Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Statistical analysis of exon lengths in various eukaryotes

This article was published in the following Dove Press journal:
Open Access Bioinformatics
18 January 2011
Number of times this article has been viewed

Alexander Kaplunovsky[1]
Anatoliy Ivashchenko[2]
Alexander Bolshoy[1]

[1]Department of Evolutionary and Environmental Biology, Genome Diversity Center, Institute of Evolution, University of Haifa, Israel; [2]Department of Biotechnology, Biochemistry, Plant Physiology, Al-Farabi Kazakh National University, Kazakhstan

Purpose: The principal goals of this research were to investigate correlations between certain properties of exons in a gene (ie, between exon density and the corresponding protein length) and to compare genomic trees obtained with different approaches of clustering based on exonic parameters. The aim was a better understanding of exon–intron structures and their origin and development. The exon–intron structures of eukaryote genes are quite different from each other, and the evolution of such structures raises many problematic questions. As a preliminary attempt to address some of these questions, we performed a statistical analysis of gene exon–intron structures.

Methods: Taking whole genomes of eukaryotes, we went through all the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. Comparing those chromosomal and genomic averages, we developed a technique of clustering based on characteristics of the exon–intron structure. This technique of clustering separates different species, grouping them according to eukaryote taxonomy.

Conclusion: Our conclusion is that the best approach is based on distances among four principal components obtained by factor analysis and followed by application of clustering algorithms, such as neighbor-joining, *k*-means, and partitioning around medoids.

Keywords: comparative genomics, exon–intron structure, eukaryotic clustering, principal component analysis

Introduction

It is no secret that people are fond of classifying things. Genomics is no exception. A lot of methods exist in comparative genomics that can be used for the purpose of genome classification.[1] The objective of cluster analysis is to divide objects into clusters in a way that similarity among the items belonging to the same group is higher than similarity among items belonging to distinct groups. In this study, we intend to show that genome clustering based on exon–intron structural characteristics is essentially accurate and reliable and expands on the results of previous studies.[2,3] This kind of clustering neither supports a widely accepted taxonomy nor argues against it. Uncovering and further analyzing the exon–intron structural properties that unify or distinguish genomes in the clustering procedure improve our understanding of the nature and evolutionary history of splicing.

One of the greatest enigmas of eukaryotic genome evolution is the widespread existence of introns. Introns have been detected in the genes of viruses,

Correspondence: Alexander Bolshoy
Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 39105, Israel
Tel +972 4824 0382
Fax +972 4824 0382
Email bolshoy@research.haifa.ac.il

chloroplasts, and mitochondria of both lower and higher eukaryotes. This study focuses on the most important type of introns, ie, the spliceosomal introns of nuclear-encoded protein genes. Here we survey some of the properties of the exon–intron structure of these genes in almost all completely sequenced eukaryotic genomes. Net and averaged exonic lengths are among the attributes considered in this study.

The exon and intron lengths vary across a broad range.[4–8] Statistical analyses of exon and intron lengths have been performed several times on different sets of eukaryotes.[2,3,5,8–15]

Previously, we have shown some genome-specific features of the exon–intron organization of eukaryotic genes using a limited set of genomes from different kingdoms.[2] We have shown that the most general feature found in all genomes is a positive correlation between the number of introns in a gene and the corresponding protein length (ie, the net length of all the exons of the gene). In addition, we have shown that the average exon length correlates negatively with the average number of exons. Recently, analyses of patterns of exon–intron architecture variation brought Zhu et al to the same conclusion.[16] One of their main observations was a decrease in average exon length as the total exon numbers in a gene increased. Although the laws of exon–intron statistics appeared to be quite general, many of the correlation parameters were genome-specific.

Intron density, which is the average number of introns per gene, is an evolutionary riddle. At first, it was thought that one could simply predict intron density from organism complexity. Initial studies supported this hypothesis, ie, *Homo sapiens* has 8.1 introns per gene on average,[17] *Caenorhabditis elegans* has 4.7,[18] *Drosophila melanogaster* has 3.4,[19] and *Arabidopsis thaliana* has 4.4.[20] In contrast, unicellular species were found to have fewer introns per gene.[21] However, further studies found significantly higher intron densities in many unicellular species,[15,22] and intron densities in Basidiomycetes and Zygomycete fungi appeared to be among the highest known for eukaryotes (4–6 per gene).[23,24] Diversity in intron densities among fungal genomes makes them extremely attractive for exploring possible answers to questions concerning exon–intron structure evolution. Indeed, fungi display a wide diversity of gene structures, ranging from less than one intron per gene for yeasts to approximately 1–2 introns per gene, on average, for many recently sequenced lower fungi (including the organisms in this study) and to roughly 5.5 introns per gene on average for some Basidiomycetes (eg, *Cryptococcus*).

Following the genome sequencing of several lower eukaryotes, it has become possible to examine exon–intron statistics with sufficiently large samples of genes. The purpose of our recent publication[3] was to determine the most appropriate approach to classify fungal chromosomes according to simple exon–intron statistics. We tested a few clustering techniques measuring distances among the chromosomes in different ways. As a result of our analysis, we commented on the consistent similarity of the partitions, resulting from different clustering methods. Clustering results[3] obtained with scaled and normalized Euclidean distances appeared to be sufficiently similar. The principal components-based clustering method, the principal directions divisive partitioning method, and the neighbor-joining algorithm produced very similar clustering results. Therefore, we propose techniques of clustering that are able to distinguish between chromosomes of different species with satisfactory results. The addition of regression parameters to averaged chromosomal parameters improves the resolution of clustering.

There is a mixture of different chromosomal characteristics in exon–intron organization. In this study, similar to our previous publications, we considered only pure exonic properties and, additionally, proportions of intron-containing genes among all protein-coding genes. We calculated and compared exonic properties, including exon densities, average exon lengths, and average net exon lengths. In this study we investigated the correlation between the number of exons in a gene (exon density) and the corresponding protein length; compared intragenomic variation with intergenomic variance of exon densities, average exon lengths, and average net exon length; compared genomic trees obtained using different approaches of clustering based on exonic parameters; and paved a road for further evolutionary in silico research of exon–intron structure and its origins and development.

Methods

Data set

The nucleotide sequences of 322 chromosomes of 32 species presented in Table 1 were obtained from the database of the Eukaryotic Genome Sequencing Projects (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). Gene annotations were used to calculate genic statistical properties. A standard gene annotation looks like the following annotation of a randomly chosen gene, NCU08052.1 of *Neurospora crassa*:

Gene <25457..>26451.
mRNA join (<25457..25690,25755..26055,26117..>26451).

Coding sequence join (25457..25690,25755..26055, 26117..26451).

The annotation means the first exon of this gene starts somewhere upstream of position 25457, and the last exon of the gene ends somewhere downstream of position 26451. For the purposes of this study, the term "exons" refers to "coding parts of exons". In other words, only those introns within coding sequences and exons without untranslated regions were used for analysis. The data related to coding parts of exons are taken from coding sequence lines. For example, the coding sequence of NCU08052.1 consists of three "exons" (25457:25690, 25755:26055, and 26117:26451) with lengths of 234 bp, 301 bp, and 335 bp, respectively. The length of the gene is greater than 995 bp, the number of exons is equal to 3, the net length of the exons (the protein size in bp) is equal to 870, and the average exon length is equal to 290.

Exon–intron structure and statistical parameters

Each gene was assigned three gene-related exonic values, ie, the net length, $L_{ex,}$ of all its exons, the number, N_{ex}, of those exons, and an average exon length, A_{ex}:

$$A_{ex} = \frac{L_{ex}}{N_{ex}}$$

For each chromosome of each genome, several absolute and averaged chromosomal characters were calculated. In addition to the three averaged characteristics of exons, the average net length, l_{ex}, of all the exons in a gene per chromosome, the average number, n_{ex}, of the exons in a gene per chromosome, the average exon length, a_{ex}, per chromosome, and the proportion of intron-containing genes, p_c, as a relevant attribute were calculated. It should be

Table 1 List of processed species and their chromosomes

Kingdom/ supergroup	Phylum	Class	Organism	Abbreviation	Chromosomes (n)
Animalia	Arthropoda	Insecta	*Drosophila melanogaster*	DM	6
	Chordata	Mammalia	*Canis familiaris*	CF	19
			Homo sapiens	HS	10
			Mus musculus	MM	10
	Nemata	Caenorhabditis	*Caenorhabditis elegans*	CE	6
Fungi	Ascomycota	Ascomycetes	*Neurospora crassa*	NC	7
		Eurotiomycetes	*Aspergillus fumigatus*	AF	8
		Saccharomycotina	*Candida glabrata*	CG	13
			Debaryomyces hansenii	DH	7
			Eremothecium gossypii	EG	7
			Kluyveromyces lactis	KL	6
			Pichia stipitis	PS	8
			Saccharomyces cerevisiaei	SC	16
			Yarrowia lipolytica	YL	6
		Sordariomycetes	*Gibberella zeae*	GZ	4
			Magnaporthe grisea	MG	7
		Taphrinomycotina	*Schizosaccharomyces pombe*	SP	3
	Basidiomycota	Agaricomycotina	*Cryptococcus neoformans*	CN	14
		Ustilaginomycotina	*Ustilago maydis*	UM	23
	Microsporidia	Apansporoblastina	*Encephalitozoon cuniculi*	EC	11
Plantae	Magnoliophyta	Liliopsida	*Oryza sativa*	OS	12
		Magnoliopsida	*Arabidopsis thaliana*	AD	5
Plantae/Viridiplantae	Chlorophyta	Prasinophyceae	*Micromonas* sp. RCC299	MS	17
			Ostreococcus_lucimarinus	OL	21
Protista/ Chromalveolata	Ciliophora	Ciliatea	*Paramecium tetraurelia*	PT	1
	Apicomplexa	Aconoidasida	*Plasmodium falciparum*	PF	14
			Plasmodium knowlesi	PK	14
			Theileria annulata	TA	3
Protista/Chromista	Cryptophyta	Cryptophyceae	*Guillardia theta*	GT	3
			Hemiselmis anderenii	HA	3
Protista/Protozoa	Euglenozoa	Kinetoplastea	*Leishmania braziliensis*	LB	35
Protista/Rhizaria	Cercozoa	Chlorarachniophycea	*Bigelowiella natans*	BN	3
Total					**322**

mentioned that a_{ex} is the mean of the A_{ex} values of individual genes per chromosome:

$$a_{ex} = \frac{1}{n}\sum_{1}^{n} A_{ex}$$

where n denotes a number of genes in the chromosome here. The measure a_{ex} defined in this is different from the average length, $\bar{a}_{ex}$, of all the exons in the chromosome, regardless of which gene(s) they belong to. The $\bar{a}_{ex}$ is calculated as the total length of all exons in a chromosome divided by the total number of all exons in a chromosome.[7] The a_{ex} usually have significantly larger values than the $\bar{a}_{ex}$ because an average length of i-th exon exponentially decreases with an index, i.[25]

We also calculated species-averaged exon parameters, ie, N_g (total number of genes per genome), AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a gene per genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in an intron-containing gene per genome), $AL0_{ex}$ (average length of an intronless gene per genome), $AL1_{ex}$ (average net length of all exons in an intron-containing gene per genome), and P_g (proportion of intron-containing genes in a genome in percent).

Distances between pairs of genomes

One of our goals was to cluster genomes using exon–intron structure parameters. We used distance-based methods of clustering, so had to define a method for distance measurement. The distance between a pair of genomes was calculated as the distance between vectors constructed from several standardized parameters defined above. The vector $\overline{x}_r$ of genomic parameters related to genome r consists of (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$), and is equal to

$$\overline{x}_r = \left\{\frac{j_{ex,r} - \mu_j}{\sigma_j}\right\},\ j \in \left\{AN_{ex}, AL_{ex}, AA_{ex}, AN1_{ex}, AL1_{ex}, AL0_{ex}\right\},$$

where μ_j is the mean value of a genomic parameter j and σ_j is its standard deviation.

Having extracted these parameters, our next task was to find an appropriate dissimilarity measure, d, such that $d(x_r, x_s)$ is small if x_r and x_s are close. The simplest dissimilarity measure is a normalized (standardized) Euclidean distance:

$$d(x_r, x_s) = \sqrt{\sum_{k=1}^{K}\left(\overline{x}_{r,k} - \overline{x}_{s,k}\right)^2}$$

Clustering of genomes

A few popular algorithms were used to cluster all 32 genomes. First of all, the well known neighbor-joining algorithm[26] was used. Using neighbor-joining, a tree that does not assume an evolutionary clock was constructed, and therefore, in effect, an unrooted tree results. We used the Neighbor of Phylip program package from the University of Washington (http://evolution.genetics.washington.edu/phylip/doc/neighbor.html), which is an implementation of neighbor-joining. Matrices of standardized distances between all pairs of chromosomes were exported to the Neighbor program. The output file was drawn by the TreeView program of Professor Rod Page (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). We also used well known k-medoid clustering by partitioning around medoids and k-means algorithms. The k-medoids and k-means algorithms are described elsewhere.[1]

Analyses of structural–functional organization of the system

One-way analysis of variance (ANOVA) was used to test for differences in exon–intron structure between several groups of species. We also used factor analysis as an integral statistical method, affording an opportunity to define and evaluate the structural–functional organization of the system. We chose principal components analysis as one of the techniques for factor analysis. This method produces a set of eigenvectors calculated from the matrix of correlations between parameters where each set represents a causal connection of elements. It is important to note that by using the technique of principal components analysis, all factors become orthogonal and are caused by different properties of the system.

Results

The aforementioned chromosomal characteristics (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) were calculated for all 322 chromosomes. As an illustration, the values of these characteristics for a randomly selected unicellular organism, *Plasmodium knowlesi*, are given in supplementary Table S1. Every column in Table S1 contains indistinguishable parameters. The intragenomic variation was found to be rather small for other unicellular organisms as well, as shown with fungi.[3]

The results of the one-way ANOVA test for differences in the first three parameters, n_{ex}, l_{ex}, and a_{ex}, of chromosomes of all genomes are presented in Table 2. In general, we found intragenomic variation in l_{ex} and a_{ex} to be quite small for almost all unicellular organisms, and this was significant in n_{ex}, l_{ex}, and a_{ex} for Plantae and Animalia (especially for

Table 2 Results of one-way ANOVA test for differences in parameters between chromosomes for several species

Organism	Kingdom/supergroup	n_{ex}	l_{ex}	a_{ex}
AD	Plantae	0.006**	0.000***	0.007**
AF	Fungi	0.211	0.097	0.431
BN	Protista/Rhizaria	0.591	0.790	0.193
CE	Animalia	0.000***	0.000***	0.000***
CF	Animalia	0.000***	0.000***	0.000***
CG	Fungi	–	0.979	0.976
CN	Fungi	0.591	0.764	0.077
DH	Fungi	–	0.190	0.058
DM	Animalia	0.000***	0.000***	0.000***
EC	Fungi	–	0.203	0.226
EG	Fungi	–	0.377	0.423
GT	Protista/Chromista	–	0.128	0.112
GZ	Fungi	0.000***	0.040**	0.000***
HA	Protista/Chromista	–	0.599	0.599
HS	Animalia	0.002**	0.123	0.000***
KL	Fungi	–	0.427	0.389
LB	Protista/Protozoa	–	0.003**	0.002**
MG	Fungi	0.045**	0.014*	0.565
MM	Animalia	0.000***	0.000***	0.000***
MS	Plantae/Viridiplantae	0.000***	0.342	0.002**
NC	Fungi	0.037*	0.009**	0.947
OL	Plantae/Viridiplantae	0.000***	0.305	0.863
OS	Plantae	0.000***	0.075	0.000***
PF	Protista/Chromalveolata	0.083	0.168	0.053
PK	Protista/Chromalveolata	0.471	0.770	0.548
PS	Fungi	0.253	0.392	0.203
PT	Protista/Chromalveolata	–	–	–
SC	Fungi	0.692	0.993	0.985
SP	Fungi	0.651	0.570	0.321
TA	Protista/Chromalveolata	0.004**	0.771	0.680
UM	Fungi	0.309	0.539	0.366
YL	Fungi	–	0.319	0.523

Notes: *significance $0.01 < P < 0.05$; **significance $0.001 < P < 0.01$; ***significance $P < 0.001$; –parameters that did not pass the Levene test of homogeneity.

D. melanogaster chromosomes, with an outstanding and short chromosome 4). Table 2 shows that the sets a_{ex}, l_{ex}, and n_{ex} in various chromosomes demonstrate significant differences. We can see that F-statistics comparing variances between and within groups of chromosomes are significant. The ANOVA method was used only for parameters that passed the Levene test of homogeneity. As can be seen, most species with a low percentage of intron-containing genes in chromosome p_c did not pass this test for n_{ex}.

Problems investigated in this study included correlations between different species-averaged parameters of exon–intron structure, clustering chromosomes of a few organisms belonging to the same kingdom (Protista, Plantae, and Animalia) by combinations of chromosome-averaged exonic characteristics, and clustering of all 32 organisms by combinations of species-averaged characteristics of exons.

Correlations among species-averaged statistical parameters

In Table 3, in addition to parameters averaged over all genes, there are data related to a set of "intron-containing" genes ($AL1_{ex}$) and to a set of "intronless" genes ($AL0_{ex}$). In the Methods section, there are descriptions and formulae for calculations of these parameters. Some putative empiric rules may be deduced from Table 3. For example, regarding average protein lengths of intron-containing and intronless genes (net length of all exons), it seems that if there is only a small amount of intron-containing genes in a genome, such proteins are shorter on average than other proteins coded by intronless genes of the same genome. This property is especially strongly expressed for some species of fungi (*Encephalitozoon cuniculi*, *Candida glabrata*, and *Kluyveromyces lactis* and also exists for *Eremothecium gossypii*, *Debaryomyces hansenii*, *Schizosaccharomyces pombe*, and *Ustilago maydis*), and for three Protista species (*Leishmania braziliensis*, *Hemiselmis anderenii*, and *Guillardia theta*). Figure 1 shows a scatter-plot of P_g versus a fraction of $AL0_{ex}/AL1_{ex}$ and is obtained from Table 3. *H. anderenii* does not appear in Figure 1 because it has no intron-containing genes. There are three main groups of points in the plot, ie, a group of genomes with a low concentration of intron-containing genes ($P_g < 10\%$), a group of genomes with a high concentration of intron-containing genes ($P_g > 70\%$), and an intermediate group. The first group is mainly characterized by a striking prevalence of longer genes among intronless genes compared with intron-containing ones. We could deduce a rule that, in genomes with a low presence of intron-containing genes, such genes are coding shorter proteins. However, there is an exception to this empiric rule, ie, *L. braziliensis*, which has a fraction $AL0_{ex}/AL1_{ex}$ similar to genomes with high P_g. An empiric rule for the second group may be formulated that there is a (linear) positive correlation between a proportion of intron-containing genes in a genome and a fraction $AL0_{ex}/AL1_{ex}$ while values of a fraction are lower than 1. Unfortunately, we have an exception to this rule as well, ie, *Bigelowiella natans*, which has a surprisingly high value of the ratio $AL0_{ex}/AL1_{ex}$. Regarding the central group, we may say only that it has an intriguing configuration that requires further investigation.

Chromosome-averaged statistical parameters

Let us consider the average parameters l_{ex}, n_{ex}, and a_{ex}. A scatter-plot of a_{ex} versus l_{ex} is shown in Figure 2B for Protista and illustrates the statement made previously that

Table 3 Species dependent exonic parameters

Organism	Kingdom/supergroup	AN_{ex}	AL_{ex}	AA_{ex}	ANI_{ex}	ALI_{ex}	P_g	ALO_{ex}
AD	Plantae	5.255	1243	412	6.404	1322	78.65	951
AF	Fungi	2.918	1465	668	3.453	1513	78.14	1289
BN	Protista/Rhizaria	4.054	960	359	4.580	907	85.37	1142
CE	Animalia	6.283	1284	213	6.428	1306	97.34	457
CF	Animalia	10.697	1696	232	11.450	1762	92.74	843
CG	Fungi	1.016	1509	1504	2.025	662	1.56	1522
CN	Fungi	6.271	1611	319	6.445	1627	96.87	1123
DH	Fungi	1.057	1387	1357	2.075	1070	5.41	1402
DM	Animalia	3.914	1824	503	4.686	2007	81.12	906
EC	Fungi	1.008	1071	1069	2.143	438	0.71	1075
EG	Fungi	1.048	1472	1452	2.035	874	4.58	1485
GT	Protista/Chromista	1.033	939	930	2.000	583	3.29	952
GZ	Fungi	3.261	1531	623	3.590	1553	82.04	1359
HA	Protista/Chromista	1.000	1019	1019	–	–	0.00	1019
HS	Animalia	8.868	1533	280	10.167	1656	85.76	790
KL	Fungi	1.025	1418	1409	2.017	760	2.47	1435
LB	Protista/Protozoa	1.012	1905	1882	2.040	3854	1.16	1882
MG	Fungi	2.844	1394	654	3.480	1468	74.34	1179
MM	Animalia	8.248	1457	302	9.658	1575	83.53	848
MS	Plantae/Viridiplantae	1.516	1488	1166	2.447	1636	34.58	1407
NC	Fungi	2.703	1476	694	3.136	1505	79.73	1366
OL	Plantae/Viridiplantae	1.279	1253	1100	2.344	1388	20.06	1222
OS	Plantae	4.846	1237	440	6.054	1348	75.96	890
PF	Protista/Chromalveolata	2.440	2238	1490	3.603	2131	55.30	2377
PK	Protista /Chromalveolata	2.591	2189	1486	4.094	2021	51.43	2373
PS	Fungi	1.408	1495	1227	2.551	1732	26.28	1409
PT	Protista/Chromalveolata	3.337	1583	583	3.803	1674	83.37	1128
SC	Fungi	1.055	1482	1444	2.035	1434	5.31	1485
SP	Fungi	1.951	1413	1040	3.098	1305	45.36	1501
TA	Protista/Chromalveolata	3.775	1581	785	4.964	1525	69.96	1716
UM	Fungi	1.782	1839	1439	3.025	1649	38.60	1961
YL	Fungi	1.158	1458	1339	2.131	1637	13.92	1428
Total		3.121	1579	1057	4.204	1606	42.97	1417

the averages of l_{ex} and a_{ex} were fairly similar for different chromosomes of the same species but, as a rule, rather distant for different species. Moreover, six separate groups of points may be observed in Figure 2B.

We colored all points using four colors relating to four Protista supergroups, ie, Chromalveolata (*Plasmodium falciparum*, *P. knowlesi*, *Paramecium tetraurelia*, and *Theileria annulata*), Chromista (*Guillardia theta, H. anderenii*), Protozoa (*L. braziliensis*), and Rhizaria (*B. natans*, see Table 1). Analyzing the contents of the groups presented in Figure 2, one can suppose that the divisions follow their taxonomy. Indeed, scatter-plots of a_{ex} vs n_{ex} (Figure 2A) and a_{ex} vs l_{ex} (Figure 2b) clearly show six separate groups of chromosomes; *B. natans* chromosomes belonging to Rhizaria form the left-most group, *G. theta* and *H. anderenii* chromosomes belonging to Chromista are located together, and Protozoa (*L. braziliensis*) form the third cluster. Chromosomes belonging to Chromalveolata form three clusters, according to their phylum and class, ie, Apicomplexa Plasmodium (*P. falciparum* and *P. knowlesi*), Apicomplexa Theileria (*T. annulata*), and a single chromosome of Paramecium (*P. tetraurelia*). These scatter-plots show that the three parameters a_{ex}, n_{ex}, and l_{ex} are sufficient for successful classification of 76 chromosomes to eight unicellular organisms.

The same conclusion regarding classification mirroring the phyla taxonomy can be made following an analysis of the matching chromosomal parameters for Animalia. Scatter-plots of a_{ex} versus l_{ex} and a_{ex} versus n_{ex} for Animalia are shown in Figure 3. Points related to averages l_{ex} and a_{ex} were related to different chromosomes of the same species and were located quite close to one another, whereas points related to chromosomes of different species are placed distant from one another. Striking exceptions are the points associated with chromosome 4 of *D. melanogaster* and chromosome 7 of *Mus musculus*. These points form clusters of a single member clearly disjointed from other

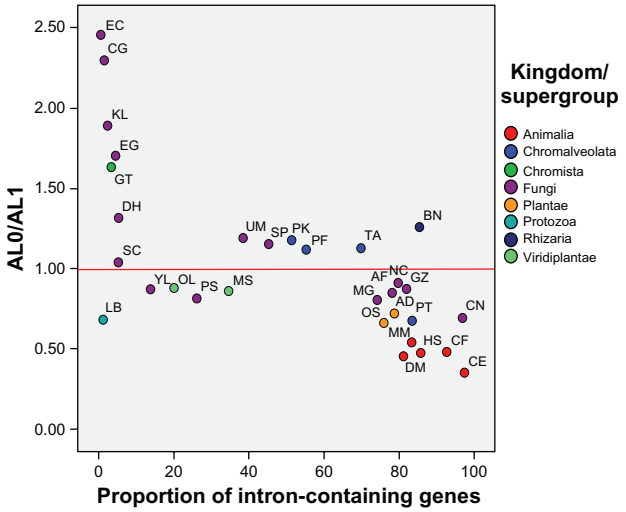

Figure 1 Scatter-plot of P_g showing the proportion of intron-containing genes in a genome on the x-axis versus the ratio between $AL0_{ex}$ (an average length of an intronless gene) and $AL1_{ex}$ (an average net length of all exons in an intron-containing gene) on the y-axis for all 32 genomes. The red line marks the level of equality of $AL0_{ex}$ and $AL1_{ex}$.

groups. All other points form three separate groups, which may be observed in Figure 3. The two parameters l_{ex} and a_{ex} separately cluster five chromosomes of *D. melanogaster* in one group, six chromosomes of *C. elegans* in another group, and all 39 chromosomes of *Canis familiaris*, *H. sapiens*, and *M. musculus* in the third group.

Let us repeat our observations deduced from Figure 3 relating to the phyla. We colored all points in three colors related to three animal phyla (see Table 1), ie, Arthropoda, Chordata, and Nemata. Figure 3a presents a scatter-plot of a_{ex} versus n_{ex} and clearly shows three separate groups of chromosomes and two outliers. *C. familiaris*, *H. sapiens*, and *M. musculus* chromosomes belonging to Chordata Mammalia form the left-most group; *C. elegans* chromosomes belonging to Nemata Caenorhabditis appear in the second left group; and the points belonging to *D. melanogaster* (Arthropoda Insecta) appear in the right group. Two chromosomes, ie, DM4 (the shortest chromosome of *D. melanogaster*) and MM07, form two separate groups, each one with a single member. The *C. elegans* chromosomes have the greatest exon density (n_{ex}) and the shortest exons (l_{ex}) among all the animal chromosomes studied.

Clustering of genomes by species-averaged statistical parameters

After the relatively satisfying success of partial clustering based on only three chromosomal characteristics, our next objective was to cluster all 32 genomes. We took seven species-averaged exon parameters mentioned previously, ie, AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a gene per genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in an intron-containing gene per genome), $AL0_{ex}$ = average (over a genome) length of an intronless gene, $AL1_{ex}$ (average net length of all exons in an intron-containing gene per genome), and P_g (proportion of intron-containing genes in a genome expressed as a percentage). The expectation was that clustering would generally follow the kingdom/supergroup/phylum classification. However, the

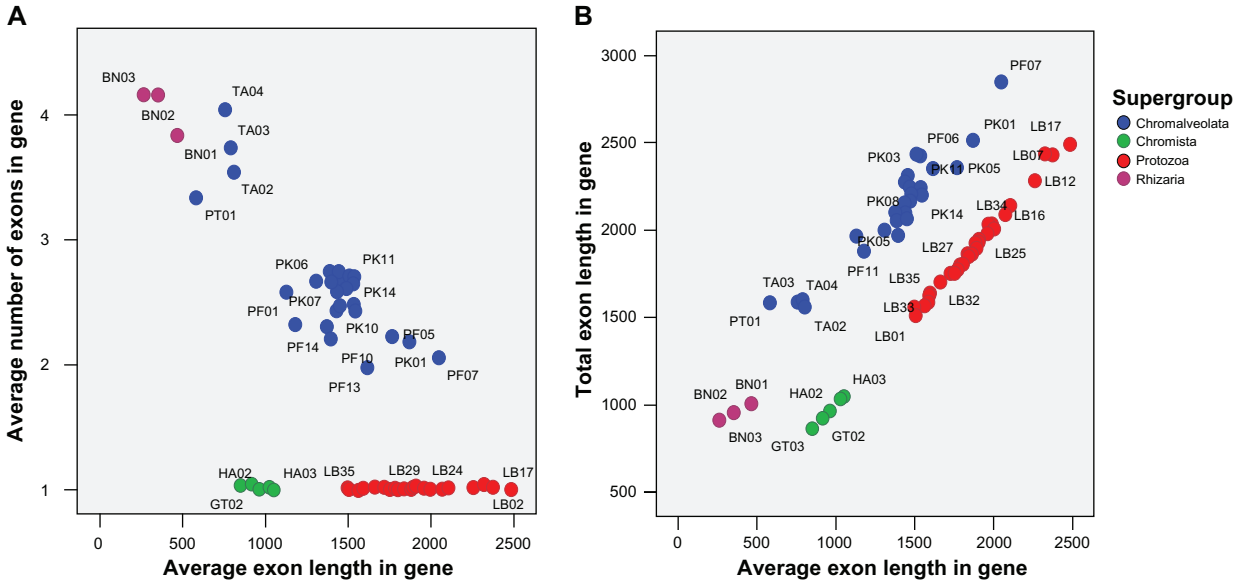

Figure 2 Scatter-plot for 76 processed chromosomes of eight Protista species, colored by four supergroups. Plot presents the average exon length per gene a_{ex} (x-axis) **A**) versus the average number of exons per gene n_{ex} (y-axis) and **B**) versus the average net exon length per gene l_{ex} (y-axis).

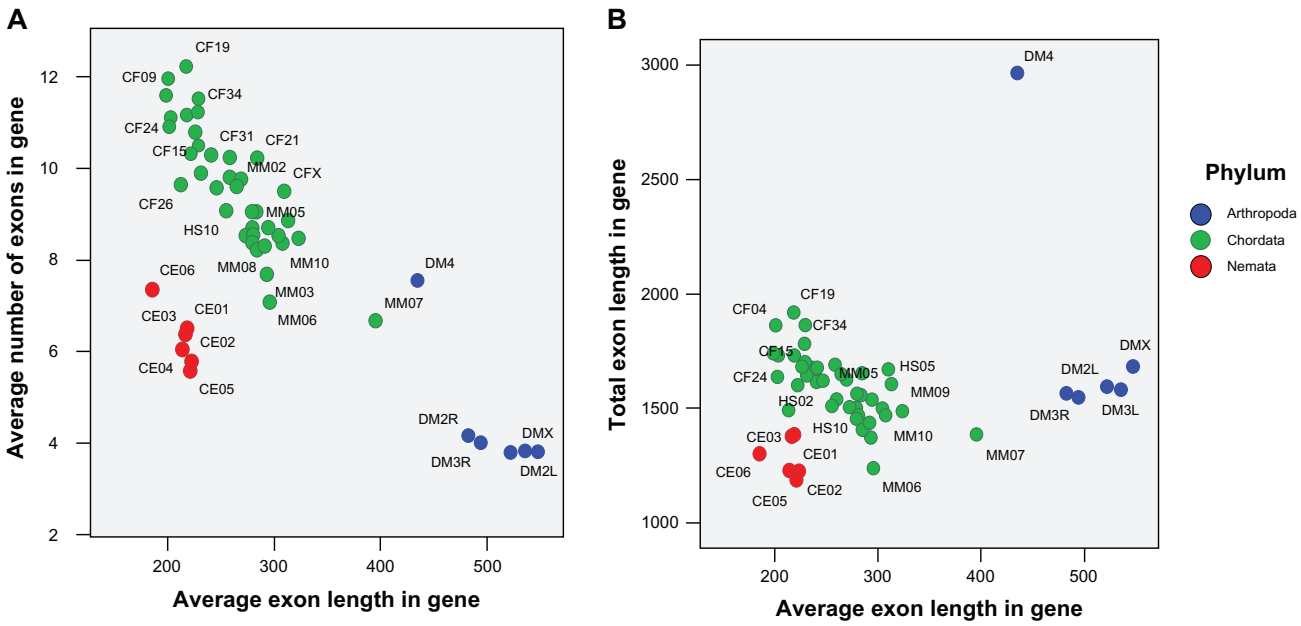

Figure 3 Scatter-plot for 59 processed chromosomes of five Animalia species, colored by three phyla. Plot presents the average exon length per gene a_{ex} (x-axis) **A**) versus the average number of exons per gene n_{ex} (*y*-axis) and **B**) versus the average net exon length per gene l_{ex} (*y*-axis).

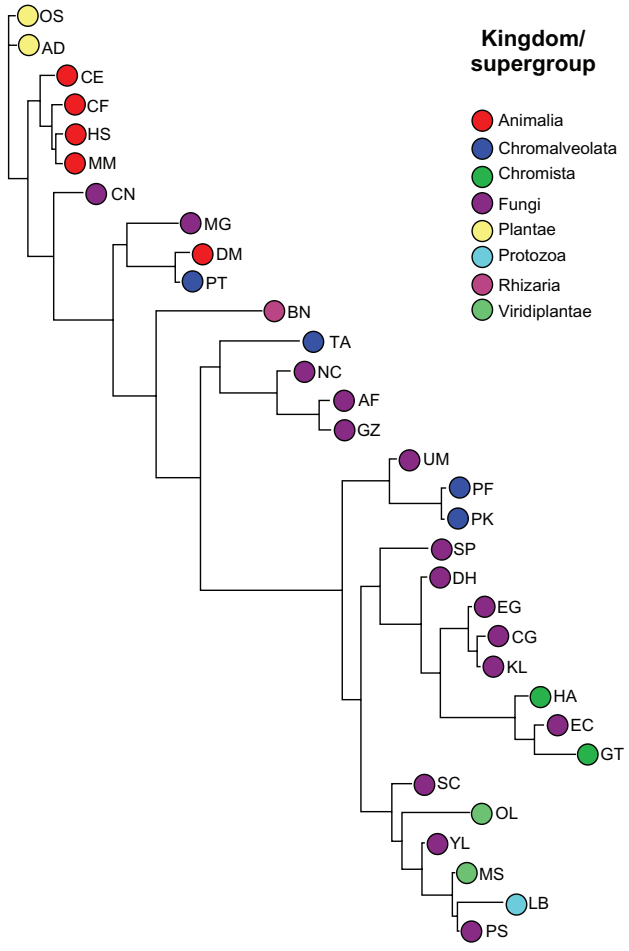

Figure 5 Dendrogram of 32 processed genomes obtained by neighbor-joining clustering technique and based on distances among four principal components obtained by factor analysis of AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$.

results were poor (data not shown). Assuming that a peculiar relationship between a parameter P_g and other parameters (see Figure 1) may negatively influence clustering, we excluded this parameter from further consideration.

At this point, we tried to cluster genomes of 32 different organisms using six parameters, namely AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$. As a first stage, we applied neighbor-joining clustering using standardized distances among the vectors (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$) and applying the Neighbor program. The dendrogram presented in Figure 4 was drawn by the TreeView program. As one can see, some organisms of the same kingdom/supergroup are distributed compactly along the tree. Nevertheless, not all species belonging to the same class form a monophyletic cluster. Mice (*M. musculus*), dogs (*C. familiaris*), and humans (*H. sapiens*) are located together, but flies (*D. melanogaster*), which form a cluster together with Protista/Chromalveolata, *T. annulata*, appear too far away from other Animalia. Viridiplantae species are placed distantly, and Protista are distributed along the tree in a strange manner. Such a classification, although better than the classification produced by seven parameters, cannot be considered adequate.

These discrepancies could be explained at least partially by the cross-dependencies of all the parameters considered. Therefore, the way to improve clustering is to replace these parameters by independent (orthogonal) parameters that could be obtained, eg, from results of a factor analysis of their correlation matrix as principal components.

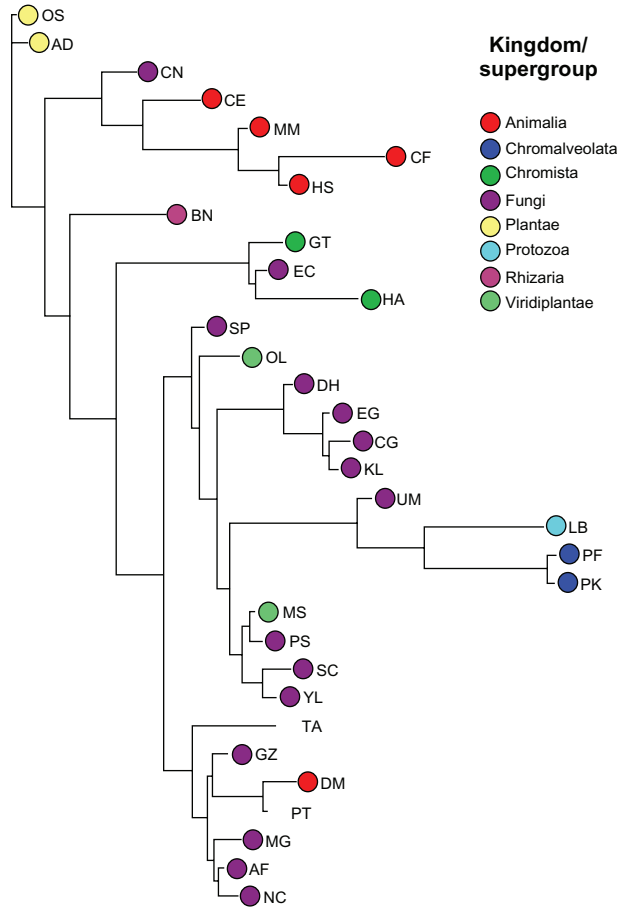

Figure 4 Dendrogram of 32 processed genomes obtained by the neighbor-joining clustering technique and based on standardized distances among parameters AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$.

The factor analysis led us to the synthesis of the following successive logical structure:

- Dividing the system into sets of "elementary" components, ie, all of the aforementioned genomic characteristics (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex,}$ $AL1_{ex}$, $AL0_{ex}$)
- Analysis of the relationships of these components in species
- Revealing system-forming relationships
- Description of the structure of the system (model) and its properties.

As shown in Table S2, four principal components are responsible for 99.4% of the organization of the whole system, and the first two describe 86.2% of the whole variability of the system. Four principal components (Table S3) have been used in genome clustering based on neighbor-joining, *k*-means, and partitioning around medoids. Results of neighbor-joining clustering are presented in Figure 5.

There are certain improvements comparing the clustering presented in Figure 4. Viridiplantae species are placed closely, and Protista are distributed along the tree less strangely than in Figure 4. However, *D. melanogaster* couples with the Protista *T. annulata* again.

Results of *k*-means (Table S4) clustering are very similar (practically identical) to the neighbor-joining results shown earlier. These *k*-means results are shown in Table S4. Results of partitioning around medoids clustering are presented in Table S5. These results are similar to neighbor-joining results as well. However, there are some additional improvements in partitioning genomes among different clusters. In general, the results show a high consistency of partitioning, in spite of differences in clustering techniques. Careful examination of Table S5 reveals hierarchic partitioning of organisms. Interestingly, partitioning around medoids clustering is not a hierarchic algorithm and should not necessarily produce any hierarchy. In our case of application of partitioning around medoids clustering to four principal components obtained by factor analysis, a strictly hierarchic structure is produced. In fact, the *k*-medoids clustering was performed for different values of *k* between 2 and 20, and it was observed that the clustering for a given value of *k* is always a strict subclustering of the clustering for *k–1*. This may be interpreted as existence of an intrinsic hierarchic structure of principal components analysis data. This may, in turn, serve as additional evidence of variance in the evolutionary nature of exon–intron structure.

Discussion

The origin of introns remains a mystery, and certain questions in molecular evolution are being investigated by in silico analysis of intron–exon structures in various organisms. To facilitate such studies, while taking advantage of the burgeoning amount of sequence data now available, we undertook a statistical analysis of the exon–intron structure for nearly all completely sequenced eukaryotic genomes in order to reveal general and genome-specific features of eukaryotic genes. We went through all of the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of exons, and the average length of an exon. Furthermore, we tried to determine the most appropriate approach to classifying eukaryotic chromosomes, according to these simple exon–intron statistics.

One of the main conclusions of the studies by Kaplunovsky et al[2] and Atambayeva et al[13] was that a positive correlation exists between the number of introns in a gene and the length of the corresponding protein (and equivalently the net length

of all the exons of the gene). Here, like Kaplunovsky et al,[3] we confirmed the observation of Ivashchenko et al[15] that, for all fungal genomes with a proportion of intron-containing genes higher than 30%, gene size and total exon length depend on the intron number in a linear manner. The correlation problem is irrelevant for organisms with an extremely low proportion of intron-containing genes, such as yeasts, Protista/Chromista, and Protista/Protozoa.

In a previous publication,[2] we reported that intragenomic variation is substantially smaller than intergenomic variance in almost all fungal genomes. In other words, we found that the laws of exon–intron statistics are specific to genomes rather than to individual chromosomes. In this respect, the similarity in exon–intron structures for dogs (*C. familiaris*), mice (*M. musculus*), and humans (*H. sapiens*) is so striking that intragenomic and intergenomic variances of the sets a_{ex}, l_{ex}, and n_{ex} in various chromosomes are practically undetectable (see Table 2). A similar statement can be made regarding two plants in this study, ie, *Arabidopsis* and rice, and thus we confirmed the observations made by Atambayeva et al.[13]

Noteworthy is the similarity in the exon–intron structures of an insect, *D*. melanogaster, and a protist, *T. annulata* (see Table 3). Neither environmental habitat factors nor the evolutionary history of organisms provide any clue to solving the mystery of the proximity of these two genomes on the genome tree based on exon–intron characteristics. Perhaps the appearance of other eukaryotes in the data set of completely sequenced genomes will provide the answer.

The main advances of this study over previous research[2,3,5,8–15] lie in the larger amount of genomes considered and the concentrated efforts made to determine the most appropriate approach for clustering based on exonic characteristics. We checked a few procedures of clustering based on exon–intron structure features averaged over intron-containing or intronless genes. As a result, we conclude that the most successful procedure should be based on distances between four principal components obtained by factor analysis and followed by application of clustering algorithms. The consistency of recovered cluster structures may be considered evidence of hidden evolutionary resemblance.

We concentrated our efforts on comparison of exonic parameters, while planning to work on intron lengths later. Clearly, the exon–intron structures of eukaryotic genes have many important parameters that we did not consider in this work, and we intend to pursue these in future research. In particular, the ratio of exon and intron lengths promises to be an important feature of a gene. In some genomes, the intron length is comparable with the exon length, ie, in unicellular eukaryotes,[4,5] plants,[5,27] and particular animals.[5–7] In general, introns are longer than exons in mammalian genes.[14] Correlations of intronic characteristics with such genomic properties as gene density would be a goal for further research as well.

Acknowledgments

We are grateful to S Hosid and Z Volkovich for technical assistance with this research.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Bolshoy A, Volkovich Z, Kirzhner V, Barzily Z. Genome clustering: From Linguistic Models to Classification of Genetic Texts. *Studies in Computational Intelligence Series*. Kacprzyk J, editor. Berlin, Heidelberg: Springer-Verlag; 2010.
2. Kaplunovsky A, Khailenko VA, Bolshoy A, Atambayeva SA, Ivashchenko AT. Statistics of exon lengths in animals, plants, fungi, and protists. *Int J Biol Life Sci.* 2009;1:139–144.
3. Kaplunovsky A, Zabrodsky D, Volkovich Z, Ivashchenko AT, Bolshoy A. Statistics of exon lengths in fungi. *Open Bioinformatics J.* 2010;4:31–40.
4. Kupfer DM, Drabenstot SD, Buchanan KL, et al. Introns and splicing elements of five diverse fungi. *Eukaryot Cell.* 2004;3:1088–1100.
5. Deutsch M, Long M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 1999;27:3219–3228.
6. Wendel JF, Cronn RC, Alvarez I, Liu B, Small RL, Senchina DS. Intron size and genome size in plants. *Mol Biol Evol.* 2002;19:2346–2352.
7. Sakharkar MK, Chow VT, Kangueane P. Distributions of exons and introns in the human genome. *In Silico Biol.* 2004;4:387–393.
8. Roy SW, Penny D. Intron length distributions and gene prediction. *Nucleic Acids Res.* 2007;35:4737–4742.
9. Naora H, Deacon NJ. Relationship between the total size of exon and introns in the protein-coding genes of higher eukaryotes. *Proc Natl Acad Sci U S A.* 1982;79:6196–6200.
10. Hawkins JD. A survey on intron and exon lengths. *Nucleic Acids Res.* 1988;16:9893–9908.
11. Kriventseva EV, Gelfand MS. Statistical analysis of the exon–intron structure of higher and lower eukaryote genes. *J Biomol Struct Dyn.* 1999;17:281–288.
12. Ivashchenko AT, Atambayeva SA. Variation in lengths of introns and exons in genes of the Arabidopsis thaliana nuclear genome. *Russ J Genet.* 2004;40:1179–1181.
13. Atambayeva SA, Khailenko VA, Ivashchenko AT. Intron and exon length variation in arabidopsis, rice, nematode, and human. *Mol Biol.* 2008;42:312–320.
14. Ivashchenko AT, Khailenko VA, Atambayeva SA. Variation of the lengths of exons and introns in human genome genes. *Russ J Genet.* 2009;45:16–22.
15. Ivashchenko AT, Tauasarova MI, Atambayeva SA. Exon–intron structure of genes in complete fungal genomes. *Mol Biol.* 2009;43:24–31.
16. Zhu LC, Zhang Y, Zhang W, Yang SH, Chen JQ, Tian DC. Patterns of exon–intron architecture variation of genes in eukaryotic genomes. *BMC Genomics.* 2009;10:12.
17. Collins FS, Lander ES, Rogers J, Waterston RH. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431:931–945.
18. Schwarz EM, Antoshechkin I, Bastiani C, et al. WormBase: Better software, richer content. *Nucleic Acids Res*. 2006;34 Database Issue: D475–D478.

19. Drysdale RA, Crosby MA, FlyBase C. FlyBase: Genes and gene models. *Nucleic Acids Res.* 2005;33:D390–D395.
20. Haas BJ, Wortman JR, Ronning CM, et al. Complete reannotation of the Arabidopsis genome: Methods, tools, protocols and the final release. *BMC Biol.* 2005;3:7.
21. Logsdon JMJ, Stoltzfus A, Doolittle WF. Molecular evolution: Recent cases of spliceosomal intron gain? *Curr Biol.* 1998;8:R560–R563.
22. Archibald JM, O'Kelly CJ, Doolittle WF. The chaperonin genes of jakobid and jakobid-like flagellates: Implications for eukaryotic evolution. *Mol Biol Evol.* 2002;19:422–431.
23. Loftus BJ, Fung E, Roncaglia P, et al. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science.* 2005;307:1321–1324.
24. Martinez D, Berka RM, Henrissat B, et al. Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nat Biotechnol.* 2008;26:553–560.
25. Gudlaugsdottir S, Boswell DR, Wood GR, Ma J. Exon size distribution and the origin of introns. *Genetica.* 2007;131:299–306.
26. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987;4:406–425.
27. Ren XY, Vorst O, Fiers MWEJ, Stiekema WJ, Nap JP. In plants, highly expressed genes are the least compact. *Trends Genet.* 2006;22:528–532.

Supplementary tables

Table S1 Exonic chromosomal parameters of *Plasmodium knowlesi*

Chromosome	n_{ex}	l_{ex}	a_{ex}	nl_{ex}	LI_{ex}	$L0_{ex}$	p_c
PK01	2.18 ± 0.32	2513 ± 373	1871 ± 311	3.79 ± 0.57	2326 ± 644	2651 ± 442	42.29
PK02	2.61 ± 0.39	2234 ± 384	1476 ± 340	3.88 ± 0.58	1991 ± 438	2541 ± 671	55.84
PK03	2.74 ± 0.36	2313 ± 346	1455 ± 252	4.16 ± 0.51	2216 ± 506	2432 ± 445	55.10
PK04	2.70 ± 0.38	2159 ± 291	1436 ± 263	4.08 ± 0.56	1893 ± 359	2487 ± 468	55.17
PK05	2.75 ± 0.38	2048 ± 304	1387 ± 277	4.30 ± 0.62	1773 ± 363	2356 ± 530	52.90
PK06	2.67 ± 0.38	1996 ± 284	1308 ± 205	4.33 ± 0.62	1972 ± 452	2020 ± 345	50.00
PK07	2.65 ± 0.28	2093 ± 218	1396 ± 189	4.07 ± 0.41	1974 ± 184	2232 ± 340	53.85
PK08	2.47 ± 0.23	2062 ± 216	1451 ± 193	3.71 ± 0.34	1746 ± 263	2436 ± 346	54.18
PK09	2.69 ± 0.24	2226 ± 203	1492 ± 175	4.15 ± 0.36	2016 ± 274	2469 ± 308	53.64
PK10	2.43 ± 0.24	2114 ± 305	1431 ± 205	3.84 ± 0.36	2109 ± 551	2119 ± 351	50.32
PK11	2.70 ± 0.26	2244 ± 330	1538 ± 203	4.44 ± 0.41	2025 ± 288	2459 ± 345	49.48
PK12	2.59 ± 0.20	2213 ± 187	1483 ± 145	4.17 ± 0.33	2119 ± 281	2308 ± 248	50.29
PK13	2.63 ± 0.29	2235 ± 225	1527 ± 186	4.30 ± 0.49	2071 ± 323	2395 ± 311	49.56
PK14	2.42 ± 0.19	2195 ± 166	1542 ± 155	4.01 ± 0.32	2062 ± 238	2315 ± 256	47.36
Total	2.59 ± 0.07	2185 ± 66	1487 ± 55	4.11 ± 0.11	2019 ± 93	2358 ± 95	51.43

Table S2 Total variance explained

Component	% of variance	Cumulative %
1	61.413	61.413
2	24.809	86.222
3	8.290	94.512
4	4.871	99.383

Table S3 Component matrix extraction method: principal component analysis with four extracted components

Component matrix (a)						
Abbr	**Kingdom/supergroup**	**Organism**	**Component**			
			1	**2**	**3**	**4**
AD	Plantae	*Arabidopsis thaliana*	–0.928	–0.314	0.069	0.172
AF	Fungi	*Aspergillus fumigatus*	–0.566	0.785	–0.065	0.236
BN	Protista/Rhizaria	*Bigelowiella natans*	–0.601	–0.545	–0.210	0.544
CE	Animalia	*Caenorhabditis elegans*	–0.929	–0.296	0.189	–0.103
CF	Animalia	*Canis familiaris*	–0.956	–0.240	–0.025	–0.119
CG	Fungi	*Candida glabrata*	0.891	–0.314	–0.242	–0.221
CN	Fungi	*Cryptococcus neoformans*	–0.986	0.040	–0.123	–0.103
DH	Fungi	*Debaryomyces hansenii*	0.978	–0.204	0.046	–0.003
DM	Animalia	*Drosophila melanogaster*	–0.747	0.478	0.345	–0.298
EC	Fungi	*Encephalitozoon cuniculi*	0.582	–0.805	–0.082	0.085
EG	Fungi	*Eremothecium gossypii*	0.946	–0.250	–0.137	–0.152
GT	Protista/Chromista	*Guillardia theta*	0.417	–0.841	0.139	0.317
GZ	Fungi	*Gibberella zeae*	–0.555	0.769	–0.276	0.096
HA	Protista/Chromista	*Hemiselmis anderenii*	0.615	–0.704	–0.198	–0.064
HS	Animalia	*Homo sapiens*	–0.967	–0.229	0.050	–0.091
KL	Fungi	*Kluyveromyces lactis*	0.914	–0.346	–0.160	–0.139
LB	Protista/Protozoa	*Leishmania braziliensis*	0.677	0.628	–0.380	0.020
MG	Fungi	*Magnaporthe grisea*	–0.776	–0.492	0.237	0.314
MM	Animalia	*Mus musculus*	–0.965	–0.255	0.037	–0.038
MS	Plantae/Viridiplantae	*Micromonas* sp. RCC299	0.794	0.472	0.363	0.121
NC	Fungi	*Neurospora crassa*	–0.186	0.899	–0.225	0.287
OL	Plantae/Viridiplantae	*Ostreococcus lucimarinus*	0.737	–0.030	0.542	0.400
OS	Plantae	*Oryza sativa*	–0.927	–0.282	0.181	0.149
PF	Protista/Chromalveolata	*Plasmodium falciparum*	0.629	0.685	–0.327	–0.168
PK	Protista/Chromalveolata	*Plasmodium knowlesi*	0.646	0.632	–0.396	–0.152
PS	Fungi	*Pichia stipitis*	0.752	0.468	0.439	0.136
PT	Protista/Chromalveolata	*Paramecium tetraurelia*	–0.706	0.642	0.179	–0.239
SC	Fungi	*Saccharomyces cerevisiaei*	–0.969	0.119	0.215	0.026
SP	Fungi	*Schizosaccharomyces pombe*	0.890	0.063	–0.234	0.381
TA	Protista/Chromalveolata	*Theileria annulata*	–0.221	0.342	–0.815	0.391
UM	Fungi	*Ustilago maydis*	0.852	0.448	–0.253	–0.097
YL	Fungi	*Yarrowia lipolytica*	0.862	0.296	0.390	0.130

Table S4 Results obtained by *k*-means clustering technique and based on four principal components obtained by factor analysis of AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MG	1	2	2	2	6	7	7	8	5	8	1	6	7	2	11	8	13	12	7
AD	1	1	3	5	4	3	2	9	3	11	6	2	6	11	16	9	11	11	8
OS	1	1	3	5	4	3	2	9	3	11	6	2	6	11	16	9	11	11	8
HS	1	1	3	5	4	3	2	9	3	11	9	8	4	10	8	15	16	1	5
CF	1	1	3	5	4	3	2	9	3	11	9	8	4	10	8	15	16	1	5
MM	1	1	3	5	4	3	2	9	3	11	9	8	4	10	8	15	16	1	5
CE	1	1	3	5	4	3	2	9	3	11	8	8	5	10	8	15	16	1	14
CN	1	1	3	5	4	3	2	9	3	11	11	8	11	10	15	16	8	10	19
AF	1	2	2	2	5	4	3	2	8	5	5	10	1	3	14	11	1	3	4
GZ	1	2	2	2	5	4	3	2	8	5	5	10	1	3	14	11	1	3	4
DM	1	2	2	2	6	7	7	8	5	8	1	6	13	2	4	3	13	8	12
NC	1	2	2	2	5	4	3	2	8	5	5	10	1	3	14	11	1	3	4
PT	1	2	2	2	6	7	7	8	5	8	1	6	13	2	4	3	13	8	12
BN	1	1	3	5	3	3	2	7	7	7	4	13	8	9	9	13	5	9	9
TA	1	2	2	2	5	4	3	5	9	3	3	9	2	12	3	6	14	17	17
CG	2	3	1	3	2	6	8	1	2	2	10	12	10	5	6	1	6	16	15
EG	2	3	1	3	2	6	8	1	2	2	10	12	10	5	6	1	6	16	15
KL	2	3	1	3	2	6	8	1	2	2	10	12	10	5	6	1	6	16	15
DH	2	3	1	3	2	6	8	1	2	2	10	12	10	4	6	2	18	5	18
EC	2	3	1	3	2	5	4	6	10	9	2	5	3	14	2	12	7	14	10
GT	2	3	1	3	2	5	4	6	10	10	2	5	3	14	2	17	4	14	6
HA	2	3	1	3	2	5	4	6	10	9	2	5	3	14	2	12	7	14	10
LB	2	3	4	4	1	1	1	3	1	1	12	4	14	7	10	4	3	2	3
MS	2	3	4	4	1	1	1	3	1	1	12	4	14	15	10	4	15	18	20
PF	2	3	4	1	1	2	6	4	4	4	7	3	12	1	5	10	10	15	2
PK	2	3	4	1	1	2	6	4	4	4	7	3	12	1	7	10	10	15	2
UM	2	3	4	1	1	2	6	4	4	4	7	3	12	1	12	10	2	15	13
PS	2	3	4	4	1	1	1	3	1	1	12	4	14	15	10	4	15	19	20
SC	2	3	4	4	1	1	5	3	6	1	12	7	9	6	10	7	17	6	16
YL	2	3	4	4	1	1	1	3	6	1	12	7	14	15	10	7	15	4	20
SP	2	3	4	1	1	6	5	1	2	6	10	1	9	8	1	5	12	7	1
OL	2	3	4	4	1	1	5	3	6	1	12	11	9	13	13	14	9	13	11

Table S5 Results obtained by partitioning around medoids clustering technique and based on four principal components obtained by factor analysis of AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MG	2	3	3	3	3	3	3	3	3	3	12	12	12	12	12	12	12	12	12
AD	2	2	2	2	2	2	2	2	2	2	2	2	2	2	16	16	16	16	16
OS	2	2	2	2	2	2	2	2	2	2	2	2	2	2	16	16	16	16	16
HS	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
CF	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
MM	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
CE	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
CN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	17	17	17	17
AF	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
GZ	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	20
NC	2	3	3	3	3	3	3	3	3	3	3	3	14	14	14	14	14	14	14
CG	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
EG	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
KL	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
DH	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	19	19
PF	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
PK	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
UM	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	18	18	18
EC	1	1	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
GT	1	1	4	4	6	6	6	6	6	6	6	6	6	15	15	15	15	15	15
HA	1	1	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
LB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SC	1	1	1	1	1	1	1	1	1	1	1	13	13	13	13	13	13	13	13
YL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DM	2	3	3	3	3	7	8	8	8	8	8	8	8	8	8	8	8	8	8
PT	2	3	3	3	3	7	8	8	8	8	8	8	8	8	8	8	8	8	8
TA	2	3	3	3	3	3	7	9	9	9	9	9	9	9	9	9	9	9	9
BN	2	2	2	2	2	2	2	7	7	7	7	7	7	7	7	7	7	7	7
SP	1	1	4	4	4	4	4	4	10	10	10	10	10	10	10	10	10	10	10
OL	1	1	1	1	1	1	1	1	1	11	11	11	11	11	11	11	11	11	11