Intron Phase Patterns in Genes: Preservation and Evolutionary Changes

A. Ruvinsky*,[1] and C. Watson[2]

[1]*The Institute for Genetics and Bioinformatics, University of New England, Armidale NSW 2351, Australia*

[2]*School of Science and Technology, University of New England, Armidale NSW 2351, Australia*

Abstract: Introns are located either between codons (phase 0) or within codons (phase 1 and 2) and their phases as well as location usually stay unchanged for a long time. A string of intron phases represents a structure which may carry useful additional information about internal rearrangements of a gene. Combined search for intron phase patterns and exon lengths serves as a helpful approach for finding conserved intragenic duplications and other rearrangements. In vertebrate genes intragenic duplications usually are more numerous than in orthologs from other animal taxons. Intron phase patterns and exon lengths are highly conservative in some genes and can be traced back to a common ancestor of mammals and nematodes. Despite this, there are orthologs which show drastic losses of intron-exon structures as found in insects and urochordata. Driving forces behind such changes in exon-intron structures remain unknown and need further investigation.

Keywords: Exon, intron, intragenic duplications, evolution.

INTRODUCTION

It is a commonly held view that the majority of introns are ancient elements and their positions usually remain unchanged [1]. There are 3 phases, in which introns can be inserted: phase 0 (between codons) and phases 1 & 2 (after the first or second nucleotides of a codon). Intron sliding or shifts of intron–exon boundary over a few nucleotides leading to a change of intron phase are real but considered as rare events [2]. It means that many introns once being inserted in a certain position retain their phase for a long evolutionary time.

Changes in intron phase patterns indicate changes occurring in genes, which may or may not affect corresponding proteins. A comparison of intron phase pattern between distant species may reveal internal duplications, deletions, and other rearrangements which occurred during evolution of a particular gene. Measuring entropy of such strings allows discrimination between random and highly organised combinations of introns and exons in studied genes.

The idea that internal gene duplications played an important role in evolution of genes has a long history [3-5]. Fedorov *et al.* [6] estimated the proportion of duplicated exons in a set of 305 human genes as at least 6%. Marcotte *et al.* [7] came to a more or less similar conclusion that duplicated sequences occur in 14% of all proteins and about 3 times are more frequent in eukaryotes than in prokaryotes. Frequency of internal duplications is correlated with organismal complexity [8] and increased during metazoan evolution until the emergence of chordates [9].

Duplications which involve an exon and sections of surrounding introns or several exon-intron pairs, if they framed by introns in the same phase, do not affect reading frame as well as exon lengths. The occurring intron-exon patterns are essentially footprints of the past events and could be quite helpful in evolutionary reconstruction. For example, a string of intron phases, like 011212111111211211211112111 111, representing a structure of human *GTF2I* gene, coding for general transcription factor 2I, may provide useful information. This string is highly organised, has a low entropy value and indicates presence of intragenic repeats. Analysis of this gene and the corresponding protein confirms presence of several intragenic duplications and shed light on the evolution of the gene. Some genes are particularly prone to internal duplications and contain several series of repeats. Clearly such genes eventually became very lengthy and their evolutionary pathways could be affected by the duplication events.

As this paper shows, comparison of intron phase string patterns of orthologs from distant taxons may reveal significant changes. In insects, for instance, a considerable fraction of introns in some genes seem to be lost, while in other groups exon-intron structure of the genes might be preserved for a long time. This raises a question about evolutionary forces, which caused such changes in structure of orthologs.

METHODS

Gene Data

Information relevant to *Arabidopsis thaliana* (*At*), *Caenorhabditis elegans* (*Ce*), *Drosophila melanogaster* (*Dm*) and *Homo sapiens* (*Hs*) was extracted from the exon-intron database (EID, version 112), which was compiled in the W. Gilbert laboratory, Department of Molecular and Cellular Biology, Harvard University (Saxonov *et al.* 2000). The initial database was extensively purged by J-V. Chamary (University of Bath, UK). The removal of potential duplicates was done after performing an all-against-all BLAST, with an expected value of $P < 0.001$ [10], and creating clusters of duplicated genes. The longest of the duplicate genes were left in the database. This procedure was based on the assumption that, in the case of alternative transcripts, the

*Address correspondence to this author at the Institute for Genetics and Bioinformatics, University of New England, Armidale 2351 NSW, Australia; Tel: (61) 267 73 3900; Fax: (61) 267 73 3275;
E-mail: aruvinsk@une.edu.au

1874-4044/07

longest is the constitutive form. Even if this is not the case, it is just an arbitrary way of selecting one duplicate. Then one from the 'longest' duplicates, if several are of the same length, was randomly chosen. The total numbers of studied genes were: *Hs*-11,315, *Dm*–8,497, *Ce*-10,312 and *At*-9914.

Ensembl genome browser ortholog predictions were used for comparisons of genes from several distant species. Relevant details are described in Tables **3-5**.

Assessing Probability of Strings Using the Sliding Frame Approach

Entropy of a string measures a degree of randomness within this string. A string consisting of one element (number), like 00000, has the lowest possible entropy and a random string has the highest entropy.

We applied the sliding frame approach for measuring entropy of individual strings. The size of sliding frame may vary. Each frame size picks different information. Frame size 1 estimates probabilities of elements {0, 1 and 2} in a string. Sliding frame size 2 estimates probabilities of different pairwise combinations like 11, 20 or any other in a string. Lengthier frames pick more complex patterns like 121 or 0112. The probability function, $p(x_i)$, is defined as number of frames with pattern x_i divided by the total number of frames in a string. This definition of probability function is used for calculation of entropy, redundancy and Z statistic explained in the following sections.

A comparison of entropy between observed intron phase strings and simulated random Bernoulli schemes (explained below) using the same size of the sliding frame can be used to detect significant non-random patterns in observed strings. As for every length of intron phase strings comparisons are distinct, here we consider an example of human genes with 32 introns. In order to make reasonable statistical comparisons the strings should be long enough and a number of observed strings of the same length must be sufficient for meaningful comparisons.

Entropy

Shannon [11] defines entropy in terms of a discrete random variable X, with possible states (or outcomes) $x_1 \ldots x_n$. For three intron phases we have to use the following formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_3 p(x_i)$$

where $p(x_i)$ is the probability of the ith outcome of X.

Entropy for individual strings is calculated using the probability distribution produced by the sliding frame approach for each frame size. The frame slides across the string of intron phases one at a time making a total number of comparisons of (I–F+1), where I is the number of introns and F is the frame size. A program was written in C which is capable of calculating the entropy for given frame size.

Redundancy

We used information redundancy of intron phase patterns in order to compare such patterns in genes of different sizes and composition. Entropy of observed or simulated sequences increases with frame size, F and the length of sequences, in this case the number of introns, I.

For frame size F and I introns, we define:

Redundancy = minimum (F, $\log_3 I$) – $H(X)$

The optimal result is for frame size 4 or 5 because the observed entropy is compared with a random sequence with the same number of introns (I) and not the frame size (F) (proof is not given here). As the size of frames increases, observed entropy tends to the maximum value for a particular string length, I. Therefore the best discrimination of repetitiveness is achieved by using a frame size that is slightly larger than the largest $\log_3 (I_{max})$, as there are three possible intron phases. I_{max} is the maximal number of introns in the string.

Bernoulli Schemes

A Bernoulli scheme was used to generate random intron phase strings, which later were compared with the observed strings. The Bernoulli scheme is defined here as a stationary stochastic process with three possible outcome {0, 1 and 2}. The Bernoulli data were simulated by a newly written C program using the rand() function applying high order bits of the returned function to generate a random sequence from {0,1,2}. The simulations were run multiple times (N>100) and this resulted in numerous random strings of a particular length. Representative distribution of such randomly generated set of strings should be at least the size of the observed set of intron strings. This and other relevant programs can be provided on request. Generation of each element of a string was made using the genome wide intron phase frequencies [12]. The end result of these simulation processes was creation of a dataset of random strings with different lengths.

This Bernoulli scheme can be approximated by a trinomial distribution for sufficiently long strings which becomes more accurate in lengthy strings and also if p_0 does not take extreme values, where p_0 is a probability of random outcome as opposed to observed outcome (p). Thus we tried to test the H_0 whether there is no bias in the observed distribution of strings compared to the Bernoulli scheme. In this case the Bernoulli scheme can be approximated by a normal distribution enabling us to make inferences about outliers using the Z statistic. In such situation, a normal curve Z-test uses the formula given by:

$$Z = \frac{|p - p_0| - 1/2n}{\sqrt{p_0(1-p_0)/n}} \quad [13]$$

In the above formula p_0 is probability of the null hypothesis (H_0), p is the observed probability and n is the length of strings in a particular set. Z is compared with a standard normal distribution. Thus comparison of Bernoulli distributions of intron strings with the observed distributions provides a simple approach for finding possible biases from randomness in the observed strings.

RESULTS

Comparisons of Entropy Between Observed Intron Strings and Simulated in Bernoulli Schemes

Fig. (**1**) demonstrates entropy distributions for the observed intron phase strings and simulated Bernoulli schemes,

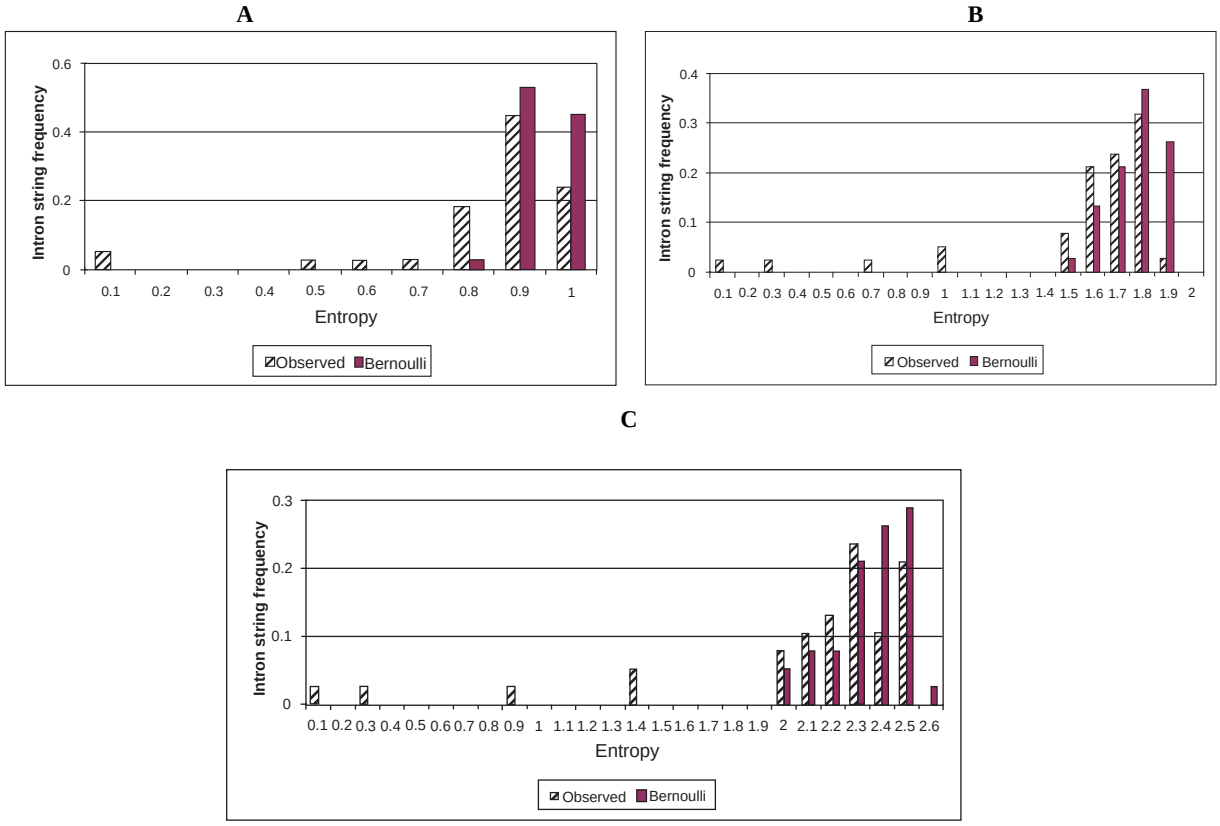

Fig. (1). Entropy distribution in observed *H. sapiens* intron phase strings with 32 introns and Bernoulli schemes were calculated for the strings of the same length. Sliding frame size: **A** (F1), **B** (F2), **C** (F3).

with length 32. This choice is dictated in part by the case study described below in section "*Intron phase patterns are helpful in phylogenetic reconstructions: human GTF2I gene*". Three frame sizes were used (1, 2 and 3). The Bernoulli schemes are significantly biased toward the maximum entropy value regardless of the size of the sliding frame, which is expected for randomly created strings. The observed distributions also have the bias but in a lesser degree than Bernoulli schemes. Typical feature of the observed distributions is a lengthy tail, which contains outliers with low entropy values located beyond several standard deviations from the median value of Bernoulli distribution. The presence of outliers is common for all observed string distributions regardless of the frame size. The genes which are represented by such outliers obviously have non-random intron phase patterns.

The Outliers are More Common in Human Genes

Table **1** shows the frequencies and numbers of outliers located beyond several increasing thresholds of *Z* distribution. The expected frequencies for random distribution (Bernoulli schemes) are also shown. In *Hs* the numbers of intron strings, which are outliers, exceed expectation in each category and are highly significant. The χ^2 comparisons of the observed and expected outliers for 2 thresholds of Z distribution ($P<0.01$ and $P<0.001$) presented at (Fig. **2**) strongly indicate that the differences are not random. Similar results

Table 1. Fraction and number of outliers observed among intron phase strings in four model species. Several threshold Z values cut 0.01, 0.001 and smaller sections of the expected normal distribution.

Species	Fraction (number) of outliers for several threshold Z values			
	Z>2.58	Z>3.29	Z>4.8	Z>6
Hs	0.032 (361)	0.011 (126)	0.001 (11)	0.0001 (7)
Dm	0.007 (58)	0.002 (18)	0.0002 (2)	0
Ce	0.019 (196)	0.006 (59)	0.001 (11)	0.0003 (3)
At	0.021 (207)	0.006 (59)	0.0008 (8)	0.0004 (4)
Normal expectation	0.01	0.001	1.00E-06	1.00E-09

were obtained for *Ce* and *At.* The number of outliers in *Dm* was lower. Thus, the data presented in Table **1** and (Figs. **1** & **2**) lead to a conclusion that in four studied model species there are many intron phase strings, and the corresponding genes, whose intron phase patterns can not be explained by random events alone.

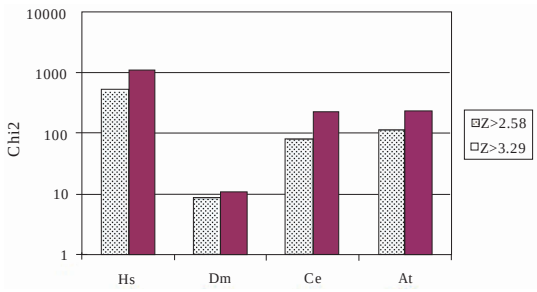

Fig. (2). Chi 2 comparisons of observed and expected outliers for 2 thresholds of Z distribution. $Z_{2.58}$ cuts 0.01 and $Z_{3.29}$ cuts 0.001 of the distribution. Y axis is in logarithmic scale.

In *Hs*, and mammals in general, such low entropy intron strings are particularly common. The frequency of outliers which are beyond $Z_{2.58}$ threshold (0.01 of the normal distribution) is 320% higher than expected and it is getting even higher for more strict *Z* thresholds. Similar trend can be seen in *At* and *Ce*, while in *Dm* the differences are more subtle. It can be assumed that numerous internal repeats, which increase redundancy and decrease entropy, are typical features of the low entropy intron strings. Thus, calculation of entropy values for individual intron phase strings provides a simple approach for preliminary identification of genes whose structure is significantly different from random. It is likely that at least some such genes evolved by internal duplications. Exon length is an additional criterion allowing identification of duplications within genes. As there is a significant variation in the length of exons, presence of several exons of the same length framed by introns in the same phase in a gene is a very unlikely result of random events. The relevant data using this approach are presented in section *Human genes with numerous internal repeats*.

Correlation Between the Number of Introns Per Gene and the Length of Coding Sequence

Fig. (**3A**) shows relations between length of coding sequence (total of all protein coding exons, CDS) and the number of introns per gene in human genes. There is an expected increase in the length of CDS as the number of introns and exons per gene is getting larger. Correlation between the lengths of CDS and the number of introns per human gene is high (r = 0.83). There is also lower variation in CDS lengths in genes with larger number of introns and vice versa. To present this observation in a more comparable way, the coefficient of variation (CV=STDEV/Mean) was calculated. As follows from (Fig. **3B**) there is a decline in CV of CDS lengths in human genes with higher number of introns. A similar trend exists in other species (data are not shown). This essentially means that correlation between the number of introns per gene and CDS length is getting stronger as number of introns increases.

A possible interpretation of this fact is that intragenic duplications are more frequent in the genes with numerous introns and, because exons are also parts of the duplications, the length of coding sequence stronger correlates with introns number. In the genes with few introns intragenic repeats are rare or absent. As the result, correlation between number of introns and CDS length is low and conversely variation of CDS length is high. Recently Chen *et al.* [9] came to a comparable conclusion studying repeats in proteins.

Human Genes with Numerous Internal Duplications

Human genes which have 7 or more exons of the same length framed by introns in the same phase are presented in Table **2**. Intron strings for each gene are shown and introns adjacent to repeated exons are printed in bold and underlined. Exon lengths for repeated exons in each gene are underlined; all other exons are presented by "x". Redundancy was calculated using the approach described in Methods.

While it seems very unlikely that numerous unrelated exons of the same size framed by the same phase introns could be found within a gene, the final conclusion can be

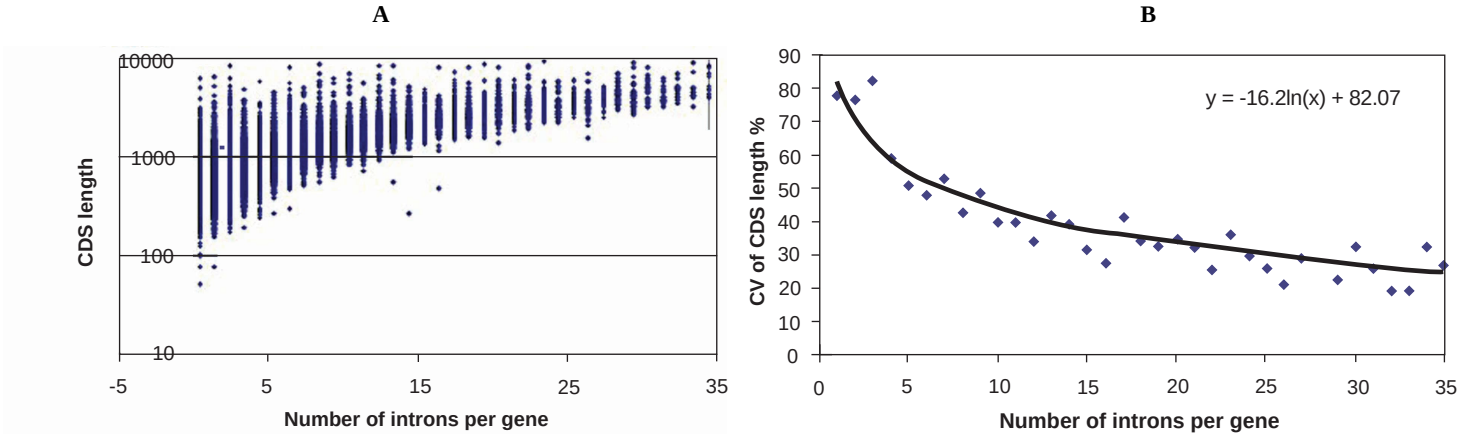

Fig. (3). Relationship between the number of introns per gene and length of coding sequences (**A**); and coefficient of variation (STDEV/MEAN) of CDS length in human genes (**B**). Y axis is on logarithmic scale in (**A**). Length of CDS is limited by 35 introns per gene as genes with higher number of introns are rare. Logarithmic trend line is added to (**B**) distribution.

Table 2. Intron strings, exon lengths and redundancy in human genes which have 7 or more repeated exons of the same length framed by introns in the same phase

Genes	Itron strings	Exon lengths	Redundancy
TARSH	0211210111**11**1**1111**1**111**11111111121111	x,x,x,x,x,x,x,x,x,x,75,75,x,75,75,75,x,75,75,75,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x	1.465
LGR4	0**222**2**2**2**22**2222**22**2221 1	x,72,72,72,x,72,x,72,72,x,72,x,72,72,x,x,x,x	1.282
SELP	011**1111111**1**1**112 u	x,x,x,x,186,186,186,186,186,186,186,x,186,x,x,x,x	1.066
KTN1	u1110002010**0200200200200200200**00000000001u	x,x,x,x,x,x,x,x,x,x,69,38,91,69,38,82,69,35,91,69,38,91,69,38,91,69,38,91,69,38,91,x,x,x,x,x,x,x,x,x,x,x,x	1.010
LGR7	u11**2222**2**22222**2111	x,x,x,x,72,72,72,72,75,72,72,72,72,72,x,x,x,x	0.637
KIDINS220	u0**000**0**00000**1112100101210000**0**0	x,x,99,99,99,x,99,198,99,99,99,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x,99,x,x	0.534
SLIT1	**222222**21**2222**21122**22**122**222**1021220212 1	x,72,72,72,72,72,72,x,x,72,72,72,144,x,x,x,x,144,144,x,x,x,72,72,72,x,x,x,x,x,x,x,x,x,x,x,x	0.493
RNH1	uu2**2222222**2	x,x,x,171,171,171,171,171,171,171,171,x	0.454
CARD4	uuu012**2222222**2	x,x,x,x,x,x,84,84,84,84,84,84,84,x	0.372
CARD15	101**2222222**2	x,x,x,x,84,84,84,84,84,84,84,x	0.342
FLJ14712	100010**11111111**10	x,x,x,x,x,x,x,x,96,96,96,96,96,96,96,x,x	0.375

*x-denotes exons which have lengths different from a core repeat, except those with double lengths; "u" represents untranslated exons; bold and underlined numbers indicate introns framing exons of the same size and these exons. Redundancy was calculated using the approach explained in Methods.

drawn from the sequence comparison of such repeated exons. The data relevant to the genes presented in Table **2** can be found online in Supplementary materials. All repeated exons for each gene in Table **2** were translated into amino acid sequences and aligned using CLUSTAL W in order to compare their sequence similarity. Those positions in the alignments, which were either strictly conservative or shown by CLUSTAL W as homologous, are highlighted in these alignments. Also there are many other positions which show a degree of similarity. In general sequence data strongly indicate that all these repeated exons are copies of ancestral exons, which undergone several intragenic duplications.

Intron Phase Patterns are Helpful in Phylogenetic Reconstructions: Human *GTF2I* Gene

As the majority of introns are indeed ancient elements of eukaryotic genes and their phase usually remained unchanged, a pattern of intron phases could be useful for studying some aspects of gene evolution. This may include finding intragenic duplications as long as they involve at least one intron and neighbouring exon. Other rearrangements such as possible intron losses and gains and more complex events also have to be taken into consideration. Certainly direct comparative analysis of DNA or amino acid sequences allow similar investigations but accumulated mutational changes may significantly complicate identification of duplicated exons and other rearrangements.

Here, as an illustration of such approach, we present analysis of human gene *GTF2I* coding for general transcription factor 2I. This gene has low entropy values; sliding frame size one, F1 is 0.6 (max entropy for F1 is 1), F2 is 1.02 (max is 2) and F3 is 1.43 (max is 3). *GTF2I* is highly significant outlier (Z = 3.834) among the genes with 32 introns. Exon-intron structure of this gene is shown at (Fig. **4**). According to the model which we used, this gene contains 33 protein coding exons and 32 introns. Observation of exon-intron structure of this gene suggests presence of intragenic duplications, which are framed by introns in phase **1** (shown in bold). Highly conservative exons located in the middle of these repeats are framed by the introns in phase *1* and *2* (shown in italic). Sequence alignment of the conservative exons confirms significant degree of sequence similarity and hence the origin from a common ancestral sequence (Supplementary materials, Fig. **1S**). All these 6 repeated exons (their positions within the gene are 4, 14, 17, 20, 25 and 29) have exactly the same length, there is no sequence gap in any of them and there are many conservative positions. The level of sequence identity varies from 66% to ~40% in 184 nucleotides.

The total number of duplication events is likely to be 5. Amino acid sequences corresponding to the compared exons also have conservative structure with the level of identity varying from 66.7 to 38.3 % in 60 amino acids. These sequences belong to a highly conserved domain (pfam 02946.12.), which is found in numerous proteins and is possibly a DNA binding domain [14]. As these conservative exons are framed by introns in phase 1 at the 5' end and in phase 2 at the 3' end, it is clear that these duplications must involve neighbouring exons. Thus, the repeated unit includes as a core conservative GTF2I exon, two surrounding introns

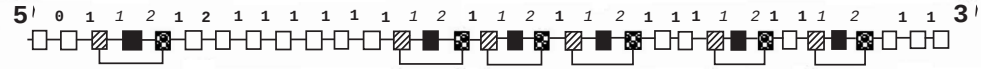

Fig. (4). Schematic presentation of exon-intron structure of human gene coding for general transcription factor 2I, *GTF2I*. Numbers indicate intron phase, bold numbers show introns involved in intragenic repeats. Rectangulars represent exons and lines represent introns. The brackets below the figure show the areas of repeats, which include highly conservative exons (black rectangulars), as well preceding (striped rectangulars) and the following exons (rectangulars with balls) and introns between them. Introns in phase **1** (bold) indicate putative ends of these 6 repeats. Conservative exons within the repeats are framed by introns in phase *1* and *2* (italic). Numeration of exons (not shown) starts from the first protein coding exon at the 5' end and finishes with the last protein coding exon at the 3' end.

and the neighbouring exons. While there is no significant variation in the core exons, the immediately neighbouring exons vary much more extensively particularly in areas remote from GTF2I exons. A general conclusion, which can be drawn from study of *GTF2I* structure, supports numerous intragenic duplication events as the essential steps in evolution of this gene.

Alignments of exon-intron structures of orthologous genes can reveal diverse evolutionary changes. Comparisons of *GTG2I* orthologs from several vertebrate species (Table **3**) show locations of duplicated regions including core conservative exon of 184 nucleotides. This also allows reasonable suggestions about the order in which duplications took place. There are 6 GTF2I units in the gene. Repeats number 3 and 4 in *Xenopus tropicalis*, *Gallus gallus* and *Homo sapiens*, if counting is done from the 5' end of the gene, are most likely the youngest and originated from repeat number 2. These repeats include 3 exons of similar sizes and framed by introns in phase 1. Both compared fish species *Danio rerio* and *Oryzias latipes* (Table **3**) do not have these repeats which could indicate that the duplications were generated in the common ancestor of Tetrapods.

Comparisons of exon-intron structures of genes from different species also shed some light on intron insertions and losses. Intron insertions are probably the cause of the steadily increasing number of exons in non-fish species, between the first and the second GTF2I repeats. Both fish species have only one relatively lengthy exon following the first GTF2I repeat, while in frogs there are 5 exons, in birds 6 and in mammals 7, all of which are rather short. A comparison of large number of species could verify these hypothetical intron insertions. Intron loss, on the contrary, is a plausible explanation for existence in both compared fish species 268 nucleotides exon (Table **3**). The corresponding position of the gene in other compared vertebrate species contain two exons of 68 and 184 nucleotides, total of which is equal to 268. Taking into consideration that the 184 nucleotide exon is an ancient element of *GTF2I* gene and surrounding introns are in the same phases, more parsimonious assumption is loss of the intron located between exons of 68 and 184 nucleotides in the common ancestor of zebrafish and medaka. An alternative explanation based on insertion of phase 1 intron in higher vertebrates is less likely.

Finally, comparisons of exon-intron structures also show shifts of reading frames. For instance, shifting of exon-intron boundary can be observed in *Xenopus tropicalis* 33 nucleotides exon (Table **3**, underlined exon). It differs from the corresponding exons in other species by 4 extra nucleotides, such addition must shift phase of the following intron from 1 to 2. This expectation is matched by the observation.

Comparison of Exon-Intron Structures of Orthologous Genes from Distant Species

Constantly growing number of studied genes from distant taxons allows comparisons of exon-intron structures of orthologs. Here we used Ensembl genome browser ortholog predictions for comparing sets of two genes (*SLIT1* and *KIDINS220*) which have significant degree of redundancy and because of that were included in Table **2**. Exon-intron structures of orthologs from eight species belonging to vertebrates, urochordates, insects and nematodes were aligned according to similarity of intron strings, exon lengths and sequence similarity.

As expected, there is a high degree of likeness between orthologs of *SLIT1* gene in vertebrate species (Table **4**). All compared genes are highly enriched by phase 2 introns, which seem to be typical for the common ancestor. Exon-introns structure of *SLIT1* from the sea squirt *Ciona intestinalis*, which belongs to urochordata and because of this is closer relative to vertebrates than other compared species, has many common features with the vertebrate orthologs. Insect genes contain twice less introns, which confirm well known fact that this group is intron poor. The nematode ortholog is different in this regard and has nearly as many introns as vertebrate orthologs. Its intron-exon structure is more similar to vertebrate genes despite longer evolutionary distance. Generally more similarities can be observed at the 5' end of the compared orthologs. The 3' end on the contrary shows less similarity. This may point out that more intron insertions and deletions as well other rearrangements took place in the 3' section of these genes since the common ancestor of the compared species.

Table 3. Exon lengths and intron phases for gene *GTF2I* and Ensembl genome browser ortholog predictions in several vertebrate species[#]. Exon lengths are shown above and intron phases below. The alignment of exons is done taking into account sequence similarity. Exons with the same lengths and position framed by introns in the same phase are shown in bold

Species*	Length in aa	Number of exons	Number of GTF2I repeats	Lengths of exons (above) and phase of introns (below)
Dr	919	26	4	126,146,140, **184**,225, 96,84,**184**,26,130,23,81,37,23,51, 66,**184**,62, 39, ***268***,29,78,193,38,116,66 0 2 **1** **2** 1 1 **1** **2** 1 2 1 1 2 1 1 **1** **2** 1 1 **2** 1 1 2 1 0
Ol	958	27	4	126,142, 63,174,**184**,269, 90,84,**184**,26,124,98,30,36,55,43, 40,**184**,62, 78, ***268***,29,78,193,38,110,69 0 1 1 **1** **2** 1 1 **1** **2** 1 2 1 1 1 2 0 **1** **2** 1 1 **2** 1 1 2 1 0
Xt	930	31	6	96,139,123, **184**, 29,43,44,78,54, 96,66,**184**,56,72,**184**,59,72,**184**,59,63, 93,66,**184**,53, 78,84,**184**,33, 5,48,80 0 1 **1** **2** 1 2 1 1 1 1 **1** **2** 1 **1** **2** 1 **1** **2** 1 1 1 **1** **2** 1 1 **1** **2** 2 1 1
Gg	984	32	6	106,139,135, **184**, 29,49,44,78,60, 63,102,66,**184**,59,72,**184**,59,72,**184**,59,84, 96,66,**184**,53,111,84,**184**,29,69,15,76 0 1 **1** **2** 1 2 1 1 1 1 **1** **2** 1 **1** **2** 1 **1** **2** 1 1 1 **1** **2** 1 1 **1** **2** 1 1 1
Hs	998	33	6	104,139,135, **184**, 29,55,44,78,60,57,63,111,66,**184**,59,72,**184**,59,72,**184**,59,75,102,66,**184**,56, 81,84,**184**,29,42,42,76 0 1 **1** **2** 1 2 1 1 1 1 1 1 **1** **2** 1 **1** **2** 1 **1** **2** 1 1 1 **1** **2** 1 1 **1** **2** 1 1 1

#Two fish species (zebrafish and medaka) were not predicted as orthologs by Ensembl genome browser.
**Danio rerio* - zebrafish; *Oryzias latipes* – medaka; *Xenopus tropicalis* - frog; *Gallus gallus* - chicken, *Homo sapiens*.
Exons with length *268* are italicized and exon 33 is underlined. Explanations are in the text.

Table 4. Exon lengths and intron phases for human gene *SLIT1* and Ensembl genome browser ortholog predictions in several animal species. Exon lengths are shown above and intron phases below. The alignment of exons is done taking into account sequence similarity. Exons with the same lengths framed by introns in the same phase are highlighted.

Sp*	Lengths of exons (above) and phase of introns (below)	Number of introns	Protein length
Hs	449,72,72,72,72,72, 72, 164,148,72,72,72,144,164, 24,145,75, 144,144,167,133,69,72,72,72,164,125, 98,140, 94,138,238,131,155, 289,212,3313 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 2 2 2 2 2 2 1 0 2 1 2 2 0 2 1 2 1	36	1534
Gg	218,72,72,72,72,72, 72, 164,148,72,72,72,144,164, 24,145,75,46, 101,144,167,133,69,72,72,72,164,125, 98,140, 94,138,238,131,89,69,289,212,415 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2 0 2 2 1 2 2 2 2 2 1 0 2 1 2 2 0 2 1 1 2 1	37	1546
Dr	71,72,72,72,72, 72, 164,148,72,72,72,144,164, 24,145,75, 144,144,167,133,69,72,72,72,164,125, 98,140, 94,138,229,131,155, 289,212,239 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 1 0 2 1 2 2 0 2 1 2 2 1	36	1465
Ci	146, 164,139,72,72,72, 72,183,141,147,72,359,145,144, 72,164,117,100,138,111,147,117,154,162, 208,207,102,135,135,137 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 2 0 0 1 1 1 1 1	30	1401
Ag	72,72,72, 144, 236, 142, 144,216,158,154,219,72,369, 72, 144,1519,126,22,143,87 2 2 2 2 1 2 2 1 2 2 2 2 2 2 0 0 1 0	19	1393
Aa	71,72,72, 144, 236, 142, 144,216,158,154,219,72,369, 216,1519,135,10,146,87 2 2 2 2 1 2 2 1 2 2 2 2 2 2 0 0 1 0	18	1393
Dm	565,72,72, 144, 236, 142, 144,216,158,154,219,72,369, 72, 144,1525,160, 143,795 2 2 2 2 1 2 2 1 2 2 2 2 2 2 0 1 0	18	1480
Ce	146,72,72,72,99,108,145,161, 72,72,380,97,201, 72,72, 149,208,97,119, 156,325, 134, 94,126,179,125,219,108,102,113, 65,365 2 2 2 2 2 2 0 2 2 2 1 2 2 2 2 1 2 0 2 2 0 2 0 0 2 1 1 1 1 0 2	31	1410

**Species: Homo sapiens, Gallus gallus, Danio rerio, Ciona intestinalis, Anopheles gambiae, Aedes aegypti, Drosophila melanogaster,* Caenorhabditis elegans.

Table 5. Exon lengths and intron phases for gene *KIDINS220* and Ensembl genome browser ortholog predictions. Exon lengths are shown above and intron phases below. The alignment of exons is done taking into account sequence similarity. Exons with the same lengths framed by introns in the same phase are highlighted.

Species	Lengths of exons (above) and phase of introns (below)	Number of introns	Protein length
Homo sapiens	121,144, 99, 99, 99,102,99,198, 99,99, 99,178,165,180,166, 152,290,141, 244, 89,145,163,179,224,114,57,132,99,237,3053 u 0 0 0 0 0 0 0 0 0 0 0 1 1 1 2 1 0 0 1 0 1 2 1 0 0 0 0 0 0	29	1777
Gallus gallus	64,148, 99, 99, 99,102,99,198, 99,99, 99,178,165,180,166, 152,290,141, 244, 89,145,163,179,227,108,147, 99,249,249 u 0 0 0 0 0 0 0 0 0 0 0 1 1 1 2 1 0 0 1 0 1 2 1 0 0 0 0 0	28	1424
Danio rerio	138,154, 99, 99, 99,102,99,198, 99,99, 99,178,165,180,166, 152,290,141, 238, 89,145,164, 39,169,221,138, 99,231,1899 u 0 0 0 0 0 0 0 0 0 0 0 1 1 1 2 1 0 0 1 0 1 0 0 1 0 0 0 u	28	1690
*Ciona intestinalis**	2072,784 2	1	952
Anopheles gambiae	189,132,399,1420, 978, 166,259,570,159 0 0 0 1 1 2 0 0	8	1424
Aedes aegypti	129,1819, 136,839, 166,955,393 0 1 2 1 2 0	6	1459
Drosophila melanogaster	156,135,198, 2765,179,236,171,225,475,267,450 0 0 0 2 1 0 0 0 1 1	10	1626
Caenorhabditis elegans	114, 66,102,174,117, 197,100,99, 166,180,139,202,107,146,311, 238,206, 561,490,121,156,57,47,101 0 0 0 0 0 2 0 0 1 1 2 0 2 1 0 1 0 0 1 2 2 2 1	23	1398

*This gene in *Ciona intestinalis* might not be the ortholog despite the Ensembl genome browser prediction.

Another studied gene, *KIDINS220*, demonstrates significantly less evolutionary stability of the exon-intron structures (Table **5**). Orthologs from insect species have about a quarter of the introns and exons typical for the vertebrate species. However, the protein lengths are not too different and some insect exons are exceptionally lengthy. These facts probably indicate intron losses in insects. The alternative explanation, an acquisition of introns by the gene in the vertebrate species, is not ruled out but looks less likely. For instance it contradicts the presence of numerous intragenenic exonic repeats which, were they unique for the vertebrate orthologs, should substantially elongate coding and hence protein sequences. This is not the case. Again, as in *SLIT1* gene, the nematode ortholog has somewhat more comparable intron-exon structure to the vertebrate genes but the similarity is more limited. The ortholog from *Ciona intestinalis*, which is expected to have alike structure to the vertebrate genes, apparently lost all introns except one and its exon-intron structure is not recognisable. The protein determined by this *Ciona intestinalis* gene is significantly shorter and this cast some doubts on its orthology. Such drastic changes in the exon-intron structure of the gene and the protein are not proportionally reflected in changes found in the primary sequence. NJ tree of the proteins coded by the *KIDINS220* orthologs supports close relations between the orthologs from *Ciona intestinalis* and the vertebrate species (Fig. **5**). Observations made in both *SLIT1* and *KIDINS220* genes do not contradict each other but rather show some differences in evolutionary pathways of the genes.

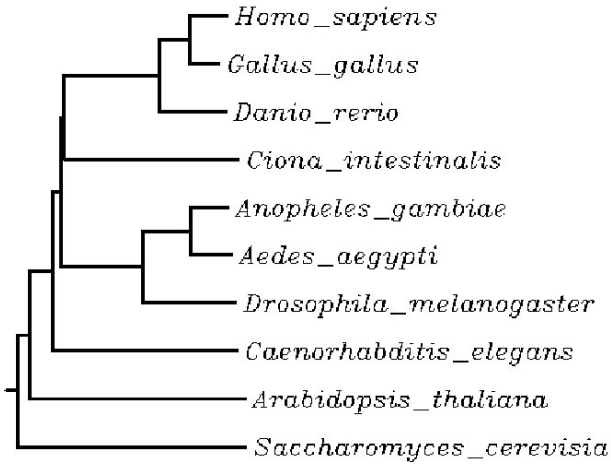

Fig. (5). NJ tree of KIDINS220 proteins from different animal species (some sequences are Ensembl genome browser ortholog predictions). Branching order of the tree does not contradict taxonomic views and show steadily increasing phylogenetic distances for more distant species.

DISCUSSION

It has been recognized for some time that intragenic duplications played a role in evolution of genes [3-5]. Study of exon-intron structures allowed estimating the proportion of duplicated exons in a set of human genes as at least 6% [6]. A similar conclusion was drawn for proteins, where duplicated sequences occur in 14% of studied eukaryotic sequences and is about 3 times more frequent than in Prokaryotes [7]. The data presented in this paper show that intragenic repeats are more frequent than expected by random occurrence (Fig. **2**); thus supporting the view that internal duplications were a common feature of evolution in some gene families. In human genes high correlation between lengths of CDS and number of introns was observed (r = 0.83). As duplication process multiplies the number of both exons and introns this could lead to the high correlation. A smaller number of outliers among intron strings typical for *Drosophila melanogaster*, might be caused by the intron losses which followed duplication process and thus masked duplication events.

There is also a correlation between frequency of internal repeats and organismal complexity [8]. As recently shown intragenic duplications were quite common during metazoan evolution until the emergence of chordates [9]. Perhaps the duplication process did not stop entirely after the emergence of chordates. As our data show in the genes with high redundancy there is a slight increase in number of repeats from fish to higher vertebrates which provide some support for the above statement (Table **3**, Supplementary materials Table **1S**).

Symmetric exons or a group of neighbouring exons framed by introns in the same phase are preferable for duplication process [15]. If the breaks occur in the surrounding introns, which have the same phase, this does not shift the reading frame and might not cause negative consequences. Several consecutive duplications create highly repetitive intron strings which can be detected by measuring their entropy. A combined search for exons with the same length and framed by introns in the same phase is the next step, which identifies intragenic duplications. Finally such intragenic duplications involving a single exon-intron pair or more complex grouping can be confirmed by the alignments of DNA and protein sequences. Long genes resulted from numerous internal duplications are relatively rare, but could be important if their proteins became “hubs” of proteome interactions [16]

The intragenic duplications can, at least in some degree, explain intron as well as exon creation process. Studies of protein domain families extend current understanding of the process and show distinct duplication patterns. Tandem repeats of certain domains can be observed in many proteins [17]. A model of gene formation based on essential role of introns in the duplication process was recently suggested [18]. Similar observation relevant to MHC-linked *tenascin-X* gene was made earlier by Hughes [19]. As the data, which we report here, show intragenic duplications were used extensively during evolution of some genes.

A comparison of repeated exon - intron units between orthologs also reveal some intron losses. Gene *SLIT1* is a good example. Beside the core 72 nucleotide repeated exons, there are also three 144 nucleotide repeats, which are most likely resulted from intron losses (see also Supplementary materials). In some cases considered in this paper, intragenic repeats have a tandem structure, which might be a product of unequal recombination. In other situations intragenic repeats are dispersed. The basic point, however, remains unchanged, intragenic repeats regardless of their lengths or positions have to be framed by introns in the same phase. This is an essential condition for successful unequal recombination; otherwise shift of reading frame is inevitable.

The question of why the basic features of intron phase patterns of orthologs from some species remained conserved, while genes from other species went through significant changes, needs clarification. Perhaps transcription itself or posttranscriptional events might be affected by the presence or absence of introns and, if so, could create certain selective pressures. Another question which still begs an answer is why intron insertions and losses seem to be more permissible during early stages of metazoan evolution. In other words the question is why such rearrangements became less common after separation of major groups.

CONCLUSION

Information carried by exon-intron structures was not widely used so far, while it could be helpful in resolving different questions relevant to gene evolution. Measuring entropy of intron strings reveals genes with numerous repeats. Comparative analysis of such genes from different species combined with a search for the same length exons framed by introns in the same phase identifies duplications, intron insertions and losses as well as shifts of exon-intron borders. It also provides additional information about timing of the rearrangements in exon-intron structures, which can not be obtained from investigations of coding sequences or proteins. This paper gives examples of usefulness of the approach. For instance, at least in some vertebrate genes intragenic duplications are more numerous than in orthologs from other animal taxons. Also intron phase patterns and exon lengths are highly conservative in some genes and can be traced back to a common ancestor of mammals and nema-

todes. In other cases, there are orthologs which show drastic losses of intron-exon structures in metazoan groups such as insects. Hopefully this type of analysis may help to understand causes for conservation and drastic changes in exon-intron structures as well time of such events.

ACKNOWLEDGEMENT

The authors are grateful to W. Ward, J-V. Chamary, A. Fedorov, L.D. Hurst and V. Kanevsky for help or advice.

REFERENCES

[1] Roy, S.W.; Gilbert, W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci., USA*, **2005**, *102*: 5773–5778.

[2] Rogozin, I.B.; Lyons-Weiler, J.; Koonin, E.V. Intron sliding in conserved gene families. *Trends Genet.*, **2000**, *16*: 430-432.

[3] Jacob, F. Molecular *tinkering in evolution*., Cambridge University Press: Cambridge, **1983**.

[4] Li., W.-H. Evolution of duplicate genes and pseudogenes. *In evolution of genes and proteins*., Nei, M.; Koehn, R.K., Sunderland, MA.: Sinauer Associates Inc, **1983**; pp 14-37.

[5] Patthy, L. Intron-dependent evolution: preferred types of exons and introns. *FEBS Lett.*, **1987**, *214*, 1-7.

[6] Fedorov, A.; Fedorova, L.; Starchenko, V.; Filatov, V.; Grigor'ev, E. Influence of exon duplication on intron and exon phase distribution. *J. Mol. Evol.*, **1998**, *46*: 263-271.

[7] Marcotte, E.M.; Pellegrini, M.; Yeates, T.O.; Eisenberg D. A census of protein repeats. *J. Mol. Biol.*, **1999**, *293*: 151-160.

[8] Lavorgna, G.; Patthy, L.; Boncinelli, E. Were protein internal repeats formed by "bricolage"? *Trends Genet.*, **2001**, *17*: 120-123.

[9] Chen, C.-C.; Li, W.-H., Sung, H.-M. Patterns of internal gene duplication in the course of metazoan evolution. *Gene*, **2007**, *396*: p.59-65

[10] Lercher, M.J.; Chamary, J.V.; Hurst, L.D. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res*., **2004**, *14*: 1002-1013.

[11] Shannon, C.E. A mathematical theory of communication. *Bell System Technical Journal*, **1948**, *27*: 379-423.

[12] Nguyen, H.D. Yoshihama, M.; Kenmochi, N., Phase distribution of spliceosomal introns: implications for intron origin. *BMC. Evol. Biol*., **2006**, *6*: 69.

[13] Hogg, R.; Tanis, E. *Probability and Statistical Inference*. 7th ed.; NJ: Pearson Prentice Hall: Upper Saddle River, **2005**.

[14] Vullhorst, D.; Buonanno, A. Multiple GTF2I-like repeats of general transcription factor 3 exhibit DNA binding properties: evidence for a common origin as a sequence-specific DNA interaction module. *J. Biol. Chem*., **2005**, *280*: 31722-31731.

[15] Long, M.; de Souza, S.; Rosenberg, C.; Gilbert, W. Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*: 219-223.

[16] Dosztányi, Z.; Chen, J.; Dunker, A.K.; Simon, I.; Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res., **2006**, *5*: 2985-2995.

[17] Björklund, Å.K.; Ekman, D.; Elofsson, A. Expansion of protein domain repeats. *PLoS. Comput. Biol*., **2006**, *2*: 0959-0970.

[18] Street, T.O.; Rose, G.D.; Barrick, D. The role of introns in repeat protein gene formation. *J. Mol. Biol*., **2006**, *360*: 258-266.

[19] Hughes, A.L. Concerted evolution of exons and introns in the MHC-linked tenascin-X gene in mammals. *Mol. Biol. Evol*., **1999**, *16*: 1558-1567.

Received: October 17, 2007 Revised: November 02, 2007 Accepted: November 13, 2007

Supplementary Data for the Paper Intron Phase Patterns in Genes: Preservation and Evolutionary Changes

A. Ruvinsky and C. Watson

1. ADDITIONAL DATA AND ALIGNMENTS FOR GENES LISTED IN TABLE 2

Below is a list of human genes presented in Table **2**. The genes have at least 7 exons of the same size and these exons are framed by introns in the same phase. All such exons for any gene in the list were translated into polypeptides and aligned using CLUSTAL W in order to check their sequence similarity. Only those positions in the alignments, which were either conservative of picked up by CLUSTAL W as homologous are highlighted. There are many other positions which have some degree of similarity.

TARSH Gene

Intron (phase: 02112101111111111111111111111121111, size: 66824, 23615, 3723, 12438, 553, 8975, 932, 574, 7884, 1971, 2447, 10370, 1140, 540, 1169, 1127, 1155, 38136, 3294, 8323, 1405, 2154, 3174, 9244, 1755, 2948, 533, 3652, 4793, 11149, 763, 870, 1112, 941);

Exon (size: 104, 173, 69, 133, 182, 53, 49, 72, 93, 78, **75, 75**, 66, **75, 75, 75**, 72, **75, 75, 75**, 60, 81, 78, 63, 78, 93, 129, 69, 210, 109, 80, 30, 162, 123, 1274)

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_15        APSETPFVPQKLEIFTSPEMQPTTP
_EXON_18        APGKTQFISLKPKIPLSPEVTHTKP
_EXON_20        VPKVPQRVTAKPKTSPSPEVSYTTP
_EXON_16        APQQTTSIPSTPKRRPRPKPPRTKP
_EXON_19        APKQTPRAPPKPKTSPRPRIPQTQP
_EXON_11        VTPETVPRSTKPTTSSALDVSETTL
_EXON_14        ATSDRILDSIPPKTSRTLEQPRATL
_EXON_12        ASSEKPWIVPTAKISEDSKVLQPQT
                ..                    .
```

Sequence similarity between exons which have length of 75 nucleotides is relatively low and might be gradually lost during evolution of the gene.

LGR4 Gene

Intron (phase: 022222222222222211, size: 59257, 20270, 1281, 5625, 845, 2079, 1221, 183, 300, 1502, 1473, 787, 2252, 298, 1152, 584, 2471);

Exon (size: 629, **72, 72, 72**, 216, **72**, 69, **72, 72**, 69, **72**, 66, **72, 72**, 126, 116, 84, 3181);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_6         LHLHNNKIRSLSQHCFDGLDNLET
_EXON_8         LGFHSNSISVIPDGAFDGNPLLRT
_EXON_13        ISLQRNQIYQIKEGTFQGLISLRI
_EXON_4         LTLQNNQLKTVPSEAIRGLSALQS
_EXON_9         IHLYDNPLSFVGNSAFHNLSDLHS
_EXON_3         LQLAGNDLSFIHPKALSGLKELKV
_EXON_11        LTLTGTKISSIPNNLCQEQKMLRT
_EXON_2         LDISMNNITQLPEDAFKNFPFLEE
_EXON_14        LDLSRNLIHEIHSRAFATLGPITN
                : :  . :  :          :
                1 3  6 8  11   16    22
```

This alignment supports sequence homology between exons which have length of 72 nucleotides. The numbers below this alignment show highly conservative and conservative positions within this repeat. The same positions are conservative in two other genes *SLIT1* and *LGR7*, which can be found in this document. BLAST analysis of a copy of 72 nucleotide exon demonstrates wide presence of homologous sequence (similar to leucine-rich repeat-containing G-protein) in many mammalian species as well as in some low vertebrates with high degree of sequence similarity (data not shown).

SELP Gene

Intron (phase: 011111111111112u, size: 10852, 1714, 3334, 471, 526, 539, 1802, 2369, 3737, 5849, 856, 1047, 967, 2148, 1223, 686);

Exon (size: 71, 91, 387, 108, **186, 186, 186, 186, 186, 186, 186**, 210, **186**, 120, 31, 55, 610);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_6     AAQCPPLKIPERGNMTCLHSAKAFQHQSSCSFSCEEGFALVGPEVVQCTASGVWTAPAPVCK
_EXON_8     AISCEPLESPVHGSMDCSPSLRAFQYDTNCSFRCAEGFMLRGADIVRCDNLGQWTAPAPVCQ
_EXON_11    AIKCPELFAPEQGSLDCSDTRGEFNVGSTCHFSCDNGFKLEGPNNVECTTSGRWSATPPTCK
_EXON_10    AIPCTPLLSPQNGTMTCVQPLGSSSYKSTCQFICDEGYSLSGPERLDCTRSGRWTDSPPMCE
_EXON_7     AVQCQHLEAPSEGTMDCVHPLTAFAYGSSCKFECQPGYRVRGLDMLRCIDSGHWSAPLPTCE
_EXON_9     ALQCQDLPVPNEARVNCSHPFGAFRYQSVCSFTCNEGLLLVGASVLQCLATGNWNSVPPECQ
_EXON_13    AVKCSELHVNKPIAMNCSNLWGNFSYGSICSFHCLEGQLLNGSAQTACQENGHWSTTVPTCQ
_EXON_5     VRECGELELPQHVLMNCSHPLGNFSFNSQCSFHCTDGYQVNGPSKLECLASGIWTNKPPQCL
            .  *  *      : *         : * * *  * : *    *   * *.   * *
```

This alignment supports sequence homology between exons which have length of 186 nucleotides.

KTN1 Gene

Intron (phase: u11100020100200200200200200200200000000001u, size:31664, 3945, 1309, 1047, 8599, 1928, 3137, 1184, 1790, 661, 414, 1337, 690, 370, 490, 156, 483, 5212, 1017, 686, 905, 514, 150, 1252, 922, 76, 473, 2445, 2413, 1051, 1818, 5628, 3398, 756, 129, 752, 227, 2822, 2433, 1120, 4413);

Exon (size: 138, 553, 138, 171, 131, 117, 141, 107, 133, 88, 167, **69**, 38, 91, **69**, 38, 82, **69**, 35, 91, **69**, 38, 91, **69**, 38, 91, **69**, 38, 91, **69**, 38, 91, 87, 90, 81, 90, 93, 72, 84, 84, 84, 93, 121, 478);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_12        DILEQNEALKAQIQQFHSQIAAQ
_EXON_27        DLQEENESLKAHVQEVAQHNLKE
_EXON_30        DVQDENKLFKSQIEQLKQQNYQQ
_EXON_24        DLKQEIKALKEEIGNVQLEKAQQ
_EXON_15        DIQNMNFLLKAEVQKLQALANEQ
_EXON_21        AIRTENSSLTKEVQDLKAKQNDQ
_EXON_18        ILNDQNKALKSEVQKLQTLVSEQ
                 :      :. .: ..      :
```

This alignment supports sequence homology between exons which have length of 69 nucleotides. There is also strong sequence similarity between exons with lengths of 38 and 91 nucleotides. It is likely that the repeat included these three exons and surrounding introns.

***LGR7* Gene**

Intron (phase: u1122222222222111, size: 57759, 20565, 5826, 5636, 2832, 4086, 89, 4768, 9614, 1746, 4695, 4531, 1180, 5577, 1652, 1297, 3039);

Exon (size: 101, 138, 99, 106, **72, 72, 72, 72**, 75, **72, 72, 72, 72, 72**, 230, 411, 219, 456,);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_5         MSLQWNLIRKLPPDCFKNYHDLQK
_EXON_12        LDLGSNKIENLPPLIFKDLKELSQ
_EXON_7         LYLSHNRITFLKPGVFEDLHRLEW
_EXON_11        LVMRKNKINHLNENTFAPLQKLDE
_EXON_13        LNLSYNPIQKIQANQFDYLVKLKS
_EXON_6         LYLQNNKITSISIYAFRGLNSLTK
_EXON_8         LIIEDNHLSRISPPTFYGLNSLIL
_EXON_10        LDLEGNHIHNLRNLTFISCSNLTV
_EXON_14        LSLEGIEISNIQQRMFRPLMNLSH
                : :    :  :    *     *
                1 3  6 8  11   16    22
```

This alignment supports sequence homology between exons which have length of 72 nucleotides. The numbers below this alignment show highly conservative and conservative positions within this repeat. The same positions are conservative in two other genes *SLIT1* and *LGR4*, which can be found in this document.

KIDINS220 Gene

Intron (phase: u00000000001112100101210000000, size: 10350, 8192, 978, 4282, 740, 6025, 3143, 2431, 2099, 1332, 2783, 2585, 1000, 1067, 2289, 175, 5926, 534, 168, 1746, 5915, 19025, 1130, 2111, 685, 10145, 2114, 974, 1461);

Exon (size: 121, 144, **99, 99, 99**, 102, **99**, **198**, **99, 99, 99**, 178, 165, 180, 166, 152, 290, 141, 244, 89, 145, 163, 179, 224, 114, 57, 132, **99**, 237, 3053);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_4          DNWTALISASKEGHVHIVEELLKCGVNLEHRDM
_Exon_8b         DGNTALMIASKEGHTEIVQDLLDAGTYVNIPDR
_EXON_7          YGTTPLVWAARKGHLECVKHLLAMGADVDQEGA
_EXON_9          SGDTVLIGAVRGGHVEIVRALLQKYADIDIRGQ
_EXON_3          CGQTPLMIAAEQGNLEIVKELIKNGANCNLEDL
_EXON_5          GGWTALMWACYKGRTDVVELLLSHGANPSVTGL
_EXON_11         DGETPLIKATKMRNIEVVELLLDKGAKVSAVDK
_EXON_8a         NSMTALIVAVKGGYTQSVKEILKRNPNVNLTDK
_EXON_10         DNKTALYWAVEKGNATMVRDILQCNPDTEICTK
                  . * *  *         *. ::      .
_EXON_28         ANINGRVLAQCNIDELKKEMNMNFGDWHLFRST
```

This alignment supports sequence homology between exons which have length of 99 nucleotides. The exon which has length 198 nucleotides most likely is the result of intron loss that joins two neighbouring 99 nucleotide repeats (8a and 8b). Exon 28 is less similar to other exons.

SLIT1 Gene

Intron (phase: 2222222122222112222122222210212202121, size: 20587, 1367, 5490, 91731, 1138, 565, 549, 2667, 481, 555, 2178, 745, 7202, 958, 138, 624, 193, 3140, 302, 2780, 2221, 3155, 2793, 777, 9343, 2142, 299, 4427, 184, 2610, 4272, 1659, 470, 1075, 334, 807);

Exon (size: 449, **72, 72, 72, 72, 72, 72**, 164, 148, **72,72, 72, 144**, 164, 24, 145, 75, **144, 144**, 167, 133, 69, **72, 72, 72**, 164, 125, 98, 140, 94, 138, 238, 131, 155, 289, 212, 3313);

CLUSTAL W (1.83) multiple sequence alignment

```
EXON_3           LQLMENQIGAVERGAFDDMKELER
EXON_6           LQLDKNQISCIEEGAFRALRGLEV
EXON_5           LDLSENAIQAIPRKAFRGATDLKN
Exon_18b         LHLTANQLESIRSGMFRGLDGLRT
EXON_4           LRLNRNQLHMLPELLFQNNQALSR
Exon_13a         LLLNANKINCIRPDAFQDLQNLSL
Exon_11          IDLSNNQIAEIAPDAFQGLRSLNS
Exon_18a         INLSNNKVSEIEDGAFEGAASVSE
Exon_24          LILSYNALQCIPPLAFQGLRSLRL
Exon_10          IRLELNGIKSIPPGAFSPYRKLRR
EXON_2           LELNGNNITRIHKNDFAGLKQLRV
Exon_19a         LMLRNNRISCIHNDSFTGLRNVRL
EXON_7           LTLNNNNITTIPVSSFNHMPKLRT
Exon_23          VDLSNNKISSLSNSSFTNMSQLTT
Exon_12          LVLYGNKITDLPRGVFGGLYTLQL
Exon_13b         LSLYDNKIQSLAKGTFTSLRAIQT
Exon_19b         LSLYDNQITTVSPGAFDTLQSLST
Exon_25          LSLHGNDISTLQEGIFADVTSLSH
                 : *  * :  :    *     :
                 1 3  6 8  11   16    22
```

This alignment supports sequence homology between exons which have length of 72 nucleotides. The exons which have length 144 nucleotides most likely are the result of intron loss that joins two neighbouring 72 nucleotide repeats (13a & 13b; 18a & 18b; 19a & 19b).

The numbers below this alignment show highly conservative and conservative positions within this repeat. The same positions are conservative in two other genes *LGR4* and *LGR7*, which can be found in this document.

***RNH1* Gene**

Intron (phase: uu22222222, size:2116, 2574, 1407, 484, 643, 81, 135, 315, 2917, 104);

Exon (size: 161, 173, 188, **171, 171, 171, 171, 171, 171, 171**, 266);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_3          RLDDCGLTEARCKDISSALRVNPALAELNLRSNELGDVGVHCVLQGLQTPSCKIQKL
_EXON_4          SLQNCCLTGAGCGVLSSTLRTLPTLQELHLSDNLLGDAGLQLLCEGLLDPQCRLEKL
_EXON_6          KLESCGVTSDNCRDLCGIVASKASLRELALGSNKLGDVGMAELCPGLLHPSSRLRTL
_EXON_7          WIWECGITAKGCGDLCRVLRAKESLKELSLAGNELGDEGARLLCETLLEPGCQLESL
_EXON_9          WLADCDVSDSSCSSLAATLLANHSLRELDLSNNCLGDAGILQLVESVRQPGCLLEQL
_EXON_5          QLEYCSLSAASCEPLASVLRAKPDFKELTVSNNDINEAGVRVLCQGLKDSPCQLEAL
_EXON_8          WVKSCSFTAACCSHFSSVLAQNRFLLELQISNNRLEDAGVRELCQGLGQPGSVLRVL
                  :  * .:   *  :.  :     : ** : .* : : *   :   :  . . :. *
```

This alignment supports sequence homology between exons which have length of 171 nucleotides.

CARD4 Gene

Intron (phase: uuu0122222222, size: 18564, 660, 2109, 1409, 2096, 2834, 1247, 742, 8484, 1491, 2818, 3638, 3663);

Exon (size: 87, 141, 89, 322, 175, 1825, **84, 84, 84, 84, 84, 84, 84**, 1184);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_11        KLGKNKITSEGGKYLALAVKNSKSISEV
_EXON_15        WLIQNQITAKGTAQLADALQSNTGITEI
_EXON_13        SLASNGISTEGGKSLARALQQNTSLEIL
_EXON_14        WLTQNELNDEVAESLAEMLKVNQTLKHL
_EXON_9         RLSVNQITDGGVKVLSEELTKYKIVTYL
_EXON_10        GLYNNQITDVGARYVTKILDECKGLTHL
_EXON_12        GMWGNQVGDEGAKAFAEALRNHPSLTTL
                 :  * :       .:  :     :  :
                 2  5 7      1314 17    23 26
```

This alignment supports sequence homology between all exons which have length of 84 nucleotides. The numbers below this alignment show highly conservative positions within this repeat. The same positions (except position 13) are conservative in gene *CARD15*, which can be found in this document.

CARD15 Gene

Intron (phase: 10122222222, size: 2171, 7900, 2597, 4213, 220, 2950, 2613, 595, 2104, 4245, 1845);

Exon (size: 178, 467, 106, 1816, **84, 84, 84, 84, 84, 84, 84**, 1331);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_9         SLVGNNIGSVGAQALALMLAKNVMLEEL
_EXON_11        KLSNNCITYLGAEALLQALERNDTILEV
_EXON_8         GFWGNRVGDEGAQALAEALGDHQSLRWL
_EXON_10        CLEENHLQDEGVCSLAEGLKKNSSLKIL
_EXON_7         RLGNNYITAAGAQVLAEGLRGNTSLQFL
_EXON_5         YLRDNNISDRGICKLIECALHCEQLQKL
_EXON_6         ALFNNKLTDGCAHSMAKLLACRQNFLAL
                 :  * :       :         :  :
                 2  5 7      14  17    23 26
```

This alignment supports sequence homology between exons which have length of 84 nucleotides.

The numbers below this alignment show highly conservative and conservative positions within this repeat. The same positions (except position 13) are conservative in gene *CARD4*, which can be found in this document.

FLJ14712 Gene

Intron(phase: 10001011111111110, size: 2051, 5835, 114, 1454, 2062, 934, 4397, 7103, 83, 1016, 928, 1641, 3049, 93, 751, 2572, 2267);

Exon (size: 848, 137, 156, 105, 115, 251, 220, 141, **96, 96, 96, 96, 96, 96, 96, 96**, 104, 313);

CLUSTAL W (1.83) multiple sequence alignment

```
_EXON_24        SICDPTCMNGGKCVGPSTCSCPSGWSGKRCN
_EXON_25        PICLQKCKNGGECIAPSICHCPSSWEGVRCQ
_EXON_19        ALCDPDCKNHGKCIKPNICQCLPGHGGATCD
_EXON_22        ALCDPVCLNGGSCNKPNTCLCPNGFFGEHCQ
_EXON_20        EHCNPPCQHGGTCLAGNLCTCPYGFVGPRCE
_EXON_23        AFCHPPCKNGGHCMRNNVCVCREGYTGRRFQ
_EXON_21        MVCNRHCENGGQCLTPDICQCKPGWYGPTCS
_EXON_18        TICKYPCGKSRECVAPNICKCKPGYIGSNCQ
                  *   * :   *   . * *   .  *   .
```

This alignment supports sequence homology between exons which have length of 96 nucleotides.

Table 1S. Number of exons per gene and lengths of proteins in orthologous genes from four species among the genes shown in Table 2, which have the highest redundancy in the humans*

Genes	*C.elegans*		*Danio rerio*		*Gallus gallus*		*Homo sapiens*	
	Protein	**Exons**	**Protein**	**Exons**	**Protein**	**Exons**	**Protein**	**Exons**
Tarsh	524	7(2)	785	21(2)	913	30(4)	1074	35(8)
LGR4	929	13(2)	907	15(6)	1038	19(9)	950	18(9)
SELP	575	12(5)	648	11(7)	610	14(6)	830	17(9)
KTN1	1133	12(2)	1236	35(3)	1368	43(7)	1357	45(7)
LGR7	un	un	539	12(6)	755	17(10)	740	18(10)
KIDINS220	1398	24(1)	1690	29(8)	1424	29(8)	1777	30(8)
SLIT1	1410	32(7)	790	22(9)	1546	39(12)	1534	37(12)
RNH1	un	un	un	un	456	10(7)	461	11(7)
CARD4	un	un	951	8(5)	951	11(7)	953	14(7)
CARD15	un	un	970	11(6)	759	10(3)	1040	12(7)
FLJ14712	un	un	837	21(7)	828	22(8)	849	18(8)

* Number in brackets shows the number of repeated exons underlined in Table 2.
un – ortholog is not known.

Fig. 1S. Multiple alignment of six conservative exons from the human *GTF2I* gene. All these exons have exactly the same length (184 nucleotides), there is no gap in any of them and there are many conservative positions (highlighted). All these conservative exons are preceded by phase 1 introns and followed by phase 2 introns.