

Adaptation in genes, duplicates, families, functional modules and genomes

François Serra

October 2011

Contents

1. Introduction	1
1.1. Adaptive changes to evolutionary speed	1
1.2. Evolution, and the detection at molecular level	1
1.3. Grouping genes and finding evolutionary patterns	1
1.4. What is DNA? How genes rose?	1
1.5. Life in DNA, from genes to repetitive elements.	1
I. Structure and dynamics of genomes	2
2. Random-like structure of DNA	3
2.1. Introduction	3
2.2. Results and Discussion	3
2.2.1. Computing genome complexity	3
2.2.2. Genome complexity and ploidy level	5
2.2.3. Chromosome complexity	9
2.2.4. Complexity in chromosome segments	9
2.3. Material and methods	16
2.3.1. The complexity ratio and complexity value	16
2.3.2. Complexity in strings	17
2.3.3. Simulations	18
3. Life inside genomes, dynamics and predictions	19
3.1. Genomic elements, dispersion and abundance	19
3.2. Species Abundance Diversity in genomes	19
3.3. Neutrality of SAD	19
3.4. Material Methods	19
3.4.1. Ecology	19
II. Detection of selective pressures in genomes	20
4. Searching for evolutionary patterns in funcionally linked group of genes	21
4.1. Introduction	21
4.2. Material and Methods	21
4.2.1. Dataset	21
4.2.2. Alignments	21

Contents

4.3. open on colocalization to not random	21
5. Tools, programs, methods	23
5.1. ETE-evol plugin	23
5.1.1. BRANCHED1	23
5.1.2. Protamines Rodents and Primates	23
5.2. Pipeline for study of adaptation at genomic scale	23
5.2.1. Selective pressure on duplicated genes in Drosophila	23
5.3. Phylemon	23
6. Conclusions	24
A. RepeatMasker summary output	A

Nomenclatura

BWT Burros-Wheeler transform

Chr Chromosome

CR Complexity Ratio

CV Complexity Value

GE Genomic Element

MTF Move To Front

1. Introduction

- 1.1. Adaptive changes to evolutionary speed**
- 1.2. Evolution, and the detection at molecular level**
- 1.3. Grouping genes and finding evolutionary patterns**
- 1.4. What is DNA? How genes rose?**
- 1.5. Life in DNA, from genes to repetitive elements.**

Part I.

Structure and dynamics of genomes

2. Random-like structure of DNA

2.1. Introduction

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 DNA 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 $p < 2.0e-16$, Figure 2.2-A. 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

2. Random-like structure of DNA

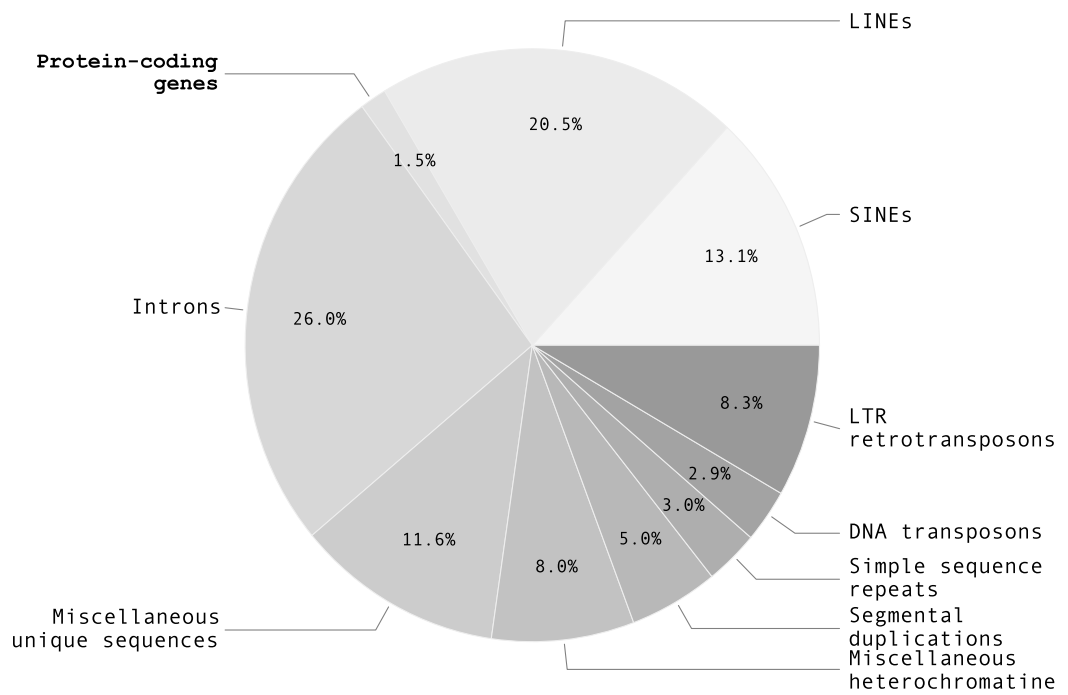


Figure 2.1.: Genomic components of human genome.

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

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 $\sim 1.69\text{e}+03$ bp) to the largest double-strand DNA genome of the short-tailed opossum (size $\sim 3.41\text{e}+09$ 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 \times), randomly generated sequences with uniform distribution of ACGT; however, CR falls exponentially when ploidy level increases reaching CR=0.25 for 10 \times 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 Species (0.309), which is comparable to the CR of a random polyploid sequence of 7 \times . 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 ($p < 2\text{e-}16$, adjusted-R2 = 0.997), while independent linear

2. Random-like structure of DNA

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	Amoebozoa	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

Table 2.1.: 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/>

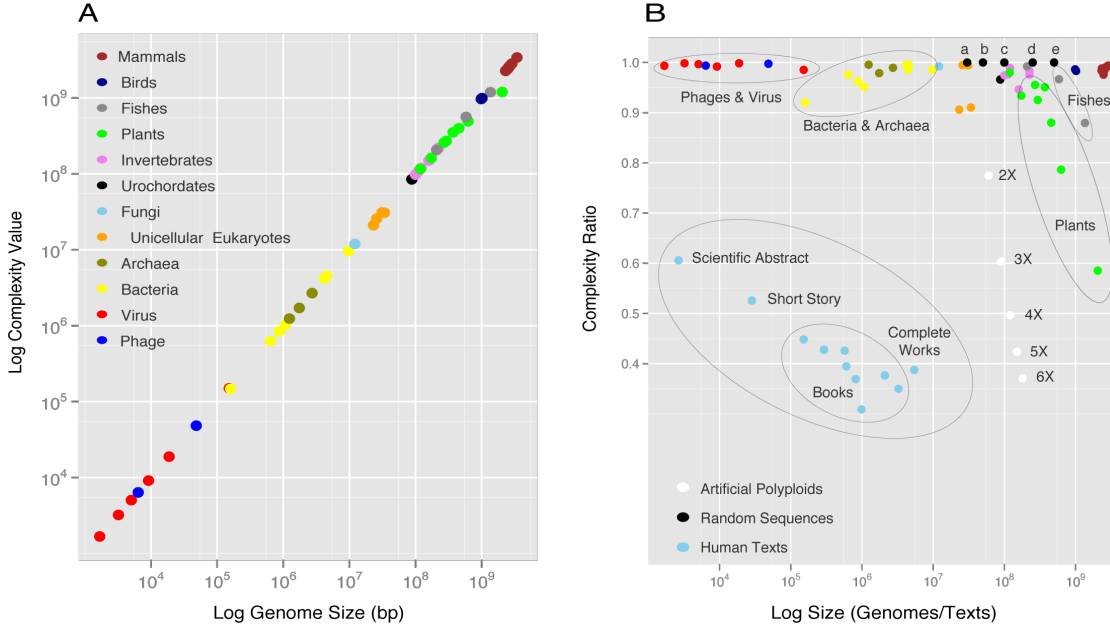


Figure 2.2.: Genome complexity value.

(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).

2. Random-like structure of DNA

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

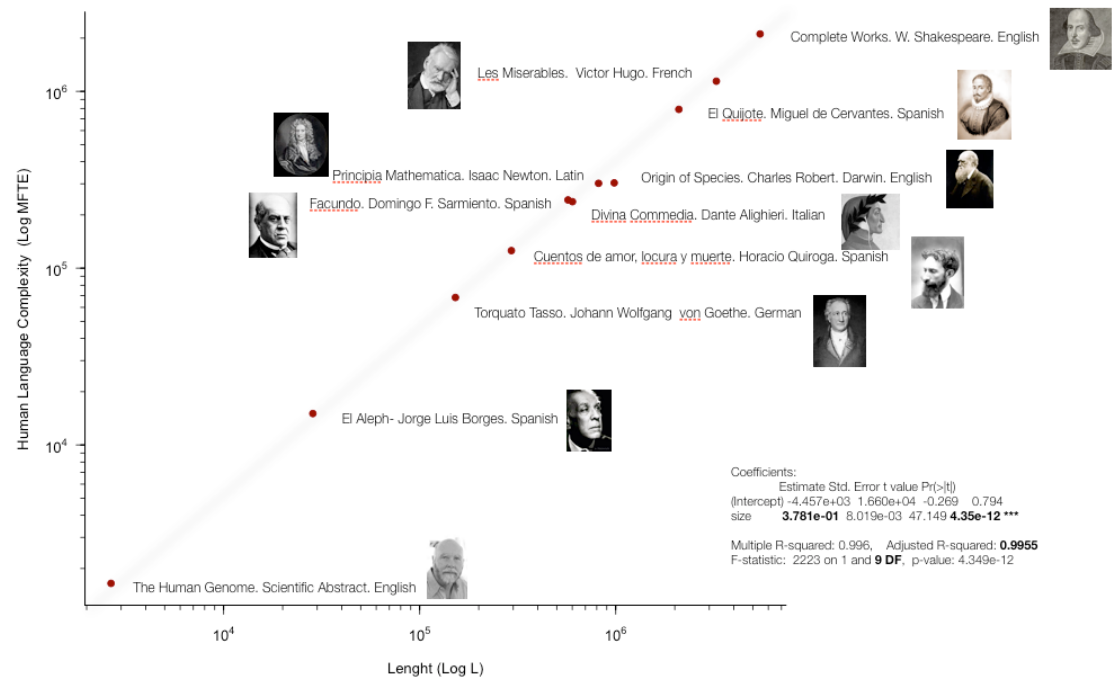


Figure 2.3.: Human language complexity.

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 $R = 0.995$).

models slopes were 0.633 ($p < 4.8\text{e-}07$, adjusted- $R^2 = 0.921$), and 0.988 ($p < 2\text{e-}16$, adjusted- $R^2 = 1.00$) 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 dataset, or excluding polyploid species revealed a very significant statistical relationship (slope = 0.924, adjusted- $R^2 = 0.989$, $p < 2\text{e-}16$, or slope = 0.951, adjusted- $R^2 = 0.999$, $p < 2\text{e-}16$, respectively). The adjustment of a linear regression model to chromosomes of polyploid species was statistically significant ($p < 2\text{e-}16$, $R^2 = 0.982$), 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 ($p < 2\text{e-}16$), 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 first 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. cerevisiae* (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

2. Random-like structure of DNA

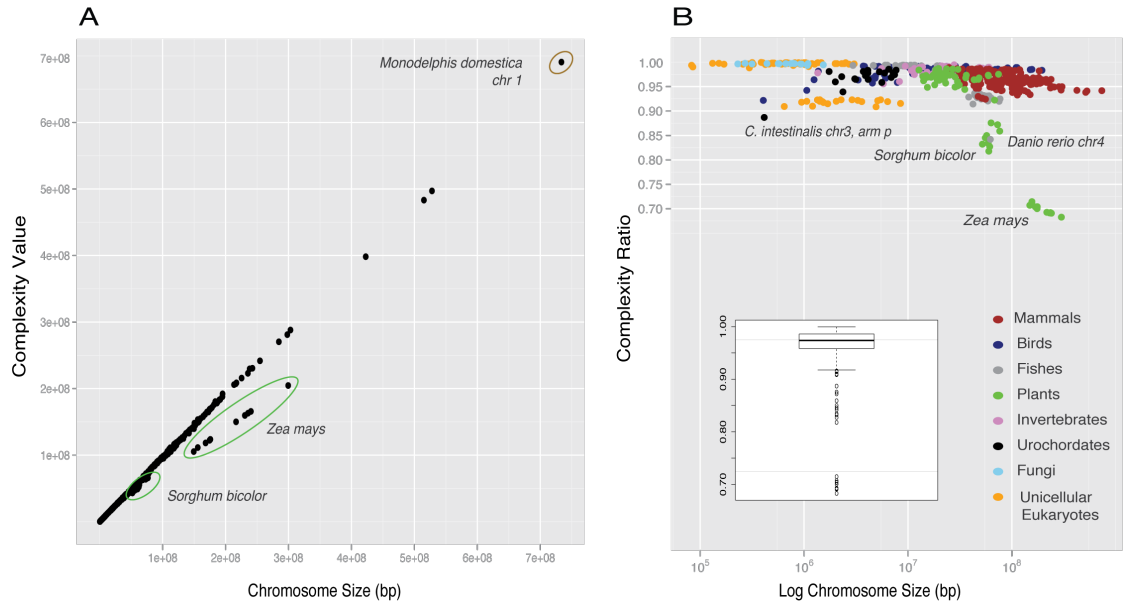


Figure 2.4.: Chromosome complexity ratio.

(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.

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].

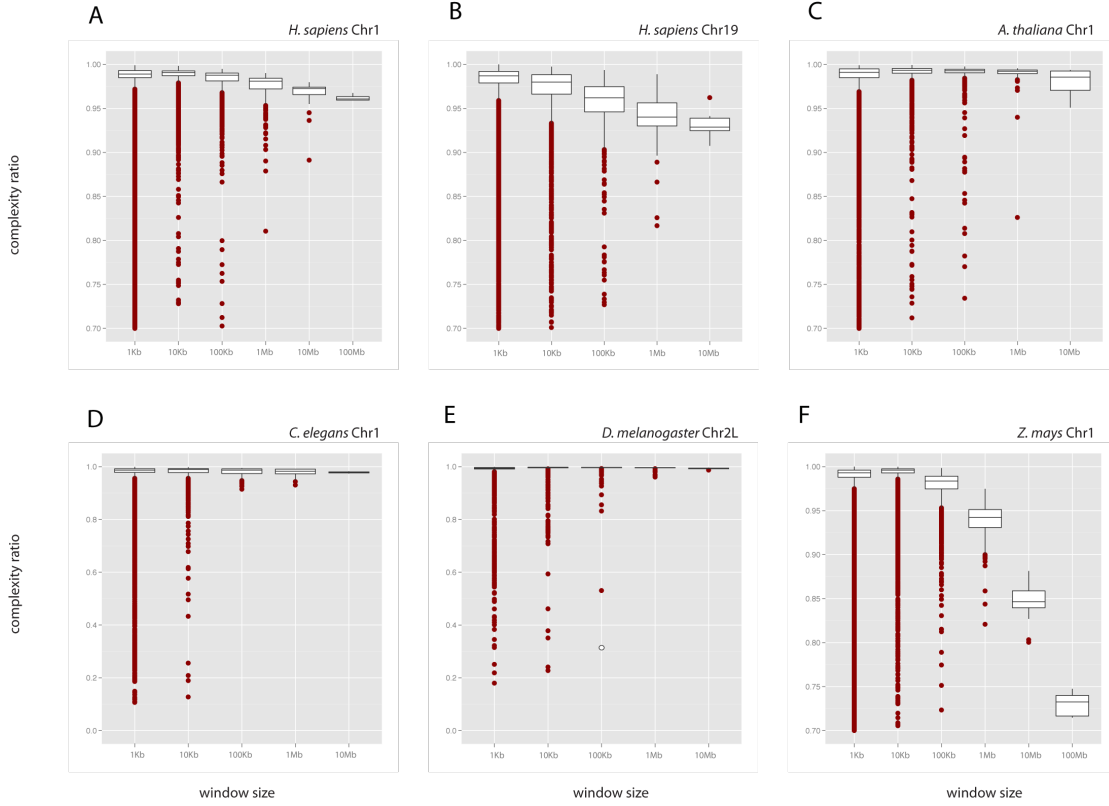


Figure 2.5.: Sliding window analysis in chromosomes.

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.

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 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.

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/> Complex-

2. Random-like structure of DNA

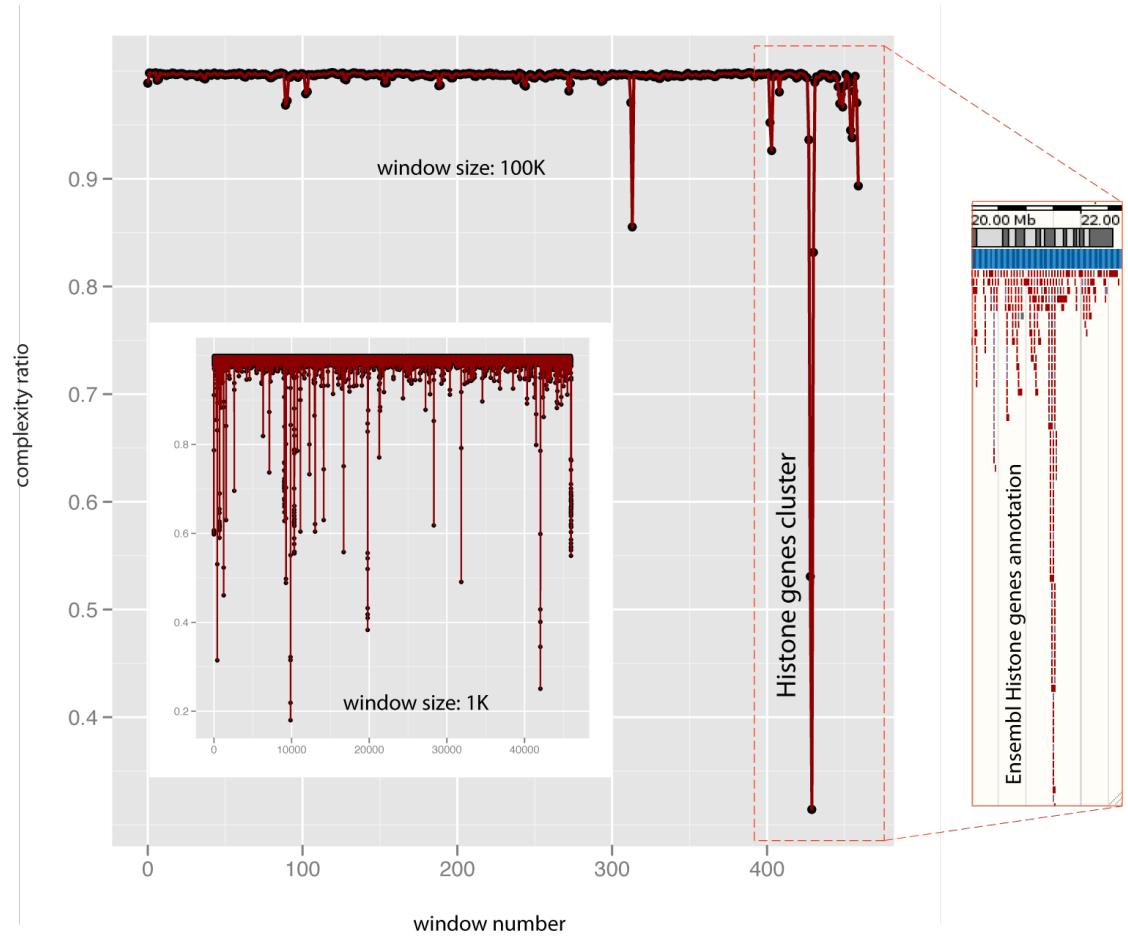


Figure 2.6.: Sliding window in a full chromosome.

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.

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

2. Random-like structure of DNA

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.

		<i>A. thaliana</i>		<i>C. elegans</i>		<i>H. sapiens</i>		<i>Z. mays</i>	
		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 single 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 $k \neq \log_4 n$, and each these short sequences occur about $n \cdot 4^{-k}$ times. As CR decreases, diversity and balanced abundance deteriorates. In particular, maximum CR is holds for some de

2. Random-like structure of DNA

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 ($CR \geq 1 - \epsilon$, for any real value ϵ 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 Burrows-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 :

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

For DNA alphabet entropy is defined as:

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

Thus the CR can be factorize as:

$$CR(s) = E(MTF(BWT(s))) \quad (2.3)$$

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

$$CV(s) = E(MTF(BWT(s))) \times length(s) \quad (2.4)$$

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, ampicomplexa, 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 $6\times$ 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 untranslated 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

2. Random-like structure of DNA

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 Gutenberg (<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 $1.0\text{e-}08$ mutations per site per generation (this value is in between the mutation rate estimated for *Homo sapiens*: $2.5\text{e-}08$ [Nachman & Crowell2000] , and *Arabidopsis thaliana*: $7.1\text{e-}09$ [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.

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 functionally linked group of genes

4.1. Introduction

4.2. Material and Methods

4.2.1. Dataset

Five mammals

Complete genomes of 5 mammals species (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus* and *Canis familiaris*) were 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]. Only groups of orthologs *one-to-one* with one representative of each species were kept in the final dataset. Figure 4.1 NUMBERS

6 Drosophila

4.2.2. Alignments

Each of the group of orthologous sequences were aligned with Muscle [Edgar2004], and, once aligned sequences were cleaned with trimAl [Capella-Gutiérrez *et al.*2009] keeping all sequences but trimming alignment columns with the euristic1 method.

4.3. open on colocalization to not random

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

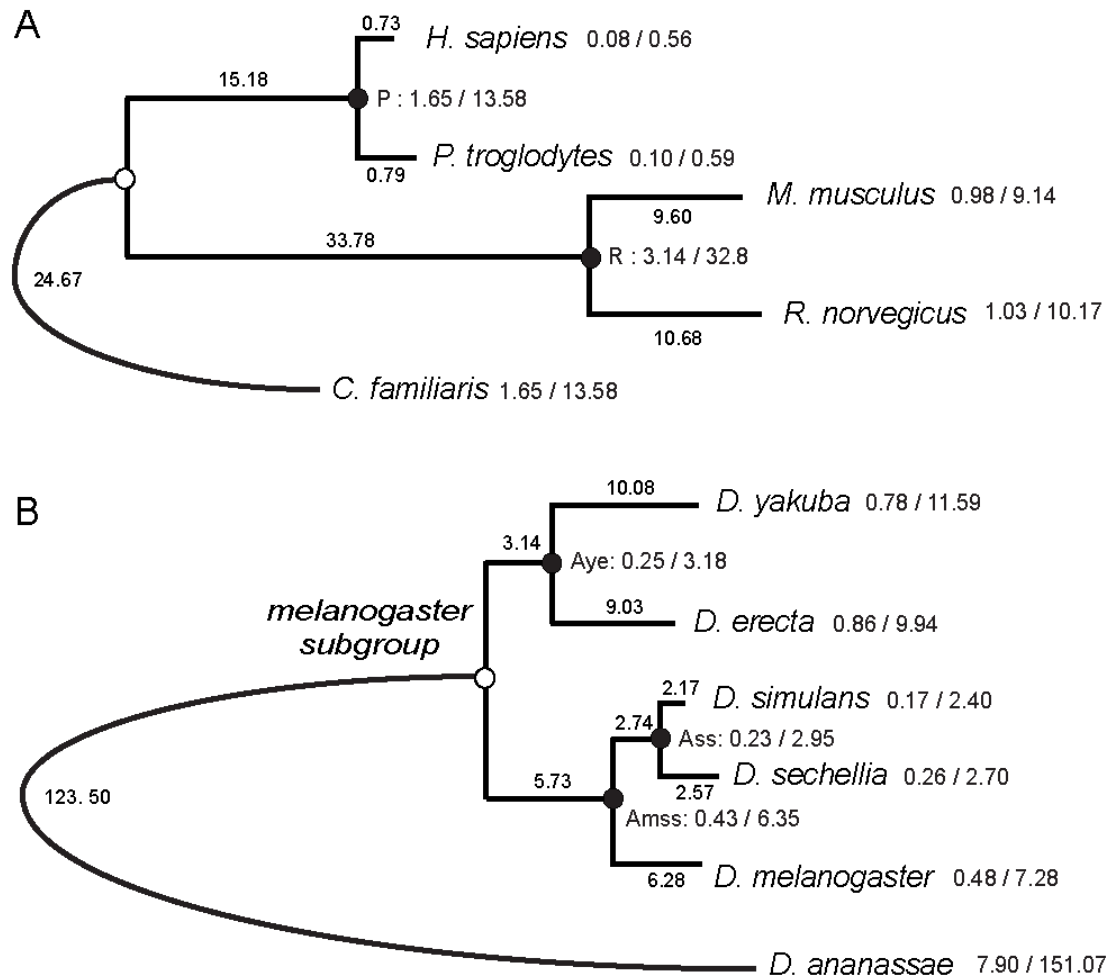


Figure 4.1.: Mammals and *Drosophila* phylogeny.

blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf blabli blob lu dkfnlskjdf

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

- [Adjero 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.
- [Becher & Heiber2011] VERÓNICA BECHER AND PABLO ARIEL HEIBER, On extending de Bruijn sequences. *Information Processing Letters* **111**(18) (2011), 930–932.
- [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).
- [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.
- [Chaitin1975] GREGORY J. CHAITIN, A Theory of Program Size Formally Identical to Information Theory. *Journal of the ACM* **22**(3) (1975), 329–340.
- [Du et al.2006] JIANG DU, JOEL S ROZOWSKY, JAN O KORBEL, ZHENG DONG 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, YUAN CHEN, PETER CLAPHAM, GUY COATES, SUSAN FAIRLEY, STEPHEN FITZGERALD, LEO GORDON, MAURICE HENDRIX, THIBAUT HOURLIER, NATHAN JOHNSON, ANDREAS KÄHÄRI, DAMIAN KEEFE, STEPHEN KEENAN, RHODA KINSELLA, FELIX KOKOCINSKI, EUGENE KULESHA, PONTUS LARSSON, IAN LONGDEN, WILLIAM McLAREN, BERT OVERDUIN, BETHAN PRITCHARD, HARPREET SINGH RIAT, DANIEL RIOS, GRAHAM R S RITCHIE, MAGALI RUFFIER, MICHAEL SCHUSTER, DANIEL SOBRAL, GIULIETTA SPUDICH, Y AMY TANG, STEPHEN TREVANION, JANA VANDROVCOVA, ALBERT J VILELLA, SIMON WHITE, STEVEN P WILDER, AMONIDA ZADISSA, JORGE ZAMORA, BRONWEN L AKEN, EWAN BIRNEY, FIONA CUNNINGHAM, IAN DUNHAM, RICHARD DURBIN, XOSÉ M FERNÁNDEZ-SUAREZ, JAVIER HERRERO, TIM J P HUBBARD, ANNE PARKER, GLENN

Bibliography

- PROCTOR, JAN VOGEL, AND STEPHEN M J SEARLE, 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, ZHENG DONG 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.
- [Gregory2005] T RYAN GREGORY, Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics* **6**(9) (2005), 699–708.
- [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.
- [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.
- [Lander *et al.*2001] E S LANDER, L M LINTON, B BIRREN, C NUSBAUM, M C ZODY, J BALDWIN, K DEVON, K DEWAR, M DOYLE, W FITZHUGH, R FUNKE, D GAGE, K HARRIS, A HEAFORD, J HOWLAND, L KANN, J LEHOCZKY, R LEVINE, P MCEWAN, K MCKERNAN, J MELDRIM, J P MESIROV, C MIRANDA, W MORRIS, J NAYLOR, C RAYMOND, M ROSETTI, R SANTOS, A SHERIDAN, C SOUGNEZ, N STANGE-THOMANN, N STOJANOVIC, A SUBRAMANIAN, D WYMAN, J ROGERS, J SULSTON, R AINSCOUGH, S BECK, D BENTLEY,

- J BURTON, C CLEE, N CARTER, A COULSON, R DEADMAN, P DELOUKAS, A DUNHAM, I DUNHAM, R DURBIN, L FRENCH, D GRAHAM, S GREGORY, T HUBBARD, S HUMPHRAY, A HUNT, M JONES, C LLOYD, A MCMURRAY, L MATTHEWS, S MERCER, S MILNE, J C MULLIKIN, A MUNGALL, R PLUMB, M ROSS, R SHOWNKEEN, S SIMS, R H WATERSTON, R K WILSON, L W HILLIER, J D MCPHERSON, M A MARRA, E R MARDIS, L A FULTON, A T CHINWALLA, K H PEPIN, W R GISH, S L CHISSOE, M C WENDL, K D DELEHAUNTY, T L MINER, A DELEHAUNTY, J B KRAMER, L L COOK, R S FULTON, D L JOHNSON, P J MINX, S W CLIFTON, T HAWKINS, E BRANSCOMB, P PREDKI, P RICHARDSON, S WENNING, T SLEZAK, N DOGETT, J F CHENG, A OLSEN, S LUCAS, C ELKIN, E UBERBACHER, M FRAZIER, R A GIBBS, D M MUZNY, S E SCHERER, J B BOUCK, E J SODERGREN, K C WORLEY, C M RIVES, J H GORRELL, M L METZKER, S L NAYLOR, R S KUCHERLAPATI, D L NELSON, G M WEINSTOCK, Y SAKAKI, A FUJIYAMA, M HATTORI, T YADA, A TOYODA, T ITOH, C KAWAGOE, H WATANABE, Y TOTOKI, T TAYLOR, J WEISSENBAACH, R HEILIG, W SAURIN, F ARTIGUENAVE, P BROTTIER, T BRULS, E PELLETIER, C ROBERT, P WINCKER, D R SMITH, L DOUCETTE-STAMM, M RUBENFIELD, K WEINSTOCK, H M LEE, J DUBOIS, A ROSENTHAL, M PLATZER, G NYAKATURA, S TAUDIEN, A RUMP, H YANG, J YU, J WANG, G HUANG, J GU, L HOOD, L ROWEN, A MADAN, S QIN, R W DAVIS, N A FEDERSPIEL, A P ABOLA, M J PROCTOR, R M MYERS, J SCHMUTZ, M DICKSON, J GRIMWOOD, D R COX, M V OLSON, R KAUL, N SHIMIZU, K KAWASAKI, S MINOSHIMA, G A EVANS, M ATHANASIOU, R SCHULTZ, B A ROE, F CHEN, H PAN, J RAMSER, H LEHRACH, R REINHARDT, W R MCCOMBIE, M DE LA BASTIDE, N DEDHIA, H BLÖCKER, K HORNISCHER, G NORDSIEK, R AGARWALA, L ARAVIND, J A BAILEY, A BATEMAN, S BATZOGLOU, E BIRNEY, P BORK, D G BROWN, C B BURGE, L CERUTTI, H C CHEN, D CHURCH, M CLAMP, R R COPLEY, T DOERKS, S R EDDY, E E EICHLER, T S FUREY, J GALAGAN, J G GILBERT, C HARMON, Y HAYASHIZAKI, D HAUSSLER, H HERMJAKOB, K HOKAMP, W JANG, L S JOHNSON, T A JONES, S KASIF, A KASPRYZK, S KENNEDY, W J KENT, P KITTS, E V KOONIN, I KORF, D KULP, D LANCET, T M LOWE, A MCLYSAGHT, T MIKKELSEN, J V MORAN, N MULDER, V J POLLARA, C P PONTING, G SCHULER, J SCHULTZ, G SLATER, A F SMIT, E STUPKA, J SZUSTAKOWSKI, D THIERRY-MIEG, J THIERRY-MIEG, L WAGNER, J WALLIS, R WHEELER, A WILLIAMS, Y I WOLF, K H WOLFE, S P YANG, R F YEH, F COLLINS, M S GUYER, J PETERSON, A FELSENFELD, K A WETTERSTRAND, A PATRINOS, M J MORGAN, P DE JONG, J J CATANESE, K OSOEGAWA, H SHIZUYA, S CHOI, Y J CHEN, AND J SZUSTAKOWKI, Initial sequencing and analysis of the human genome. *Nature* **409**(6822) (2001), 860–921.
- [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.
- [Nachman & Crowell2000] M W NACHMAN AND S L CROWELL, Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1) (2000), 297–304.
- [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 WARTHMAN, RICHARD M CLARK, RUTH G SHAW, DETLEF

Bibliography

- 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.
- [Ryabko1980] B YA RYABKO, Data Compression by Means of a 'Book Stack'. *Problems Information Transmission* **16**(4) (1980), 16–21.
- [Shannon1948] C E SHANNON, *A mathematical theory of communication*, vol. 5. The Bell System Technical Journal, 1948.
- [Smit *et al.*2010] A F A SMIT, R HUBLEY, AND P GREEN, RepeatMasker Open-3.0. (2010), p. <http://www.repeatmasker.org>.
- [Team2011] R DEVELOPMENT CORE TEAM, R: A Language and Environment for Statistical Computing, 2011.
- [Venter *et al.*2001] J C VENTER, M D ADAMS, E W MYERS, P W LI, R J MURAL, G G SUTTON, H O SMITH, M YANDELL, C A EVANS, R A HOLT, J D GOCAYNE, P AMANATIDES, R M BALLEW, D H HUSON, J R WORTMAN, Q ZHANG, C D KODIRA, X H ZHENG, L CHEN, M SKUPSKI, G SUBRAMANIAN, P D THOMAS, J ZHANG, G L GABOR MIKLOS, C NELSON, S BRODER, A G CLARK, J NADEAU, V A MCKUSICK, N ZINDER, A J LEVINE, R J ROBERTS, M SIMON, C SLAYMAN, M HUNKAPILLER, R BOLANOS, A DELCHER, I DEW, D FASULO, M FLANIGAN, L FLOREA, A HALPERN, S HANNENHALLI, S KRAVITZ, S LEVY, C MOBARRY, K REINERT, K REMINGTON, J ABU-THREIDEH, E BEASLEY, K BIDDICK, V BONAZZI, R BRANDON, M CARGILL, I CHANDRAMOULISWARAN, R CHARLAB, K CHATURVEDI, Z DENG, V DI FRANCESCO, P DUNN, K EILBECK, C EVANGELISTA, A E GABRIELIAN, W GAN, W GE, F GONG, Z GU, P GUAN, T J HEIMAN, M E HIGGINS, R R JI, Z KE, K A KETCHUM, Z LAI, Y LEI, Z LI, J LI, Y LIANG, X LIN, F LU, G V MERKULOV, N MILSHINA, H M MOORE, A K NAIK, V A NARAYAN, B NEE-LAM, D NUSSKERN, D B RUSCH, S SALZBERG, W SHAO, B SHUE, J SUN, Z WANG, A WANG, X WANG, J WANG, M WEI, R WIDES, C XIAO, C YAN, A YAO, J YE, M ZHAN, W ZHANG, H ZHANG, Q ZHAO, L ZHENG, F ZHONG, W ZHONG, S ZHU, S ZHAO, D GILBERT, S BAUMHUETER, G SPIER, C CARTER, A CRAVCHIK, T WOODAGE, F ALI, H AN, A AWE, D BALDWIN, H BADEN, M BARNSTEAD, I BARROW, K BEESON, D BUSAM, A CARVER, A CENTER, M L CHENG, L CURRY, S DANAHAR, L DAVENPORT, R DESILETS, S DIETZ, K DODSON, L DOUP, S FERRIERA, N GARG, A GLUECKSMANN, B HART, J HAYNES, C HAYNES, C HEINER, S HLADUN, D HOSTIN, J HOUCK, T HOWLAND, C IBEGWAM, J JOHNSON, F KALUSH, L KLINE, S KODURU, A LOVE, F MANN, D MAY, S MCCAWLEY, T MCINTOSH, I MCMULLEN, M MOY, L MOY, B MURPHY, K NELSON, C PFANNKOCHE, E PRATTS, V PURI, H QURESHI, M REARDON, R RODRIGUEZ, Y H ROGERS, D ROMBLAD, B RUHFEL, R SCOTT, C SITTER, M SMALLWOOD, E STEWART, R STRONG, E SUH, R THOMAS, N N TINT, S TSE, C VECH, G WANG, J WETTER, S WILLIAMS, M WILLIAMS, S WINDSOR, E WINN-DEEN, K WOLFE, J ZAVERI, K ZAVERI, J F ABRIL, R GUIGÓ, M J CAMPBELL, K V SJOLANDER, B KARLAK, A KEJARIWAL, H MI, B LAZAREVA, T HATTON, A NARECHANIA, K DIEMER, A MURUGANUJAN, N GUO, S SATO, V BAFNA, S ISTRAIL, R LIPPERT, R SCHWARTZ, B WALENZ, S YOOSEPH, D ALLEN, A BASU, J BAXENDALE, L BLICK, M CAMINHA, J CARNES-STINE, P CAULK, Y H CHIANG, M COYNE, C DAHLKE, A MAYS, M DOMBROSKI, M DONNELLY, D ELY, S ESPARHAM, C FOSLER, H GIRE, S GLANOWSKI, K GLASSER, A GLODEK, M GOROKHOV, K GRAHAM, B GROPMAN, M HARRIS, J HEIL, S HENDERSON, J HOOVER, D JENNINGS, C JORDAN, J JORDAN, J KASHA, L KAGAN, C KRAFT, A LEVITSKY, M LEWIS, X LIU, J LOPEZ, D MA, W MAJOROS, J MCDANIEL, S MURPHY, M NEWMAN, T NGUYEN,

- N NGUYEN, M NODELL, S PAN, J PECK, M PETERSON, W ROWE, R SANDERS, J SCOTT, M SIMPSON, T SMITH, A SPRAGUE, T STOCKWELL, R TURNER, E VENTER, M WANG, M WEN, D WU, M WU, A XIA, A ZANDIEH, AND X ZHU, 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.
- [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 WERNEGREN, Genome evolution in bacterial endosymbionts of insects. *Nature reviews. Genetics* **3**(11) (2002), 850–61.
- [Wolfe2001] K H WOLFE, Yesterday’s polyploids and the mystery of diploidization. *Nature reviews. Genetics* **2**(5) (2001), 333–41.

List of Figures

2.1. Genomic components of human genome	4
2.2. Genome complexity value	7
2.3. Human language complexity	8
2.4. Chromosome complexity ratio	10
2.5. Sliding window analysis in chromosomes	11
2.6. Sliding window in a full chromosome	12
4.1. Mammals and <i>Drosophila</i> phylogeny	22

List of Tables

2.1. Genomes Complexity.	6
2.2. Human language Complexity	8
2.3. Mean complexity ratio of some genome components	13
2.4. Complexity ratio of genome classes concatenated and shuffled	14

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 elements*	length occupied	percentage 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 %
ERV	108675	55875700 bp	1.80 %
ERV-L-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 rmbblastn version : 2.2.23+
RepBase Update 20110419, RM database version 20110419