EVOLUTION OF EXONIC AND INTRONIC REGIONS IN PRIMATES

Andrés Arturo Lanzós Camaioni
e- mail: andreslanzos@gmail.com

Final Degree Project
Tutors:
- Carlos Canchaya
- David Posada
Departamento de Bioquímica, Genética e Inmunología
Facultad de Biología
Universidad de Vigo.

SUMMARY

Technological advances in the past few decades have created multiple possibilities for comparative genomics. One of these is the phylogenetic study of whole genomes or phylogenomics. This thesis focuses on the evolutionary comparison between exons and introns of the X chromosome in primates. The results suggest that in the case of the X chromosome, introns are better suited for the study of the evolutionary history of the primates.

INTRODUCTION

Key advances in genomic technologies have resulted in a massive accumulation of biological information in recent years. This accumulation of data has prompted the raise of bioinformatics, which is the enforcement of the computational technology to the analysis and management of biological data. Among the many areas of application of bioinformatics, we can found the comparative study of the genetic material of multiple species. In this case, some of the resources and tools available are databases like Ensembl (Flicek *et al.,* 2012; *http://www.ensembl.org/index.html*) or those at the NCBI (*http://www.ncbi.nlm.nih.gov/*), programming languages such as Perl or Python, sequence alignment algorithms like BLAST (Altschul *et al.,* 1990) or EPO (Paten *et al.,* 2008a,b) and a lot of statistical models developed by researchers around the world and accessible throughout the world wide web. Indeed, genomics has largely benefited from bioinformatics during the last decade. For this reason, researchers have been able to work on genes sequences, protein domains, protein structures, chromosomes, transcripts and whole genomes (Teufel *et al.,* 2006; Margulies *et al.,* 2007).

The availability of multiple genome sequences allows to address many fundamental evolutionary questions on a genomic scale. One of these is the molecular evolution of exonic and intronic gene regions in eukaryotes. Also known as exons and introns, the second are the regions eliminated during the maturation of eukaryotic RNA in the process of RNA splicing (Figure 1).

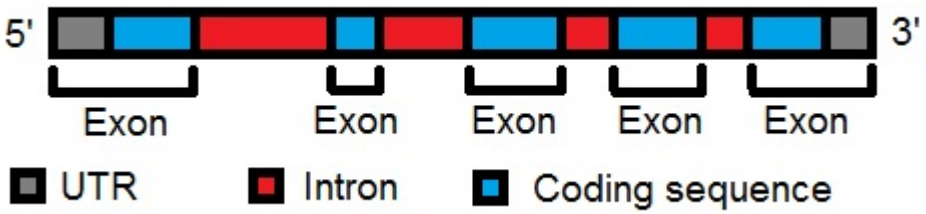

Figure 1. General scheme of eukaryotic pre-ARN with exons and introns.

Different researchers have showed that the structure and the evolution of exons and introns follow patterns that can be characterized (Rogozin *et al.,* 2005; Zhu *et al.,* 2009; Koralewski *et al.,* 2011). For example, the evolutionary rate of introns is bigger than exons (Gelfman et al, 2012), because exons contain coding sequences that are strongly conserved through evolution.

However, one of the questions that has not been addressed yet is whether exons and introns show similar nucleotide substitution patterns. Related to this, the phylogenetic utility of exons and introns, as far

as I know, have never been explicitly addressed. Here, I tried to delve into these two questions through the comparison of exons and introns of six species of primates: human (*Homo sapiens sapiens*), gorilla (*Gorilla gorilla*), chimpanzee (*Pan troglodytes*), orangutan (*Pongo abelii*), macaque (*Macaca mulatta*) and marmoset (*Callithrix jacchus*).

Objectives

The general aim of this thesis is the comparison of the molecular evolution of the exonic and intronic regions in primates. The particular objectives are:

1) To establish the average differences between exons and introns regarding genomic features like alignment length, GC content and nucleotide diversity.

2) To characterize nucleotide substitution patterns in exonic and intronic regions using a statistical model selection framework.

3) To compare the different phylogenetic hypothesis generated from exons and introns.

METHODS

Data Mining

The *Enredo, Pecan* and *Ortheus pipeline* (EPO; Paten *et al.,* 2008a,b) was used in order to gather genomic data from the six primate species of interest. *Enredo* produces collinear segments from genomes handling both rearrangements, deletions and duplications and *Pecan* is a global aligner used for these segments (Paten *et al.,* 2008a). Finally, *Ortheus* provides genome-wide ancestral sequence reconstructions (Paten *et al.,* 2008b). The EPO sequence alignments are available at the Ensembl Compara database of the European Bioinformatic Institute and the Wellcome Trust Sanger Institute (*http://www.ensembl.org/info/docs/api/compara/index.html*).

The '6-primates-EPO' (human, chimpanzee, gorilla, orangutan, macaque and marmoset) genomic alignments were programmatically accessed through the *Compara* API. Due to the enormous size of these data, the analysis was restricted to the chromosome X. For this, a Perl script was specifically designed to extract all the genomic blocks from the chromosome X of the EPO-6 data set. Note that these blocks can contain 2-6 species, as not all genomic blocks occur in all primates. In addition, different filters were applied to select only the common regions between the exons and introns within these blocks:

- Blocks with one or more sequences without annotated genes were eliminated.
- Blocks with more than one sequence per species (due to duplication) were eliminated.
- Aligned genic regions not present in all the species in the block were eliminated. Genes from different species do not always coincide exactly, so only the overlapping gene regions were selected (Figure 2).
- Within the overlapping genic regions, exons and introns were identified.
- Exonic regions common to the all species were again selected and saved to a SQLite database. The same procedure was applied to the intronic regions (Figure 3).

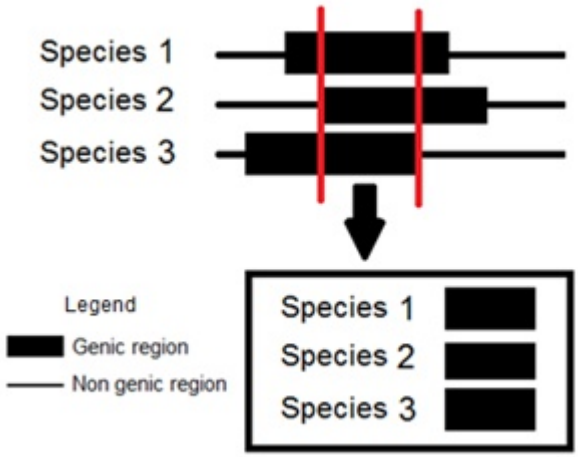

Figure 2. Selection of common regions for genes. Only the genic region common to the all sequences in the genomic align block were selected.

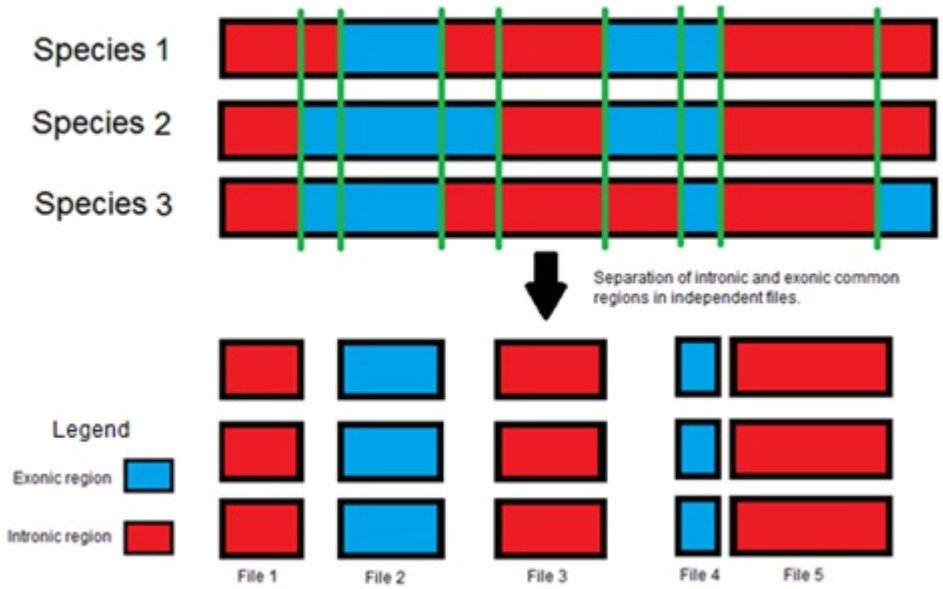

Figure 3. Selection of common regions for exons and introns.

Statistical analysis with R

The statistical comparison of exonic and intronic features was performed with the R statistical environment (R Development Core Team, 2013). Exonic and intronic alignments were exported from the SQLite database as FASTA files, and imported into R using the *seqinr* package (Charif, 2013). For each alignment, the nucleotide diversity, GC content and length were measured using *ape* (Paradis, 2013a). The nucleotide diversity is the average number of nucleotide differences per site between two randomly chosen DNA sequences (Nei *et al.,* 1979). Differences between exons and introns were assessed using linear regression and t-tests.

Model selection

Nucleotide substitution models are used to explain the probabilities of change among the different nucleotides according to the equilibrium nucleotide frequencies. The best-fit nucleotide substitution model (Posada *et al.,* 2001) was estimated for every alignment using the R packages *phangorn* (Schliep, 2013) and *pegas* (Paradis *et al.,* 2013b), according to the Bayesian Information Criterion (BIC; Schwarz, 1978). The best-fit models was selected among a set of 6 potential candidates with increasing complexity (Table 1).

Table 1. Substitution models examined

Substitution models	Number of parameters	Base frequencies	Transition rates
JC	1	A=C=G=T	AC=AG=AT=CG=CG=CT=GT
F81	4	A≠C≠G≠T	AC=AG=AT=CG=CG=CT=GT
K80	2	A=C=G=T	(AG=CT)≠(AC=AT=CG=GT)
HKY	6	A≠C≠G≠T	(AG=CT)≠(AC=AT=CG=GT)
SYM	6	A=C=G=T	AC≠AG≠AT≠CG≠CG≠CT≠GT
GTR	9	A≠C≠G≠T	AC≠AG≠AT≠CG≠CG≠CT≠GT

Phylogenetic analysis

Finally, maximum likelihood phylogenetic trees were estimated for every alignment, and their phylogenetic distance (number of different nodes) to the *'known'* primate tree (Figure 5) was calculated using the R package *phangorn*.

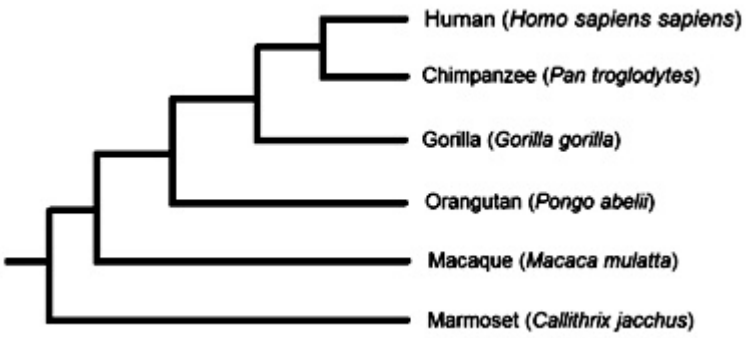

Figure 4. Putatively 'known' primate tree used in this thesis.

Training

In order to be able to handle the methodology just described, before anything I had to acquire computational skills that are routine in so-called "dry" computational biology and bioinformatic laboratories. This entitled:

- Shell environment: how to use the Linux operative system and its shell and how to connect to a remote computational node.

- Programming: I learned different aspects of the programming language Perl to be able to efficiently handle and analyze genomic data. I examined several online tutorials (http://es.tldp.org/Tutoriales/PERL/tutoperl-print.pdf) and books (Tidall, 2001, 2003).

- Data mining: to extract the genomic information from the Ensembl database, I also learned the specific application programming interface (API) of Ensemble. Generally speaking, an API specifies how some software components should interact with each other. The Ensembl API is a library that includes specifications for routines, data structures, object classes, and variables.

- Databases: In order to keep organized all the genomic information I learned how to create and use a SQLite database.

- Statistical analysis: I learned how to use R (R Development Core Team, 2013) to carry out automated statistical tests. I used different tutorials and online books (R Development Core Team, 2000; Paradis, 2003; Gentleman, 2008; Krijnen, 2009; Martinez, 2009).

RESULTS

A total of 1265 genomic blocks were downloaded from Ensembl. After filtering and dividing them into segments, the final number of alignments obtained were 7354, of which 2868 corresponded to exons and 4486 to introns. The average number of species present in the exon and intron alignments was very similar (5.76 and 5.78, respectively).

Genomic features

The average nucleotide diversity and alignment length was significantly lower (t-test p-value < 0.001) (Table 2 and Figure 6) for exons than for introns. On the contrary, the GC content for exons was significantly bigger than for introns (t-test p-value < 0.001) (Table 2 and Figure 7).

Table 2. Summary of genomic features for exons and introns.

	Min		Max		Median		Mean	
	Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
Nucleotide Diversity	0.0012	0.0012	0.2203	0.3059	0.0191	0.0490	0.0241	0.0499
Alignment length	103	103	3161	223700	159	1678	265	6462
GC content	0.256	0.178	0.823	0.853	0.489	0.407	0.496	0.428

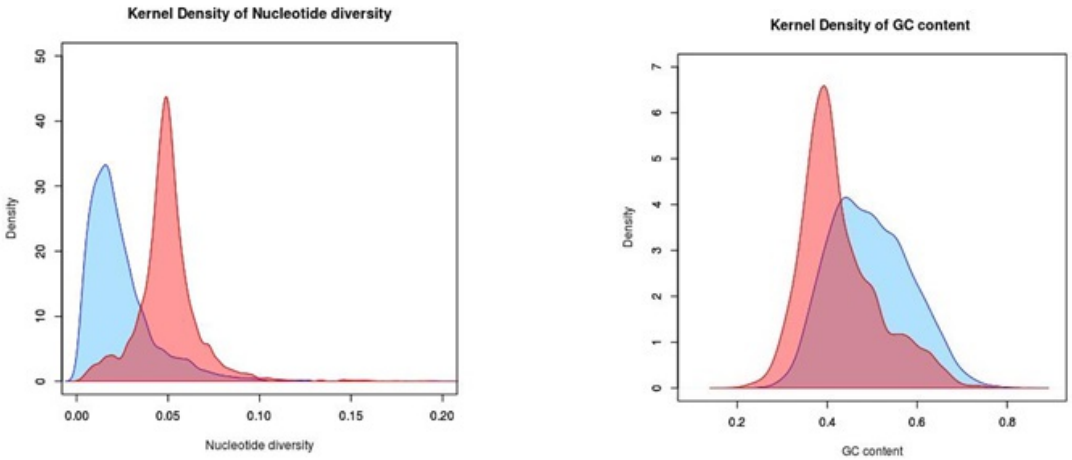

Figure 5. Density distribution of the nucleotide diversity (a) and GC concent (b) of exons (blue) and introns (red).

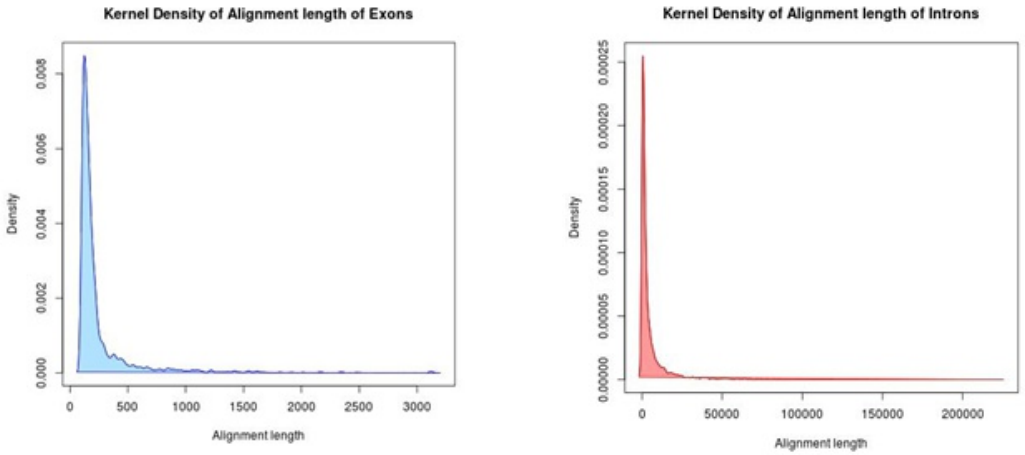

Figure 6. Density distribution of the alignment length of exons and introns. In this case they were plotted separately because both distributions have very different scales.

Best-fit substitution models

In general, best-fit substitution models for exons were simpler (included less parameters) than for introns (Figure 7). The most frequent models for exons were K80 and JC. And the most frequent models for introns were HKY and GTR.

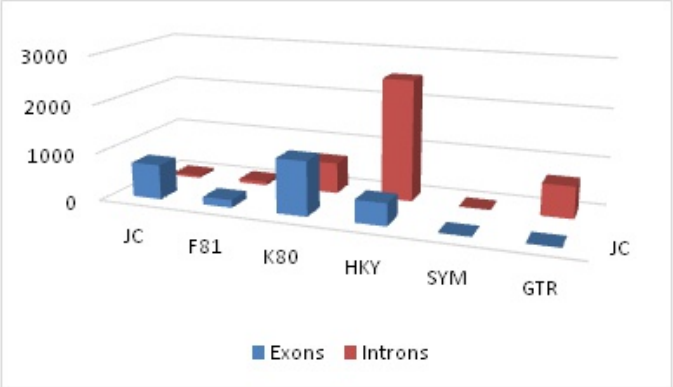

Figure 7. Distribution of best-fit nucleotide substitution models in exons and introns.

Phylogenetic trees

Remarkably, the phylogenetic distances between the estimated trees and the putative primate phylogeny were significantly higher (t-test p-value < 0.001) for exons than for introns (average distance of 2.99 and 0.76, respectively). In fact, intron trees were often identical to the 'known' primate tree (i.e., a phylogenetic distance of 0 in Figure 8).

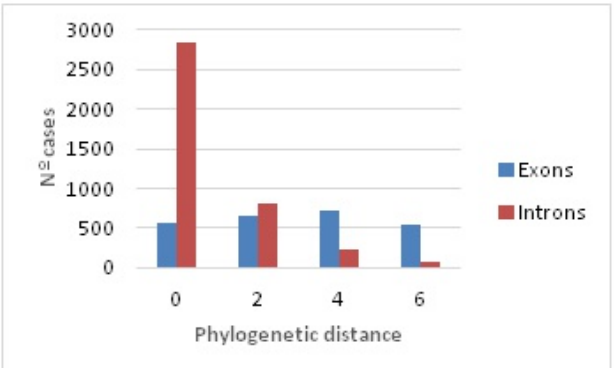

Figure 8. Distribution of the phylogenetic distances in exons and introns.

As the number of species grew, the phylogenetic distance got bigger too, as expected. Moreover, exonic trees were usually less resolved (Figures 9 and 10).

Figure 9. Example of poorly resolved exonic tree. Notice the lack of differentiation between human and chimpanzee.

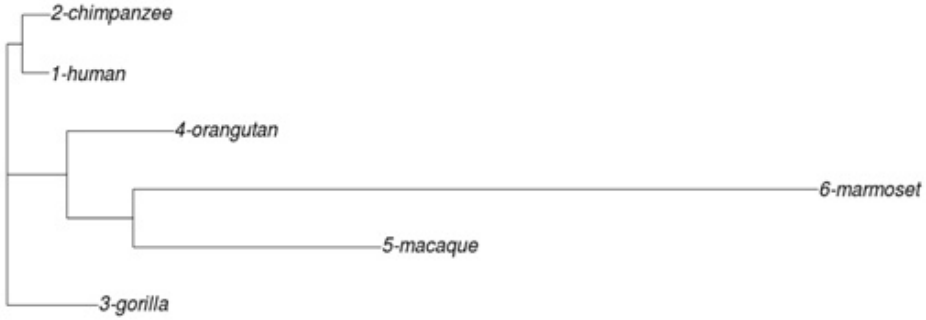

Figure 10. Example of a well resolved intronic tree with phylogenetic distance of 0 to the known primate tree.

Correlation among estimated parameters

Despite the lower Adjusted R-squares, all the relationship were found as significant among any pair of genomic parameters measure, number of species or phylogenetic distance (Table 3).

Table 3. Linear regressions.

Parameter·X	Parameter·Y	Adjusted·R-squared		P-value	
		Exons	Introns	Exons	Introns
N°·Species	Phylogenetic·distance	0.1262	0.0115	<0.001	<0.001
Nucleotide·diversity	GC·content	0.0094	0.0267	<0.001	<0.001
Nucleotide·diversity	Alignment·length	0.0041	0.0010	<0.001	0.019
GC·content	Alignment·length	0.0014	0.0219	0.027	<0.001

DISCUSSION

The function of the exons as coding sequences explains why they are more conserved in all the primate species and why its nucleotide diversity is half the nucleotide diversity of introns. In this case, purifying selection is stronger in exons than introns and exonic regions were smaller than introns (265 vs 6462 bp), as previous studies have showed (Hawkins, 1998; Sakharkar *et al.,* 2004). Finally, the difference found regarding the larger GC content in exons than introns also agrees with previous studies which suggest that this difference is in fact important to the definition of introns and exons during splicing (Amit *et al.,* 2012). Best-fit substitution models were simpler for exons than for introns, which again is the expected result given the smaller nucleotide diversity of the former. Indeed, if the number of changes is different in exons and introns, then the models needed to explain their evolution should be simpler.

In principle, we might expect that phylogenetic trees derived from exons should be more reliable than those estimated from introns, and in fact, many phylogenetic studies rely only on exonic regions. However, in this case I observed the opposite, as the intronic trees were closer to the putative primate tree. One explanation for this result is that the low nucleotide diversity and the short alignment length of exons limits their phylogenetic informativeness. This is congruent with exonic trees tending to have smaller branch lengths resulting in less resolved trees (Figure 10). Previous studies found correlations between parameters as GC content and alignment length (Gazave *et al.,* 2007; Zhu *et al.,* 2009), here I found that all linear correlation between any of the estimated parameters is significant, despite the low Adjusted R-squared values.

A final consideration is that it would be convenient to extend this study to the whole genome, apart from the X chromosome. It is assumed that X chromosome is the most conserved chromosome in mammals (Murphy *et al.,* 1999), so the study of other chromosomes could offer a different view. It is possible that exons from more variable chromosomes are better suited than the X chromosome exons for phylogenetic analysis.

CONCLUSIONS

1. The estimated genic features showed differences between exons and introns. The average nucleotide diversity and alignment length was significantly lower for exons than for introns. On the contrary, the GC content for exons was significantly bigger than for introns.

2. Best-fit substitution models for exons were simpler than for introns. The most frequent model for exons was K80, while for introns it was HKY. Both models imply that separate evolutionary rates for transitions and transversions should be taken into account.

3. The phylogenetic distances between the estimated trees and the putative primate phylogeny were significantly higher for exons than for introns.

4. The results suggest that in the case of the X chromosome, introns are better suited for the study of primate's evolutionary history than exons.

5. It would be convenient to extend this analysis to whole genomes, in order to obtain more general results.

LITERATURE CITED

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Potoslosky, B., Pupko, T., Ast, G. (2012). Differential GC content between exons and introns establishes distinct strageties of splice-site recognition. Cell Rep. 1(5):543-556.

Charif, D., Lobry, J.R., Necsulea, A., Palmeira, L., Penel, S., Perriere, G. (2013). Package "seqinr". Retrieved on June, 2013 from http://cran.r-project.org/web/packages/seqinr/seqinr.pdf

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Singh Riat, H., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suárez, X.M., Harrow, J., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., Searle, S.M.J. (2012). Ensembl 2012. Nucl. Acids Res. 40(D1): D84-D90.

Gazave, E., Marqués-Bonet, T., Fernando, O., Charlesworth, B., Navarro, A. (2007). Patterns and rates of intrón divergence between humans and chimpanzees. Genome Biol. 8:R21

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. 22:35-50.

Gentleman, R. (2008). R Programming for Bioinformatics. USA: Chapman & Hall.

Hawkins, J.D. (1988). A survey on intron and exon lengths. Nucleic Acids Res. 16(21):9893-9908.

Koralewski, T.E., Krutovsky, K.V. (2011). Evolution of Exon-Intron Structure and Alternative Splicing. PLoS ONE. 6(3): e18055.

Krijnen, W.P. (2009). Applied Statistics for Bioinformatics using R. Retrieved on May, 2013 from http://cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J.I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J.B., Bickel, P.,

Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Stone, E.A., Rosenbloom, K.R., Kent, W.J., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Hinrichs, A., Trumbower, H., Clawson, H., Zweig, A., Kuhn, R.M., Barber, G., Harte, R., Karolchik, D., Field, M.A., Moore, R.A., Matthewson, C.A., Schein, J.E., Marra, M.A., Antonarakis, S.E., Batzoglou, S., Goldman, N., Hardison, R., Haussler, D., Miller, W., Pachter, L., Green, E.D., Sidow, A. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 17:760-774.

Martínez, M. (2009). R for Biologists. Retrieved on May, 2013 from http://cran.r-project.org/doc/contrib/Martinez-RforBiologistv1.1.pdf.

Murphy, W.J., Sun, S., Chen, Z-Q., Pecon-Slattery, J., O'Brien, S.J. (1999). Extensive Conservation of Sex Chromosome Organization Between Cat and Human Revealed by Parallel Radiation Hybrid Mapping. Genome Res. 9(12):1223-1230.

Nei, M., Li, W-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci.. 76(10):5269-5273.

Paradis, E. (2003). R para Principiantes. Retrieved on May, 2013 from http://cran.r-project.org/doc/contrib/rdebuts_es.pdf

Paradis, E. (2013). Package "pegas". Retrieved on June, 2013 from http://cran.r-project.org/web/packages/pegas/pegas.pdf

Paradis, E., Bolker, B., Claude, J., Sien Cuon, H., Desper, R., Durand, B., Dutheil, J., Gascuel, O., Heibl, C., Lawson, D., Lefort, V., Legendre, P., Lemon, J., Noel, Y., Nylander, J., Opgen-Rhein, R., Popescu, A.A., Shliep, K., Strimmer, K., de Vienne, D. (2013). Retrieved on June, 2013 from http://cran.r-project.org/web/packages/ape/ape.pdf

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., Birney, E. (2008a). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 18:1814-1828.

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., Birney, E. (2008b). Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 18:1829-1843.

Posada, D., Crandall, K.A. (2001). Selecting the Best-Fit Model of Nucleotide Substitution. Syst. Biol. 50(4):580-601.

R Development Core Team. (2000). Introducción a R. Retrieved on May, 2013 from http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf.

R Development Core Team. (2013). R: A Language and Environment for Statistical Computing. http://www.r-project.org./

Rogozin, I.B., Sverdlov, A.V., Babenko, V.N., Koonin, E.V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. Brief. Bioinform. 6(2):118-134.

Sakharkar, M.K., Chow, V.T., Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Bio-. 4(4):387-393.

Schliep, K.P. (2013). Package "phangorn". Retrieved on June, 2013 from http://cran.r-project.org/web/packages/phangorn/phangorn.pdf

Schwarz, G.E. (1978). Estimating the dimension of a model. Annals of Statistics. 6(2):461-464.

Teufel, A., Krupp, M., Weinmann, A., Galle, P. R. (2006). Current bioinformatics tools in genomic biomedical research (Review). Int. J. Mol. Med. 17:967-973.

Tidall, J. (2001). Beginning Perl for Bioinformatics. USA: O'Reilly.

Tidall, J. (2003). Mastering Perl for Bioinformatics. USA: O'Reilly.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J-Q., Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 10:47.