on pseudogenes and nonfunctional sequences support the hypothesis that sites under low functional constraints tend to increase AT content (*see however* ref. 104). This AT bias has been observed for eukaryotic pseudogenes [105, 106] and thoroughly investigated at the genomic level [99, 107, 108].

In summary, by comparing the entropy and GC content of the functional gene *Est-6* and of the putative pseudogene ψ*Est-6* in seven species of the *D. melanogaster* subgroup we were able to define a number of gene- and pseudogene-specific features:

1. The structural entropy analysis reveals significantly lower structural regularity and higher structural divergence for ψ*Est-6* than for *Est-6*, as expected if ψ*Est-6* is a pseudogene or nonfunctional gene. The accumulation of structural entropy is not however completely random but it is nucleotide-dependent and related to the GC content of the genes.
2. Total GC content is significantly lower in ψ*Est-6* than in *Est-6*, mostly due to GC3, the third codon position.
3. The dispersion of the GC values between functional regions (exon I and exon II) is lower for ψ*Est-6* than for *Est-6*. For *Est-6* there is pronounced difference between GC3 and GCc in exon I; for ψ*Est-6* the difference is much less pronounced. Furthermore, the GC distribution in different codon positions is more uniform in the pseudogene that in the functional gene.

4. There is correlation in GC content between exon I and exon II for the putative pseudogene ψ*Est-6* but not for the functional gene *Est-6*. Analogously, there is correlation between GC content and the length of sequence for the pseudogene but not for the functional gene.
5. Codon bias is significantly higher for the functional gene *Est-6* than for the pseudogene ψ*Est-6*.

Thus, we demonstrate that both entropy and GC content analyses are effective approaches for the identification of pseudogenes and for distinguishing them from functional genes. The observed differences in entropy and GC content may reflect an evolutionary shift associated with ψ*Est-6* pseudogenization and consequent functional divergence of ψ*Est-6* and *Est-6*. The evidence is inconsistent with the hypothesis that ψ*Est-6* is a neutrally evolving, nonfunctional pseudogene, but it is also inconsistent with the hypothesis that ψ*Est-6* functions as a duplicate of *Est-6*, or even that it is a fully functional gene. We conclude that ψ*Est-6* exhibits features of both functional and nonfunctional genes [24, 25, 35, 55, 56, 63, 65]. Consequently, a sharp division between genes and pseudogenes may not be appropriate in this and other contexts. A pseudogene may lose some specific function(s) but retain other(s) or acquire new one(s). There are many examples of functional or "active" pseudogenes (see references above). The term "potogene" may be appropriate for ψ*Est-6*, following Brosius and Gould [109], who have pointed out that the products of a gene duplication, including those that become pseudogenes, may eventually acquire distinctive functions, and thus might be called potogenes to call attention to their potentiality for becoming new genes or acquire new functions [24, 25, 35, 55, 56]. The data on ψ*Est-6* are in accordance with a general picture observed for many other pseudogenes, including almost all known pseudogenes in *Drosophila* [24, 25, 35, 56], while opposed to the traditional view that defines pseudogenes as sequences of genomic DNA that are nonfunctional ("junk" DNA) and consequently free of selection. Our results help to understand why eukaryote genomes contain many pseudogenes that appear to have avoided full degeneration. Furthermore, our data show that intergenic epistatic selection may play an important role in the evolution of the β-*esterase* gene cluster, preserving ψ*Est*-6 from degenerative destruction and reflecting its functional interaction with *Est-6*. The β-*esterase* gene cluster of *D. melanogaster* may represent an example of a functionally interacting complex ("intergene") in which two components (*Est-6* and ψ*Est-6*) or more are required to perform the final function. Hence, the example provided by *Est-6* and ψ*Est-6* as well as by many other genes (*reviewed in* ref. 24) indicate that pseudogenes are an important part of the genome representing a repertoire of sequences often involved in the function of their parental sequences, jointly representing indivisible functionally interacting entities ("intergenic complexes"

Table 3
Contrasting characteristics of *Est-6* and ψ*Est-6* which could be used as an approximate of "identification keys" to differentiate functional genes from putative pseudogenes

Characteristic	Functional gene *Est-6*	Putative pseudogene ψ*Est-6*
Nucleotide variability	Lower	Higher
Population recombination rate[a]	Higher	Lower
Intragenic gene conversion	Lower	Higher
Linkage disequilibrium	Lower	Higher
Entropy	Lower	Higher
GC content	Higher	Lower
GC content dispersion between functional region	Higher	Lower
GC content correlation between functional region	Lower	Higher
Codon bias	Higher	Lower
Number of sites under positive selection	Higher	Lower

"Lower" denotes lower value of a particular characteristic in functional gene or putative pseudogene
"Higher" denotes higher value of a particular characteristic in functional gene or putative pseudogene. Their values are relative to one another for any given characteristic
[a]The population recombination rate is similar for ψ*Est-6* and *Est-6* in the ancestral population (Africa), but significantly different in derived populations (non-African samples)

or "intergenes"), in which each single part cannot successfully accomplish the final functional role without the other(s) [25, 55, 56, 63].

Taking into account the functional importance of pseudogenes [24–28], it is hardly possible to expect complete loss of constraints in pseudogene sequences and, consequently, it is hardly possible to elaborate a method for ultimately unambiguous separation of pseudogenes from functional genes. However, the comparative data described above, along with some other distinctive differences between *Est-6* and ψ*Est-6* inferred from the genetic analysis of *D. melanogaster* populations [35, 55, 56, 63, 65], allow us to define a set of characteristics which clearly distinguish *Est-6* and ψ*Est-6* and could be used as a sort of "identification keys" to delimitate functional genes and putative pseudogenes (Table 3). This set of characteristics reflects the data obtained for the functional gene *Est-6* and the putative pseudogene ψ*Est-6* in the seven species of the *D. melanogaster* subgroup. We, however, believe that these characteristics might also be useful for other species and genes, in order to formulate at least initial hypotheses concerning gene–pseudogene differentiation (classification), identification of pseudogenes, and comparative analysis of gene–pseudogene evolution.

Acknowledgment

We are grateful to Elena Balakireva for encouragement and help.

References

1. Harrison PM, Echols N, Gerstein MB (2001) Digging for dead genes: analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res 29:818–830
2. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 12:272–280
3. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M (2003) Identification of pseudogenes in the *Drosophila melanogaster* genome. Nucleic Acids Res 31:1033–1037
4. Sakai H, Koyanagi KO, Itoh T, Imanishi T, Gojobori T (2003) Detection of processed pseudogenes based on cDNA mapping to the human genome. Genome Informatics 14: 452–453
5. Coin L, Durbin R (2004) Improved techniques for the identification of pseudogenes. Bioinformatics 20(Suppl 1):i94–i100
6. Zhang Z, Carriero N, Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20:62–67
7. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M (2006) PseudoPipe: an automated pseudogene identification pipeline. Comput Appl Biosci 22:1437–1439
8. Bischof JM, Chiang AP, Scheetz TE, Stone EM, Casavant TL, Sheffield VC, Braun TA (2006) Genome-wide identification of pseudogenes capable of disease-causing gene conversion. Hum Mutat 27:545–552
9. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7(Suppl 1):S4
10. Menashe I, Aloni R, Lancet D (2006) A probabilistic classifier for olfactory receptor pseudogenes. BMC Bioinformatics 7:393
11. Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7(Suppl 1):S10–S12
12. van Baren MJ, Brent MR (2006) Iterative gene prediction and pseudogene removal improves genome annotation. Genome Res 16:678–685
13. Zheng D, Gerstein MB (2006) A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol 7(Suppl 1):S13–S20
14. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17: 839–851
15. Ortutay C, Vihinen M (2008) PseudoGeneQuest: service for identification of different pseudogene types in the human genome. BMC Bioinformatics 9:299
16. Molineris I, Sales G, Bianchi F, di Cunto F, Caselle M (2010) A new approach for the identification of processed pseudogenes. J Comput Biol 17:755–765
17. Rouchka EC, Cha IE (2009) Current trends in pseudogene detection and characterization. Curr Bioinformatics 4:112–119
18. Chen S-M, Ma K-Y, Zeng J (2011) Pseudogene: lessons from PCR bias, identification and resurrection. Mol Biol Rep 38:3709–3715
19. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 35:D55–D60
20. Pavlicek A, Paces J, Zika R, Hejnar J (2002) Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene 300:189–194
21. Balakirev ES, Ayala FJ (1996) Is esterase-P encoded by a cryptic pseudogene in *Drosophila melanogaster*? Genetics 144:1511–1518
22. Leveugle M, Prat K, Perrier N, Birnbaum D, Coulier F (2003) ParaDB: a tool for paralogy mapping in vertebrate genomes. Nucleic Acids Res 31:63–67

23. Sakharkar KR, Chaturvedi I, Chow VT, Kwoh CK, Kangueane P, Sakharkar MK (2005) u-Genome: a database on genome design in unicellular genomes. In Silico Biol 5: 611–615
24. Balakirev ES, Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet 37:123–151
25. Balakirev ES, Ayala FJ (2003) Pseudogenes are not junk DNA. In: Wasser SP (ed) Evolutionary theory and processes: modern horizons. Kluwer, The Netherlands, pp 177–193
26. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? RNA 17:792–798
27. Tutar Y (2012) Pseudogenes. Comp Funct Genomics 2012:424526. doi:10.1155/2012/424526
28. Wen Y-Z, Zheng L-L, Qu L-H, Ayala FJ, Lun Z-R (2012) Pseudogenes are not pseudo any more. RNA Biol 9:27–32
29. Lewin B (2007) Genes IX. Oxford University Press, Oxford, NY
30. Lobzin VV, Chechetkin VR (2000) Order and correlations in genomic DNA sequences. The spectral approach. Physics–Uspekhi 43: 55–78
31. Trifonov EN (2011) Thirty years of multiple sequence codes. Genomics Proteomics Bioinformatics 9:1–6
32. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R (1997) Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 13:263–270
33. Yin C, Yau SS-T (2007) Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol 247: 687–694
34. Holste D, Weiss O, Grosse I, Herzel H (2000) Are noncoding sequences of *Rickettsia prowazekii* remnants of "neutralized" genes? J Mol Evol 51:353–362
35. Balakirev ES, Chechetkin VR, Lobzin VV, Ayala FJ (2003) DNA polymorphism in the β-*esterase* gene cluster of *Drosophila melanogaster*. Genetics 164:533–544
36. Vetsigian K, Goldenfeld N (2009) Genome rhetoric and the emergence of compositional bias. Proc Natl Acad Sci U S A 106:215–220
37. Li W (2011) On parameters of the human genome. J Theor Biol 288:92–104
38. Tillo D, Hughes TR (2009) G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10:442
39. Mann S, Chen Y-PP (2010) Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics 95:7–15
40. Dutta C, Paul S (2012) Microbial lifestyle and genome signatures. Curr Genomics 13: 153–162
41. Wu H, Zhang Z, Hu S, Yu J (2012) On the molecular mechanism of GC content variation among eubacterial genomes. Biol Direct 7:2
42. Hildebrand H, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6:e1001107. doi:10.1371/journal.pgen.1001107
43. Raghavan R, Kelkar YD, Ochman H (2012) A selective force favoring increased G+C content in bacterial genes. Proc Natl Acad Sci U S A 109:14504–14507
44. Illingworth RS, Bird AP (2009) CpG islands: 'A rough guide'. FEBS Lett 583:1713–1720
45. Bell CG, Wilson GA, Butcher LM, Roos C, Walter L, Beck S (2012) Human-specific CpG "beacons" identify loci associated with human-specific traits and disease. Epigenetics 7:1188–1199
46. Collet C, Nielsen KM, Russell RJ, Karl M, Oakeshott JG, Richmond RC (1990) Molecular analysis of duplicated esterase genes in *Drosophila melanogaster*. Mol Biol Evol 7:9–28
47. Oakeshott JG, Collet C, Phillis R, Nielsen KM, Russell RJ, Chambers GK, Ross V, Richmond RC (1987) Molecular cloning and characterization of esterase 6, a serine hydrolase from Drosophila. Proc Natl Acad Sci U S A 84:3359–3363
48. Richmond RC, Nielsen KM, Brady JP, Snella EM (1990) Physiology, biochemistry and molecular biology of the *Est-6* locus in *Drosophila melanogaster*. In: Barker JSF, Starmer WT, MacInture RJ (eds) Ecological and evolutionary genetics of Drosophila. Plenum, New York, pp 273–292
49. Oakeshott JG, van Papenrecht EA, Boyce TM, Healy MJ, Russell RJ (1993) Evolutionary genetics of Drosophila esterases. Genetica 90:239–268
50. Oakeshott JG, Boyce TM, Russell RJ, Healy MJ (1995) Molecular insights into the evolution of an enzyme; esterase 6 in *Drosophila*. Trends Ecol Evol 10:103–110
51. Richmond RC, Gilbert DG, Sheehan KB, Gromko MH, Butterworth FM (1980) Esterase 6 and reproduction in *Drosophila melanogaster*. Science 207:1483–1485
52. Gromko MH, Gilbert DF, Richmond RC (1984) Sperm transfer and use in the multiple

mating system of Drosophila. In: Smith RL (ed) Sperm competition and the evolution of animal mating systems. Academic, New York, pp –426

53. Dumancic MM, Oakeshott JG, Russell RJ, Healy MJ (1997) Characterization of the *EstP* protein in *Drosophila melanogaster* and its conservation in Drosophilids. Biochem Genet 35:251–271
54. Healy MJ, Dumancic MM, Oakeshott JG (1991) Biochemical and physiological studies of soluble esterases from *Drosophila melanogaster*. Biochem Genet 29:365–388
55. Balakirev ES, Ayala FJ (2003) Molecular population genetics of the β-*esterase* gene cluster of *Drosophila melanogaster*. J Genet 82:115–131
56. Balakirev ES, Ayala FJ (2004) The β-*esterase* gene cluster of *Drosophila melanogaster*: Is ψ*Est-6* a pseudogene, a functional gene, or both? Genetica 121:165–179
57. Yenikolopov GN, Malevantschuk OA, Peunova NI, Sergeev PV, Georgiev GP (1989) *Est* locus of *Drosophila virilis* contains two related genes. Dokl Acad Nauk SSSR 306:1247–1249 (in Russian)
58. Brady JP, Richmond RC, Oakeshott JG (1990) Cloning of the esterase-5 locus from *Drosophila pseudoobscura* and comparison with its homologue in *D. melanogaster*. Mol Biol Evol 7:525–546
59. East PD, Graham A, Whitington G (1990) Molecular isolation and preliminary characterization of a duplicated esterase locus in *Drosophila buzzatii*. In: Barker JSF, Starmer WT, MacInture RJ (eds) Ecological and evolutionary genetics of Drosophila. Plenum, New York, pp 389–406
60. King LM (1998) The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. Genetics 148:305–315
61. Balakirev ES, Balakirev EI, Rodríguez-Trelles F, Ayala FJ (1999) Molecular evolution of two linked genes, *Est-6* and *Sod*, in *Drosophila melanogaster*. Genetics 153:1357–1369
62. Balakirev ES, Balakirev EI, Ayala FJ (2002) Molecular evolution of the *Est-6* gene in *Drosophila melanogaster*: contrasting patterns of DNA variability in adjacent functional regions. Gene 288:167–177
63. Balakirev ES, Anisimova M, Ayala FJ (2006) Positive and negative selection in the β-*esterase* gene cluster of the *Drosophila melanogaster* subgroup. J Mol Evol 62:496–510
64. Balakirev ES, Ayala FJ (2003) Nucleotide variation of the *Est-6* gene region in natural populations of *Drosophila melanogaster*. Genetics 165:1901–1914
65. Balakirev ES, Chechetkin VR, Lobzin VV, Ayala FJ (2005) Entropy and GC content in the β-*esterase* gene cluster of *Drosophila melanogaster* subgroup. Mol Biol Evol 22:2063–2072
66. Thompson JD, Higgins DJ, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
67. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
68. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452
69. Filatov DA (2002) PROSEQ: a software for preparation and evolutionary analysis of DNA sequence data sets. Mol Ecol Notes 2:621–624
70. Chechetkin VR, Turygin AY (1994) On the spectral criteria of disorder in non-periodic sequences: application to inflation models, symbolic dynamics and DNA sequences. J Phys A Math Gen 27:4875–4898
71. Chechetkin VR, Turygin AY (1995) Search of hidden periodicities in DNA sequences. J Theor Biol 175:477–494
72. Chechetkin VR, Lobzin VV (1998) Nucleosome units and hidden periodicities in DNA sequences. J Biomol Struct Dyn 15: 937–947
73. Chechetkin VR, Lobzin VV (1996) Levels of ordering in coding and non-coding regions of DNA sequences. Phys Lett A 222:354–360
74. Chechetkin VR (2011) Spectral sum rules and search for periodicities in DNA sequences. Phys Lett A 375:1729–1732
75. Kravatskaya GI, Chechetkin VR, Kravatsky YV, Tumanyan VG (2013) Structural attributes of nucleotide sequences in promoter regions of supercoiling-sensitive genes: how to relate microarray expression data with genomic sequences. Genomics 101(1):1–13. doi:10.1016/j.ygeno.2012.10.003, http://dx.doi.org
76. Lemeunier F, David JR, Tsacas L, Ashburner M (1986) The *melanogaster* species group. In: Ashburner M, Carson HL, Thompson JN Jr (eds) The genetics and biology of Drosophila, vol 3e. Academic, London, pp 147–256
77. Cariou M-L (1987) Biochemical phylogeny of the eight species in the *Drosophila melanogaster* subgroup, including *D. sechellia* and *D. orena*. Genet Res 50:181–185

78. Lachaise D, Cariou M-L, David JR, Lemeunier F, Tsacas L, Ashburner M (1988) Biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol 22:159–225
79. Lachaise D, Harry M, Solignac M, Lemeunier F, Benassi V, Cariou M-L (2000) Evolutionary novelties in islands: *Drosophila santomea*, a new *melanogaster* sister species from Sao Tome. Proc R Soc Biol Sci 267:1487–1495
80. Ko W-Y, David RM, Akashi H (2003) Molecular phylogeny of the *Drosophila melanogaster* species subgroup. J Mol Evol 57:562–573
81. da Lage JL, Kergoat GJ, Maczkowiak F, Silvain JF, Cariou ML, Lachaise D (2007) A phylogeny of Drosophilidae using the *Amyrel* gene: questioning the *Drosophila melanogaster* species group boundaries. J Zool Syst Evol Res 45:47–63
82. Drosophila 12 genomes consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218
83. Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM (2012) Estimating divergence dates and substitution rates in the Drosophila phylogeny. Mol Biol Evol 29:3459–3473
84. Yang Y, Hou Z-C, Qian Y-H, Kang H, Zeng Q-T (2012) Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the *Drosophila melanogaster* species group (Drosophilidae, Diptera). Mol Phylogenet Evol 62:214–223
85. Johnson NL, Leone FC (1977) Statistics and experimental design in engineering and the physical sciences, vol II. John Wiley, New York, Ch. 13
86. Shields DC, Sharp PM, Higgins DJ, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716
87. Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29
88. Morton BR (1993) Chloroplast DNA codon usage: evidence for selection at the *psb A* locus based on tRNA availability. J Mol Evol 37: 273–280
89. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907
90. Moriyama EN, Hartl DL (1993) Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134:847–858
91. Heger A, Ponting CP (2007) Variable strength of translational selection among 12 Drosophila species. Genetics 177:1337–1348
92. Vicario S, Moriyama EN, Powell JR (2007) Codon usage in twelve species of *Drosophila*. BMC Evol Biol 7:226
93. de Procé SM, Zeng K, Betancourt AJ, Charlesworth B (2012) Selection on codon usage and base composition in *Drosophila americana*. Biol Lett 8:82–85
94. Starmer WT, Sullivan DT (1989) A shift in the third-codon-position nucleotide frequency in alcohol dehydrogenase genes in the genus *Drosophila*. Mol Biol Evol 6:546–552
95. Moriyama EN, Gojobori T (1992) Rates of synonymous substitutions and base composition of nuclear genes in Drosophila. Genetics 130:855–864
96. Currie PD, Sullivan DT (1994) Structure, expression and duplication of genes which encode phosphoglyceromutase of *Drosophila melanogaster*. Genetics 138:353–363
97. Sullivan DT, Starmer WT, Curtiss SW, Menotti-Raymond M, Yum J (1994) Unusual molecular evolution of an *Adh* pseudogene in *Drosophila*. Mol Biol Evol 11:443–458
98. Ramos-Onsins S, Aguadé M (1998) Molecular evolution of the *Cecropin* multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics 150:157–171
99. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 30: 2515–2523
100. Kliman RM, Hey H (1994) The effects of mutation and natural selections on codon bias in the genes of Drosophila. Genetics 137: 1049–1056
101. Epstein RJ, Lin K, Tan TW (2000) A functional significance for codon third bases. Gene 245:291–298
102. Lin K, Tan SB, Kolatkar PR, Epstein RJ (2003) Nonrandom intragenic variations in patterns of codon bias implicate a sequential interplay between transitional genetic drift and functional amino acid selection. J Mol Evol 57:538–545
103. Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci U S A 96:4482–4487
104. Duret L, Hurst LD (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol Biol Evol 18:757–762

105. Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369
106. Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71
107. Alvarez-Valin F, Lamolle G, Bernardi G (2002) Isochores, GC_3 and mutation biases in the human genome. Gene 300:161–168
108. Zhang Z, Gerstein M (2003) Patterns of nucleotide substitutions, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res 31: 5338–5348
109. Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci U S A 89: 10706–10710