Modularity and Neutrality in Genomes

DNA Structure, Components Dynamics and Functionally Related Proteins

François Serra

October 2011

Contents

Nomenclatura

bp DNA base-pair

BWT Burros-Wheeler transform

CDS DNA coding sequence

Chr Chromosome

CR Complexity Ratio

CV Complexity Value

dN Rate of non-synonymous mutations

dS Rate of synonymous mutations

GE Genomic Element

GSA Gene-Set Analysis

GSEA Gene-Set Enrichment Analysis

GSSA Gene-Set Selection Analysis

MTF Move To Front

PSG Positively selected genes

SH Significantly high

SL Significantly low

1. Introduction

1.1. What is DNA? How genes rose?

1.2. Definition of neutrality

1.2.1. Neutrality in modularity

Explanation of protein networks by to parameters probability of edge deletion δ and probability of link creation α after a single gene duplication [Solé & Valverde2008]

1.3. Life in DNA, from genes to repetitive elements.

1.4. Adaptive changes to evolutionary speed

1.5. Evolution, and the detection at molecular level

1.6. Grouping genes and finding evolutionary patterns

Part I.

Structure and dynamics of genomes

2. Random-like structure of DNA

2.1. Background

From a biological perspective it seems obvious that DNA is something else than random mix of A, T, G and C nucleotides. Genomes are composed of functional elements as can be protein-coding genes, or promoters but also by non-functional elements like repetitive elements that by definition can not be random when taken together. However to what extent can we state that genomes are not a random soup of 4 letters? From the first analysis of the human genome [Lander *et al.*2001] we have some idea of the proportion of each of the *families* of elements Figure 2.1. Intuitively we could assume that the structure of DNA is different in those families of genomic elements (GE). The sequence of a protein-coding gene would represent a specific selection of nucleotides with surely the highest informational content, while introns would tend more to random assembly and finally we can easily imagine that simple repeats present some biases towards 2 or 3 nucleotides (e.g.: CpG islands).

[Gregory2005]

This question could be solved in some sense by measuring genomes entropy. This measure presents the disadvantage that extreme cases of high entropy could correspond to *a*) **a specially high content of information**, entropy-based algorithms are actually used to predict or confirm automatic detection of genes [Du *et al.*2006, Gerstein *et al.*2007], *b*) **an exact random structure**, some work in the sense of testing the random structure of DNA have been done using entropy [Loewenstern & Yianilos1999]. However this characteristic of entropy could be only a semantic problem if we use it as a measure of relative variation in DNA complexity in genomes, and try to discern statistical patterns in the DNA sequences of different genomic element such as interspersed repeats or functional element (like protein-coding genes). This kind of description of 1DNA sequence complexity was already done by [Holste *et al.*2001], but only in human chromosome 22.

2.2. Results and Discussion

2.2.1. Computing genome complexity

Complexity value (CV) of complete genomes of 54 species of 20 major systematic groups of organisms, ranging 3.4Gb to 1.6Kb genome size was computed Table 2.1. The distribution of these complexity values showed an accurate fit to a linear regression model when genome size was used as the independent variable, with $p < 2.0e\text{-}16$, Figure 2.2-A.

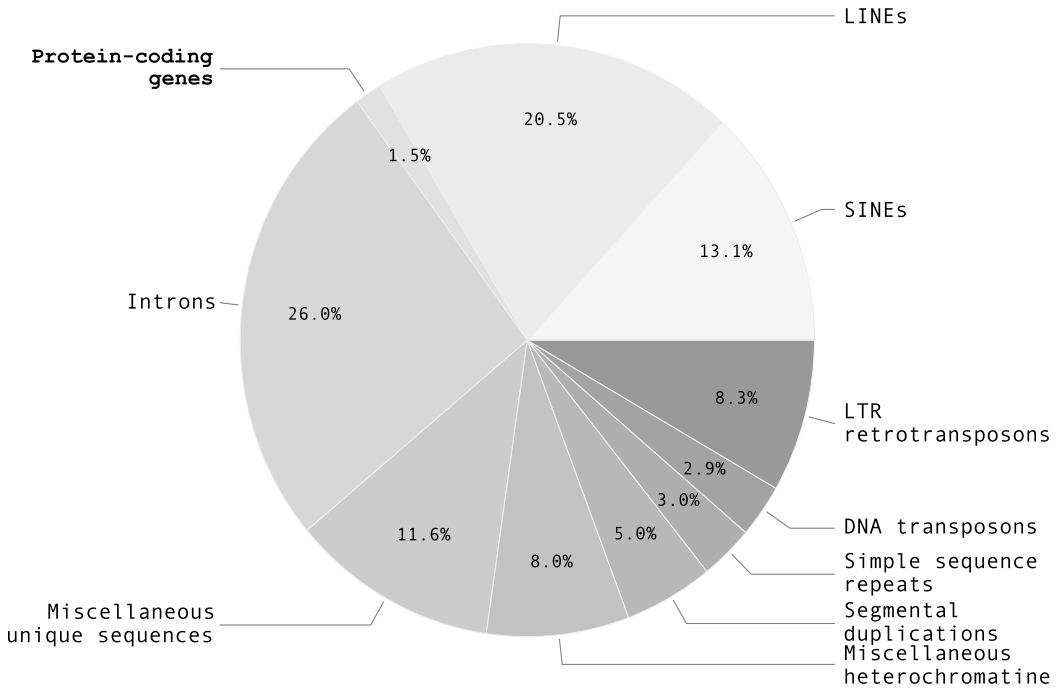


Figure 2.1.: Genomic components of human genome.
Proportion of the major families of different genomic elements (GE) in the human genome according to [Lander *et al.*2001]

The slope (alpha) of the regression (alpha=0.967), was very close to the maximum complexity slope (alpha = 1). Residual variation around the fitted regression was almost null (adjusted-R2 = 0.987). The fit of genomes to almost maximum complexity slope is remarkable considering that the linear model covers six order of magnitude of genome size along all diversity of life. From the shortest single-strand RNA genome of *Hepatitis D* virus (size $\sim$ 1.69e+03 bp) to the largest double-strand DNA genome of the short-tailed opossum (size $\sim$ 3.41e+09 bp). Obligate endosymbionts bacteria with extreme reduction of genome size (*Carsonella ruddii*, *Buchnera aphidicola*, and *Ureaplasma urealyticum*) [Wernegreen2002]; parthenogenetic crustaceans with ubiquitous duplications of genes (*Daphnia pulex*), archean organisms living in extreme environmental conditions (*Sulfolobus islandicus*, *Methanocaldococcus vulcanius*, *Thermococcus sibiricus*), eukaryotes with a variable number of repetitive families, as well as the first synthetic organism made by humans (*Synthetic mycoplasma mycoides*) [Gibson *et al.*2010], fit the slope of the linear regression model.

We studied deviations of complexity value to the slope (alpha) by computing the complexity ratio (CR), and the deviation to the maximum ratio (Dmax = 1- CR). According to Table 2.1, only ten species showed Dmax > 0.05. These are: six ancient or recent polyploid species; the most extreme case of genome reduction in bacteria; the explosive case of gene expansion in Daphnia, and two unicellular eukaryotes.

The highest CR=1 was obtained for non-polyploid (1×), randomly generated sequences with uniform distribution of ACGT; however, CR falls exponentially when ploidy level increases reaching CR=0.25 for 10× Table 2.2. Differences in CR were calculated for polyploids after log transformation and linear regression model adjustment, providing a slope (alpha) = -0.81 (adjusted-R2 = 0.97, p << 0.0001).

Complexity ratios of complete genomes, random sequences of different ploidy and human language texts are displayed in Figure 2.2-B. Maximum CR corresponds to random sequence of lengths ranging from 5 Kb to 2.5 Gb (a, b, c, d and e). Non-polyploid genomes showed CR > 0.90. Within polyploids the lowest ratio corresponds to *Z. mays* with CR=0.58, and the next to the lowest ratio, its closest relative *S. bicolor* with CR=0.78. Overall strings analyzed, the lowest CR was obtained in human language texts. CR of 11 human texts of different sizes and languages, from short scientific abstract to the complete works of William Shakespeare, are also depicted Figure 2.2-B and Figure 2.3. CR diminishes as texts size increases, due to the limited lexicon and the fixed language grammar. Complexity reached the lowest ratio in Darwin's Origin of

Table 2.1. *(following page)*: Genomes Complexity.
Genomes size (GS), genomes complexity (GC), genome complexity ratio ($GCR = \frac{GC}{GS}$), and deviation from the maximum GCR (Dmax=1-GCV) for 54 species of different taxa. NCBI accession number or Ensembl (*E!*) version (ACN-EV). ***Features***: **AP**: Ancient Polyploid; **DSD**: Double-Strand DNA; **EE**: Extreme Environment; **GE**: Gene Expansion; **IP**: Intracellular Parasite; **LBG**: Largest Bacterial Genome; **LGS**: Largest Genome Sequenced; **RG**: Reduced Genome; **RNA**: RNA Virus; **RP**: Recent Polyploid; **SBG**: Shortest Bacterial Genome; **SGS**: Shortest Genome Sequenced; **SL**: Synthetic Life; **SSD**: Single-Strand DNA; **UE**: Unicellular Eukaryote. ***Notes***: -1-: `http://www.hgsc.bcm.tmc.edu/ftp-archive/Tcastaneum/Tcas3.0/`

Features	Species	ACN-EV	Clade	GS	GC	GCR	Dmax
RNA	Hepatitis B	NC3977.1	Virus	1,682	1,671	1	0
SGS-RNA	Hepatitis D	D01075.1	Virus	3,215	3,210	0.9984	0.0016
SSD	Tomato mosaic	NC010836 & NC10835.1	Virus	5,058	5,040	0.9964	0.0036
SSD	Enterobacteria phage m13	V00604	Phage	6,407	6,367	0.9938	0.0062
RNA	HIV 1	NC001802	Virus	9,181	9,105	0.9917	0.0083
RNA	Sudan ebolavirus	NC006432	Virus	18,875	18,842	0.9983	0.0017
DSD	Enterobacteria phage lambda	NC001416	Phage	48,502	48,381	0.9975	0.0025
DSD	Human herpesvirus1	NC001806	Virus	152,261	150,036	0.9854	0.0146
SBG-IP-RG	Carsonella ruddii	NC008512	Bacteria	159,662	146,930	0.9203	0.0797
IP-RG	Buchnera aphidicola	AE013218.1	Bacteria	642,122	626,533	0.9757	0.0243
IP-RG	Ureaplasma urealyticum	CP001184	Bacteria	873,755	840,812	0.9623	0.0377
SL	Synthetic mycoplasma mycoides	CP002027.1	Bacteria	1,078,809	1,026,444	0.9515	0.0485
EE	Thermococcus sibiricus	CP001463.1	Archaea	1,242,891	1,237,320	0.9955	0.0045
EE	Methanocaldococcus vulcanius	CP001787.1	Archaea	1,746,040	1,708,968	0.9788	0.0212
EE	Sulfolobus islandicus	CP001731.1	Archaea	2,722,004	2,692,455	0.9891	0.0109
	Bacillus subtilis	*E!* Bacteria 9	Bacteria	4,215,606	4,198,057	0.9958	0.0042
	Mycobacterium tuberculosis	*E!* Bacteria 9	Bacteria	4,411,532	4,348,606	0.9857	0.0143
	Escherichia coli	CP001396.1	Bacteria	4,578,159	4,551,258	0.9941	0.0059
LBG	Burkholderia xenovorans	NC007951-3	Bacteria	9,731,138	9,593,486	0.9859	0.0141
AP	Saccharomyces cerevisiae	*E!* Fungi 3	Fungi	12,070,898	11,974,342	0.992	0.008
UE	Plasmodium falciparum	*E!* Protists 9	Apicomplexa	23,263,332	21,070,640	0.9057	0.0943
UE	Phaeodactylum tricornutum	*E!* Protists 9	Heterokonta	25,805,651	25,667,448	0.9946	0.0054
UE	Dictyostelium discoideum	*E!* Protists 9	Amebozoa	31,199,234	31,023,020	0.9944	0.0056
UE	Thalassiosira pseudonana	*E!* Protists 9	Heterokonta	33,919,934	30,877,496	0.9103	0.0897
	Ciona intestinalis	*E!* 62	Urochordate	87,649,861	84,674,396	0.9661	0.0339
	Caenorhabditis elegans	*E!* Metazoa 9	Invertebrates	100,272,217	97,720,472	0.9746	0.0254
	Tribolium castaneum	-1-	Invertebrates	112,129,668	109,424,212	0.9759	0.0241
AP-RG	Arabidopsis thaliana	*E!* Plants 9	Plants	118,960,082	116,563,556	0.9799	0.0201
	Drosophila melanogaster	*E!* Metazoa 9	Invertebrates	120,290,887	118,973,632	0.989	0.011
GE	Daphnia pulex	*E!* Metazoa 9	Invertebrates	158,632,523	150,111,316	0.9463	0.0537
AP	Arabidopsis lyrata	*E!* Plants 9	Plants	173,245,910	161,798,504	0.9339	0.0661
AP	Tetraodon nigroviridis	*E!* 62	Fishes	208,708,313	207,067,712	0.9921	0.0079
	Apis mellifera	*E!* Metazoa 9	Invertebrates	224,750,524	219,278,732	0.9757	0.0243
	Anopheles gambiae	*E!* Metazoa 9	Invertebrates	225,028,531	221,180,624	0.9829	0.0171
AP	Brachypodium distachyon	*E!* Plants 9	Plants	270,058,956	257,893,524	0.955	0.045
AP	Oryza sativa	*E!* Plants 9	Plants	293,104,375	271,137,108	0.9251	0.0749
AP	Populus trichocarpa	*E!* Plants 9	Plants	370,421,283	352,063,876	0.9504	0.0496
AP	Physcomitrella patens	*E!* Plants 9	Bryophyta	453,927,385	399,508,556	0.8801	0.1199
AP	Sorghum bicolor	*E!* Plants 9	Plants	625,636,188	491,993,216	0.7864	0.2136
AP	Oryzias latipes	*E!* 62	Fishes	582,126,393	562,662,192	0.9666	0.0334
	Gallus gallus	*E!* 62	Birds	984,855,151	971,359,304	0.9863	0.0137
	Taeniopygia guttata	*E!* 62	Birds	1,013,982,659	996,918,996	0.9832	0.0168
AP	Danio rerio	*E!* 62	Fishes	1,354,636,069	1,191,452,752	0.8795	0.1205
AP-RP	Zea mays	*E!* Plants 9	Plants	2,045,697,632	1,197,255,904	0.5853	0.4147
	Canis familiaris	*E!* 62	Mammals	2,309,875,279	2,272,374,188	0.9838	0.0162
	Equus caballus	*E!* 62	Mammals	2,335,454,424	2,307,202,104	0.9879	0.0121
	Bos taurus	*E!* 62	Mammals	2,466,956,401	2,406,743,280	0.9756	0.0244
	Rattus norvegicus	*E!* 62	Mammals	2,477,053,718	2,430,894,052	0.9814	0.0186
	Mus musculus	*E!* 62	Mammals	2,558,509,481	2,521,038,616	0.9854	0.0146
	Pan troglodytes	*E!* 62	Mammals	2,598,733,311	2,566,544,200	0.9876	0.0124
	Macaca mulatta	*E!* 62	Mammals	2,646,263,164	2,621,196,144	0.9905	0.0095
	Pongo abelii	*E!* 62	Mammals	2,722,968,487	2,697,592,876	0.9907	0.0093
	Homo sapiens	*E!* 62	Mammals	2,858,658,095	2,841,049,052	0.9938	0.0062
LGS	Monodelphis domestica	*E!* 62	Mammals	3,412,593,369	3,402,944,248	0.9972	0.0028

Species (0.309), which is comparable to the CR of a random polyploid sequence of 7×. Observe that text sizes are contained in the range of phages, virus and bacteria genome sizes. Details of complexities of human writings are in shown Table 2.2.

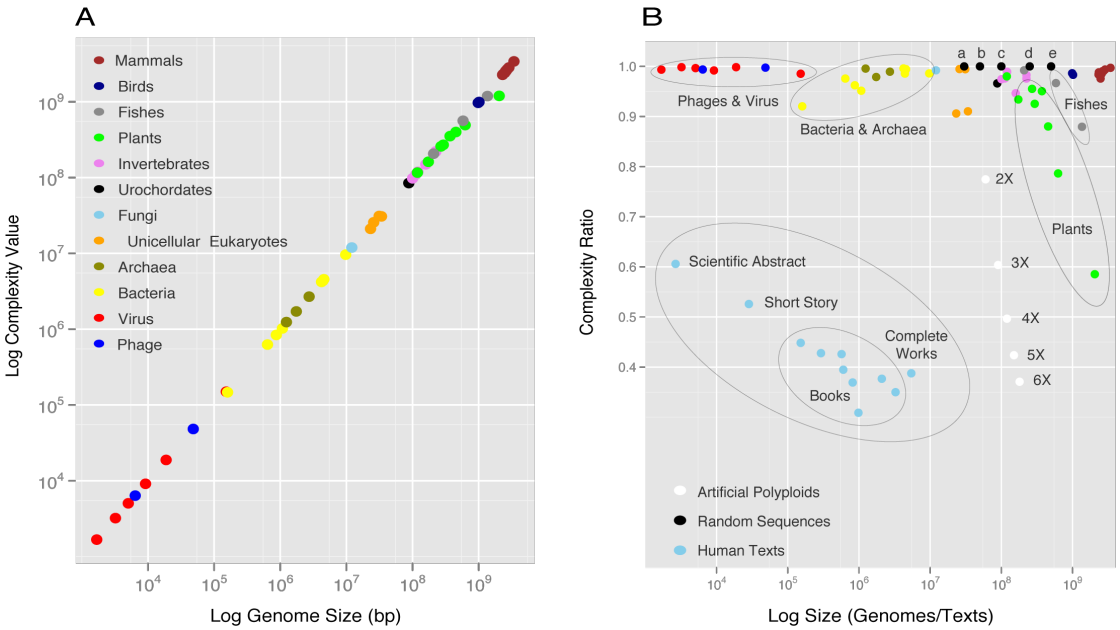
2.2.2. Genome complexity and ploidy level

Recent polyploid species as maize and sorghum exhibited noticeably low complexity ratios, however, ancient polyploids and non-polyploids had indistinguishable complexity ratios. We tested the hypothesis that the observed genome complexity values are correlated with size and ploidy level. A categorical variable divided polyploid (ancient or recent), and non-polyploid species described in Table 2.1. The size-interaction term provided significant deviations ($p < 2e\text{-}16$, adjusted-R2 = 0.997), while independent linear models slopes were 0.633 ($p < 4.8e\text{-}07$, adjusted-R2 = 0.921), and 0.988 ($p < 2e\text{-}16$, adjusted-R2 = 1.00) for polyploid and non-polyploid genomes.

2.2.3. Chromosome complexity

Complexity value of each eukaryote chromosome (567 autosomes of 31 species) was computed and plotted against size Figure 2.4-A. Linear regression models considering the full

Figure 2.2. ***(following page)*****: Genome complexity value.**
(A) Complexity values and genome size of 54 genomes. Log scales are used to display species diversity. Species listed by genome size increase are (see Table 2.1 for details): *Hepatitis D* (V), *Hepatitis B* (V), *Tomato mosaic* (V), *Enterobacteria phage m13* (Ph), *Hiv 1* (V), *Sudan ebolavirus* (V), *Enterobacteria phage lambda* (Ph), *Human herpesvirus1* (V), *Carsonella ruddii* (Ba), *Buchnera aphidicola* (Ba), *Ureaplasma urealyticum* (Ba), *Synthetic mycoplasma mycoides* (Ba), *Thermococcus sibiricus* (Ar), *Methanocaldococcus vulcanius* (Ar), *Sulfolobus islandicus* (Ar), *Bacillus subtilis* (Ba), *Mycobacterium tuberculosis* (Ba), *Escherichia coli* (Ba), *Burkholderia xenovorans* (Ba), *Saccharomyces cerevisiae* (Fu), *Plasmodium falciparum* (Ue), *Phaeodactylum tricornutum* (Ue), *Thalassiosira pseudonana* (Ue), *Dictyostelium discoideum* (Ue), *Ciona intestinalis* (Ur), *Caenorhabditis elegans* (I), *Tribolium castaneum* (I), *Arabidopsis thaliana* (Pl), *Drosophila melanogaster* (I), *Daphnia pulex* (I), *Arabidopsis lyrata* (Pl), *Tetraodon nigroviridis* (Fi), *Apis mellifera* (I), *Anopheles gambiae* (I), *Brachypodium distachyon* (Pl), *Oryza sativa* (Pl), *Populus trichocarpa* (Pl), *Physcomitrella patens* (Pl), *Oryzias latipes* (Fi), *Sorghum bicolor* (Pl), *Gallus gallus* (Bi), *Taeniopygia guttata* (Bi), *Danio rerio* (Fi), *Zea mays* (Pl), *Canis familiaris* (M), *Equus caballus* (M), *Bos taurus* (M), *Rattus norvegicus* (M), *Mus musculus* (M), *Pan troglodytes* (M), *Macaca mulatta* (M), *Pongo abelii* (M), *Homo sapiens* (M), *Monodelphis domestica* (M). V: Virus, Ph: Phage, Ba: Bacteria, A: Archaea, Fu: Fungi, Ue: Unicellular eukaryote, Ur: Urochordate, I: Invertebrate, Pl: Plants, Fi: Fish, Bi: Bird, M: Mammal. **(B)** Most genomes have complexity ratio (CR) between 0.90 and 1.0. Four polyploid species have CR < 0.9: P. patens (0.880), *D. rerio* (0.879), *S. bicolor* (0.786) and *Z. mays* (0.585). a, b, c, d, e correspond to random [ACGT] strings of 30, 50, 100, 250 and 500 Mb length, respectively. 2× to 6× correspond to random polyploids [ACGT] sequences where 1× is "a". Changes in sequence length due to polyploidy produce no change in complexity ratio (see Table 2.2). Notice the low CR of human texts (see Table S3 for details).



Features	Author - Writings	Language	L	C	CR
SA	C. Venter. The human genome (abstract)	English	2,662	1,613	0.6059
SS	J. L. Borges. El Aleph	Spanish	28,507	14,991	0.5259
B	A. Von Goethe. Torcuato Tasso	German	152,104	68,187	0.4483
B	H. Quiroga. Cuentos amor, locura y muerte	Spanish	293,482	125,552	0.4278
B	D. F. Sarmiento. Facundo	Spanish	601,477	242,982	0.4259
B	D. Alighieri. Divina Commedia	Italian	570,480	301,609	0.3692
B	I. Newton. Principia Mathematica	Latin	817,032	237,558	0.395
B	B C. Darwin. The Origin of species	English	981,958	303,503	0.3091
B	B M. Cervantes. El Quijote	Spanish	2,097,943	790,702	0.3769
B	B V. Hugo. Les Miserables	French	3,259,269	1,141,378	0.3502
CW	W. Shakespeare	English	5,447,165	2,111,425	0.3876

Table 2.2.: Human language Complexity
Work length (L), complexity (C), complexity ratio (CR), and deviations from the maximum ratio of complexity (Dmax=1- CR) for 11 human writings in six different languages. Features: SA: Scientific abstract, SS: Short story; B: Book, CW: Complete Work

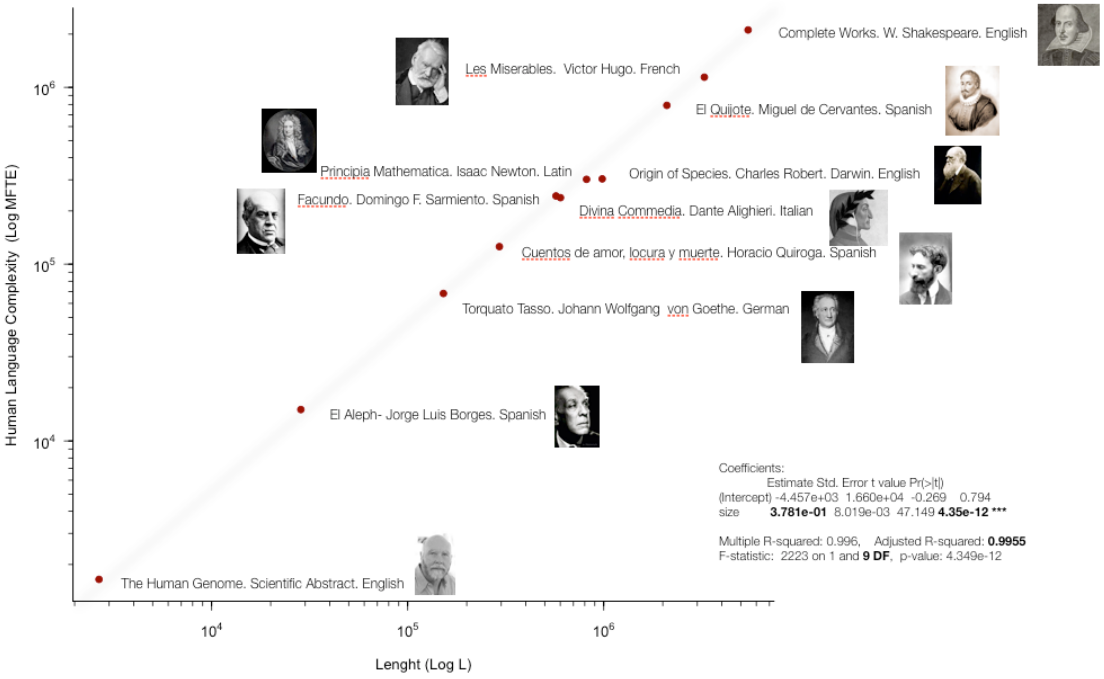


Figure 2.3.: Human language complexity.
Complexity in human writings shows a constant increase with text length. Regression analysis shows that in contrast to genomes, human language is highly repetitive. While genomes match an almost perfect regression of slope 1, human language complexity fits a linear regression model with slope alpha=0.378, (adjusted R = 0.995).

dataset, or excluding polyploid species revealed a very significant statistical relationship (slope = 0.924, adjusted-R2 = 0.989, p < 2e-16, or slope = 0.951, adjusted-R2 = 0.999, p < 2e- 16, respectively). The adjustment of a linear regression model to chromosomes of polyploid species was statistically significant (p < 2e-16, R2 = 0.982), while their complexity values exhibited a lower slope (alpha= 0.696), than the complexity values of chromosomes of non- polyploid species. Again, as was observed in genomes, the size-interaction term was statistically significant (p < 2e-16), suggesting that complexity and size deviates differently for chromosomes of polyploid and non-polyploid species. Notice that for non-polyploid species the slope of their chromosome complexity values against size almost coincides with the slope of their genome complexity values (alpha = 0.989, 0.988, respectively). Figure 2.4-B displays CR for chromosomes. The boxplot inside shows the distribution of CR for all chromosomes. The fist quartile of the full sample indicates that 75% of the data are above 0.958, while the median and mean was 0.974 and 0.964. The minimum CR value corresponds to maize chromosome 10 (0.683), and maximum to *P. tricornutum* chromosome 28 (0.999). Opossum chromosome 1 (the largest chromosome) has a CR of 0.942. Mean CR of maize's chromosomes was 0.698, while maize genome CR was 0.585. The difference suggests extensive duplicated regions in maize chromosomes, which was previously described in [Weber & Helentjaris1989, Gaut2001] and attributed to a tetraploid event occurred in the origin of maize 11.4 My ago [Gaut & Doebley1997, Wolfe2001]. However, differences between mean chromosome to genome CR were observed in different species with variable deviations: sorghum (0.854:0.786), zebrafish (0.924:0.879), *A. lyrata* (0.966:0.934), *P. trichocarpa* (0.971:0.950), *S. cerevisae* (0.996:0.992), and *A. thaliana* (0.986:0.980), *M. domestica* (0.944:0.997), *M. musculus* (0.959: 0.985), and *H. sapiens* (0.960:0.993). Appendix A gives the values for the full data set. Further insights on chromosome and genome CR differences are discussed in the section on polyploid and return to maximum complexity, below.

2.2.4. Complexity in chromosome segments

Chromosomes were split in overlapping windows of various sizes (from 1 Kb to 100 Mb) and complexity ratio in these windows was computed. Figure 2.5 shows boxplots of six selected chromosomes, at different scales, all having extreme CR. Median values of CR over all windows of *H. sapiens* Chr1 Figure 2.5-A, *A. thaliana* Chr1 Figure 2.5-C, *C. elegans* Chr1 Figure 2.5-D, and *D. melanogaster* Chr2L Figure 2.5-E were above 0.97. Lower values were obtained in *Z. mays* Chr 1 Figure 2.5-F and in *H sapiens* Chr19 Figure 2.5-B for large windows sizes; in particular, for windows larger than 1Mb, CR noticeably fell down. The reasons for this fall are different in the two cases: while maize Chr1 is tetraploid, human Chr19 contains the highest number of Alu sequences reported in human chromosomes [Venter *et al.*2001].

In general, for all chromosomes, the larger the window size, the lower the median CR value. This pattern can be explained by existence of repeats, which can only be detected when the window size is large enough. In addition, CR dispersion decreases when window size increases, a fact that is explained by the substantial DNA combinatorial variation in

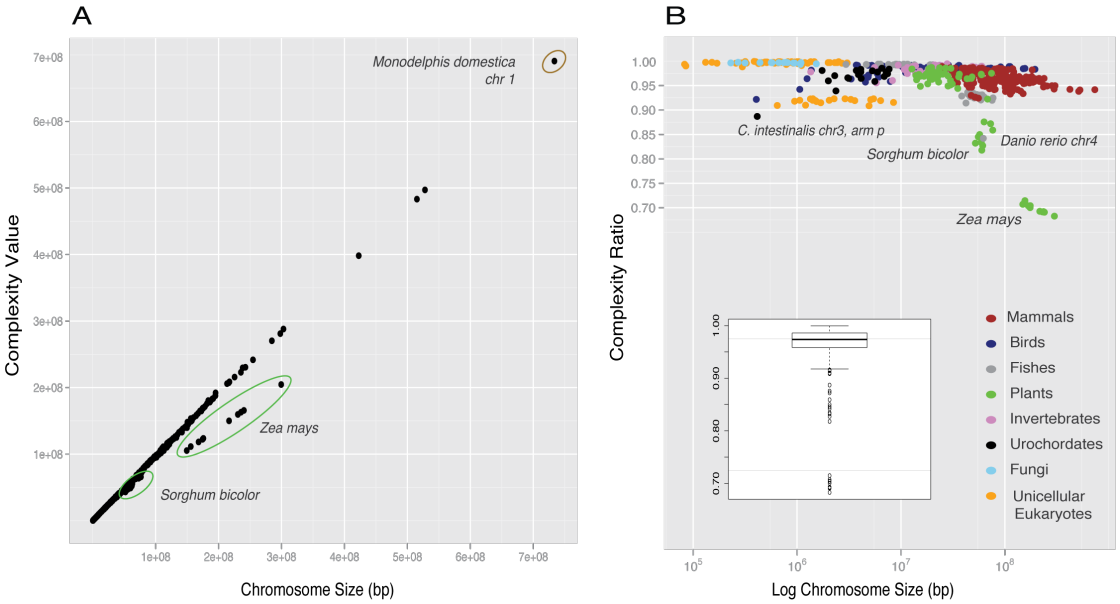


Figure 2.4.: Chromosome complexity ratio.
(A) Complexity ratio and chromosome size of 31 eukaryote species (567 chromosomes). Notice how far chromosomes of Z. mays, and in minor degree S. bicolor (both recent polyploid species) depart for the general trend. **(-B)** Most chromosomes (96.2%) have complexity ratios ranging 0.9 to 1.0, as observed for complete genomes Figure 2.2B. Boxplot inside shows the distribution of CR of all chromosomes.

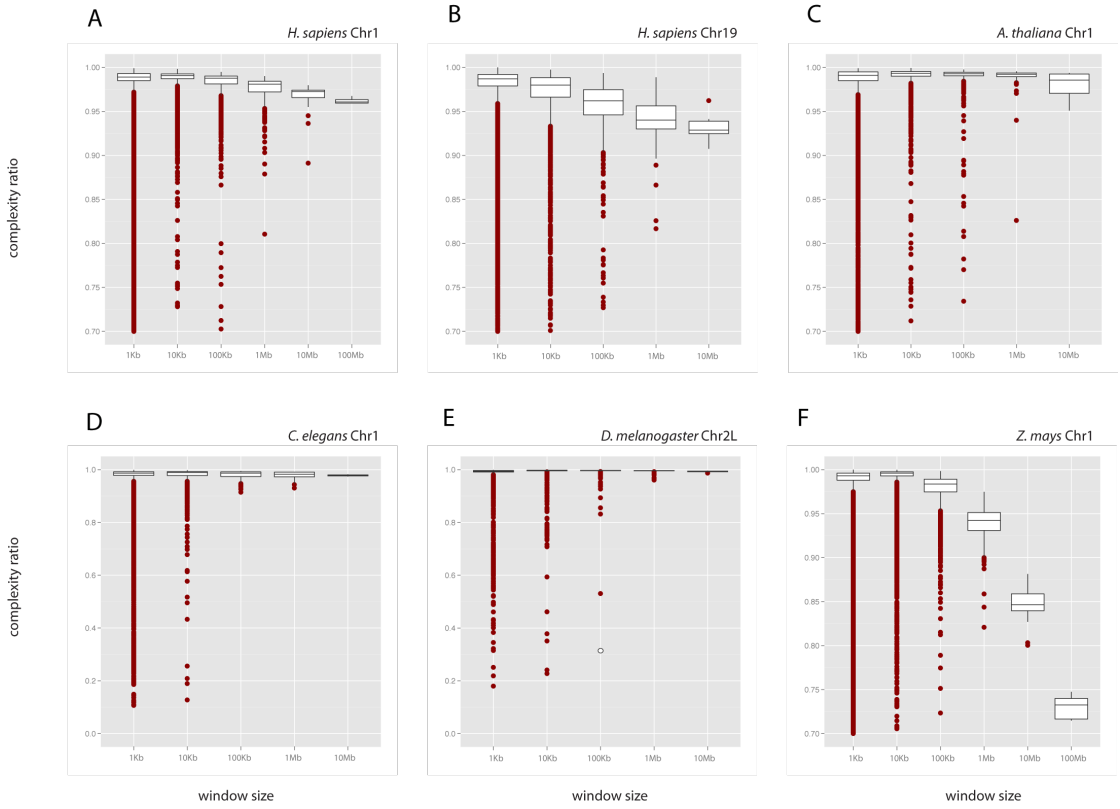


Figure 2.5.: Sliding window analysis in chromosomes.
Boxplots show results of sliding window analyses in six selected chromosomes (A-F). Most chromosomes have median CR higher than 0.975 independently of window size. White dot in the 100Kb window size chart of D. melanogaster Chr 2L (E) corresponds to the "deep spike" displayed in Figure 2.6. Scales were selected to enlarged differences in CR.

large chromosome windows. This effect is shown in Figure 2.6 for window sizes of 1Kb and 100 Kb in *D. melanogaster* Chr 2L, with a rugged versus smooth CR profiles. The outstanding minimum CR 0.31 was for 100 Kb-window size. This sudden decrement in CR occurred at 21,400 - 21,550 Mb where the histone cluster with more than 100 genes of the family locates.

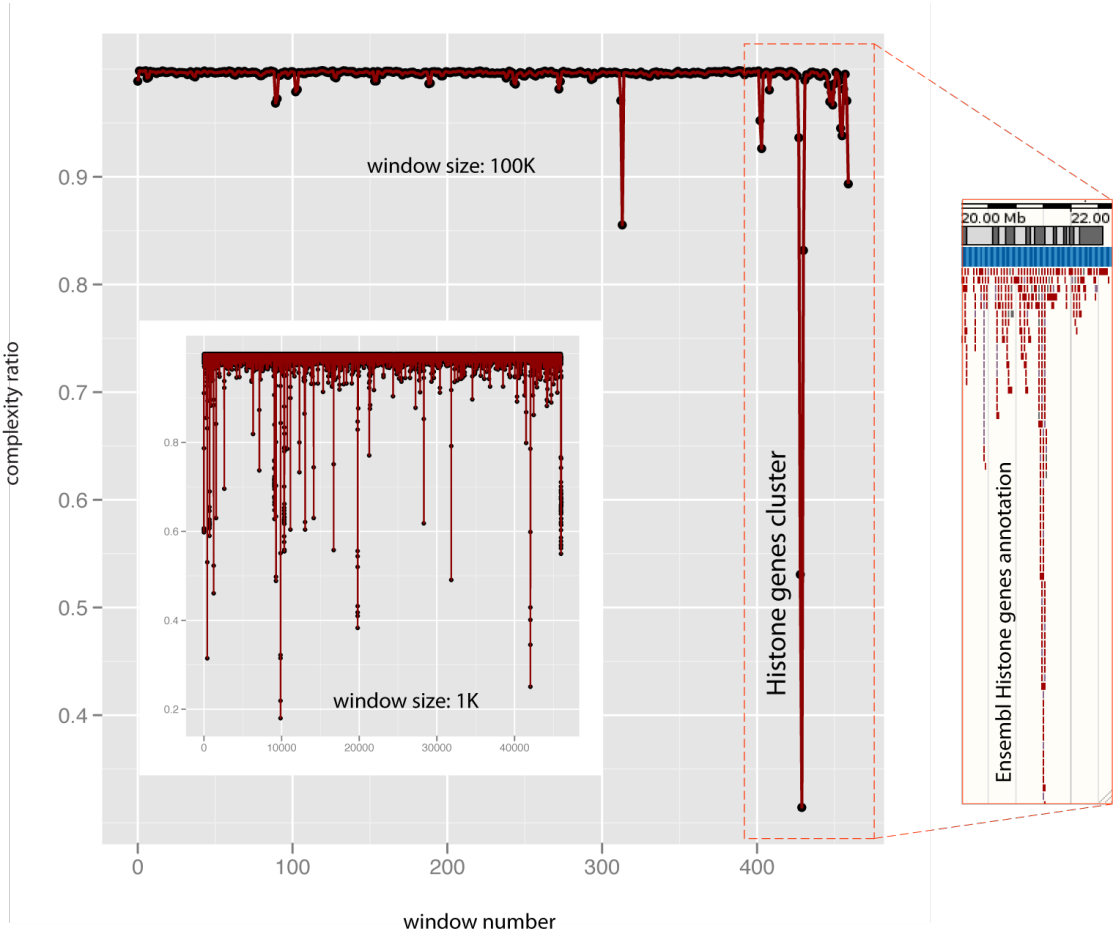


Figure 2.6.: Sliding window in a full chromosome.
Complexity ratio along *D. melanogaster* chromosome 2L is displayed at two window size scales. Ensembl annotation of the histone genes cluster is shown in the left box. See associated DAS server displaying CR along chromosomes for 15 eukaryote species.

The right picture shows Ensembl annotation for the histone genes cluster. The complexity values correspondind to an exhaustive sliding-window scan of chromosomes of fifteen different species is available in a DAS server at `http://bioinfo.cipf.es/das/` Complexity in repetitive elements and genes Eukaryote genome structure is generally sketched out by the massive presence of non-functional repetitive elements (RE-s) spread out all over the genome, and a tiny portion of singular functional elements covering the rest. To get insights into the statistical structure of these contrasting regions of genomes we com-

puted the complexity ratio of genes and of each of the main families of RE's (as DNA-T, LTR, LINE, SINE and satellite). To do this, for each family, all units were concatenated in their original order in chromosomes, after scanning them with RepeatMasker (see details of RepeatMasker output for individual species in Document S1).

Genes showed, as expected, the highest CR among all classes analyzed, independently of the species. When genes were split in their two main components, exons showed even a higher CR. Unexpectedly, high values of CR were also obtained in LINE, LTR and DNA-T (Table 2.3). In constrast, SINE and satellites showed the lowest CR. The low complexity ratio observed in SINE and satellites is mainly due to their repetitive structure, in the form of orderly arranged short sized repetitions. The high CR associated to LINE, LTR and DNA-T (DNA-T) is explained by their larger length and their high internal variability in units of the families. In mammals DNA-T and LTR elements exhibited higher CR than LINE elements. This is not the case for fishes, some invertebrates and plants. In plants, LINE has the highest CR after genes (Table 2.3 and see Figure 2.3 for comparison among all eukaryote species analyzed).

Table 2.3.: Mean complexity ratio of some genome components in different species.

Species	**Satellite**	**SINE**	**LINE**	**LTR**	**DNA-T**	**Genes**	**Introns**	**Exons**
H. sapiens	0.485	0.437	0.881	0.922	0.962	0.953	0.952	0.985
P. troglodytes	0.491	0.442	0.885	0.926	0.962	0.967	0.965	0.993
R. norvegicus	0.539	0.586	0.668	0.912	0.975	0.977	0.976	0.992
M. musculus	0.595	0.576	0.74	0.875	0.973	0.973	0.97	0.991
C. familiaris	0.6	0.487	0.911	0.974	0.982	0.982	0.98	0.993
T. nigroviridis	—	0.585	0.903	—	—	0.994	0.993	0.993
D. rerio	0.628	0.43	0.796	0.791	0.824	0.942	0.936	0.988
C. intestinalis	0.644	0.537	0.836	0.937	0.801	0.968	0.957	0.994
C. elegans	0.52	0.401	0.93	0.94	0.827	0.978	0.957	0.99
A. gambiae	0.232	0.438	0.805	0.902	0.771	0.992	0.992	0.9
D. melanogaster	0.548	—	0.81	0.744	0.81	0.985	0.982	0.99
Z. mays	0.337	0.531	0.906	0.495	0.7223	0.962	0.956	0.975
S. bicolor	0.345	0.619	0.966	0.602	0.757	0.99	0.991	0.988
A. thaliana	0.467	0.675	0.971	0.84	0.896	0.989	0.986	0.988
A. lyrata	0.417	0.457	0.928	0.772	0.826	0.994	0.988	0.996

For each family, we used the complexity ratio to describe the disposition of the elements inside a chromosome. CR of linearly arranged elements was compared to the CR of shuffled elements. Table 2.4 shows these values for eight selected chromosomes of different species. CR in the linear arrangement was much lower in SINE and satellites than in the rest of the classes. This reveals a structure of identical or very similar repeats along neighbor chromosome segments. This pattern did not showed up in the other families. The notable exception was LTR of the maize chromosome, known to have expanded dramatically in recent evolutionary times [Blanc & Wolfe2004]. All shuffled classes (including SINE and satellites) had a CR equal to one, or very close to one. This entails an almost uniform statistical distribution of DNA sequences in that class. This result point outs that genomes are plenty of genetic variation, even in regions where

the expected pattern is the homogeneous repetition of almost indistinguishable units of RE's.

Table 2.4.: Size and complexity ratio of different genome classes concatenated (CON), and shuffled (SHU) for selected chromosomes. Size in Mb.

		A. thaliana		***C. elegans***		***H. sapiens***		***Z. mays***	
		Chr 1	Chr 5	Chr 1	Chr 2	Chr 1	Chr 21	Chr 1	Chr 10
Satellite	SIZE	0.476	0.147	0.159	0.149	0.172	0.118	0.48	0.288
	CON	0.223	0.299	0.489	0.547	0.519	0.567	0.325	0.309
	SHU	0.889	0.968	0.962	0.975	0.972	0.987	0.961	0.948
SINE	SIZE	0.023	0.023	0.009	0.007	35.782	3.979	0.051	0.023
	CON	0.69	0.682	0.367	0.402	0.439	0.433	0.525	0.531
	SHU	0.976	0.975	0.956	0.943	0.925	0.942	0.945	0.951
LINE	SIZE	0.121	0.146	0.039	0.026	26.321	3.778	1.454	0.739
	CON	0.975	0.972	0.93	0.982	0.874	0.905	0.899	0.916
	SHU	1.000	1.000	0.999	1.000	0.999	1.000	1.000	1.000
LTR	SIZE	0.914	0.944	0.022	0.013	10.474	2.11	115.466	57.56
	CON	0.811	0.809	0.98	0.984	0.906	0.93	0.47	0.513
	SHU	0.998	0.998	1.000	1.000	0.999	1.000	0.993	0.995
DNA-T	SIZE	0.68	0.541	0.704	0.518	3.734	0.552	6.066	3.109
	CON	0.883	0.887	0.81	0.84	0.95	0.98	0.7	0.74
	SHU	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000
GENES	SIZE	18.242	16.312	10.77	9.918	140.258	21.909	37.623	16.759
	CON	0.988	0.989	0.975	0.981	0.951	0.964	0.956	0.967
	SHU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
INTRON	SIZE	5.318	4.73	6.074	4.94	130.429	20.696	22.229	9.735
	CON	0.985	0.986	0.95	0.963	0.95	0.964	0.948	0.966
	SHU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
EXON	SIZE	12.925	11.582	4.694	4.979	9.829	1.213	15.394	7.024
	CON	0.988	0.989	0.991	0.991	0.983	0.99	0.972	0.976
	SHU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Polyploidy and return to maximum complexity Evolution erodes ancient footprints of genome polyploidy and diploidization (the process by which a polyploid genome turns into a diploid one) proceeds during time [Wolfe2001]. As shown in previous sections, CR of recent polyploids is much lower than in non-polyploid, or in ancient polyploid species. Diploidization can be achieved by multiple mechanisms [Wolfe2001], being the gradual disintegration of the duplicated genetic material by random mutation. This is the simplest form. However, more dramatic mechanisms such as massive deletion, and transpositions of genetic material was reported in *A. thaliana* [Hu *et al.*2011]. We tested the hypothesis that the complexity ratio of polyploid genomes increases along the diploidization process.

Polyploid origin and posterior decay of genetic redundancy was simulated by means of mutations and transpositions in random sequences of different lengths and ploidy levels, and in *Z. mays* Chr1 and *S. bicolor* Chr1 (Figure 5). In all cases, sequences under

random mutation and transposition reached maximum CR=1 after a number of generations large enough. Larger sequences representing genomes or chromosomes increased their CR faster than shorter sequences. This is as expected in probability theory since each sigle choice (introduced by a random mutation or a transposition) in a large set is more informative than in a smaller set, because it makes a selection in a bigger space of possibilities. The dynamics of CR increase was identical for the maize and sorghum chromosomes and the simulated random sequences (Figure 5A). Figure 5B shows that genomes and chromosomes reached maximum CR=1 after many cycles of transpositions. Using a simulated genome with tetraploid structure, transposition preserved the relation that chromosome CR is higher than genome CR, along all generations up to convergence to maximum CR=1. This feature was reported above for maize and sorghum (see discussion on chromosome complexity ratio).

Once CR reached almost maximum complexity any signal of polyploidy is finally lost, and DNA structure is indistinguishable from diploid genomes. High complexity and random-like structure of DNA Excluding recent polyploids, high CR (almost maximum) was observed in complete genomes of organisms sampled in all diversity of life, in their chromosomes and along large enough chromosomes segments, and in shuffled arrangements of elements in individual genetic classes. We conjecture this is a universal feature of all genomic sequences. Polyploids were the only DNA sequences on which we obtained low complexity ratios, hence, they are the only DNA sequences with the distinguishing feature low CR: Low CR corresponds to a simple combinatorial structure of the sequence.

The combinatorial structure of a sequence is a description of the observed arrangement of the symbols among all possible permutations of the same length. Sequences with many long repeats have low CR. Sequences with the minimum CR=0 consist of a single symbol (as $AAAAAAAAAA$) -this is the simplest combinatorial structure- so they are plainly compressible. Polyploid genomes of maize and sorghum have CR=0.585, and CR=0.786, respectively. Values that were close to the simulated tetraploid genomes (Fig 5B). It is also possible to achieve low CR in sequences without any long repeats, but with an orderly arrangement of the symbols. Although we have not found this phenomena in natural DNA, we constructed de Bruijn sequences (these are the mathematically defined sequences with perfect equifrequency of subsequences: every possible sequence of logarithmic length appears exactly once as a sequence of consecutive symbols) [de Bruijn1946, Becher & Heiber2011] with low CR. The regularity in the combinatorial structure of these particular de Bruijn sequences is captured by the MTF algorithm. See Appendix I for examples on short sequences. High complexity ratio implies the following properties on the combinatorial structure of DNA sequences: High CR corresponds to high diversity and balanced abundance of short repeats. Maximum CR=1 is reached by sequences a sequence of length n if it contains full diversity of length k, for k ! log4 n, and each these short sequences occur about n*4-k times. As CR decreases, diversity and balanced abundance deteriorates. In particular, maximum CR is holds for some de Bruijn sequences [de Bruijn1946, Becher & Heiber2011]. Also maximum CR=1 occurs in randomly generated sequences with uniform distribution of A, C, G, T. For genomic sequences [Liu *et al.*2008] reported that more than 98% of 12 bp oligomers appear in vertebrate genomes while less than 2% of 19 bp oligomers are present. For the human

genome we computed all maximal exact repeats over 30 bp, and counted their diversity and quantity [Nies2009]. We observed that the largest correlations in the human genome are intrachromosomal, the actual largest exact repeat is 67,632bp long and it occurs just twice inside Chr 1, while the largest inter-chromosomal perfect correlation is 21,865 bp occurring just once Chr 1 and once in Chr 5.

High CR corresponds to random-like sequences. Intuitively, a non random sequence will exhibit some significant regularity that can be used to compress the sequence. The mathematical underpinning relies on the theory of pure randomness [Chaitin1975, Nies2009], which states that an infinite sequence is random when its initial segments are incompressible. Up to some deviations, for finite sequences and particular compression methods, the identification between statistical randomness and incompressibility holds. The complexity ratio (CR) expresses incompressibility by the BWT-MTF scheme and further encoding as determined by Shannon's entropy. High complexity ratios correspond to highly incompressible sequences, which are sequences with a random-like structure. As in statistical randomness, the number of sequences with high CR grows exponentially with the sequence length. Thus, each genome is a singular instance out of the extraordinary many combinatorial variants of the same length with the same high complexity rate. A lower bound of the number of sequences with high CR (CR " 1 - !, for any real value ! between 0 and 1) is proved Appendix II.

2.3. Material and methods

2.3.1. The complexity ratio and complexity value

Complexity Ratio (CR) is defined by a classical formula used in data compression [Adjeroh *et al.*2008], the Burros-Wheeler transform BWT [Burrows & Wheeler1994], followed by the Move To Front (MTF) [Ryabko1980] and finally resume this to one value using Shannon's entropy [Shannon1948]. Thus the CR is Shannon's entropy of a transformation or digestion of the sequence. The purpose of this transformation is to reveal the regularities in a sequence. In case the original sequence has no significant regularities, all numbers will be at the same rate; else, some numbers will be in excess and others in shortage (compression algorithms use this to obtain a short output). Shannon's entropy is zero -this is the minimum- only when all numbers are zero. This occurs when a sequence consists just of a single repeated symbol, which is the simplest possible combinatorial structure. When, at the other edge, entropy is equal to one (the maximum entropy), then all numbers have exactly the same frequency, and it indicates that the sequence has a random-like combinatorial structure. Algorithmically, the BWT of a given sequence is a permutation of the symbols in the sequence that represents the lexicographic order of all possible rotations of the sequence. The MTF transforms a given sequence into a sequence of numbers, operating from left to right, and maintaining a stack of recently used symbols. Each number is an index in the stack and denotes an alphabet symbol. Shannon's entropy maps a sequence into a real number between zero and one. It weights the frequency of the alphabet symbols in a given sequence. For each symbol i in the alphabet, let $p_{(i)}$ be the probability of finding i in the sequence s; N_i

the number occurrences of i in s and $length(s)$ the total length of the sequence s:

$$p_{(i)} = \frac{N_i}{length(s)} \tag{2.1}$$

For DNA alphabet entropy is defined as:

$$E(s) = -\sum_{i=0}^{\exists} p_{(i)} \times log_4(p_{(i)}) \tag{2.2}$$

Thus the CR can be factorize as:

$$CR(s) = E(MTF(BWT(s))) \tag{2.3}$$

The complexity value (CV) of a sequence is its CR times the number of characters in this sequence (here s):

$$CV(s) = E(MTF(BWT(s))) \times length(s) \tag{2.4}$$

As the CV of a sequence depends on the transformation of the MTF applied to the whole sequence, its computation impede the use of parts of the sequence independently.

2.3.2. Complexity in strings

Complete genomes of 54 species were download from NCBI and Ensembl Genome Project [Flicek *et al.*2011]. Fourteen major groups of taxa were selected: virus, phages, bacteria, archaea, fungi, amplicomplexa, heterokonta, amebozoa, urochordates, invertebrates, plants, fishes, birds, and mammals. Species among taxa were chosen to the interest as model species and the presence of particular biological features such as: variation in genome size, ancestral and recent polyploidy, living in extreme environments, living as intracellular parasites, gene expansion, genome reduction, RNA or single-strand DNA genomes, and synthetic genomes Table 2.1. Eukaryote genomes with coverage of 6× or greater were chosen. Sexual chromosomes were excluded from the analysis, and ambiguous "N" characters were removed from sequences, and not taken into account when computing chromosome length. Eukaryote chromosomes were concatenated in genomes to estimate genome complexity. Interspersed repeats and low complexity DNA sequences were screened and mapped in chromosomes of thirty different eukaryotes using RepeatMasker [Smit *et al.*2010]. Complexity of major families of repetitive elements such as DNA transposons, LTR, LINE, SINE, satellites and exons, introns, and complete genes (considering unstranslated regions) was computed after concatenation of all elements in chromosomes excluding. Random sequences with different ploidy levels were generated in python. Complexity value of biological sequences and random sequences was computed with the DNA alphabet of four letters. Complexity in biological sequences was computed in the +1 strand. Analyses of -1 strand provided no differences in results. Short stories, books and complete works in its original languages were downloaded from Project Gutember (`http://www.gutenberg.org/`). To automatically detect the alphabet size

in texts (including mathematical and punctuation symbols) we run COMPL program with "auto" option. To study complexity along chromosomes, a sliding window method shifting along chromosomes in overlapping units of 1.0 Kb to 100 Mb was performed. Linear models, and linear models with interactions were run in R language [Team2011].

2.3.3. Simulations

We performed four kinds of experiments where complexity value and ratio were computed. First: random polyploid construction of sequences of various sizes and ploidy levels (one to ten). Second: the evolution along 40 million generations by constant neutral mutation rate of 1.0e-08 mutations per site per generation (this value is in between the mutation rate estimated for *Homo sapiens*: 2.5e-08 [Nachman & Crowell2000] , and *Arabidopsis thaliana*: 7.1e-09 [Ossowski *et al.*2010]) on random sequence, and on chromosomes of *Zea mays* and *Sorghum bicolor*. Third: the evolution along 50,000 generations of random polyploid genomes of different sizes (100Kb, 1Mb, 10Mb) by transpositions of a fixed length (1.0 Kb) between chromosomes. The number of transposition per generation was set as a constant function of genome size (genome size/1,000). Last: the concatenation and shuffling (computed with the python base function: "shuffle") of all repetition instances in chromosomes for main repetitive families, and genes were considered. Complexity value and ratio were computed every 100 generations.

2.4. Discussion

However it is striking that whatever subgroup of elements we selected (*Simple repeats*, *Satellites*...), no perfect random structure was found in any of the steps of life complexity. Sequences can be divided into modules of nucleotides as mentioned in [Wagner *et al.*2007] or defined by Herzel and collaborators [Herzel *et al.*1995] when it fulfill 3 requirements, concisely:

- **Requirement 1**: must occur more frequently than expectation by random.
- **Requirement 2**: must not be included into a larger module.
- **Requirement 3**: must be non-overlapping between them.

3. Life inside genomes, dynamics and predictions

3.1. Genomic elements, dispersion and abundance

3.2. Species Abundance Diversity in genomes

3.3. Neutrality of SAD

3.4. Material Methods

3.4.1. Ecology

Part II.

Detection of selective pressures in genomes

4. Searching for evolutionary patterns in funcionally linked group of genes

4.1. Background

The analysis of adaptation at large or complete genome scale is currently based on concepts and methods developed for single genes analyzes [Arbiza *et al.*2006, Bakewell *et al.*2007, Bustamante *et al.*2005, Clark *et al.*2003, Nielsen *et al.*2005]. Statistical methods to test for neutrality [Nielsen2001], are currently used without considering if genes works independently or associated to others to produce a single phenotypic response. In this sense we are applying pre-genomics concepts and methods to genomics data. The current paradigm for large scale analysis of adaptation consists in a two steps framework: first, the search for a list of genes (in a gene-by-gene framework analysis) with a statistical significant signal of positive selection ($\omega > 1$), and second, the search for over-represented functional classes of genes in this list. Although it is logically consistent, it has been noted that this kind of strategy causes an enormous loss of information due to the large number of false negatives that are accepted in order to preserve a low ratio of false positives necessary when genomics data is considered [Al-Shahrour *et al.*2007, Al-Shahrour *et al.*2005, Al-Shahrour *et al.*2006, Subramanian *et al.*2005].

Genes do not operate alone within the cell, but in a intricate network of interactions that we have only recently started to envisage [Stelzl *et al.*2005]. It is a widely accepted fact that coexpressing genes tend to be fulfilling common roles in the cell [Lee & Sonnhammer2003]. Moreover, coexpression seems to occur, in many cases, in contiguous chromosomal regions [Caron *et al.*2001] and furthermore, recent evidences suggest that functionally related genes map close in the genome, even in higher eukaryotes [Hurst *et al.*2004]. Many higher-order levels of interaction are continuously being discovered and even complex traits, including diseases, have started to be considered from a systems biology perspective [Ideker & Sharan2008, Vamathevan *et al.*2008].

Recent methodology was proposed to circumvent the classical two-step analysis as a new attempt to test for selective signatures across species at genome-scale level [Shapiro & Alm2008] Using the deviations of the expected rates of evolution for a large group of genes in a group of gamma proteobacteria, the authors conclude that the coherence of selective patterns suggests that the genomic landscape is organized into functional modules even at the level of natural selection.

The hypothesis we aim to test in this study is not about individual genes, but about functional classes. Mutations occur on single genes but natural selection acts on phenotypes by operating on whole sub-cellular systems. Mutations in genes either remain finally fixed or disappear because of their beneficial or disadvantageous effect on indi-

vidual fitness, respectively. This effect on the function of individual proteins can only be understood in the context of the system (e.g. a pathway, GO functional roles, etc.) in which the proteins are involved. If a list of genes arranged by some parameter that accounts for their evolutionary rates is examined, it is expected that genes belonging to pathways or functional classes favored or disfavored by selection will tend to appear towards the extremes.

This approach circumvents the implicit assumption posed by the two-step analysis described above assuming by that the gene is the only target of selection. If natural selection works by means of minor quantitative effects of many different changes distributed along different gene products most of them working together in a few number of systems (GO functional terms, biochemical pathways and/or interactome modules) we expect to find: 1- correlated nonsynonymous rate changes associated to these functions , 2- synonymous rate changes not necessarily associated to the same functions, 3- a higher number of significant functions than those discovered in the classical two-step approach.

In the first part of this paper we extend the classical two-step approach previously reported by us for human and chimp [Arbiza *et al.*2006], to rat and mouse now considering a set of XXXX orthologous genes of human, chimpanzee, mouse, rat and dog. The objective is to compare the classical two-step approach with the new system approach developed in the second part of the paper. In both cases we search for differences in evolutionary rates differentiating positive selection from relaxation along the branches of the phylogeny of the species.

4.2. Results

4.2.1. Gene-set selection analysis on functional modules

Mammals, represented by human, chimpanzee, rat and mouse, and five Drosophila genomes were studied. For each species, genes were ranked into four lists according to the estimation of i- synonymous (dS), ii- nonsynonymous (dN) rates of substitution, iii- selective pressures (ω = dN/dS), and iv- the change of selective pressures between (A) ancestor and (D) descendant species ($\Delta\omega$D = ωD-ωA) along the phylogeny Figure 4.2. Maximum likelihood (ML) estimates of evolutionary variables were performed using a free-ratio branch model [Yang2007]. As such, four lists containing 12,543 and 9,240 orthologous genes in mammals in Drosophila species were obtained for the analyses, respectively. GSSA was conducted using a total of 1,394/199 and 1,331/116 GO/KEGG terms in mammals and Drosophila species respectively. GSSA is performed in five different steps (S1 to S5 in Figure 4.3 on page 32 in Section 4.3). First, the method ranks all genes within a genome (G) according to one of the alternative evolutionary variables (dS, dN, ω and $\Delta\omega$). Second, genes are associated (dark dots) to different functional categories (GO or any other functional term). Note that a single gene can be associated with multiple functions (yellow bar in Figure 2). Third, for each functional category a total of 30 partitions are established along the list of ranked values [Al-Shahrour *et al.*2007], [Al-Shahrour *et al.*2005]. Fourth, for each partition GSSA computes a two-

tailed Fisher's exact test and reports significant over or under represented functional classes comparing the upper side (A) and the lower side (B) of the list. Finally, p-values are corrected for multiple testing (FDR). Throughout the manuscript only p-values for partitions with the highest confidence were reported after FDR.

The application of GSSA to lists of genes ranked by dS, dN, ω and the $\Delta\omega$ values yielded a large number of functional modules (defined by GO and KEGG annotations) with rates that were significantly skewed toward the extremes of the lists (Table 4.1) in mammal and *Drosophila* species. For instance, 11% of GO terms, and 15% of KEGG pathways contain genes with biased distribution of rates towards the top of the ranked list, and found statistically significant at high ω ratio (SHω, 5% false-discovery rate, FDR) in mammals. Alternatively, 4.1% and 2.6% of GO terms and KEGG pathways were found with significantly high values of ω (SHω) in *Drosophila*, respectively.

		SH*		**SL***	
		KEGG	**GO**	**KEGG**	**GO**
Mammals	dS	15 (1.9)	187 (3.3)	12 (2.1)	364 (6.5)
	dN	145 (18.2)	708 (12.6)	230 (28.9)	1,839 (32.9)
	ω	123 (15.5)	649 (11.6)	206 (25.9)	1,675 (30.0)
	$\Delta\omega$	64 (8.0)	421 (7.5)	107 (13.4)	818 (14.7)
Drosophilas	dS	18 (3.1)	104 (1.5)	26 (4.5)	1,263 (18.9)
	dN	31 (5.3)	276 (4.1)	26 (4.5)	2,097 (31.5)
	ω	15 (2.6)	213 (4.1)	24 (4.1)	1,321 (19.8)
	$\Delta\omega$	2 (0.3)	143 (2.1)	7 (1.2)	184 (2.8)

Table 4.1.: Numbers and percentages of functional modules with significant results after GSSA. For Significantly High (SH) and Significantly Low (SL) results.

Table 4.1 also reveals that functional modules with genes changing at significantly low ω ratios (SLω), and therefore showing a distribution shifted towards the bottom of the ranked list (see Figure 2), were more frequent than modules under the significantly high ω (SHω). This observation is in agreement with the fact that purifying selection is the predominant form of selection in biological systems. Moreover, in support of the slightly neutral character of synonymous mutations, and the effects of population size in the final outcome of selection [Lynch2007] GSSA results show a higher number of significant deviations of dS in *Drosophila* rather than in mammals.

Only a minor proportion of functional terms changed significantly at higher or lower rates relative to estimates of the corresponding ancestral lineages. Specifically, increased or decreased ω values on the external branches (recorded by positive and negative values of $\Delta\omega$) were observed for only half of the cases where a significant increase or decrease of ω was identified in mammals and Drosophilas. This observation points out the conservative character of the selective constraints in functional related groups of genes during evolution.

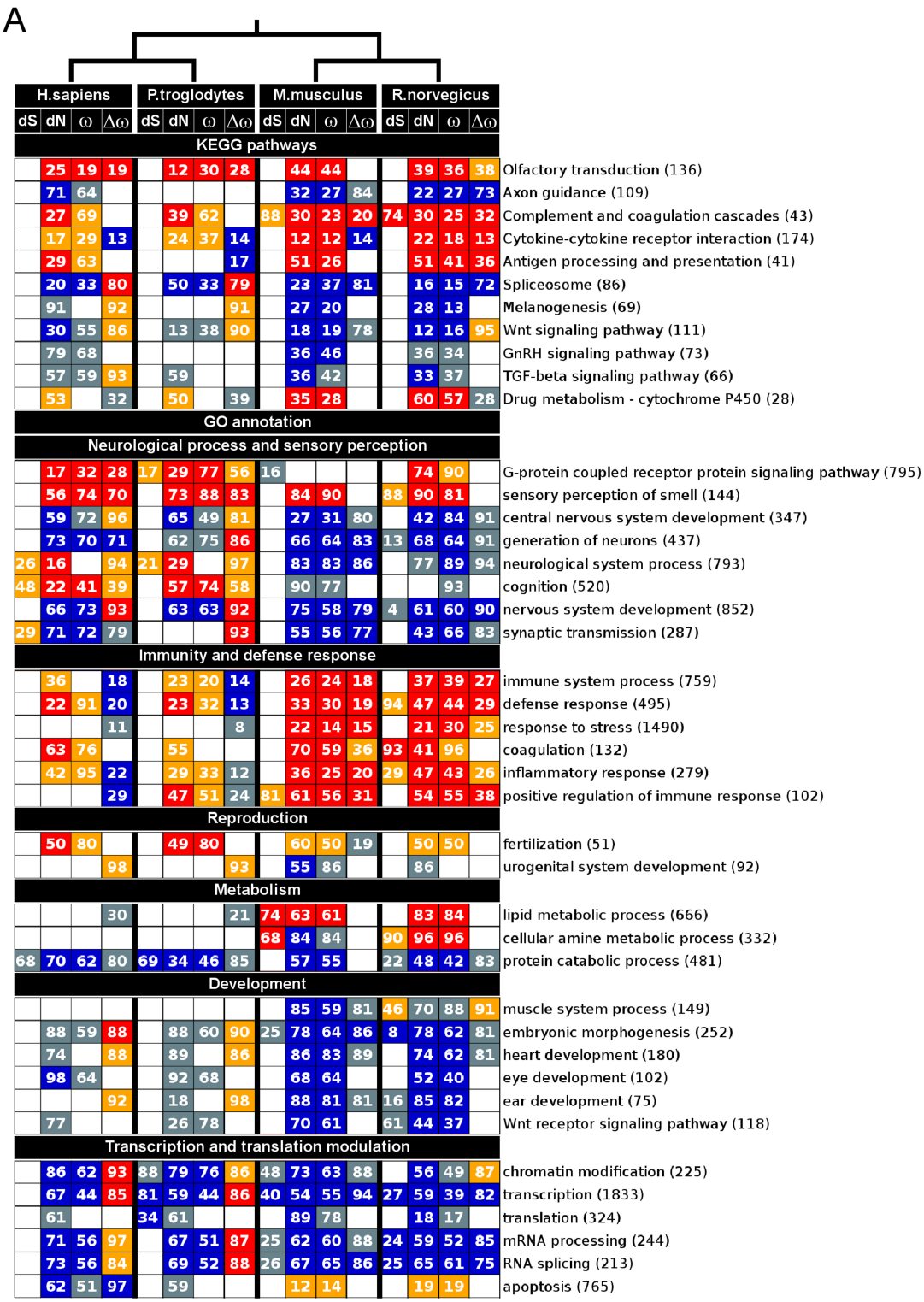
A summary of the results of the GSSA for mammals and *Drosophilas* is shown in Figure 4.1 (see Figures S1 to S4 for a complete description of results after GSSA in

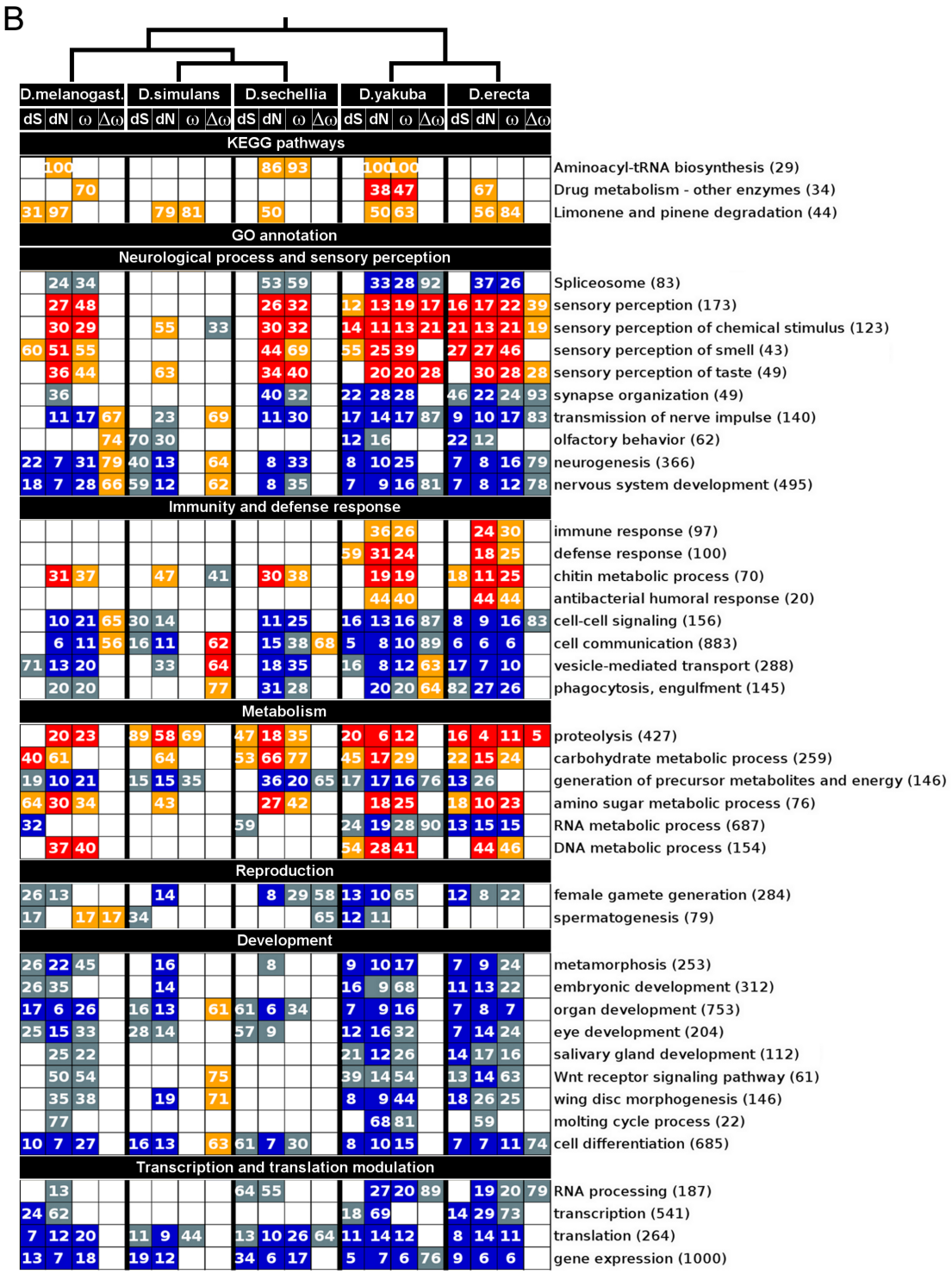
mammals and *Drosophila* species). The figure shows that GSSA has the power to detect many functional changes in evolutionary rates within a substantial number of functional categories. Although the rough pattern shows similar evolutionary constraints in groups of genes between the two main clusters of species, important differences were also detected within them. For instance, functional terms associated to neurological process and sensory perception clearly contrasted between primates and rodents (Figure 4.1-A). While most of these terms are associated to a significant relative increase in rates from the common ancestor of primates ($+\Delta\omega$), all the changes observed in rodents were due to the relative increase of the selective constraints ($-\Delta\omega$) probably due to the effects of purifying selection from the common ancestor. Alternatively, functional modules associated to Immunity and Defense response evolved at significantly higher rates than expected in rodents, but decreased significantly in relation to the ancestral rates in primates. Such functional differences between primates and rodents were previously observed when pooling groups of species [Kosiol *et al.*2008]. Other functional modules such as *Development*, and *Transcription/Transduction* comparatively evolved at very low dN and ω ratio but experienced a higher relaxation of the ancestral constraints ($+\Delta\omega$) in primates than in rodents. Moreover, significant differences in rates can be detected between human and chimpanzee (Ha04360: *Axon guidance*, Ha04610: *Antigen processes and presentation*, GO0007268: *synaptic transmission*, among others), and between mouse and rat (GO0007186: *G-protein coupled receptor protein signaling pathway*, and Ha04310: *Wnt signaling pathway*, among others).

In addition, most of the GO terms significantly associated to high dN and Ï in Drosophilas were unevenly distributed within the two clusters of the phylogeny (Figure 4.1-B). GO terms such as sensory perception, defense response, immune response and metabolic process, among others, presented a remarkable divergence in the monophyletic groups of *D. erecta* and *D. yakuba* but they were not observed in *D. sechellia*, *D. melanogaster* and *D. simulans*. Most of GO terms from *Development*, *Transcription* and *Translation* (Figure 4.1-A and -B) were significantly accumulated towards the extremes of the lists corresponding to the lowest rates of substitutions, suggesting they are significantly constrained by strong purifying selection (5% FDR) in both taxa.

The fact that most of the functional modules under selection (SHω and SLω) correlate with changes in dN, suggests that selective pressures are mainly driven by nonsynonymous rather than by synonymous substitutions during evolution. Moreover, according to the expectation of the nearly neutral theory, a low but still considerable number of

Figure 4.1. *(following page)*: GSSA of evolutionary variables.
The figure shows a selection of GO terms and KEGG pathways with significant and not significant deviations after GSSA of evolutionary rates in mammals (A) and Drosophila (B) species. Colored boxes represent functional modules with genes significantly accumulated at the corresponding extremes of the ranked list as explained in Figure 2. The number inside each box represents the percentage of the total number of genes of the functional module (in parenthesis) that contribute to its significance. Here we reported the numbers of the first significant partition after FET and FDR. Topologies represent the phylogenetic relationships of species.


A
H.sapiens
P.troglodytes
M.musculus
R.norvegicus
dS dN ω Δω
KEGG pathways
Olfactory transduction (136)
Axon guidance (109)
Complement and coagulation cascades (43)
Cytokine-cytokine receptor interaction (174)
Antigen processing and presentation (41)
Spliceosome (86)
Melanogenesis (69)
Wnt signaling pathway (111)
GnRH signaling pathway (73)
TGF-beta signaling pathway (66)
Drug metabolism - cytochrome P450 (28)
GO annotation
Neurological process and sensory perception
G-protein coupled receptor protein signaling pathway (795)
sensory perception of smell (144)
central nervous system development (347)
generation of neurons (437)
neurological system process (793)
cognition (520)
nervous system development (852)
synaptic transmission (287)
Immunity and defense response
immune system process (759)
defense response (495)
response to stress (1490)
coagulation (132)
inflammatory response (279)
positive regulation of immune response (102)
Reproduction
fertilization (51)
urogenital system development (92)
Metabolism
lipid metabolic process (666)
cellular amine metabolic process (332)
protein catabolic process (481)
Development
muscle system process (149)
embryonic morphogenesis (252)
heart development (180)
eye development (102)
ear development (75)
Wnt receptor signaling pathway (118)
Transcription and translation modulation
chromatin modification (225)
transcription (1833)
translation (324)
mRNA processing (244)
RNA splicing (213)
apoptosis (765)


B
D.melanogast.
D.simulans
D.sechellia
D.yakuba
D.erecta
dS dN ω Δω
KEGG pathways
Aminoacyl-tRNA biosynthesis (29)
Drug metabolism - other enzymes (34)
Limonene and pinene degradation (44)
GO annotation
Neurological process and sensory perception
Spliceosome (83)
sensory perception (173)
sensory perception of chemical stimulus (123)
sensory perception of smell (43)
sensory perception of taste (49)
synapse organization (49)
transmission of nerve impulse (140)
olfactory behavior (62)
neurogenesis (366)
nervous system development (495)
Immunity and defense response
immune response (97)
defense response (100)
chitin metabolic process (70)
antibacterial humoral response (20)
cell-cell signaling (156)
cell communication (883)
vesicle-mediated transport (288)
phagocytosis, engulfment (145)
Metabolism
proteolysis (427)
carbohydrate metabolic process (259)
generation of precursor metabolites and energy (146)
amino sugar metabolic process (76)
RNA metabolic process (687)
DNA metabolic process (154)
Reproduction
female gamete generation (284)
spermatogenesis (79)
Development
metamorphosis (253)
embryonic development (312)
organ development (753)
eye development (204)
salivary gland development (112)
Wnt receptor signaling pathway (61)
wing disc morphogenesis (146)
molting cycle process (22)
cell differentiation (685)
Transcription and translation modulation
RNA processing (187)
transcription (541)
translation (264)
gene expression (1000)

significant associations of functional modules to dS were found in *Drosophila* (19.5%) and rodents (11.3%), while in primates (6.4%), where population sizes are known to be smaller, the number of significant modules was smaller [Petit & Barbadilla2009].

The strategy presented here lead to detect significant patterns of increments and decrements modeled by natural selection in evolutionary rates of functional groups of genes. This pattern is consistent with the hypothesis that natural selection acts on phenotypes by the combined action of many functional related genes. Moreover, this functionally based approach identified with statistical significance, and on individual species, all the functional modules previously found significantly enriched by positively selected genes and therefore the main targets of adaptive biological functions in species (Table 2) (see Supplementary Table S3 for a complete list of terms). Although GSSA is not a test for positive selection, it is evident that functional modules containing PSGs can be significantly detected by this method on individual species. In the next section we will analyze the relative contribution of PSGs to the statistical differentiation of functional modules in genomes.

4.3. Material and Methods

4.3.1. Orthology prediction

Complete genomes of 5 mammals species (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus* and *Canis familiaris*) where retrieved from *Ensembl* [Flicek *et al.*2011]. Also orthology prediction between each pair of species possibly done between human and the others was retrieved from *Ensembl Compara* [Vilella *et al.*2009] using biomart [Kinsella *et al.*2011] and taking human as seed species. Only groups of orthologs *one-to-one* with one representative of each species where kept in the final dataset Figure 4.2-A.

The same procedure was applied for *melanogaster* group, including 6 species namely, *Drosophila melanogaster* (as seed-species), *Drosophila sechelia*, *Drosophila simulans*, *Drosophila yakuba*, *Drosophila erecta* and, as outgroup, *Drosophila ananassae* (see Figure 4.2-B).

4.3.2. Alignments refinement and filters

DNA coding sequences (CDS) were aligned according to protein translation pattern using *Muscle* version 3.7 [Edgar2004] embedded into the *CDS-Protal* utility in *Phylemon 2.0* [Sánchez *et al.*2011], and to avoid the presence of badly aligned regions alignments were cleaned using *TrimAl* [Capella-Gutiérrez *et al.*2009] keeping all sequences but trimming alignment columns with the heuristic method *automated-1*. Additionally, alignments smaller than 100 bp were excluded from the analysis.

In mammals, the upper limit for dN and dS considered was those of the human interferon γ (dN = 3.06) and the relaxin protein [Graur & Li2000] (dS = 6.39 substitutions per site per 1e9 years). Assuming the human-mouse, mouse-rat and human-chimp differentiation times to be about 80, 70 and 5 million years [Blair Hedges & Kumar2003], re-

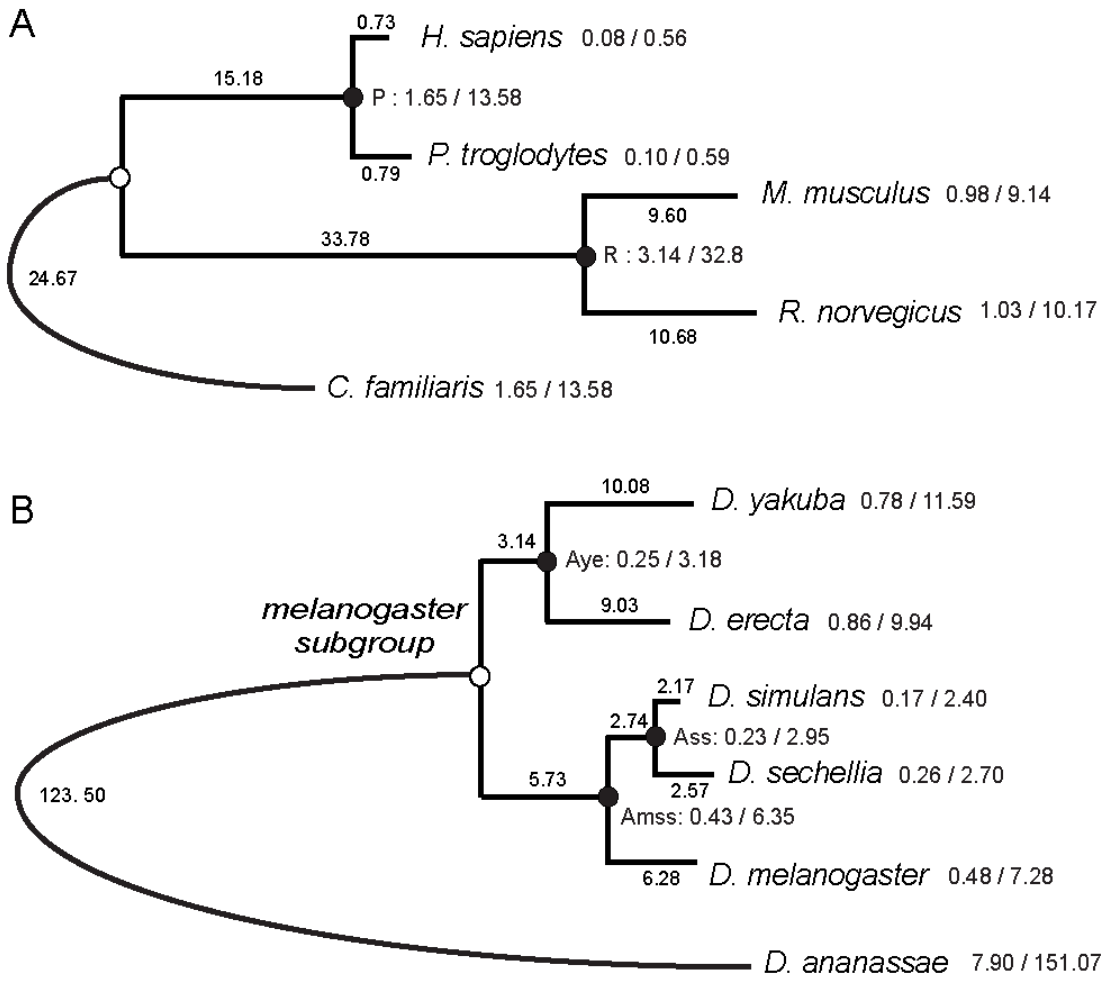


Figure 4.2.: Mammals and *melanogaster* group phylogeny.
Numbers on internal and external nodes represent the median number of nonsynonymous and synonymous substitutions per codon (dN/dS) estimated from all the coding sequences compared in mammal (A) and Drosophila (B) genomes. Branch lengths and rates were multiplied by 100. Ancestral estimation of parameters was done in primates (P), rodents (R), D. yakuba and D. erecta (Aye), D. simulans and D. sechellia (Ass), and D. melanogaster, D. simulans and D. sechellia (Amss). C. familiaris and D. ananassae were chosen as outgroup species in the corresponding tree.

spectively, ortholog comparisons between primates and rodents with dS≥1 and dN≥0.5, rodents with dS≥0.256, dN≥0.122, and primates with dS≥0.064 and dN≥0.030 substitutions/site were excluded.

The number of orthologs kept for analysis after filtering steps, is 12,453 for mammals, and 9,240 for flies.

4.3.3. Evolutionary analysis

Maximum likelihood estimation of dN, dS, and ω was computed using CodeML program from PAML [Yang2007]. Evolutionary rates were computed in orthologous sequences according to the free-ratio branch model assuming independent ω ratio for each branch of the tree of mammals and Drosophila species (see raw values of rates in Table S1 and S2). Evolutionary rates (dN, dS), its ratio (ω), and its difference between ancestral and descendant species ($\Delta\omega$) were ranked along all genes of genomes and further analyzed by GSSA.

External branches of Figure 1 were labeled as foreground to test for positive selection using branch-site models in Test I and Test II [Zhang *et al.*2005]. Positive results of relaxation of selective constraints (or weak signals of positive selection) were discarded [Arbiza *et al.*2006]. To quantify the relative contribution of PSGs in functional modules showing SHω and SLω results in GSSA, a t-test (from R package [Ihaka & Gentleman1996]) with the mean number of PSGs per functional modules was computed in primates, rodents, mammals and Drosophila species. An independent set of PSGs was collected to test the robustness of our results in mammals [Kosiol *et al.*2008], and Drosophila species [Clark *et al.*2007].

4.3.4. GSSA, evolutionary and statistical simulations

Gene-set selection analysis across lists of genes ranked by different evolutionary rate parameters (dS, dN, ω and $\Delta\omega$) was computed using the program Babelomics [Al-Shahrour *et al.*2008]. This program implements a version of GSA [Al-Shahrour *et al.*2005] which can be applied to any list of ranked genes regardless of the initial experimental design [Dopazo,Huang *et al.*2009]. The aim of the test is to find functional classes, namely blocks of genes that share some functional property, showing a significant asymmetric distribution towards the extremes of a list of ranked genes. This is achieved by means of a segmentation test, which consists on the sequential application of a Fisher's exact test over the contingency tables formed with the two sides of different partitions (A and B in Figure 4.3) made on an ordered list of genes. The two-tailed Fisher's exact test finds significantly over or under represented functional classes when comparing the upper side to the lower side of the list, as defined by any partition (in Figure 4.3, four of the five partitions show significant differences). Similarly to other equivalent gene-set analyses, the outcomes are those modules (GO and KEGG) significantly associated to high or low values of the evolutionary parameter used to rank the genes. Previous results showed that a number between 20 and 50 partitions often gives optimal results in terms of sensitivity and results recovered [Al-Shahrour *et al.*2005]. Here we applied 30 partitions

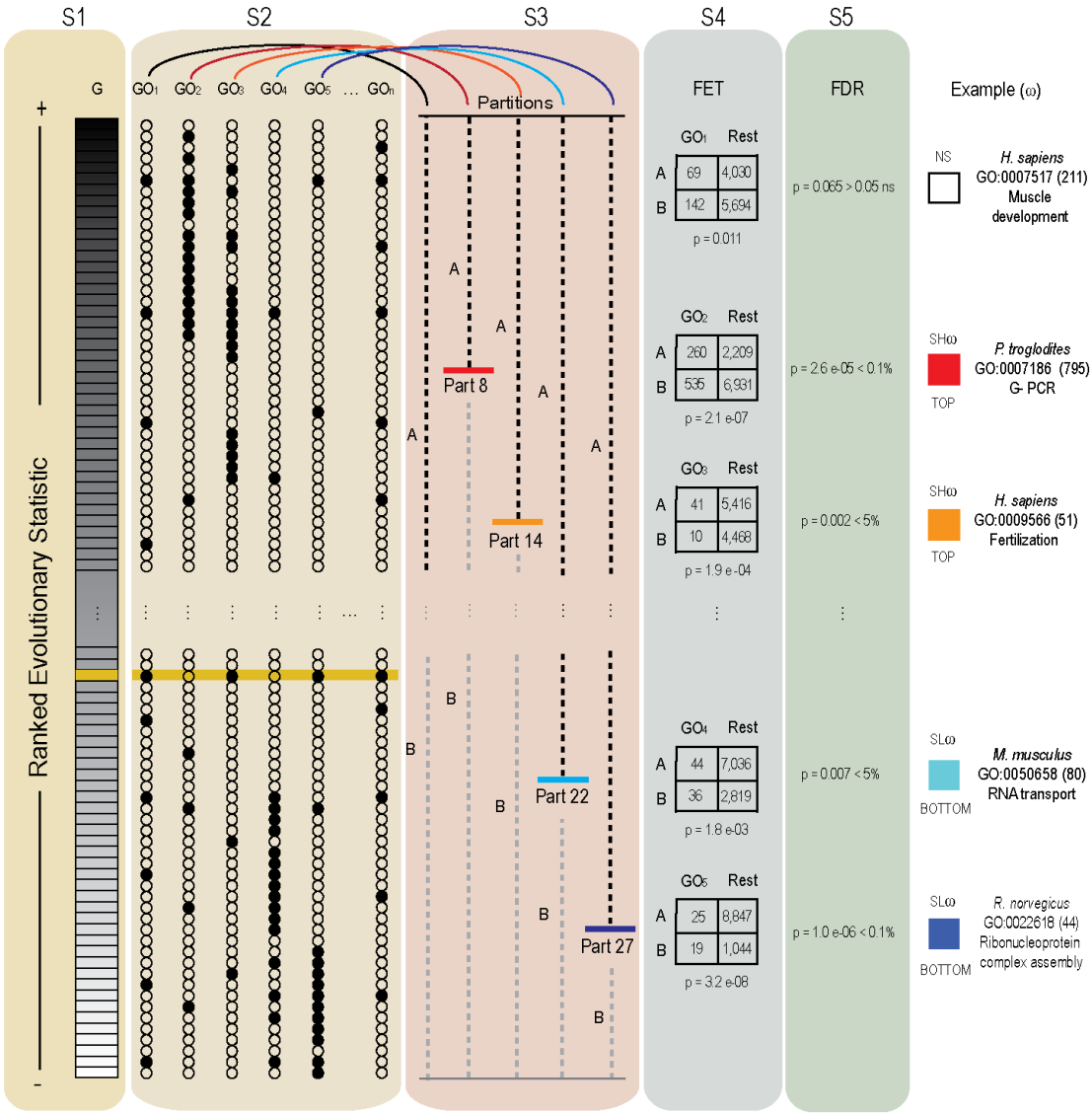
along all the GSSA performed. Given that multiple functional classes (C) are tested in multiple partitions (P), the unadjusted p-values for a total of C · P tests were corrected by the widely accepted FDR method [Benjamini *et al.*2001].

Originally, 1,394/1,331 GO terms, and 199/116 KEGG pathways were analyzed in mammals and Drosophila species respectively. The global GO directed acyclic graph was processed with Blast2GO [Conesa *et al.*2005] to extend the annotation at missing parental nodes, discarding GO levels out of 2 to 8 for mammals, and 2 to 12 for Drosophilas. The final set of GO and KEGG terms used in the GSSA corresponds to those containing a minimum number of 15 genes. To test possible biases attributed to the size of the functional category, the magnitude of change in evolutionary rate or the proportion of genes experiencing a rate change we randomized the original assignation of ENSG's to the list of ranked values and functional annotation (see Figure 4.4-A). For each evolutionary variable and species 10.000 randomizations and the corresponding GSSA were performed. The proportion of false positives (significant results after GSSA) was computed for each evolutionary variable and plotted along the size of functional categories (from 20 to 1,400 with intervals of 20). Because this proportion never reached values higher than 0.5% (FDR) we rejected the possibility that either group size or rate distribution biased GSSA results in our data set (see Figure 4.4-B and Figure 4.4-C).

Finally, in order to validate the independence of the GSSA from the effects of alternative evolutionary constraints we simulated selective regimes (purifying selection, positive selection and relaxation of selective constraints) using branch-site models. Here we addressed the possibility of a variation in the representation of significant results after GSSA (see Figure 4.5). The pipeline described here, shows three different areas:

- **Real Data**: the dark yellow area describes the steps used to reach to results

Figure 4.3. *(following page)*: Summary of the steps developed by the GSSA.
GSSA can be roughly described in a series of five steps (S1 to S5). S1: rank genes of a genome according to an evolutionary variable, S2: assign functional classes to all the listed genes, S3: apply a fixed number of partitions on the ranked list, S4: proceeds with a Fisher exact test (FET) for each partition, S5: adjust p-values by FDR. See text for a full description. Colored boxes (red, orange, cyan and blue) represent functional modules with genes significantly accumulated (0.1% FDR and 5% FDR) at the corresponding extremes of a list (top and bottom), and therefore with significantly high (SH) and low (SL) values of the evolutionary variable (ω) respectively. White represents a non-significant association (NS). Examples show five alternative GO categories with significant and non-significant distributions of the ω statistic. In parenthesis, the total number of genes corresponding to the GO term is shown. For GO1, the function seems to be uncorrelated with the arrangements of the genes. In the example (GO:0007517) partition 16 in human (not shown in the picture) reported the lowest p-value (p = 0.011) although it was not significant after FDR correction (FDR = 0.065). Upper (A) and lower (B) sides of the ranked list (S3) represent both sides of the specified partition number. Remainder GO categories (GO2 to GO5) show the association of dark dots with values located at the top (significant high ω values -SHω), and at the bottom (significant low ω values -SLω) of the list (for GO2-GO3 and GO4-GO5, respectively). In examples, FETs found the most significant p-value for partitions 8, 14, 22 and 27 for GO:0007517, GO:0007186, GO:0009566, GO:0050658 and GO:0022618 in chimpanzee, human, mouse and rat genome, respectively.


S1
S2
S3
S4
S5
G
GO1 GO2 GO3 GO4 GO5 … GOn
Partitions
Ranked Evolutionary Statistic
+
-
Part 8
Part 14
Part 22
Part 27
A
B
FET
FDR
Example (ω)
GO1 Rest
A 69 4,030
B 142 5,694
p = 0.011
GO2 Rest
A 260 2,209
B 535 6,931
p = 2.1 e-07
GO3 Rest
A 41 5,416
B 10 4,468
p = 1.9 e-04
GO4 Rest
A 44 7,036
B 36 2,819
p = 1.8 e-03
GO5 Rest
A 25 8,847
B 19 1,044
p = 3.2 e-08
p = 0.065 > 0.05 ns
p = 2.6 e-05 < 0.1%
p = 0.002 < 5%
p = 0.007 < 5%
p = 1.0 e-06 < 0.1%
NS
H. sapiens
GO:0007517 (211)
Muscle development
SHω
TOP
P. troglodites
GO:0007186 (795)
G- PCR
SHω
TOP
H. sapiens
GO:0009566 (51)
Fertilization
SLω
BOTTOM
M. musculus
GO:0050658 (80)
RNA transport
SLω
BOTTOM
R. norvegicus
GO:0022618 (44)
Ribonucleoprotein complex assembly

described in the manuscript. The light yellow area describes the use of the CodeML program from PAML package (reference 15 in the ms) to extract -from the original set of sequences -the evolutionary parameters to simulate new sequences under purifying selection (PF), positive selection (PS) and relaxation of the selective constraints (RX) using branch-site models (see model description below). Human, mouse, *D. erecta* and *D. melanogaster* were used as foreground species in the corresponding models.

- **Simulated Data**: Evolver (PAML program) simulates sequences using parameters (codon frequencies and branch lengths) from the empirical data. We checked the desired characteristics of positive selection (PS) and relaxation of selective constraints (RX) on the set of the simulated sequences Table 4.2. Evolutionary variables (dS, dN, ω and $\Delta\omega$) were estimated from simulated sequences by means of a free-ratio branch model (CodeML). The complete pipeline of the GSSA was applied in the simulated data.

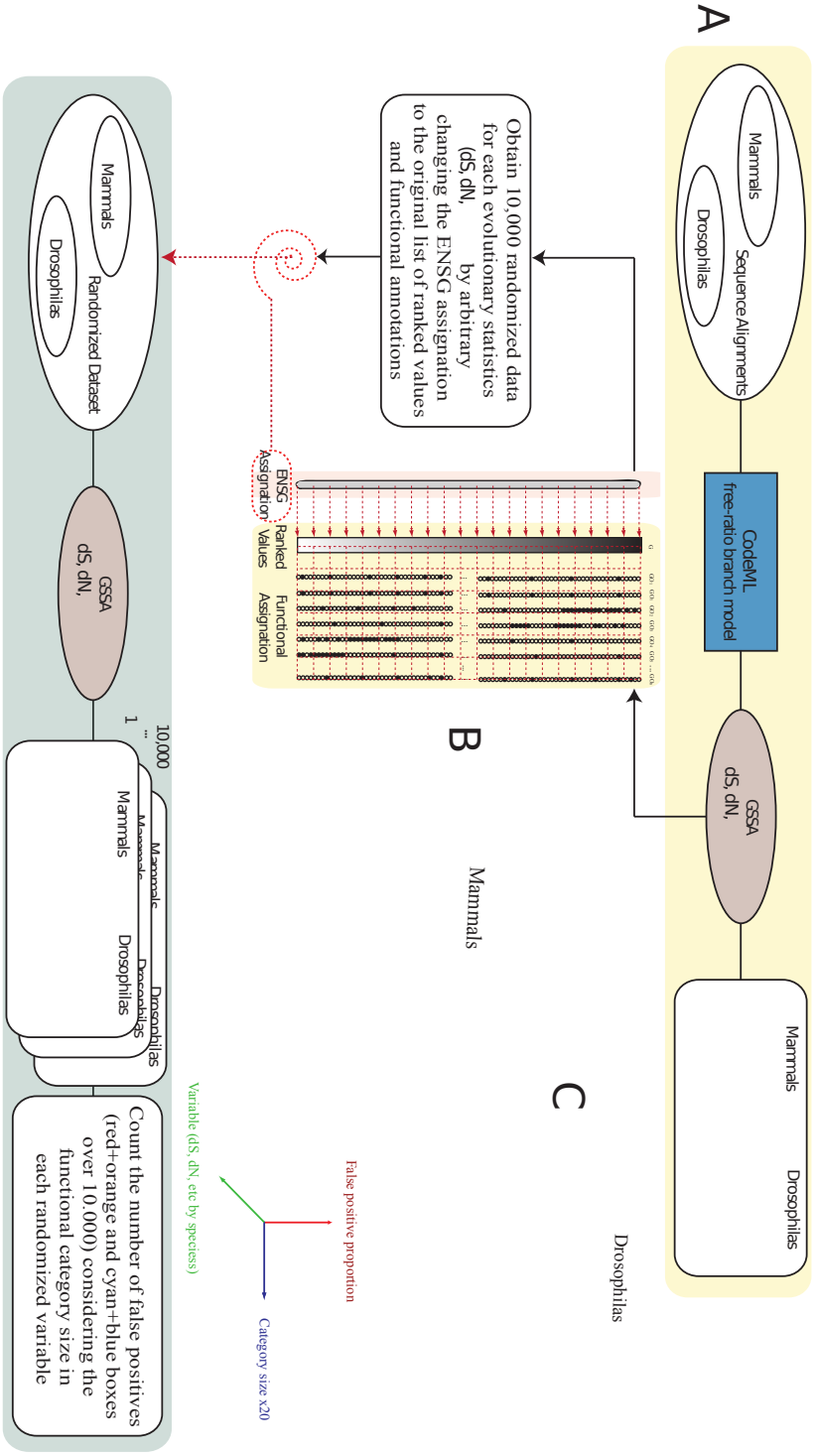
- **Testing simulation**: The odd-ratio of the values observed on the contingency table of each significant functional term after GSSA was computed. Values higher and lower than one contribute to the total number of functional modules with significant high and low ω values. To test the statistical contribution of these functional modules to these extremes on the simulated regimes (PS, RX and PF) the log odd-ratios were compared using t-test.

	PS		RX		PF	
	# PSG	# RXG	# PSG	# RXG	# PSG	# RXG
Homo sapiens	658	1640	11	1939	0	1
Mus musculus	1500	954	14	1565	1	0
D. melanogaster	736	630	25	1104	0	0
D. erecta	778	1292	26	1713	2	1

Table 4.2.: Number of PSG and relaxed genes (RXG) in each of the simulated evolutionary scenarios.

Our results showed that in spite of the alternative evolutionary scenarios no significant differences were observed between log odd-ratios distribution (p¡0.05). This result is exactly what we expected. The average effect of PF, and RX-PS is the proportional decrease and increase of the mean value of ω on sequences, respectively. This change has minor effects (if any) in the relative position of genes in the ranked list of genes of

Figure 4.4. ***(following page)*****: Randomisation experiment.**
(A) The pipeline shows the steps followed to tests possible biases attributed to the size of the functional category, the magnitude of change in evolutionary rate and the proportion of genes experiencing a rate change in the GSSA. The proportion of false positive results never reached 5% (FDR) in mammals (B) and Drosophila (C).


A
Mammals
Drosophilas
Sequence Alignments
CodeML
free-ratio branch model
GSSA
dS, dN,
Mammals
Drosophilas
Obtain 10,000 randomized data
for each evolutionary statistics
(dS, dN, by arbitrary
changing the ENSG assignation
to the original list of ranked values
and functional annotations
ENSG
Assignation
Ranked
Values
Functional
Assignation
B
Mammals
Drosophilas
C
False positive proportion
Category size x20
Variable (dS, dN, etc by species)
Mammals
Drosophilas
Randomized Dataset
GSSA
dS, dN,
10,000
1
Mammals
Drosophilas
Count the number of false positives
(red+orange and cyan+blue boxes
over 10,000) considering the
functional category size in
each randomized variable

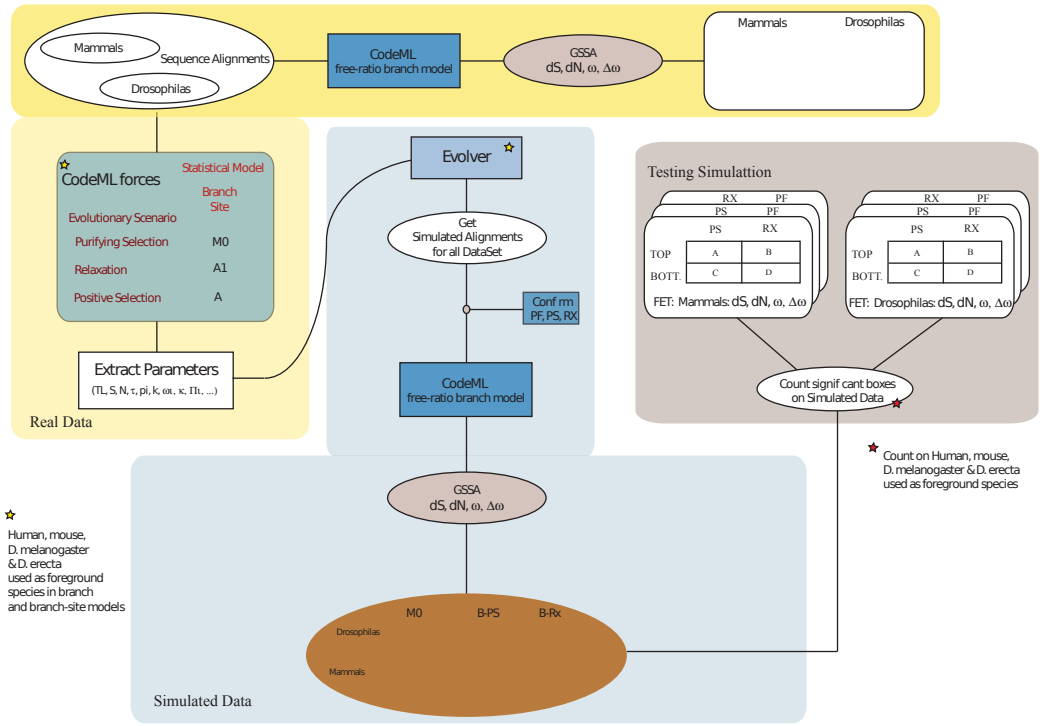


Figure 4.5.: Evolutionary and statistical simulation of GSSA.
The pipeline shows the steps taken along three different spaces of analysis, the real data, the simulated data and the testing block. See Supplementary Results for a complete explanation of methods and results.

a genome. Accordingly, since no net differences were produced after ranking genes, no significant differences are expected after the t-test (PS-RX: p= 0.99, PS-PF: p= 0.45, and RX-PF: p= 0.46). The fact that basically the same number of significant results was observed in each evolutionary scenario confirmed this prediction Table 4.3. We conclude that neither of the selective regimes simulated produce significant differences or biases in the GSSA of ω values.

	PS	RX	PF
PS	—	92.50%	98.50%
RX	91.10%	—	99.00%
PF	88.90%	90.60%	—

Table 4.3.: Proportion of significant functional categories that are still significant (identical signs of odd-ratios) under a different evolutionary scenario.

4.4. open on colocalization to not random

5. Tools, programs, methods

5.1. ETE-evol plugin

5.1.1. BRANCHED1

5.1.2. Protamines Rodents and Primates

5.2. Pipeline for study of adaptation at genomic scale

5.2.1. Selective pressure on duplicated genes in Drosophila

5.3. Phylemon

6. Conclusions

Bibliography

[Adjeroh *et al.*2008] Donald Adjeroh, Tim Bell, and Amar Mukherjee, The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. In *ACM SIGACT News*, vol. 41, 21–24. Springer US, Boston, MA, 2008.

[Al-Shahrour *et al.*2005] Fátima Al-Shahrour, Ramón Díaz-Uriarte, and Joaquín Dopazo, Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics (Oxford, England)* **21**(13) (2005), 2988–93.

[Al-Shahrour *et al.*2006] Fátima Al-Shahrour, Pablo Minguez, Joaquín Tárraga, David Montaner, Eva Alloza, Juan M Vaquerizas, Lucía Conde, Christian Blaschke, Javier Vera, and Joaquín Dopazo, BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic acids research* **34**(Web Server issue) (2006), W472–6.

[Al-Shahrour *et al.*2007] Fátima Al-Shahrour, Leonardo Arbiza, Hernán Dopazo, Jaime Huerta-Cepas, Pablo Mínguez, David Montaner, and Joaquín Dopazo, From genes to functional classes in the study of biological systems. *BMC bioinformatics* **8** (2007), p. 114.

[Al-Shahrour *et al.*2008] Fátima Al-Shahrour, José Carbonell, Pablo Minguez, Stefan Goetz, Ana Conesa, Joaquín Tárraga, Ignacio Medina, Eva Alloza, David Montaner, and Joaquín Dopazo, Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic acids research* **36**(Web Server issue) (2008), W341–6.

[Arbiza *et al.*2006] Leonardo Arbiza, Joaquín Dopazo, and Hernán Dopazo, Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS computational biology* **2**(4) (2006), p. e38.

[Bakewell *et al.*2007] Margaret A Bakewell, Peng Shi, and Jianzhi Zhang, More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**(18) (2007), 7489–94.

[Becher & Heiber2011] Verónica Becher and Pablo Ariel Heiber, On extending de Bruijn sequences. *Information Processing Letters* **111**(18) (2011), 930–932.

[Benjamini *et al.*2001] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **125**(1-2) (2001), 279–284.

[Blair Hedges & Kumar2003] S Blair Hedges and Sudhir Kumar, Genomic clocks and evolutionary timescales. *Trends in genetics : TIG* **19**(4) (2003), 200–6.

[Blanc & Wolfe2004] Guillaume Blanc and Kenneth H Wolfe, Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* **16**(7) (2004), 1667–78.

[de Bruijn1946] N G de Bruijn, A combinatorial problem. *Koninklijke Netherlands: Academe Van Wetenschappen* **49** (1946), 758–764.

[Burrows & Wheeler1994] Michael Burrows and David J Wheeler, A block-sorting lossless data compression algorithm. *Digital SRC Research Report* **124** (1994).

[Bustamante *et al.*2005] Carlos D. Bustamante, Adi Fledel-Alon, Scott Williamson, Rasmus Nielsen, Melissa Todd Hubisz, Stephen Glanowski, David M. Tanenbaum, Thomas J. White, John J. Sninsky, Ryan D. Hernandez, Daniel Civello, Mark D. Adams, Michele Cargill, and Andrew G. Clark, Natural selection on protein-coding genes in the human genome. *Nature* **437**(7062) (2005), 1153–1157.

[Capella-Gutiérrez *et al.*2009] Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)* **25**(15) (2009), 1972–3.

[Caron *et al.*2001] Huib Caron, B van Schaik, M van Der Mee, Frank Baas, Gregory Riggins, P van Sluis, M C Hermus, R van Asperen, Kathy Boon, P A Voûte, S Heisterkamp, A van Kampen, and R Versteeg, The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science (New York, N.Y.)* **291**(5507) (2001), 1289–92.

[Chaitin1975] Gregory J. Chaitin, A Theory of Program Size Formally Identical to Information Theory. *Journal of the ACM* **22**(3) (1975), 329–340.

[Clark *et al.*2003] Andrew G. Clark, Stephen Glanowski, Rasmus Nielsen, Paul D Thomas, Anish Kejariwal, Melissa A Todd, David M Tanenbaum, Daniel Civello, Fu Lu, Brian Murphy, Steve Ferriera, Gary Wang, Xianqgun Zheng, Thomas J White, John J Sninsky, Mark D Adams, and Michele Cargill, Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science (New York, N.Y.)* **302**(5652) (2003), 1960–3.

[Clark *et al.*2007] Andrew G. Clark, Michael B Eisen, Douglas R Smith, Casey M Bergman, Brian Oliver, et al., Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167) (2007), 203–18.

[Conesa *et al.*2005] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)* **21**(18) (2005), 3674–6.

[Dopazo] Joaquin Dopazo, Formulating and testing hypotheses in functional genomics. *Artificial intelligence in medicine* **45**(2-3) , 97–107.

[Du *et al.*2006] Jiang Du, Joel S Rozowsky, Jan O Korbel, Zhengdong D Zhang, Thomas E Royce, Martin H Schultz, Michael Snyder, and Mark Gerstein, A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics (Oxford, England)* **22**(24) (2006), 3016–24.

[Edgar2004] Robert C Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5) (2004), 1792–7.

[Flicek *et al.*2011] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, et al., Ensembl 2011. *Nucleic acids research* **39**(Database issue) (2011), D800–6.

[Gaut2001] B S Gaut, Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome research* **11**(1) (2001), 55–66.

[Gaut & Doebley1997] B S Gaut and J F Doebley, DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the United States of America* **94**(13) (1997), 6809–14.

[Gerstein *et al.*2007] Mark B Gerstein, Can Bruce, Joel S Rozowsky, Deyou Zheng, Jiang Du, Jan O Korbel, Olof Emanuelsson, Zhengdong D Zhang, Sherman Weissman, and Michael Snyder, What is a gene, post-ENCODE? History and updated definition. *Genome research* **17**(6) (2007), 669–81.

[Gibson *et al.*2010] Daniel G Gibson, John I Glass, Carole Lartigue, Vladimir N Noskov, Ray-Yuan Chuang, Mikkel A Algire, Gwynedd A Benders, Michael G Montague, Li Ma, Monzia M Moodie, Chuck Merryman, Sanjay Vashee, Radha Krishnakumar, Nacyra Assad-Garcia, Cynthia Andrews-Pfannkoch, Evgeniya A Denisova, Lei Young, Zhi-Qing Qi, Thomas H Segall-Shapiro, Christopher H Calvey, Prashanth P Parmar, Clyde A Hutchison, Hamilton O Smith, and J Craig Venter, Creation of a bacterial cell controlled by a chemically synthesized genome. *Science (New York, N.Y.)* **329**(5987) (2010), 52–6.

[Graur & Li2000] Dan Graur and Wen-Hsiung Li, *FUNDAMENTALS OF MOLECULAR EVOLUTION*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, second edi edition, 2000.

[Gregory2005] T Ryan Gregory, Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics* **6**(9) (2005), 699–708.

[Herzel *et al.*1995] Hanspeter Herzel, Werner Ebeling, Ivo Grosse, and Armin O. Schmitt, Statistical Analysis of DNA Sequences. In *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism*, ed. Dietmar Schomburg and Uta Lessel, chapter 3, 29–46. Wiley-VCH Verlag GmbH, Weinheim, Germany, 1995.

[Holste *et al.*2001] Dirk Holste, Ivo Grosse, and Hanspeter Herzel, Statistical analysis of the DNA sequence of human chromosome 22. *Physical Review E* **64**(4) (2001), 1–9.

[Hu *et al.*2011] Tina T Hu, Pedro Pattyn, Erica G Bakker, Jun Cao, Jan-Fang Cheng, Richard M Clark, Noah Fahlgren, Jeffrey a Fawcett, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Jesse D Hollister, Stephan Ossowski, Robert P Ottilar, Asaf a Salamov, Korbinian Schneeberger, Manuel Spannagl, Xi Wang, Liang Yang, Mikhail E Nasrallah, Joy Bergelson, James C Carrington, Brandon S Gaut, Jeremy Schmutz, Klaus F X Mayer, Yves Van de Peer, Igor V Grigoriev, Magnus Nordborg, Detlef Weigel, and Ya-Long Guo, The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics* **43**(5) (2011), 476–81.

[Huang *et al.*2009] Da Wei Huang, Brad T Sherman, and Richard A Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**(1) (2009), 1–13.

[Hurst *et al.*2004] Laurence D Hurst, Csaba Pál, and Martin J Lercher, The evolutionary dynamics of eukaryotic gene order. *Nature reviews. Genetics* **5**(4) (2004), 299–310.

[Ideker & Sharan2008] Trey Ideker and Roded Sharan, Protein networks in disease. *Genome research* **18**(4) (2008), 644–52.

[Ihaka & Gentleman1996] Ross Ihaka and Robert Gentleman, R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**(3) (1996), p. 299.

[Kinsella *et al.*2011] Rhoda J. Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek, Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database : the journal of biological databases and curation* **2011** (2011), p. bar030.

[Kosiol *et al.*2008] Carolin Kosiol, Tomás Vinar, Rute R da Fonseca, Melissa J Hubisz, Carlos D Bustamante, Rasmus Nielsen, and Adam Siepel, Patterns of positive selection in six Mammalian genomes. *PLoS genetics* **4**(8) (2008), p. e1000144.

[Lander *et al.*2001] Eric S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, et al., Initial sequencing and analysis of the human genome. *Nature* **409**(6822) (2001), 860–921.

[Lee & Sonnhammer2003] Jennifer M Lee and Erik L L Sonnhammer, Genomic gene clustering analysis of pathways in eukaryotes. *Genome research* **13**(5) (2003), 875–82.

[Liu *et al.*2008] Zhandong Liu, Santosh S Venkatesh, and Carlo C Maley, Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. *BMC genomics* **9** (2008), p. 509.

[Loewenstern & Yianilos1999] David M Loewenstern and Peter N Yianilos, Significantly lower entropy estimates for natural DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **6**(1) (1999), 125–42.

[Lynch2007] Michael Lynch, *The Origins of Genome Architecture*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 2007.

[Nachman & Crowell2000] M W Nachman and S L Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1) (2000), 297–304.

[Nielsen2001] R Nielsen, Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**(Pt 6) (2001), 641–7.

[Nielsen *et al.*2005] Rasmus Nielsen, Carlos Bustamante, Andrew G. Clark, Stephen Glanowski, Timothy B Sackton, Melissa J Hubisz, Adi Fledel-Alon, David M Tanenbaum, Daniel Civello, Thomas J White, John J Sninsky, Mark D Adams, and Michele Cargill, A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS biology* **3**(6) (2005), p. e170.

[Nies2009] André Nies, *Computability and Randomness*. Oxford University Press, macintyre, edition, 2009.

[Ossowski *et al.*2010] Stephan Ossowski, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M Clark, Ruth G Shaw, Detlef Weigel, and Michael Lynch, The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science (New York, N.Y.)* **327**(5961) (2010), 92–4.

[Petit & Barbadilla2009] N Petit and A Barbadilla, Selection efficiency and effective population size in Drosophila species. *Journal of evolutionary biology* **22**(3) (2009), 515–26.

[Ryabko1980] B Ya Ryabko, Data Compression by Means of a 'Book Stack'. *Problems Information Transmission* **16**(4) (1980), 16–21.

[Sánchez *et al.*2011] Rubén Sánchez, François Serra, Joaquín Tárraga, Ignacio Medina, José Carbonell, Luis Pulido, Alejandro de María, Salvador Capella-Gutíerrez, Jaime Huerta-Cepas, Toni Gabaldón, Joaquín Dopazo, and Hernán Dopazo, Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic acids research* **39 Suppl 2**(Web Server issue) (2011), W470–4.

[Shannon1948] C E Shannon, *A mathematical theory of communication*, vol. 5. The Bell System Technical Journal, 1948.

[Shapiro & Alm2008] B.J. Shapiro and E.J. Alm, Comparing patterns of natural selection across species using selective signatures. *PLoS genetics* **4**(2) (2008), e23+.

[Smit *et al.*2010] A F A Smit, R Hubley, and P Green, RepeatMasker Open-3.0, 2010.

[Solé & Valverde2008] Ricard V Solé and Sergi Valverde, Spontaneous emergence of modularity in cellular networks. *Journal of the Royal Society, Interface / the Royal Society* **5**(18) (2008), 129–33.

[Stelzl *et al.*2005] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E Wanker, A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**(6) (2005), 957–68.

[Subramanian *et al.*2005] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**(43) (2005), 15545–50.

[Team2011] R Development Core Team, R: A Language and Environment for Statistical Computing, 2011.

[Vamathevan *et al.*2008] Jessica J Vamathevan, Samiul Hasan, Richard D Emes, Heather Amrine-Madsen, Dilip Rajagopalan, Simon D Topp, Vinod Kumar, Michael Word, Mark D Simmons, Steven M Foord, Philippe Sanseau, Ziheng Yang, and Joanna D Holbrook, The role of positive selection in determining the molecular cause of species differences in disease. *BMC evolutionary biology* **8**(1) (2008), p. 273.

[Venter *et al.*2001] J Craig Venter, M D Adams, E W Myers, P W Li, R J Mural, et al., The sequence of the human genome. *Science (New York, N.Y.)* **291**(5507) (2001), 1304–51.

[Vilella *et al.*2009] Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**(2) (2009), 327–35.

[Wagner *et al.*2007] Günter P Wagner, Mihaela Pavlicev, and James M Cheverud, The road to modularity. *Nature reviews. Genetics* **8**(12) (2007), 921–31.

[Weber & Helentjaris1989] D Weber and T Helentjaris, Mapping RFLP loci in maize using B-A translocations. *Genetics* **121**(3) (1989), 583–90.

[Wernegreen2002] Jennifer J Wernegreen, Genome evolution in bacterial endosymbionts of insects. *Nature reviews. Genetics* **3**(11) (2002), 850–61.

[Wolfe2001] Kenneth H Wolfe, Yesterday's polyploids and the mystery of diploidization. *Nature reviews. Genetics* **2**(5) (2001), 333–41.

[Yang2007] Ziheng Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8) (2007), 1586–91.

[Zhang *et al.*2005] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang, Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution* **22**(12) (2005), 2472–9.

List of Figures

List of Tables

Glossary

seed -*sequence* of a gene or a protein, is the sequence used as starting point in the search of homologous sequences within a given set of entries. Extending this concept at genomic level, we can talk about *seed-genome* or *seed-species*. ***Note:*** In a phylome, it is expected to observe an over-representation of proteins belonging from the seed-species. 28

A. RepeatMasker summary output

```
==================================================
file name: Homo_sapiens.all_chromosomes.fasta
sequences:            24
total length: 3095677412 bp  (2858660140 bp excl N/X-runs)
GC level:         Unknown %
bases masked: 1412780617 bp ( 45.64 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:            1658864    385270856 bp   12.45 %
      ALUs        1136457    306395826 bp    9.90 %
      MIRs         517233     78244089 bp    2.53 %

LINEs:            913889    609952196 bp   19.70 %
      LINE1       539553    503348534 bp   16.26 %
      LINE2       319303     93411598 bp    3.02 %
      L3/CR1       42713     10009516 bp    0.32 %

LTR elements:     487433    259122242 bp    8.37 %
      ERVL        108675     55875700 bp    1.80 %
      ERVL-MaLRs  247590    108138874 bp    3.49 %
      ERV_classI  109816     82706444 bp    2.67 %
      ERV_classII   7480      8820605 bp    0.28 %

DNA elements:     383832     95646896 bp    3.09 %
     hAT-Charlie  214295     43419001 bp    1.40 %
     TcMar-Tigger  82218     33550442 bp    1.08 %

Unclassified:       9962      5418573 bp    0.18 %

Total interspersed repeats:1355410763 bp   43.78 %


Small RNA:         13482      1443809 bp    0.05 %

Satellites:         4502     12381861 bp    0.40 %
Simple repeats:   403012     25937716 bp    0.84 %
Low complexity:   393080     17947554 bp    0.58 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The query species was assumed to be homo sapiens
RepeatMasker version open-3.3.0 , default mode

run with rmblastn version : 2.2.23+
RepBase Update 20110419, RM database version 20110419
```