Informational, Ecological and Systemic approaches for the analysis of Genomes

François Serra

October 2011

Contents

Acknowledgments

merci merci merci merci merci merci

Read this thesis

Some particularities about this thesis:

- **The glossary**: some words are underline with dotted lines like this. A short definition of these words can be found in the Glossary. At the end of each definition, appears the page number of all occurrences of the word defined.
- **The bibliography**: the bibliography is quite standard. In the text references appears between square brackets [like this]. In the Bibliography the number of authors is limited to 15. At the end of each reference appear the page numbers corresponding to their citation in the main text (for example, [Alonso *et al.* 2006] cited 3 times: ↪ *pages 10, 66 and 69*).

Nomenclature

bp DNA base-pair

BWT Burrows-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

FDR False Discovery Rate

GE Genetic Element

GSA Gene-Set Analysis

GSEA Gene-Set Enrichment Analysis

GSSA Gene-Set Selection Analysis

H Shannon's Entropy

LRT Likelihood Ratio Test

LTR Long Terminal Repeat

MTF Move To Front

PSG Positively selected genes

RSA Relative Species Abundance

SH Significantly High

SL Significantly Low

TE Transposable Element

UNTB Unified Neutral Theory of Biodiversity

WGD Whole Genome Duplication

1. Introduction

1.1. The definition of neutrality

A key concept in the study of evolution is the definition of neutrality. *"Nothing in Biology Makes Sense Except in the Light of Evolution [Dobzhansky 1973]"* the famous quote of Theodosius Dobzhansky many times cited and modified, can be extended to *nothing in evolution makes sense except over the denominator of neutrality.* Whatever biological field from ecology to (population) genetics through behavioral biology the study of adaptation is stacked to a descriptive stage until the definition of a neutral model. Or in Mayr's words:

> *When one attempts to determine for a given trait whether it is the result of natural selection or of chance (the incidental byproduct of stochastic processes), one is faced by an epistemological dilemma. Almost any change in the course of evolution might have resulted by chance. Can one ever prove this? Probably never.* [Mayr 1983]

Chance here can be understand as all events occurring in the context of the evolution of an entity that can be explained or reproduced by a neutral model. An algorithmic definition of a neutrality is given by Dennet in the first chapter "Natural Selection as an Algorithmic Process" of his book *Darwin's Dangerous Idea: Evolution and the Meanings of Life* [Dennett 1995]:

> *substrate neutrality: The procedure for long division works equally well with pencil or pen, paper or parchment, neon lights or skywriting, using any symbol system you like. The power of the procedure is due to its logical structure, not the causal powers*

> *of the materials used in the instantiation, just so long as those causal powers permit the prescribed steps to be followed exactly.*

All this thesis and the work that was done during the years preceding its publication is grounded on the definition, and posterior detection, of the neutral processes beyond the formation of the genomes and in the phenotypic consequences of its translation into proteins.

1.2. What is DNA? – Defining a genome.

A genome represents the overall hereditary information of an organism. As suggested by its etymology –the blend of the words *gene* and *chromosome*– a genome includes all genes but also non-coding sequences in the ensemble of chromosomes. The information conveyed by a genome is encoded either in DNA or in RNA, in all cases, this biological codex is constituted of four letters.

Also the codification of biological information is universal among all living species, variation in structure and amount are broad.

1.2.1. Biological complexity versus genome size

Darwinian evolution do not encompass directional change or global adaptive improvement. And in terms of biological complexity, this statement is confirmed by the diversity of life remaining at the lower levels of complexity. Also it is true that among the whole range of living species, from most ancestral bacteria found in the Chloroflexi's phylum [Cavalier-Smith 2006] to mammals (Figure 1.1), some branches suffered directional gains in complexity, the neutrality of these changes are hardly rejectable [McShea 1996].

Leaving this philosophical problem, at genomic level, it would be expected that the quantity of hereditary information would be proportional to the level of complexity of organisms. However, since the first measures of genome size [Vendrely & Vendrely 1948] it was quickly rejected that the quantity of DNA or C-value (amount of DNA found in an haploid nucleus) correlates with either organism complexity nor even with the number of genes [Mirsky & Ris 1951]. This contradiction was referred to as **C-value paradox** and is still a very hot problem with very few advances since the

Figure 1.1.: Overview of the tree of life.
Picture from the Tree of Life Web Project ©2007 [Maddison & Schulz 2007]. In this picture we see how biological complexity can raises from theoretical unicellular Universal Common Ancestor (bottom of the tree), to tips representing living species, with, in some cases, the acquisition of multicellularity, division of labor, evolution of meiosis, sexual reproduction, cell differentiation, early arrest of reproductive cells, cooperation among units of evolution...

discovery of non-coding DNA and the observation that some amoebas are holding 200 times more DNA than humans [Thomas 1971]. Recent most famous work being done by Gregory [Gregory 2001] with the introduction of a nuance preferring the term of **C-value enigma** to refer to it. With higher number of species measured [Gregory 2012], the paradox is nowadays more conspicuous than ever (the spectrum of C-values now extends between 0.002pg for parasitic microsporidium *Encephalitozoon intestinalis* and 1,400pg for the free-living amoeba *Chaos chaos*), that is, almost 7 orders of magnitude.

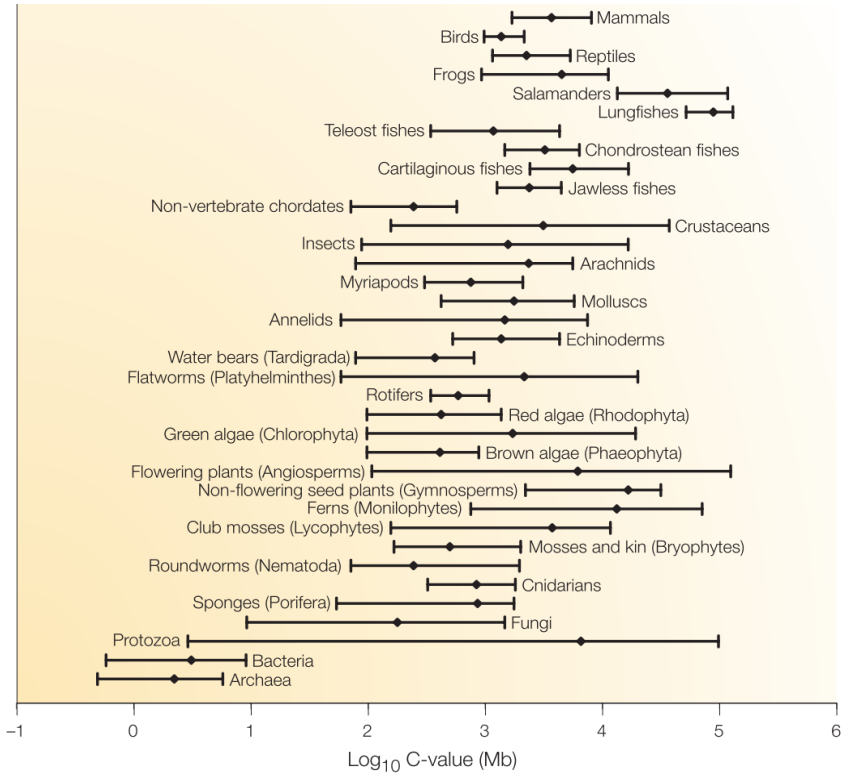


Figure 1.2.: C-values of the main groups of life.
Variation in genome size within and among the main groups of life from [Gregory 2005].

The main questions raised by the imbalance between C-value and num-

ber of genes, biological complexity or even clade specificity (see Figure 1.2) can be summarized as this [Gregory 2005]:

1. What types of non-coding DNA are found in eukaryotic genomes, and in what proportions?

2. From where does this non-coding DNA come, and how is it spread and/or lost from genomes over time?

3. What effects, or perhaps even functions, does this non-coding DNA have for chromosomes, nuclei, cells, and organism phenotypes?

4. Why do some species (for example birds) exhibit remarkably streamlined chromosomes, while others possess massive amounts of non-coding DNA (like salamanders)?

Methodologically these questions appears to be divisible in two problems, the search of patterns in the informational content of genomes along the diversity of life, and the dynamics underlying the distribution and appearance of non-coding DNA.

1.2.2. Informational content of DNA

From a biological perspective it is almost impossible to imagine DNA as a random mix of A, T, G and C nucleotides. Genomes contains the entirety of our hereditary information code either into DNA or RNA for some viruses. They 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 are the exact opposite of a random structure in terms of computational. However to what extent can we state that genomes are not a random soup of 4 letters? Intuitively we could assume that DNA presents a much simpler structure in those families of genetic 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).

One point of the C-value paradigm is that some species with similar level of complexity or number of genes present large differences in genome

size. These differences are usually explained by the spread of repetitive genetic-elements or large/small-scale genomic duplications [Gregory 2005]. Thus, one could expect that, reducing the importance of repetition in the measure of genome size would correct the defective correlation between DNA content and organism complexity. A quick way of achieving this would be to use the size of compressed genomes as was proposed by Taft [Taft *et al.* 2007], as data compression algorithms take advantage of the presence of repetitions.

Difference between compressed genome size and direct genome size, could then be viewed as genome complexity. In contrast with biological complexity, the measure of DNA complexity is generally better accepted, and give us a perfect tool in order to try to decipher DNA structure. Some work have been already done using different measures of DNA complexity, Taft et al. [Taft *et al.* 2007] proposed a relationship between biological complexity and compressed genome size hypothesizing that simple organisms compress better than complex ones. An other study conducted by Liu et al. [Liu *et al.* 2008] deduced from the comparison of the occurrence of n-mers across genomes of 7 species that human genome is not perfectly random as it lacks some fixed-length sequence (mandatory in a random context). Exact opposite conclusion was raised in the study of the statistical structure of one piece of human chromosome [Azbel' 1995] and of bacteriophage lambda. Here the main conjecture raised by the work was that statistical structure of DNA is universal for life.

The first part of this thesis stands also ahead of calculating DNA complexity, but, as main point it takes advantage on many recently sequenced genomes and on getting a global measure, getting one single value per genome instead of estimating it from many local estimations. The complexity value that we define to describe the statistical structure of DNA is based on a classical method for data compression [Adjeroh *et al.* 2008], it is able to detect regularities due to repetitions among data. Almost same measure was already used in [Holste *et al.* 2001] but only applied to DNA chunks of human chromosome 22.

The results presented in section 3.1 show that along all the diversity of life covered by our set of species, the ratio between the complexity and the sequence size is almost maximal in genomes and chromosomes. Notable exceptions are recent polyploids.

1.2.3. Dynamics of genetics elements

From the first analysis of the human genome [Lander *et al.* 2001] we know of the proportion of each of the *families* of elements Figure 1.3.

These proportions of genetic elements in the human genome were contrasted in the years following 2001 with others genomes sequenced coming to the picture of their important variation among organisms. As an example, Figure 1.4 shows the variation in proportions of the 2 major families of transposable elements in different eukaryotic species.

In the second part of this thesis we wanted to focus on the dynamics beyond the dispersion of the genetic elements in genomes of different eukaryotic species. Defining genetic elements as all kinds of non-coding DNA, using classic families of repetitive elements, and also coding sequences classified into biotypes.

Simile with ecology

When describing ecosystems, ecologists usually focus on living species' natural environment, and their distribution and abundances within it. One common pattern that raises when studying different ecosystems, is that whatever environment studied, at whatever trophic level, it seems to be universal the case where few species are "dominating" the ecosystem, comprising the majority of the individuals [Preston 1948, Fisher *et al.* 1943]. Thus, the question that raises is:

> What mechanisms control this uneven distribution of species abundance in ecological communities?

This problem, reduced to the study of species diversity and abundance, is one of the oldest and more active topic in ecology [McGill *et al.* 2007]. Roughly speaking, ecological models of species abundance are of two kinds: descriptive (statistical-based) or mechanistic (niche-based or neutrals). While many mechanistic approaches assume ecological niche differences as the main cause driving community composition, neutral models assume that all niche differences are null [Magurran 2004]. The unified neutral theory of biodiversity (UNTB) [Hubbell 2001, Rosindell *et al.* 2011] is a neutral-stochastic theory originally inspired in population genetic [Kimura 1985, Wright 1931]. It assumes that individuals among trophically similar species are ecologically identical. This provocative assumption means

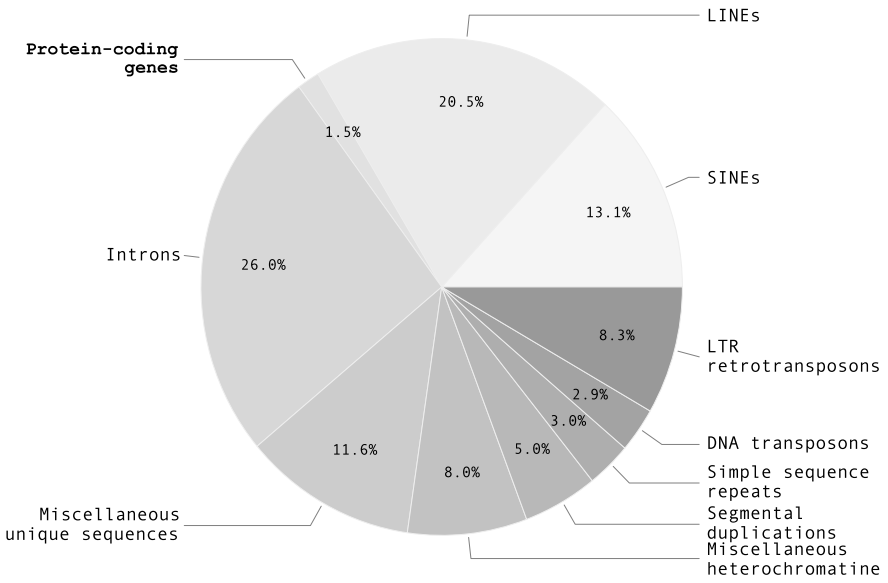


Figure 1.3.: Genomic components of human genome.
Proportion of the major families of different genomic elements (GE) in the human genome according to [Lander *et al.* 2001]

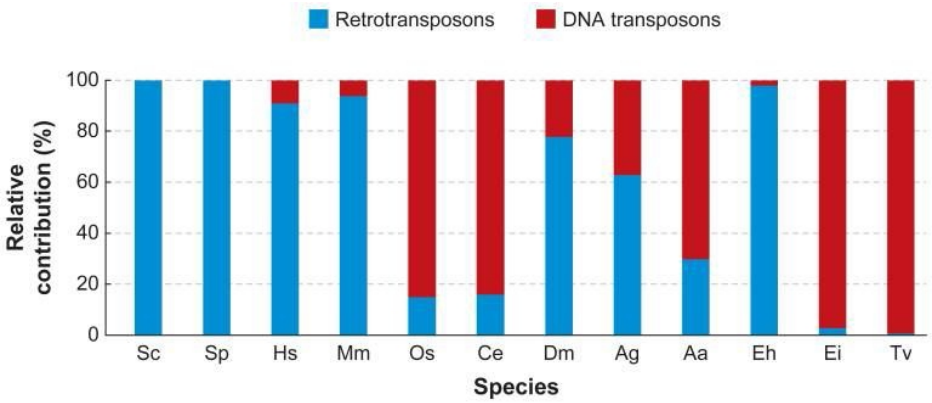


Figure 1.4.: Relative amount of retrotransposons and DNA transposons in diverse eukaryotic genomes.
This graph shows the contribution of DNA transposons and retrotransposons in percentage relative to the total number of transposable elements in each species. Species abbreviations: **Sc**: Saccharomyces cerevisiae; **Sp**: Schizosaccharomyces pombe; **Hs**: Homo sapiens; **Mm**: Mus musculus; **Os**: Oryza sativa; **Ce**: Caenorhabditis elegans; **Dm**: Drosophila melanogaster; **Ag**: Anopheles gambiae, malaria mosquito; **Aa**: Aedes aegypti, yellow fever mosquito; **Eh**: Entamoeba histolytica; **Ei**: Entamoeba invadens; **Tv**: Trichomonas vaginalis. Adapted from [Pray 2008]

that these individuals, regardless of the species specificity, are controlled by a common birth, death, dispersal, and speciation rates. The model is thus able to predict the species diversity pattern according to very few parameters. Indeed, the observed values of number of species, the total number of individuals, and 2 extra parameters describing the species richness and the migration rate are sufficient to model the species abundance diversity in a neutral context. The most important being the fundamental biodiversity number (θ). θ is analogous to $4N\mu$ of population genetics, it governs species richness in spatial and temporal scale. Given a sample size and a number of species, θ is sufficient to model the species diversity and relative abundances. Under neutral model an other parameter can also be estimated in order to take into account a specific migration rates m (see subsection 2.2.5 for details about variations in Hubbell's neutral model).

A community is then a group of species whose competitive interaction strengths are determined by their niche overlaps, and new species originate through adaptation to new niches. This view was challenged by MacArthur and Wilson with their equilibrium theory of island biogeography [MacArthur & Wilson 1967], which was extended by Hubbell [Hubbell 2001].

In ecology, neutral model is thus a useful null model against to test alternative biological hypotheses of relative species abundance distribution [Volkov *et al.* 2003, Alonso *et al.* 2006]. In non-neutral model, species are considered to be ecologically different, with more or less different niches. One simple step to move away from neutrality is, for example, to assume that species richness affects the fitness or death rate of a species (see addition of δ parameter to neutral model [Jabot & Chave 2011]).

Also dynamical ecological models of genomes were already formalized and simulated [Abrusán & Krambeck 2006, Leonardo & Nuzhdin 2002, Le Rouzic *et al.* 2007]. Some complex models parallel interactions like parasitism, competition and cooperation between different families of transposable elements (TEs). However, ideal models of genomics would consider not only TEs, but also all diversity of GEs populating eukaryote genomes: satellites sequences, DNA-transposons, LTRs-retrotransposons, LINEs, SINEs (retroposons), mi-RNA, rRNA, tRNA, and genes among the many functional and non-functional elements. Such model does not exist for genomes.

The definition of "species"

In biology, species are defined as the basic unit of biological classification. The limit between one species and an other is classically defined from the observation of sexual capabilities, or reproductive isolation. Thus, one species is defined as the ensemble of individuals able to engender fertile offspring by interbreeding [Mayr 1942]. Also this definition is hard to apply literally for most of the living organisms, considering asexual reproduction but also within the range of sex "quantity" variation from cases of "too little sex" (thelytoky) up to "too much sex" (hybridization) [Templeton 1989].

Likewise ecological communities, eukaryotic genomes contain a variable number of more or less abundant elements of different genetic classes: transposon-derived elements, satellite repetitive sequences, and their less abundant functional sequences such as RNA or genes. Here, in order to follow up our simile with ecology we had to decide what should be consider as a "species" in genomes. The decision here was difficult, but we decided to use the last level (with lower number of individuals) of classification that allows a functional definition of the sample, with no direct description of the sequence itself. We believe that the level of hierarchy that allow us to do so is the "family" or "class" level in the Repbase ontology used through RepeatMasker. For simplicity we talk of genetic species (GSs) of different classes, putting together all repetitive elements and biotypes, as individuals of different species.

Application of ecologist' methodology on genomic data

Here, taking advantage of the methods and models developed by ecologists we ask: is there a common pattern behind the relative abundance and diversity of GES in genomes? And, in the case that such pattern exists, is it sufficient to explain together diversities of functional and non-functional components in eukaryote genomes? To what extent abundance and diversity of genome components reflects adaptive or stochastic outcomes? Here we test the statistical adjustment of the UNTB predictions 31 different eukaryote genomes.

To achieve this objective we discuss results in three different sections. First we analyze genomes and chromosomes by virtue of relative species abundance (RSA) curves, classical graphical tools used in ecology to know

if genomes and chromosomes display uneven species distributions as is universally observed in ecological communities. Second we simulate the random distribution of all elements of genomes in all their chromosomes to statistically test the role of chance in chromosome design. Third, we test the statistical adjustment of the neutral ecological theory of biodiversity to the relative abundance and diversity of functional and non-functional elements of eukaryote chromosomes.

We conclude that abundances and diversity of GEs in most chromosomes is predicted by the stochastic dynamics of a model for which the principle of functional equivalence among elements is the primary assumption. Finally, we present a strong test to the hypothesis. If functional and non-functional GEs are distributed stochastically in chromosomes, their length must be predicted by demographic parameters only. Ecologists assert that effects of natural selection are dispensable to model abundance and species diversity in tropical forest [Jabot & Chave 2011]. Paralleling ecological communities Darwinian dynamics seems to be irrelevant to define abundance and diversity of genome components.

1.3. Genomic study of selective pressures in set of genes

In past years, with the development of genomic data in close related species, an effort was made in order to detect signals of selective pressure by applying methods that have been developed since [Kimura 1985] based on observations of significant deviation from neutrality for a given gene. These methodologies conceived for the study of a single gene were successfully able to detect genes escaping neutrality ($\omega \neq 1$ see Equation 1.1 page 13), and in particular positively selected genes (with $\omega > 1$) [Arbiza *et al.* 2006, Bakewell *et al.* 2007, Bustamante *et al.* 2005, Clark *et al.* 2003, Nielsen *et al.* 2005]. However none of those works were able to find, within the groups of genes detected to be under positive selection, a significant enrichment of a given functional trait. It is true that taking all those results together, assiduous readers could perceive functional patterns raising from all those studies. Functional terms related to *Sensory perception*, *Immune response* or *Regulation of transcription* were present in almost all genomic studies of positive selection conducted in primates

or rodents.

1.3.1. Detection of adaptation at molecular level – single gene approach

The result of selective pressures on DNA sequences are usually inferred by comparing the number of changes supposed to have a functional impact, with the number of changes observed in regions supposed to be escaping natural selection.

In the specific context of coding regions of the genome, changes occurring at nucleotide level can be divided in two kinds depending on whether they will be reflected in translated protein sequences or not (respectively non-synonymous and synonymous changes). Even if several works outlined the footprint of natural selection in biases of synonymous changes through codon usage (see reviews [Hershberg & Petrov 2008, Plotkin & Kudla 2011]), it is still assumed that its stranglehold on those silent sites is weak [Yang & Nielsen 2008] and its usage as proxy for neutral mutation rate is used since 1980 [Miyata *et al.* 1980].

On the other hand, the rate of non-synonymous mutations is theoretically assumed to be subject to selective pressure as mutations occurring at those sites may have functional consequence by changing the protein sequence. Moreover its rate is significantly lower and present more variation from one gene to another when compared to the rate of silent mutations in consequence of the amount of purifying selection [Kimura 1985].

Thus, assuming the proxy that silent mutations are neutral, the comparison of synonymous and non-synonymous mutation rates makes protein coding regions a perfect case in point for measuring the impact of natural selection within DNA sequences. Selective pressure can therefore be directly deduced from the ratio of non-synonymous mutation rate (dN) over synonymous mutation rate (dS), this value being associated to ω:

$$\omega = \frac{dN}{dS} \tag{1.1}$$

widely used [Pond *et al.* 2005]

Measured in coding regions using the Statistical methods to test for neutrality [Nielsen 2001], 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 & Sonnhammer 2003]. 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 & Sharan 2008, 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 & Alm 2008] 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 individual 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.

1.4. Side-products – Implementation of software and pipeline

1.4.1. Ecolopy

Since the definition of neutral models in ecology [Hubbell 2001, Volkov *et al.* 2003] computational tools were developed in order to manipulate data collected by ecological sampling, and to apply specific statistical test over them. The main tools developed in this sense are 3.

The more complete tool takes its name directly from the neutral model it allows to test, the package untb [Hankin 2007]. This package is implemented in R language [Team 2011] and stands ahead of other R packages allowing to deal with ecological data and to execute most classical statis-

tical analysis that ecologist have been developing [Borcard *et al.* 2011]. These packages have as main advantage to be usable all together and to make available a great fan of statistical tools. The main default of the package being relate to the language used that lacks of computational efficiency.

Another suite of scripts has been implemented by Etienne in the context of its publication [Etienne 2005]. These scripts or little programs were implemented in the unique context of testing the UNTB and are actually mainly taken up by the untb R package. The main advantage of this tool comes down to its simplicity of use and its speed. Still it is designed for limited samples, and the calculation of the fit to neutral model of one community of 200 species and 20 thousands individuals can take up to one day. Also the use of PARI/GP language is perhaps not the best choice in the context of bioinformatics where taking advantage and merging the work of the community is one of the main source of advances.

Finally the programs that allows to compute the fit of neutral models, and works faster are Tetame and Parthy implemented by Jabot and Chave [Jabot & Chave 2011,Jabot *et al.* 2008]. Those programs are implemented in C++ language and works very fast, their only default are the lack of flexibility, it is complicated to extract from the data more information that the one thought by Jabot and Chave.

All these packages allow to get the fit of sampling data to the UNTB plus in some cases the possibility to "browse" most classical descriptive ecological statistics. But at the time to test for the UNTB in genomic data none was able to deal with the high numbers of the samplings in genomes and chromosomes. And this was the main purpose of developing a new package able to deal with such large numbers.

Ecolopy is the name of the package developed for this purpose. It is a fully developed package, but still lacks of the whole set statistical tools available in R. However its design and the use of python [Van Rossum & Drake 2003] were though in order to provide scalable program architecture. Python language, is a programming language that offers a strong support for integration with other languages and tools, and whose popularity is raising among the bioinformatics community [Bassi 2007b]. Alternatively Ecolopy takes advantage of the GNU Multiple Precision (GMP) [Granlund 2000] library, and the Multiple-Precision Binary Floating-Point (MPFR) [Fousse *et al.* 2007] library, through the fast multiprecision GMPY module.

In this thesis we are dedicating a short section to go through some of the main features and advances of the program.

1.4.2. Pipeline for the detection of molecular evolution

In the last section of this thesis I am going to emphasize on different tools implemented in the context of improving and facilitating the different steps conducting to the detection of selective pressures at gene level. Also I will quickly present the most relevant works that took advantage of these tools.

Detecting selective pressures at genomic level consists in 6 main steps itemized as:

- **Species selection**: first of all we need to define a seed species or sequence. Starting from this seed, we have to select a set of sequences not to distant from our seed in order to avoid saturation of synonymous changes [Gojobori 1983, Smith & Smith 1996] (as an example, from human, we should remain within mammals). Usually it is recommended to have at least 4 sequences [Yang 2009].

- **Homologous sequences retrieval**: once selected our set of species the next step consists in retrieving homologous sequences. Most popular alternatives for doing this are *Ensembl* [Flicek *et al.* 2011] or the database resources of the NCBI [Sayers *et al.* 2011].

- **Selection of isoforms**: This step can be archived before or after the retrieval of homologous sequences. The easiest and fastest way being to select isoforms over a given set of homologous genes. The implementation of the selection of isoform was necessary since most of the public databases are using longest transcript of each gene in their analysis, and this can be an error given that many times transcripts are lost in several species.

- **Alignment**: alignment step is usually consider the simplest step and it is one the fastest. However when measuring selective pressure, and in particular for the detection of positive selection, misalignment have a high weight in the proportion of false positives. Most popular tools used here are MUSCLE [Edgar 2004], MAFFT [Katoh *et al.* 2005], Dialign [Subramanian *et al.* 2008] or T-COFFEE [Notredame 2010]. Alignments, once calculated, are usually trimmed

with Gblocks [Talavera & Castresana 2007] or Trimal [Capella-Gutiérrez *et al.* 2009] in order to abnormally divergent remove columns from the alignments.

- **Phylogenetic reconstruction**: this step is unnecessary when species tree is known (what now represents most of the cases). In any case, we will see how to construct fairly phylogenetic trees using model testing [Posada 2008, Abascal *et al.* 2005] and PhyML [Guindon & Gascuel 2003].

- **Identification of selective pressure and testing of evolutionary hypotheses**: this is the final step of our analysis. It concentrate most of the computation, and was a real challenge to automatize. The programs used to calculate selective pressures and fit evolutionary models are SLR [Massingham & Goldman 2005] and CodeML from PAML package [Yang 2007]. We used them through the ETE package [Huerta-Cepas *et al.* 2010] and developed for this a plugin able to bind SLR and CodeML and summarize their outputs.

We will see how to build this pipeline, and some of its applications.

2. Material and Methods

2.1. Measuring DNA complexity

2.1.1. The complexity ratio and complexity value

Complexity Ratio (CR) is defined by a classical formula used in data compression [Adjeroh *et al.* 2008]. It is the result of three transformation steps of a given sequence. First, the Burrows-Wheeler transform (BWT) [Burrows & Wheeler 1994], second the Move To Front (MTF) [Ryabko 1980] algorithm and finally the summarizing of the unexpected dispersion of the values obtained through Shannon's entropy [Shannon 1948] (see Table 2.1 for an example of the process). 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. Shannon's entropy is zero –this is the minimum– only when a sequence consists just of a single repeated symbol, which is the simplest possible combinatorial structure. Conversely, when entropy is equal to one (the maximum entropy), it indicates that the sequence has a random-like combinatorial structure.

Algorithmically, the BWT of a given sequence is a permutation of the symbols that represents the lexicographic order of all possible rotations of the sequence. The MTF transforms a given sequence into a list 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:

Given a sequence, $seq = AACCTTCGTAGCATGG$:

#	Rotating sequence	*I.*	BWT	Char. list				MTF
0	AACCTTCGTAGCATG**G**|	0	G	**G**	a	t	c	0
1	ACCTTCGTAGCATGG|**A**	1	A	G	**A**	t	c	1
2	CCTTCGTAGCATGG|A**A**	5	T	A	g	**T**	c	2
3	CTTCGTAGCATGG|AA**C**	7	C	T	a	g	**C**	3
4	TTCGTAGCATGG|AAC**C**	15	G	C	t	a	**G**	3
5	TCGTAGCATGG|AACC**T**	13	A	G	c	t	**A**	3
6	CGTAGCATGG|AACCT**T**	6	T	A	g	c	**T**	3
7	GTAGCATGG|AACCTT**C**	11	C	T	a	g	**C**	3
8	TAGCATGG|AACCTTC**G**	12	G	C	t	a	**G**	3
9	AGCATGG|AACCTTCG**T**	2	A	G	c	t	**A**	3
10	GCATGG|AACCTTCGT**A**	9	T	A	g	c	**T**	3
11	CATGG|AACCTTCGTA**G**	4	C	T	a	g	**C**	3
12	ATGG|AACCTTCGTAG**C**	3	G	C	t	a	**G**	3
13	TGG|AACCTTCGTAGC**A**	14	T	G	c	**T**	a	2
14	GG|AACCTTCGTAGCA**T**	10	A	T	g	c	**A**	3
15	G|AACCTTCGTAGCAT**G**	8	C	A	t	g	**C**	3

$\Rightarrow CR(seq) = E(MTF(BWT(seq))) = E(0, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3) = 0.593$

Table 2.1.: CR explained by example.

These three tables summarizes the steps needed to obtain the final sequence of number from which we will finally compute Shannon's entropy. 1) The table on the left corresponds to the BWT. Original sequence is rotated sequentially (first character moved to back) resulting in different strings, as many as characters in the sequence. The resulting sequences are then sorted in lexicographic order. The "*I.*" column corresponds to the Index of this ordering (e.g.: the third sequence here in original order "#" takes the fifth position in lexicographic order). 2) The table in the center corresponds to the result of the BWT, that is the last character of previous sequences ordered as explain. 3) The table on the right corresponds to the application of the MTF algorithm. Starting from a sequence of all character named here "Char. list" (our four nucleotides in this case), the MTF will get the index of the current nucleotide from the BWT (upper case bold letter) in the "Char. list". In a second step, for the next iteration, MTF will transform the "Char. list" bringing to front the character present in previous BWT result (upper case letter).

Finally, we compute the Shannon's entropy of the values obtained by the MTF that results in our CR (the CV is obtained by multiplying CR by the length of the sequence).

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

For DNA alphabet entropy is defined as:

$$E(s) = -\sum_{i=0}^{3} p_{(i)} \times log_4(p_{(i)}) \tag{2.2}$$

With i the index of characters used, for nucleotides from 0 to 3. 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.

2.1.2. Complexity in strings

Genomic sequences

Complete genomes of 54 species were downloaded from *NCBI* database resource [Sayers *et al.* 2009] 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 their interest as model species or the presence of particular biological features such as: variation in genome size, ancestral or recent polyploidy, living in extreme environments, living as intracellular parasites, gene expansion, genome reduction, RNA or single-strand DNA genomes, and synthetic genomes Table 3.1. 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 excluded in the measure of chromosome length. Eukaryote's genome complexity was calculated over concatenated chromosomes.

Complexity in biological sequences was computed in the +1 strand. Analysis of -1 strand provided no differences in results.

Random sequences with different ploidy levels were also needed for the study, they were generated with *python* [Van Rossum & Drake 2003]. Complexity value of biological sequences and random sequences was computed with the DNA alphabet of four letters.

Annotation of repetitive elements

Interspersed repeats and low complexity DNA sequences were screened and mapped in genomes of all our 54 species using RepeatMasker program [Smit *et al.* 2010] (see details of RepeatMasker output for individual species in Appendix A). Libraries of genetic elements were retrieved from RepBase [Jurka *et al.* 2005], the last version available at time was used (20^{th} of September 2011). An example of summary file given by RepeatMasker is shown in Appendix A.

Complexity of major families of repetitive elements such as DNA transposons, LTRs, LINEs, SINEs, satellites and exons, introns, and complete genes (considering untranslated regions) was computed after concatenation of all elements conserving their original order in chromosomes.

Human texts

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, that takes into account all characters found, including mathematical symbols, and different punctuation signs.

Complexity in windows

To study complexity along chromosomes, a sliding window method shifting along chromosomes in overlapping units of 1.0 Kb to 100 Mb was performed. Standard linear models and linear models with interactions were run in R language [Team 2011].

2.1.3. Simulations

We performed four kinds of experiments where CV and CR were computed. First: random polyploid construction of sequences of various sizes and ploidy levels (1× to 10×). 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 & Crowell 2000] , and *Arabidopsis thaliana*: $7.1e^{-09}$ [Ossowski *et al.* 2010]) over random sequences, and 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 1.0 Kb transpositions between chromosomes. The number of transposition per generation was set as a constant function of genome size (genome size over 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.

2.2. Measuring dynamics of genetic species

2.2.1. Genomes

For the study of dynamics of genomic elements Genomic sequences of 31 species from unicellular eukaryotes to mammals were used, all were extracted from previous work see subsection 2.1.2 and Table 3.1. The complete list of species used is: *1) Gallus gallus* (Birds) *2) Taeniopygia guttata* (Birds) *3) Danio rerio* (Fishes) *4) Oryzias latipes* (Fishes) *5) Tetraodon nigroviridis* (Fishes) *6) Saccharomyces cerevisiae* (Fungi) *7) Anopheles gambiae* (Invertebrates) *8) Caenorhabditis elegans* (Invertebrates) *9) Drosophila melanogaster* (Invertebrates) *10) Tribolium castaneum* (Invertebrates) *11) Bos taurus* (Mammals) *12) Canis familiaris* (Mammals) *13) Equus caballus* (Mammals) *14) Homo sapiens* (Mammals) *15) Macaca mulatta* (Mammals) *16) Monodelphis domestica* (Mammals) *17) Mus musculus* (Mammals) *18) Pan troglodytes* (Mammals) *19) Pongo abelii* (Mammals) *20) Rattus norvegicus* (Mammals) *21) Arabidopsis lyrata* (Plants) *22) Arabidopsis thaliana* (Plants) *23) Brachypodium distachyon* (Plants) *24) Oryza sativa* (Plants) *25) Populus trichocarpa* (Plants) *26) Sor-*

ghum bicolor (Plants) *27) Zea mays* (Plants) *28) Dictyostelium discoideum* (unicellular Eukaryotes) *29) Plasmodium falciparum* (unicellular Eukaryotes) *30) Thalassiosira pseudonana* (unicellular Eukaryotes) and *31) Ciona intestinalis* (Urochordate) .

2.2.2. Mining of Genetic Species

For this study, we define genetic species (GSs) as the conjunction of 2 categories, repetitive elements and functional elements.

Repetitive Elements

Repetitive elements were mapped following the methodology explained in section 2.1.2.

To measure dynamics of genetic elements we had to define a level to consider as "species" in the RepBase ontology (see Introduction 1.2.3). We decided to consider "species" those class of repeats that can be defined functionally. What would correspond to superfamilies of transposable elements according to [Wicker *et al.* 2007] or, also, to RepBase classification [Kapitonov & Jurka 2008].

Complete list of superfamilies mapped is shown in Table 2.2

Functional Elements

Functional elements correspond to biotypes category of the genes according to Ensembl [Flicek *et al.* 2011] nomenclature. They were retrieved using the Biomart API [Kinsella *et al.* 2011]. The non-redundant list of function elements across all species is shown in Table 2.3.

Note that pseudogenes were not included in that list in order to keep the functional aspect of this family of GSs.

2.2.3. Randomization of genetic elements

In order to test for the random distribution of GSs among chromosomes of each genomes, we generated 1,000 genomes, corresponding to each species, with a random distribution of GSs. Taking only in consideration chromosome length.

Family	Superfamilies
ARTEFACT	ARTEFACT
DNA	DNA, Chapaev, Chapaev-Chap3, Crypton, En-Spm, Ginger, Harbinger, Kolobok-T2, Maverick, Merlin, Mirage, MuDR, NOF, P, PiggyBac, TcMar, TcMar-ISRm11, TcMar-Mariner, TcMar-Pogo, TcMar-Stowaway, TcMar-Tc1, TcMar-Tc2, TcMar-Tc4, TcMar-Tigger, Tourist, Transib, Zator, hAT, hAT-Ac, hAT-Blackjack, hAT-Charlie, hAT-Pegasus, hAT-Tag1, hAT-Tip100, hAT-Tol2, hAT-hobo
LINE	LINE, CR1, DRE, Dong-R4, I, Jockey, L1, L1-Tx1, L2, LOA, Penelope, R1, R2, R2-Hero, RTE, RTE-BovB, RTE-X, telomeric
LTR	Caulimovirus, Copia, DIRS, ERV, ERV1, ERVK, ERVL, ERVL-MaLR, Gypsy, Pao
Low complexity	Low complexity
Other	Other, Composite, centromeric, subtelomeric
RC	Helitron
RNA	RNA
SINE	SINE, 5S, Alu, B2, B4, BovA, CORE, Deu, ID, L1, MIR, RTE, RTE-BovB, V, tRNA, tRNA-CR1, tRNA-Glu, tRNA-Lys, tRNA-RTE
Satellite	Satellite, W-chromosome, Y-chromosome, acro, centr, macro, telo
Simple repeat	Simple repeat
Unknown	Unknown, Y-chromosome
rRNA	rRNA
scRNA	scRNA
snRNA	snRNA
srpRNA	srpRNA
tRNA	tRNA

Table 2.2.: Superfamilies of repetitive elements and description.
Superfamilies mapped by RepeatMasker using RepBase library.

Biotype	Description
IG C gene	Immunoglobulin constant segment
IG D gene	Immunoglobulin diversity segment
IG J gene	Immunoglobulin joining segment
IG V gene	Immunoglobulin variable segment
IG Z gene	Immunoglobulin gene found in Zebrafish
MRP RNA	mitochondrial RNA-processing RNA
RNase MRP RNA	enzymatically active ribonucleoprotein
RNase P RNA	enzymatically active ribonucleoprotein
SRP RNA	signal recognition particle RNA
TR C	T cell receptor constant domain
TR J	T cell receptor joining domain
TR V	T cell receptor variable domain
class I RNA	class of small non-coding RNA
class II RNA	class of small non-coding RNA
lincRNA	large intervening non-coding RNA (multiexonic non-coding RNA)
miRNA	micro RNA
misc RNA	miscellaneous RNA
ncRNA	non-coding RNA
processed transcript	Non-coding transcript without open reading frame (ORF).
protein coding	Contains an open reading frame (ORF)
rRNA	Ribosomal RNA
retrotransposed	non-coding pseudogene produced by integration of a reverse transcribed mRNA into the genome
snRNA	small nuclear RNA
snlRNA	small nuclear like RNA
snoRNA	small nucleolar RNA, involved in modifications of other RNAs
tRNA	transfer RNA
transposable element	transposable element

Table 2.3.: Biotype and description.
Summary table of the biotypes used, and short description retrieved from *Ensembl* glossary [Flicek *et al.* 2011] and from Sequence Ontology browser [Eilbeck *et al.* 2005].

In order to discard centromeric regions, chromosome sizes were not directly inferred from sequence length. Here, we define the size of a chromosome as the sum of all 10 kilobase windows containing at least 1 GS (see Table 2.4).

Then GSs of each genome were distributed among chromosomes, according to a probability dependent of the size of the chromosome. As an example, in Human, it was around 6 times likely for a GS to belong to chromosome 1 than chromosome 22 (respective lengths are 225 megabases and 35 megabases)

Chr	Chr length	Corrected length	Percentage left
1	249,240,621	225,200,000	90.35%
2	243,188,741	237,670,000	97.73%
3	197,961,181	194,230,000	98.12%
4	191,044,271	187,270,000	98.02%
5	180,901,928	177,090,000	97.89%
6	171,048,878	167,050,000	97.66%
7	159,128,663	154,640,000	97.18%
8	146,302,151	142,290,000	97.26%
9	141,151,937	120,130,000	85.11%
10	135,524,747	131,040,000	96.69%
11	134,946,516	130,310,000	96.56%
12	133,841,891	129,970,000	97.11%
13	115,109,733	95,500,000	82.96%
14	107,289,415	87,910,000	81.94%
15	102,521,389	81,520,000	79.52%
16	90,290,985	78,640,000	87.10%
17	81,195,208	77,700,000	95.70%
18	78,017,245	74,580,000	95.59%
19	59,118,983	55,460,000	93.81%
20	62,962,324	59,430,000	94.39%
21	48,119,895	35,100,000	72.94%
22	51,244,541	34,790,000	67.89%
X	155,260,558	150,230,000	96.76%
Y	59,033,288	22,520,000	38.15%

Table 2.4.: Transformation of Chromosome size. Example of changes in estimation of chromosome length after removing regions with no GSs for Human chromosome 1.

2.2.4. Ecology

Several packages or programs were already developed in order to deal with species abundances data, implementing statistical functions in order to fit

data in ecological models and even able to test for neutrality [Jabot & Chave 2011, Etienne 2007, Hankin 2007]. However none of those programs were able to deal with genomic data, with abundances in the order of the million of individuals specially in the case of Etienne's model where computation of $K(D, A)$ uses stirling numbers (see section 2.2.5, equation Equation 2.10). In order to adapt the algorithm to genomic dataset we developed the *Ecolopy* package, that, as a main point, uses the GMP [Granlund 2000] and MPFR [Fousse *et al.* 2007] libraries through GMPY biding [Martelli 2007]. Other improvements specific to Ecolopy, and needed for dealing with genomic dataset where done.

Ecolopy is entirely written in Python [van Rossum & de Boer 1991], a programming language that offers a strong support for integration with other languages and tools, and whose popularity is raising among the bioinformatics community [Bassi 2007a]. Ecolopy is still a fully ripened package, but it was designed to provide a scalable program architecture.

More details about the package will be addressed in 6.3.

2.2.5. Neutral Ecological models

Ewens sampling formula

Ewens sampling formula [Ewens 1972] (Equation 2.6) was originally designed in order to describe the number of different alleles expected to be observed in a given sample. However, the formula can be applied to other fields. In the context of the study of ecological communities its application was first suggested was suggested by Tavaré and Ewens [Tavaré & Ewens 1997] and finally implemented by Hubbell [Hubbell 2001]. Hubbell proposed a model defining the Fundamental Biodiversity parameter θ (Equation 2.5) given the speciation rate ν and J_M the size of the metacommunity.

$$\theta = 2J_M\nu \tag{2.5}$$

The estimation of θ alone is sufficient to apply directly Ewens Sampling formula (Equation 2.6), and to compute its likelihood for given a community (Equation 2.7).

$$Pr\{S, n1, n2, \ldots, n_S|\theta\} = \frac{J_M!\theta^S}{1^{\phi 1}2^{\phi 2}\cdots J_M^{\phi J_M}\phi_1!\phi_2!\cdots\phi_{J_M}!\prod_{k=1}^{J_M}(\theta+k-1)} \tag{2.6}$$

Here n_i corresponds to the abundance of species i and ϕ_a the number of species with abundance a.

$$\mathcal{L} = \frac{\theta^S}{\prod_{k=1}^{J_M}(\theta+k-1)} \tag{2.7}$$

Etienne sampling formula

The main problem with Hubbell's model using Ewens sampling formula is the assumption that migration is unlimited ($m = 1$). However a new sampling formula was presented recently [Etienne 2005] including cases where $m < 1$, taking into account the number of immigrants I depending on the sample size J:

$$m = \frac{I}{I+J-1} \tag{2.8}$$

Given this Etienne's sampling formula is postulated as:

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J}\phi_J!}\frac{\theta^S}{(I)_J}\sum_{A=S}^{J} K(D, A)\frac{I^A}{(\theta)_A} \tag{2.9}$$

with $K(D, A)$ as:

$$K(D, A) := \sum_{\{a_1,\ldots,a_s|\sum_{i=1}^{S} a_i = A\}}\prod_{i=1}^{S}\frac{\bar{s}(n_i, a_i)\bar{s}(a_i, 1)}{\bar{s}(n_i, 1)} \tag{2.10}$$

Calculation of stirling numbers here are the main computational bottle neck as mentioned in [Etienne 2005]. A solution was given by [Jabot *et al.* 2008] and implemented in the Tetame program takes advantage of the recurrence function (Equation 2.11), that allows to build a table of values, given the dispersion of the ranked abundance of species, in stead of computing them directly for each pair of values. On top of this strategy,

some improvements were developed in order to deal with genomic data (section 6.3).

$$S_{(n,m)} = S_{(n-1,m-1)} - (n-1) \times S_{(n-1,m)} \tag{2.11}$$

Once computed $K(D, A)$ we are able to optimize the likelihood of the model (Equation 2.12) by varying the values of the parameters θ and m (see Model optimization subsection 2.2.6) for a given dataset.

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{J=1}^{J} \Phi_J!} \frac{\theta^S}{(\theta)_J} \times \sum_{A=S}^{J} \left(K(D, A) \frac{(\theta)_J}{(\theta)_A} \frac{I^A}{(I)_J} \right) \tag{2.12}$$

2.2.6. Model optimization

Models where optimized through different optimization strategies depending on the model selected. In the case of the Ewens' formula, θ is the only parameter to take into account, and its estimation is achieved with the *golden* optimization strategy [Jones *et al.* 2001]. For Etienne's model, two parameters were optimized, θ and m, using the best solution of different optimization strategies (see 6.3 for more details).

Optimization step being critical specially under Etienne's model, the likelihood surface of the model given a range of values of θ and m was drawn for some of the chromosomes in our dataset. This procedure allows us to estimate graphically the best solution for both parameters. The solution found by this methodology was then compared to the optimization result, in order to validate them (see Figure 2.1 as an example of this validation step). Computation time needed to generate such likelihood contour plots prevents using it for all chromosomes, but, for the 5 chromosomes tested, results were congruent.

2.2.7. Model testing

In order to compare and test the fit of the two models computed, a likelihood ratio test [Wilks 1938] was conducted thanks to the fact that Etienne's model is nested into Ewens'. Etienne's model has two free parameters (FP) while Ewens' only one (under this model m parameter is fixed to

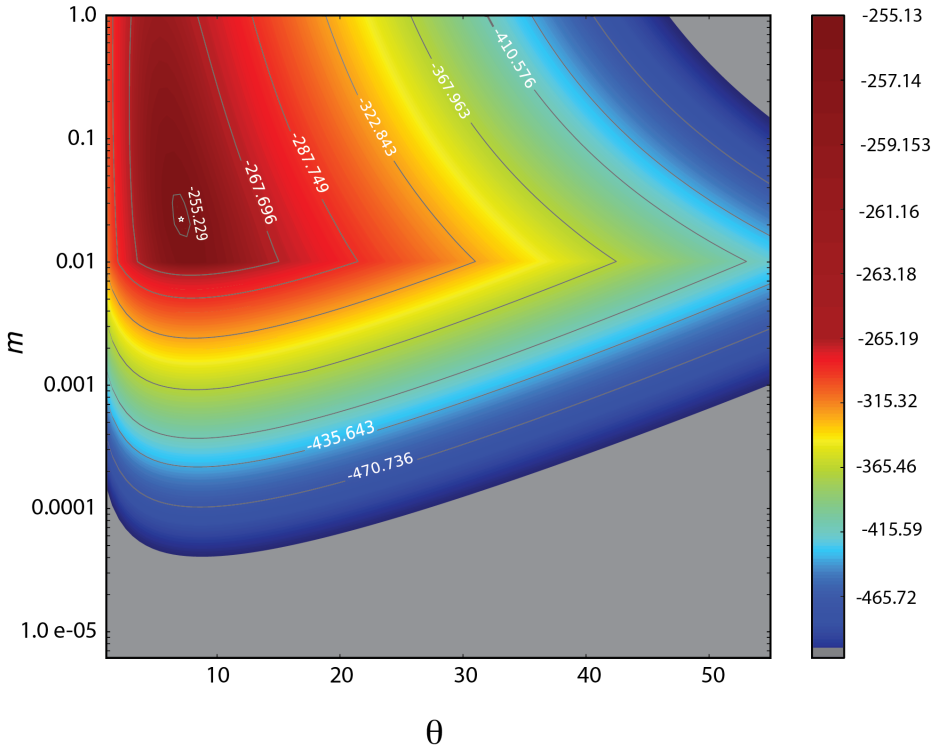


Figure 2.1.: Maximum likelihood inference of neutral parameters. Log likelihood surface as a function of migration rate (m), and the fundamental biodiversity number (θ) for *D. rerio* chromosome 19. Dark red color shows regions of the surface where parameters maximize the probability to explain abundances and diversity of genetic elements in the chromosome. Likelihood ratio tests favored Etienne in contrast to Ewens sampling formula to explain the observed data in the chromosome.

1), thus the number of degrees of freedom for the chi-squared distribution is 1 ($df = FP_{Etienne} - FP_{Ewens} = 1$).

Etienne's model was thus kept as best fit model for chromosomes that pass the LRT computed between the two models, otherwise the null model using Ewens' formula was selected.

2.2.8. Testing UNTB

In the last years at least two tests were developed in order to accept or reject the neutrality of a given community. Both tests are based on the comparison of a given number of random neutral community to the observed distribution of abundances. Random neutral communities being generated using the parameters estimated (see subsection 2.2.6) for the real data under a given neutral model (either Ewens or Etienne model). The comparison of the random neutral communities with the observed distribution of abundances, is the key point to test for neutrality.

The first of these tests [Etienne 2007] consists in comparing the distribution of likelihoods to fit neutral model. This corresponding distribution of random neutral abundances is then compared to the likelihood of the observed data. The major problem of this test is technical, the computation time needed to optimize the parameters of each abundance distribution and get the likelihoods is unrealistically too high when dealing with GSs.

The second test [Jabot & Chave 2011] uses, instead of likelihood, the comparison of Shannon's entropy [Shannon 1948] of each distribution of abundances, and is much faster as random neutral communities do not need to be fitted into a neutral model.

Thus, from the neutral parameters obtained for each chromosome, we simulated 10,000 distributions of abundances of GSs and computed, for each, its Shannon's entropy (H). Chromosomes were considered significantly non-neutral when the H of their abundances was below 95% of the 10,000 random neutral H values. As an example, Figure 2.2 shows the distribution of H for 10,000 random neutral communities generated under Etienne's model with S, J fixed to the observed numbers and θ and m corresponding to optimized values for *Anopheles gambiae's* chromosome 2L. In this figure the empirical value of H is below the 95% of the random neutral distribution, than this chromosome is considered to be non-neutral.

Additionally, given the large number of test performed (one per chro-

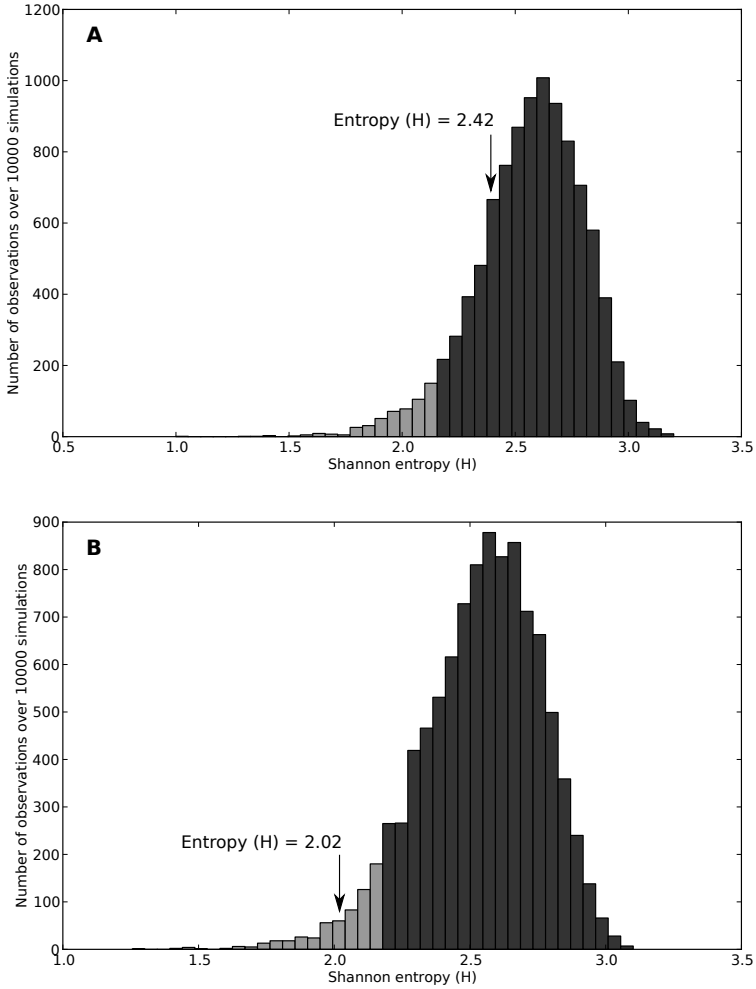


Figure 2.2.: Comparing simulated and empirical evenness.
Neutrality test statistically compares simulated null distribution of H with the empirical value. Here, the null distribution of H values were derived from 10,000 neutral simulations of (A) *H. sapiens* chromosome 1 and (B) *A. gambiae* chromosome 2L, with parameters (θ and m) optimized by ML using Etienne sampling formula. Light and dark gray bars display 5% and 95% of the simulated data, respectively. Although neutrality was not rejected in **B** (p= 0.291 and p= 0.041 for **A** and **B** respectively), posterior correction by multiple testing favored the neutral hypothesis in both cases (q= 0.609 and q= 0.159 for **A** and **B** respectively).

mosome, that is 548 tests), statistical significances were corrected by false discovery rate (FDR) [Benjamini *et al.* 2001]. Following with the example of Figure 2.2, after correction by FDR, *Anopheles gambiae's* chromosome 2L is brought back to neutrality.

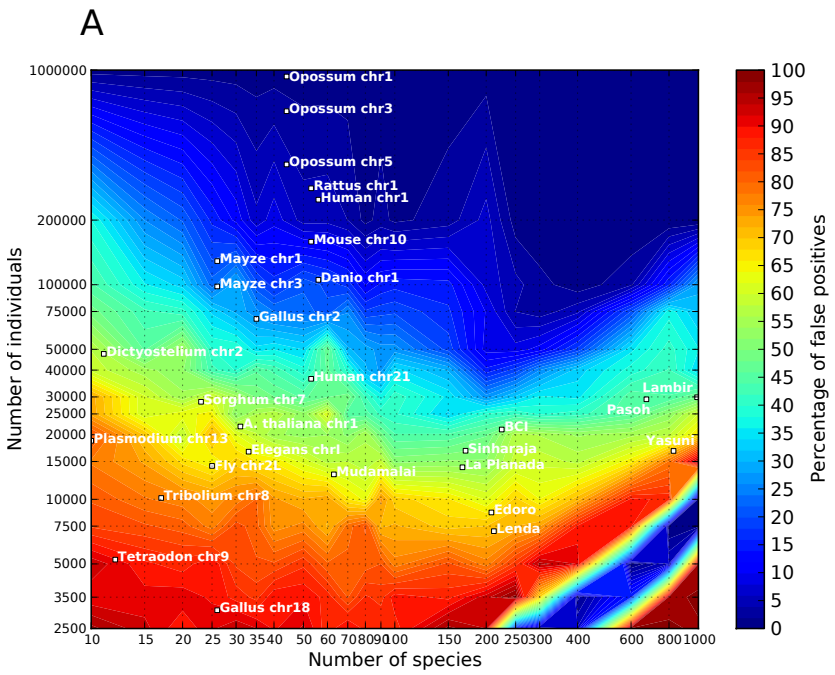
Given the lack of differences among results presented in the text of the paper, we replicated this test by fixing the number of species (S) according to the observed value of each chromosome. No differences were observed in relation to the number of chromosomes fitting in neutral models.

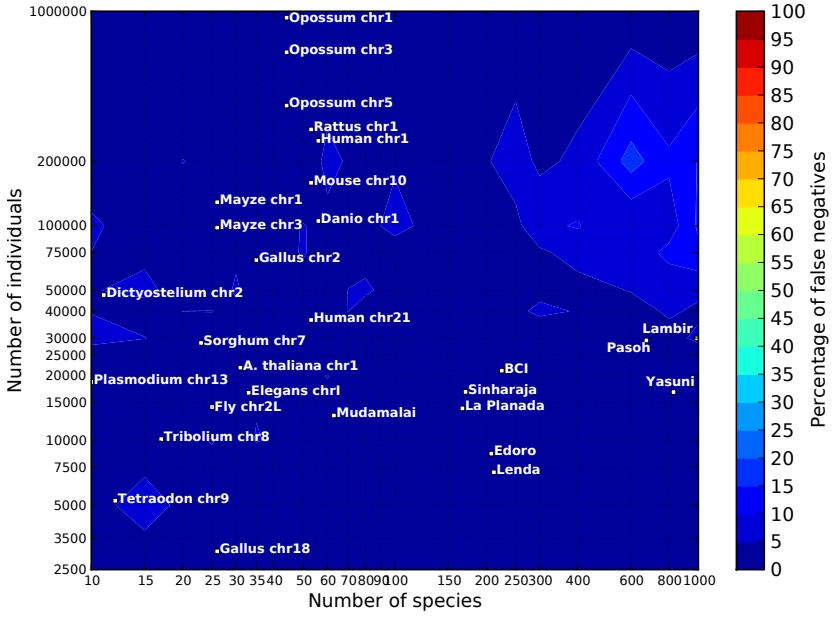
Power and specificity of neutral test

In order to validate the test of neutrality, we computed the proportion of false and true positives generating respectively random log-normal distributions and random neutral distributions. The results of the test of neutrality applied over log-normal or neutral random distributions are shown in Figure 2.3, respectively.

According to this result, we validate the power given that in the whole range of S and J, the proportion of true positives was very high (Figure 2.3B. Nevertheless these simulation pointed out come difficulties to differentiate log-normal distributions from neutral distributions. Specifically, when J is lower than 100,000 individuals, the proportion of false positives is never lower than 50%. However we decided to get through as this raise in false positive rate only affect the smallest chromosomes, and also because lognormal distributions are really close to neutral distributions, being able to differentiate them is not obvious at all [McGill *et al.* 2006].

Figure 2.3. ***(following page)*: Type I and Type II errors of the neutral test.**
False positive and false negative results of the neutral test were assessed for a variable range of species (S) and individuals (J). Neutral and log-normal distributions were assumed as null and alternative hypotheses, respectively. Panel A describes the proportion of times the test rejected the null hypothesis being true. Red regions describe the space where the proportion of false positive is too high. This is a dangerous area to test for neutrality. Panel B shows the percentage of times the test failed to reject the null hypothesis being false. Data of chromosomes and ecological communities15 are pointed in both panels.


A
Number of individuals
Number of species
Percentage of false positives
Opossum chr1
Opossum chr3
Opossum chr5
Rattus chr1
Human chr1
Mouse chr10
Mayze chr1
Mayze chr3
Danio chr1
Gallus chr2
Dictyostelium chr2
Human chr21
Sorghum chr7
A. thaliana chr1
Plasmodium chr13
Elegans chrI
Fly chr2L
Mudamalai
Tribolium chr8
Tetraodon chr9
Gallus chr18
BCI
Sinharaja
La Planada
Edoro
Lenda
Lambir
Pasoh
Yasuni


B
Number of individuals
Number of species
Percentage of false negatives
Opossum chr1
Opossum chr3
Opossum chr5
Rattus chr1
Human chr1
Mouse chr10
Mayze chr1
Mayze chr3
Danio chr1
Gallus chr2
Dictyostelium chr2
Human chr21
Sorghum chr7
A. thaliana chr1
Plasmodium chr13
Elegans chrI
Fly chr2L
Mudamalai
Tribolium chr8
Tetraodon chr9
Gallus chr18
BCI
Sinharaja
La Planada
Edoro
Lenda
Lambir
Pasoh
Yasuni

2.3. Detection of selective pressure at molecular level

2.3.1. 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 the seed species (*H. sapiens* in the case of mammals) and each one of the rest of the species was retrieved from *Ensembl Compara* [Vilella *et al.* 2009] using biomart [Kinsella *et al.* 2011] (see Figure 2.4 to have an insight of the phylogenies and distances). Only groups of orthologs annotated as "one-to-one" –with only one representative of each species– where kept in the final dataset.

The same procedure was applied for *melanogaster* group, including 6 species namely, *Drosophila melanogaster* (taken as seed-species), *Drosophila sechellia*, *Drosophila simulans*, *Drosophila yakuba*, *Drosophila erecta* and, as outgroup, *Drosophila ananassae* (see Figure 2.4-B).

2.3.2. Alignments refinement and filters

DNA coding sequences (CDS) were aligned according to protein translation pattern using *Muscle* version 3.7 [Edgar 2004] 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 & Li 2000] (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 & Kumar 2003], respectively, ortholog comparisons between primates and rodents with dS$\geq$1 and dN$\geq$0.5, rodents with dS$\geq$0.256, dN$\geq$0.122, and primates with dS$\geq$0.064 and dN$\geq$0.030 substitutions/site were excluded.

The number of orthologs kept for analysis after filtering steps, is 12,453

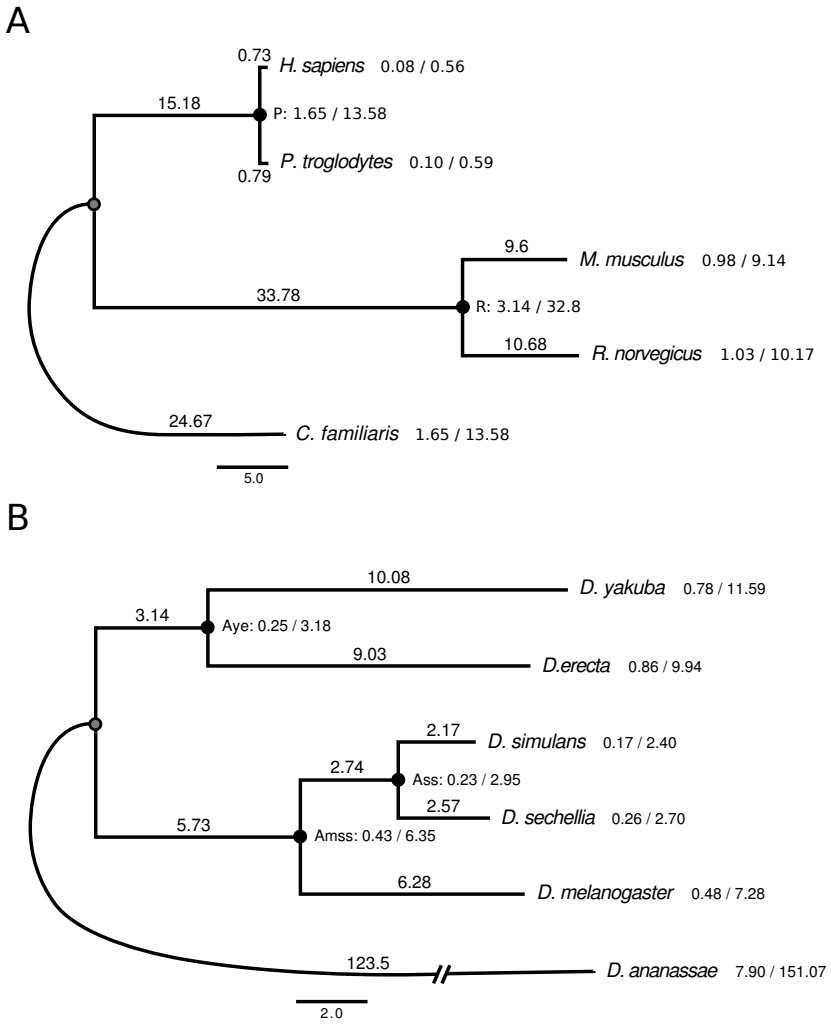


Figure 2.4.: Mammals and *melanogaster* group phylogeny.
Numbers on internal and external nodes represent the median number of non-synonymous and synonymous substitutions per codon (dN/dS) estimated from all coding sequences compared in mammalian (A) and Drosophila (B) genomes. Branch lengths and rates were multiplied by 100. Ancestral estimations 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.

for mammals, and 9,240 for flies.

2.3.3. Evolutionary analysis

Maximum likelihood estimation of dN, dS, and ω was computed using CodeML program from PAML [Yang 2007]. 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 & Gentleman 1996]) 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].

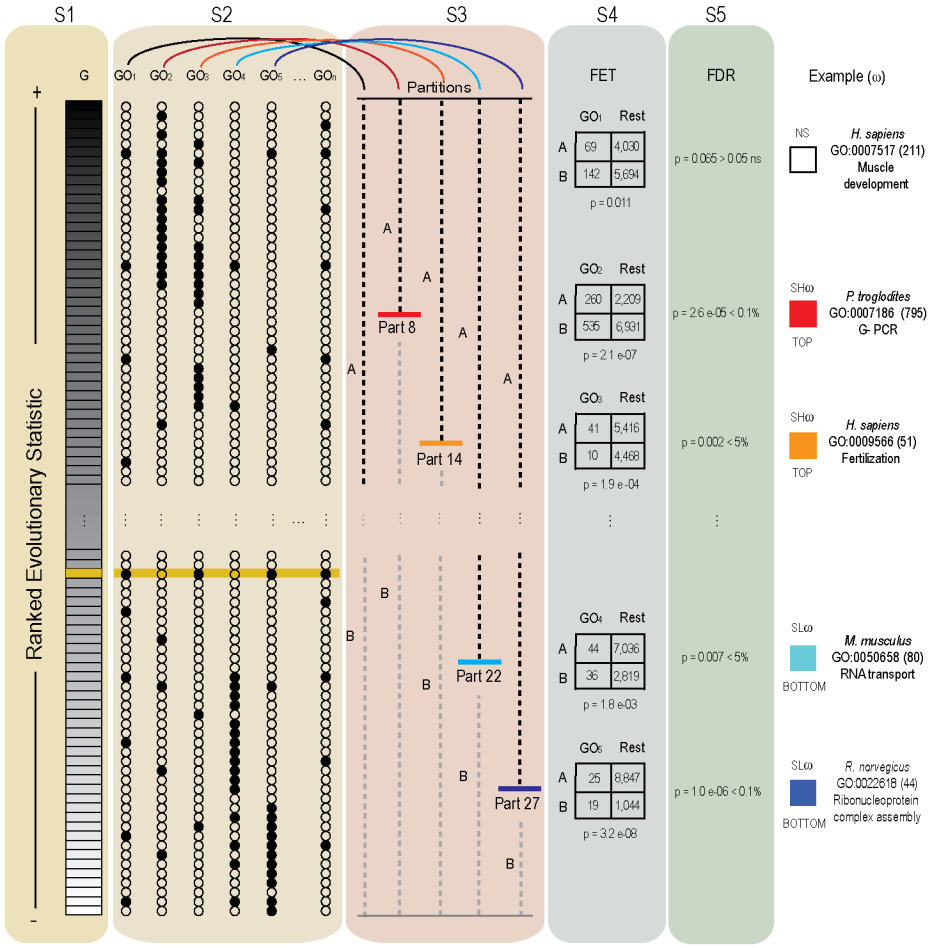
2.3.4. GSSA, evolutionary and statistical simulations

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 2.5) 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 2.5, 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 $C \cdot P$ tests were corrected by the widely accepted FDR method [Benjamini *et al.* 2001].

Figure 2.5. ***(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.

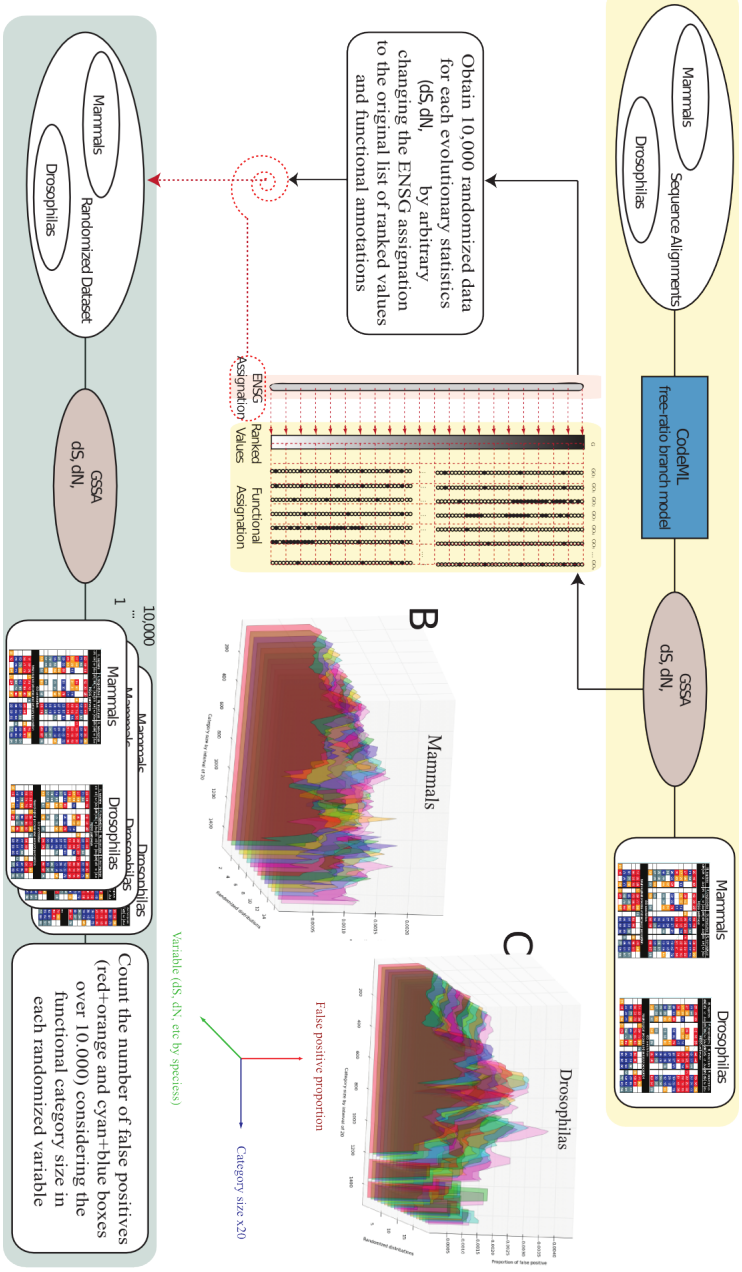

S1
S2
S3
S4
S5
G
GO1 GO2 GO3 GO4 GO5 … GOn
Partitions
FET
FDR
Example (ω)
Ranked Evolutionary Statistic
+
-
Part 8
Part 14
Part 22
Part 27
GO1 Rest
A 69 4,030
B 142 5,694
p = 0.011
GO2 Rest
A 260 2,209
B 535 6,931
p = 2.1 e-07
GO3 Rest
A 41 5,416
B 10 4,468
p = 1.9 e-04
GO4 Rest
A 44 7,036
B 36 2,819
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

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 2.6-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 2.6-B and Figure 2.6-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 2.7). The pipeline described here, shows three different areas:

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

Figure 2.6. ***(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
C
Drosophilas
False positive proportion
Category size x20
Variable (dS, dN, etc by species)
10,000
1
Mammals
Drosophilas
Mammals
Drosophilas
Randomized Dataset
GSSA
dS, dN,
Count the number of false positives
(red+orange and cyan+blue boxes
over 10,000) considering the
functional category size in
each randomized variable

lection (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 2.5. 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 2.5.: 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 a genome. Accordingly, since no net differences were produced after ranking

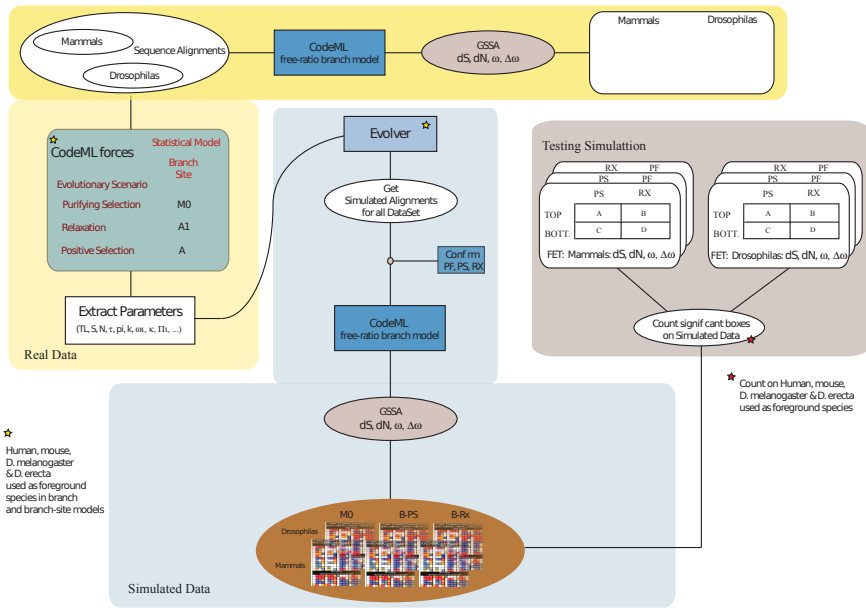


Figure 2.7.: Evolutionary and statistical simulation of GSSA.
The pipeline shows the steps taken along three different spaces of analysis, the real data, the simulated data and the testing block. See Supplementary Results for a complete explanation of methods and results.

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 2.6. 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 2.6.: Proportion of significant functional categories that are still significant (identical signs of odd-ratios) under a different evolutionary scenario.

Part I.

Structure and dynamics of genomes

3. Random-like structure of DNA

3.1. Results

3.1.1. Computing genome complexity

AMONG 20 major systematic groups we pick 54 species and computed the complexity value (CV) of their genomes with sizes ranging from 3.4Gb to 1.6Kb Table 3.1. The first striking observation was the degree of direct correlation observed between genome size and CV (Figure 3.1-A) with a slope of the regression equal to 0.967 very close to 1, that would imply a maximum complexity for all genomes. The residual variation around the fitted regression and along the 6 orders of magnitude, was almost null ($adjusted - R^2 = 0.987$).

Given the slope and the degree of adjustment that shows our set of living species, 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 *Monodelphis domestica* (size $\sim$ 3.41e+09 bp) as well as our *wired* organisms as:

- obligate endosymbionts bacteria with extreme reduction of genome size (*Carsonella ruddii*, *Buchnera aphidicola*, and *Ureaplasma urealyticum*) [Wernegreen 2002]
- parthenogenetic crustaceans with ubiquitous duplications of genes (*Daphnia pulex*)
- archean organisms living in extreme environmental conditions (*Sulfolobus islandicus*, *Methanocaldococcus vulcanius*, *Thermococcus sibiricus*)
- the first synthetic organism made by humans (*Synthetic mycoplasma mycoides*) [Gibson *et al.* 2010]

All of them fit the slope of the linear regression model.

In order to better contrast deviations from maximum complexity we computed the complexity ratio (CR), and the deviation to the maximum ratio (Dmax = 1- CR) for each species. According to Table 3.1, 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 [Colbourne *et al.* 2011], and two unicellular eukaryotes that curiously correspond to the 2 genomes sequenced with higher proportion of A + T, Plasmodium [Gardner *et al.* 2002] (A + T content around 81%) and Dictyostelium [Eichinger & Noegel 2003] (A + T content around 78%).

The highest CR=1 was obtained for randomly generated sequences with uniform distribution of ACGT. To simulate events of polyploidization random sequences were duplicated up to 5 times (corresponding to ×a 10X). The fall in CR reaching CR=0.25 for 10× could thus be compared to CR values of real polyploids placing maize genomes at the level of a perfect triploid and sorghum to the level of a perfect diploid. On top of the representation of CR of genomes and random sequences with different level of "ploidization" results of the computation of CR of human texts were added to Figure 3.1-B (values used can be found in Table 3.2).

All together, complexity ratios of complete genomes, random sequences of different ploidy and human language texts were computed. Maximum

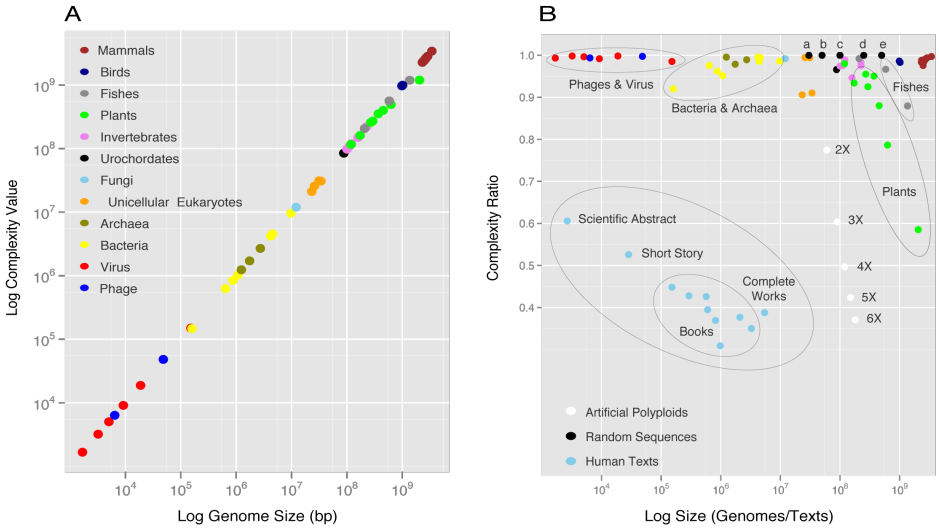
Table 3.1. ***(following page)*: Genomes Complexity.**
Genomes size (GS), genomes complexity (GC), genome complexity ratio ($GCR = \frac{GC}{GS}$), and deviation from the maximum GCR (Dmax=1-GCV) for 54 species of different taxa. NCBI accession number or Ensembl (*E!*) version (ACN-EV). ***Features***: **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/. (*) Clades are abbreviated as: *Vi*: Virus; *Ph*: Phage; *Ba*: Bacteria; *Ar*: Archaea; *Fu*: Fungi; *Ap*: Apicomplexa; *Am*: Amebozoa; *He*: Heterokonta; *Ur*: Urochordate; *In*: Invertebrates; *Pl*: Plants; *Fi*: Fishes; *Br*: Bryophyta; *Ma*: Mammals;

Feat.	Species ^{clade*}	ACN-EV	GS	GC	GCR	Dmax
RNA	Hepatitis B^{Vi}	NC3977.1	1,682	1,671	1	0
SGS RNA	Hepatitis D^{Vi}	D01075.1	3,215	3,210	0.9984	0.0016
SSD	Tomato mosaic^{Vi}	NC010836 NC10835.1	5,058	5,040	0.9964	0.0036
SSD	Enterobacteria phage m13^{Ph}	V00604	6,407	6,367	0.9938	0.0062
RNA	HIV 1^{Vi}	NC001802	9,181	9,105	0.9917	0.0083
RNA	Sudan ebolavirus^{Vi}	NC006432	18,875	18,842	0.9983	0.0017
DSD	Enterobacteria phage lambda^{Ph}	NC001416	48,502	48,381	0.9975	0.0025
DSD	Human herpesvirus1^{Vi}	NC001806	152,261	150,036	0.9854	0.0146
SBG IP RG	Carsonella ruddii^{Ba}	NC008512	159,662	146,930	0.9203	0.0797
IP RG	Buchnera aphidicola^{Ba}	AE013218.1	642,122	626,533	0.9757	0.0243
IP RG	Ureaplasma urealyticum^{Ba}	CP001184	873,755	840,812	0.9623	0.0377
SL	Synthetic mycoplasma mycoides^{Ba}	CP002027.1	1,078,809	1,026,444	0.9515	0.0485
EE	Thermococcus sibiricus^{Ar}	CP001463.1	1,242,891	1,237,320	0.9955	0.0045
EE	Methanocaldo-coccus vulcanius^{Ar}	CP001787.1	1,746,040	1,708,968	0.9788	0.0212
EE	Sulfolobus islandicus^{Ar}	CP001731.1	2,722,004	2,692,455	0.9891	0.0109
	Bacillus subtilis^{Ba}	*E!* Bacte. 9	4,215,606	4,198,057	0.9958	0.0042
	Mycobacterium tuberculosis^{Ba}	*E!* Bacte. 9	4,411,532	4,348,606	0.9857	0.0143
	Escherichia coli^{Ba}	CP001396.1	4,578,159	4,551,258	0.9941	0.0059
LBG	Burkholderia xenovorans^{Ba}	NC007951-3	9,731,138	9,593,486	0.9859	0.0141
AP	Saccharomyces cerevisiae^{Fu}	*E!* Fungi 3	12,070,898	11,974,342	0.992	0.008
UE	Plasmodium falciparum^{Ap}	*E!* Proti. 9	23,263,332	21,070,640	0.9057	0.0943
UE	Phaeodactylum tricornutum^{He}	*E!* Proti. 9	25,805,651	25,667,448	0.9946	0.0054
UE	Thalassiosira pseudonana^{He}	*E!* Proti. 9	31,199,234	31,023,020	0.9944	0.0056
UE	Dictyostelium discoideum^{Am}	*E!* Proti. 9	33,919,934	30,877,496	0.9103	0.0897
	Ciona intestinalis^{Ur}	*E!* 62	87,649,861	84,674,396	0.9661	0.0339

Feat.	Species ^{clade}	ACN-EV	GS	GC	GCR	Dmax
	Caenorhabditis elegans^{In}	*E!* Meta. 9	100,272,217	97,720,472	0.9746	0.0254
	Tribolium castaneum^{In}	-1-	112,129,668	109,424,212	0.9759	0.0241
AP RG	Arabidopsis thaliana^{Pl}	*E!* Plants 9	118,960,082	116,563,556	0.9799	0.0201
	Drosophila melanogaster^{In}	*E!* Metaz. 9	120,290,887	118,973,632	0.989	0.011
GE	Daphnia pulex^{In}	*E!* Metaz. 9	158,632,523	150,111,316	0.9463	0.0537
AP	Arabidopsis lyrata^{Pl}	*E!* Plants 9	173,245,910	161,798,504	0.9339	0.0661
AP	Tetraodon nigroviridis^{Fi}	*E!* 62	208,708,313	207,067,712	0.9921	0.0079
	Apis mellifera^{In}	*E!* Metaz. 9	224,750,524	219,278,732	0.9757	0.0243
	Anopheles gambiae^{In}	*E!* Metaz. 9	225,028,531	221,180,624	0.9829	0.0171
AP	Brachypodium distachyon^{Pl}	*E!* Plants 9	270,058,956	257,893,524	0.955	0.045
AP	Oryza sativa^{Pl}	*E!* Plants 9	293,104,375	271,137,108	0.9251	0.0749
AP	Populus trichocarpa^{Pl}	*E!* Plants 9	370,421,283	352,063,876	0.9504	0.0496
AP	Physcomitrella patens^{Br}	*E!* Plants 9	453,927,385	399,508,556	0.8801	0.1199
AP	Sorghum bicolor^{Pl}	*E!* Plants 9	625,636,188	491,993,216	0.7864	0.2136
AP	Oryzias latipes^{Fi}	*E!* 62	582,126,393	562,662,192	0.9666	0.0334
	Gallus gallus^{Bi}	*E!* 62	984,855,151	971,359,304	0.9863	0.0137
	Taeniopygia guttata^{Bi}	*E!* 62	1,013,982,659	996,918,996	0.9832	0.0168
AP	Danio rerio^{Fi}	*E!* 62	1,354,636,069	1,191,452,752	0.8795	0.1205
AP RP	Zea mays^{Pl}	*E!* Plants 9	2,045,697,632	1,197,255,904	0.5853	0.4147
	Canis familiaris^{Ma}	*E!* 62	2,309,875,279	2,272,374,188	0.9838	0.0162
	Equus caballus^{Ma}	*E!* 62	2,335,454,424	2,307,202,104	0.9879	0.0121
	Bos taurus^{Ma}	*E!* 62	2,466,956,401	2,406,743,280	0.9756	0.0244
	Rattus norvegicus^{Ma}	*E!* 62	2,477,053,718	2,430,894,052	0.9814	0.0186
	Mus musculus^{Ma}	*E!* 62	2,558,509,481	2,521,038,616	0.9854	0.0146
	Pan troglodytes^{Ma}	*E!* 62	2,598,733,311	2,566,544,200	0.9876	0.0124
	Macaca mulatta^{Ma}	*E!* 62	2,646,263,164	2,621,196,144	0.9905	0.0095
	Pongo abelii^{Ma}	*E!* 62	2,722,968,487	2,697,592,876	0.9907	0.0093
	Homo sapiens^{Ma}	*E!* 62	2,858,658,095	2,841,049,052	0.9938	0.0062
LGS	Monodelphis domestica^{Ma}	*E!* 62	3,412,593,369	3,402,944,248	0.9972	0.0028

CR corresponds to random sequence of lengths ranging from 5 Kb to 2.5 Gb (a, b, c, d and e). In the case of biological sequences, non-polyploid genomes showed CR > 0.90. Conversely, polyploids showed CR below 0.95, with the lowest ratio for *Z. mays* (CR=0.58), and the next to the lowest ratio, its closest relative *S. bicolor* (CR=0.78). Overall non-random 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

Figure 3.1. ***(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 3.1 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. $2\times$ to $6\times$ correspond to random polyploids [ACGT] sequences where $1\times$ is "a". Changes in sequence length due to polyploidy produce no change in complexity ratio (see Table 3.2). Notice the low CR of human texts (see Table S3 for details).



Kind	Author - Writings	Lang.	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 de 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 3.2.: 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. Kinds: SA: Scientific abstract, SS: Short story; B: Book, CW: Complete Work

3.1-B and Figure 3.2. 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×. 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 3.2.

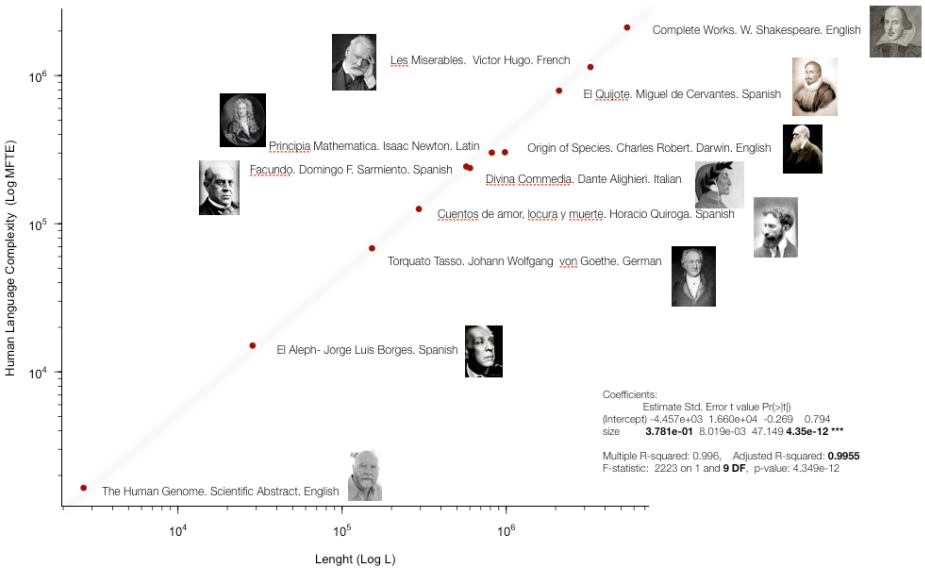


Figure 3.2.: Human language complexity.
Complexity in human writings shows a constant increase with text length. Regression analysis shows that in contrast to genomes, human language is highly repetitive. While genomes match an almost perfect regression of slope 1, human language complexity fits a linear regression model with slope alpha=0.378, (adjusted R = 0.995).

3.1.2. Genome complexity and ploidy level

Analysis of CR Figure 1.2-B reveals a clear segregation of species' genomes by their level of polyploidy, most recent polyploids like maize and sorghum exhibit lowest CR. However this trend seems to be quickly lost as ancient polyploids are hardly distinguishable from non-polyploids. The nicest example here can be found within *Arabidopsis* clade where the 2

close relatives *thaliana* and *lyrata* seem to have followed different routes after their whole genome duplication. While *Arabidopsis thaliana* suffered a drastic genome reduction after polyploidization (mainly due to hundreds of thousands of small deletions), its relative *Arabidopsis lyrata* remained complete [Hu *et al.* 2011]. And this is reflected in the differences in CR between those two species CR=0.9339 for *lyrata* and CR=0.9799 for *thaliana*.

In order to confirm the trend observed around level of polyploidy and CR value, 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 3.1. 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.

3.1.3. Chromosome complexity

Following the same methodology used for genomes, we computed CR of individual chromosomes (567 autosomes of 31 species). CV obtained were normalized by chromosome size resulting in a CR and can be seen in Figure 3.3-A. As previously statistics were very convincing, the slope of the relationship between chromosome size and CV around 0.924, and can be increased excluding polyploid species to 0.951 (alone, polyploid species exhibit a low slope = 0.696). And again, the size-interaction term was indisputable ($p < 2e\text{-}16$)

Notice that when considering non-polyploid species, the slope of the correlation CV-size is almost common for chromosomes and genomes (slope = 0.989 and 0.988, respectively).

The boxplot inside Figure 3.3-B summarizes the distribution of chromosomes' CR. 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 & Helentjaris 1989, Gaut 2001] and

attributed to a tetraploid event occurred in the origin of maize 11.4 My ago [Gaut & Doebley 1997, Wolfe 2001].

What brought out when comparing CR of genomes and of their corresponding chromosome, is that even if the global picture was conserved, some differences raised. Namely, here are some of those differences (genomes' CR – mean chromosomes' CR): 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).

As it appears, those differences are either positive or negative. The falls in genomes' CR, occurs generally in polyploid species and can be explained by a broader definition of "repetitive elements" as a consequence of the consideration of a wider window (see next subsection 3.1.4). In contrast, raises in genome's CR is a more sensitive problem further discussed in the subsection 3.1.6 on polyploid and return to maximum complexity.

3.1.4. Complexity in chromosome segments

In order to fully understand how CR ratios works, and how good is the idea to work with full genomes or chromosomes, we decided to apply our algorithm to windows of different sizes. Chromosomes were thus, split in overlapping windows of various sizes (from 1 Kb to 100 Mb) and CR in each of these windows was computed. Figure 3.4 shows boxplots of six selected chromosomes, at different scales, all having extreme CR. As a first trend we observed that median values of CR over all windows of *H. sapiens* Chr1 Figure 3.4-A, *A. thaliana* Chr1 Figure 3.4-C, *C. elegans* Chr1 Figure 3.4-D, and *D. melanogaster* Chr2L Figure 3.4-E were above 0.97. Whereas lower values were obtained in *Z. mays* Chr 1 Figure 3.4-F and in *H sapiens* Chr19 Figure 3.4-B. The fall in CR in these last chromosome is more dramatic for large windows sizes (1Mb). The reasons for this fall are different in the two cases: while maize Chr1 is tetraploid, human Chr19 is known to contain the highest number of Alu sequences over human chromosomes [Venter *et al.* 2001].

Among all chromosomes presented here (Figure 3.4), but also in the rest of chromosomes analyzed (available through this link: `http://bioinfo.cipf.es/das/`), the observation that larger the window size, the lower the median

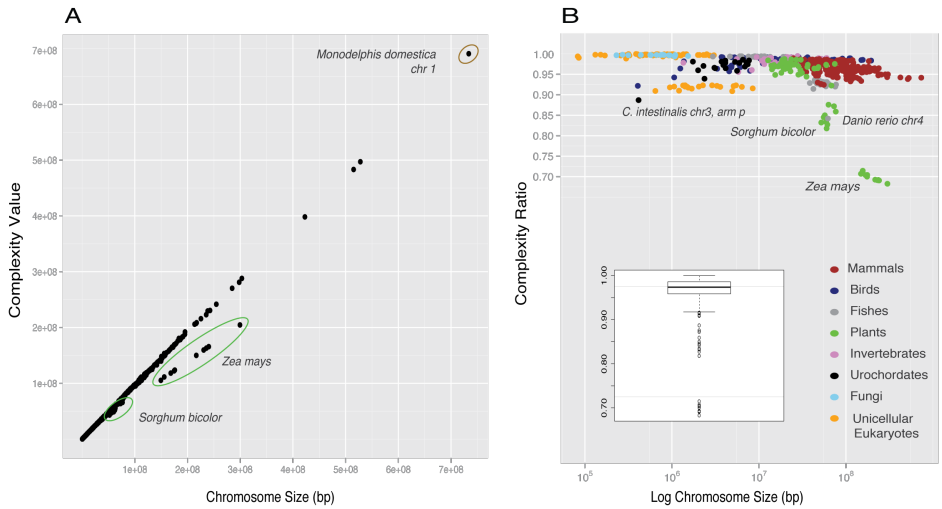


Figure 3.3.: Chromosome complexity ratio.
(A) Complexity ratio and chromosome size of 31 eukaryote species (567 chromosomes). Notice how far chromosomes of Z. mays, and in minor degree S. bicolor (both recent polyploid species) depart for the general trend. **(-B)** Most chromosomes (96.2%) have complexity ratios ranging 0.9 to 1.0, as observed for complete genomes Figure 3.1B. Boxplot inside shows the distribution of CR of all chromosomes.

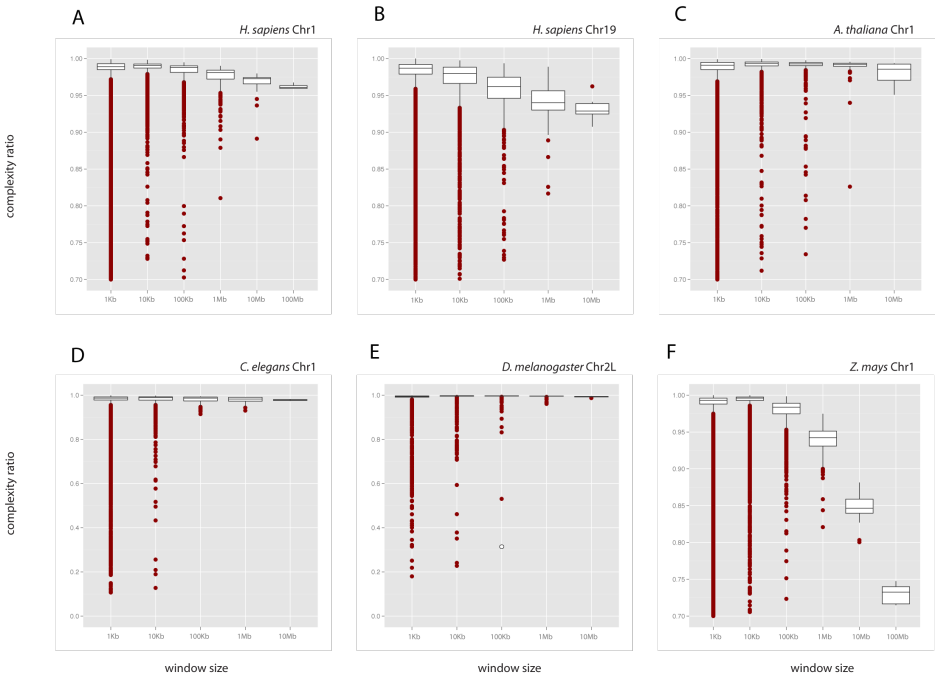


Figure 3.4.: Sliding window analysis in chromosomes.
Boxplots show results of sliding window analysis in six selected chromosomes (A-F). Most chromosomes have median CR higher than 0.975 independently of window size. White dot in the 100Kb window size chart of D. melanogaster Chr 2L (E) corresponds to the "deep spike" displayed in Figure 3.5. Scales were selected to enlarged differences in CR.

CR value was prevalent. This pattern can be explained by existence of repeats, which can only be detected when the window size is large enough.

As a last example of how window size affects the detection of regions with low entropy stands Figure 3.5. This figure represents the entropy-shape of *D. melanogaster* chromosome 2 for two sizes of windows. For small windows (1K) the rugged pattern drawn by the values of CR across the chromosome is hardly interpretable, we can only figure that each fall in CR correspond to regions with a high number of small repetitions of one or two nucleotides. By contrast, when windows size reaches 100K, the shape of CR is much smoother, revealing only one main peak of low CR. Interestingly corresponding to an histone cluster with more than 100 genes of the family locates. The picture on the right of the figure shows Ensembl annotation for the histone genes cluster.

3.1.5. Complexity in repetitive elements and genes – low and high?

Eukaryote genome structure is generally sketched out by the massive presence of non-functional repetitive elements (REs) 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). This could be achieved, for each family, by concatenating all units in their original order in chromosomes.

Genes, characterized by a specially high content of information, showed, as expected, the highest CR among all classes analyzed, independently of the species – this characteristic of coding regions, is important for entropy-based algorithms that predicts or confirm automatic detection of genes [Du *et al.* 2006, Gerstein *et al.* 2007], but it is important here to have echoes of our premiere methodology taking all genes together. Going deeper in the analysis, when genes were split in their two main components, exons showed even a higher CR.

For repetitive elements the expectation was that CR would fall due to the low complex nature of the elements. And indeed this result was found in SINE and satellites. However for LINE, LTR and DNA-T (Table 3.3) unexpectedly high values were observed. This result can be explained by

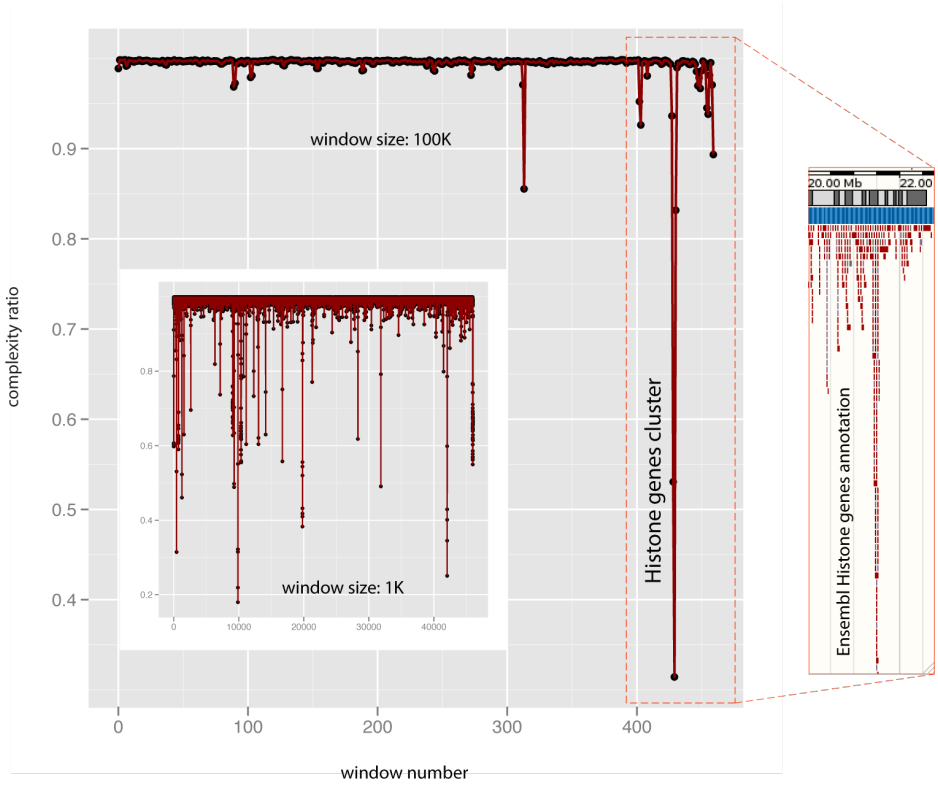


Figure 3.5.: Sliding window in a full chromosome.
Complexity ratio along *D. melanogaster* chromosome 2L is displayed at two window size scales. Ensembl annotation of the histone genes cluster is shown in the left box. See associated DAS server displaying CR along chromosomes for 15 eukaryote species.

the larger length of these elements and their high internal variability among families.

We also noticed some interesting clade specificity regarding the relative CR of RE families. For example, in mammals DNA-T and LTR elements exhibited higher CR than LINE elements, while this is not the case for fishes, some invertebrates and plants. Moreover, in plants, LINE has the highest CR after genes (see Table 3.3 and Figure 3.2 for comparison among all eukaryote species analyzed).

Table 3.3.: 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

TODO: POURQUOI CR DIMINU QUAND UN SHUFFLE????

In order to test the hypothesis that the order in which RE are placed in chromosome influences our measures, we compared the CR values of the RE in "natural" versus random order. Table 3.4 shows these values for eight selected chromosomes of different species. Curiously, CR of elements in natural order 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 & Wolfe 2004]. 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 3.4.: Size and complexity ratio of different genome classes in natural order (NAT), 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
	NAT	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
	NAT	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
	NAT	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
	NAT	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
	NAT	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
GENE	SIZE	18.242	16.312	10.77	9.918	140.258	21.909	37.623	16.759
	NAT	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
	NAT	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
	NAT	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

3.1.6. Polyploidy and return to maximum complexity

Evolution erodes ancient footprints of genome polyploidy and diploidization proceeds during time [Wolfe 2001]. As shown in previous sections, CR of recent polyploids is much lower than in non-polyploid, or in ancient polyploid species. "Erosion" of polyploids can be achieved by multiple mechanisms [Wolfe 2001]. The most simple, perhaps, being the gradual disintegration of the duplicated genetic material by random mutation.

Other mechanisms, more dramatic, also participate in the loss of polyploid footprints, such as massive deletion and transpositions of genetic material as 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.

In order to better understand the decay of genetic redundancy after polyploidization, the action of two mechanisms –mutation and transposition– were simulated over repeated random sequences of different lengths. The first process (mutation) was also applied to 2 chromosomes of our most recent polyploids *Z. mays* Chr1 and *S. bicolor* Chr1.

In all cases, sequences under random mutation (Figure 3.6-A) or transposition (Figure 3.6-B) reached maximum CR=1 after a number of generations large enough. A general observation is that the lower is the CR the more affected it is by the first changes (either mutation or transpositions). Exactly 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. For real polyploids (sorghum and maize), the dynamics of CR increase was identical (Figure 3.6-A). Figure 3.6-B 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 property was reported above for maize and sorghum (see discussion on subsection 3.1.3).

As CR get closer to 1 through time, DNA structure of polyploids become indistinguishable from diploid genomes.

3.1.7. High complexity and random-like structure of DNA

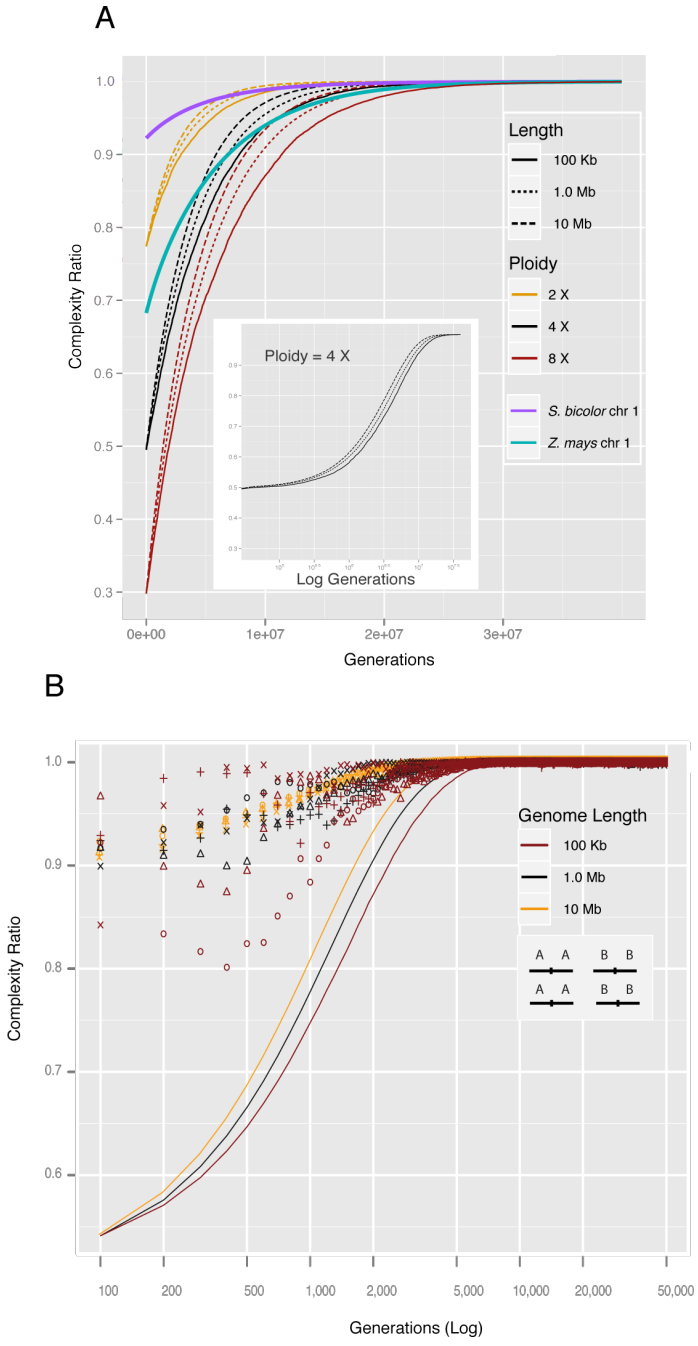
Excluding recent polyploids, high CR (almost maximum) was observed in complete genomes of organisms sampled in all diversity of life, in their chromosomes and along large enough chromosomes segments. We conjecture this is a universal feature of all genomic sequences.

3.1.8. 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 (see subsection 2.1.2 to get a complete picture on how the CR works). Polyploid genomes of maize and sorghum have CR=0.585, and CR=0.786, respectively. Values that were respectively close to diploid and triploid simulated genomes (Figure 3.1-B). 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 [de Bruijn 1946, Becher & Heiber 2011] with low CR. See Table 2.1 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 a sequence of length n if it contains full diversity of length k, for $k \leq log_4 n$, and each these short sequences occurs about $n{\cdot}4^{-k}$ times. As CR decreases, diversity and balanced abundance deteriorates. In particular, maximum

Figure 3.6. ***(following page)*****: Return to maximum complexity after polyploidization.**
A. Random genomes of different lengths and ploidy levels experienced increase of complexity ratio by constant accumulation of random mutation (1e-08 mutations per site, per generation). Chromosomes of maize and sorghum are included in the simulation. The picture inside displays the log transformation of the x axis to compare the 4X polyploid sequences under mutation with translocation process. B. Two random sequences (A & B) were arranged in chromosomes to simulate tetraploid genomes of different length. Translocations of 1Kb length occurred at a constant rate of genome size over 1,000, during 50,000 generations, and plotted each 100 generations. Dots (circles, triangles, "+" and "×") represent individual chromosomes while lines display genome complexity after chromosome concatenation. Before rearrangements, tetraploid genomes and chromosomes yielded identical CR independently of genome or chromosome sizes (not shown to override redundancy). CR of genomes and chromosomes were 0.49 and 0.77, respectively. Notice these values are close to the observed in maize (Fig1B, Fig2B).


A
Complexity Ratio
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0e+00
1e+07
2e+07
3e+07
Generations
Length
100 Kb
1.0 Mb
10 Mb
Ploidy
2 X
4 X
8 X
S. bicolor chr 1
Z. mays chr 1
Ploidy = 4 X
Log Generations
B
Complexity Ratio
1.0
0.9
0.8
0.7
0.6
100
200
500
1,000
2,000
5,000
10,000
20,000
50,000
Generations (Log)
Genome Length
100 Kb
1.0 Mb
10 Mb
A A
B B
A A
B B

CR appears for some de Bruijn sequences [de Bruijn 1946, Becher & Heiber 2011]. Also maximum CR=1 occurs in randomly generated sequences with uniform distribution of A, C, G and 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 [Nies 2009]. We observed that the largest correlations in the human genome are chromosome specific, the actual largest exact repeat is 67,632 bp long and it corresponds to a single duplicated region inside Chr 1, while the largest inter-chromosomal perfect correlation is 21,865 bp also appearing only once in Chr 1 and Chr 5.

3.1.9. 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 [Chaitin 1975, Nies 2009], 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, statistical randomness and compressibility are antonyms. Thus, the complexity ratio (CR) can also be understood as the inverse of the degree of compressibility of a sequence. 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.

3.2. Discussion

3.2.1. Universal structure of DNA

Up today no conclusive work on statistical property of DNA was conducted in full genomes. After having a broad look to our result, the most remarkable result is certainly that whatever genome taken at whatever level of magnification, genome chromosome or windows, DNA seems to be

strongly attracted to maximum complexity. At genomic and chromosomal level, recent polyploid species were the only outliers observed. Aside from these exceptions, CR was stacked to 1 in the whole range of diversity of life browsed, from viruses to mammals. Even the most expected "simple" genomes like higher eukaryotes, with around a half of their genomes composed by repetitive sequences presented a CR>0.95. Even if we found that SINE and satellites were indeed presenting lower CR, the other families of REs showed too high degrees of variability to be able to lower their CR.

3.2.2. Mechanisms of genome amplification and divergence

From our observations, genome size enlargement by duplications result in lowering CR. However maximum CR is rapidly recover, ancient polyploids in our dataset were all presenting high values of CR. Also the simulations conducted showed that after 30 millions generations is enough to recover a CR>0.95, and this taking only into account single mutations at a mean rate. In this context we envision growth in genome complexity as successive falls and growth of CR, or a raise in CV following a stepped curve. It is though that during this process, intermediate states with genomes of constant sizes suffering mutations and rearrangements could give birth to new functional sequences, thus providing raw material of species divergence, and biological complexity growth [Lynch 2000].

3.2.3. Genome size reduction

Theoretically, genome size reduction events are not expected to lower the CR, because any region of the genome large enough showed an almost random structure. Natural selection will ultimately determine the success of genome segments losses, but as we observed, intracellular bacterial parasites seemed to moved down (along the straight line of Figure 3.1-A) from larger genomes sizes to shorter genomes by fitting almost maximum CR during this process. Moreover, ancient polyploid species as *Arabidopsis thaliana* [Hu *et al.* 2011] seemed to reach longer genome sizes after polyploidy, and posterior genome size reduction during diploidization process, yet showing almost maximum CR.

Thus in the case of gradual sharp or genome reduction, the pattern drawn by the evolution of CR values is expected to be smoother than in

the case of genome amplification.

3.2.4. Limits of CR space

We speculate that the space of complexity ratios filled in human writings Figure 3.1-B, is a neglected region for life. A non-random combinatorial structure is impossible for DNA of virus, phages and prokaryotes having small genome sizes. By the effects of natural selection, simple forms of life –with genomes dimensions varying from the size of a few text paragraphs to the complete works of William Shakespeare– are probably forced to have a random like-structure.

3.2.5. Hypotheses

All together these observations lead us to hypothesize that:

- A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes along all diversity of life.
- Polyploid genomes will reach almost maximal complexity in the structure of their DNA with enough time.
- Since the DNA combinatorial structure is quasi-random, genome complexity will only increase by DNA amplification, and posterior divergence of duplicates during evolution

These hypotheses can be falsified in some specific cases:

- Genomes of recent polyploid species evidencing a quasi-random DNA structure (high CR).
- Non-polyploid genomes evidencing a non-random DNA structure (low CR), due for example to a very strong GC content bias.

In this manuscript we described the combinatorial DNA structure of genomes. Ultimately we hypothesized a universal random-like structure along all diversity of life. It is very hard to think that such a structure is adaptive in its origin. However, far from being biologically irrelevant, useful properties may freely emerge from such random-like combinatorial structure in genomes. After all, exons, the main functional pieces of genes,

are the elements with the most random-like DNA structure.

A simple law controlling genome statistical design for all kind of organisms makes nature modest and beautiful. Although it is hard to argue that by rolling a dice millions times a functional genome will suddenly emerge, perhaps the fixation of some trial by natural selection let us imagine that it is feasible.

4. Life inside genomes, dynamics and predictions

TO INSERT:

- "At the beginning of the 20th century, communities were viewed as a superorganism that develops in a particular and fixed way to form a well-established climax community." [Alonso *et al.* 2006]
- Explain why using functional elements: develop a model that can explain dynamics and evolution of TEs would have to take into account, not only the interaction with other TEs but also with genes, as forbidden regions or as regulator [Le Rouzic *et al.* 2007]
- Discussion: non-neutral models like symetric [Alonso *et al.* 2008] model developped by Jabot and Chave [Jabot & Chave 2011], including one new parameter, δ that ponderates the death probability according to richness of species.

4.1. Results

4.1.1. Genetic elements, dispersion and abundance

Ecologists frequently use RSA curves to compare the richness, the degree of dominance, and the number of rare species in communities. The raw data used in these plots is the total number of individuals per species sampled in the ecosystem. The most interesting property of RSA curves is that species are unlabeled in the ranking order; hence ecosystems can be compared whatever the species they contain.

Taking advantage of the current automatic methods of genome annotation and string recognition, all GEs belonging to different functional and non-functional classes were counted to build RSA curves in genomes and chromosomes. These numbers represent censuses of GEs analogous to those sampled in ecosystems.

Figure 4.1 display RSA curves for a selected group of genomes and their largest chromosome respectively. Curves differ in many ways although two patterns are evident: 1- RSA curves of genomes and chromosomes are very similar for species, 2- all RSA curves display the universal S-shape observed in ecological environments [McGill *et al.* 2007, Hubbell 2001]. Both observations suggest a common mechanism of distribution of GEs in genomes and chromosomes.

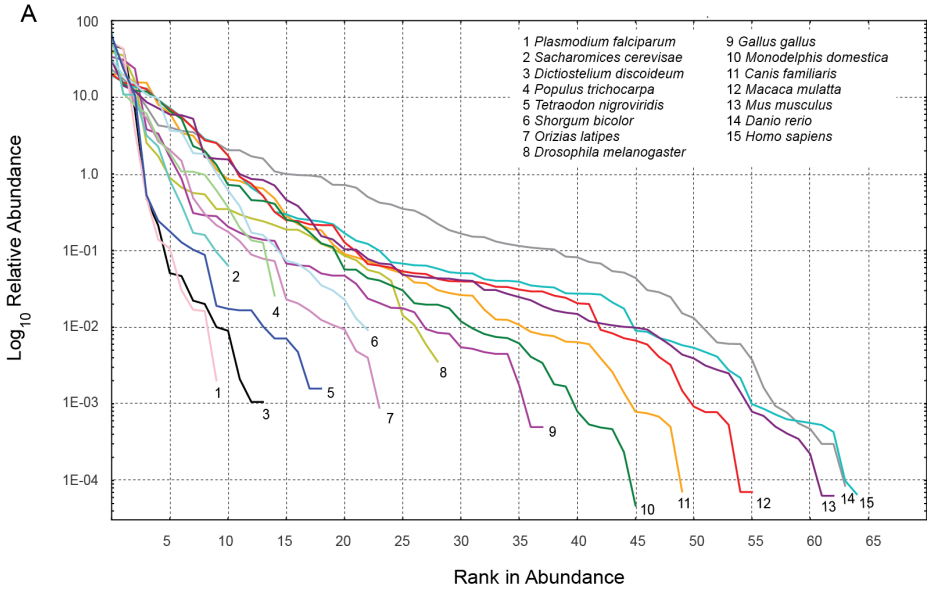
4.1.2. Counterbalanced species abundances in genomes

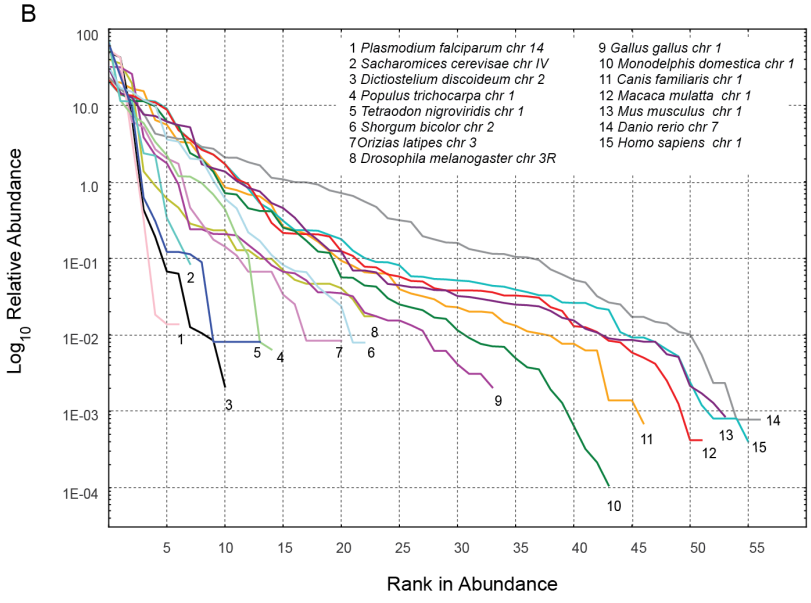
To what extent chromosome's RSA curves represent the random distribution of the complete set of elements of the genome? To answer, we simulate the random distribution of the full set of GEs reported in genomes in their corresponding chromosomes. After one thousand simulations the mean expected abundance and its standard deviation were computed for all genetic classes in chromosomes. These values were used to plot random expected RSA curves for chromosomes.

Statistical tests (t-test, $FDR < 0.05$) established that less than 1% of all genetic classes in all chromosomes tested showed abundances according to their random expected distribution. This homogeneous process therefore, does not account for the observed RSA curves in chromosomes. However, another kind of arbitrary process is suggested for chromosomes if simulated and observed RSA curves are superimposed Figure 4.2.

This notable adjustment is due to the fact that genetic classes are unlabeled in the ranking order, and changes in the order of abundances are counterbalanced in chromosomes. For instance, the classes of functional tRNA and satellite elements are at position 33 and 42 of the ranking of

Figure 4.1. ***(following page)*****: RSA curves.**
Relative species abundances for some selected genomes (A) and their corresponding largest chromosomes (B)


A
Log10 Relative Abundance
100
10.0
1.0
1E-01
1E-02
1E-03
1E-04
5
10
15
20
25
30
35
40
45
50
55
60
65
Rank in Abundance
1 Plasmodium falciparum
2 Sacharomices cerevisae
3 Dictiostelium discoideum
4 Populus trichocarpa
5 Tetraodon nigroviridis
6 Shorgum bicolor
7 Orizias latipes
8 Drosophila melanogaster
9 Gallus gallus
10 Monodelphis domestica
11 Canis familiaris
12 Macaca mulatta
13 Mus musculus
14 Danio rerio
15 Homo sapiens


B
Log10 Relative Abundance
100
10.0
1.0
1E-01
1E-02
1E-03
1E-04
5
10
15
20
25
30
35
40
45
50
55
Rank in Abundance
1 Plasmodium falciparum chr 14
2 Sacharomices cerevisae chr IV
3 Dictiostelium discoideum chr 2
4 Populus trichocarpa chr 1
5 Tetraodon nigroviridis chr 1
6 Shorgum bicolor chr 2
7Orizias latipes chr 3
8 Drosophila melanogaster chr 3R
9 Gallus gallus chr 1
10 Monodelphis domestica chr 1
11 Canis familiaris chr 1
12 Macaca mulatta chr 1
13 Mus musculus chr 1
14 Danio rerio chr 7
15 Homo sapiens chr 1

abundances respectively in human chromosome 1 Figure 4.2-A. However, according to their random distribution the expected values in the ranking are 43 and 23 respectively. That is, tRNA and satellite elements show higher and lower abundances than the expected by random distribution. Over and under abundances of different genetic classes counterbalance each other in the same chromosome leading to an almost perfect fit between observed and expected RSA curves. We only found significant differences between observed and expected RSA curves for less than 100 of most 540 chromosomes tested (KS test, $p < 0.05$).

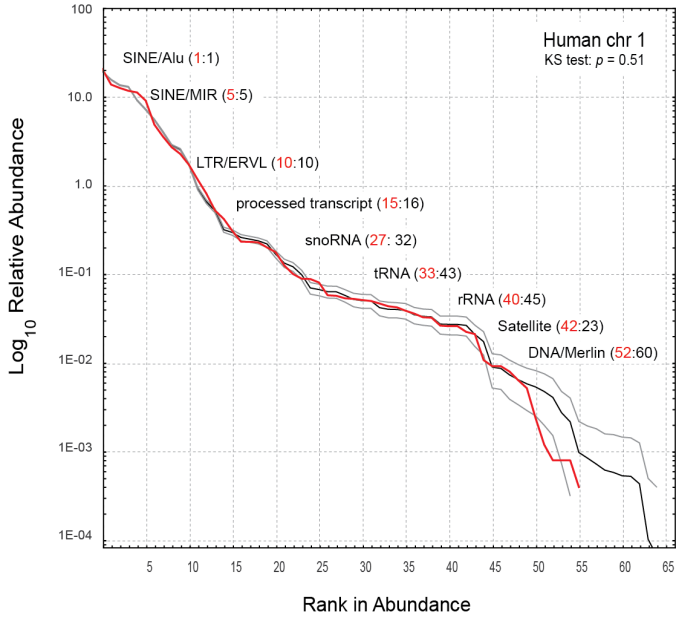
What is the mechanism contributing to this counterbalanced dynamics of GEs in eukaryote's chromosomes? Next, we test the neutral theory of biodiversity as the main explanation accounting for such events in genomes.

4.1.3. Neutrality of SAD

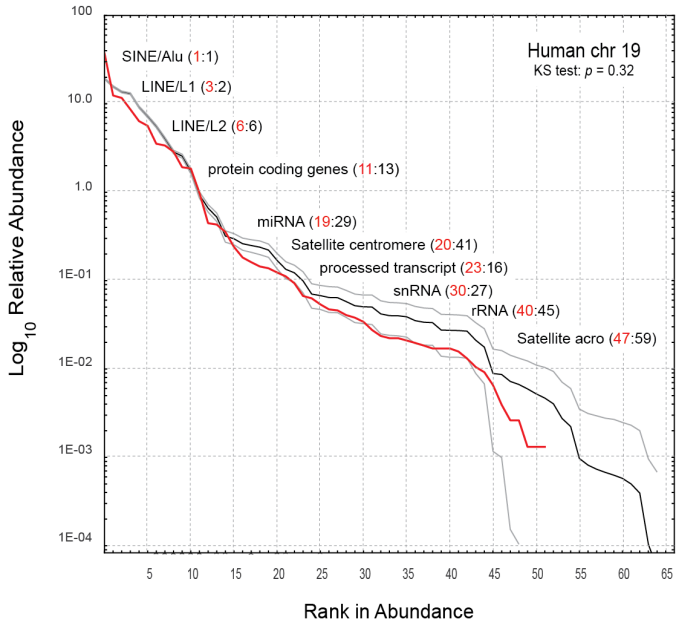
Similar to the kinetic theory of ideal gases in physics the neutral theory of biodiversity is a stochastic theory assuming equivalence among interacting individuals. The theory assumes that diversity in a local community of individuals is maintained by migration from the metacommunity at a constant rate (m). Births and deaths in the local community occur at constant rates during generation regardless the species. The metacommunity dynamics is controlled by speciation at a single constant rate (ν) [Rosindell *et al.* 2011, Alonso *et al.* 2006].

Figure 4.2. *(following page)*: Relative species abundance curves for human chr 1 (A) and chr 19 (B).
Relative species abundance curves for human chr 1 (A) and chr 19 (B). Red and black lines display observed and simulated RSA curves for all functional and non-functional genetic classes in chromosomes respectively. Grey lines show two standard deviations around the mean of the simulated data. The absence of statistical differences between RSA curves is mainly due to the frequent counterbalanced changes observed in the ranking of abundances of genetic classes. Numbers in parenthesis depict the observed (red) and the expected value (black) in the ranking of abundances for few genetic classes in both chromosomes. Differences between numbers point out over and under abundances in chromosomes. Note the higher than the expected number of SINE/Alu elements in human chromosome 19 (class 1 in the ranking).

A


Human chr 1
KS test: p = 0.51
SINE/Alu (1:1)
SINE/MIR (5:5)
LTR/ERVL (10:10)
processed transcript (15:16)
snoRNA (27: 32)
tRNA (33:43)
rRNA (40:45)
Satellite (42:23)
DNA/Merlin (52:60)
Log10 Relative Abundance
Rank in Abundance

B


Human chr 19
KS test: p = 0.32
SINE/Alu (1:1)
LINE/L1 (3:2)
LINE/L2 (6:6)
protein coding genes (11:13)
miRNA (19:29)
Satellite centromere (20:41)
processed transcript (23:16)
snRNA (30:27)
rRNA (40:45)
Satellite acro (47:59)
Log10 Relative Abundance
Rank in Abundance

For genomes, we realized that each chromosome is the physical arena where GEs die and are replaced by other elements of the same or different species. These GEs could come from the same chromosome, or from any other chromosome of the genome. We assume that each chromosome represents a local community of J elements and S different genetic classes (species) while the rest of chromosomes correspond to the metacommunity of size J_M. Thus, given the total number of functional and non-functional elements in each chromosome we optimized by maximum likelihood (ML) the neutral theory's parameters m and θ ($= 2J_M\nu$) using Ewens and Etienne's sampling formula (Equation 2.6) (see example Figure 2.1).

Deviations of neutrality were detected in 33 out of 578 (5.7%) chromosomes. However, deviations vanished at all after multiple testing correction ($FDR < 0.05$, Table). We conclude that Hubbell's neutral model fits abundance and diversity of GEs in all the chromosomes of the 31 eukaryotes genomes analyzed.

4.1.4. Diversity and chromosome length

4.2. Discussion

Abundance and diversity of selfish DNA, or selfish GEs [Doolittle & Sapienza 1980, Orgel & Crick 1980] results from millions of years of close interaction with their host, the genome. Population dynamics models dealing with transposable elements (TE) has been formalized and reviewed [Charlesworth & Charlesworth 2009, Charlesworth *et al.* 1994, Le Rouzic & Deceliere 2005]. Transposition and excision rates, as well as host fitness impact are some of their most significant parameters. While the predicted deleterious effects of transposition have been confirmed repeatedly in Drosophila (citar...), almost a neutral accumulation of TEs is expected in mammals due to the larger genomes and smaller population size [Lynch & Conery 2003]. An important concern with these models is the explicit absence of relationships of TEs with other genetic components of the genome.

The general transposition-selection based model predicts an adaptive equilibrium of TE abundance. However, abundances of the same TE differ between population and species (citar). Additionally, models c 9-11. Importantly variation between individuals was observed... [Brookfield 2005]

However, a former question is mandatory: Did TE's diversity and abundance features shaped by random processes in genomes [Lynch & Conery 2003, Venner *et al.* 2009].

Nature seems to play forest and genomes with the same dice. Functional elements, as expected deviates expectations If functional elements are excluded from chromosomes, neutrality was rejected in a single chromosome (D. rerio chr2, p= 0.04, q= 0.36).

Part II.

Detection of selective pressures in genomes

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

5.1. 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 2.4. Maximum likelihood (ML) estimates of evolutionary variables were performed using a free-ratio branch model [Yang 2007]. As such, four lists containing 12,543 and 9,240 orthologous genes in mammals in Drosophila species were obtained for the analyzes, 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 2.5 on page 40 in section 2.3). 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 5.1) 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 5.1.: Numbers and percentages of functional modules with significant results after GSSA. For Significantly High (SH) and Significantly Low (SL) results.

Table 5.1 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 [Lynch 2007] 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 5.1 (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 5.1-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

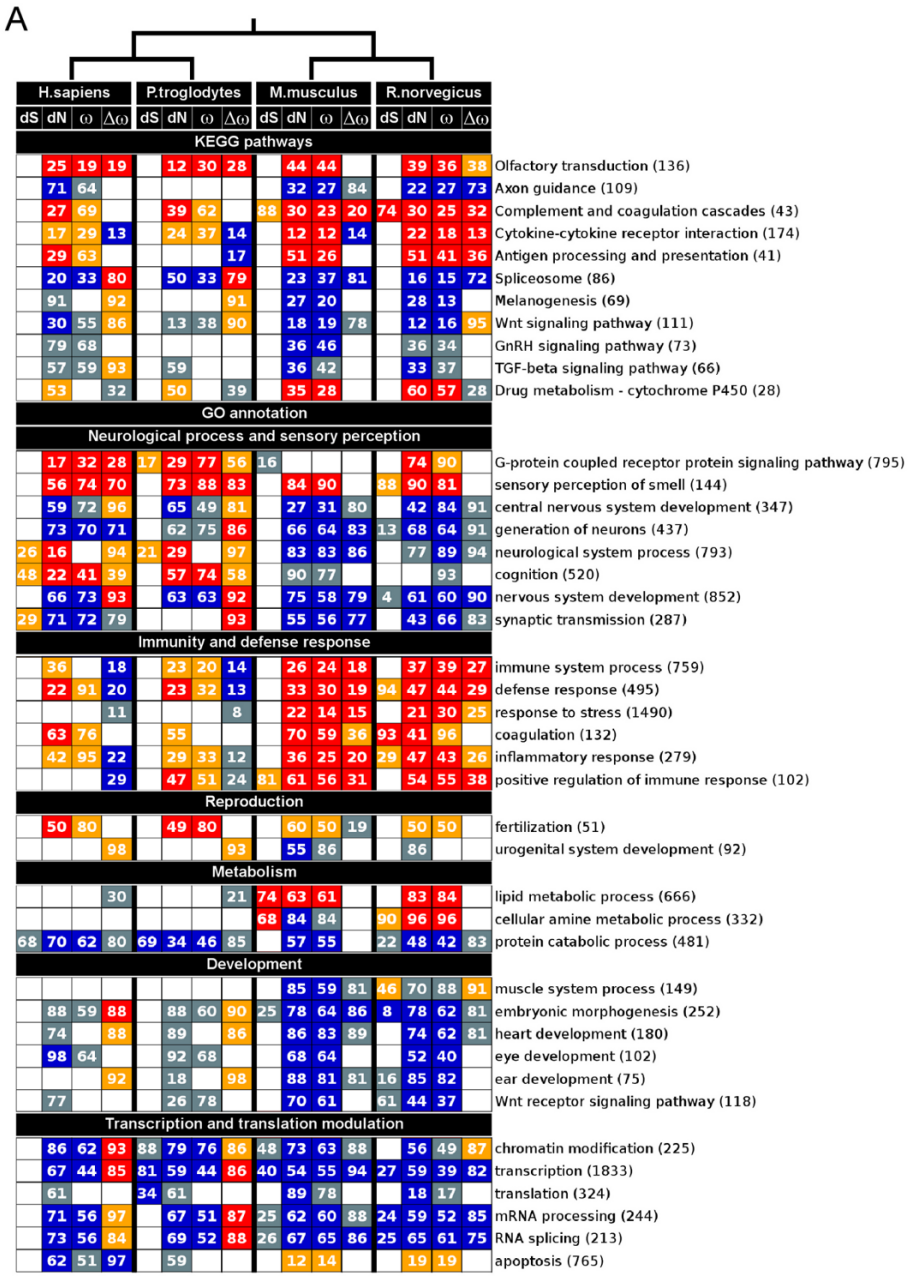
and rat (GO0007186: *G-protein coupled receptor protein signaling pathway*, and Ha04310: *Wnt signaling pathway*, among others).

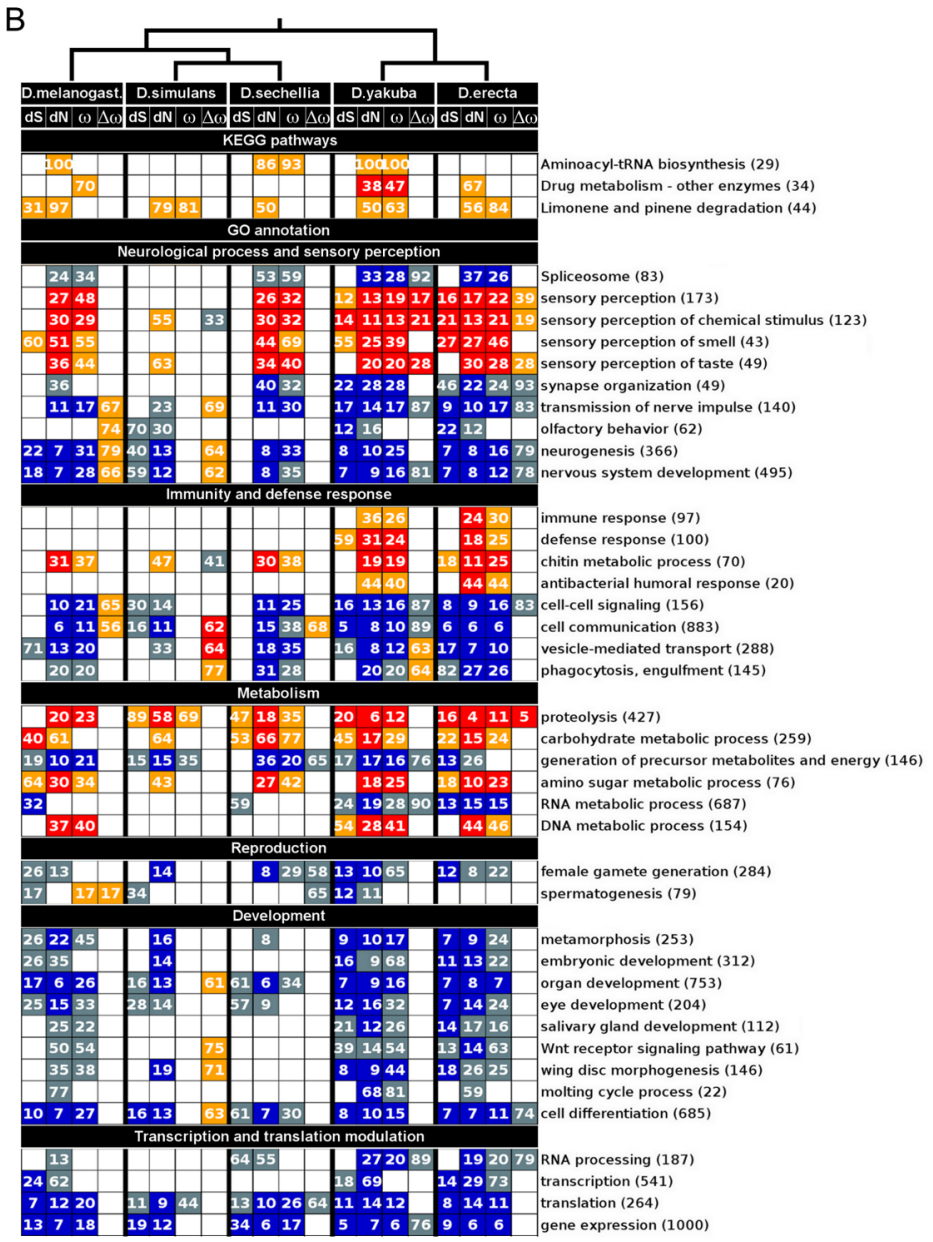
In addition, most of the GO terms significantly associated to high dN and ω in Drosophilas were unevenly distributed within the two clusters of the phylogeny (Figure 5.1-B). GO terms such as sensory perception, defense response, immune response and metabolic process, among others, presented a remarkable 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 5.1-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 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 & Barbadilla 2009].

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

Figure 5.1. ***(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)

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 5.2. 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.

5.2. Positively selected genes and the evolution of functional modules

GSSA tests for difference in rates over functional related groups of genes. To what extent genes under positive selection contribute to the significance of functional modules in mammals and *Drosophila* species after GSSA? To answer this question, branch-site (the most sensitive) test of positive selection was conducted on terminal branches of phylogenies Figure 2.4. We found 715 PSGs in mammals and 626 in *Drosophila*. Figure 5.2-A shows the distribution of the mean evolutionary rates (dN and dS) of functional modules providing significant and not significant results after GSSA of the ω ratio. When considering the total number of the functional modules

Table 5.2. ***(following page)*: Functional enrichment results using gene-by-gene and gene-set approaches.**
The table depicts some selected biological functions enriched by PSGs as cited in references 1 to 7, and the corresponding significant result observed after GSSA of ω values. References 1 to 7 correspond to cites 6, 7, CSAC, 4, 5, 9 and 8 in the manuscript, respectively. Abbreviations: SHv: statistically significant high v values; SLv: statistically significant low v values; H: *H. sapiens*; C: *P. troglodytes*; Pr: primates; M: *M. musculus*; R: *R. norvegicus*; Ro: rodents; mel: *D. melanogaster*; sim: *D. simulans*; sec: *D sechelia*; yak: *D. yakuba*; ere: *D. erecta*; Ds: *Drosophila* species.
*: p=0.05;
** p=0.001. CSAC: Chimpanzee Sequencing and Analysis Consortium, Nature. 2005 vol. **437** (7055) pp. 69-87.

Biological process	Enrichment in PSGs							GSSA: Functional category with significantly
	1	2	3	4	5	6	7	HIGH rates of omega
Olfaction/Sensory perception of smell	H	Pr**				Pr**		H**, C**, M**, R**, mel*, sec*, ere**, yak**
Chemosensory perception	H	Pr**						H**, C**, M**, R**, mel**, sec**, ere**, yak**
G-protein-mediated signaling	H				H	Pr**		H**, C**, R*
DNA /nucleic acid metabolism				C			Dr	C*, M**, R**, mel**, yak**, ere*
Amino acid metabolism	H, C						Dr	M**, R**
Proteolysis							Dr	M**, R**, mel**, sim*, sec*, yak**, ere**
Fatty acid/Lipid metabolism					H		Dr	M**, R**
Carbohydrate metabolism							Dr	sec*, yak*, ere*
Adult reproduction and gametogenesis							Dr	sec*
Spermatogenesis and motility		Pr*	Pr					H*, M*, mel*
Immune response		Pr**		H, C		Ro**		C*, M**, R**, yak*, ere*
Inflammatory response						Ro**		H*, C*, M**, R**
Defense response						Ro**		H*, C*, M**, R**, yak**, ere*
Response to wounding						Ro**		H*, M**, R**
Hummoral imm. resp. mediated by circulating Ig						Ro**		M**, R**
T-cell-mediated immunity		Pr**						M*
Natural killer-cell-mediated immunity		Pr*						R*
B-cell- and antibody-mediated immunity		Pr*						M**, R**
Response to pest, pathogen, or parasite				H				C*, M**, R**, yak*, ere*
Stress response					C	Ro**		M**, R**
Response to external stimulus						Ro**		M**, R*
Sensory Perception	H	Pr**		H		Pr**		H**, C**, M*, mel**, sec**, yak**, ere**
Cell surface receptor-mediated signal trasnduction	H					Pr**		C*
Cell adhesion	H							R*
Amino acid transport			Pr					M*
Protein amino acid glycosylation			Pr					M*
Amino acid transport	C							M*
								LOW rates of omega
Sensory Perception	H	Pr**		H		Pr**		R*
Cell surface receptor-mediated signal trasnduction	H					Pr**		mel*, yak*, ere*
Cell adhesion	H							H**, C**, mel**, ere*
Amino acid transport			Pr					R*
Protein amino acid glycosylation			Pr					H*
Amino acid transport	C							C*
Hearing / Perception of sound	H		Pr					M*, R*
Neurological process						Pr**		M**, R**, yak*, ere*
Synaptic transmission			Pr					H**, M**, R**, mel**, sec**, ere**, yak**
Signal transduction/intracellular signaling cascade	H, C		Pr				Dr	H**, C**, M**, R**, dmel**, dsec*, dyak**, dere**
Ion transport	H				H		Dr	H*, M**, R**, mel*, sec*, ere*
Potassium ion transport			Pr					H*, C*, M**, R**
Inorganic anion transport			Pr					M*, R*
Intracellular protein traffic	H							H**, C**, M**, R**, mel*, sec**, yak**, ere*
Transport							Dr	mel**, sec**, ere**, yak**
Protein transport				H			Dr	H*, C**, M**, R**, mel**, sim*, sec**, ere**, yak**
Metabolism of cyclic nucleotides	H							M*, R*
Protein metabolism & modification				H, C	C		Dr	H**, C**, M**, R**, ere*, yak*
Phosphate metabolism/phosphorylation				H, C			Dr	H*, C*, M**, R**, mel*, sec**, yak**, ere*
Purine metabolism	C							M*, R*, sec**
Carbohydrate biosynthesis			Pr					M**, R*
Cation transport	H							H*, M**, R**
Nervous system development							Dr	H*, M**, R**, mel**, sec*, yak**, ere**
Skeletal development	C							M**, R**
Organ development							Dr	H*, M**, R**, mel**, sec*, yak**, ere**
Post-embryonic development							Dr	M*, mel*, yak**, ere*
Embryonic development							Dr	H**, C*, M**, R**, yak*, ere*
Ectoderm development					H			C*, M*, R*, mel*, yak*, ere*
Cell proliferation and differentiation	C						Dr	H**, C*, M**, R**, mel**, sec*, yak**, ere**
Cell cycle							Dr	H*, M*, R*, mel**, sec**, yak**, ere**
Cell structurre/morphogenesis	C						Dr	H**, C*, M**, R**, mel**, sec*, yak**, ere**
Cell structure and motility	C							H*, M**, R**, sec*
Inhibition of apoptosis		Pr*						H*, yak*
Cell-cell signalling							Dr	H**, C*, M**, R**, mel**, sec**, ere**, yak**
Regulation of nucleobase				H, C				H**, C**, M**, R**, ere*
Translation							Dr	M*, R*, mel**, sim*, sec**, yak**, ere**
Transcription				H, C	C		Dr	H**, C**, M**, R**, ere*
Protein catabolism				H, C	C			H**, C**, M**, R**
Interferon-mediated immunity		Pr*						

with PSGs, 55%, 53%, and 42% of these original functional categories observed with SH, SL and NS results after GSSA (ω values) still remained Figure 5.2-B. This suggests that: 1- evolution of many of the functional modules changing at SHω ratios in the genome is not driven by a considerable accumulation of PSGs. Functional modules such as *complement and coagulation cascades* in human, *gonad development* in chimpanzee, *regulation of innate immune response* in mouse, *primary immunodeficiency* in rat, and *spermatid differentiation* in *D. melanogaster* are examples of functional modules evolving at significantly elevated ω ratio without any PSGs; 2- molecular adaptation takes place in functional modules under strong selective constraints (see last part of Table 5.2). For instance, *apoptosis* in human, *generation of neurons* in chimpanzee, *tissue development* in mouse, *Wnt signaling pathway* in rat, *eye development* in *D. melanogaster*, *wing disc development* in *D. yakuba*, and *generation of neurons* in *D. erecta* are some of the functional modules evolving at SLω ratios in the corresponding genomes that contain PSGs; and finally, 3- an important number of functional modules without significant differences in ω ratios (grey dots in Figure 5.2) still contain genes under positive selection. For instance, *homologous recombination* in humans, brain *development* in chimpanzee, *female or male sex differentiation* in mouse, *regulation of mitotic cell cycle* in rat, *chromatin modification* in *D. sechellia*, and *oogenesis* in *D. melanogaster*.

These results are in agreement with previous observations in *Drosophila* were it was emphasized that not every mutation under positive selection responds to a change in selection [Mustonen & Lässig 2009]. Beneficial changes could occur at evolutionary equilibrium, repairing previous deleterious changes and restoring existing functions [Mustonen & Lässig 2009].

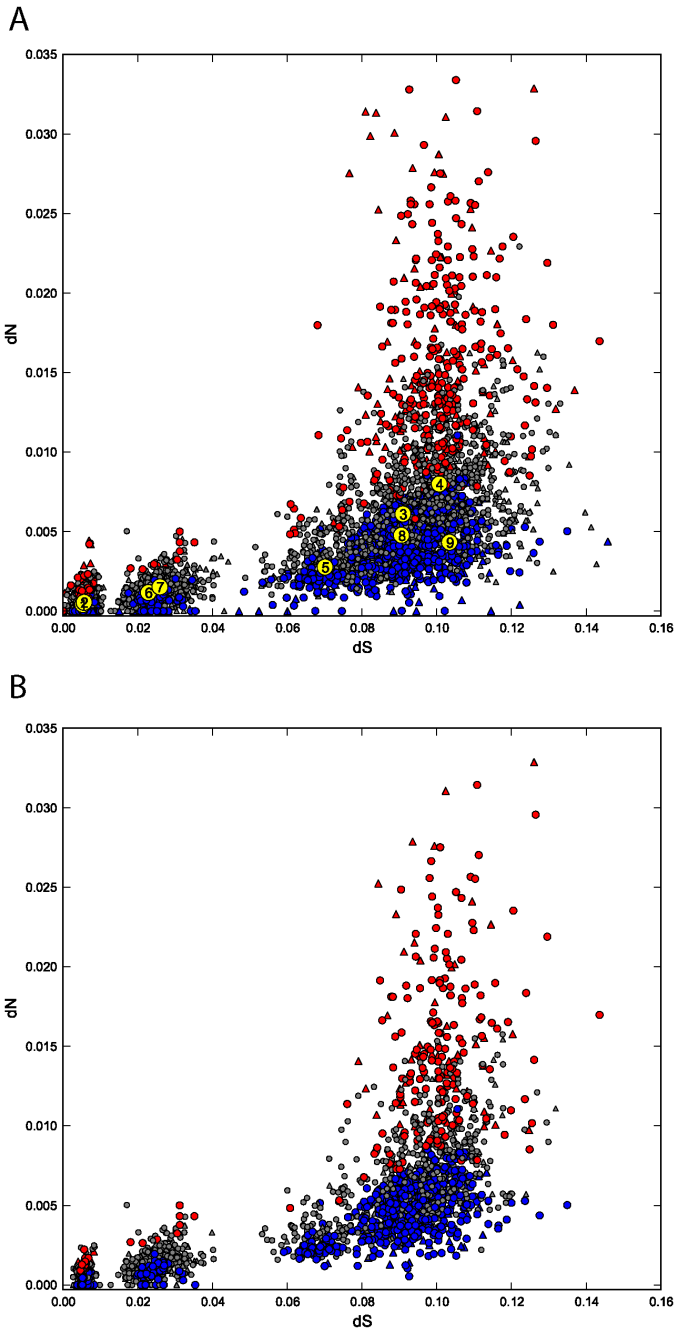
Finally, we ask if PSGs preferentially concentrate in functional modules evolving at faster rates in different genomes. For doing that we computed the mean number of PSGs in functional modules with SHω and SLω results (red and blue dots in Figure 5.2-B. As expected, functional modules evolving at high ω ratio contain higher numbers of PSGs in rodents ($p<0.001$), mammals ($p<0.001$), and *Drosophila* ($p<0.001$) species. For primates however, it was not significant ($p = 0.47$), indicating that PSGs are distributed almost evenly in functional modules evolving at significantly high and low values of ω in human and chimpanzee.

To contrast these results, PSGs from previous works in mammal and

Drosophila species were collected [Clark *et al.* 2007], [Kosiol *et al.* 2008]. The pattern of distribution of PSGs in functional modules was in agreement with the mentioned results: significantly skewed ($p<0.001$) towards higher numbers of PSGs in mammals, rodents, and *Drosophila* species, but showing no differences in primates ($p = 0.73$).

In summary, PSGs are frequently observed in functional modules evolving under a wide range of evolutionary scenarios; however, they concentrate more frequently in functional groups of genes changing at elevated rates in rodents and *Drosophila* species. Alternatively, PSGs were evenly distributed in functional modules changing at the extreme rates of evolution in primates. This observation suggests that a more complex scheme than the cumulative differences of PSGs must rely on the observed adaptive differences in human and chimpanzee genomes. The search for integrative factors taking into account the action of multiple genes other than only those which have been targeted by positive selection [He *et al.* 2010], could provide a more accurate view for the analysis of the integrated framework underlying adaptation in complete genomes.

Figure 5.2. ***(following page)*****: Positive selection and evolution of functional modules.**
Circles and triangles represent the median values of dN and dS for KEGG pathways and GO terms (level 6-7), respectively in mammals, and in the *Drosophila* species. Functional modules with SHω and SLω results after GSSA are shown in red and blue. Those modules without statistical differences are gray. Yellow dots depict the median dS and dN values for *H. sapiens* (1), *P. troglodytes* (2), *M. musculus* (3), *R. norvegicus* (4), *D. simulans* (5), *D. sechellia* (6), *D. melanogaster* (7), *D. yakuba* (8) and *D. erecta* (9). (B) In this case, circles and triangles represent a subset (of A) with modules containing at least one PSG. Note that they are distributed along a wide range of values of dS and dN and in functional categories with significant (red/blue), and non-significant (gray) results after the GSSA (ω ratio).


A
B
dN
dS
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
1
2
3
4
5
6
7
8
9

6. Tools, programs, methods

6.1. Pipeline for study of adaptation at genomic scale

THE detection of selective pressures in one gene by comparing the rates of changes in its sequence is done over aligned sequences. This step is extremely sensible to misalignment, one column

6.1.1. Alignment

6.1.2. Testing substitution models

6.1.3. Phylogenetic reconstruction

6.1.4. ETE-evol plugin

6.1.5. BRANCHED1

6.1.6. Protamines Rodents and Primates

6.1.7. Selective pressure on duplicated genes in Drosophila

6.2. Phylemon

6.3. Ecolopy

Given the sampling formula (Equation 2.6) Ecolopy is able to generate random neutral species abundances distributions given a sample size J and a value of θ. The number of species generated is free according to the formula but can be fixed by keeping only those random abundances generated with the desired number of species. The likelihood function

(Equation 2.7) is also integrated in the program, and used for optimization of θ parameter (see subsection 2.2.6).

Models where optimized through different optimization strategies depending on the model selected. In the case of the Ewens' formula, θ is the only parameter to take into account, and its estimation is achieved with the *golden* optimization strategy [Jones *et al.* 2001]. For Etienne's model, two parameters were optimized, θ and m, using the best solution of the *downhill simplex algorithm* [Nelder & Mead 1965], *L-BFGS-B algorithm* [Byrd *et al.* 1995], *truncated Newton algorithm* [Nash 1984] and *Sequential Least SQuares Programming* all implemented in *Scipy* [Jones *et al.* 2001].

7. Conclusions

Bibliography

[Abascal *et al.* 2005] Federico Abascal, Rafael Zardoya, and David Posada, ProtTest: selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* **21**(9) (2005), 2104–5. ↪ *page 18*

[Abrusán & Krambeck 2006] György Abrusán and Hans-Jürgen Krambeck, Competition may determine the diversity of transposable elements. *Theoretical population biology* **70**(3) (2006), 364–75. ↪ *page 10*

[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. ↪ *pages 6 and 19*

[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. ↪ *pages 14, 38, 39, and 81*

[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. ↪ *page 14*

[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. ↪ *pages 14 and 81*

[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. ↪ *page 38*

[Alonso *et al.* 2006] David Alonso, Rampal S Etienne, and Alan J McKane, The merits of neutral theory. *Trends in ecology & evolution* **21**(8) (2006), 451–7. ↪ *pages 10, 71, and 74*

[Alonso *et al.* 2008] David Alonso, Annette Ostling, and Rampal S Etienne, The implicit assumption of symmetry and the species abundance distribution. *Ecology letters* **11**(2) (2008), 93–105. ↪ *page 71*

[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. ↪ *pages 12, 15, and 38*

[Azbel' 1995] Mark Ya Azbel', Universality in a DNA statistical structure. *Physical review letters* **75**(1) (1995), 168–171. ↪ *page 6*

[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. ↪ *page 12*

[Bassi 2007a] Sebastian Bassi, *Python for Bioinformatics*. CHAPMAN & HALL/CRC Mathematical and Computational Biology Series, 2007a. ↪ *page 28*

[Bassi 2007b] Sebastian Bassi, A primer on python for life science researchers. *PLoS computational biology* **3**(11) (2007b), p. e199. ↪ *page 16*

[Becher & Heiber 2011] Verónica Becher and Pablo Ariel Heiber, On extending de Bruijn sequences. *Information Processing Letters* **111**(18) (2011), 930–932. ↪ *pages 65 and 67*

[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. ↪ *pages 34 and 39*

[Blair Hedges & Kumar 2003] S Blair Hedges and Sudhir Kumar, Genomic clocks and evolutionary timescales. *Trends in genetics : TIG* **19**(4) (2003), 200–6. ↪ *page 36*

[Blanc & Wolfe 2004] 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. ↪ *page 62*

[Borcard *et al.* 2011] Daniel Borcard, Francois Gillet, and Pierre Legendre, *Numerical Ecology with R*. Springer, 2011. ↪ *page 16*

[Brookfield 2005] John F Y Brookfield, The ecology of the genome - mobile DNA elements and their hosts. *Nature reviews. Genetics* **6**(2) (2005), 128–36. ↪ *page 76*

[de Bruijn 1946] N G de Bruijn, A combinatorial problem. *Koninklijke Netherlands: Academe Van Wetenschappen* **49** (1946), 758–764. ↪ *pages 65 and 67*

[Burrows & Wheeler 1994] Michael Burrows and David J Wheeler, A block-sorting lossless data compression algorithm. *Digital SRC Research Report* **124** (1994). ↪ *page 19*

[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. ↪ *page 12*

[Byrd *et al.* 1995] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu, A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* **16**(5) (1995), p. 1190. ↪ *page 94*

[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. ↪ *pages 18 and 36*

[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. ↪ *page 14*

[Cavalier-Smith 2006] Thomas Cavalier-Smith, Rooting the tree of life by transition analyses. *Biology direct* **1**(1) (2006), p. 19. ↪ *page 2*

[Chaitin 1975] Gregory J. Chaitin, A Theory of Program Size Formally Identical to Information Theory. *Journal of the ACM* **22**(3) (1975), 329–340. ↪ *page 67*

[Charlesworth & Charlesworth 2009] Brian Charlesworth and Deborah Charlesworth, The population dynamics of transposable elements. *Genetical Research* **42**(01) (2009), p. 1. ↪ *page 76*

[Charlesworth *et al.* 1994] Brian Charlesworth, P Sniegowski, and W Stephan, The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**(6494) (1994), 215–20. ↪ *page 76*

[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, et al., Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science (New York, N.Y.)* **302**(5652) (2003), 1960–3. ↪ *page 12*

[Clark *et al.* 2007] Andrew G. Clark, Michael B Eisen, Douglas R Smith, Casey M Bergman, Brian Oliver, Therese A Markow, Thomas C Kaufman, Manolis Kellis, William Gelbart, Venky N Iyer, Daniel A Pollard, Timothy B Sackton, Amanda M Larracuente, Nadia D Singh, Jose P Abad, et al., Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167) (2007), 203–18. ↪ *pages 38 and 90*

[Colbourne *et al.* 2011] John K. Colbourne, Michael E. Pfrender, Donald Gilbert, W. Kelley Thomas, Abraham Tucker, Todd H. Oakley, Shinichi Tokishita, Andrea Aerts, Georg J. Arnold, Malay Kumar Basu, Darren J. Bauer, Carla E Cáceres, Liran Carmel, Claudio Casola, Jeong-Hyeon Choi, et al., The ecoresponsive genome of Daphnia pulex. *Science (New York, N.Y.)* **331**(6017) (2011), 555–61. ↪ *page 50*

[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. ↪ *page 41*

[Dennett 1995] Daniel C Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Penguin Books, 1995. ↪ *page 1*

[Dobzhansky 1973] Theodosius Dobzhansky, Nothing in biology makes sense except in the light of evolution. *American Biology Teacher* **35**(March 1973) (1973), 125–129. ↪ *page 1*

[Doolittle & Sapienza 1980] W F Doolittle and C Sapienza, Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**(5757) (1980), 601–3. ↪ *page 76*

[Dopazo] Joaquin Dopazo, Formulating and testing hypotheses in functional genomics. *Artificial intelligence in medicine* **45**(2-3) , 97–107. ↪ *page 38*

[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. ↪ *page 60*

[Edgar 2004] Robert C Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5) (2004), 1792–7. ↪ *pages 17 and 36*

[Eichinger & Noegel 2003] Ludwig Eichinger and Angelika A Noegel, Crawling into a new era-the Dictyostelium genome project. *The EMBO journal* **22**(9) (2003), 1941–6. ↪ *page 50*

[Eilbeck *et al.* 2005] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner, The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**(5) (2005), p. R44. ↪ *page 26*

[Etienne 2005] Rampal S Etienne, A new sampling formula for neutral biodiversity. *Ecology Letters* **8**(3) (2005), 253–260. ↪ *pages 16 and 29*

[Etienne 2007] Rampal S Etienne, A neutral sampling formula for multiple samples and an 'exact' test of neutrality. *Ecology letters* **10**(7) (2007), 608–18. ↪ *pages 28 and 32*

[Ewens 1972] Waren J Ewens, The sampling theory of selectively neutral alleles. *Theoretical population biology* **3**(1) (1972), 87–112. ↪ *page 28*

[Fisher *et al.* 1943] RA Fisher, AS Corbet, and C.B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* **12**(1) (1943), 42–58. ↪ *page 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, et al., Ensembl 2011. *Nucleic acids research* **39**(Database issue) (2011), D800–6. ↪ *pages 17, 21, 24, 26, and 36*

[Fousse *et al.* 2007] Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier, and Paul Zimmermann, MPFR: A multiple-precision binary floating-point library with correct rounding. *ACM Transactions on Mathematical Software (TOMS)* **33**(2) (2007), p. 13. ↪ *pages 16 and 28*

[Gardner *et al.* 2002] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan a Eisen, Kim Rutherford, Steven L Salzberg, et al., Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**(6906) (2002), 498–511. ↪ *page 50*

[Gaut 2001] 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. ↪ *page 56*

[Gaut & Doebley 1997] 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. ↪ *page 57*

[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. ↪ *page 60*

[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, et al., Creation of a bacterial cell controlled by a chemically synthesized genome. *Science (New York, N.Y.)* **329**(5987) (2010), 52–6. ↪ *page 49*

[Gojobori 1983] T. Gojobori, Codon substitution in evolution and the" saturation" of synonymous changes. *Genetics* **105**(4) (1983), p. 1011. ↪ *page 17*

[Granlund 2000] Torbjörn Granlund, GMP: The GNU Multiple Precision Arithmetic Library, http://gmplib.org/, 2000. ↪ *pages 16 and 28*

[Graur & Li 2000] Dan Graur and Wen-Hsiung Li, *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, second edi edition, 2000. ↪ *page 36*

[Gregory 2001] T Ryan Gregory, Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological reviews of the Cambridge Philosophical Society* **76**(1) (2001), 65–101. ↪ *page 4*

[Gregory 2005] T Ryan Gregory, Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics* **6**(9) (2005), 699–708. ↪ *pages 4, 5, and 6*

[Gregory 2012] T Ryan Gregory, Animal Genome Size Database, http://www.genomesize.com, 2012. ↪ *page 4*

[Guindon & Gascuel 2003] Stéphane Guindon and Olivier Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* **52**(5) (2003), 696–704. ↪ *page 18*

[Hankin 2007] Robin K S Hankin, Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. *Journal of Statistical Software* **22**(12) (2007), 1–15. ↪ *pages 15 and 28*

[He *et al.* 2010] Xionglei He, Wenfeng Qian, Zhi Wang, Ying Li, and Jianzhi Zhang, Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nature genetics* **42**(3) (2010), 272–6. ↪ *page 90*

[Hershberg & Petrov 2008] Ruth Hershberg and Dmitri a Petrov, Selection on codon bias. *Annual review of genetics* **42**(iv) (2008), 287–99. ↪ *page 13*

[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. ↪ *page 6*

[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, et al., The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics* **43**(5) (2011), 476–81. ↪ *pages 56, 64, and 68*

[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. ↪ *page 38*

[Hubbell 2001] Stephen P Hubbell, *The unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001. ↪ *pages 7, 10, 15, 28, and 72*

[Huerta-Cepas *et al.* 2010] Jaime Huerta-Cepas, Joaquín Dopazo, and Toni Gabaldón, ETE: a python Environment for Tree Exploration. *BMC bioinformatics* **11**(1) (2010), p. 24. ↪ *page 18*

[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. ↪ *page 14*

[Ideker & Sharan 2008] Trey Ideker and Roded Sharan, Protein networks in disease. *Genome research* **18**(4) (2008), 644–52. ↪ *page 14*

[Ihaka & Gentleman 1996] Ross Ihaka and Robert Gentleman, R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**(3) (1996), p. 299. ↪ *page 38*

[Jabot & Chave 2011] Franck Jabot and Jérôme Chave, Analyzing Tropical Forest Tree Species Abundance Distributions Using a Nonneutral Model and through Approximate Bayesian Inference. *The American naturalist* **178**(2) (2011), E37–47. ↪ *pages 10, 12, 16, 28, 32, and 71*

[Jabot *et al.* 2008] Franck Jabot, Rampal S Etienne, and Jérôme Chave, Reconciling neutral community models and environmental filtering: theory and an empirical test. *Oikos* **117**(9) (2008), 1308–1320. ↪ *pages 16 and 29*

[Jones *et al.* 2001] Eric Jones, Travis Oliphant, Pearu Peterson, and Others, Scipy: Open source scientific tools for Python, http://www.scipy.org/, 2001. ↪ *pages 30 and 94*

[Jurka *et al.* 2005] J Jurka, V V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**(1-4) (2005), 462–7. ↪ *page 22*

[Kapitonov & Jurka 2008] Vladimir V Kapitonov and Jerzy Jurka, A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature reviews. Genetics* **9**(5) (2008), 411–2; author reply 414. ↪ *page 24*

[Katoh *et al.* 2005] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research* **33**(2) (2005), 511–8. ↪ *page 17*

[Kimura 1985] Motoo Kimura, *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, UK, 1985. ↪ *pages 7, 12, and 13*

[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. ↪ *pages 24 and 36*

[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. ↪ *pages 38, 83, and 90*

[Lander *et al.* 2001] Eric 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, et al., Initial sequencing and analysis of the human genome. *Nature* **409**(6822) (2001), 860–921. ↪ *pages 7 and 8*

[Le Rouzic & Deceliere 2005] Arnaud Le Rouzic and Grégory Deceliere, Models of the population genetics of transposable elements. *Genetical research* **85**(3) (2005), 171–81. ↪ *page 76*

[Le Rouzic *et al.* 2007] Arnaud Le Rouzic, Thibaud S Boutin, and Pierre Capy, Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences of the United States of America* **104**(49) (2007), 19375–80. ↪ *pages 10 and 71*

[Lee & Sonnhammer 2003] Jennifer M Lee and Erik L L Sonnhammer, Genomic gene clustering analysis of pathways in eukaryotes. *Genome research* **13**(5) (2003), 875–82. ↪ *page 14*

[Leonardo & Nuzhdin 2002] Teresa E Leonardo and Sergey V Nuzhdin, Intracellular battlegrounds: conflict and cooperation between transposable elements. *Genetical research* **80**(3) (2002), 155–61. ↪ *page 10*

[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. ↪ *pages 6 and 67*

[Lynch 2000] M. Lynch, The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290**(5494) (2000), 1151–1155. ↪ *page 68*

[Lynch 2007] Michael Lynch, *The Origins of Genome Architecture*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 2007. ↪ *page 83*

[Lynch & Conery 2003] Michael Lynch and John S Conery, The origins of genome complexity. *Science (New York, N.Y.)* **302**(5649) (2003), 1401–4. ↪ *pages 76 and 77*

[MacArthur & Wilson 1967] Robert H. MacArthur and Edward O. Wilson, *The Theory of Island Biogeography*. Princeton University Press, 1967. ↪ *page 10*

[Maddison & Schulz 2007] D. R. Maddison and K.-S. Schulz, The Tree of Life Web Project, http://tolweb.org, 2007. ↪ *page 3*

[Magurran 2004] Anne E. Magurran, *Measuring Biological Diversity*. Blackwell Science Ltd, 2004. ↪ *page 7*

[Martelli 2007] Alex Martelli, GMPY Multiprecision arithmetic for Python, http://code.google.com/p/gmpy/, 2007. ↪ *page 28*

[Massingham & Goldman 2005] Tim Massingham and Nick Goldman, Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**(3) (2005), 1753–62. ↪ *page 18*

[Mayr 1942] Ernst Mayr, *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press, 1942. ↪ *page 11*

[Mayr 1983] E Mayr, How to carry out the adaptationist program? *American Naturalist* **121**(3) (1983), 324–334. ↪ *page 1*

[McGill *et al.* 2006] Brian J McGill, Brian a Maurer, and Michael D Weiser, Empirical evaluation of neutral theory. *Ecology* **87**(6) (2006), 1411–23. ↪ *page 34*

[McGill *et al.* 2007] Brian J McGill, Rampal S Etienne, John S Gray, David Alonso, Marti J Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J Enquist, Jessica L Green, Fangliang He, Allen H Hurlbert, Anne E Magurran, Pablo a Marquet, Brian a Maurer, Annette Ostling, et al., Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters* **10**(10) (2007), 995–1015. ↪ *pages 7 and 72*

[McShea 1996] Daniel W. McShea, Perspective: Metazoan Complexity and Evolution: Is There a Trend? *Evolution* **50**(2) (1996), p. 477. ↪ *page 2*

[Mirsky & Ris 1951] A E Mirsky and H Ris, The desoxyribonucleic acid content of animal cells and its evolutionary significance. *The Journal of general physiology* **34**(4) (1951), 451–62. ↪ *page 2*

[Miyata *et al.* 1980] Takashi Miyata, Teruo Yasunaga, and Toshiro Nishida, Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proceedings of the National Academy of Sciences of the United States of America* **77**(12) (1980), 7328–32. ↪ *page 13*

[Mustonen & Lässig 2009] Ville Mustonen and Michael Lässig, From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in genetics : TIG* **25**(3) (2009), 111–9. ↪ *page 89*

[Nachman & Crowell 2000] M W Nachman and S L Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1) (2000), 297–304. ↪ *page 23*

[Nash 1984] Stephen G. Nash, Newton-Type Minimization via the Lanczos Method. *SIAM Journal on Numerical Analysis* **21**(4) (1984), p. 770. ↪ *page 94*

[Nelder & Mead 1965] J A Nelder and R Mead, A Simplex Method for Function Minimization. *The computer journal* **7**(4) (1965), 308–313. ↪ *page 94*

[Nielsen 2001] R Nielsen, Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**(Pt 6) (2001), 641–7. ↪ *page 13*

[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. ↪ *page 12*

[Nies 2009] André Nies, *Computability and Randomness*. Oxford University Press, macintyre, edition, 2009. ↪ *page 67*

[Notredame 2010] Cedric Notredame, Computing multiple sequence/structure alignments with the T-coffee package. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 3** (2010), Unit 3.8.1–25. ↪ *page 17*

[Orgel & Crick 1980] L E Orgel and F H Crick, Selfish DNA: the ultimate parasite. *Nature* **284**(5757) (1980), 604–7. ↪ *page 76*

[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. ↪ *page 23*

[Petit & Barbadilla 2009] N Petit and A Barbadilla, Selection efficiency and effective population size in Drosophila species. *Journal of evolutionary biology* **22**(3) (2009), 515–26. ↪ *page 84*

[Plotkin & Kudla 2011] Joshua B Plotkin and Grzegorz Kudla, Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics* **12**(1) (2011), 32–42. ↪ *page 13*

[Pond *et al.* 2005] Sergei L Kosakovsky Pond, Simon D W Frost, and Spencer V Muse, HyPhy: hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)* **21**(5) (2005), 676–9. ↪ *page 13*

[Posada 2008] David Posada, jModelTest: phylogenetic model averaging. *Molecular biology and evolution* **25**(7) (2008), 1253–6. ↪ *page 18*

[Pray 2008] Leslie A Pray, Transposons: The Jumping Genes. *Nature Education* **1**(1) (2008). ↪ *page 9*

[Preston 1948] FW Preston, The Commonness, And Rarity, of Species. *Ecology* **29**(3) (1948), p. 254. ↪ *page 7*

[Rosindell *et al.* 2011] James Rosindell, Stephen P Hubbell, and Rampal S Etienne, The Unified Neutral Theory of Biodiversity and Biogeography at Age Ten. *Trends in ecology & evolution* **26**(7) (2011). ↪ *pages 7 and 74*

[van Rossum & de Boer 1991] G van Rossum and J de Boer, Interactively testing remote servers using the python programming language. *CWI Quarterly* **4**(4) (1991), 283–303. ↪ *page 28*

[Ryabko 1980] B Ya Ryabko, Data Compression by Means of a 'Book Stack'. *Problems Information Transmission* **16**(4) (1980), 16–21. ↪ *page 19*

[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**(Web Server issue) (2011), W470–4. ↪ *page 36*

[Sayers *et al.* 2009] Eric W Sayers, Tanya Barrett, Dennis a Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, et al., Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **37**(Database issue) (2009), D5–15. ↪ *page 21*

[Sayers *et al.* 2011] Eric W Sayers, Tanya Barrett, Dennis a Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M Fingerman, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, et al., Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **39**(Database issue) (2011), D38–51. ↪ *page 17*

[Shannon 1948] C E Shannon, *A mathematical theory of communication*, vol. 5. The Bell System Technical Journal, 1948. ↪ *pages 19 and 32*

[Shapiro & Alm 2008] B.J. Shapiro and E.J. Alm, Comparing patterns of natural selection across species using selective signatures. *PLoS genetics* **4**(2) (2008), e23+. ↪ *page 14*

[Smit *et al.* 2010] A F A Smit, R Hubley, and P Green, RepeatMasker Open-3.0, http://www.repeatmasker.org, 2010. ↪ *page 22*

[Smith & Smith 1996] J M Smith and N H Smith, Synonymous nucleotide divergence: what is "saturation"? *Genetics* **142**(3) (1996), 1033–6. ↪ *page 17*

[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, et al., A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**(6) (2005), 957–68. ↪ *page 14*

[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. ↪ *page 14*

[Subramanian *et al.* 2008] Amarendran R Subramanian, Michael Kaufmann, and Burkhard Morgenstern, DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for molecular biology : AMB* **3** (2008), p. 6. ↪ *page 17*

[Taft *et al.* 2007] Ryan J Taft, Michael Pheasant, and John S Mattick, The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays : news and reviews in molecular, cellular and developmental biology* **29**(3) (2007), 288–99. ↪ *page 6*

[Talavera & Castresana 2007] Gerard Talavera and Jose Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* **56**(4) (2007), 564–77. ↪ *page 18*

[Tavaré & Ewens 1997] Simon Tavaré and Waren J Ewens, Multivariate Ewens Distribution. In *Discrete Multivariate Distributions*, ed. N Johnson, S Kotz, and N Balakrishnan, chapter 41, 232–246. John Wiley & Sons, 1997. ↪ *page 28*

[Team 2011] R Development Core Team, R: A Language and Environment for Statistical Computing, http://www.r-project.org, 2011. ↪ *pages 15 and 22*

[Templeton 1989] Alan R Templeton, The meaning of species and speciation: a genetic perspective. In *The Units of evolution: essays on the nature of species*, chapter 9, 159–183. Bradford, 1989. ↪ *page 11*

[Thomas 1971] C A Thomas, The genetic organization of chromosomes. *Annual review of genetics* **5** (1971), 237–56. ↪ *page 4*

[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. ↪ *page 14*

[Van Rossum & Drake 2003] Guido Van Rossum and F L Drake, *Python language reference manual*. Network Theory Ltd., 2003. ↪ *pages 16 and 22*

[Vendrely & Vendrely 1948] R Vendrely and C Vendrely, La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales. *Cellular and Molecular Life Sciences* **4**(11) (1948), 434–6. ↪ *page 2*

[Venner *et al.* 2009] Samuel Venner, Cédric Feschotte, and Christian Biémont, Dynamics of transposable elements: towards a community ecology of the genome. *Trends in genetics : TIG* **25**(7) (2009), 317–23. ↪ *page 77*

[Venter *et al.* 2001] J Craig Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, Hamilton O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, et al., The sequence of the human genome. *Science (New York, N.Y.)* **291**(5507) (2001), 1304–51. ↪ *page 57*

[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. ↪ *page 36*

[Volkov *et al.* 2003] Igor Volkov, Jayanth R Banavar, Stephen P Hubbell, and Amos Maritan, Neutral theory and relative species abundance in ecology. *Nature* **424**(6952) (2003), 1035–7. ↪ *pages 10 and 15*

[Weber & Helentjaris 1989] D Weber and T Helentjaris, Mapping RFLP loci in maize using B-A translocations. *Genetics* **121**(3) (1989), 583–90. ↪ *page 56*

[Wernegreen 2002] Jennifer J Wernegreen, Genome evolution in bacterial endosymbionts of insects. *Nature reviews. Genetics* **3**(11) (2002), 850–61. ↪ *page 49*

[Wicker *et al.* 2007] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H Schulman, A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics* **8**(12) (2007), 973–82. ↪ *page 24*

[Wilks 1938] S. S. Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* **9**(1) (1938), 60–62. ↪ *page 30*

[Wolfe 2001] Kenneth H Wolfe, Yesterday's polyploids and the mystery of diploidization. *Nature reviews. Genetics* **2**(5) (2001), 333–41. ↪ *pages 57 and 63*

[Wright 1931] S Wright, Evolution in Mendelian Populations. *Genetics* **16**(2) (1931), 97–159. ↪ *page 7*

[Yang 2007] Ziheng Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8) (2007), 1586–91. ↪ *pages 18, 38, and 81*

[Yang 2009] Ziheng Yang, User Guide PAML : Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **3**(September) (2009). ↪ *page 17*

[Yang & Nielsen 2008] Ziheng Yang and Rasmus Nielsen, Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution* **25**(3) (2008), 568–79. ↪ *page 13*

[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. ↪ *page 38*

List of Figures

List of Tables

Glossary

Burrows-Wheeler *-transform*, also called block-sorting compression, is an algorithm used in data compression techniques such as bzip2. It is based on sorting all possible rotations of a given string, sort the result in lexicographic order and finally take the last character of each rotated string. 19

de Bruijn *-sequence* these is the mathematically defined string of characters with perfect equal frequency of sub-sequences: every possible combination of logarithmic length appears exactly once as a sequence of consecutive symbols. 65, 67

ecological niche The role of a species of organisms in an ecological community,defined by the resources that the species requires from its environment. The "competitive exclusion principle" implies that species can only stably coexist if they have different ecological niches. 7

LINE A long interspersed element sequence - typically used for non-long terminal repeat retrotransposons. 10, 22, 60, 62

LTR Long Terminal Repeat, a kind of retrotransposons with direct repeats of 300-500bp of DNA at each end of the element. These sequences resemble the integrated proviruses of retroviruses. 10, 22, 60, 62

retroposon A mobile DNA sequence that can move to new locations through an RNA intermediate. XXII, 10

retrotransposon An autonomous transposable element that can move to a new location through an RNA intermediate. Two major classes of retrotransposons exist, with or without long terminal repeats (see LTR and non-LTR). 9, 10

script It is a program written for a software environment that automate the execution of tasks which could alternatively be executed one by one by a human operator. 16

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.
-*species* in the case of ortholog retrival a *seed species* is the equivalent of a *seed sequence*. 17, 36

selfish DNA Sequences of DNA that accumulate in the genome through non-selective means, and which have a negative effect on the fitness of their hosts. 76

SINE A short interspersed element sequence - this is a retroposon sequence of less than 500 bp in length that does not encode the protein activities required for its movement. 10, 22, 60, 62, 68, 74

superfamily *Transposable elements'*-, it corresponds to the fourth level in the classification of transposable elements according to http://www.bioinformatics.org/wikiposon/doku.php?id=main. This level will gather elements based on the structure of the internal sequence. 24

transposon A mobile DNA sequence that moves to new genomic locations through a DNA route, rather than through an RNA intermediate. This movement is catalysed by the action of a transposase protein that is encoded by an autonomous element. 9–11, 22

trophic Of or involving the feeding habits or food relationship of different organisms in a food chain. 7

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
```