SUPPLEMENTARY MATERIALS FOR

Comparison of splice sites in mammals and chicken

Josep F. Abril, Robert Castelo and Roderic Guigó[1]

Abstract

We have carried out an initial analysis of the dynamics of the recent evolution of the splice sites sequences on a large collection of human, rodent (mouse and rat), and chicken introns. Our results indicate that the sequences of splice sites are largely homogeneous within tetrapoda. We have also found that orthologous splice signals between human and rodents and within rodents are more conserved than unrelated splice sites, but the additional conservation can be explained mostly by background intron conservation. In contrast, additional conservation over background is detectable in orthologous mammalian and chicken splice sites. Our results also indicate that the U2 and U12 intron classes seem to have evolved independently since the split of mammals and birds; we have not been able to find a convincing case of interconversion between these two classes in our collections of orthologous introns. Similarly, we have not found a single case of switching between AT-AC and GT-AG subtypes within U12 introns, suggesting that this event has been a rare occurrence in recent evolutionary times. Switching between GT-AG and the non-canonical GC-AG U2 subtypes, on the contrary, does not appear to be unusual; in particular, T to C mutations appear to be relatively well tolerated in GT-AG introns with very strong donor sites.

Keywords: comparative genomics, human, mouse, rat, chicken, U2/U12 splice sites, splice sites evolution, non-canonical splicing, GT-AG, GC-AG, AT-AC, splice signals subtype switching, comparative pictogram

Contents

[1]To whom correspondence should be addressed: rguigo@imim.es

1 Materials and Methods

Complete data sets and further information will be available at:
`http://genome.imim.es/datasets/hmrg2004/`

1.1 Orthologous mammalian REFSEQ introns

Figure 1 illustrates the protocol initially devised to filter orthologous introns from human, mouse and rat orthologous REFSEQ genes. Only coding exons were taken into account from the very beginning. Those orthologous sets for which genes have not same exon number were discarded—for instance, due to misannotated exons as shown in Figure 1.A, or because of intron insertion/deletion as pinpointed in Figure 1.B. From the resulting set, those exonic structures whose exon phase was not conserved across each splice site, were also thrown away—as the one in the Figure 1.C. The final orthologous introns set was enriched with introns from coding regions and most of the UTR introns were lost. Having the same number of exons and phase conserved does not guarantees that all the real orthologous introns were found.

In order to obtain a set of confident orthologous introns, another approach was taken. The consecutive exons for each gene were aligned with `t_coffee` (Notredame et al., 2000), using default parameters, to all the consecutive exon pairs on the orthologous genes. We were able to recover a larger set of orthologous introns. In example, for the genes shown in Figure 1.A, three orthologous introns of a total of four introns were retrieved; with respect to the genes appearing in Figure 1.B, eight introns from a total of nine; and finally six extra introns were captured from the available seven on the set from Figure 1.C.

1.2 Comparison of splice site sequence patterns

Since comparative pictograms are highly interpretative we have attempted a more rigorous comparison. We assessed how many of the observed differences in the comparative pictograms were significant by calculating confidence intervals for the ratios of nucleotide proportions. The ratio of proportions, or *relative risk*, of nucleotide usage is the ratio $p_{n|1}/p_{n|2}$ where $p_{n|1}$ and $p_{n|2}$ are the empirical probabilities of nucleotide n estimated from organisms 1 and 2, respectively. In this analysis we assume that the DNA sites follow a Position Weight Matrix (PWM) model where nucleotides occur independently on each position along the splice site leading to one nucleotide probability distribution and four relative risks, for each of these independent positions. We considered donor sites with 3+GT+4 positions and acceptor sites with 18+AG+3 positions. In a comparative pictogram we represent the proportions of nucleotides observed at each position for two sets of aligned DNA sites (in this case splice sites). The proportions of the two species are drawn side by side at each position to facilitate their comparison and the background occupied by each nucleotide is colored according to the corresponding relative risk. The yellow color is assigned to the ratio of 1 (0 in log-scale) and it approaches red when the relative risk becomes large and approaches green when it becomes small, using some saturation value specified in a scale that is part of the graphical representation.

For each pair of organisms, the PWM model requires the calculation of $7 \times 4 = 28$ relative risks for donors and $21 \times 4 = 84$ for acceptors. These relative risks were calculated with their respective confidence intervals and therefore in order to achieve an overall confidence level of 95% for each species comparison, the significance level of the each confidence interval was adjusted accordingly (see second column on Table 1). Relative risks were calculated in $\log_2$-scale and their corresponding confidence intervals following Agresti (1990, pg. 56):

$$\log_2 \frac{p_{n|1}}{p_{n|2}} \pm z_{\alpha/2} \frac{1}{\log 2} \left[\left(c_{n|1} + \frac{1}{2} \right)^{-1} - \left(c_{+|1} + \frac{1}{2} \right)^{-1} + \left(c_{n|2} + \frac{1}{2} \right)^{-1} - \left(c_{+|2} + \frac{1}{2} \right)^{-1} \right]^{1/2}, \qquad (1)$$

where $c_{n|i}$ are the counts of occurrences of nucleotide n for organism i, and $c_{+|i} = \sum_n c_{n|i}$ is the total number of nucleotides. When the confidence interval in (1) did not include the value 0 (value 1 when exponentiating endpoints) then we considered the observed difference in nucleotide usage (the relative risk) to be significant. Table 1 shows the percentage of how many of the relative risks, within the corresponding comparison, represent significant differences, which is quite homogeneous within tetrapoda and substantially smaller than in fish and insects.

Besides the assessment of relative risks of nucleotide usage we wanted to further analyze the distributional differences of splice sites between species. A standard way of comparing probability distributions is by performing a Chi-squared goodness-of-fit test under the null hypothesis that the two tested samples belong to the same probability distribution. However, when we performed this test between the distributions of nucleotides determined by the PWM, the sample size has been large enough that differences in less than 1% in the use of nucleotides become significant, although such differences are not meaningful for us. More importantly, these discrepancies are difficult to compare due to the different splice-site sample sizes available for each organism. For this reason, we make the analogous question on whether the distributional characteristics of splice sites depend on the organism to where they belong to, i.e., the *site-species* dependence. Again, due to the large sample size all pairs of compared organisms show a significant dependence relationship but, in this case, we can quantify the degree of dependence that allow us to decide what pairs of organisms have a similar degree of discrepancy between their respective splice site sequence patterns.

To assess (in)dependence we have used the *likelihood-ratio* G^2-statistic which asymptotically follows a Chi-squared distribution (see for instance, Agresti, 1990, pg. 48):

$$G^2 = 2 \sum_i \sum_n c_{in} \log \frac{c_{in}}{(c_{i+} c_{+n})/c_{++}} \,, \tag{2}$$

where c_{in} are the counts of occurrences of nucleotide n with organism i, $c_{i+} = \sum_n c_{in}$, $c_{+n} = \sum_i c_{in}$ and $c_{++} = \sum_{in} c_{in}$ which is the number of nucleotides on both organisms $i \in \{1, 2\}$ for the corresponding distribution. The formulation in (2) is convenient for us because it can be decomposed in the product of a constant by an information entropy term describing the degree of dependence, as follows:

$$\begin{aligned} G^2 &= 2 \cdot \sum_i \sum_n c_{in} \log \frac{c_{in} c_{++}}{(c_{i+} c_{+n} c_{++})/c_{++}} \\ &= 2 \cdot \sum_i \sum_n c_{in} \log \frac{p_{in}}{p_{i+} p_{+n}} \\ &= 2 \cdot c_{++} \cdot \sum_i \sum_n p_{in} \log \frac{p_{in}}{p_{i+} p_{+n}} \\ &= 2 \cdot c_{++} \cdot \mathrm{D}(p_{in} || p_{i+} p_{+n}) \,, \end{aligned} \tag{3}$$

where p_{in} is the empirical probability of observing organism i and nucleotide n, p_{i+} is the marginal probability of organism i, p_{+n} is the marginal probability of nucleotide n and the term $\mathrm{D}(p_{in}||p_{i+}p_{+n})$ corresponds to the relative entropy (also known as Kullback-Leibler divergence, Kullback, 1951) of a site-species independence model $p_{i+}p_{+n}$ with respect to the site-species dependence model p_{in}. When presenting this quantity in Table 2 we used the logarithm in base 2 and hence this term can be interpreted as the number of bits required to reconstruct the distribution given by p_{in} from the one given by $p_{i+}p_{+n}$. This is a non-negative quantity that takes the 0 value when $p_{in} = p_{i+}p_{+n}$, i.e., when the distributional characteristics of splice sites are independent of the species being compared, and increases as the degree of dependence grows large. In column 1 within donor and acceptor sites in Table 2 we have the relative entropy of the site-species dependence and in column 2 the resulting G^2-statistic (here $\mathrm{D}(p_{in}||p_{i+}p_{+n})$ uses the natural logarithm to preserve the asymptotic equivalence to Chi-squared) which resulted significant in all cases for the corresponding degrees of freedom. Under the PWM model we obtain a G^2-statistic for each position. The independence of the positions allows one to obtain a relative entropy value for the splice site by adding up the relative entropies of every position along the signal. Analogously, the reproductive property of Chi-squared (see Agresti, 1990, pg. 43) permits obtaining a G^2-statistic for the splice site by adding up the values of each position and where the resulting degrees of freedom is the sum of the degrees of freedom of each position, i.e., 21 for donors and 63 for acceptors. As one can see, the degree of site-species dependence, measured by the relative entropy of the dependence model, remains relatively bounded for comparisons within tetrapoda but jumps one or two orders of magnitude for comparisons between tetrapoda and fish, and between tetrapoda and invertebrates.

Finally, we also computed the *site relative entropy* which corresponds to a distance measure between the splice site sequence pattern of two organisms. Under the PWM we calculated for each position the relative entropy of the distribution of nucleotides for organism 2 with respect to organism 1:

$$\mathrm{D}(p_1||p_2) = \sum_{n\in\{\mathrm{A,C,G,T}\}} p_{n|1} \log_2 \frac{p_{n|1}}{p_{n|2}}\,, \tag{4}$$

where, for any given position along the signal, $p_{n|1}$ and $p_{n|2}$ are the empirical probabilities of nucleotide n in organism 1 and 2, respectively. This value, using logarithm in base 2, can be interpreted as the number of bits required to build the nucleotide distribution of organism 1 from the distribution of organism 2. The relative entropies at each position were summed up to provide a single value for the site relative entropy of organism 2 with respect to organism 1. Since the relative entropy is not symmetric, we calculated $\mathrm{D}(p_2||p_1)$, and the value $\mathrm{D}(p_1||p_2)+\mathrm{D}(p_2||p_1)$ is the one reported in the third column of donors and acceptors in Table 2. As we see, this value correlates well with the phylogenetic distance of the compared organisms.

In summary, these results lend quantitative support to the observation using comparative pictograms that, even though subtle changes accumulate through evolution (as follows from the site relative entropies) splice site sequence patterns are largely homogeneous within tetrapoda, and noticeable distinct from those of other vertebrate and invertebrate taxa.

2 Supplementary Tables

splice site	signif. level	comparison					
		mouse-rat	human-mouse	human-rat	human-chicken	human-zebrafish	human-fruitfly
donor	$0.05/28$	32%	29%	54%	39%	93%	93%
acceptor	$0.05/84$	27%	56%	69%	44%	98%	94%

Table 1: Proportion of confidence intervals for the relative risk of nucleotide usage that exclude the ratio of 1. For each species comparison, and splice site, significance level was adjusted to achieve an overall level of $\alpha = 0.05$, i.e., a 95% confidence interval.

compared species	donor sites			acceptor sites		
	site-species dependence		site relative entropy ($\times 10^{-4}$)	site-species dependence		site relative entropy ($\times 10^{-4}$)
	rel. entropy ($\times 10^{-4}$)	G^2-stat.		rel. entropy ($\times 10^{-4}$)	G^2-stat.	
mouse-rat	4.93	121.13	54.80	23.81	584.86	264.54
human-mouse	9.68	453.54	80.82	23.99	1,124.05	199.86
human-rat	12.77	432.44	182.28	54.69	1,851.70	773.49
human-chicken	18.53	550.54	724.73	47.15	1,400.78	1,824.91
human-zebrafish	743.46	37,033.00	6,106.08	2,046.98	102,007.40	16,858.03
human-fruitfly	467.35	14,985.70	8,599.28	1,627.44	52,121.24	28,844.58

Table 2: Site-species dependence (relative entropy and G^2-statistic) at columns 1 and 2 within donors and acceptors measures how much the nucleotide composition of splice sites depends on the compared species: the relative entropy is the number of bits required to rebuild the dependence model from the independence model, and the G^2-statistic is calculated following Equation 3, it is asymptotically equivalent to Chi-squared and gives a significant p-value on the dependence relationship in all comparisons. Site relative entropy at column 3 within donors and acceptors measures the distance in bits between the nucleotide composition of splice sites of the two compared species.

3 Supplementary Figures

Figure 1: **Classifying REFSEQ orthologous genes.**

A) Missing exons: Annotation errors, alternative splicing, or pseudogenes were not taken into account. In the example three red arrows highlight human 3′UTRs that are matching coding exons in rodents, with significant TBLASTX hits (human *NM000897* / mouse *NM008521* / rat *NM053639*). This suggests the human gene is not well described.
B) Missing introns: Intron insertions or deletions were also discarded. In the figure, a red arrow points out the position in which the intron loss appears in the genic structure, while the red circle shows the same in the protein Exonic Structure Alignment (ESA; human *NM002198* / mouse *NM008390* / rat *NM012591*). As in the previous case, the valid pair-wise sequences were reassigned to the corresponding human-mouse, human-rat, or mouse-rat pair-wise sets.
C) Phases of coding exons not conserved: The red asterisk and the circle indicate that the mouse intron is out of phase with the human and rat ortholog (human *NM001333* / mouse *NM009984* / rat *NM013156*). These cases have also been discarded.
D) Exonic structure was conserved: Conservation of exonic structure does not necessarily mean exact conservation of exon lengths. In this case, the red ellipse pinpoints a region in which orthologous exon lengths have not been conserved. Intron positions along the alignment, and intron phases, however are still conserved. Thus, this case would have been retained (human *NM000139* / mouse *NM013516* / rat *NM012845*).
Colors in the ESA alignments correspond to consecutive exons. Genic structure figures were obtained using `gff2ps` (Abril and Guigó, 2000). All sequences are shown in forward strand and coordinates are relative to their first nucleotide. Black boxes represent the non coding regions of exons (UTRs). Coding exons were split in two halves, the left was color filled according to exon frame, while the right one shows the color of the remainder (the frame of the next downstream exon as a function of the current exon frame and its length). The frame color code is as follows: blue, red and green, for frames 0, 1, and 2 respectively. Introns are shown as red boxes (darker red for UTR introns and light red for introns within the CDS). Purple, orange, and brown boxes below each gene structure are projections of the WU-TBLASTX (W. Gish, unpublished, `http://blast.wustl.edu`) hits using genomic sequences from human, mouse and rat as target sequences, respectively. The height of those boxes is proportional to the score of the corresponding TBLASTX hit.

Figure 2: **Human, mouse, rat and chicken U12 orthologous intron sets.**

Ungapped alignments of the donor (-10 to +16 around the 5′ splice sites) and the acceptor (-30 to +10 around the 3′ splice sites) sequences for all the U12 orthologous intron sets were drawn using `TeXshade` (Beitz, 2000). Splice sites core signals are highlighted in a black box, the U12 conserved donor sequence (+3 to +8) is marked in green, sequence hits to the U12 branch point are colored in red, while conserved nucleotides at a given position are shown over a blue background.
Only the four species U12 orthologous introns were displayed. Further orthologous sets, including pair-wise and triads, are available at the supplementary materials web page (`http://genome.imim.es/datasets/hmrg2004/`).
It is worth to note that in the fourth example, the intron 16th of mouse gene *NM_007459* seems to be not conforming to the U12 donor pattern. But it is not a case of conversion between U2 and U12 splice sites, just displacing the splice sites two nucleotides upstream we recover the U12 donor pattern and the overall alignment of the exonic regions improves.

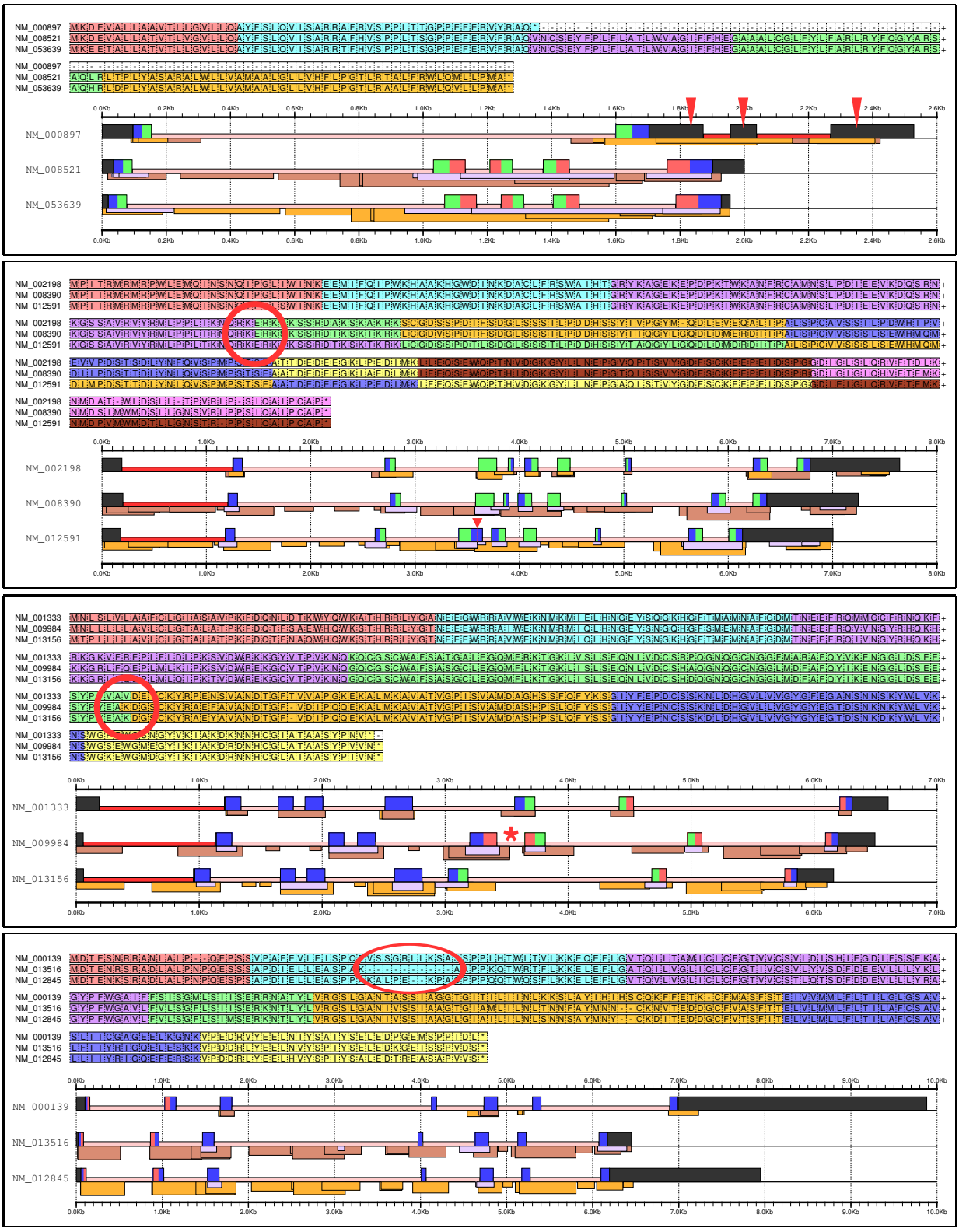
A
B
C
D
NM_000897
NM_008521
NM_053639
NM_002198
NM_008390
NM_012591
NM_001333
NM_009984
NM_013156
NM_000139
NM_013516
NM_012845

Figure 1

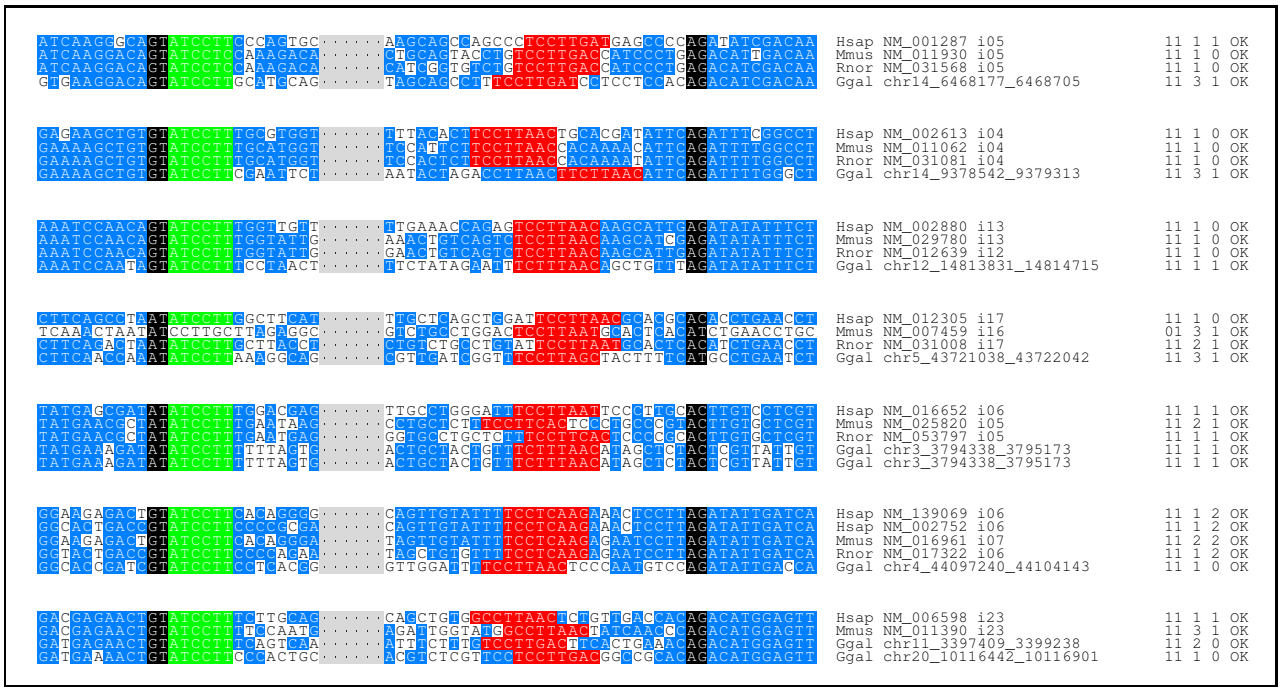
Hsap NM_001287 i05 11 1 1 OK
Mmus NM_011930 i05 11 1 0 OK
Rnor NM_031568 i05 11 1 0 OK
Ggal chr14_6468177_6468705 11 3 1 OK
Hsap NM_002613 i04 11 1 0 OK
Mmus NM_011062 i04 11 1 0 OK
Rnor NM_031081 i04 11 1 0 OK
Ggal chr14_9378542_9379313 11 3 1 OK
Hsap NM_002880 i13 11 1 0 OK
Mmus NM_029780 i13 11 1 0 OK
Rnor NM_012639 i12 11 1 0 OK
Ggal chr12_14813831_14814715 11 1 1 OK
Hsap NM_012305 i17 11 1 0 OK
Mmus NM_007459 i16 01 3 1 OK
Rnor NM_031008 i17 11 2 1 OK
Ggal chr5_43721038_43722042 11 3 1 OK
Hsap NM_016652 i06 11 1 1 OK
Mmus NM_025820 i05 11 2 1 OK
Rnor NM_053797 i05 11 1 1 OK
Ggal chr3_3794338_3795173 11 1 1 OK
Ggal chr3_3794338_3795173 11 1 1 OK
Hsap NM_139069 i06 11 1 2 OK
Hsap NM_002752 i06 11 1 2 OK
Mmus NM_016961 i07 11 2 2 OK
Rnor NM_017322 i06 11 1 2 OK
Ggal chr4_44097240_44104143 11 1 0 OK
Hsap NM_006598 i23 11 1 1 OK
Mmus NM_011390 i23 11 3 1 OK
Ggal chr11_3397409_3399238 11 2 0 OK
Ggal chr20_10116442_10116901 11 1 0 OK

Figure 2

References

Abril, J. F. and Guigó, R. (2000). `gff2ps`: Visualizing genomic annotations. *Bioinformatics*, 16(8):743–744.

Agresti, A. (1990). *Categorical data analysis*. Wiley-Interscience, New York.

Beitz, E. (2000). `TeXshade`: shading and labeling of multiple sequence alignments using `LaTeX2e`. *Bioinformatics*, 16:135–139.

Kullback, S. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). `T-Coffee`: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217.