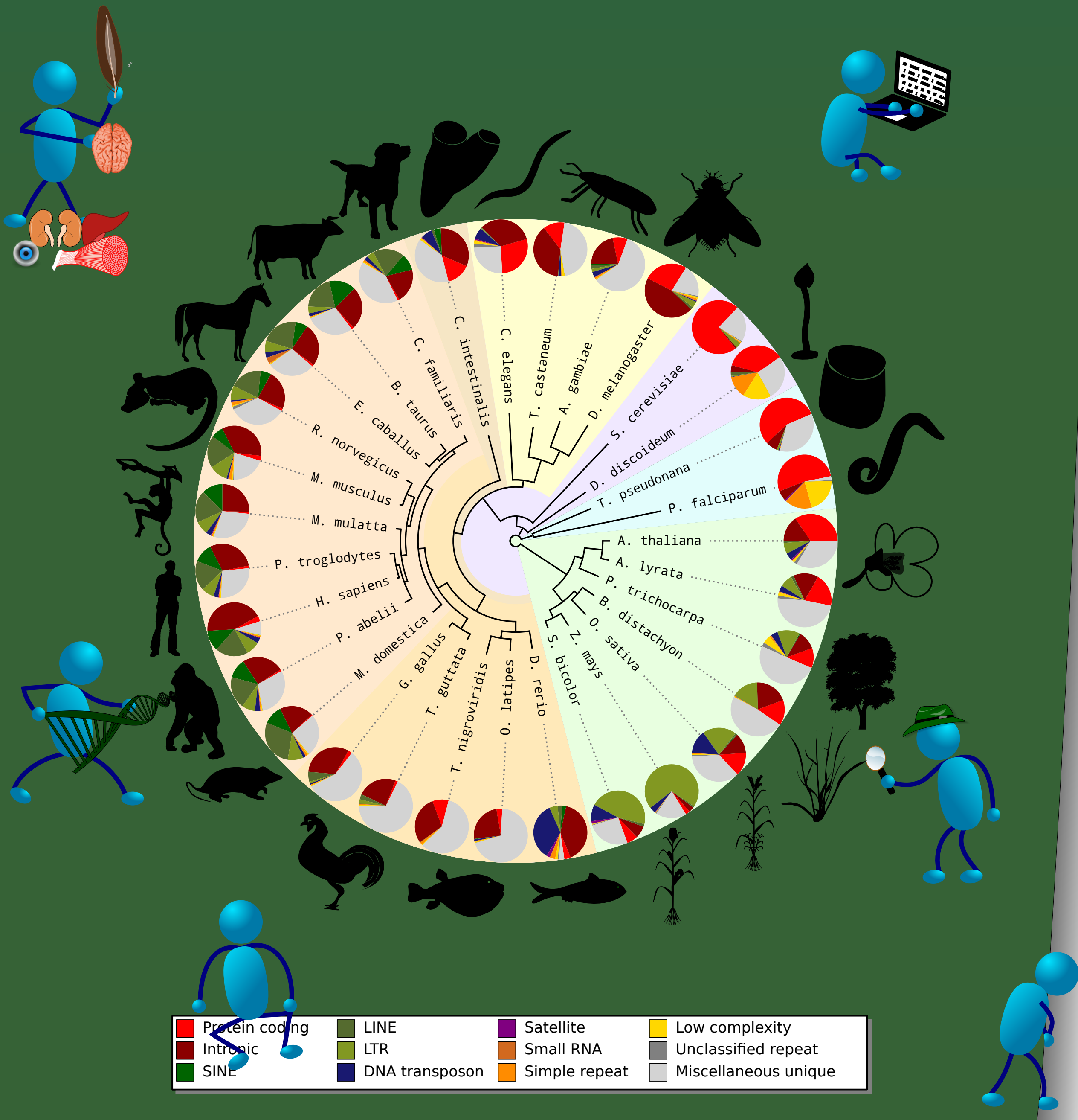
Informational, Ecological and System Approaches for Complete Genome Analysis



François Serra

VNIVERSITAT
DE VALÈNCIA

Programa de Doctorado de Biotecnología

PhD Thesis

Informational, Ecological and System Approaches for Complete Genome Analysis

Author:
François Serra

Supervisor:
Dr. Hernán Dopazo

Tutor:
Dr. Amparo Latorre

July 2012

Informe favorable

El Dr. Hernán Dopazo, investigador en el Departamento de Ecología, Genética y Evolución de la Facultad de Ciencias Exactas y Naturales en la Universidad de Buenos-Aires (Argentina).

CERTIFICA

Que la presente Tesis Doctoral, titulada *"Informational, Ecological and System Approaches for Complete Genome Analysis"*, ha sido realizada, bajo su dirección, por François Serra, licenciado en Biología por la Universidad *Université Louis Pasteur* de Estrasburgo (Francia); y que, habiendo revisado el trabajo, considera que reúne las condiciones necesarias para optar al grado de Doctor.

Y para que conste a los efectos oportunos, se expide la siguiente certificación.

Valencia, Julio de 2012

Acknowledgements

En primer lugar me gustaría agradecer a mi director de tesis, y amigo, Hernán Dopazo. A lo largo de estos años de tesis doctoral, consiguió transformar el mal estudiante que era en un apasionado investigador ávido por conocer y entender los patrones y procesos evolutivos que moldean... Lo ves?? Gracias por esto y por todas las oportunidades que me distes. Gracias por la emoción transmitida, la generosidad, la disponibilidad... Me lo pase muy bien en estos años, y sin duda tienes un gran parte de responsabilidad.

Gracias a los bioinfo, frikis y normales, los que apenas conocí y los que apenas habré conocido. Especialmente gracias a Fat, Naxo, David, Ximo, Lucía, Eva, Jordi, Joaquín, Marc, Toni por haber me ayudado a dar los primeros en bioinformática, en los bares, en las playas y en el baloncesto... A los que llegaron un pelín después Josete, Paco, Ana, Patricia (la que habla mucho), Patricia (la otra), Martina, Arturo, Kike, Marta, Luz, Jorge, Rodrigo, Miguel-Ángel, Roberto, Alejandro, David, Davide, Paco (el vikingo), Sonia, Marina, Salva. Gracias a los que me olvido, que no por ello son menos importantes, de hecho seguramente lo son mucho más! A Rubén por soportar mis ideas de bombero. A Juán por tener ideas de bombero. Gracias a Leo por ser un gran compañero y por ser bastante malo con los malabares. A Emidio por ensordecernos con su risa estruendosa. A Peio por tener un nombre tan gracioso. A Jaime por su pelo largo. A Pablo enseñarme tantos tacos. A Stefan... pues por muchísimas cosas la verdad, pero creo que sobre todo por el súper robot araña. A Pablo, Pilar, Peligro, Carolina, Rafa, el Seco, Imelda, Jota, Juan-Buena-Honda por hacer que me cueste tanto recordar mis dos primeros años en Valencia.

Merci à ceux qui, en dépit de l'horrible climat, sont venu me voir et apporter leur soutiens (j'ai pas le correcteur orthographique là... ça va déchirer), Ibt, Nadège, Léna, Snoop Damien, Marco, Dany, Isabelle, Simon, Marie-Paule... À ceux de la fac et particulièrement à Nico, Angy et Mathieux.

Gracias a la familia de Yasmina a Toni, Vicen y Alexis por ser tan bondadosos. A las amigas biólogas y a sus novios.

À la famille et particulièrement à José et Tati.
À Alexandre et Phillipe pour si bien reconnaître que j'ai toujours raison.
Merci à papa et maman pour m'avoir tout appris et être toujours là même si je n'ai jamais besoin de rien :)

Gracias a Yasmina por este trozo de camino que empezamos a recorrer y que me hace tan feliz.

À mes parents
À mes petits frêres
A Yasmina

Contents

List of Figures

List of Tables

Reading this thesis

Some particularities about this thesis:

- **The glossary**: some words are underlined with dotted lines like this. A short definition of these words can be found in the Glossary. The page numbers for all occurrences of the defined word appear at the end of each definition.

- **The bibliography**: the bibliography is quite standard. In the text references appear between square brackets [like this]. In the Bibliography, the number of authors listed is limited to 15. The page numbers corresponding to citations of references in the main text appear at the end of each reference (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
CNE	Constructive neutral evolution
CR	Complexity Ratio
CV	Complexity Value
dN	Rate of non-synonymous substitutions
dS	Rate of synonymous substitutions
FDR	False Discovery Rate
FET	Fisher Exact Test
GE	Genetic Element
GS	Genetic Species
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
My	Million years
pg	picogram
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. Evolution only makes sense in light of neutrality

A key concept in the study of evolution is the definition of neutrality. Conceptually, neutral evolution is simple – all evolutionary changes that do not imply variation in fitness are considered to be neutral. The identification of neutral changes is a key point in the understanding of the evolutionary forces that mold biological systems, as their theoretical ensemble represents the narrow area that segregates the two most famous categories of evolutionary changes – deleterious changes and advantageous changes. Once the conceptual importance of neutral evolution is appreciated, it may seem less surprising to find the first reference to it in *The Origin of Species*, where the prevalence of advantageous changes over deleterious ones is the theoretical demonstrable keystone of natural selection process.

> *"Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions."* [Darwin 1872]

However, from this historical starting point, real neutral processes have been considered as, at most, minor contributors to evolution. This is because it has been assumed that phenotypical changes that become fixed in species or populations are the direct consequence of a specific selection process, guided by an implied rise in fitness. Consequently, in any biological field, from ecology to (population) genetics through behavioral biology, the inferences made about the relative irrelevance of neutral processes when compared to directional changes, have been accepted, in spite of a lack of statistical proof. A context that may have pushed Gregory C. Williams to formulate the doctrine in *Adaptation and Natural Selection*:

> *"[...] adaptation is a special and onerous concept that should be used only where it is really necessary."* [Williams 1966]

In a sense, William's rule implies a rigorous definition of neutral changes in order to ensure the identification of real adaptive processes.

From an historical perspective, the first to design mathematical tools allowing estimation of the impact of neutral evolutionary changes was Ronald A. Fisher [Fisher 1930, Leigh 2007]. However, his work led him to consider as negligible the importance of neutral changes in the process of adaptation through natural selection. At the time, this view was only questioned by his contemporary, Sewall Wright, who, in contrast, emphasized the importance of drift or random moves in the race bringing species to states of higher fitness [Wright 1932, Frank 2012]. Nevertheless, this view on the fact that neutral changes may play an important role in the process of adaptation was hardly accepted.

Another important step forward in calling into question the adaptationist view was made by John Burdon Sanderson Haldane with his work on *The cost of natural selection* [Haldane 1957]. In his article, the cost of natural selection was formally stated in terms of the number of selective deaths necessary for spreading of a newly formed allele in a given population. This calculation led Haldane to establish an upper limit for the amount of selective deaths a species would be able to afford (around 1 gene substitution per 300 generations).

Even with the feeling that the amount of changes observed in nature, and attributed to adaptation, would hardly fit into Haldane's prediction, Wright's idea on the crucial role of neutral changes, had to wait until the first measures of molecular evolution. These measures, far higher than expected (1 substitution per 2 generations), were the "straw that broke the camel's back" and finally prompted Motoo Kimura and, independently Jack Lester King and Thomas H. Jukes [King & Jukes 1969], to present what Kimura called *the neutral theory of molecular evolution* [Kimura 1968] as the only possible explanation able to balance Haldane's books.

Though neutral theory was finally able to account for the constancy of the molecular clock and the amount of observed mutations, the theory faced some empirical problems [Ridley 2004] ***1)*** the influence of generation time over the rate of synonymous mutations was stronger than for non-synonymous ones; ***2)*** the constancy of the molecular clock was lower than expected; ***3)*** the observed levels of heterozygosity between species were too similar and the levels of heterozygosity in species with large population sizes were lower than expected; and ***4)*** levels of genetic variation not related to evolutionary rates. In response to these problems, Tomoko Ohta proposed the *nearly neutral theory of molecular evolution* [Ohta 1992], where the strict neutral theory was extended with a new class of mutations, nearly neutral. In this way, adaptation could play a role in selection on such nearly neutral mutations when population size was sufficiently large.

In ecology, although statistical early proposed models that lacked biological concepts were able to accurately describe the observed patterns of distribution of species abundances over ecosystems [Motomura 1932, Fisher *et al.* 1943, Preston 1948], elucidation of the underlying neutral processes arrived later than in population genetics. Indeed, as genetic adaptationism was populating bibliographies since Fisher; in ecology, the observed patterns of species richness and abundances based on the concept of the ecological niche [Hutchinson 1959] and the principle of competitive exclusion [Gause 1934] convinced most ecologists that species differences were mainly adaptive. Ecological communities were viewed as groups of species competing for their niches with a strength proportional to observed niche overlaps. In this scenario, new species originated through adaptation to new niches [Alonso *et al.* 2006].

This vision that most species differences were adaptive was challenged in the late sixties by Robert H. MacArthur and Edward O. Wilson with their *theory of island biogeography* [MacArthur & Wilson 1967]. Under this theory, differences between species could be the result of random processes. Indeed, as its base, all species are considered identical, with equal probabilities of extinction or migration from one island to another. Although, according to Stephen Hubbell [Hubbell 2001], this model met all the requirements of being neutral, the first formal neutral models were those proposed by [Watterson 1974, Caswell 1976], which were based directly on the calculation of neutral drift in population genetics, as defined by Ewens sampling formula [Ewens 1972].

Finally, these works, together with *the theory of island biogeography*, were recovered and extended by Stephen Hubbell in *The Unified Neutral Theory of Biodiversity and Biogeography* [Hubbell 2001], by integrating the random dispersion parameter to the ecological neutral model [Alonso *et al.* 2006].

Ultimately, molecular evolution, population genetics and ecology were finally integrating neutral models. Geneticists and ecologists began to picture the effect of neutral processes on the total amount of evolutionary changes and, consequently, to distinguish changes truly attributable to the process of adaptation.

In both ecology and molecular evolution, neutral processes were found to be responsible for most of the observed evolutionary changes. Or, in other words, and referring to Pierre Teilhard de Chardin's famous quote:

> *"(Evolution) general condition to which all theories, all hypotheses, all systems must bow and which they must satisfy henceforward if they are to be thinkable and true. Evolution is a light which illuminates all facts, a curve that all lines must follow."* [de Chardin 1955],

the *"curve that all lines must follow"* would indeed represent mostly neutral processes; directional changes being relegated to the condition of outliers. The differentiation of events appearing by successive random walks from those favored by natural selection, which only became possible after the definition of a neutral model, now became classical, or even compulsory, in the study of evolutionary changes in both genetics and ecology.

This thesis analyzes three different biological approximations in complete genomes. Each of which expresses explicitly a neutral model and its corresponding deviations. The expression of this common denominator all through the three main chapters contributes to the explanatory value of this thesis; without it, conclusions may be irrelevant. These approximations are:

- **Informational**: this is perhaps the simplest view of a genome. It considers only the sequence of A, C, G and T nucleotides as the fundamental and independent units constituting genomes. Under this perspective, genomes are formed of a simple sequence with a given informational content. Defining a common structure of genomes across the diversity of life is our first objective, which incorporates as a first step an outline of the genomic substrate. We show how genomes are characterized by the universal quasi-random structure of their genetic information.

- **Ecological**: while the description of abundances and diversity of genetic elements in genomes (mostly transposable elements [Venner *et al.* 2009]) is still in its infancy, ecologists are already implementing neutral models to predict ecosystem composition. In this context, it may seem natural to apply ecological models [Hubbell 2001,Etienne 2005] to the different genetic elements populating genomes in order to expose the neutral patterns that explain their dynamics.

- **Systems**: this last perspective of genomes that we analyze, is phenotypical. In particular, we focus on proteins working together, to complete a biochemical pathway, or to fulfill a molecular function. In the case of protein-coding genes, neutrality is precisely defined [Kimura 1985], and it is possible to test with precision for any deviations. However, at the level of functionally-related genes, classical methodologies fail to find the fingerprint of natural selection. In response to this difficulty, we present an alternative strategy, specifically implemented to detect the effect of selective pressures at the level in question.

1.2. Genome content and information – Challenging the C-value paradox

A genome represents the overall genetic information of an organism – it provides all of the guidelines required by an organism to operate. But, contrary to what is suggested by the word's etymology, i.e. the combination of the words *gene* and *chromosome*, a genome also includes a wide variety of non-coding sequences distributed along the ensemble of its chromosomes. The information conveyed by a genome is encoded either in DNA or in RNA, and in all cases this biological codex is comprised of four nucleotidic bases, conventionally represented by the letters A, C, G and T/U.

This codification of biological information is universal for all living species, even though variation in structure and information quantity is vast.

1.2.1. Biological complexity versus genome size

Darwinian evolution does not encompass directional change or global adaptive improvement. In terms of biological complexity, this statement may be supported 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 the most basal bacteria in the tree of life (which, for example, according to Thomas Cavalier-Smith could be found in Chloroflexi's phylum [Cavalier-Smith 2006]) to mammals (Figure 1.1), some evolutionary paths underwent undeniable directional gains in complexity. However, even where gradual acquisition of an increasing number of complex traits can be shown, the neutrality of these changes can hardly be rejected [McShea 1996]. The ratchet-like model of constructive neutral evolution (CNE) [Gray *et al.* 2010,Stoltzfus 1999] can be considered an example of a test for neutral increase in complexity, which has been successfully applied to explain the appearance of complexity in small-scale systems like RNA editing [Speijer 2011]. Up to now, this model still suffers some criticisms, being considered as too simple [Speijer 2011] (see also the reply [Doolittle *et al.* 2011]).

The main problem when trying to define if natural selection favors (or not) incremental biological diversity is precisely the definition of the complexity of organisms. At the 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 amount of DNA or the C-value (amount of DNA found in a haploid nucleus) correlates with either organism complexity or even with the number of genes [Mirsky & Ris 1951]. This contradiction was referred to as the **C-value paradox** [Thomas 1971], with the most striking example being *Amoebae dubia* which possesses 200 times more DNA than humans [Friz 1968,McGrath & Katz 2004]. Recently, famous work is being done by T. Ryan Gregory [Gregory 2001], who introduced a nuance to the contradiction and prefers the term **C-value enigma**. Utilizing a higher number of species measurements [Gregory 2012], the paradox is now more conspicuous than ever (the spectrum of C-values now extends from 0.002pg for parasitic microsporidium *Encephalitozoon intestinalis* and 1,400pg for the free-living amoeba *Chaos chaos* – that is almost six orders of magnitude difference, Figure 1.2).

Questions raised by the imbalance between the C-value and the number of genes, or the biological complexity or even clade specificity (see Figure 1.2) can be summarized in these points [Gregory 2005]:

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

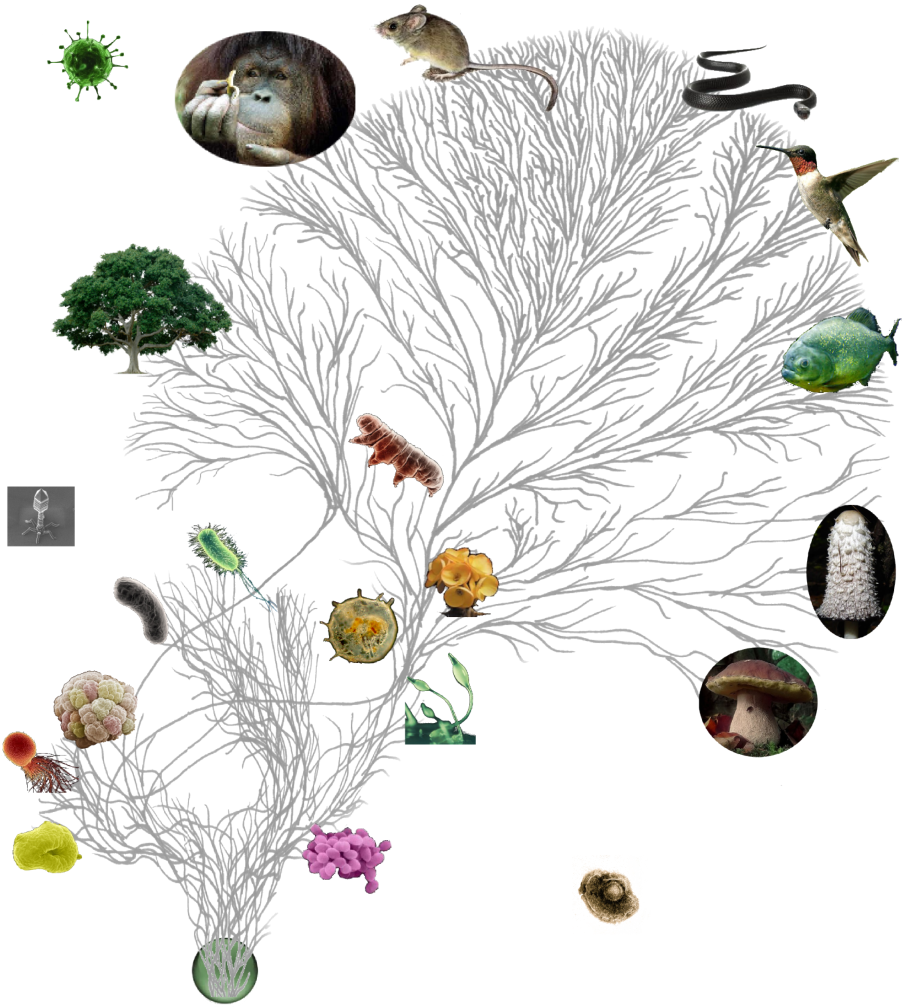


Figure 1.1.: Overview of the tree of life.
Picture adapted from the Tree of Life Web Project ©2007 [Maddison & Schulz 2007]. This picture shows how biological complexity can develop from a unicellular Universal Common Ancestor (bottom of the tree), to tips representing living species, with, in some cases, the acquisition of multicellularity, division of labor, evolution of meiosis, sexual reproduction, cell differentiation, early arrest of reproductive cells, etc.

what proportions?

2. What is the origin of non-coding DNA 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?

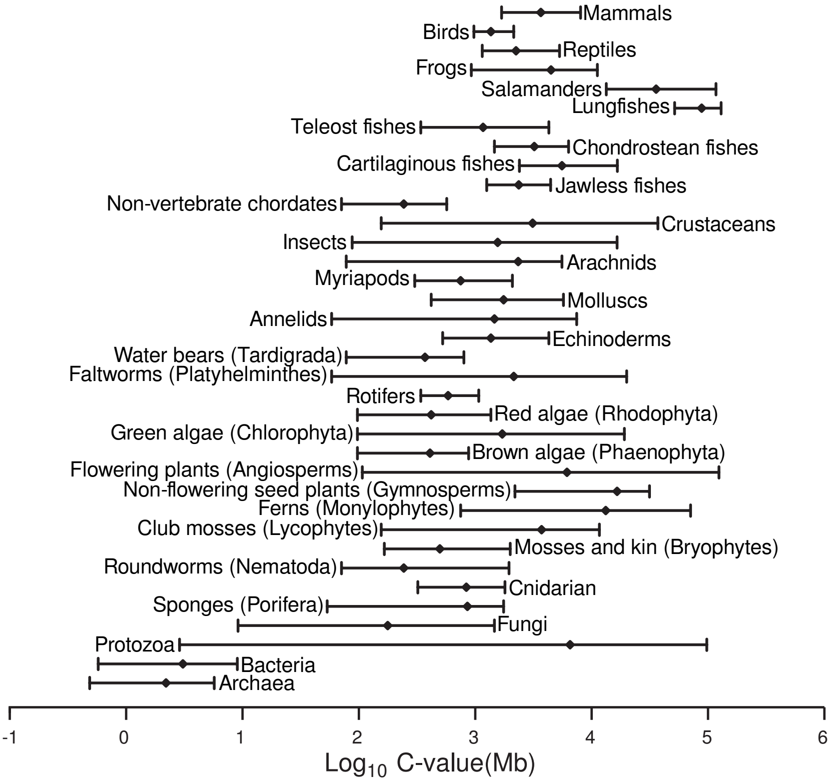


Figure 1.2.: C-values of the main groups of life.
Variation in genome size within and among the main groups of life, adapted from [Gregory 2005].

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 into two problems; the search of patterns in the informational content of genomes throughout the diversity of life and the dynamics underlying the distribution and appearance of non-coding DNA. These two questions constitute the third and fourth chapters of this thesis.

1.2.2. Informational content of DNA

From a biological perspective it is almost impossible to imagine DNA as a random mix of A, C, G and T nucleotides. Genomes contain the entirety of our hereditary information coded either into DNA (or RNA for some viruses). Genomes are composed of functional elements such as 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. However, and challenging quite established concepts, we could ask: *To what extent can we state that genomes are not a random soup of four letters?* Intuitively, the assumption is that, at least in Eukaryotes, DNA presents a quite simple structure weighed down by the bulk of repetitive genetic elements (GE) that populate genomes. Actually, if sequences can be divided into functional categories, a fair assumption would be that protein-coding genes would present a specific selection of nucleotides with surely the highest informational content, while introns would tend more to random assembly, and lastly simple repeats would present some biases towards 2 or 3 nucleotides (e.g.: CpG islands) surely lowering dramatically sequence complexity.

One relevant point in the C-value paradigm is that some species, with similar levels of complexity or numbers of genes, present large differences in genome size. These differences may be explained by the spread of repetitive genetic elements and large- or small-scale genomic duplications [Gregory 2005]. Thus, it could be expected that by reducing the importance of repetition in measurements of genome size, the defective correlation between DNA content and organism complexity could be corrected. Methodologically, a simple way of achieving this would be to use compressed genomes sizes, as proposed by Ryan J. Taft et al. [Taft *et al.* 2007], since data compression algorithms take advantage of the presence of repetitions. The difference between compressed genome size and real genome size could then be viewed as genome complexity.

In contrast with biological complexity, measurements of DNA complexity are generally more accepted, and provide us an appropriate tool to decipher DNA structure. Here, the different measures of DNA complexity by Taft et al. [Taft *et al.* 2007] are worth mentioning. In their publication, they proposed an association between biological complexity and compressed genome size, hypothesizing that simple organisms compress better than complex ones. Another study conducted by Zhandong Liu et al. [Liu *et al.* 2008] deduced, from the comparison of the occurrences of n-mers across the genomes of seven species, that the human genome is not perfectly random as it lacks some fixed-length sequences (which would be mandatory in a random context). In contrast with this observation, a study on the statistical structure of one piece of human a chromosome [Azbel' 1995] and on bacteriophage lambda, resulted in the suggestion that the statistical structure of DNA is universal for life.

The third chapter of this thesis focusses on calculating DNA complexity, but this time, supported by many recently sequenced genomes. The work presented in this context also takes advantage of a global measure, i.e. getting one single value per genome instead of using many local estimations. The complexity value that we define to describe the statistical structure of DNA is based on classical methods for data compression [Adjeroh *et al.* 2008], and it is able to detect the typical regularities caused by repetitive sequences among the data. A very similar measure was recently used in [Holste *et al.* 2001], but it was only applied to human chromosome 22 in a sliding-window analysis.

The related results presented in this thesis (chapter 3: Random-like structure of DNA) show that throughout the entire diversity of life covered by the set of target species, the ratio between complexity and sequence size is almost maximal in both genomes and chromosomes, with a notable exception of recent polyploids.

1.2.3. Dynamics of genetics elements

From the first analysis of the human genome [Lander *et al.* 2001], there was a good approximation of the relative proportions corresponding to each of the families of genetic elements (Figure 1.3).

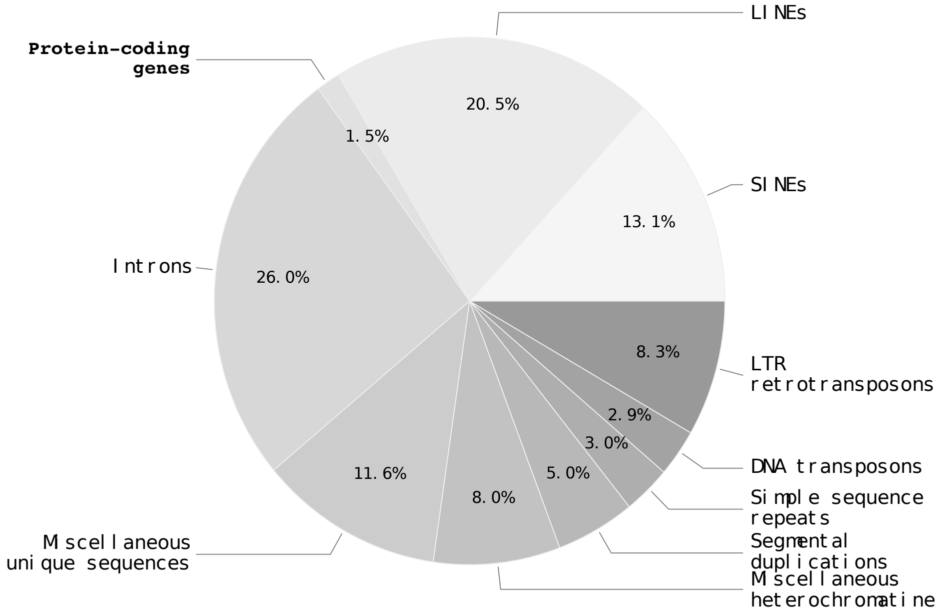


Figure 1.3.: Genomic components of the human genome.
Relative proportions of major families of different genomic elements (GE) in the human genome according to [Lander *et al.* 2001].

In the years following 2001 the arrival of other sequenced genomes revealed a significant variation in the proportions of families of genetic elements between species. As an example, Figure 1.4 shows the variation in proportions of the two major families of transposable elements (TEs) in different eukaryotic species.

In the fourth chapter of this thesis, we focus on the dynamics beyond the distribution of genetic elements in a broad array of eukaryotic genomes. We define "genetic species" as all kinds of non-coding DNA according to classic families of

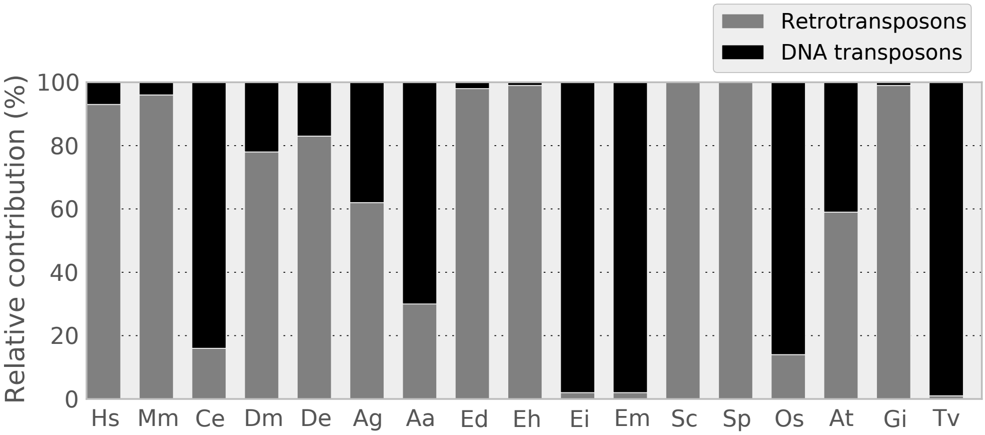


Figure 1.4.: Relative occurrence of retrotransposons and DNA transposons in diverse eukaryotic genomes.
This graph shows the contribution of DNA transposons and retrotransposons in percentage relative to the total number of transposable elements in each species. Species abbreviations: **Hs**=*Homo sapiens*, **Mm**=*Mus musculus*, **Ce**=*Caenorhabditis elegans*, **Dm**=*Drosophila melanogaster*, **De**=*Drosophila erecta*, **Ag**=*Anopheles gambiae*, **Aa**=*Aedes aegypti*, **Ed**=*Entamoeba dispar*, **Eh**=*Entamoeba histolytica*, **Ei**=*Entamoeba invadens*, **Em**=*Entamoeba moshkovskii*, **Sc**=*Saccharomyces cerevisiae*, **Sp**=*Schizosaccharomyces pombe*, **Os**=*Oryza sativa japonica*, **At**=*Arabidopsis thaliana*, **Gi**=*Giardia lamblia*, **Tv**=*Trichomonas vaginalis*. Adapted from [Pritham 2009].

repetitive elements [Wicker *et al.* 2007,Kapitonov & Jurka 2008], and also coding sequences classified into biotypes [Flicek *et al.* 2011].

Analogy with ecology

When Laurent Keller tackles the problem of selfish genetic elements in his book *Levels of Selection in Evolution*, he mentions that for the study of selective forces acting on them, a solution might be to reuse the well-established methodology of the ecology of social life:

> *"Understanding the evolution of selfish genetic elements and their importance in causing intragenomic conflict does not require any novel concepts. The same logic used to understand social interactions between separate organisms applies to the evolution of cooperation and conflict between genes within an organism"* [Keller 1999]

Following on from this thought, in an attempt to elucidate the processes governing the distribution of genetic elements in eukaryotic genomes, we apply recognized ecological methodologies.

When describing ecosystems, ecologists usually focus on the living species' natural environment, and their distribution and abundances within it. One common pattern that arises when studying different ecosystems is that, whatever environment studied and at whatever trophic level, it seems to be universal that a few species "dominate" so that they comprise the majority of the individuals in the ecosystem, while the remaining species are relatively rare [Preston 1948, Fisher *et al.* 1943]. This comprehensive pattern raises a natural question: *What mechanisms control this uneven distribution of species abundance in ecological communities?*

This problem, which can be reduced to the study of species diversity and abundance, is one of the oldest and more active topics in ecology [McGill *et al.* 2007], or again citing Darwin:

> *"When we look at the plants and bushes clothing an entangled bank, we are tempted to attribute their proportional numbers and kinds to what we call chance. But how false a view is this!"* [Darwin 1872]

To evaluate the accuracy of this vision and to actually infer the influence of chance in the increase or decrease in abundance of some species, ecologists have been implementing increasingly complex models that may or may not include parameters related to species fitness. Roughly speaking, there are two kinds of ecological models of species abundance: descriptive (statistical-based) or mechanistic (niche-based or neutrals). While many mechanistic approaches assume ecological niche differences as the main cause driving community composition, the latter models assume that niche differences are null [Magurran 2004].

So, a community consists of a group of species whose strengths of competitive interaction strengths are determined by their niche overlaps, and new species originate through adaptation to new niches. This view was challenged by Robert H. MacArthur and Edward O. Wilson with their equilibrium theory of island biogeography [MacArthur & Wilson 1967], which was finally extended by Stephen P. Hubbell [Hubbell 2001].

The unified neutral theory of biodiversity (UNTB) [Hubbell 2001, Rosindell *et al.* 2011] is a neutral-stochastic theory originally inspired by population genetics [Kimura 1985, Wright 1931]. It assumes that individuals of trophically-similar species are ecologically identical. This provocative assumption implies that individuals, regardless of their species-specificity, are controlled by common birth, death, dispersal, and speciation rates. Such a model is therefore able to predict species diversity patterns according to very few parameters. Indeed, the observed values of number of species, the total number of individuals, and two extra parameters describing the species richness and the migration rate are sufficient to model species abundance diversity in a neutral context. Of these parameters the most important may be the fundamental biodiversity number (θ). θ is analogous to the $4N\mu$ of population genetics, and it governs species richness in the spatial and temporal scales. The other parameter that may be considered and estimated is the migration rates m (see **Fitting Neutral Ecological models**, page 29, for details about variations in Hubbell's neutral model).

In ecology, the neutral model is a useful null model against which alternative biological hypotheses of relative species abundance distributions can be tested [Volkov *et al.* 2003, Alonso *et al.* 2006]. One simple step to move away from UNTB's definition of neutrality (while remaining within the scope of neutrality) is, for example, to assume that the fitness or death rate of a species is dependent on its abundance (see addition of δ parameter to neutral model [Jabot & Chave 2011]).

Since the first mention of selfish DNA [Dawkin 1976, Doolittle & Sapienza 1980, Orgel & Crick 1980], the idea that TEs or genes might be considered as living entities is recurrent in bibliographies. For example, referring to TEs as:

> *"Tiny organisms [...] that survive by spreading their progeny on host chromosomes."* [Leonardo & Nuzhdin 2002]

As well as this example, attention is usually only placed on TEs, considering as incidental interactions with the remaining types of repetitive sequences or even genes. Nonetheless, dynamical ecological models have been successfully applied to genomes considering TEs interactions [Abrusán & Krambeck 2006, Leonardo & Nuzhdin 2002, Le Rouzic *et al.* 2007a]. Some of these models, the most complex, account for interactions like parasitism, competition and cooperation between different families of TEs.

The exclusive use of TEs in these models is certainly a consequence of the foggy or unresolved biological relations that may exist among the remaining genetic species, and also of the difficulty of considering each as "living organism" (as is feasible for TEs, see Table 1.1).

However, ideal genomic models would not only consider TEs, but also all diversity of genetic elements (GEs) populating eukaryote genomes: satellites sequences, DNA-transposons, LTRs-retrotransposons, LINEs, SINEs (retroposons), miRNA, rRNA, tRNA, and genes among the many functional and non-functional elements. Such a model, although already conceived [Le Rouzic *et al.* 2007b], does not yet exist for genomes.

The definition of "genetic species"

In biology, species are defined as the basic unit of biological classification. The limit between one species and another is classically defined from the observation of sexual capabilities or reproductive isolation. Thus, one species may be defined as the ensemble of individuals able to engender fertile offspring by interbreeding

Population genetics	Ecology
Host species	ecological niche
Host individual	patch
Host genome	Habitat
A TE family	A species
Autonomous TE	Host
Non-autonomous TE	Parasite
Reproduction	Migration
Copy number in a genome	Number of individuals in a patch
Transposition rate	Birth (growth) rate
Deletion rate	Death rate
Natural selection	Density-dependent mortality
TE sequence	Genome of an individual

Table 1.1.: Analogies between population genetic models of TEs dynamics and inter-specific relationships in ecology.
Table reproduced from [Le Rouzic *et al.* 2007b].

[Mayr 1942]. It is well known that this definition is almost impossible to apply literally for most living organisms, as they reproduce asexually. Moreover, within sexual species, at both ends of the range of sex "quantity" variation, from cases of "too little sex" (e.g. the thelytoky observable in arthropods or some lizards) up to "too much sex" (hybridization), the classical definition of species runs into some difficulties [Templeton 1989].

Likewise for 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 less abundant functional sequences such as RNA or genes. Here, in order to follow the analogy with ecological systems, we had to decide which entity should be considered as a "species" in genomes. The decision was neither trivial nor categorical, but we decided to use the lowest level of classification (with the lower number of individuals) that allows a functional definition of the sample (with no direct description of the sequence itself). We suggest that the level of hierarchy that corresponds to this condition is the "family" or "class" level according to the RepBase ontology, also referred to as "superfamily" according to the International Committee on the Classification of Transposable Elements (`http://girinst.org/conference/ICCTE.html`) [Kapitonov & Jurka 2008]. For simplicity, we refer to genetic species (GSs), putting together all repetitive elements and also biotypes (a transcript classification including protein-coding, pseudogene, and non-coding RNAs). Each of the elements belonging to a GS is referred to as an individual or a genetic element.

Application of an ecological methodology to genomic data

Here, taking advantage of the methods and models developed by ecologists, we propose three main questions: ***1)*** *is there a common pattern behind the relative abundance and diversity of GSs in genomes?*; ***2)*** in the case that such a pattern exists, *is it sufficient to explain, the diversities of functional and non-functional components in eukaryote genomes?*; and ***3)*** *to what extent does abundance and diversity of genome components reflect adaptive or stochastic outcomes?*

In order to answer these questions, we tested the statistical adjustment of UNTB predictions in 31 eukaryotic genomes. We present the results of these analyses in three different sections:

- First, we describe the shape of the distribution of GSs in genomes and chromosomes using relative species abundance (RSA) curves; classical graphical tools used in ecology. We also test the role of chance in the rise of these observed shapes through simulation of random distributions of GSs among chromosomes.
- Second, we apply another classical ecological methodology to calculate the species-area (or GSs-chromosome length) relationship.
- Third, we test the statistical adjustment of the neutral ecological theory of biodiversity to the relative abundance and diversity of GSs of eukaryote chromosomes.

We conclude that the abundance and diversity of GSs in most of the chromosomes studied is predicted by the stochastic dynamics of a model for which the principle of functional equivalence amongst elements is the primary assumption. Additionally, we extend this observation through a test of neutrality, confirming that a strong neutral component is behind the distribution and diversity of GSs in chromosomes. Finally we hypothesize that at large temporal and spatial scales across all classes of GSs, an overarching neutral or nearly neutral process governs the evolution of abundance and diversity of GSs in eukaryote genomes.

1.3. Genomic study of selective pressures in set of genes

In recent years, with the development of genomic data, computational evolutionary researchers have been trying to detect signals of selective pressure among a growing set of genomes. Overall, the methodologies applied have been based on a measure of significant deviation from neutrality for a given gene (methodologies that have been used since [Kimura 1985]). This approach, conceived for the study of a single gene, is successfully able to detect genes escaping from neutrality ($\omega \neq 1$ see Equation 1.1 page 16) and, in particular, positively-selected genes (PSGs) (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 these studies was able to find significant enrichment of a given functional trait among the groups of genes detected to be under positive selection. Nevertheless, taking all these results together, assiduous readers might perceive some comprehensive patterns

(Figure 1.5). For instance, functional terms related to *Sensory perception*, *Immune response* or *Regulation of transcription*, were present in almost all genomic studies of positive selection conducted in primate or rodent genomes.

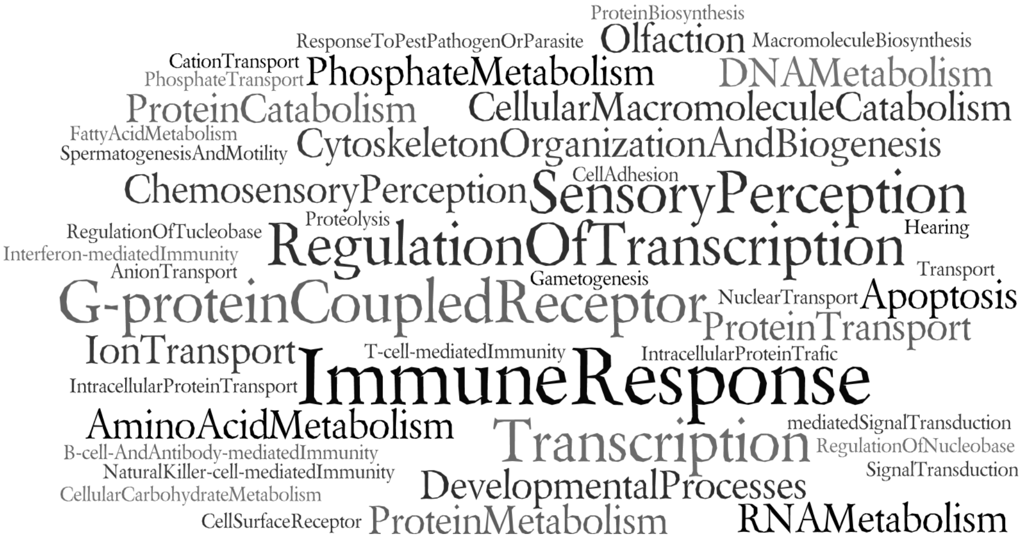


Figure 1.5.: Cloud of functional categories enriched in PSGs in initial genomic studies.
Cloud obtained by localizing nearly significant functional categories (e.g. significant before correcting for multiple testing) from initial genomic studies conducted in primates and rodents [Arbiza *et al.* 2006, Bakewell *et al.* 2007, Clark *et al.* 2003, Nielsen *et al.* 2005].

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

The amount of selective pressure acting on a given DNA sequence is usually inferred by comparing the number of changes reflected in phenotypes with the number of changes observed in regions known to escape from natural selection.

In the specific context of coding regions of the genome, changes occurring at nucleotide level can be divided into two categories depending on whether they are reflected in the translated protein sequences or not (respectively called non-synonymous or synonymous changes). Even though several studies have 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 the stranglehold of natural selection on those silent sites is weak [Yang & Nielsen 2008] and its use as a proxy for neutral mutation rate has successfully been used since the eighties [Miyata *et al.* 1980].

On the other hand, the rate of non-synonymous mutations is assumed to be related to selective pressure as mutations occurring at those sites may have functional consequence through changing protein structures or biochemical properties. Moreover, the non-synonymous mutation rate is significantly lower and more heterogeneous than the rate of silent mutations. The observation of these differences in rates is indeed the typical footprint of purifying selection [Kimura 1985], as most deleterious mutations, all of which appear in non-synonymous sites, are expected to be purged by natural selection.

Thus, assuming the proxy that silent mutations are neutral, the comparison of synonymous and non-synonymous mutation rates makes codons a perfect case in point for measuring the effect 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), namely, the ω ratio:

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

This ratio is estimated in coding regions and used to test for neutrality through different statistical methods [Nielsen 2001]. However, in the context of genomic studies, it is important to point out that the methodology does not take into account if genes works independently or in association with others to produce a single phenotypic response. In this sense, we are applying pre-genomics concepts and methods to genomics data.

1.3.2. Identification of selective pressures in the genomic era

The current paradigm for large scale analysis of adaptation consists of a two-step framework: **first**, the definition of a list of genes (in a gene-by-gene framework analysis) with a statistically-significant signal of positive selection ($\omega > 1$), and **second**, the search for over-represented functional classes within this list of genes. Although it is logically consistent, it has been noted that this kind of strategy results in a significant loss of information due to the large number of false negatives that are accepted in order to preserve a low ratio of false positives, which is necessary when thousands of tests are involved [Al-Shahrour *et al.* 2007, Al-Shahrour *et al.* 2005, Al-Shahrour *et al.* 2006, Subramanian *et al.* 2005].

Recently, a new methodology was proposed to improve the classical two-step analysis in an attempt to find selective signatures across species. This methodology simply consists of grouping the signal observed in related species in order to raise the statistical power, thereby expecting that the nearly significant functional terms reported in previous studies (Figure 1.5) would hold enough genes to reach significance. This approach has been successfully conducted in flies, mammals and a group of gamma proteobacteria [Shapiro & Alm 2008, Clark *et al.* 2007, Kosiol *et al.* 2008].

- Grouping PSGs of flies from the *melanogaster* group, the authors [Clark *et al.* 2007] were able to identify functional categories showing significant deviations. These categories included *Defense response*, *Proteolysis*, *DNA metabolic process*, and *Odorant binding*, among others.

- In mammals, the most representative functional categories found after pooling together all PSGs (400 genes) found in primates and rodents [Kosiol *et al.* 2008] were respectively, *Chemosensory perception* and *Defense/Immunity.*

- In a group of gamma proteobacteria, a classical test for positive selection was not applied. The authors [Shapiro & Alm 2008] used the deviations from the expected rates of evolution for a large group of genes, to infer selective pressures. The main conclusion was that the coherence of selective patterns suggested that the genomic landscape is organized into functional modules that were independently subjected to natural selection.

Results found by way of grouping the PSGs (or fast-evolving genes in the case of proteobacterias) of related species together, have allowed the statistical power of enrichment tests to be increased. Functional categories that, until then, were barely probable candidates (Figure 1.5) finally reached significance; although this was achieved at the expense of species-specificity.

1.3.3. Implementation of a new methodology

The hypothesis we aim to test in this study is not about individual genes, but about functional classes. Single genes undergo mutations, but natural selection acts on phenotypes by operating on entire sub-cellular systems [Oster & Alberch 1982]. Under a Darwinian view, mutations in genes either remain fixed or disappear, depending on their beneficial or disadvantageous effect on individual fitness respectively. This effect on the functioning of individual proteins can only be understood in the context of the system (e.g. a pathway, 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 extremities. This methodology is called the gene set selection analysis (GSSA).

This approach circumvents the implicit assumption posed by the two-step analysis described above; where it is assumed that the gene is specifically targeted by natural selection. However, if natural selection works by means of the minor quantitative effects of many different changes distributed along different gene products, most of them working together in a limited number of systems (GO functional terms, biochemical pathways), we than expect to find: ***1)*** correlated non-synonymous rate changes associated with these functions; ***2)*** synonymous rate changes not necessarily associated with the same functions; and *3)* a higher number of significant functions than those discovered in the classical two-step approach.

In the first part of chapter: **Searching for evolutionary patterns in functionally linked group of genes** (page 75) we extend the classical two-step approach previously reported by several authors for humans and chimps [Arbiza *et al.* 2006,Bakewell *et al.* 2007,Bustamante *et al.* 2005,Clark *et al.* 2003,Nielsen *et al.* 2005], to rats and mouses now considering a set of 12,453 orthologous genes from the human, chimpanzee, mouse, rat and dog genomes. In order to validate GSSA methodology outside of mammals, we also apply the GSSA to the *melanogaster* group of *Drosophila*, over the 9,240 orthologous genes between *D. yakuba*, *D. erecta*, *D. simulans*, *D. sechellia*, *D. melanogaster* and *D. ananassae*.

The objective of this part of the study is to find functionally-related groups of genes subjected to a common selective pattern, either conserved or accelerated. We identify these functional categories and discuss their similarities and differences in relation to the trends reported from the classical two-step approach (Figure 1.5). Finally, as GSSA do not directly involve PSGs, we test their significance and localization in the results.

1.4. Implementation of software and protocol

1.4.1. Ecolopy – A package to test for neutrality in genome ecosystems

Since the definition of neutral models in ecology [Hubbell 2001,Volkov *et al.* 2003] computational tools have been developed in order to manipulate data collected by ecological sampling and to apply specific statistical tests over them. There have been three main tools developed in this sense:

- The most complete tool takes its name directly from the neutral model it tests, i.e. the package untb [Hankin 2007]. This package is implemented in R language [Team 2011] and can be used together with other R packages that deal with ecological data and for executing most of the classical statistical analyses developep by and for ecologists [Borcard *et al.* 2011]. The main restriction of the package is related to the language used, which somewhat lacks computational efficiency.

- Another suite of scripts has been implemented by Rampal S. Etienne in the context of his publication [Etienne 2005]. These scripts are implemented in PARI/GP language, in the unique context of testing the UNTB and are now mainly adopted by the untb R package. The main advantage of this tool comes down to its simplicity of use and its speed.

- Finally, the programs that allow most efficient computation of the fit of neutral models are Tetame and Parthy, implemented by Franck Jabot and Jérôme Chave [Jabot & Chave 2011, Jabot *et al.* 2008]. Both are implemented in C++ language and are very fast. The only criticism we can suggest is their lack of flexibility, as a drawback of their computational efficiency.

All these packages allow determination of the fit of sampling data to the UNTB plus, in some cases, the possibility to generate the most classical ecological statistics. However, none of these packages could deal with the high numbers corresponding to the sampling of genomes and chromosomes to test for the UNTB in genomic data. For this reason, we started developing a new package "Ecolopy" in order to be able to deal with genomic data.

Ecolopy is a fully functional package for testing the UNTB, but still lacks the entire set of statistical tools available in R. However, its design and the use of Python [Van Rossum & Drake 2003] were employed in order to provide a scalable program architecture and to allow, if necessary, integration of some algorithms developed in R (through the RPy binding [Moreira & Warnes 2004] for example).

In this thesis, a short section is dedicated to go through some of the main features and advances that the program offers.

1.4.2. Computational molecular evolution

Classically, the detection of selective pressures in protein-coding genes can be achieved through five steps (see Appendix A for a short review), categorized as:

- **Definition of a set of species**: first of all, a seed species or sequence need to be defined. Starting from this sequence, a set of species needs to be selected that is not too distant from the seed in order to avoid saturation of synonymous changes [Gojobori 1983, Smith & Smith 1996] (taking humans as an example, the selected set of species should only come from other mammals). Usually it is recommended to have at least four sequences [Yang 2009].

- **Homologous sequences retrieval**: once the set of species is selected, the next step consists of retrieving homologous sequences; the most popular options being Ensembl [Flicek *et al.* 2011], the database resources of the NCBI [Sayers *et al.* 2011] or the UCSC database [Fujita *et al.* 2011].

- **Alignment**: the importance of this step is often underestimated. However, when measuring selective pressure and, in particular, for the detection of positive selection, misalignments generate a high proportion of false positives. The most popular tools used here are MUSCLE [Edgar 2004], MAFFT

[Katoh *et al.* 2005], Dialign [Subramanian *et al.* 2008] or T-COFFEE [Notredame 2010]. Once calculated, alignments are usually trimmed with Gblocks [Talavera & Castresana 2007] or Trimal [Capella-Gutiérrez *et al.* 2009] in order to remove abnormally divergent columns (sites) from the alignments.

- **Phylogenetic reconstruction**: this step may be unnecessary if the species tree is already known (as is generally the case nowadays). Otherwise, it involves the construction of a phylogenetic tree, paying particular attention to the use of model testing [Posada 2008, Abascal *et al.* 2005].

- **Identification of selective pressure and testing of evolutionary hypotheses**: this is the final step of the analysis. It requires the most computation time, and proved a real challenge to automate. Among the most popular programs used to calculate selective pressures and fit evolutionary models are SLR [Massingham & Goldman 2005] and CodeML from PAML package [Yang 2007].

Along this pipeline for detecting selective pressures in protein-coding genes, the last step that consists of formulating and testing evolutionary hypothesis is generally the most complicated to achieve and automate in the context of, for example, genomic studies. In this section of the thesis (in section **The "Evol" extension**, page 96), the ETE-Evol extension is presented; a tool that allows the formulation and testing of evolutionary hypotheses. ETE-Evol is an extension of the ETE program [Huerta-Cepas *et al.* 2010]; a tool for the manipulation, analysis and visualization of phylogenetic trees. This extension was originally designed in order to ease the integration of evolutionary hypothesis testing within a genomic pipeline, and also presents a broad range of features in order to interactively study and visualize the evolutionary history of a single group of homologous genes.

1.4.3. Phylemon web server

When Phylemon was first published [Tárraga *et al.* 2007], no other web platform that offered the possibility of conducting all the steps making up a complete phylogenetic analysis or to test for adaptation at the molecular level was available was available. Each one of the steps described above are indeed part of Phylemon. In addition to these utilities, this online platform offered a complete set of features allowing concatenation, storage and even visualization of a wide range of analyses.

Subsequently, other online platforms for phylogenetic analysis have emerged that present Phylemon-like features, namely Datamonkey [Delport *et al.* 2010], MobylePasteur [Néron *et al.* 2009] or Phylogeny.fr [Dereeper *et al.* 2008]. There are small differences in the tools proposed and the connections between them.

Bolstered by our experience when listening to users comments and teaching phylogenetic courses, we decided to release a new version of Phylemon; adding new tools but, and above all, emphasizing the integration of the tools and developing complete documentation illustrated by examples.

As the main features, Phylemon 2.0 proposes ***1)*** an integrated environment that enables the concatenation of evolutionary analyses and the storage of results; ***2)*** the concatenation of the tools allowing users to follow in their analysis the steps proposed by the server; and, finally, ***3)*** an enhanced "pipeliner" that permits to complete analyses, involving multiple tools, to be built graphically, and also allows to save, load or even share these pipelines.

In the last part of this thesis, I briefly describe these improvements, and discuss their implication for the scientific community.

2. Material and Methods

2.1. Measuring DNA complexity

2.1.1. The complexity ratio and complexity value

We define the complexity ratio (CR) in terms of classical formulae used in data compression [Adjeroh *et al.* 2008]. Its computation consists of three transformation steps for a given sequence. **First**, the Burrows-Wheeler transform (BWT) [Burrows & Wheeler 1994], **second** the Move To Front (MTF) [Ryabko 1980] algorithm and, **finally**, a summary 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 (i.e. the minimum possible value) only when a sequence consists solely of a single repeated symbol, which is the simplest possible combinatorial structure. Conversely, when entropy is equal to one (the maximum entropy value), it indicates that the sequence has a random-like combinatorial structure.

Algorithmically, the BWT of a given sequence summarizes all its lexicographically -sorted permutations. The MTF transforms a given sequence into a list of numbers. The higher the number, the less the character was used in the previous part of the sequence of length equal to the number of characters found in the sequence (in the case of DNA, this stack contains 4 characters). The MTF operates from left to right. Each generated number is an index in the stack and denotes an alphabetical symbol. Shannon's entropy converts a sequence into a real number between zero and one. It weights the frequency of the alphabetical 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; where N_i is the number of occurrences of i in s and $length(s)$ is the total length of the sequence s:

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

For the DNA alphabet, entropy is defined as:

$$E(s) = -\sum_{i=A,C,G,T} p_{(i)} \times log_4(p_{(i)}) \tag{2.2}$$

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

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

Table 2.1.: CR explained by an example.

These 3 tables summarize the steps followed to obtain the final sequence of numbers from which Shannon's entropy is computed. 1) The table on the left corresponds to the Burrows-Wheeler transform (BWT). The original sequence is rotated sequentially (so that the first character moves to the back), resulting in different strings and as many strings 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 sequence number 2 of the original order "#" takes the position 5 in lexicographic order). 2) The table in the center corresponds to the result of the BWT, which comprises the last character of each of the previously ordered sequences. 3) The table on the right corresponds to the application of the MTF algorithm. Starting from a sequence of all characters (named here "Char. list" and with four nucleotides in this case), the Move-to-front (MTF) algorithm calculates the index of the BWT nucleotide (upper case bold letter) in the "Char. list". In a second step, for the next iteration, MTF transforms the "Char. list", bringing to the front the corresponding BWT character (upper case letter). Finally, Shannon's entropy of these latter values is calculated (line below tables), which generates the CR (the CV is obtained by multiplying CR by the length of the sequence).

With i being the index of characters used, which for nucleotides ranges from 0 to 3. Thus, the CR can be given as:

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

The complexity value (CV) of a sequence is its CR multiplied by the number of characters in the 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 entire sequence, sequences can't be split for the analysis.

2.1.2. Complexity in strings

Genomic sequences

Complete genomes of 54 species were downloaded from the NCBI database resource [Sayers *et al.* 2009] and Ensembl Genome Project [Flicek *et al.* 2011]. Fourteen major groups of taxa were selected: viruses, phages, bacteria, archaea, fungi, amplicomplexa, heterokonta, amebozoa, urochordates, invertebrates, plants, fish, birds, and mammals. Species were chosen based on 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 a coverage of 6× or greater were chosen. Sexual chromosomes were excluded from the analysis, and ambiguous "N" characters were removed from sequences, and thus, excluded when measuring chromosome length. Genome complexity was calculated over concatenated chromosomes.

Complexity in biological sequences was computed in the +1 strand. Analysis of -1 strands did not generate significantly different results.

Random sequences with different ploidy levels were also needed for the study and these were generated with the Python base function: "random" [Van Rossum & Drake 2003]. The complexity value of biological and random sequences was computed with the DNA alphabet of four letters.

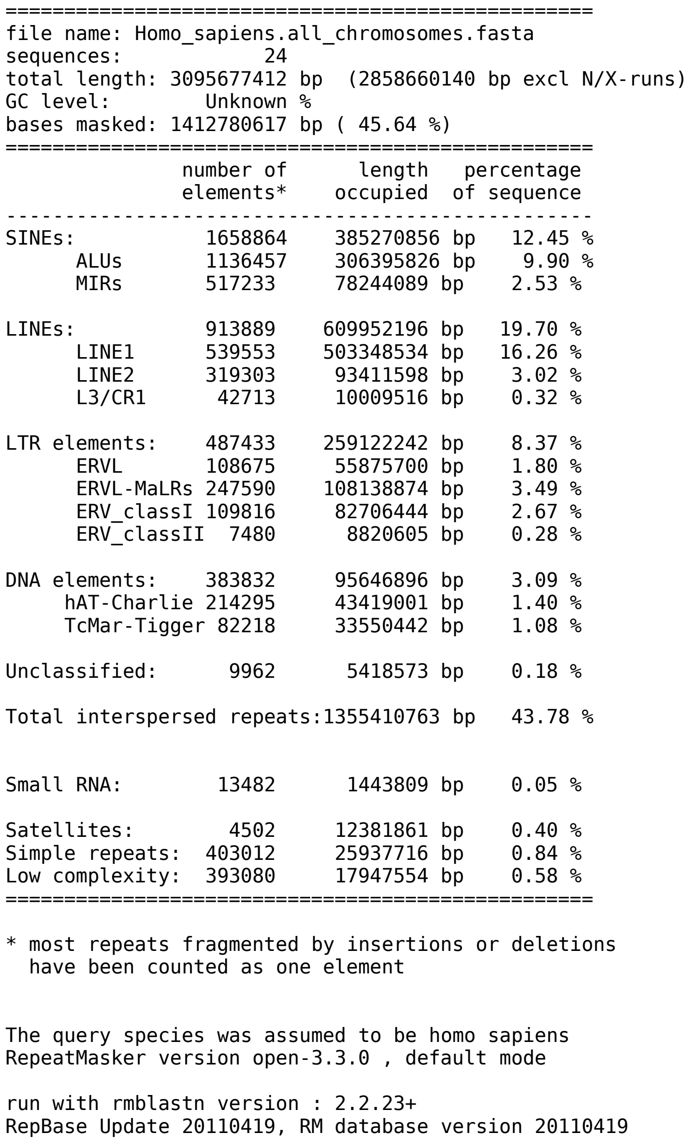
Annotation of repetitive elements

Interspersed repeats and low complexity DNA sequences were screened and mapped in the genomes of all 54 selected species using the RepeatMasker program [Smit *et al.* 2010]. Libraries of genetic elements were retrieved from RepBase (Release 20110419) [Jurka *et al.* 2005]. An example of a summary file generated by RepeatMasker is shown in Figure 2.1.

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

Human texts

Short stories, books and complete works in their original languages were downloaded from Project Gutenberg (`http://www.gutenberg.org/`). To automatically detect



```
==================================================
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
```

Figure 2.1.: RepeatMasker summary output file.
File returned by RepeatMasker for the human genome. It contains the proportion of each family and superfamily of genetic elements found by RepeatMasker, in relation to sequence size (in this case, the entire genome).

the sizes of the alphabets size in texts (including mathematical and punctuation symbols) we run the COMPL program (`http://kapow.dc.uba.ar/compl`) with the "auto" option, that takes into account all characters found, including mathematical symbols, and different punctuation signs.

Estimation of complexity in sliding and overlapping windows along chromosomes

To study complexity along chromosomes, a sliding window method that moves along chromosomes in overlapping units of 1.0 Kb to 100 Mb was performed.

2.1.3. Simulation of "polyploid" random sequence degeneracy through mutation and translocation

We performed four kinds of experiments, in which CV and CR were computed. **First**: the random polyploid construction of sequences of various sizes and ploidy levels (1× to 10×). **Second**: evolution over 40 million generations with a constant neutral mutation rate of $1.0e^{-08}$ mutations per site per generation (this value lies between the mutation rate estimated for *Homo sapiens*: $2.5e^{-08}$ [Nachman & Crowell 2000] , and *Arabidopsis thaliana*: $7.1e^{-09}$ [Ossowski *et al.* 2010]) for random sequences, and chromosomes of *Zea mays* and *Sorghum bicolor*. **Third**: evolution over 50,000 generations for random polyploid genomes of different sizes (100Kb, 1Mb, 10Mb) with 1.0 Kb translocations between chromosomes. The number of translocations per generation was set as a constant function of genome size (genome size divided by 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. The CV and CR values were calculated every 100 generations.

2.2. Measuring dynamics of genetic species

2.2.1. Genomes

For the study on the dynamics of genetic elements, the genomic sequences of 31 species from unicellular eukaryotes to mammals were used. These genomes correspond to a subset of the 54 genomes presented in the previous section (see subsection 2.1.2). References can be found in Table 3.1, page 49. The complete list of species used is: ***1)*** *Gallus gallus* (Bird) ***2)*** *Taeniopygia guttata* (Bird) ***3)*** *Danio rerio* (Fish) ***4)*** *Oryzias latipes* (Fish) ***5)*** *Tetraodon nigroviridis* (Fish) ***6)*** *Saccharomyces cerevisiae* (Fungi) ***7)*** *Anopheles gambiae* (Invertebrate) ***8)*** *Caenorhabditis elegans* (Invertebrate) ***9)*** *Drosophila melanogaster* (Invertebrate) ***10)*** *Tribolium castaneum* (Invertebrate) ***11)*** *Bos taurus* (Mammal) ***12)*** *Canis familiaris* (Mammal) ***13)*** *Equus caballus* (Mammal) ***14)*** *Homo sapiens* (Mammal) ***15)*** *Macaca mulatta* (Mammal) ***16)*** *Monodelphis domestica* (Mammal) ***17)*** *Mus musculus*

(Mammal) ***18)*** *Pan troglodytes* (Mammal) ***19)*** *Pongo abelii* (Mammal) ***20)*** *Rattus norvegicus* (Mammal) ***21)*** *Arabidopsis lyrata* (Plant) ***22)*** *Arabidopsis thaliana* (Plant) ***23)*** *Brachypodium distachyon* (Plant) ***24)*** *Oryza sativa* (Plant) ***25)*** *Populus trichocarpa* (Plant) ***26)*** *Sorghum bicolor* (Plant) ***27)*** *Zea mays* (Plant) ***28)*** *Dictyostelium discoideum* (unicellular Eukaryote) ***29)*** *Plasmodium falciparum* (unicellular Eukaryote) ***30)*** *Thalassiosira pseudonana* (unicellular Eukaryote) and ***31)*** *Ciona intestinalis* (Urochordate).

2.2.2. Mining Genetic Species

For this study, we define genetic species (GSs) as being the aggregate of superfamilies of repetitive elements and functional elements grouped into biotypes, each of which are described below.

Repetitive Elements

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

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

A complete list of the mapped superfamilies is shown in Table 2.2

Functional Elements

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

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

2.2.3. Simulated random distribution of genetic elements among chromosomes

In order to test for the random distribution of GSs among the chromosomes of each genome, we generated 1,000 genomes corresponding to each species with a random distribution of GSs for each generated genome.

Consequently, the GSs of each genome were distributed among the chromosomes according to a probability dependent on the size of the chromosome. As

Family	Superfamilies
ARTEFACT	ARTEFACT
DNA	Academ, Chapaev, Chapaev-Chap3, Crypton, En-Spm, Ginger, Harbinger, Kolobok-Hydra, Kolobok-IS4EU, Kolobok-T2, Maverick, Merlin, Mirage, MuDR, NOF, Novosib, P, PiggyBac, Sola, TcMar, TcMar-Ant1, TcMar-Fot1, TcMar-Gizmo, TcMar-ISRm11, TcMar-Mariner, TcMar-Mogwai, TcMar-Pogo, TcMar-Stowaway, TcMar-Tc1, TcMar-Tc2, TcMar-Tc4, TcMar-Tigger, TcMar-m44, Tourist, Transib, Zator, hAT, hAT-Ac, hAT-Blackjack, hAT-Charlie, hAT-Gulliver, hAT-Pegasus, hAT-Restless, hAT-Tag1, hAT-Tip100, hAT-Tol2, hAT-hAT1, hAT-hAT5, hAT-hATm, hAT-hATw, hAT-hATx, hAT-hobo
LINE	CR1, CRE, DIRS, DRE, Dong-R4, Genie, I, Jockey, L1, L1-Tx1, L2, L2-Hydra, LOA, Odin, Penelope, Proto1, Proto2, R1, R2, R2-Hero, RTE, RTE-BovB, RTE-RTE, RTE-X, Rex-Babar, Tad1, Zorro, telomeric
LTR	Caulimovirus, Copia, Copia(Xen1), DIRS, ERV, ERV-Foamy, ERV-Lenti, ERV1, ERVK, ERVL, ERVL-MaLR, Gypsy, Gypsy-Troyka, Ngaro, Pao, TATE, Viper
Low complexity	Low complexity
Other	Composite, DNA virus, centromeric, subtelomeric
RC	Helitron
RNA	RNA
SINE	5S, 7SL, Alu, B2, B4, BovA, C, CORE, Deu, Dong-R4, I, ID, L1, L2, MIR, Mermaid, R1, R2, RTE, RTE-BovB, Salmon, Sauria, V, tRNA, tRNA-7SL, tRNA-CR1, tRNA-Glu, tRNA-L2, tRNA-Lys, tRNA-R2, tRNA-RTE
Satellite	W-chromosome, Y-chromosome, acromeric, centromeric, macro, subtelomeric, telomeric
Simple repeat	Simple repeat
Unknown	Y-chromosome, centromeric
rRNA	rRNA
scRNA	scRNA
snRNA	snRNA
tRNA	tRNA

Table 2.2.: Superfamilies of repetitive elements and description. Complete list of the superfamilies of repetitive elements mapped by RepeatMasker using the RepBase library (Release 20110419).

an example, in the human genome, it was around six times more likely for a GS to occur in chromosome 1 than chromosome 22 (chromosome lengths were 225 megabases and 35 megabases for chromosome 1 and 22 respectively).

Since chromosome size is critical to the randomization process, we decided to remove chromosome regions where no GSs would be found (e.g. centromeric regions or incompletely sequenced regions) from the computation. For this purpose, we defined chromosome size as being the sum of all 10 kilobase regions containing at least one GS (see Table 2.4).

2.2.4. Fitting 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

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.
List of biotypes used a short description retrieved from the Ensembl glossary [Flicek *et al.* 2011] and from the Sequence Ontology browser [Eilbeck *et al.* 2005].

Chromosome	Chromosome length	Corrected length	Percentage remaining
1	249,240,621	225,200,000	90.35%
2	243,188,741	237,670,000	97.73%
6	171,048,878	167,050,000	97.66%
10	135,524,747	131,040,000	96.69%
12	133,841,891	129,970,000	97.11%
15	102,521,389	81,520,000	79.52%
20	62,962,324	59,430,000	94.39%
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.: A sample of human chromosome size changes after removing regions without genetic elements (GEs).

given sample. However, the formula can be applied to other scientific fields. In the context of the study of ecological communities, its application was first suggested by Tavaré and Ewens [Tavaré & Ewens 1997] and finally implemented by

Hubbell [Hubbell 2001]. Hubbell proposed a model that defined the fundamental biodiversity parameter θ (Equation 2.5), given the speciation rate ν and J_M as the size of the metacommunity.

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

The estimation of θ alone is sufficient to directly apply 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's 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], which includes cases of $m < 1$, taking into account the number of immigrants I depending on the sample size J:

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

Etienne's sampling formula is then 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:

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

Once $K(D, A)$ has been computed, the likelihood of the model can be optimized (Equation 2.9) by varying the values of the parameters θ and m (see Model optimization subsection 2.2.5) for a given dataset.

2.2.5. Model optimization

Models where optimized using different optimization strategies depending on the model selected. In the case of Ewens formula, θ is the only parameter taken into account, and its estimation is achieved with a single optimization step. For Etienne's model, two parameters were optimized, θ and m, using the best solution of different optimization strategies (see chapter **Ecolopy**, page 89 for more details).

A way to ensure that the optimized parameters of Etienne's model truly point to the maximum likelihood consists of placing them over a likelihood surface corresponding to a range of θ and m values. The computation time needed to generate such likelihood contour plots prevented its application to all chromosomes in the dataset. Nevertheless, for the five chromosomes tested, the optimal values of θ and m visually represented in the contour plots were congruent with the results of the optimization (see Figure 2.2 as an example of this validation step).

2.2.6. Model testing – Likelihood-ratio test

In order to compare and test the fit of a given distribution in the two models computed, a likelihood-ratio test [Wilks 1938] can be conducted between them. Etienne's model has two free parameters (FP) while Ewens' model only has one. Thus, the number of degrees of freedom for the chi-squared distribution is 1 ($df = FP_{Etienne} - FP_{Ewens} = 1$).

2.2.7. Testing UNTB

In recent years, at least two tests have been developed in order to accept or reject the neutrality of a given ecological community. Both these tests are based on the comparison of a given number of random neutral communities (or replicates) to the observed distribution of abundances. Since replicates are generated using the parameters estimated (see subsection 2.2.5) for the real data under a given neutral model (either Ewens or Etienne's model), we expected that they would be very close to the original distribution of abundances. The distances between replicates and the original data is thus a measure of how well the data fits a neutral model.

The first test [Etienne 2007] consists of comparing the likelihoods of data fitting the neutral model. Random neutral abundances are fitted to the model, and their likelihoods are used to build a distribution of values. Then, this distribution is compared to the likelihood of the real data. The major problem with this test is technical; the computation time needed to optimize the parameters of each simulated distribution is unrealistically high when dealing with genomic data.

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

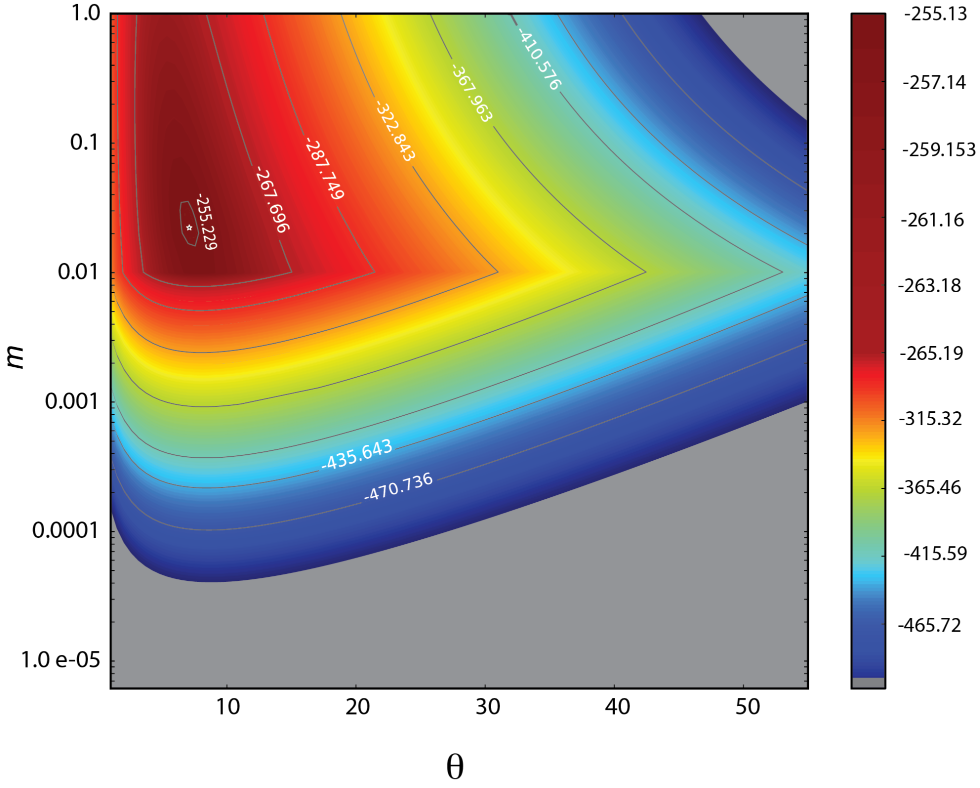


Figure 2.2.: Maximum likelihood inference of neutral parameters.
Log likelihood surface as a function of the migration rate (m), and the fundamental biodiversity number (θ) for *D. rerio* chromosome 19. Dark red color shows regions of the surface where parameters maximize the probability of explaining the abundance and diversity of observed genetic elements in the chromosome. Likelihood-ratio tests favored Etienne's sampling formula over that of Ewens to explain the observed data in the chromosome.

Due to the extent of the dataset in this study, the results presented here are generated from the comparison of Shannon's entropies.

From the neutral parameters obtained for each chromosome, we simulated 10,000 replicates and computed, for each, their Shannon's entropy (H). Chromosomes were considered significantly non-neutral when the H of their abundances was below 95% of the 10,000 simulated H values. Figure 2.3 shows the distribution of H for 10,000 simulations under Etienne's model with S, J fixed for the observed numbers and θ and m corresponding to optimized values for two chromosomes.

Additionally, given the large number of test performed (one for each of the 548 chromosomes), we corrected statistical significances by the false discovery rate (FDR) method [Benjamini *et al.* 2001]. In Figure 2.3B, chromosome 2L of

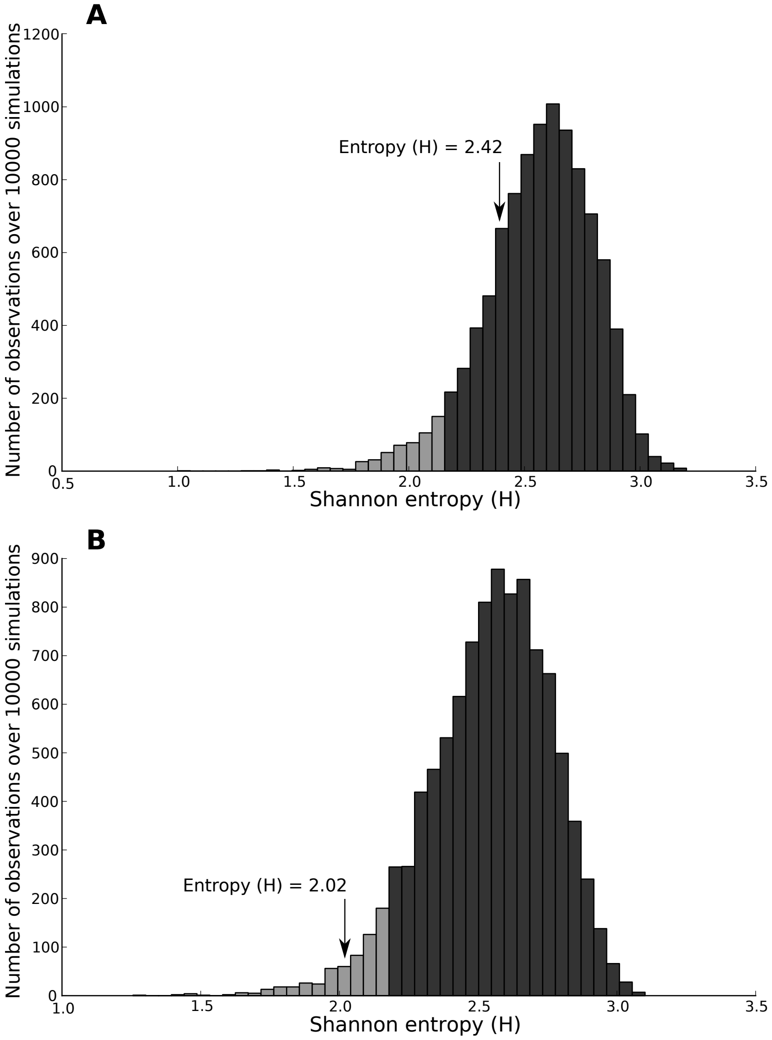


Figure 2.3.: Comparing simulated and empirical evenness.
The neutrality test compares simulated null distribution of H with the empirical value. The null distribution of H values corresponds to 10,000 neutral simulations of (A) *H. sapiens* chromosome 1 and (B) *A. gambiae* chromosome 2L, with parameters (θ and m) optimized according to Etienne's model. 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).

Anopheles gambiae's is deemed neutral only after correction by FDR.

Given the lack of differences among the results presented in section: **Neutrality of species abundances and diversity** (page 72), we replicated this test and fixed the number of species (S) to the observed values in chromosomes. No differences were observed in relation to the number of chromosomes that fitted neutral models.

Power and specificity of the neutral test

In order to validate the test of neutrality, we computed the proportion of false and true positives by 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 2.4 panel **A** and **B**, respectively.

Accordingly, we can validate the the results given that, in the entire range of S and J values, the proportion of true positives was very high (Figure 2.4B. Nevertheless these simulations highlighted some difficulties in differentiating log-normal distributions from neutral distributions. Specifically, when $J < 100,000$ individuals, the proportion of false positives is higher than 50%. However, we decided to overlook this as the increase in the false positive rate only affects the smallest chromosomes, and also because the log-normal distributions used here as alternatives, are known to be barely distinguishable from neutral distributions [McGill *et al.* 2006].

2.3. Selective pressure at molecular level

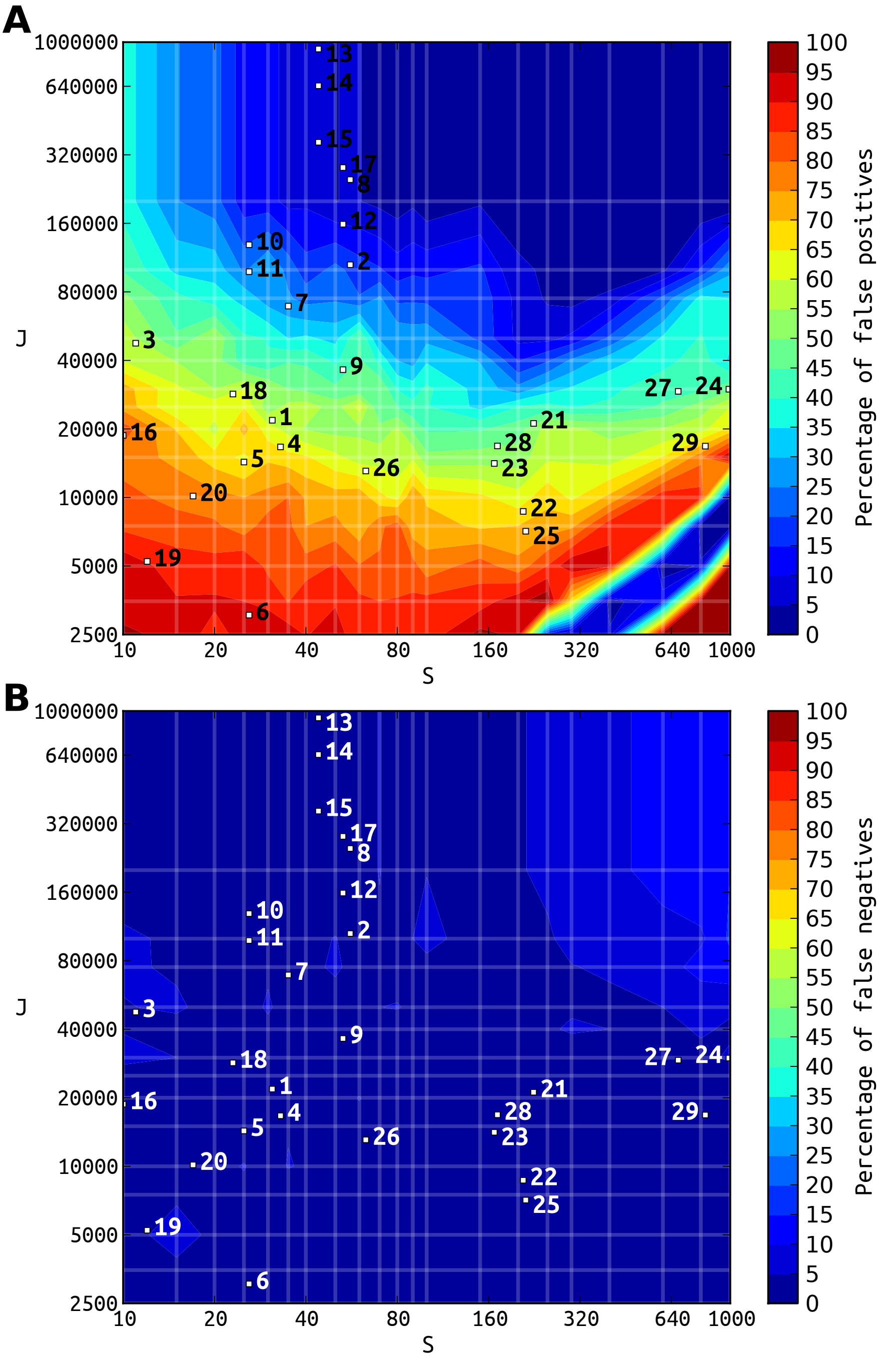
2.3.1. Orthology prediction

The complete genomes of five mammal species (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus* and *Canis familiaris*) were retrieved from Ensembl [Flicek *et al.* 2011]. Orthology predictions between seed genes and genes from other species (the seed species was *H. sapiens* in the case of mammals) were retrieved from Ensembl Compara [Vilella *et al.* 2009] using Biomart [Kinsella *et al.* 2011] (see Figure 2.5 to have an insight of the phylogenies and distances). Of the 23,438 seed genes, only groups of orthologs annotated as “one-to-one”, i.e. with only one representative of each species, were kept in the final dataset.

The same procedure was applied in six species of *Drosophila*; namely *D. melanogaster* as seed-species, *D. sechellia*, *D. simulans*, *D. yakuba* and *D. erecta* with *D. ananassae* as the outgroup (see Figure 2.5-B). Here, the starting number of seed genes was 14,076.

2.3.2. Alignments, refinement and filters

DNA coding sequences (CDS) were aligned according to the protein translation pattern using Muscle version 3.7 [Edgar 2004], embedded into the CDS-Protal


A
B
J
S
1000000
640000
320000
160000
80000
40000
20000
10000
5000
2500
10
20
40
80
160
320
640
1000
Percentage of false positives
Percentage of false negatives
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

utility in Phylemon 2.0 [Sánchez *et al.* 2011] (see appendix section: **The alignment** at page 133). Poorly aligned regions were removed using TrimAl [Capella-Gutiérrez *et al.* 2009] keeping all sequences but checking the quality of 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 were those of the human interferon γ ($dN = 3.06$) and the relaxin protein [Graur & Li 2000] ($dS = 6.39$ substitutions per site per $1e^{+09}$ years). Assuming the human-mouse, mouse-rat and human-chimp speciation times to be about 80, 70 and 5 million years [Blair Hedges & Kumar 2003] respectively, orthologous comparisons between primates and rodents with $dS \geq 1$ and $dN \geq 0.5$, between rodents with $dS \geq 0.256$ and $dN \geq 0.122$, and between primates with $dS \geq 0.064$ and $dN \geq 0.030$ substitutions per site were excluded.

In *Drosophila*, genes were also filtered by high dN and dS values using the fast evolving gene *1G5* as a relaxed reference for both dN and dS [Schmid & Tautz 1997].

The final number of orthologs kept was 12,453 for mammals and 9,240 for *Drosophila*.

2.3.3. Estimating evolutionary rates in protein-coding genes

Maximum likelihood estimations of dN, dS, and ω (the ratio of dN/dS) and tests of positive selection were computed using the CodeML program from the PAML package [Yang 2007] through the ETE program [Huerta-Cepas *et al.* 2010] (see section: **The "Evol" extension**, page 96 for a description of this specific methodology). Evolutionary rates were computed in orthologous sequences according to the free-ratio branch model that assumes independent ω ratio for each branch of the mammal and *Drosophila* species trees. Evolutionary rates (dN,

Figure 2.4. *(preceding page)*: Type I and Type II errors of the neutral test in ranges of S and J.

Panels describe the proportion of times the test (**A**) rejected the null hypothesis when it was true (red regions indicating areas prone to Type I error), and (**B**) failed to reject the null hypothesis when it was false (light blue regions indicating areas more prone to Type II error). The numbers in both panels are chromosomes: **1-** *A. thaliana* chr1; **2-** *D.rerio* chr1; **3-** *D.discoideum* chr2; **4-** *C.elegans* chrI; **5-** *D.melanogaster* chr2L; **6-** *G.gallus* chr8; **7-** *G.gallus* chr2; **8-** *H.sapiens* chr1; **9-** *H.sapiens* chr1; **10-** *Z.mays* chr1; **11-** *Z.mays* chr3; **12-** *M.musculus* chr0; **13-** *M.domesticus* chr1; **14-** *M.domesticus* chr3; **15-** *M.domesticus* chr5; **16-** *P.falciparum* chr3; **17-** *R.norvegicus* chr1; **18-** *S.bicolor* chr7; **19-** *T.nigroviridis* chr9; **20-** *T.castaneum* chr8; and ecosystems [Jabot & Chave 2011]: **21-** BCI; **22-** Edoro; **23-** La Planada; **24-** Lambir; **25-** Lenda; **26-** Mudamalai; **27-** Pasoh; **28-** Sinharaja; **29-** Yasuni.

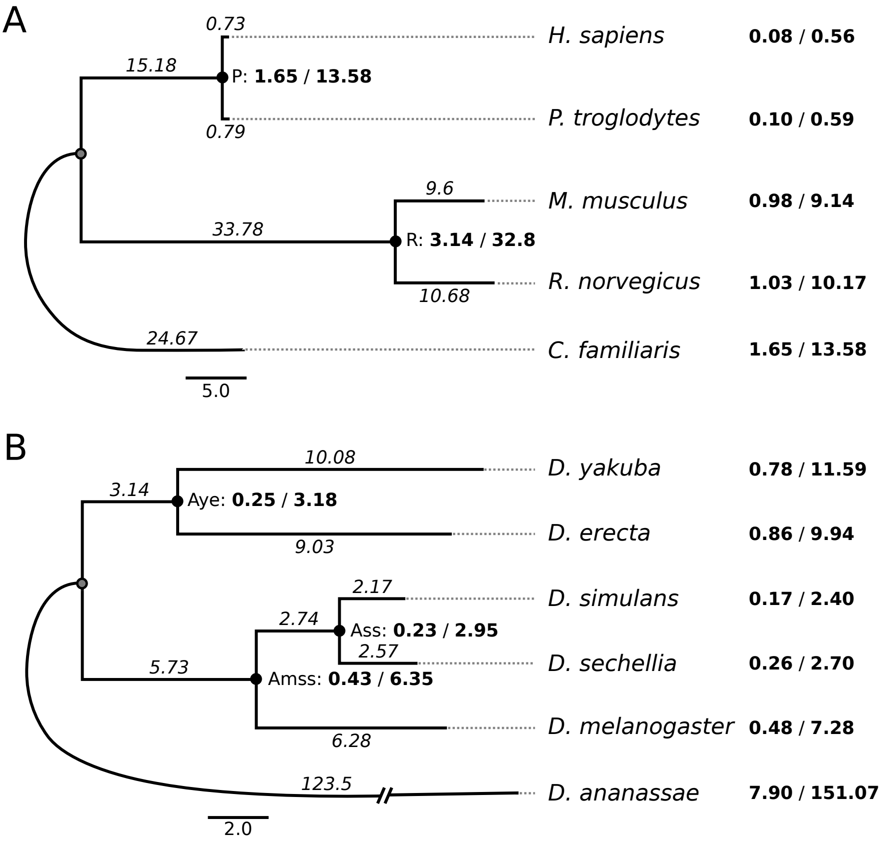


Figure 2.5.: Phylogenies of mammals and species of the *melanogaster* group of *Drosophila*.
Bold numbers represent the median rates of non-synonymous and synonymous substitutions (dN/dS) estimated from all coding sequences compared in mammals (**A**) and *Drosophila* (**B**). Numbers in italics are median branch lengths. Branch lengths and rates were multiplied by 100. Ancestral estimations were 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* in (**A**), and *D. ananassae* in (**B**), were chosen as outgroup species.

dS), their ratio (ω) and the difference between ancestral and descendant species ($\Delta\omega$) were ranked for all genes of the genomes and analyzed further by gene set selection analysis (GSSA).

External branches in Figure 2.5 were marked as foreground to test for positive and relaxed selection using branch-site models in Test I and Test II [Zhang *et al.* 2005] (see appendix section: **Testing for evolutionary scenarios in protein-coding genes**, page 137, for more complete overview of these tests). Positive results for the relaxation of selective constraints (or weak signals of positive selection) were discarded [Arbiza *et al.* 2006]. To quantify the relative contribution of positively selected genes (PSGs) in functional modules showing significantly high values of ω (SHω) and significantly low values of ω (SLω), a t-test (from R package [Ihaka & Gentleman 1996]) with the mean number of PSGs per functional modules was computed for primates, rodents, mammals and *Drosophila* species. An independent set of PSGs was collected to test the robustness of the results in mammals [Kosiol *et al.* 2008], and *Drosophila* species [Clark *et al.* 2007].

2.3.4. Gene set selection analysis (GSSA): evolutionary and statistical simulations

Gene set selection analysis works over lists of genes ranked by different evolutionary rate parameters (in this case dS, dN, ω and $\Delta\omega$). Internally, it uses the FatiScan tool [Al-Shahrour *et al.* 2007]. FatiScan is a version of gene set enrichment analysis (GSEA) [Al-Shahrour *et al.* 2005] that 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 extremities of a list of ranked genes. This is achieved by means of a segmentation test, which consists of 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 2.6) made on an ordered list of genes.

The two-tailed Fisher's exact test (FET) finds significantly over- or under- represented functional classes (GO and KEGG) when comparing the two sides of the list ranked by an evolutionary variable (in Figure 2.6, 4 of the 5 partitions show significant differences). Previous results showed that a number between 20 and 50 partitions gives optimal results in terms of sensitivity and results recovered [Al-Shahrour *et al.* 2005]. Here, we applied 30 partitions for all the GSSA performed. Given that multiple functional classes (C) are tested in multiple partitions (P), the unadjusted p-values for a total of $C \times P$ tests were corrected by FDR [Benjamini *et al.* 2001].

Originally, 1,394/1,331 GO terms, and 199/116 KEGG pathways were analyzed in mammals/*Drosophila* species, respectively. The global GO-directed acyclic graph was processed with Blast2GO [Conesa *et al.* 2005] to extend the annotation to missing parental nodes, keeping only GO terms between levels 2 and 8 for mammals, and between levels 2 and 12 for *Drosophila*. The final set of GO and KEGG terms used was also reduced to those containing at least 15 genes.

Some evolutionary rates presented discontinuous distributions, in particular the ω ratio. Partitioning a list by values can be pointless if this list scales from 0 to infinity. This is the case for ω values that, according to their distribution, would generate many empty partitions in between $100 < \omega < 900$. In order to partition more accurately we used the rank order instead of the direct value (see Table 2.5).

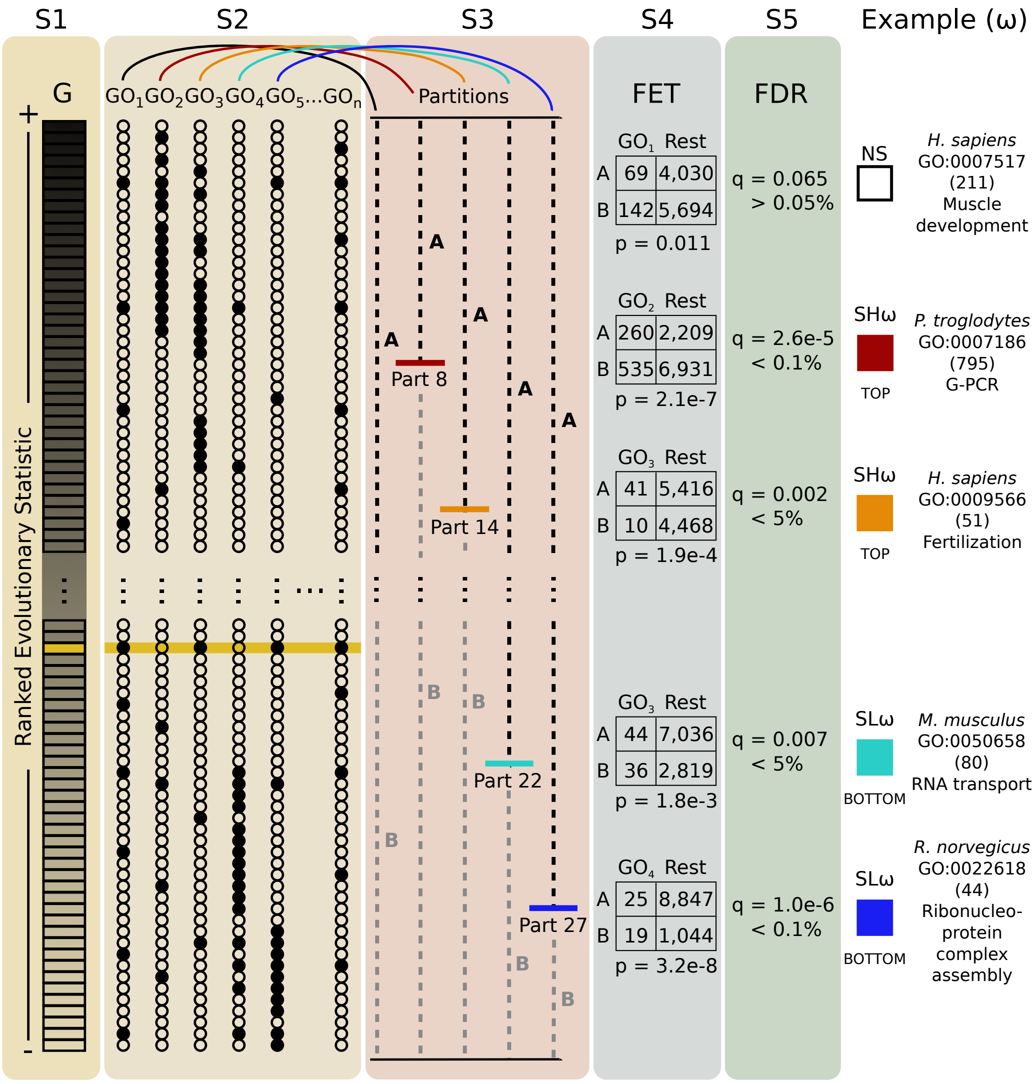
	Direct partitioning		**Partitioning by rank**	
Gene	**ω rate**	**Partition**	**rank**	**Partition**
ENSG00000211454	999.0000	A	1	A
ENSG00000169084	999.0000	A	1	A
ENSG00000159433	999.0000	A	1	A
ENSG00000162430	200.2600	B	4	B
ENSG00000176390	0.2520	C	5	B
ENSG00000156291	0.0520	C	6	C
ENSG00000176711	0.0259	C	7	C
ENSG00000166287	0.0123	C	8	C
ENSG00000174788	0.0067	C	9	C

Table 2.5.: Comparison of partitioning strategies.
A list of genes ranked by ω, divided into three partitions (A, B and C). When values are scaled from 999 to 0, "Direct partitioning" would group genes within ranges of 300 resulting in few genes with ω between 300 and 600, while"Partitioning by rank" would lead to a more continuous distribution.

Methodological validation of the GSSA

To test possible biases attributed to the size of the functional category or the magnitude of change in evolutionary rate, we randomized the values (ω, $\Delta\omega$, dN and dS) assigned to the list of genes. Functional categories should point to the same set of genes, thus conserving all structural characteristics of the data, but

Figure 2.6. *(following page)*: Summary of the steps developed by the GSSA.
GSSA can be described in a series of five steps (S1 to S5). S1: rank genes of a genome according to an evolutionary variable (e.g.: ω), S2: assign functional categories, S3: partition the ranked list, S4: proceeds with a Fisher exact test for each partition, S5: adjust p-values by FDR. Colored boxes represent the final result: functional categories found to have values of ω that are a) significantly high (SH) appear in red or orange (0.1% and 5% FDR respectively), b) significantly low (SL) in blue and cyan (0.1% and 5% FDR respectively) c) not significant (NS) in white. The example shows five GO terms with significantly and NS-biased distributions of ω. The number of genes annotated with the GO term is indicated in brackets. GO:0007517 is NS; although, in partition 16 in humans (not shown) its p-value was low, it was NS after FDR correction (q = 0.065). Upper (A) and lower (B) sides of the list (S3) represent both sides of the specified partition number. The remaining GO terms (GO_2 to GO_5) show the association of dark dots with values located at the top (SHω), and at the bottom (SLω) of the list (for GO_2-GO_3 and GO_4-GO_5, respectively). In examples, Fisher exact tests found the most significant p-value for partitions 8, 14, 22 and 27 for GO:0007186, GO:0009566, GO:0050658 and GO:0022618 in the chimpanzee, human, mouse and rat genomes, respectively.


S1
S2
S3
S4
S5
Example (ω)
G
+
-
Ranked Evolutionary Statistic
$GO_1 GO_2 GO_3 GO_4 GO_5 \ldots GO_n$
Partitions
A
B
Part 8
Part 14
Part 22
Part 27
FET
FDR
GO_1 Rest
A 69 4,030
B 142 5,694
p = 0.011
q = 0.065
> 0.05%
GO_2 Rest
A 260 2,209
B 535 6,931
p = 2.1e-7
q = 2.6e-5
< 0.1%
GO_3 Rest
A 41 5,416
B 10 4,468
p = 1.9e-4
q = 0.002
< 5%
GO_3 Rest
A 44 7,036
B 36 2,819
p = 1.8e-3
q = 0.007
< 5%
GO_4 Rest
A 25 8,847
B 19 1,044
p = 3.2e-8
q = 1.0e-6
< 0.1%
NS
H. sapiens
GO:0007517
(211)
Muscle development
SHω
TOP
P. troglodytes
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)
Ribonucleo-protein complex assembly

suppressing the biological relevance of evolutionary rates (see Figure 2.7).

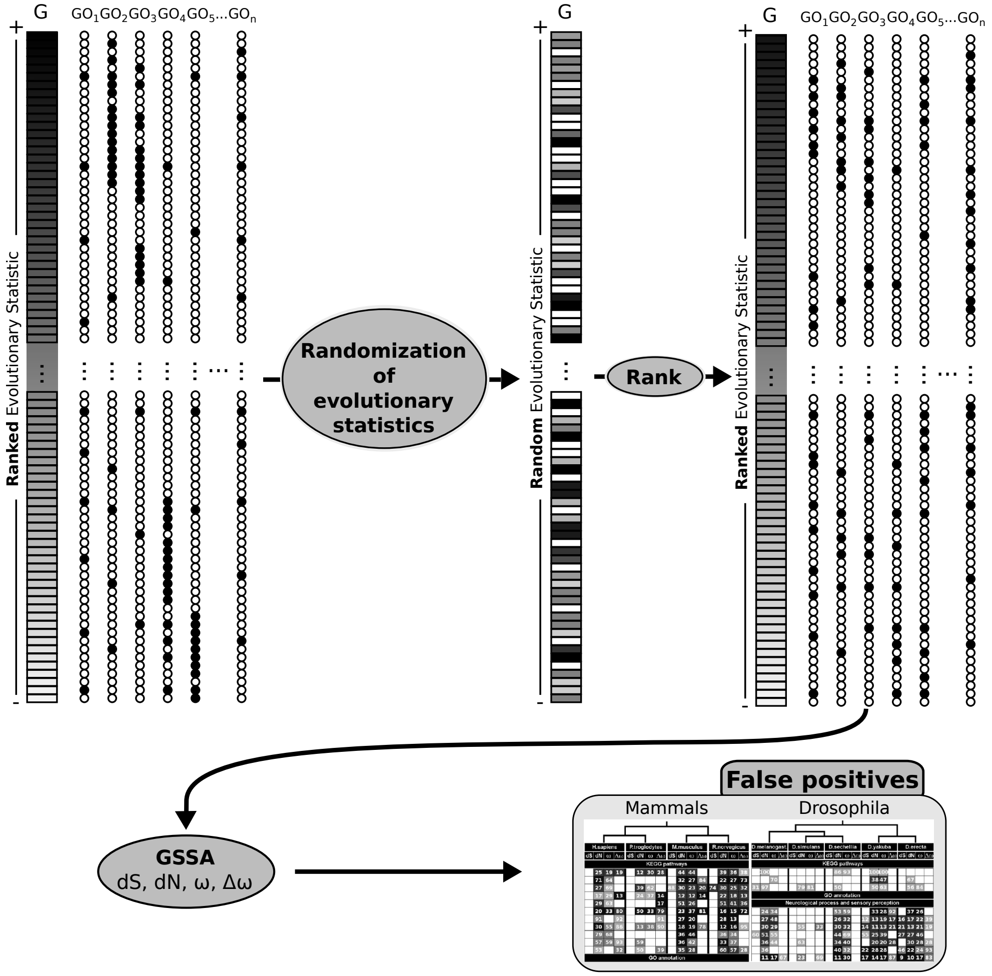
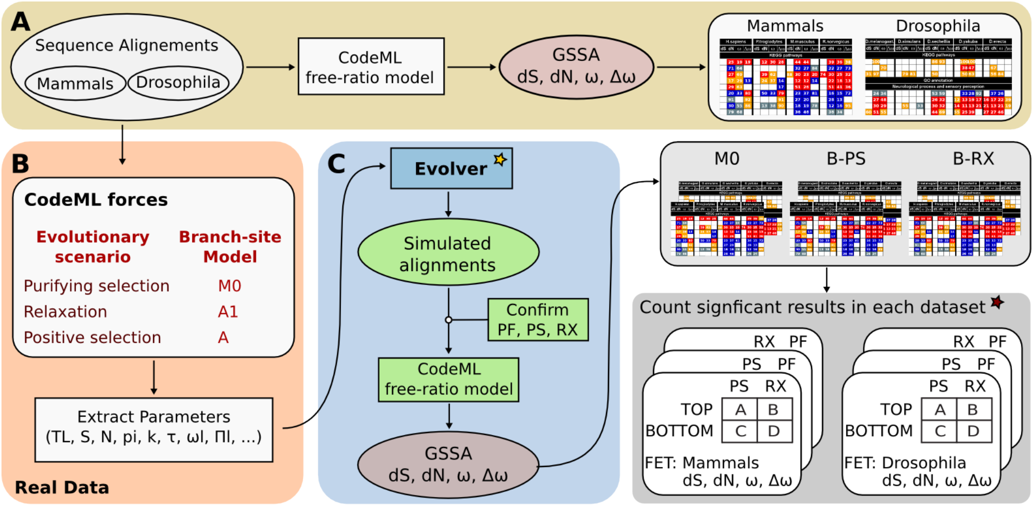

G
GO1GO2GO3GO4GO5...GOn
+
-
Ranked Evolutionary Statistic
Randomization of evolutionary statistics
Random Evolutionary Statistic
Rank
GSSA
dS, dN, ω, Δω
False positives
Mammals
Drosophila

Figure 2.7.: Randomization experiment diagram.
The diagram shows the steps followed to test possible biases attributable to the size of functional categories. Genes are randomly re-arranged according to their evolutionary statistics. Finally, genes are ranked according to their new randomly-assigned values. The results of the enrichment analysis over this dataset are considered false positives.

This methodology facilitates testing of the effect of functional category sizes, and to ensure that the distribution of evolutionary rates does not affect the experiment. Each evolutionary variable was randomized 10,000 times for each species. The proportions of false positives (GSSA significant results) were plotted against the size of functional categories (from 0 to 1,500 with intervals of 20). As these proportions never attained values higher than 0.05% FDR, we rejected the possibility that either, group sizes or rate distributions, biased the GSSA results in the dataset (see Figure 2.8).

In order to better understand the results of the GSSA, a final experiment was necessary since, at this point, the possibility that results with significantly high ω were brought about only by genes under positive or relaxed selection could not be discounted.

Thus, to validate the independence of the GSSA from the effects of alternative evolutionary constraints, we simulated different selective regimes (purifying, positive and relaxed selection) using branch-site models. Here, we addressed the possibility of a variation in the representation of significant results after GSSA. The protocol diagram described in Figure 2.9 shows three different areas:



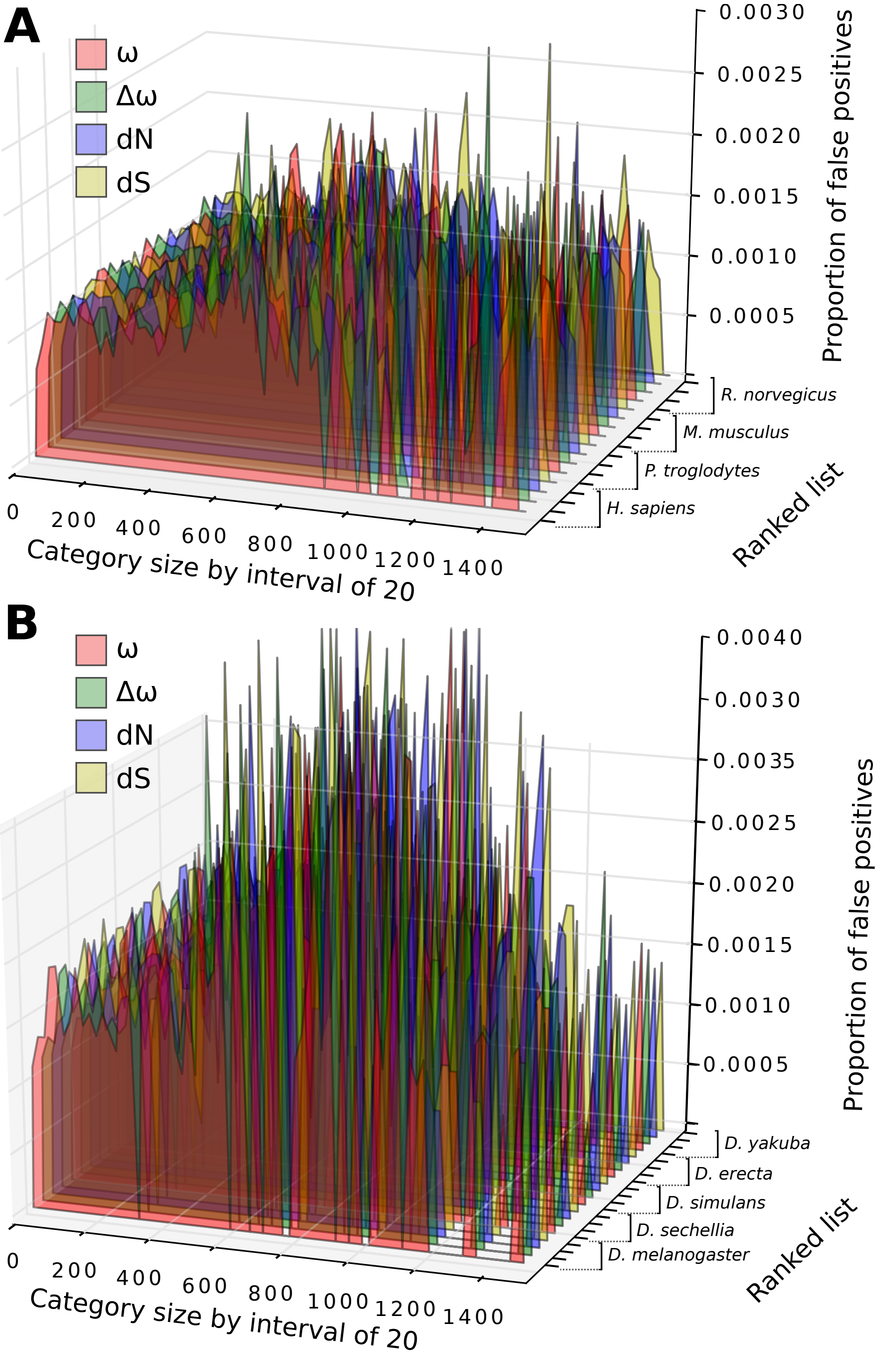
☆ Human, mouse, *D. melanogaster* and *D. erecta* used as foreground species in branch and branch-site models.

★ Count done over foreground species.

Figure 2.9.: Evolutionary and statistical simulation of GSSA.
The protocol diagram shows the steps taken along three different analytical spaces, the real data, the simulated data and the testing block. See main text for a complete explanation.

- **Real Data**: the light yellow area (**A**) describes the steps of the GSSA. The orange area (**B**) describes the use of the CodeML program from PAML package [Yang 2007] to extract from the original set of sequences all evolutionary parameters needed to simulate new sequences under purifying selection (PF), positive selection (PS) or relaxation of selective constraints (RX) according to branch-site models. Human, mouse, *D. erecta* and *D. melanogaster* were used as foreground species in the corresponding models.

- **Simulated Data**: in the light blue area (**C**), the Evolver program (also from the PAML package [Yang 2007]) simulates sequences evolving under


A
ω
Δω
dN
dS
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
Proportion of false positives
R. norvegicus
M. musculus
P. troglodytes
H. sapiens
Ranked list
0
200
400
600
800
1000
1200
1400
Category size by interval of 20
B
ω
Δω
dN
dS
0.0040
0.0030
0.0035
0.0025
0.0020
0.0015
0.0010
0.0005
Proportion of false positives
D. yakuba
D. erecta
D. simulans
D. sechellia
D. melanogaster
Ranked list
0
200
400
600
800
1000
1200
1400
Category size by interval of 20

the given parameters (codon frequencies and branch lengths) estimated from the empirical data. We checked the desired characteristics of PS and RX on the set of the simulated sequences Table 2.6. The evolutionary variables (dS, dN, ω and $\Delta\omega$) were estimated from simulated sequences using a free-ratio branch model. The complete GSSA protocol was applied over the simulated data.

- **Testing simulation**: the last part of the diagram represents the calculation of the odd-ratios corresponding to a classification of the GSSA results over all datasets. Significant categories are counted for the contingency tables, with either SHω or SLω. and belonging to two of the three simulated selective regimes (PS, RX and PF). Odd-ratio values represent the association between different selective regimes, simulated according to their proportions of SH and SL functional categories. Statistical contributions of the simulated regimes (PS, RX and PF) to the GSSA results were tested by comparing log odd-ratios with a t-test (results in Table 2.7).

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

Table 2.6.: Number of genes under positive and relaxed selection in each of the simulated evolutionary scenarios.

The results showed that, in spite of the alternative evolutionary scenarios, no significant differences were found between log odd-ratio distributions ($p<0.05$). The average effect of PF and RX/PS is a proportional decrease and increase of the mean ω value on sequences, respectively. This change has minor effects (if any) on the relative position of genes in the ranked list of genes of the genomes. Accordingly, since no net differences were produced after ranking genes, no signif-

Figure 2.8. *(preceding page)*: Randomization experiment results.
These graphics show the proportion of false positives within the results of an enrichment analysis conducted over lists of genes ranked by a shuffled evolutionary variable (see main text for details). Results are segregated into ***1)*** ranges for the number of genes belonging to a functional category in order to discard the effect of category size on the proportion of false positives, and ***2)*** evolutionary variables (red, green, blue and yellow for ω, $\Delta\omega$, dN and dS, respectively) and species in order to discard biases due to the specific distribution of one of the variables in a given species. Randomizations were conducted in mammals (**A**) and *Drosophila* (**B**).

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

Table 2.7.: Proportion of significant functional categories coinciding for two simulated evolutionary scenario, or retaining identical signs of odd-ratios under a different evolutionary scenario.

icant 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 2.7. We conclude that none of the simulated selective regimes produce significant differences or biases in the GSSA of ω values.

3. Random-like structure of DNA

3.1. Computing genome complexity

Of the 20 major systematic groups we picked 54 species and computed the complexity value (CV) of their genomes with sizes ranging from 1.6Kb to 3.4Gb 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. This first result implies maximum complexity for all genomes. The residual variation around the fitted regression and along the 6 orders of magnitude, was almost zero ($adjusted - R^2 = 0.987$).

The slope and degree of adjustment for the set of species that shows are quite surprising, given the diversity of living forms used in this analysis. The chosen array of organisms range from the shortest single-stranded RNA genome of the *Hepatitis D* virus (size $\sim 1.69e^{+03}$ bp) to the largest double-stranded DNA genome of the short-tailed opossum *Monodelphis domestica* (size $\sim 3.41e^{+09}$ bp), and even included selected organisms having peculiar genomes, such as:

- obligate endosymbiont bacteria with extremely reduced genome sizes (*Carsonella ruddii*, *Buchnera aphidicola*, and *Ureaplasma urealyticum*) [Wernegreen 2002].
- parthenogenetic crustaceans with ubiquitous gene duplications (*Daphnia pulex*).
- archean organisms living in extreme environmental conditions (*Sulfolobus islandicus*, *Methanocaldococcus vulcanius*, *Thermococcus sibiricus*).
- the first synthetic organism (*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 were:

- six ancient or recent polyploid species.

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	Methanocaldococcus 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
	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

Feat.	Species ^{clade*}	ACN-EV	GS	GC	GCR	Dmax
	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

Table 3.1.: Genome complexity.

Genomes size (GS), genome 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 numbers or Ensembl (*E!*) version (ACN-EV) are given. ***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;

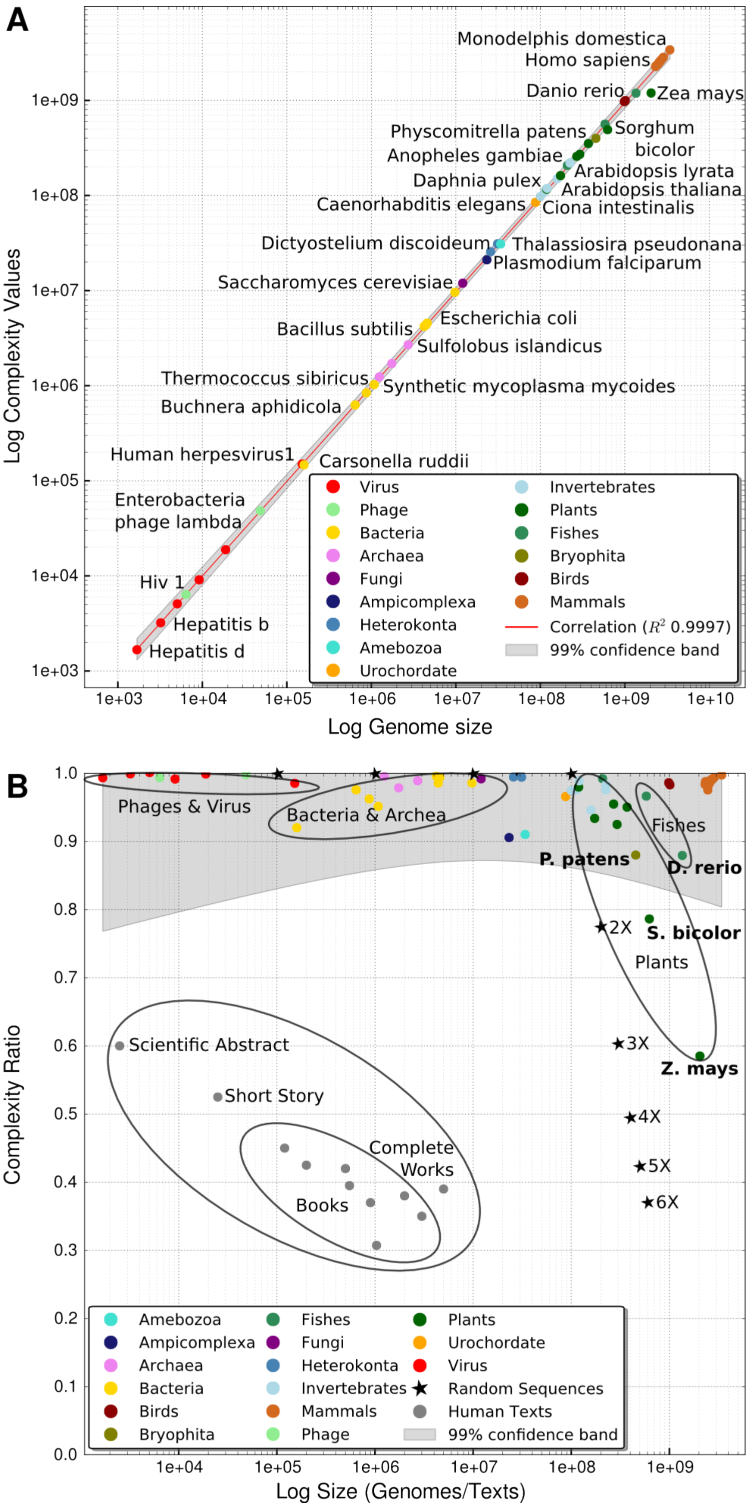
- the most extreme case of genome reduction in bacteria, *Carsonella ruddii.*
- the explosive case of gene expansion in Daphnia [Colbourne *et al.* 2011].
- two unicellular eukaryotes that, curiously, correspond to the two genomes sequenced with the highest proportions of A + T; *Plasmodium* [Gardner *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 a uniform distribution of A, C, G and T. To simulate events of polyploidization random sequences were duplicated up to five times (corresponding to a 10×). The assiocated decrease in CR, reaching CR=0.25 for 10×, could then be compared to the CR values of real polyploids, which placed maize genomes at the level of a perfect triploid and sorghum at the level of a perfect diploid.

In addition to genomes and random sequences, the results of a computation of CR for human texts were added to Figure 3.1-B (values used can be found in Table 3.2).

The complexity ratios of complete genomes, random sequences of different ploidy and human language texts were computed together. Maximum CR corresponds to random sequences of lengths ranging from 5 Kb to 2.5 Gb. In the case of biological sequences, non-polyploid genomes showed CR > 0.90. Conversely, polyploids showed CR values below 0.95, with the lowest ratio for *Z. mays* (CR=0.58), and the second lowest ratio for its closest relative *S. bicolor* (CR=0.78). Overall, for the non-random strings analyzed, the lowest CR was obtained for human language texts. The CR of 11 human texts of different sizes and languages, from short scientific abstract to the complete works of William Shakespeare, are also depicted in Figure 3.1-B. The CR diminishes as text size increases, due to the limited lexicon and the fixed language grammar. Complexity was lowest in Darwin's *Origin of Species* (≈ 0.309), which is comparable to the CR of a random polyploid

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. Some relevant species are labeled (for the complete list see Table 3.1). **(B)** Most genomes have a complexity ratio (CR) between 0.90 and 1.0. Four polyploid species have CR < 0.9 (shown in bold in the figure): P. patens (0.880), *D. rerio* (0.879), *S. bicolor* (0.786) and *Z. mays* (0.585). The stars with CR = 1 correspond to random [A, C, G, T] strings of 30, 50, 100, 250 and 500 Mb length, respectively. Others stars with lower CR values correspond to the 500 Mb random string repeated from 2× to 6× to simulate perfect polyploids. Changes in sequence length due to polyploidy produce no change in the CR. Notice the low CR of human texts (see Table 3.2). Confidence bands in both plots correspond to the 99% certainty of containing the best regression line.


A
Log Complexity Values
Log Genome size
1e+03
1e+04
1e+05
1e+06
1e+07
1e+08
1e+09
1e+10
Monodelphis domestica
Homo sapiens
Danio rerio
Zea mays
Physcomitrella patens
Sorghum bicolor
Anopheles gambiae
Arabidopsis lyrata
Daphnia pulex
Arabidopsis thaliana
Caenorhabditis elegans
Ciona intestinalis
Dictyostelium discoideum
Thalassiosira pseudonana
Plasmodium falciparum
Saccharomyces cerevisiae
Escherichia coli
Bacillus subtilis
Sulfolobus islandicus
Thermococcus sibiricus
Synthetic mycoplasma mycoides
Buchnera aphidicola
Human herpesvirus1
Carsonella ruddii
Enterobacteria phage lambda
Hiv 1
Hepatitis b
Hepatitis d
Virus
Phage
Bacteria
Archaea
Fungi
Ampicomplexa
Heterokonta
Amebozoa
Urochordate
Invertebrates
Plants
Fishes
Bryophita
Birds
Mammals
Correlation (R^2 0.9997)
99% confidence band
B
Complexity Ratio
Log Size (Genomes/Texts)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1e+04
1e+05
1e+06
1e+07
1e+08
1e+09
Phages & Virus
Bacteria & Archea
Fishes
P. patens
D. rerio
★2X
S. bicolor
Plants
★3X
Z. mays
★4X
★5X
★6X
Scientific Abstract
Short Story
Complete Works
Books
Amebozoa
Ampicomplexa
Archaea
Bacteria
Birds
Bryophita
Fishes
Fungi
Heterokonta
Invertebrates
Mammals
Phage
Plants
Urochordate
Virus
Random Sequences
Human Texts
99% confidence band

Type	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 text complexity

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

sequence of 7×. It is noteworthy that the sizes of the human texts analyzed are within the range of phage, virus and bacterial genome sizes. The complexities of human texts are detailed in Table 3.2.

3.1.1. Genome complexity and ploidy level

Analysis of CR Figure 1.2-B reveals a clear segregation of species' genomes by their level of ploidy, within the most recent polyploids like maize and sorghum exhibiting the lowest CR values. However this trend seems to be quickly lost, as ancient polyploids are hardly distinguishable from non-polyploids. A very illustrative example can be found within the *Arabidopsis* genus, where the two close relatives *A. thaliana* and *A. lyrata* (which diverged 10 My ago [Hu *et al.* 2011]) seem to have followed different evolutionary routes after their whole genome duplication (< 70 My ago [Proost *et al.* 2011]). While *A. thaliana* suffered a drastic genome reduction after polyploidization (mainly due to hundreds of thousands of small deletions), its relative, *A. lyrata*, remained complete [Hu *et al.* 2011]. This is reflected in the difference in CR between these two species CR=0.9339 for *A. lyrata* and CR=0.9799 for *A. thaliana*, the later undergoing a faster incremental increase in CR.

In order to confirm the trend observed for levels of polyploidy and CR values, we tested the hypothesis that the observed genome complexity values correlated with size and ploidy level. A categorical variable divided polyploid (ancient or recent), and non-polyploid species, as described in Table 3.1. The size-interaction variable provided significant deviations ($p < 2e^{-16}$, adjusted-$R^2 = 0.997$), while independent linear models slopes were 0.633 ($p < 4.8^{e-07}$, adjusted-$R^2 = 0.921$), and 0.988 ($p < 2e^{-16}$, adjusted-$R^2 = 1.00$) for polyploid and non-polyploid genomes, respectively.

3.2. Chromosome complexity

Following the same methodology used for genomes, we computed the CR values of individual chromosomes (567 autosomes of 31 species). The CV values obtained were normalized by chromosome size, thereby providing the CR values seen in Figure 3.2. As previously, the statistics were very convincing. The slope of the relationship between chromosome size and CV was around 0.924, and could be increased to 0.951 if polyploid species were excluded (alone, polyploid species exhibited a low slope = 0.696). Again, the significance of the size-interaction variable was indisputable ($p < 2e^{-16}$).

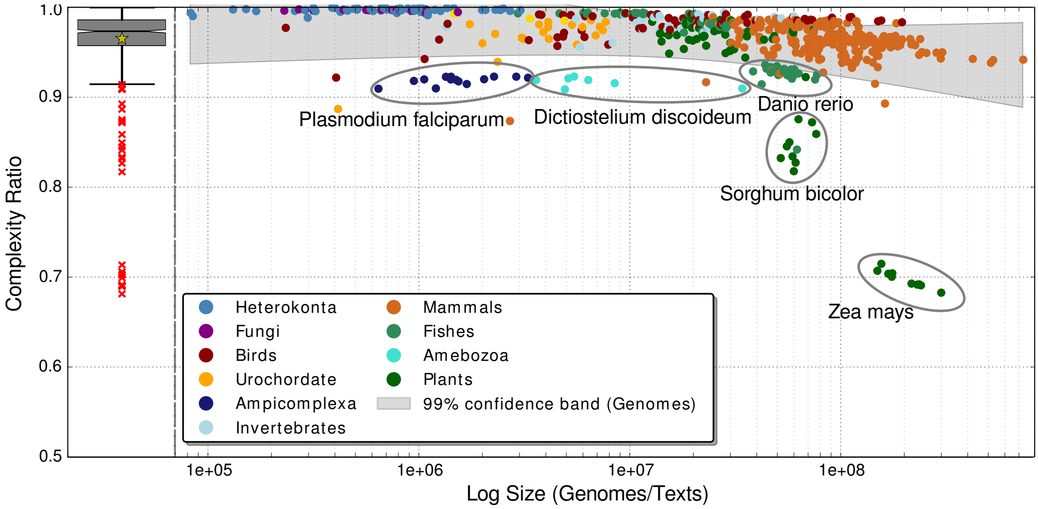


Figure 3.2.: Chromosome complexity ratio.
Most chromosomes (96.2%) have complexity ratios ranging from 0.9 to 1.0, as observed for complete genomes Figure 3.1B. The boxplot on the left shows the distribution of CR values for all chromosomes (outliers are shown in red, and the yellow star corresponds to the mean value). Notice how deviant the chromosomes of *Z. mays*, and to a lesser extent *S. bicolor* (both recent polyploid species), are from the general trend.

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

The boxplot inside Figure 3.2 summarizes the distribution of chromosomal CR values. The first quartile of the full sample set indicates that 75% of the values are above 0.958, while the median and mean are 0.974 and 0.964, respectively. The minimum CR value corresponds to maize chromosome 10 (0.683), and the maximum CR values is for *P. tricornutum* chromosome 28 (0.999). Opossum chromosome 1 (the largest chromosome) has a CR value of 0.942. Mean CR for the set of maize chromosomes was 0.698, while the overall CR value for the maize genome was 0.585. This difference suggests extensive duplicated regions in maize chromosomes, which have been previously described [Weber & Helentjaris 1989, Gaut 2001] and attributed to a tetraploid event that occurred when maize evolved 11 My ago [Gaut & Doebley 1997, Wolfe 2001].

Although, in general, comparisons of overall genome CR and mean chromosomal CR values for species were concordant, some differences were notable, namely (overall genomic CR – mean chromosomal CR): *S. bicolor* (0.854 – 0.786), *D. rerio* (0.924 – 0.879), *A. lyrata* (0.966 – 0.934), *P. trichocarpa* (0.971 – 0.950), *S. cerevisiae* (0.996 – 0.992), *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).

A smaller genomic CR value (relative to the mean chromosomal value) generally occurs in polyploid species and can be explained in a broader sense in terms of "repetitive elements" and as a consequence of considering a wider window of analysis (see next subsection 3.4.1). In contrast, the reasons for a higher genomic CR value are more complex and are discussed further in the subsection 3.4.2 on polyploids and their return to maximum complexity.

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

Eukaryote genome structure is generally characterized by the extensive presence of non-functional repetitive elements (REs) spread out throughout the genome, and a tiny portion of singular functional elements covering the rest. To get insights into the statistical structure of these contrasting genomic regions, 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 was achieved, for each family, by concatenating all units in their original order into chromosomes.

Genes characterized by especially high information content, as expected, showed the highest CR values among all classes analyzed, independently of the species. Indeed, the typical structure of genic regions is important for entropy-based algorithms that predicts or confirm automatic gene detection [Du *et al.* 2006, Gerstein *et al.* 2007]. However, here it is important to remember that the principle methodology used considered all genes together, instead of using the typical sliding window. Deeper analysis showed that when genes were split into their two main components, exons possessed higher CR values than introns.

For repetitive elements, we hypothesized that CR values would be lower due to the limited complexity of these elements. This was indeed the result for SINE and satellites. However, for LINE, LTR and DNA-T (Table 3.3), unexpectedly high values were observed. This result can be explained by the greater length of these elements and their high internal variability among families.

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

Table 3.3.: Mean complexity ratio of some genetic components.

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

In order to test the hypothesis that the order in which RE were placed in the chromosome influenced the results, 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 values for elements in their natural order was much lower in SINEs and satellites than in the remaining classes, suggesting a structure of identical or very similar repeats along neighboring chromosome segments. This pattern did not show up in the other families of RE. The notable exception was LTRs of the maize chromosome, which is known to have expanded dramatically in recent evolutionary times [Blanc & Wolfe 2004]. All shuffled classes (including SINEs and satellites) had a CR value very close to one. This result indicates that genomes have extensive genetic variation, even in regions where the expected pattern is the homogeneous repetition of almost indistinguishable units of RE's.

		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

Table 3.4.: Complexity ratio of concatenated and shuffled genomic classes. Size and complexity ratio of different genome classes in natural order (NAT), and shuffled (SHU) for selected chromosomes. SIZE represents the length in Mb of the concatenated elements.

3.4. Methodological validation and interpretation – Understanding the CR

3.4.1. Complexity in chromosome segments

In order to fully understand how CR ratios work, and how applicable they are for working with full genomes or chromosomes, we decided applied the algorithm to windows of different sizes. Chromosomes were. therefore, split into overlapping windows of various sizes (from 1 Kb to 100 Mb) and the CR values in each of these windows was computed. Figure 3.3 shows boxplots of six selected chromosomes, at different scales, all having extreme CR values.

As an initial trend, we observed that the median values of CR, over all windows, for *A. thaliana* Chr1, *C. elegans* ChrI, *D. melanogaster* Chr2L and *H. sapiens* Chr1 were all above 0.97 (Figure 3.3). However. lower values were obtained for *Z. mays* Chr1 and in *H sapiens* Chr19. 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

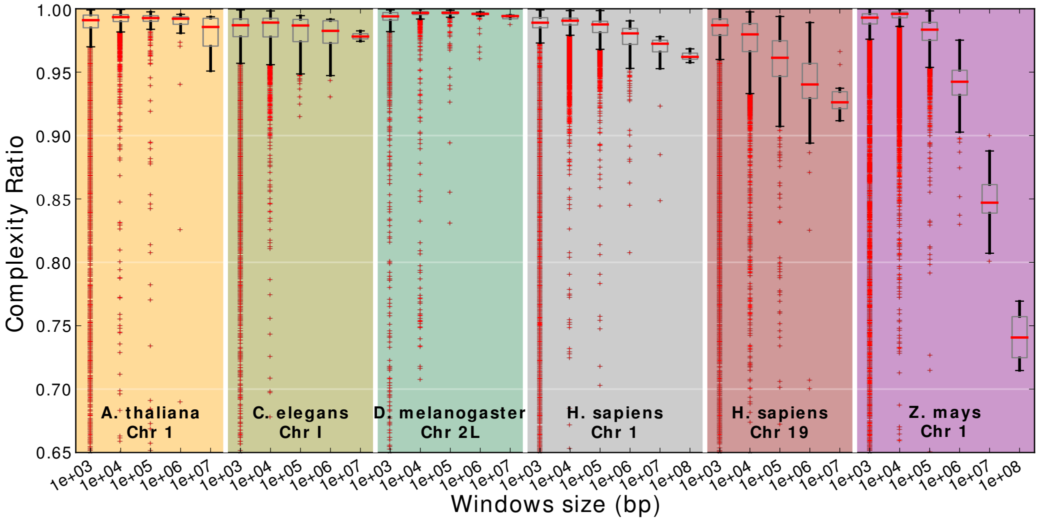


Figure 3.3.: Sliding window analysis of chromosomes.
Boxplots show results of sliding window analysis in six selected chromosomes. Most chromosomes have median CR values > 0.975, independently of window size.

to contain the highest number of Alu sequences over human chromosomes [Venter *et al.* 2001].

Among all chromosomes presented here (Figure 3.3), 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 CR value was prevalent. This pattern is explained by the existence of repeats, which can only be detected when the window size is large enough.

A final example of how window size affects the detection of regions with low entropy is shown in Figure 3.4. This figure represents the entropy-shape of *D. melanogaster* chromosome 2L for two window sizes. In the case of small windows (1Kb), the variable pattern of CR values across the chromosome is not really interpretable; we can only assume that each decrease in CR corresponds to a region with a high number of small repetitions of one or two nucleotides. In contrast, when the window size is 100Kb, the shape of the CR curve is much smoother, revealing only one notable peak of low CR. Interestingly, this peak corresponds to a histone cluster of more than 100 genes of the same family. The plot at the top of the figure shows the Ensembl gene annotation for this chromosome with the location of the cluster of histone genes also highlighted.

These examples reveal how the measure of entropy is affected by the size of the sequence measured. The most striking example that shows the importance of selecting large windows in order to include a maximum amount of information is the case of maize chromosome 1, where the mean CR values demonstrate a

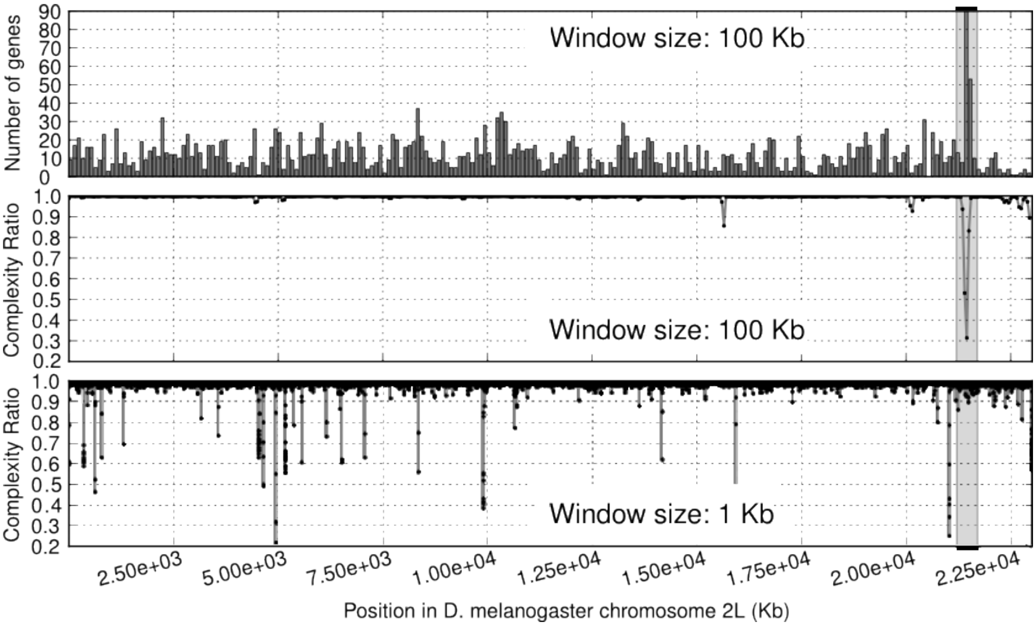


Figure 3.4.: Sliding window analysis for an entire chromosome.
All three plots correspond to a sliding window analysis of the complete *D. melanogaster* chromosome 2L. The top graph represents a gene count in a window of 1e^{+05}Kb (data retrieved from Ensembl [Flicek *et al.* 2011]). The middle and bottom graphs represent the complexity ratio displayed at window sizes of 100Kb and 1Kb, respectively. Ensembl annotation of the histone genes cluster with more than 100 histone genes is highlighted in gray in the three plots.

pronounced reduction when the size of the analysis window increased.

3.4.2. Polyploidy and return to maximum complexity

Evolution erodes the ancient footprints of genome polyploidy and diploidization events over time [Wolfe 2001]. As shown in previous sections, the CR of recent polyploids is much lower than that of non-polyploids, or of ancient polyploid species. The "Erosion" of polyploid genomes 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 more dramatic mechanisms also participate in the loss of polyploid footprints, such as massive deletion and transpositions of genetic material, as reported for *A. thaliana* [Hu *et al.* 2011]. We tested the hypothesis that the complexity ratio of polyploid genomes increases with their "maturity".

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

In all cases, sequences undergoing random mutation (Figure 3.5-A) or translocations (Figure 3.5-B) reached maximum CR=1 after a sufficient number of generations. A general observation is that the lower the CR value the more sensitive the sequence is to changes (either through mutations or translocations). This is exactly as expected by probability theory, since each change (introduced by a random mutation or a translocations) in a larger dataset is more informative than in a smaller dataset, because it allows selection over a bigger range of possibilities. For the real polyploids (sorghum and maize), the dynamics of the CR increase was identical (Figure 3.5-A). Figure 3.5-B shows that genomes and chromosomes reached maximum CR=1 after many cycles of translocations. Using a simulated genome with a tetraploid structure, translocations preserved the relationship where chromosomal CR is higher than genomic CR, for all generations up to convergence at maximum CR=1. This property was previously reported for maize and sorghum (see discussion on section 3.2).

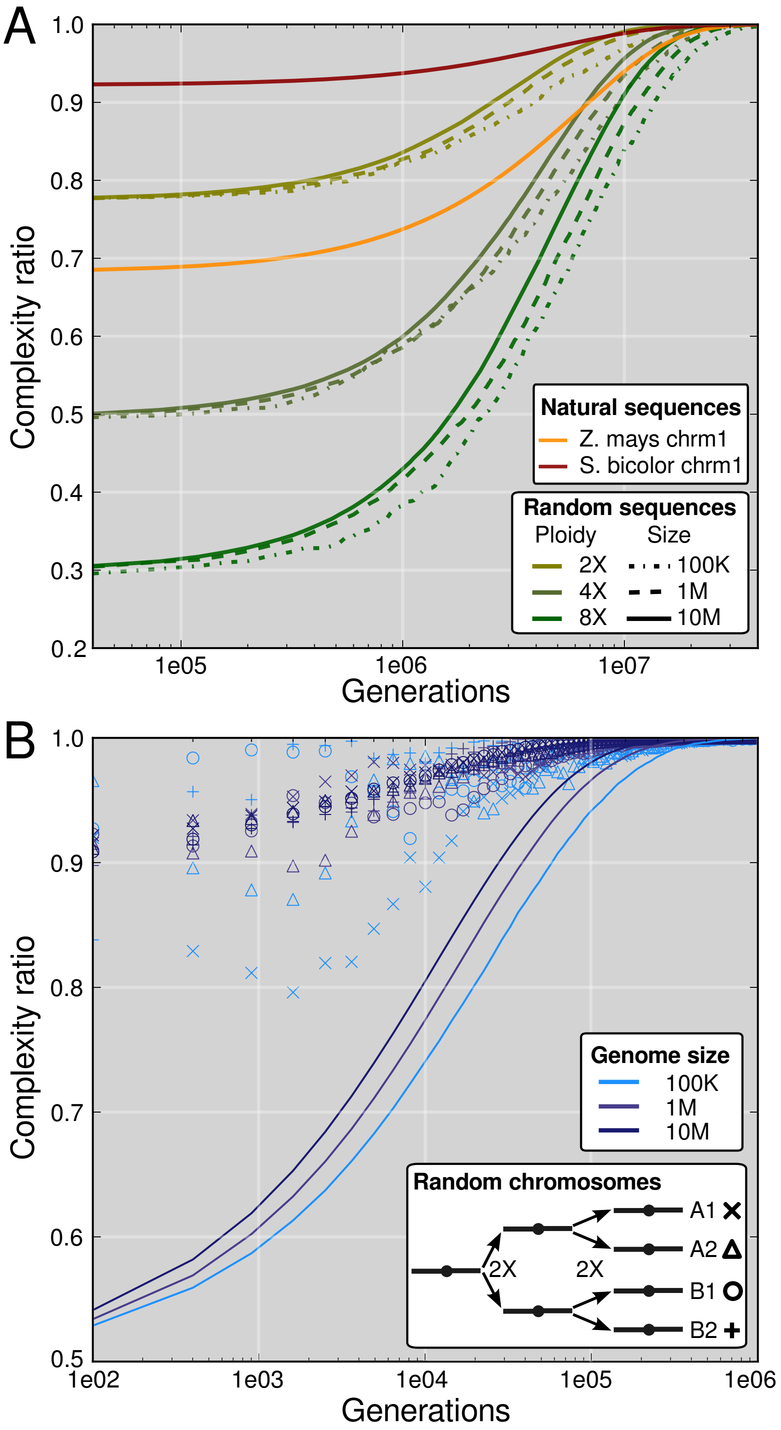
Thus, as the CR values gets closer to 1 through time, the DNA structure of polyploids become indistinguishable from diploid genomes.

3.4.3. Low CR corresponds to a simple combinatorial structure for sequences

The combinatorial structure of a sequence is a description of the observed arrangement of the symbols among all the possible permutations of the same length. Sequences with many long repeats have a low CR (see subsection 2.1.2, in Material and Methods, to get a complete picture of 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 simulated diploid and triploid genomes (Figure 3.1-B). It is also possible to achieve low CR in sequences without long repeats, but having an orderly arrangement of symbols. Although we did not observe this phenomenon for natural DNA, we constructed de Bruijn sequences [de Bruijn 1946, Becher & Heiber 2011] with low CR (see Table 2.1 in Material and Methods for examples of short sequences).

3.4.4. High CR corresponds to random-like sequences

High CR implies both high diversity and a balanced abundance of short repeats in DNA sequences. Maximum CR=1 is attained by a sequence of length n if it contains full diversity of length k, for $k \leq log_4 n$, and each of these short sequences occurs about $n \cdot 4^{-k}$ times.


A
Complexity ratio
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1e05
1e06
1e07
Generations
Natural sequences
Z. mays chrm1
S. bicolor chrm1
Random sequences
Ploidy
Size
2X
4X
8X
100K
1M
10M
B
Complexity ratio
1.0
0.9
0.8
0.7
0.6
0.5
1e02
1e03
1e04
1e05
1e06
Generations
Genome size
100K
1M
10M
Random chromosomes
2X
2X
A1 ✕
A2 △
B1 ○
B2 ✚

Intuitively, a non-random sequence will exhibit some significant regularity that can be used to compress the sequence. The mathematical concept underpinning this idea 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. Excluding some deviations, for finite sequences and particular compression methods, statistical randomness is the exact inverse of compressibility. Thus, 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 increasing sequence length. Thus, each genome is a singular instance out of the extraordinary number of combinatorial variants of the same length with the same high complexity rate.

3.5. Discussion

3.5.1. Universal informational structure of DNA

To date, no conclusive work on the statistical properties of DNA has been conducted on complete genomes. Taking a broad look at the results presented here, the most remarkable feature is certainly that, whatever genome is chosen at whatever level of magnification (genome chromosome or windows), DNA seems to be strongly selected towards maximum complexity. At the genomic and chromosomal levels, recent polyploid species were the only outliers observed. Aside from these exceptions, CR was close to 1 for the whole range of diversity of life examined, from viruses to mammals. Even the genomes expected to be simplest, like those of higher eukaryotes for which half of their genomes are composed of repetitive sequences, presented a CR>0.95. When genetic elements (GEs) were analyzed separately, only SINEs and satellites presented lower CR values, whereas all other families of GEs showed high degrees of variability.

Figure 3.5. *(preceding page)*: Return to maximum complexity after polyploidization.
(**A**) Random genomes of different lengths and ploidy levels demonstrate an increase in CR through the accumulation of random mutations ($1e^{-08}$ mutations per site, per generation). Chromosomes of maize and sorghum are included in the simulation. **(B)** Starting from one random sequence, we simulated two rounds of diploidization, thereby simulating a tetraploid genome. The experiment was repeated for different lengths. Translocations of 1Kb occurred at a constant rate over 10,000 generations (plotted for each 100 generations). "○", "△", "+" and "×" represent four simulated chromosomes, while the different coloured lines their concatenations.

3.5.2. Mechanisms of genome amplification and divergence

Genome size enlargement by duplications results in a reduction in CR. However, maximum CR is rapidly recovered, as evidenced by ancient polyploids in the dataset that all presented high values of CR. Also, the simulations conducted showed that 30 million generations is sufficient to recover a CR>0.95 in the tested polyploids, even only taking into account single mutations at a mean rate. In this context, we view the evolution of genome complexity as successive decreases and increases in CR. It is thought that during this process, intermediate states with genomes of constant sizes suffering mutations and rearrangements could give birth to new functional sequences, thus providing the raw material for species divergence and growth in biological complexity [Lynch 2000].

3.5.3. Genome size reduction

Theoretically, genome size reduction events are not expected to lower the CR, because any sufficiently large region of the genome showed an almost random structure. Natural selection will ultimately determine the extent of losses of genome segments but, as observed, intracellular bacterial parasites seemed to move (along the straight line of Figure 3.1-A) from larger genomes sizes to shorter genomes along a trajectory that fit almost maximum CR during this process. An other example is given by the ancient polyploid species *Arabidopsis thaliana* [Hu *et al.* 2011], which recovers maximum CR after diploidization and genome reduction. This is in contrasting to close relative *Arabidopsis lyrata*, which still possesses a high degree of genomic redundancy.

Thus, in the case of both gradual and sharp genomic reductions, the pattern demonstrated by the evolution of CR values is expected to be smoother than in the case of genome amplification.

3.5.4. Limits of CR space

We speculate that the theoretical space of complexity ratio values filled by human texts Figure 3.1-B, is a region neglected by life. A non-random combinatorial structure for DNA is inconceivable for organisms with small genomes like viruses, phages or prokaryotes. Through the effects of natural selection, simple forms of life –with genome sizes ranging from the equivalent of a few paragraphs of text to the complete works of William Shakespeare – are probably forced to have a random like-structure, thereby limiting, their alternatives to a high genomic complexity.

3.5.5. Hypotheses

Together, these observations lead to the hypotheses that:

- A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes throughout the diversity of life.

- The fate of polyploid genomes is to reach almost maximal complexity in their DNA structure, increases in CR occur as a function of time.

- Since the DNA combinatorial structure is quasi-random, genome complexity only increases through DNA amplification followed by the divergence of duplicates during evolution.

However, these hypotheses can be nullified in some specific cases:

- Genomes of recent polyploid species evidencing a quasi-random DNA structure (high CR), as species that have undergone significant genome reduction.

- Genomes evidencing a non-random DNA structure (low CR), for example, due to a very strong A + T content bias.

In this chapter, the combinatorial DNA structure of genomes has been described. We hypothesized a universal random-like structure throughout the diversity of life. It is very hard to believe that such a structure is adaptive in origin. However, far from being biologically-irrelevant, useful properties may emerge from such a random-like combinatorial structure in genomes. After all, exons, the main functional units of genes, are the elements with the most random-like DNA structure.

A simple pattern controlling the genome's statistical design for all kinds of organisms makes nature modest and beautiful.

4. Ecology of genome elements

In the previous chapter, we showed that the intrinsic structure of genomes is nearly random in informational terms or, in other words, that any aleatory combination of the four nucleotides would have (nearly) the same properties as a real genome. When studying the combinatorial structure of genetic elements (GEs) (see **Complexity in repetitive elements and genes – low and high?** in section 3.3), we also demonstrated that the complexity of some families of GEs was slightly lower when considered separately. However, given that the complexity ratio is almost identical for all chromosomes of a given genome, we can therefore suggest that GEs are, in some way, homogeneously distributed in chromosomes.

This hypothesis, however, contradicts the observed proportion of GEs in closely related species. For example, in the Introduction (**Dynamics of genetics elements**), we reported how the dynamics of TEs led them to present diametrically opposed proportions of DNA transposons and retrotransposons in Eukaryotes (Figure 1.4).

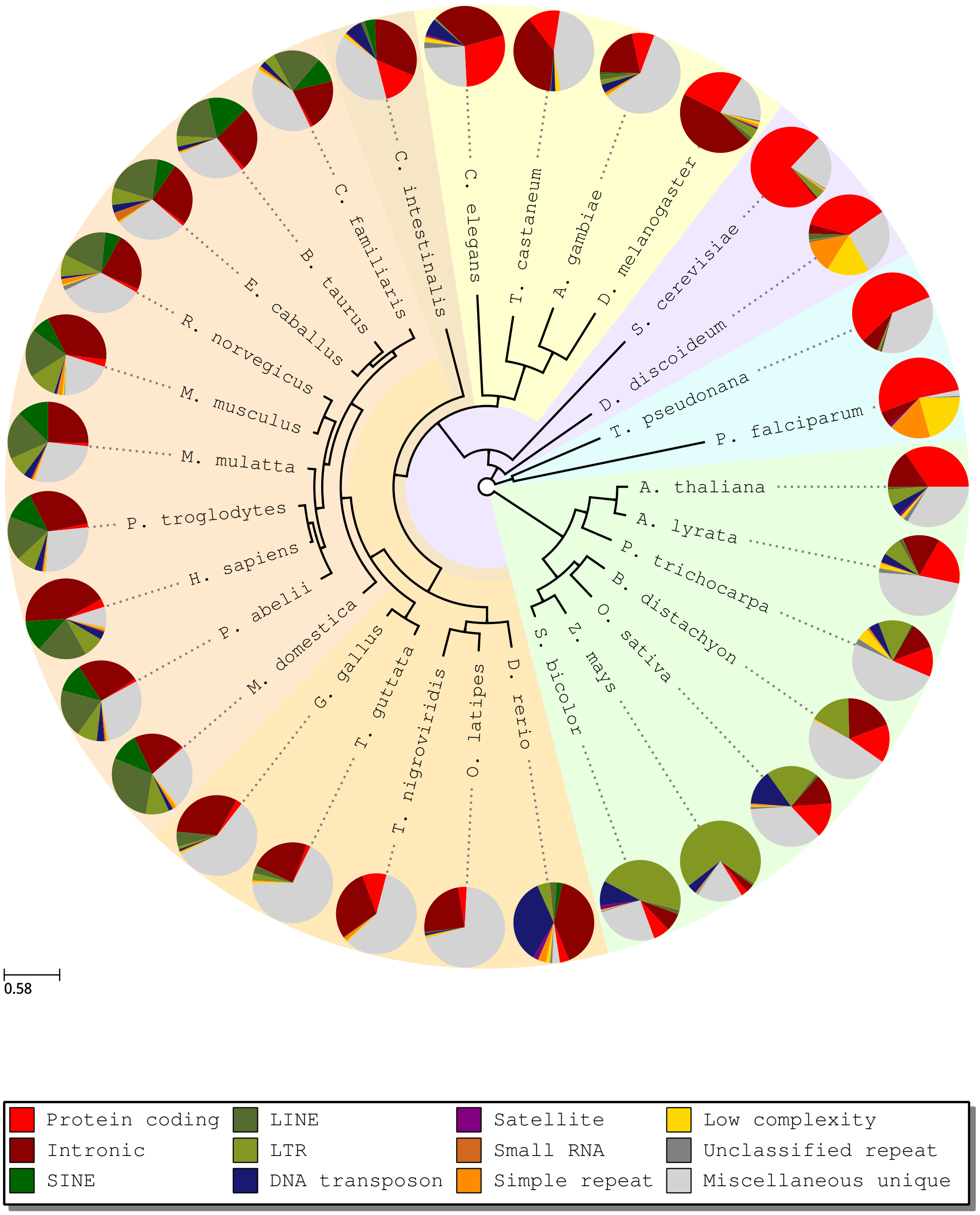
As initial results and in order to get a broader view of the differences in proportions of GEs in eukaryotes, we wanted to report the proportions of repetitive elements and genic regions found in the 31 eukaryotic species (see **Mining Genetic Species**, page 28). The resulting proportions are summarized in the introductory Figure 4.1. In addition, some clade-specific trends can be observed; in mammals, for example, there were significant differences between species.

In this context, we will analyze the diversity and abundance of groups or families of GEs – which we refer to as genetic species (GSs) in this chapter – with simple statistical tests against a hypothesis of random distribution.

4.1. Non-random distribution of genetic elements

The simplest hypothesis in relation to the distribution of GSs is that they are randomly distributed throughout genomes. In order to test this hypothesis, we simulated a thousand random distributions of GSs among the chromosomes of each genome, and tested if these proportions were conserved in each chromosome by applying a one sample t-test.

The number of t-test computed here was significant, for each of the 548 chromosomes among the set of 31 genomes, we compared the observed abundance for each of the GSs to the distribution of the corresponding simulated data. After a


C. intestinalis
C. elegans
T. castaneum
A. gambiae
D. melanogaster
S. cerevisiae
D. discoideum
T. pseudonana
P. falciparum
A. thaliana
A. lyrata
P. trichocarpa
B. distachyon
O. sativa
Z. mays
S. bicolor
D. rerio
O. latipes
T. nigroviridis
T. guttata
G. gallus
M. domestica
P. abelii
H. sapiens
P. troglodytes
M. mulatta
M. musculus
R. norvegicus
E. caballus
B. taurus
C. familiaris
0.58
Protein coding
Intronic
SINE
LINE
LTR
DNA transposon
Satellite
Small RNA
Simple repeat
Low complexity
Unclassified repeat
Miscellaneous unique

necessary correction for multiple testing, less than 4% of all GSs of the chromosomes tested showed abundances according to their genomic mean (assuming that the result of the simulation tends to a genomic mean).

This result, while negative, was expected given the fast dynamic that seems to underlie the distribution of GSs in eukaryotic genomes (Figure 4.1). Moreover, random distributions are the exception in nature, even in the case of species distributions and abundances that are governed by stochastic processes. Here, we use statistical tools developed by ecologists in order to decipher the dynamics underlying the diversity and abundance of GSs.

4.2. Counterbalanced species abundances in genomes

4.2.1. Genetic species: diversity and abundance

Ecologists frequently use relative species abundance (RSA) curves to compare the richness, the degree of dominance and the number of rare species among communities. The raw data used in such plots is the total number of individuals per species sampled in the ecosystem. The most interesting property of RSA curves is that species are not labeled in the ranking order; hence ecosystems can be compared, whatever species they contain.

Here, RSA curves were built using the full set of GSs for each of the 548 chromosomes, and also for the corresponding 31 complete genomes. The raw data thus, represent a census of GSs.

Figure 4.2 displays RSA curves for a selected group of genomes and their largest chromosome. The curves differ in many ways, although two patterns are evident: ***1)*** the RSA curves of genomes and chromosomes are very similar – the only noticeable difference being a reduction in the number of GSs in chromosomes, and ***2)*** all RSA curves (from both genomes and chromosomes) display the universal S-shape also observed in ecological environments [McGill *et al.* 2007, Hubbell

Figure 4.1. *(preceding page)*: Proportions of the major groups of genetic elements in 31 eukaryotes.
The phylogeny is adapted from [Huerta-Cepas *et al.* 2012] with a correction for the amoeboid *D. discoideum* [Roger & Simpson 2009]. Pie charts display the proportions of the various genetic elements. Terms in the legend correspond to: ***1)*** **Protein coding** sequences, ***2)*** **Intronic** sequences and untranslated regions, ***3)*** **LINE**, ***4)*** **SINE**, ***5)*** **LTRs** all three retrotransposons types, ***6)*** **DNA transposon**, ***7)*** **Satellite** long tandem repeats, ***8)*** **Small RNA** mostly tRNA or snRNA pseudogenes, ***9)*** **Simple repeat** or microsatellites, ***10)*** **Low complexity** poly-purine or poly-pyrimidine (AT or GC rich), ***11)*** **Unclassified repeat** i.e. not yet characterized repetitive elements, ***12)*** **Miscellaneous unique** basically what remains after identification of all previous elements.

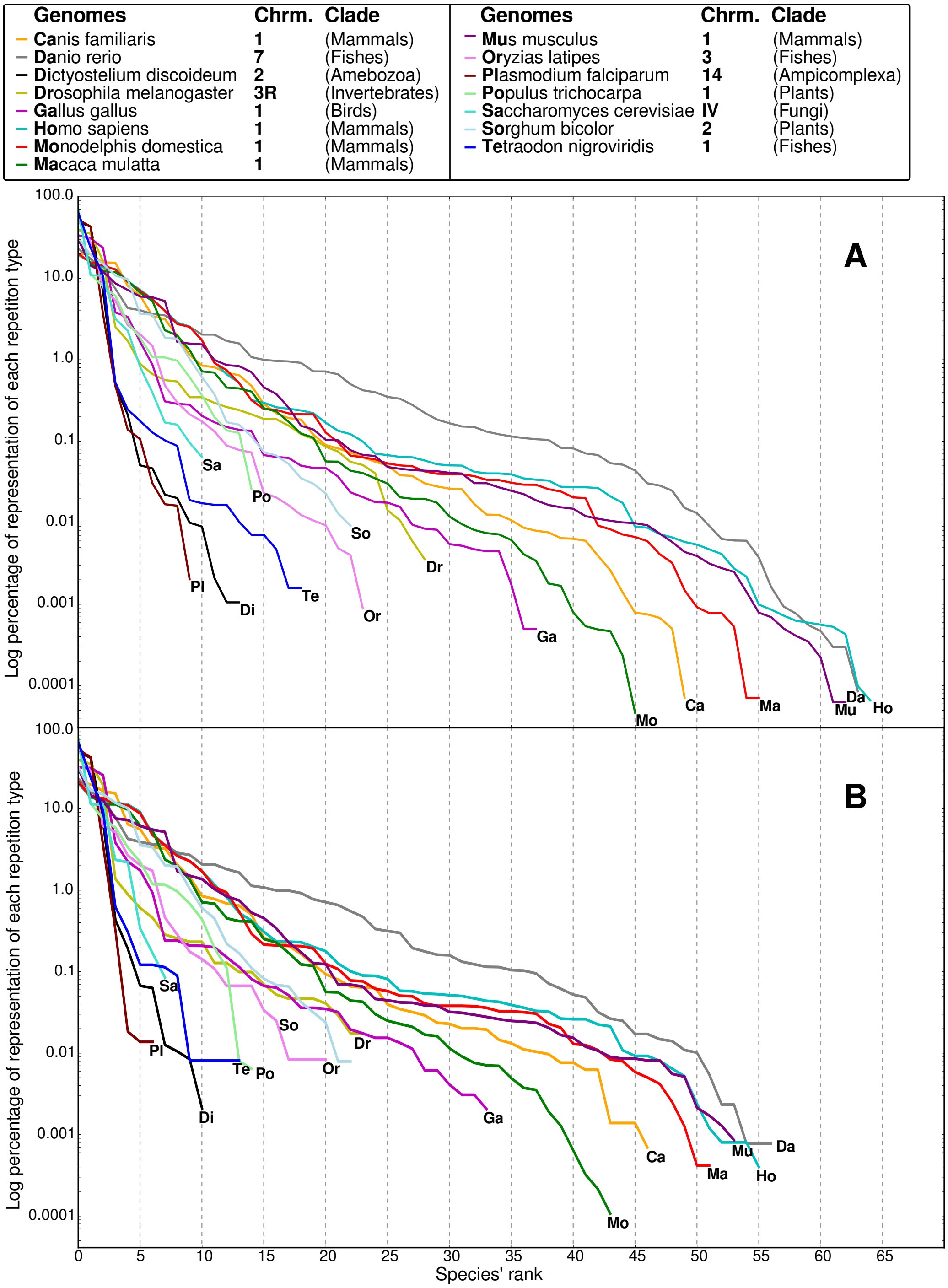


Figure 4.2.: Relative Species Abundance (RSA) curves.
RSA for some selected genomes **(A)** and their corresponding largest chromosomes **(B)**.

2001]. Both observations suggest a common mechanism of distribution of GSs in genomes and chromosomes.

4.2.2. Relative species abundance curves in genomes and chromosomes

To what extent do chromosomal RSA curves represent the random distribution of the complete set of elements of the genome? To answer this question, we used the same simulated data used in section **Non-random distribution of genetic elements**. The mean expected abundance and standard deviation of this simulated data, were used to plot random expected RSA curves for chromosomes.

According to **Non-random distribution of genetic elements** and to results just published for TE's in *D. melanogaster* [Bartolomé *et al.* 2002, Rizzon *et al.* 2002], an homogeneous random process, cannot account for the observed abundances of genetic species in chromosomes. However, if observed and simulated chromosomic RSA curves are superimposed, a notable concordance is visible. Figure 4.3 shows this concordance for two chromosomes. This remarkable concurrence is permitted by removing the GSs labels in the RSA plots and, hence, by the shift this facilitates in the ranking order of abundances. For instance, tRNA and satellite elements, respectively occupy positions 43 and 23 of the overall ranking of abundances in the human genome Figure 4.3-A. However, in chromosome 1, their ranking positions are 33 and 42, respectively. That, means that tRNA and satellites elements show higher and lower abundances, respectively, than that expected by random distribution.

In order to test the degree of adjustment for the superimposed simulated and observed data (red and grey curves in Figure 4.3), a Kolmogorov-Smirnov test was conducted. Overall, and after the necessary correction for multiple testing, statistical differences between observed and simulated data were detected for only 76 out of 548 chromosomes tested (KS-test, $P < 0.05$); that 86% of chromosomes' RSA were found to be finely superimposed with simulated ones. This result, of obtained with the same data but summarizing species identity to a local ranking, contrasts greatly with the previous finding that only 4% of the GSs were found in the proportions expected by chance.

Thus, given the 86% agreement of chromosomes tested here, GS distribution actually does appear to be governed by some kind of stochastic process.

4.3. Genetic elements: diversity and chromosome length

If a purely stochastic process controls the abundance and diversity of genetic elements in chromosomes, as suggested by the previous results, it is expected that S, the number of GSs present in a given chromosome, will increase with

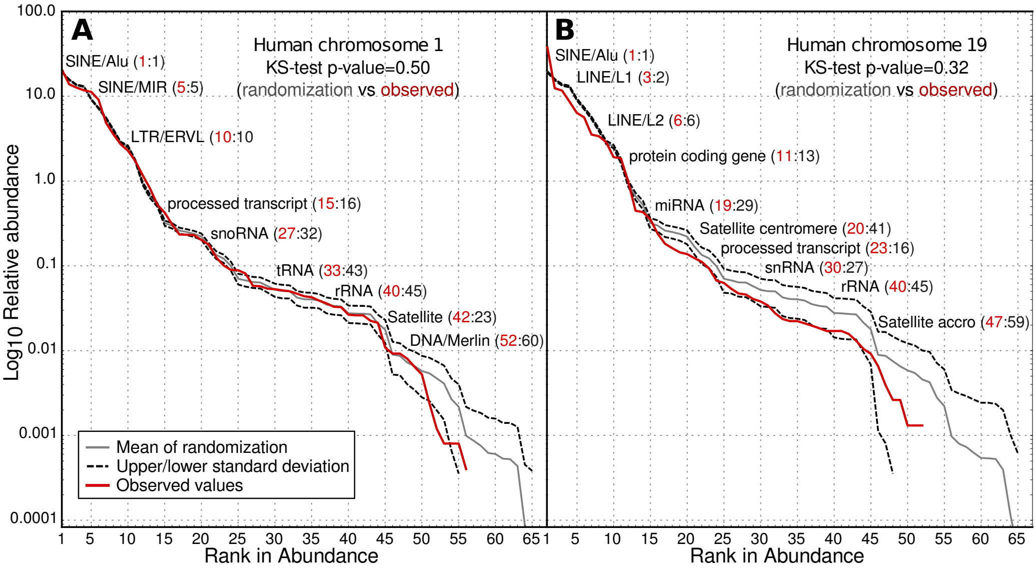


Figure 4.3.: Relative Species Abundance curves for human chromosome 1 (A) and chromosome 19 (B).

Red and grey lines represent observed and simulated values, respectively, for all genetic species in the chromosomes. Dotted lines are two standard deviations around the mean simulated values. Numbers in parenthesis depict the observed (red) and the expected (grey) values in the ranking of abundances for a few classes of GSs in both chromosomes. Note the higher than expected number of SINEs/Alu elements in human chr 19.

chromosome length. In ecology, it is universally observed that larger areas contain more species. Does this pattern hold true for chromosomes?

The standard species-area relationship in ecology is the Arrhenius power law [Arrhenius 1921] $S = cAz$, where S is the number of species, A is the area and c and z are constants. Following in from this comparison, species areas would be analogous to chromosome length. Figure 4.4 displays the correlation between the number of GSs and the chromosome size in a log-log transformation for polyploids and all chromosomes.

Although there is good correlation between the number of GSs and chromosome length, the first discrepancy in the analogy being used here between genomes and ecosystems becomes apparent. The plotted polyploid chromosomes seem to display a slower increase in GSs along the axis of chromosome length. As this observation is completely predictable, given that genomes subjected to polyploidization should theoretically contain the same amount of GSs despite their larger size, we computed the statistical fit of the species-area relationship solely for non-polyploid eukaryotic species.

After a least square fit of the power function, $c = 0.28$, $z = 0.27$ ($R^2 = 0.64$, $n =$

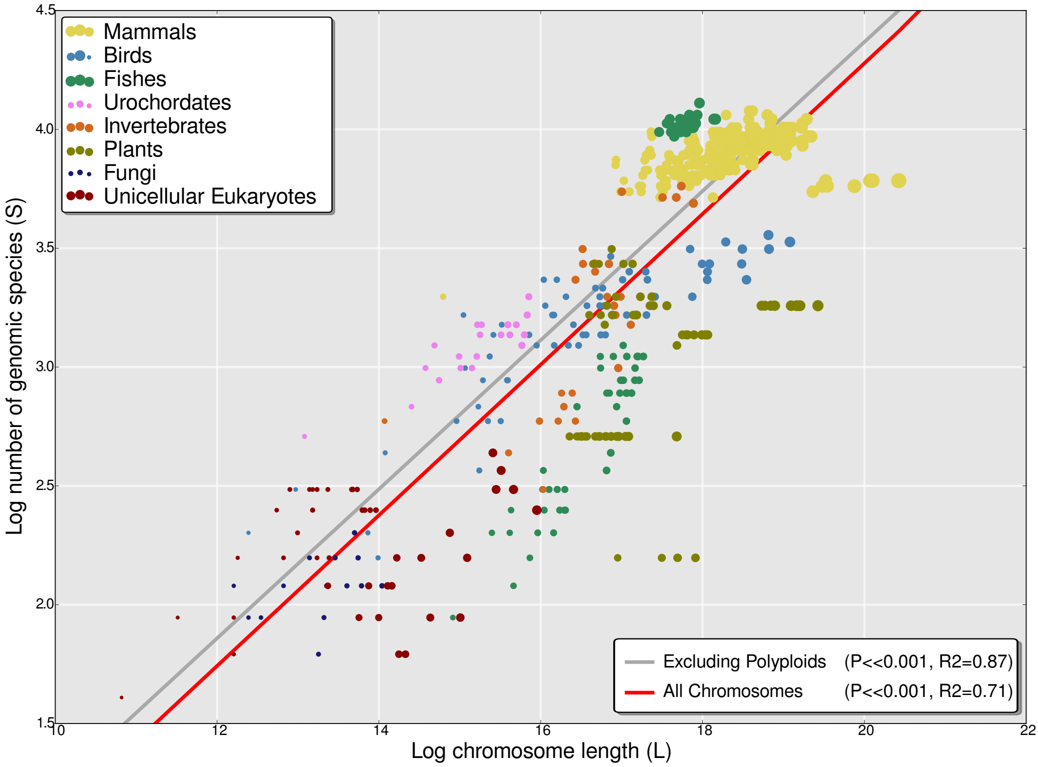


Figure 4.4.: Species chromosome size relationship

This plot represents the correlation between the number of genetic species (GSs) and the chromosome size in a log-log transformation. As additional information, the size of the dots is a function of the number of individuals belonging to a given species.

548) for all chromosomes studied (including for polyploid fish and plant species), and $c = 0.50$, $z = 0.25$ ($R^2 = 0.81$, $n = 412$) (excluding polyploids as mentioned above).

In both cases, the adjustment was statistically significant in a linear regression model ($P << 0.001$). Thus, and as in community ecology, eukaryote chromosomes display the universal species-area relationship with z values corresponding to regional spatial scales [Rosenzweig 1995].

Thus, we believe there is strong evidence that the distribution of GSs among chromosomes is characterized by a strong stochastic component that explains: *1)* the observations raised by the comparison of RSA curves of simulated and observed distributions of GSs (see previous section **Relative species abundance curves in genomes and chromosomes**), and *2)* this last result, showing how the number of GSs present in a chromosome is strongly correlated with its length. In order to improve and more accurately test this observation, we investigated if a neutral dynamic model could predict this shared demographic pattern of genomes.

4.4. Neutrality of species abundances and diversity

Similar to the kinetic theory of ideal gases in physics, the unified neutral theory of biodiversity (UNTB [Hubbell 2001]) 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 per generation, regardless of the species. Metacommunity dynamics are controlled by speciation at a single constant rate (ν) [Rosindell *et al.* 2011, Alonso *et al.* 2006].

For genomes, we realized that each chromosome is the physical arena in which GSs die and are replaced by other elements of the same or different species. These GSs may 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 the genome corresponds to the metacommunity of size J_M. Thus, we used the sample of the total number of functional and non-functional elements in each chromosome as raw-data to optimize by maximum likelihood (ML) the neutral model's parameters m and θ $(= 2J_M\nu)$ using Ewens and Etienne's sampling formula (Equation 2.6).

Deviations from neutrality were detected in 33 out of 548 (6.0%) chromosomes. However, these deviations vanished after correction for multiple tests correction ($FDR < 0.05$, see Table 4.1 for a summary of the results). We conclude that Hubbell's neutral model fits the abundance and diversity of GSs in all the chromosomes of the 31 eukaryote genomes analyzed.

4.5. Discussion

Almost one hundred years ago, ecologists recognized the universally uneven distribution of species abundances, and the fact of species increments in larger areas [Magurran 2004]. Just a decade ago, however, neutral demographic processes emerged as the simplest mechanical explanation behind both patterns in communities [Hubbell 2001]. More recently, Michael Lynch and John S. Conery [Lynch & Conery 2003] hypothesized that the complexity of eukaryote genomes emerged passively during evolution as a consequence of population size reduction. Here, we have demonstrated that a simple stochastic process associated with to a few parameters fits the pattern of abundance and diversity of genetic species for a great diversity of eukaryote genomes.

An example of the implementation of a neutral model to explain the distribution of genetic elements in genomes can be found very recently in the work conducted

Species	**Ch**	***J***	***S***	**ΔH**	**θ**	***m***	**P (Q)-val**	**Model**
Tribolium castaneum	7	7,865	18	- 0.97	2.12	—	0.01 (0.66)	Ewens
Anopheles gambiae	X	21,215	42	- 0.58	6.97	0.037	0.03 (0.66)	Etienne
Gallus gallus	9	6,621	32	- 0.56	4.28	—	0.05 (0.66)	Ewens
Drosophila melanogaster	X	20,787	26	- 0.51	2.86	—	0.09 (0.66)	Ewens
Tetraodon nigroviridis	3	8,505	17	- 0.49	1.96	—	0.11 (0.66)	Ewens
Mus musculus	14	143,018	59	- 0.27	6.58	0.149	0.15 (0.66)	Etienne
Populus trichocarpa	2	32,946	15	-0.37	2.23	0.009	0.16 (0.66)	Etienne
Oryzias latipes	19	7,223	21	-0.35	2.57	—	0.17 (0.66)	Ewens
Homo sapiens	17	93,105	52	- 0.24	6.51	0.065	0.18 (0.66)	Etienne
Macaca mulatta	16	84,626	50	- 0.19	5.95	0.119	0.23 (0.66)	Etienne
Saccharomyces cerevisiae	II	640	9	- 0.25	1.37	—	0.26 (0.66)	Ewens
Dictyostelium discoideum	1	26,650	14	- 0.24	1.36	—	0.27 (0.66)	Ewens
Danio rerio	1	105,305	56	- 0.11	8.17	0.016	0.29 (0.66)	Etienne
Canis familiaris	1	144,103	47	-0.10	5.28	0.093	0.32 (0.66)	Etienne
Plasmodium falciparum	13	18,738	10	-0.11	0.95	—	0.38 (0.66)	Ewens
Monodelphis domestica	2	675,788	44	-0.02	4.46	0.031	0.43 (0.66)	Etienne
Sorghum bicolor	1	37,626	23	0.19	2.86	0.067	0.68 (0.79)	Ewens

Table 4.1.: Result of the fit to the UNTB and the test of neutrality for selected chromosomes

The table depicts the parameters and statistics estimated for a selection of chromosomes of different species. Chromosomes are arranged according to p-value, from the least to the most neutral. J = the total number of genetic elements; S = the number of genetic species; ΔH = the difference between the observed and the expected evenness (Shannon's diversity index); θ = the fundamental diversity number; m = the migration rate; P and Q-val = statistical significances of the neutral test before and after a false-discovery rate correction. The last column shows the model (Ewens or Etienne's) that best fitted the empirical distribution of genetic elements in the chromosome after likelihood-ratio test ($p < 0.05$, df = 1). None of the 548 chromosomes from the 31 eukaryote genomes showed significant deviations from neutrality (Q-val<0.05).

by Bart Haegeman and Joshua S. Weitz on six bacterial genomes [Haegeman & Weitz 2012]. In this work the authors defined a neutral model for the evolution of the genomes that was able to predict the proportions of: ***1)*** the genes shared by all the genomes, ***2)*** genes absent from some genomes, and ***3)*** species-specific genes. This model, which combined birth and death process among individuals of the same species, and gene transfer between species, was able to reproduce the observation that most genes are either specific to one genome or shared by all species. Although the model proposed by the authors of that work is only applicable to bacteria, we think that their results are along the same lines as ours in that the appearance and conservation of genes in genomes follows a neutral process.

We are certainly aware that the statistical fit of a neutral pattern does not necessarily imply the existence of a neutral process behind the pattern (see **Power and specificity of the neutral test** section 2.2.7), but the excellent and taxonomically-broad fit of neutral theory to genetic element diversity and abundance reported here raises the question as to *why there is not a stronger signature of natural se-*

lection in ecological communities or in genomes at larger scales? Ecologists have acknowledge the existence of many kinds of trade-offs in community ecology; for instance, species with high dispersal rates are not good competitors. However, it is not yet known to what extent such trade-offs maintain diversity or are consistent with, and permit, neutral dynamics. For genomes, the mechanisms that maintain element diversity, and whether these involve trade-offs, are not yet understood. Which mechanisms operate will also depend on whether genome size is under strong or weak selection [Cavalier-Smith 2005]. More likely, genetic element diversity arises from some combination of neutral drift and selection on different genetic species [Mustonen & Lässig 2009]. Independently of the answer to this question, the model tested here should be employed as the null hypothesis to test for alternative mechanisms explaining species abundance and diversity in eukaryote genomes.

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

In previous chapters, common genomic patterns in terms of informational content (chapter 3) and common demographic and ecological patterns explained with significant accuracy, the composition of eukaryotic genomes, were described (chapter 4). In both cases, the results presented here allow us to conceive biological diversity and physiology in a way that completely omits the process of adaptation through natural selection.

In this chapter, we turn our attention to this “forgotten” directional parameter, raising the level of study to functional systems and overlooking the genic level. We analyze the effect of natural selection on the full set of protein-coding genes in two groups of organisms, mammals and *Drosophila*.

5.1. Gene set selection analysis on functional modules

We studied mammals (represented by human, chimpanzee, rat and mouse genomes) and five *Drosophila* species. For each genome, genes were ranked into four categories according to the estimation of ***1)*** synonymous (dS) and ***2)*** non-synonymous (dN) rates of substitution; ***3)*** selective pressures ($\omega = dN/dS$); and ***4)*** the change of selective pressures between (A) ancestor and (D) descendant species ($\Delta\omega_D = \omega_D - \omega_A$) in the phylogeny Figure 2.5 (see section: **Gene set selection analysis (GSSA): evolutionary and statistical simulations**, page 39 for details on the methodology).

The application of the GSSA over the lists of genes ranked by dS, dN, ω and the $\Delta\omega$ values yielded a large number of functional modules with rates that were significantly skewed towards the extremities of the lists (Table 5.1) for both mammal and *Drosophila* species. In mammals for instance, 11% of the GO terms, and 15% of the KEGG pathways tested were found to be significantly enriched in genes with high ω rates (SHω, 5% false-discovery rate, FDR). In *Drosophila*, slightly lower proportions were found, with 4.1% and 2.6% of GO terms and KEGG pathways respectively.

Table 5.1 also reveals that functional modules with genes changing at significantly low ω ratios (SLω), and therefore showing a distribution that is biased towards the bottom of the ranked list (see Figure 5.1), were more frequent than

		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.

Significantly High (SH) and Significantly Low (SL) results after correction for multiple testing (5%FDR) are shown.

modules with a 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 to the neutral character of synonymous mutations, and with the effects of population size on the final outcome of selection [Lynch 2007], GSSA results show a higher number of significant deviations of dS in *Drosophila* than in mammals.

When contrasted with the ω rate of ancestral sequences, the observed tendency is for only a minor proportion of functional terms to be under significantly high or low selective pressures. Specifically, increased or decreased ω values up to the external branches (marked 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 descendants. This observation highlights the conservative character of selective constraints in functionally-related groups of genes.

Results of the GSSA for mammals and *Drosophila* are summarized in Figure 5.1A and Figure 5.1B, respectively. These figures show a selection of functional terms with significantly high or low rates for each of the evolutionary variable considered. The first striking point, also indicated in Table 5.1, is the number of significant results. By considering the full set of genes, the GSSA is able to detect functional biases with much more statistical power than if genes belonging to a given evolutionary scenario are only considered.

5.1.1. Clade specific patterns

The first pattern that stands out when looking at Figure 5.1 is the differentiation of clades: human together with chimpanzee, mouse with rat and, among the *melanogaster subgroup*, *D. yakuba* and *D. erecta* also show similar patterns. For instance, functional terms associated with neurological processes and sensory perception clearly contrasted between primates and rodents (Figure 5.1-A). This segregation by clade is to be expected, taking into account the extent of evolutionary history since the ancestral state in the final count of synonymous and non-synonymous changes. As an example, we focus on the terms related to neurological processes in mammals. In both humans and chimpanzees, neurological processes and sensory perception show a significant increase in ω when compared to their common ancestor: evidenced by the statistical significance of $\Delta\omega$. In contrast, in rodents, the values of $\Delta\omega$ are significantly low for the functional terms related to neurological processes.

Alternatively, functional modules associated with *"Immunity"* and the *"Defense response"* evolved at significantly higher rates than expected in rodents, but decreased significantly in comparison to ancestral rates in primates. Such functional differences between primates and rodents have previously been observed when groups of species were pooled [Kosiol *et al.* 2008] (see section: Genomic study of selective pressures in set of genes, page 14). Aside from these fast-evolving categories, other functional modules such as *"Development"* and *"Transcription/-Transduction"* evolved at comparatively very low dN and ω ratios but experienced a stronger relaxation of ancestral constraints ($+\Delta\omega$) in primates than in rodents.

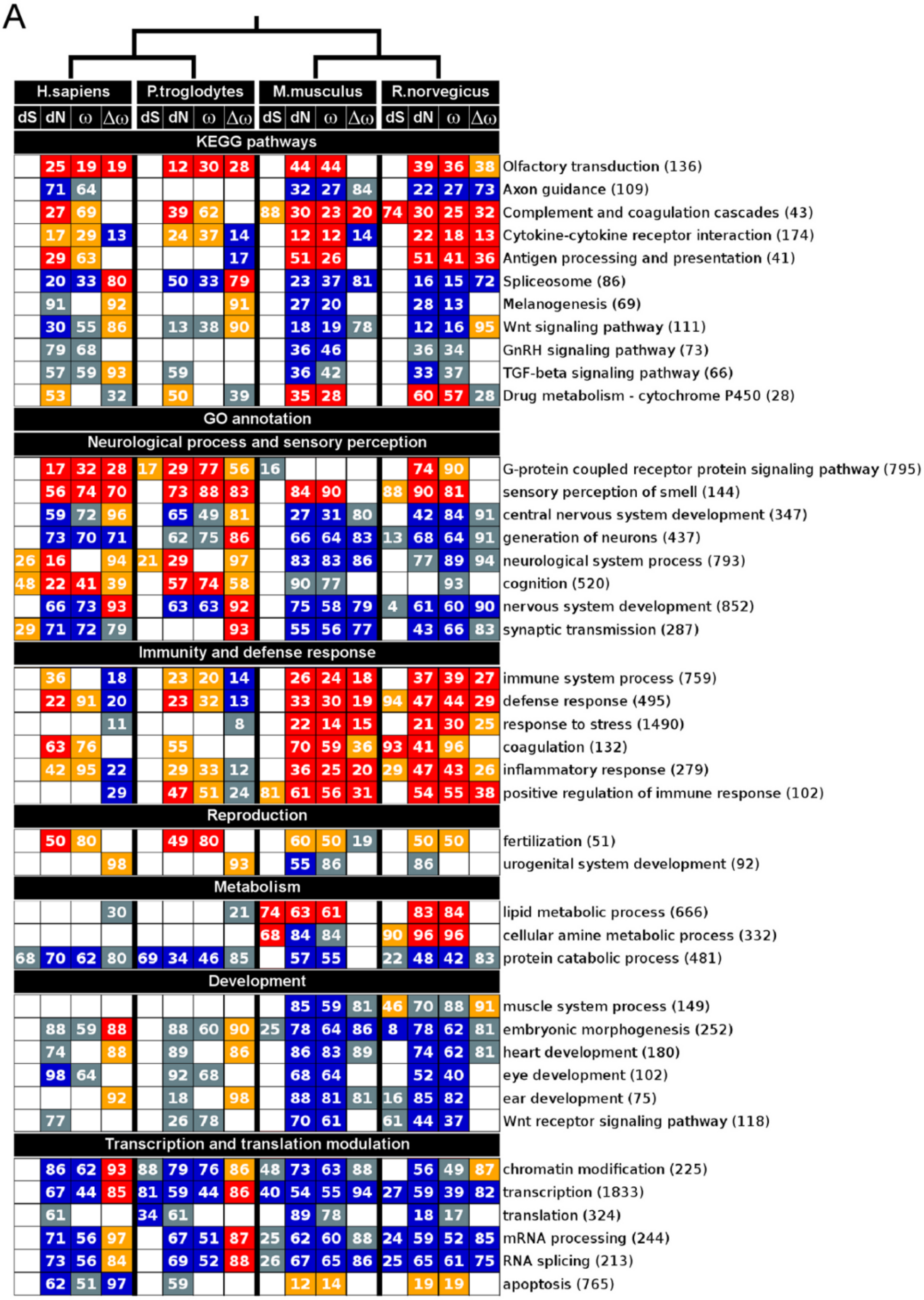
In *Drosophila*, most of the GO terms significantly associated with high dN and ω were also unevenly distributed within the two clusters of the phylogeny (Figure 5.1-B). GO terms such as *"Sensory perception"*, *"Defense response"*, *"Immune response"* and *"Metabolic process"* among others, presented remarkable divergence in the monophyletic groups of *D. erecta* and *D. yakuba*, but they were not significant in *D. sechellia*, *D. melanogaster* and *D. simulans.* Most of the GO terms related to *"Development"*, *"Transcription"* and *"Translation"* (Figure 5.1-A and -B) were found to be constrained by purifying selection with significantly low rates of ω (5% FDR) in both taxa.

5.1.2. Species-specific enrichment

Going further into the analysis of the results, there is evidence of species-specific functional enrichment. Following on from the example of sensory perception in rodents, the *"G-protein coupled receptor protein signaling pathway"* has a signifi-

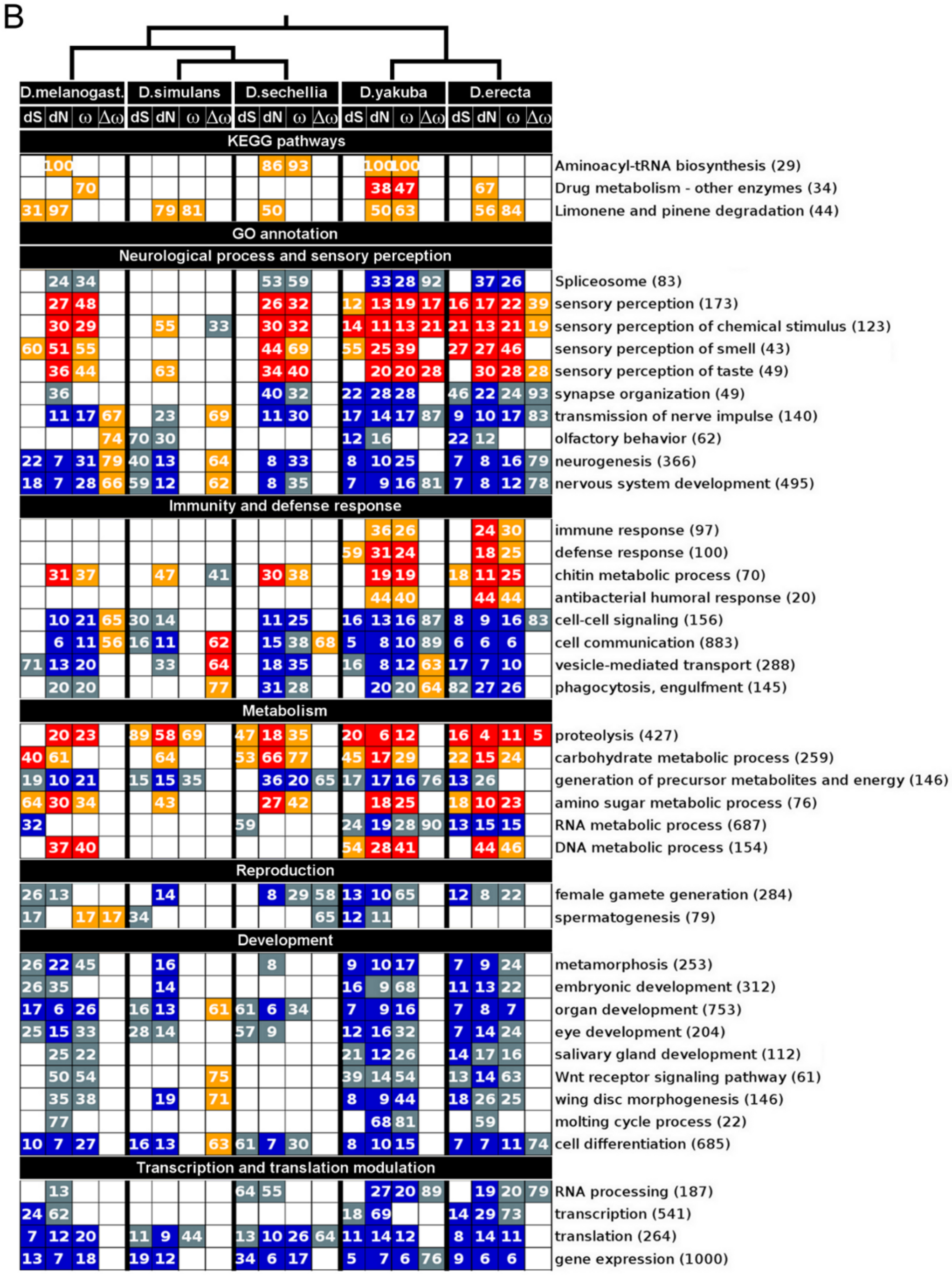
Figure 5.1. *(following page)*: GSSA of evolutionary variables.
The figure shows a selection of GO terms and KEGG pathways with significant and non-significant deviations following GSSA of evolutionary rates in mammal (**A**) and Drosophila (**B**) species. Colored boxes represent functional modules, with genes significantly grouped at the corresponding extremities of the ranked list, as explained in Figure 2.6. The number inside each box represents the percentage of the total number of genes of the functional module (in parenthesis) that contributes 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	dS	dN	ω	Δω	dS	dN	ω	Δω	dS	dN	ω	Δω	
KEGG pathways																
	25	19	19		12	30	28		44	44			39	36	38	Olfactory transduction (136)
	71	64							32	27	84		22	27	73	Axon guidance (109)
	27	69			39	62		88	30	23	20	74	30	25	32	Complement and coagulation cascades (43)
	17	29	13		24	37	14		12	12	14		22	18	13	Cytokine-cytokine receptor interaction (174)
	29	63					17		51	26			51	41	36	Antigen processing and presentation (41)
	20	33	80		50	33	79		23	37	81		16	15	72	Spliceosome (86)
	91		92				91		27	20			28	13		Melanogenesis (69)
	30	55	86		13	38	90		18	19	78		12	16	95	Wnt signaling pathway (111)
	79	68							36	46			36	34		GnRH signaling pathway (73)
	57	59	93		59				36	42			33	37		TGF-beta signaling pathway (66)
	53		32		50		39		35	28			60	57	28	Drug metabolism - cytochrome P450 (28)
GO annotation																
Neurological process and sensory perception																
	17	32	28	17	29	77	56	16					74	90		G-protein coupled receptor protein signaling pathway (795)
	56	74	70		73	88	83		84	90		88	90	81		sensory perception of smell (144)
	59	72	96		65	49	81		27	31	80		42	84	91	central nervous system development (347)
	73	70	71		62	75	86		66	64	83	13	68	64	91	generation of neurons (437)
26	16		94	21	29		97		83	83	86		77	89	94	neurological system process (793)
48	22	41	39		57	74	58		90	77				93		cognition (520)
	66	73	93		63	63	92		75	58	79	4	61	60	90	nervous system development (852)
29	71	72	79				93		55	56	77		43	66	83	synaptic transmission (287)
Immunity and defense response																
	36		18		23	20	14		26	24	18		37	39	27	immune system process (759)
	22	91	20		23	32	13		33	30	19	94	47	44	29	defense response (495)
			11				8		22	14	15		21	30	25	response to stress (1490)
	63	76			55				70	59	36	93	41	96		coagulation (132)
	42	95	22		29	33	12		36	25	20	29	47	43	26	inflammatory response (279)
			29		47	51	24	81	61	56	31		54	55	38	positive regulation of immune response (102)
Reproduction																
	50	80			49	80			60	50	19		50	50		fertilization (51)
			98				93		55	86			86			urogenital system development (92)
Metabolism																
			30				21	74	63	61			83	84		lipid metabolic process (666)
								68	84	84		90	96	96		cellular amine metabolic process (332)
68	70	62	80	69	34	46	85		57	55		22	48	42	83	protein catabolic process (481)
Development																
									85	59	81	46	70	88	91	muscle system process (149)
	88	59	88		88	60	90	25	78	64	86	8	78	62	81	embryonic morphogenesis (252)
	74		88		89		86		86	83	89		74	62	81	heart development (180)
	98	64			92	68			68	64			52	40		eye development (102)
			92		18		98		88	81	81	16	85	82		ear development (75)
	77				26	78			70	61		61	44	37		Wnt receptor signaling pathway (118)
Transcription and translation modulation																
	86	62	93	88	79	76	86	48	73	63	88		56	49	87	chromatin modification (225)
	67	44	85	81	59	44	86	40	54	55	94	27	59	39	82	transcription (1833)
	61			34	61				89	78			18	17		translation (324)
	71	56	97		67	51	87	25	62	60	88	24	59	52	85	mRNA processing (244)
	73	56	84		69	52	88	26	67	65	86	25	65	61	75	RNA splicing (213)
	62	51	97		59				12	14			19	19		apoptosis (765)

B



D.melanogast.				D.simulans				D.sechellia				D.yakuba				D.erecta				
dS	dN	ω	Δω	dS	dN	ω	Δω	dS	dN	ω	Δω	dS	dN	ω	Δω	dS	dN	ω	Δω	
KEGG pathways																				
	100								86	93			100	100						Aminoacyl-tRNA biosynthesis (29)
		70											38	47			67			Drug metabolism - other enzymes (34)
31	97				79	81			50				50	63			56	84		Limonene and pinene degradation (44)
GO annotation																				
Neurological process and sensory perception																				
	24	34							53	59			33	28	92		37	26		Spliceosome (83)
	27	48							26	32		12	13	19	17	16	17	22	39	sensory perception (173)
	30	29			55		33		30	32		14	11	13	21	21	13	21	19	sensory perception of chemical stimulus (123)
60	51	55							44	69		55	25	39		27	27	46		sensory perception of smell (43)
	36	44			63				34	40			20	20	28		30	28	28	sensory perception of taste (49)
	36								40	32		22	28	28		46	22	24	93	synapse organization (49)
	11	17	67		23		69		11	30		17	14	17	87	9	10	17	83	transmission of nerve impulse (140)
			74	70	30							12	16			22	12			olfactory behavior (62)
22	7	31	79	40	13		64		8	33		8	10	25		7	8	16	79	neurogenesis (366)
18	7	28	66	59	12		62		8	35		7	9	16	81	7	8	12	78	nervous system development (495)
Immunity and defense response																				
													36	26			24	30		immune response (97)
												59	31	24			18	25		defense response (100)
	31	37			47		41		30	38			19	19		18	11	25		chitin metabolic process (70)
													44	40			44	44		antibacterial humoral response (20)
	10	21	65	30	14				11	25		16	13	16	87	8	9	16	83	cell-cell signaling (156)
	6	11	56	16	11		62		15	38	68	5	8	10	89	6	6	6		cell communication (883)
71	13	20			33		64		18	35		16	8	12	63	17	7	10		vesicle-mediated transport (288)
	20	20					77		31	28			20	20	64	82	27	26		phagocytosis, engulfment (145)
Metabolism																				
	20	23		89	58	69		47	18	35		20	6	12		16	4	11	5	proteolysis (427)
40	61				64			53	66	77		45	17	29		22	15	24		carbohydrate metabolic process (259)
19	10	21		15	15	35			36	20	65	17	17	16	76	13	26			generation of precursor metabolites and energy (146)
64	30	34			43				27	42			18	25		18	10	23		amino sugar metabolic process (76)
32								59				24	19	28	90	13	15	15		RNA metabolic process (687)
	37	40										54	28	41			44	46		DNA metabolic process (154)
Reproduction																				
26	13				14				8	29	58	13	10	65		12	8	22		female gamete generation (284)
17		17	17	34							65	12	11							spermatogenesis (79)
Development																				
26	22	45			16				8			9	10	17		7	9	24		metamorphosis (253)
26	35				14							16	9	68		11	13	22		embryonic development (312)
17	6	26		16	13		61	61	6	34		7	9	16		7	8	7		organ development (753)
25	15	33		28	14			57	9			12	16	32		7	14	24		eye development (204)
	25	22										21	12	26		14	17	16		salivary gland development (112)
	50	54					75					39	14	54		13	14	63		Wnt receptor signaling pathway (61)
	35	38			19		71					8	9	44		18	26	25		wing disc morphogenesis (146)
	77												68	81			59			molting cycle process (22)
10	7	27		16	13		63	61	7	30		8	10	15		7	7	11	74	cell differentiation (685)
Transcription and translation modulation																				
	13							64	55				27	20	89		19	20	79	RNA processing (187)
24	62											18	69			14	29	73		transcription (541)
7	12	20		11	9	44		13	10	26	64	11	14	12		8	14	11		translation (264)
13	7	18		19	12			34	6	17		5	7	6	76	9	6	6		gene expression (1000)

cantly high value of ω in the rat, but not in the mouse.

An important result is also the significant differences observed between the human and chimpanzee genomes in a few neurological processes such as the KEGG pathway Ha04360: *Axon guidance* and the GO term GO0007268: *synaptic transmission.* In both cases, genes related to these functional terms are more conserved in humans.

In *Drosophila*, clade consistency in the *D. erecta/D. yakuba* group was higher. However, some interesting differences are apparent, such as: ***1)*** *"Drug metabolism - other enzymes"* that appears more conserved in *D. erecta*, or ***2)*** *Neurogenesis* and *Nervous system development* that are conserved in *D. sechellia* and in *D. melanogaster*, while without signal in *D. simulans.* Also in this case we can contrast an other difference between these conserved categories in *D. sechellia* and *D. melanogaster* where, in the case of the second, although presenting significantly low ω is significantly less conserved than its ancestor (as indicated by the high $\Delta\omega$).

5.1.3. Comparison of the evolutionary variables

The fact that most of the functional modules with SHω and SLω correlate with changes in dN suggests that selective pressures are mainly driven by non-synonymous rather than by synonymous substitutions. Moreover, according to the expectation of the nearly neutral theory, a low but still considerable number of significant associations of functional modules with 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 lower [Petit & Barbadilla 2009].

5.2. Positively selected genes within fast and slowly evolving functional modules

We have demonstrated how GSSA is able to find significant functional enrichment towards the extremities of a list of genes ranked by an evolutionary variable. However, in some way, this result contradicts the lack of significance observed when only analyzing sets of genes under positive selection. This raises the question: *To what extent do genes under positive selection contribute to the significance of functional modules in mammals and Drosophila species after GSSA?*

To answer this question, a branch-site test of positive selection (the most sensitive and widespread) was conducted on the terminal branches of both phylogenies (Figure 2.5). Overall, 715 positively selected genes (PSGs) were detected in mammals and 626 in *Drosophila.* In this last section, we examine how PSGs fit with previous results.

5.2.1. Descriptive analysis

The idea here was to describe the distribution of PSGs among the functional modules with either SH, SL or NS rates of ω. Thus, we first plotted the distribution of all functional categories (putting together the KEGG pathways and GO terms of both mammals and *Drosophila*) according to their mean values of dN and dS. Representated in this way, functional categories detected as having SHω were expected to be depicted above a line passing through the origin with a slope representative of a given cut-off value of ω ($\frac{dN}{dS}$). This is indeed appreciable in Figure 5.2-A that shows functional modules with significant and not significant results after GSSA of the ω ratio.

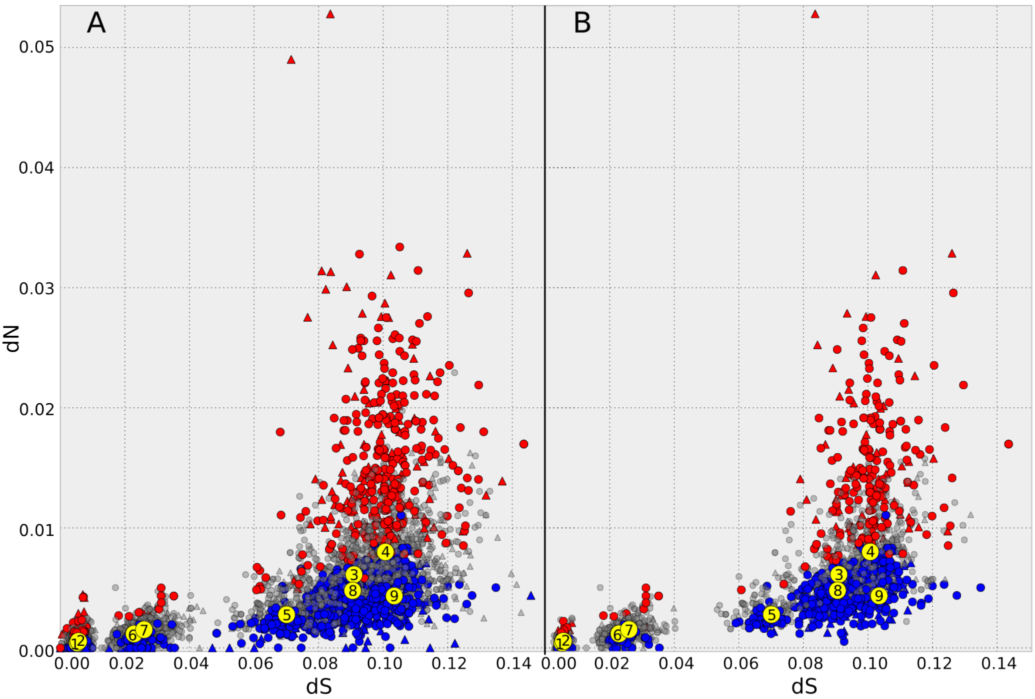


Figure 5.2.: Positive selection in the evolutionary scenarios of functional modules.
Circles and triangles represent the median values of dN and dS for functional categories in mammals and *Drosophila* species, respectively. SHω, SLω and NSω results after GSSA are indicated in red, blue and grey, respectively. Yellow dots depict the genomic median 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). (**A**) represents all functional categories, while (**B**) shows only modules containing at least 1 PSG. Note that PSGs are distributed along a wide range of dS and dN and in functional categories with significant (red/blue), and even NS (gray) results after the GSSA (ω ratio).

Biological process	Enrichment in PSGs							GSSA
	1	2	3	4	5	6	7	LOW ω rates
Sensory Perception	H	Pr†		H		Pr†		R†
Cell surface receptor-mediated signal transduction	H					Pr†		me† ya† er†
Cell adhesion	H							H‡ C‡ me‡ er†
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‡ ya† er†
Synaptic transmission			Pr					H‡ M‡ R‡ me‡ se‡ er‡ ya†
Signal transduction/intracel. signaling cascade	H C		Pr				Dr	H‡ C‡ M‡ R‡ me‡ se† ya‡ er†
Ion transport	H				H		Dr	H† M‡ R‡ me† se† er†
Potassium ion transport			Pr					H† C† M‡ R†
Inorganic anion transport			Pr					M† R†
Intracellular protein traffic	H							H‡ C‡ M‡ R‡ me† se‡ ya‡ er†
Transport							Dr	me‡ se‡ er‡ ya†
Protein transport				H			Dr	H† C‡ M‡ R‡ me‡ si† se‡ er‡ ya†
Metabolism of cyclic nucleotides	H							M† R†
Protein metabolism & modification				H C	C		Dr	H‡ C‡ M‡ R‡ er† ya†
Phosphate metabolism/phosphorylation				H C			Dr	H† C† M‡ R‡ me† se‡ ya‡ er†
Purine metabolism	C							M† R† se†
Carbohydrate biosynthesis			Pr					M‡ R†
Cation transport	H							H† M‡ R†
Nervous system development							Dr	H† M‡ R‡ me‡ se† ya‡ er†
Skeletal development	C							M‡ R†
Organ development							Dr	H† M‡ R‡ me‡ se† ya‡ er†
Post-embryonic development							Dr	M† me† ya‡ er†
Embryonic development							Dr	H‡ C† M‡ R‡ ya† er†
Ectoderm development					H			C† M† R† me† ya† er†
Cell proliferation and differentiation	C						Dr	H‡ C† M‡ R‡ me‡ se† ya‡ er†
Cell cycle							Dr	H† M† R† me‡ se‡ ya‡ er†
Cell structure/morphogenesis	C						Dr	H‡ C† M‡ R‡ me‡ se† ya‡ er†
Cell structure and motility	C							H† M‡ R‡ se†
Inhibition of apoptosis		Pr†						H† ya†
Cell-cell signalling							Dr	H‡ C† M‡ R‡ me‡ se‡ er‡ ya†
Regulation of nucleobase				H C				H‡ C‡ M‡ R‡ er†
Translation							Dr	M† R† me‡ si† se‡ ya‡ er†
Transcription				H C	C		Dr	H‡ C‡ M‡ R‡ er†
Protein catabolism				H C	C			H‡ C‡ M‡ R†
Interferon-mediated immunity		Pr†						

Biological process	**Enrichment in PSGs**							**GSSA**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**HIGH ω rates**
Olfaction/Sensory perception of smell	H	Pr†				Pr†		H‡ C‡ M‡ R‡ me† se† er‡ ya†
Chemosensory perception	H	Pr†						H‡ C‡ M‡ R‡ me‡ se‡ er‡ ya†
G-protein-mediated signaling	H				H	Pr†		H‡ C‡ R†
DNA/nucleic acid metabolism				C			Dr	C† M‡ R‡ me‡ ya‡ er†
Amino acid metabolism	H C						Dr	M‡ R†
Proteolysis							Dr	M‡ R‡ me‡ si† se† ya‡ er†
Fatty acid/Lipid metabolism					H		Dr	M‡ R†
Carbohydrate metabolism							Dr	se† ya† er†
Adult reproduction and gametogenesis							Dr	se†
Spermatogenesis and motility		Pr†	Pr					H† M† me†
Immune response		Pr†		H C		Ro†		C† M‡ R‡ ya† er†
Inflammatory response						Ro†		H† C† M‡ R†
Defense response						Ro†		H† C† M‡ R‡ ya‡ er†
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‡ ya† er†
Stress response					C	Ro†		M‡ R†
Response to external stimulus						Ro†		M‡ R†
Sensory Perception	H	Pr†		H		Pr†		H‡ C‡ M† me‡ se‡ ya‡ er†
Cell surface receptor-mediated signal transduction	H					Pr†		C†
Cell adhesion	H							R†
Amino acid transport			Pr					M†
Protein amino acid glycosylation			Pr					M†
Amino acid transport	C							M†

Table 5.2.: Functional enrichment results using gene-by-gene and gene-set approaches.

The table depicts 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 citations [Clark *et al.* 2003], [Nielsen *et al.* 2005], [Mikkelsen *et al.* 2005], [Arbiza *et al.* 2006], [Bakewell *et al.* 2007], [Kosiol *et al.* 2008] and [Clark *et al.* 2007] 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; me: *D. melanogaster*; si: *D. simulans*; se: *D sechelia*; ya: *D. yakuba*; er: *D. erecta*; Ds: *Drosophila* species.

† p=0.05; ‡ p=0.001

By combining the dataset formed by PSGs with all functional categories analyzed, the total number of functional modules containing at least one PSG represented 55%, 53%, and 42%, respectively, of the functional categories with SHω, SLω and NSω results after GSSA. Figure 5.2-B is a graphical representation of this result; it is an exact replicate of Figure 5.2-A, but only keeping functional categories that include at least one PSG.

The proportion of functional categories containing PSGs suggests that:

- The accumulation of PSGs is not the main driver in the evolution of functional modules changing at SHω ratios in the genome. Functional modules such as *"Complement and coagulation cascades"* in human, *"Gonad development"* in chimpanzee, *"Regulation of innate immune response"* in the mouse, *"Primary immunodeficiency"* in the rat or *"Spermatid differentiation"* in *D. melanogaster* are examples of functional categories evolving at significantly elevated ω values without any PSG.

- Molecular adaptation also takes place in functional modules under strong selective constraints with SLω (see the first part of Table 5.2). For instance, *"Apoptosis"* in human, *"Generation of neurons"* in chimpanzees, *"Tissue development"* in the mouse, *"Wnt signaling pathway"* in the 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 that simultaneously evolve at SLω and carry PSGs.

- A significant number of functional modules without significant differences in ω ratios (grey dots in Figure 5.2-B) still contain genes under positive selection. For instance, *"Homologous recombination"* in humans, brain *"Development"* in chimpanzee, *"Female or male sex differentiation"* in the mouse, *"Regulation of mitotic cell cycle"* in the rat, *"Chromatin modification"* in *D. sechellia*, and *"Oogenesis"* in *D. melanogaster*.

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

5.2.2. Statistical approach

Finally, we wondered if PSGs preferentially concentrate in the functional modules evolving at faster rates in different genomes. In this sense, we computed the mean number of PSGs in functional modules with SHω and SLω results (red and blue dots in Figure 5.2-B). The expectation that functional modules evolving at high

ω ratios posses higher numbers of PSGs was confirmed for rodents ($p < 0.001$), all mammals combined ($p < 0.001$) and *Drosophila* ($p < 0.001$). However, the result was far from significant when primates were considered separately ($p = 0.47$), indicating a random distribution of PSGs among functional categories with SHω and SLω. Note that, as a consequence of the larger number of PSGs in rodents, the result for primates is not sufficient to lower the reported significance of the test among mammals.

To corroborate these results, the same analysis was conducted with PSGs detected in previous works ([Kosiol *et al.* 2008] for mammals and [Clark *et al.* 2007] for *Drosophila*). The pattern of distribution of PSGs in functional modules was in exact agreement with the results described previously: there was a significant bias (p<0.001) towards higher numbers of PSGs in functional modules with high ω ratios 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, but they concentrate more frequently in functional groups of genes changing at elevated rates in rodents and *Drosophila* species. However, PSGs were evenly distributed in functional modules changing at the extreme rates of evolution in primates. This observation suggests that adaptive differences in the human and chimpanzee genomes may rely on a more complex scheme than the cumulative differences of PSGs. The search for integrative factors, taking into account the action of multiple genes other only those targeted by positive selection [He *et al.* 2010], could provide a more accurate view of the integrated framework underlying adaptation in complete genomes.

5.3. Discussion

Evolutionary biologists recognize that natural selection works on phenotypes indirectly by changing the frequency of genes in populations [Lewontin 1974]. Since the revolution of molecular techniques and its use in evolutionary genetics, the statistical search for adaptation at a gene level has superseded the complexity of measuring fitness in nature [Endler 1986]. Nowadays, to measure the influence of natural selection on phenotypes we typically look for adaptive evidences on genes and then search for over-represented functional modules enriched with the PSGs found in the genomes. This approach consists, of two independent steps, and disregards the cooperative action of the network of genes underlying phenotypes [He *et al.* 2010, Alvarez-Ponce *et al.* 2009].

The aim of the GSSA is not to test for evolutionary constraints on individual genes, as has been addressed in several previous studies. GSSA tests for significant differences in rates over functionally related groups of genes and, therefore, the relative contribution of a gene inside a functional category is dependent on the genomic distribution of evolutionary rates.

The results produced in this study completely confirm the trends observed in previous works focusing on PSGs [Arbiza *et al.* 2006, Clark *et al.* 2003, Shapiro & Alm 2008, Kosiol *et al.* 2008] (see Figure 1.5). The Observation that most of these candidate functional categories were found with significantly high values of ω can be explained by the fact that the amount of PSGs needed to approach significance in previous works certainly contributes to raising the ω value of the entire functional class. Moreover, this functionality-based approach was able to identify biological functions in individual species as the main targets of adaptive changes Table 5.2.

By defining functional modules submitted to specific selective pressures, this study represents a clear step forward in demonstrating the hypothesis that phenotypes change during evolution by the coordinated action of genes. Although GSSA is not a test for positive selection, it is evident that functional modules containing PSGs can be significantly detected by this method. Our findings fit perfectly with the results and trends reported in previous works (see Figure 1.5 and Table 5.2).

Some comment is warranted with respect to the presence of PSGs in functional categories with either significantly low or non-significant rates of ω. It is important to bear in mind that genes are not usually annotated to only one functional category. Thus, in a sense, it is not surprising to find SLω or even NSω functional categories in Figure 5.2-B. Nevertheless, the biological meaning of this multiple annotation is important. A protein involved in both a conserved and an accelerated pathway would, intuitively, be submitted to strong selective pressure. Taking this into account, the presence of PSGs in conserved functional modules can not be just a methodological artifact. Moreover, the statistical test conducted shows that, even assuming that multiple annotation would contribute to lowering its power, the tendency towards significance is beyond doubt (either in the case of significant bias toward SHω or when no significance was detected in primates). Thus, the existence of many PSGs in functional modules evolving at significantly low (or non-significant) ω ratios does not represent a false positive result in the analysis of molecular adaptation. This result, observed here and also reported in previous publications, simply suggests that PSGs are frequently recruited for purposes other than those giving rise to the classical increase in evolutionary rates of functional sets of genes involved in adaptive processes, such as evolutionary arms-races. A possible explanation is that many of the PSGs in genomes change in association with the constraints imposed by the architecture of the gene interaction network [Alvarez-Ponce *et al.* 2009], or adjust to the deleterious mutations of other genes in the network, just to maintain their phenotypic function. In this sense, it is true that adaptation requires positive selection, but the reciprocal that would suggest that every mutation under positive selection contributes to the adaptive dynamic process of species evolution is a dangerous inference [Mustonen & Lässig 2009].

Currently, with the possibility of conducting analysis at the level of the genome, evolutionary biology cannot disregard major aspects of systems biology approaches that consider the modular organization of genes. With the testing strategy presented here, we increased the statistical power for the evolutionary analysis of individual genomes and suggest that PSGs could have additional roles in the genome other than the adaptive evolutionary change of phenotypes.

6. Tools and programs

Throughout this thesis, different methods of sequence analysis, statistical inference, phylogenetic reconstruction and evolutionary hypothesis testing were implemented. Being aware of their usefulness in the fields of genomics, bioinformatics and evolution, we took advantage of these scripts or simple programs and gradually amalgamated them into documented functional packages or utilities in order to make them available for the wider community. We present this tools in this chapter.

6.1. Ecolopy

Ecolopy was originally designed to test for the unified theory of biodiversity (UNTB) [Hubbell 2001] in genomes. Other packages designed for ecologists were not able to deal efficiently with genomic data. In this section, the implementation of the package and its utility is demonstrated, giving simple examples with classical ecological datasets, such as a sample of trees from a given region.

6.1.1. Implementation

Ecolopy is a Python package specifically designed to test the UNTB over large datasets, such as the census of genetic elements in a genome. The work presented in chapter 4: **Ecology of genome elements** is a good example of its use, and of its capabilities.

With regard to the architecture of the package, Ecolopy produces two main Python objects, defined by the classes *Community* and *EcologicalModel* Table 6.1.

Class	Description	Instantiations
Community	Represents an ecosystem using the number of species in it and their abundances	Species abundance distributions Random distribution of species abundances
EcologicalModel	An ecological model able to describe an ecosystem	Model according to Ewens formula Model according to Etienne's formula Model according to Log-normal

Table 6.1.: *Community* and *EcologicalModel* classes.
Main classes of the Ecolopy package and their most useful instantiations.

Solutions for genomic data

As the "raison d'etre" of Ecolopy is to be able to deal with genomic data, some changes to its the classical implementation [Jabot & Chave 2011, Etienne 2007, Hankin 2007] were essential.

Over the models implemented, the main computational bottleneck lies in the resolution of Etienne's formula [Etienne 2005] (Equation 2.9), and particularly in the calculation of the stirling numbers in the $K(D, A)$ equation (see Equation 2.10). A first solution, given by [Jabot *et al.* 2008] and implemented in the Tetame program, consists of taking advantage of the recurrence function (Equation 6.1). This function allows the building of a table of precomputed values, thus skipping the direct computation of stirling numbers for each pair of values.

$$S_{(n,m)} = S_{(n-1,m-1)} - (n-1) \times S_{(n-1,m)} \tag{6.1}$$

However, at the expense of the improvment in computation time, the size of the table of values created was excessive in the case of genomic data. A solution was to recursively remove the stirling numbers not needed by the computation.

Finally, given the order of magnitude values of stirling numbers can reach ($> 1e^{+1000}$ for medium size chromosomes), an other adjustment was necessary. The GMP [Granlund 2000] and MPFR [Fousse *et al.* 2007] libraries (through GMPY binding [Martelli 2007]) were used in order to cope with such a sizeable dataset.

Optimization strategies

In Ecolopy, models can be optimized through different strategies depending on the model selected. In the case of Ewens' formula, θ is the only parameter taken into account, and this one-dimensional optimization is easily achieved with the "golden section search" optimization strategy [Kiefer 1953, Jones *et al.* 2001]. For Etienne's model, two parameters are optimized, θ and m. In this case, the optimization step is more sensitive to local maximums, so several optimization strategies (all implemented in SciPy [Jones *et al.* 2001]) are proposed:

- "fmin": the downhill simplex algorithm [Nelder & Mead 1965]
- "l-bfgs-b": a limited version of the Broyden, Fletcher, Goldfarb, and Shanno method for unconstrained optimization [Byrd *et al.* 1995, Zhu *et al.* 1994].
- "tnc": minimize using gradient information in a Truncated Newton Conjugate-gradient [Nocedal & Wright 2000].
- "slsqp": Sequential Least SQuares Programming [Kraft 1988].

Obviously, the best methodology consists of verifying that all optimization strategies converge. However, the use of the downhill simplex algorithm plus one of the other methodologies (using bounds) should be sufficient (if the results are found to converge). This last strategy was used in chapter 4. Note that different starting values of θ and m may also be used through the algorithm in order to ensure that the global maximum is found. Finally, a last step can be performed if computation time is not critical. This consists of drawing a likelihood surface for a given range of θ and m values (see Figure 2.2 for an example).

6.1.2. General usage

In this section, some of the features of Ecolopy are summarized, with as an example, the analysis of a sample community similar to the BCI dataset [Hubbell *et al.* 2005] referred to as the "BCI-like" dataset. The ultimate goal being to test the UNTB [Hubbell 2001].

The first step, in the analysis of a community with Ecolopy consists in creating a *Community* object, representing the data (called "com"):

```
from ecolopy import Community

abd_list = [1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,
            2,2,2,3,3,3,3,3,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,5,6,7,7,7,
            7,7,7,8,8,8,9,9,10,10,10,10,10,10,10,12,12,12,12,12,12,12,
            13,13,13,13,13,14,14,14,15,15,15,16,16,16,17,18,19,20,21,21,
            21,21,22,22,22,23,23,23,23,25,25,25,25,26,26,26,26,27,27,27,
            28,28,28,29,29,30,31,32,33,33,33,33,33,36,38,38,39,39,40,41,
            43,43,45,45,47,49,50,51,52,52,54,55,55,55,58,61,63,63,64,67,
            67,68,68,70,76,78,80,81,82,85,85,87,88,92,92,93,98,98,98,99,
            100,101,111,118,121,143,147,149,156,163,164,167,177,184,188,
            201,203,218,229,236,236,244,248,264,285,288,289,294,322,325,
            345,346,364,376,379,381,617,644,681,724,755,788,983,1681,1717]
com = Community(abd_list)
```

Descriptive analysis

Some functions are available to overview the data. The first, accessible through the "`print`" function, displays a summary about the number of species and individuals:

```
print com
```

This gives the following output:

```
  Community (object)
    Number of individuals (J) : 21457
    Number of species (S)     : 225
    Shannon entropy (shannon) : 4.2704
    Metacommunity size (j_tot): 64371
    Models computed           :
    Model loaded              : None
```

This summary shows the number of individuals (J), of species (S) and the size of the metacommunity that, by default is set to three times the size of the community. The value of Shannon's entropy, corresponding to the distribution of abundances, is also provided. Finally, this output also shows the number of models that are fitted to this community (in this case, none).

Continuing with the descriptive analysis, another interesting feature, widely used in ecology, is the relative species abundance plot (see Figure 4.2 and Figure 4.3 for an example with genomic and chromosomic data). These curves can be displayed with:

```
com.draw_rsa()
```

This generates a figure such as Figure 6.1.

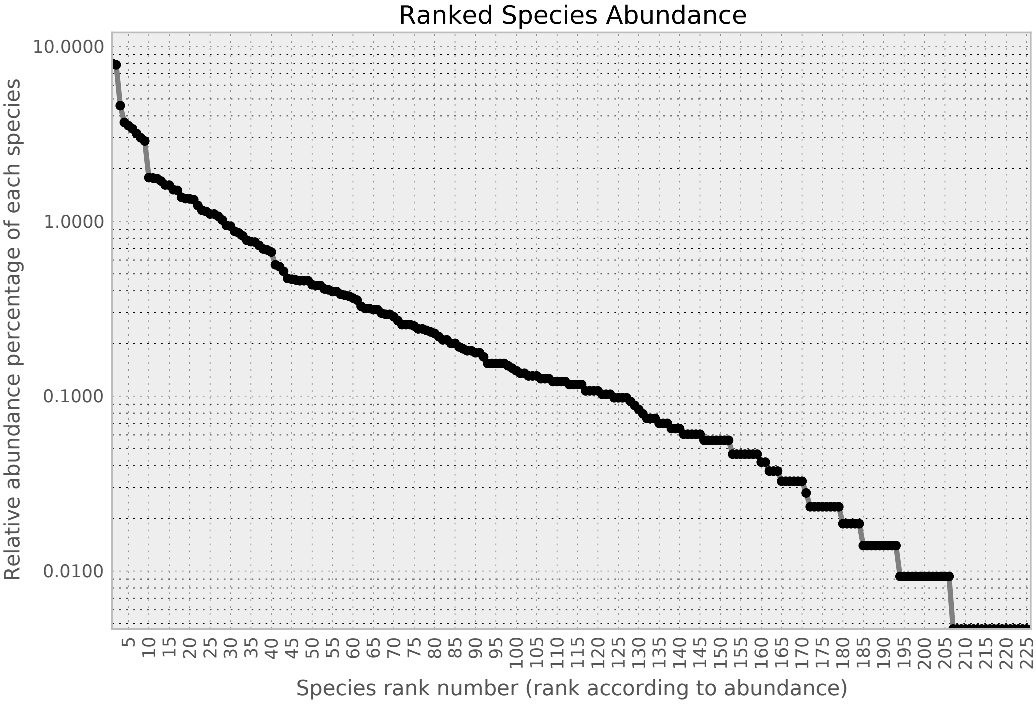


Figure 6.1.: Example of RSA with the BCI-like dataset

This is a representation of the relative species abundance (RSA) of the BCI-like dataset [Hubbell *et al.* 2005]; an output file of Ecolopy. The abscissa represents the rank of the species sorted by their abundance, while the ordinate is the percentage representation of each species according to the total number of individuals in the ecosystem (the first species on the left includes around 60% of the individuals sampled).

Ecological model optimization

Before testing for the UNTB, the ecological models should be optimized in order to get the values of θ and m corresponding to the data. To determine if the estimation of m through maximum likelihood is necessary, both, Etienne's and Ewens models to the data:

```
com.fit_model('ewens')
com.fit_model('etienne')
```

Note that these calls are the simplest (with default parameters) and as seen in the previous section, in the case of Etienne's model different optimization strategies should be used.

Likelihood ratio test

Ewens and Etienne's formulas are nested models; while the first only optimizes the value of θ, the second optimizes both, m and θ. So we can apply a likelihood ratio test (LRT) with one degree of freedom to evaluate the importance of considering the m parameter. If the resulting p-value is below 0.05, the most accurate model would be to the alternative; in this case, the one derived from Etienne's formula. Ecolopy's computation of the LRT is not automatic, as this step may be subject to posterior corrections, such as the application of FDR adjustment.

Continuing with the BCI-like dataset, now there are two models "attached" to the community. The result of a "`print`" command would now contain both "ewens" and "etienne" under "Models computed". However, there would still be no model in the "Model loaded", as it still has to be determined which fits better. With the purpose of identifying this, the LRT can be achieved as follows:

```
pval = com.lrt('ewens', 'etienne')
print pval
```

In this case, the p-value returned is $6.8e^{-06}$, so Ewens model can be rejected. Finally, Etienne's model can be loaded as the default (or current) model:

```
com.set_current_model('etienne')
print com
```

This resulting in:

```
Community (object)
    Number of individuals (J) : 21457
    Number of species (S)     : 225
    Shannon's index (shannon) : 4.2704
    Metacommunity size (j_tot): 64371
    Models computed           : ewens, etienne
    Model loaded              : EtienneModel
        I                       : 2211.10111912
        theta                       : 47.6743015824
        m                           : 0.0934250928321
        lnL                         : 308.72540670819615
```

Test UNTB

Finally, after optimization and selection of the most accurate model, a test needs to be done in order to validate that the neutral model is able to accurately predict the observed distribution of species' abundances.

Over recent years, two methodologies have been proposed in order to determine the goodness of fit of the UNTB. Both are based on a comparison between random distributions of species abundances generated according to neutral models and original data. A full description of these tests can be found in **Testing UNTB**, page 32.

Both of these methodologies are implemented in Ecolopy, and can be run using the "`test_neutrality`" function:

```
pval, neut_h  = com.test_neutrality(model='etienne',gens=10000, full=True)
```

The p-value generated in this case around 0.15, which means that we can not reject neutrality for the dataset.

Note that the function is run for 10,000 generations, each corresponding to a random neutral community generated according to the optimized Etienne's model. Also, an extra argument can be set here; the "`full`" parameter when set to "`true`" will complete the output with the entropies corresponding to each random neutral community, allowing their distribution to be plotted:

```
from ecolopy.utils import draw_shannon_distrib

draw_shannon_distrib(neut_h, com.shannon)
```

This resulting in Figure 6.2

More information about the different functions, together with a complete description of the parameters and options, as well as a quick tutorial can be found at `http://bioinfo.cipf.es/ecolopy/tutorial/load_abundance.html`. The source can be downloaded from `https://gitorious.org/ecolopy`.

6.1.3. UNTBGen Web server

With the purpose of making Ecolopy available for a wider range of students or researchers, a web server (`http://bioinfo.cipf.es/apps/untbgen/`) has been developed, that allows neutrality tests to be conducted over communities, following the steps described above.

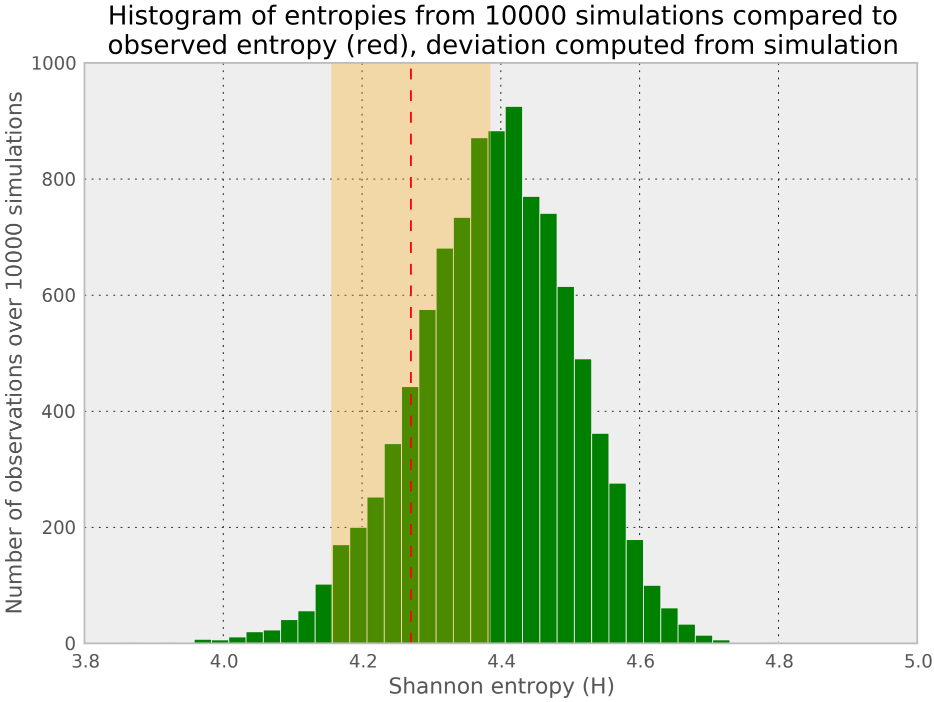


Figure 6.2.: Distribution of Shannon's entropies corresponding to simulated communities.
The histogram represents the distribution of Shannon's entropies corresponding to communities simulated under Etienne's model with optimized parameters corresponding to the BCI-like dataset. The red vertical dotted line shows the entropy of the original dataset. The orange shade is a deviation inferred from the simulated communities.

All the tests, or even descriptive analyses as mentioned earlier in this section, can be conducted in the web server (see Appendix B):

- Draw an (interactive) RSA curve corresponding to the data loaded.
- Conduct both tests of neutrality (based on the distribution of either Shannon's entropies or likelihoods corresponding to simulated random neutral communities). Additionally, users have the possibility to fix or relax the number of species in simulated communities.
- Compute a LRT between Ewens and Etienne's models.
- Draw a contour plot of the likelihood surface according to a range of θ and m values.

Besides the user-friendly aspect the web server offers, as an additional advantage, data and results can be stored in the user's account.

6.2. The “Evol” extension

When testing for evolutionary scenarios, a vast range of models can be used through diferent softwares (see A, for a description of different models and softwares), however it is important to note that these models are standards, and other programs or packages offer different implementations protocols to carry out similar tests [Knight *et al.* 2007, Pond *et al.* 2005].

This section is dedicated to what can be considered either a genomic solution or, at least, a simplification of the methodology for the determination of selective pressures and the test of evolutionary hypotheses. This solution uses ETE [Huerta-Cepas *et al.* 2010] and, more particularly, the “Evol” extension that allows CodeML and SLR programs to be run. The results of such analyses are embedded in ETE’s Tree objects and, from there, can be contrasted (evolutionary model comparison), visualized or summarized.

6.2.1. Implementation

ETE is a Python package originally designed for the manipulation, analysis and visualization of phylogenetic trees. Of the most general Tree class, the best may be PhyloTree (from the “Phylo” extension), which implements specific algorithms to deal with phylogenetic trees. Amongst its most useful functions are the possibility to: ***1)*** link a tree to an alignment, ***2)*** infer evolutionary events (speciation or duplication) through different algorithms, ***3)*** relatively date nodes of a tree, and ***4)*** automatically root a gene tree (according to a given species tree).

In order to take advantage of these functions, and also for coherence, the Evol extension was implemented as a specific case of the Phylo extension. Or, in more computational language, the EvolTree class inherits from the PhyloTree class (note that in the same way, PhyloTree inherits from the main Tree class).

The Evol extension contains two main classes: the EvolTree that inherits from the PhyloTree classes and the Model classes that represents a given evolutionary model (for now, it can be either one of the models proposed by CodeML or by SLR).

The Evol extension is currently available as a branch of ETE at `https://github.com/jhcepas/ete/tree/evoltree`. Some documentation can be found at `http://bioinfo.cipf.es/fransua/ete-evol/tutorial/tutorial_adaptation.html`.

6.2.2. General usage

In this section a quick overview the different functionalities of the Evol extension is given.

Evol trees and alignments can be loaded in the same way as PhyloTrees. Below is a short example of how to load a tree, together with its alignment, and run the free-ratio model:

```
from ete2 import EvolTree

tree = EvolTree("(Orangutan,Human,Chimp);")
tree.link_to_alignment("""
>Chimp
CCC GCA CGA TGG CTC AAT GTA AAG TTA AGA TGC GAA TTG AGA ACA CTA AAA AAA
TTG GGA CTG GAC GGC TAC AAG GCA GTA AGT CAA TAC GTT AAA GGT CGT GCG ATT
>Orangutan
GAT GCA CGA TGG ATC AAT CCA AAG TTA AGA TGC GAA TTG AGA ACT CTG AAA AAA
TTG GGA CTG GAC GGC TAC AAG GCA GCA AGT CAA TAC GTT AAA GGT CGT AGC TCT
>Human
TAC GCA CGA TGG CTC AAC GTA AAA TTA AGA TGT GAA TTA AGG ACG CTC AAA AAA
TTG GGA CTG GAC GGC TAC AAG GCA GTA AGT CAA TAC GTT CAA GGT CGT GCC AGT
""")
tree.run_model("fb")
tree.show()
```

The "show" command here works as for PhyloTrees but, additionally, it displays a summary of the selective pressures acting on branches (colored circles appearing at each node in Figure 6.3, for which colors and sizes are a function of the corresponding values of ω estimated by the model).

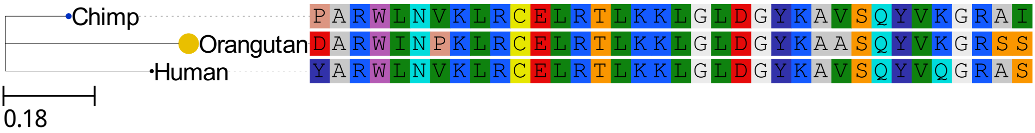

Chimp
Orangutan
Human
0.18
P A R W L N V K L R C E L R T L K K L G L D G Y K A V S Q Y V K G R A I
D A R W I N P K L R C E L R T L K K L G L D G Y K A A S Q Y V K G R S S
Y A R W L N V K L R C E L R T L K K L G L D G Y K A V S Q Y V Q G R A S

Figure 6.3.: Sample representation of the free-ratio model in an EvolTree. Default representation of an EvolTree. Node sizes and colors are functions of the values of ω for the given branch. For this example, the values found for the sequences under the free-ratio model are, Orangutan $\omega = 1.29$, Chimp $\omega = 0.28$ and Human $\omega = 0.08$.

Although the example above may be "too simple" when compared to the number of parameters that have to been set when directly using the CodeML program, the extension allows each one of the proposed CodeML parameters to be modified. As an example, below is the list of default parameters set in the Evol extension for the free-ratio model:

```
      aaDist = 0       |         icode = 0    |          noisy = 0
   fix_alpha = 1       |     fix_kappa = 0    |        NSsites = 0
       alpha = 0.0     |         kappa = 2    |      fix_omega = 0
 fix_blength = 0       |        Malpha = 0    |          omega = 0.7
   cleandata = 0       |        method = 0    |   RateAncestor = 0
       clock = 0       |         Mgene = 0    |        runmode = 0
   CodonFreq = 2       |         model = 1    |        seqtype = 1
       getSE = 0       |         ncatG = 8    |     Small_Diff = 1e-6
```

Changing one of these parameters can easily be done. Suppose a different starting value for the optimization of ω needs to be set (as shown in previous box, by default it is set to 0.7):

```
tree.run("fb", omega=1.2)
```

In the context of site analysis, the Evol extension also has some features in order to visually summarize the shape of selective pressure along sites. In order to evaluate the selective pressures among sites, the positive selection model needs to be validated as being the model with the best fit:

```
tree.run_model("M2")
tree.run_model("M1")
tree.get_most_likely("M2", "M1")
```

Continuing with the example, the result of the "`get_most_likely`" (i.e. the p-value of the LRT between the models M1a and M2a) is below 0.05 (p=0.014). Thus, "M2a" has a better fit, and we can accept the assumption that some sites are evolving at $\omega > 1$. As previously mentioned, another advantage of the Evol extension lies in the possibility of getting an overview of the ω shape varying along the alignment and according to a given evolutionary model (see upper plot in Figure 6.4).

In order to confirm the positively-selected sites detected by the CodeML program, a good option is to use another methodology such as the site-wise likelihood-ratio implemented in the SLR program. Using the Evol extension, running the SLR program, and displaying a summary result of both the M2a model of CodeML and that of SLR generates something like this output:

```
tree.run_model("SLR")
tree.show (histfaces=["M2", "SLR"])
```

Thus, according to the p-value of the LRT between models M2a and M1a, and to Figure 6.4, it can be deduced that for the sample alignment two sites are under positive selection, of which one is confirmed by the SLR program.

More complex branch models can also easily be defined in order to test different evolutionary hypotheses. For example, to test for significant differences in ω rates between branches, the tree has to be labeled and specific branch models has to be computed, as explained in appendix section: **Branch models** page 137.

Each node of an Evol tree is labeled with an ID that corresponds to the identifier given by CodeML (namely the "`node_id`"). These labels are useful for labeling a tree. Labels corresponds to CodeML convention, that is a hash symbol (#) followed by a number. To test the hypothesis that the orangutan sequence is evolving at a significantly higher rate than that of the human and chimp together:

```
tree.mark_tree ([(tree & 'Orangutan').node_id], marks=['#1'])
tree.write()
```

The "`write`" command is used here to check that the tree has actually been labelled (for this example it should be: "`(Chimp,Orangutan #1,Human);`"). The

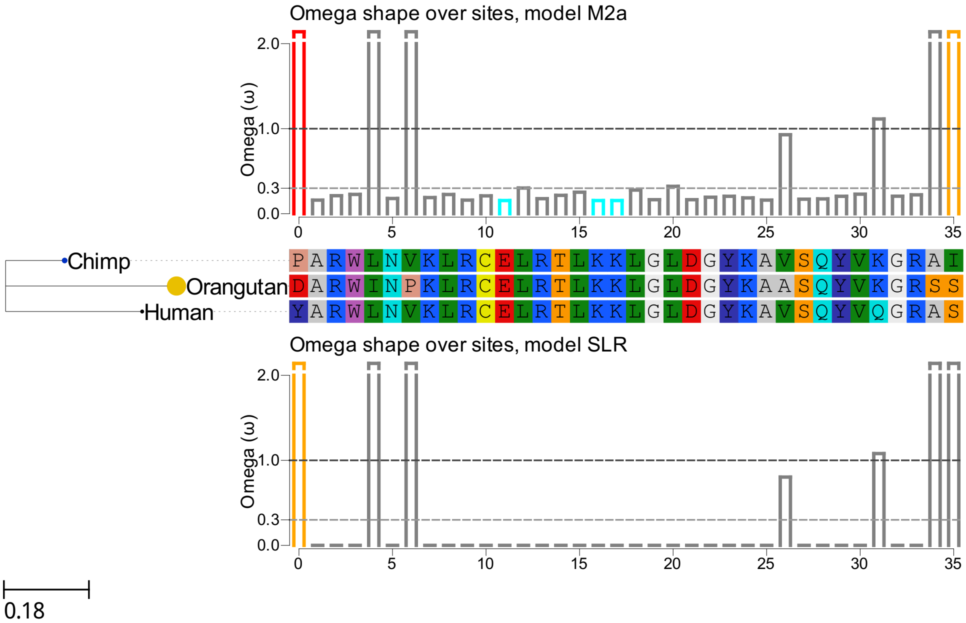


Figure 6.4.: Representation of model M2a with the CodeML and SLR results. Same representation as in Figure 6.3, with two additional bar plots representing the values of ω for each site of the alignment. The upper plot represents the values of ω computed under the M2a model with the CodeML program, while the lower plot represents the values of ω calculated by the SLR program. Colors of the bars represent the significance of belonging to a given class of sites [**red**: class of sites with $\omega > 1$ (probability>0.99), **orange**: class of sites with $\omega > 1$ (probability>0.95), **blue/cyan**: class of sites with $\omega < 1$ (probability>0.95)].

next step consists of fitting this labeled tree to a branch model, and comparing it to a null model (such as the M0 model where all branches have the same value of ω):

```
tree.run_model("b_free")
tree.mark_tree ([(tree & 'Orangutan').node_id], marks=[''])
tree.run_model("M0")

tree.get_most_likely("b_free", "M0")
```

In this example, the p-value of this LRT would be below 0.05 (p=0.021), thus orangutan sequence seems to be evolving at a different rate than that of the other species. In order to quickly see the values estimated for each part of the tree, a summary can be displayed by printing (through the "print" function) the branch model:

```
print tree.get_evol_model("b_free")
```

This results in the following output:

```
 Evolutionary Model b_free:
      log likelihood       : -244.804192
      number of parameters : 6
      sites inference      : None
      sites classes        : None
      branches             :
         mark: #1  , omega: 1.39095890112, nodes paml_ids: 3
         mark: #0  , omega: 0.123723533087, nodes paml_ids: 1 2
```

According to this summary, and to the result of the LRT, branches marked with "#1" have a significantly higher value of ω.

Finally, the branch-site test (usually considered the ultimate test of positive selection) can be computed in a similar way. In this case, the tree also needs to be labeled in order to differentiate which branches should be considered as foreground or background branches in branch-site model A (bsA) and branch-site model A1 (bsA1) (see section A.4.2). Lastly, a LRT can be performed between these optimized models. As an example, the branch-site test (Test II, see appendix section: **Testing for evolutionary scenarios in protein-coding genes** for details) can be applied in order to see if some sites of the orangutan sequence are under positive selection.

```
tree.mark_tree ([(tree & 'Orangutan').node_id], marks=['#1'])
tree.run_model("bsA")
tree.run_model("bsA1")
tree.get_most_likely("bsA", "bsA1")
```

The p-value of this LRT is 0.005, demonstrating the presence of sites evolving at $\omega > 1$ specifically in the orangutan sequence.

6.3. Phylemon2.0

More than a side-product Phylemon arose from the need of evolutionary biologists to easily run, store and compare computational analyses on their data. However, in its last version [Sánchez *et al.* 2011], some improvements were introduced based on, ***1)*** the feedback from users and colleagues, ***2)*** the advent of new tools in the fields of phylogenetics and the study of adaptation, ***3)*** experience of the entire protocol leading to the detection of selective pressures in protein-coding genes.

As with previous the version, Phylemon2.0 is organized into sections that, in general, adopt the protocol for the detection of selective pressures in protein-coding genes described above (section: **Testing for evolutionary scenarios in protein-coding genes**, page 134). In this final part of the thesis, we quickly to review some of the advances in this new version of Phylemon.

6.3.1. Alignment

The main improvements in Phylemon2.0 for aligning sequences basically consists of the addition of the Lagan and M-Lagan tools [Brudno *et al.* 2003] particularly useful when dealing with long genomic sequences.

Also, an effort was made to accommodate the recommendations mentioned in the appendix section: **The alignment**, page 133. In order to ease the execution of this step, a new version of the CDS-protAl tool has been made available, completely re-written in order to be more efficient and stable and to offer more options to users. Through the user-friendly web interface, users can now align coding sequences according to their translated amino-acid sequences. It is important to note that the options proposed in Phylemon2.0 are few in comparison of what the program can do through the command line. CDS-protAl is able to apply the full methodology described below, allowing alignment of a given set of sequences with different tools, merging of resulting alignments, and finally cleaning them using TrimAl. It is available at `https://gitorious.org/aligner`.

6.3.2. Model selection

In the context of model testing, Phylemon2.0 makes several tools available to users; the most famous being ModelTest's [Posada & Crandall 1998] direct successors, jModelTest [Posada 2008] for nucleotide sequences and ProtTest [Abascal *et al.* 2005] in the case of amino-acids. Additionally a new tool, PhyML-Best-AIC-tree, has been introduced for the particular needs of genomic studies. PhyML-Best-AIC-tree is basically a simplification of jModelTest and ProtTest combined; it can deal either with nucleotide or amino acid sequences, but only computes AIC scores. It is specifically designed for integration into a pipeline. As its main feature, it has the option to search for the best substitution model in a "clever" way. As it is described the appendix section: **Model testing and phylogenetic inference** page 133, a good approximation in the search for the best substitution model, is to test the fit of each model over a fixed tree. PhyML-Best-AIC-tree can be used in order to compute this fast approximation first for all models and, a second time, it can run a more precise analysis that involves the optimization of the tree topology but only for the models summing a given weight. The weight of a model being defined as in jModelTest or ProtTest. As for CDS-protAl, more options are available through the command line. Moreover, in the case of PhyML-Best-AIC-tree, special attention was paid to easing its integration into a pipeline, also all functions can be called independently from another Python program, thereby avoiding the use of the command line. The program source code is available at `https://github.com/fransua/pmodeltest`

6.3.3. Phylogeny

Here, the main advances were made for PhyML [Guindon & Gascuel 2003] and MrBayes [Ronquist & Huelsenbeck 2003] and consisted mainly in an extension of the forms with more options to run these tools and a better integration of their output files in the Phylemon framework. MrBayes, in particular, now has the option to be run non-interactively, and also to build a command block through Phylemon's form. Output trees can now be browsed directly either in ETE [Huerta-Cepas *et al.* 2010] or in Archaeopteryx [Zmasek 2012].

6.3.4. The Pipeliner

This is perhaps one of the greatest improvements in this new release of Phylemon. Phylemon's Pipeliner is a tool designed to enable users to develop their own pipelines in a friendly modular environment for running multiple gene analyses. The pipeline covers, with a selection of tools, all the steps needed to transform a given set of sequences into one, or several (as it accepts multiple input files), accurate phylogenetic trees. Moreover, any pipeline created by placing and linking different tools on the Pipeliner "playground" can be saved for further analysis, or even exchanged with a collaborator.

6.4. Discussion

In this last chapter the main contributions, in terms of bioinformatic tools, derived from the work presented in this thesis were reviewed. Firstly, in the context of the characterization of the dynamics of genetic elements, the program Ecolopy was presented, and second, in the fields of comparative and evolutionary genomics, the Evol extension and the Phylemon web server.

These tools were implemented based on their possible relevance to the scientific community, and thus a constant effort was made in three aspects:

- each of the tools presented in this chapter are related to extensive documentation with comprehensive tutorial. In Phylemon, a full help section is available, since version 2.0, describing all the tools and proposing simple exercises in order to better understand the possibilities and needs of each of them. For the ETE Evol extension, which is now fully integrated into ETE, and Ecolopy, extended documentation with specific tutorials is also available and, moreover, its source code is fully commented in order to ease the implementation of future extensions and for it to be called from other programs.

- in terms of scalability, and future growth or integration into other programs, but also regarding the amount of input data.

- the creation of web servers in order to offer solutions to two kinds of problems: first, the fact that with the development of genomic data, bioinformatic analyses involve a growing need for computing resources that may not be compatible with personal computers and second, the intention to propose user-friendly tools for students and researchers with less computer skills.

The choice of the programming language used for the implementation of these tools also reflects the desire to facilitate their reusability. Python [Van Rossum & Drake 2003] was chosen as a result of a balanced decision between these four major points: ***1)*** the interaction with packages implemented in R [Team 2011] – Python allows the R function to be called easily [Moreira & Warnes 2004], ***2)*** computation speed versus flexibility – Python a good compromise between the performance of C or java and the flexibility of R or perl [Fourment & Gillings 2008], ***3)*** interactivity – Python shell, asfor the R shell, allows an analysis to be built line by line, avoding the need to write a complete working script, and lastly ***4)*** its increasing popularity among the bioinformatics community [Bassi 2007] .

Finally, we would like to emphasize some contributions that these tools have brought about specific studies:

- In the context of testing evolutionary hypothesis with the Evol extension:
  - studying the strength of selection in a gene family [Lavagnino *et al.* 2012].
  - contrasting selective pressure after gene duplication, through branch models (see section A.4.2) [Martín-Trillo *et al.* 2011].
  - in the study of selective pressures in protamines of rodents and primates [Lüke *et al.* 2011, Serra *et al.* 2012b].
  - the work presented in the fourth chapter of this thesis, published in 2011 [Serra *et al.* 2011].
- In the study of the evolutionary history of genes through phylogenetic reconstruction with tools in Phylemon 2.0 (CDS-ProtAl and PhyML-Best-AIC-tree) [Gonçalves *et al.* 2011].
- In the study of the distribution and abundance of different families of genetic elements with Ecolopy [Serra *et al.* 2012a].

7. Conclusions

1. In the whole diversity of life, from viruses to mammals, informational content of the genomes exhibits quasi-maximum values. Only dramatic changes such as polyploidization events, or strong biases in nucleotide contents, are able to lower genome entropy.

2. According to the observed universal adjustment of genomes to maximum complexity, we hypothesize that increases in biological complexity are the consequence of genome expansions events through duplications or polyploidization.

3. Similarly to the biological species in ecosystems, eukaryotic genomes present an heterogeneous distribution of families or "species" of genetic elements: a few are very abundant, others quite frequent, and the majority rare.

4. Likewise for ecological species-area correlation graph, we observe that, along a great diversity of eukaryote genomes, the total number of genetic species in chromosomes is proportional to chromosome length.

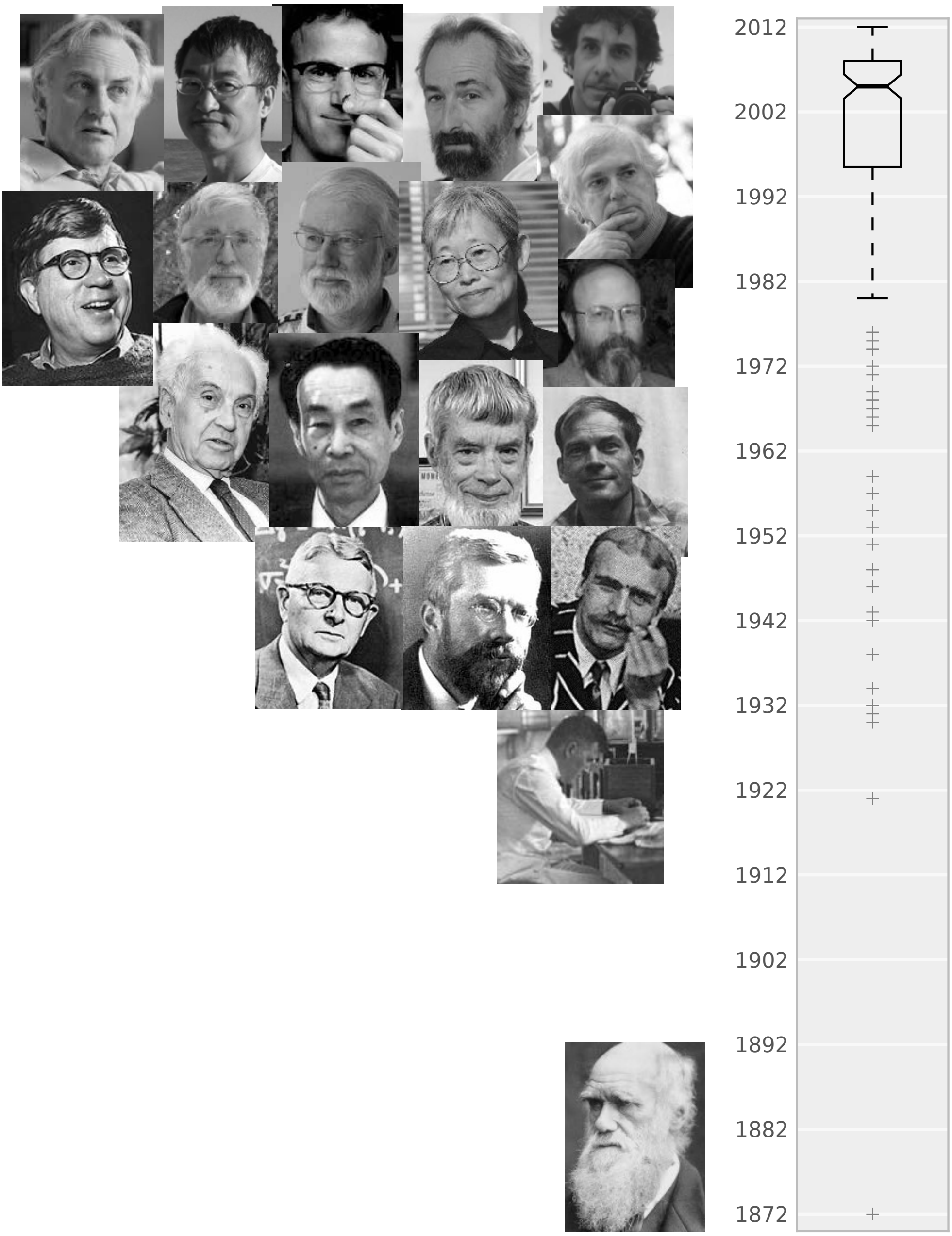
5. The distributions and abundances of families of genetic elements in eukaryotic genomes, either functional or repetitive, follow the expectations of the unified neutral theory of biodiversity (UNTB).

6. GSSA allows testing for functional biases within fast or slowly-evolving genes. This methodology successfully identified all previously reported candidate functional categories as important targets of natural selection. Moreover, given that the GSSA is not limited by the compulsory presence of positively-selected genes, it extended the list of phenotypical targets to previously undetectable ones.

7. Genes under positive selection were found to be present in functional categories evolving rapidly, slowly or without a significant trend (ω). However a significant bias towards fast evolving categories was found in rodents and *Drosophila*. Regarding to the even distribution found in primates, we hypothesize that it may be the result of small population sizes limiting the influence of natural selection; just as suggested by the theory of slightly deleterious mutations.

8. We believe that the role of genes under positive selection not only consists of the adaptive evolutionary changes of phenotypes but may also be related to other processes such as the adjustment to the deleterious mutations of other genes in a given network.

9. Throughout this thesis three primary bioinformatics tools were implemented with the idea of facilitating and extending future researches of the scientific community. These software are in line with the fields of ecology (of genomes), phylogeny, phylogenomics and the testing of evolutionary hypotheses.


2012
2002
1992
1982
1972
1962
1952
1942
1932
1922
1912
1902
1892
1882
1872

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. ↪ *pages 20 and 101*

[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 12*

[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 8 and 23*

[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 16 and 39*

[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 16*

[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 16 and 39*

[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 3, 12, and 72*

[Alvarez-Ponce *et al.* 2009] David Alvarez-Ponce, Montserrat Aguadé, and Julio Rozas, Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. *Genome Research* **19**(2) (2009), 234–42. ↪ *pages 85 and 86*

[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 14, 15, 18, 39, 83, and 86*

[Arrhenius 1921] Olof Arrhenius, Species and area. *Journal of Ecology* **9**(1) (1921), 95–99. ↪ *page 70*

[Azbel' 1995] Mark Ya Azbel', Universality in a DNA statistical structure. *Physical review letters* **75**(1) (1995), 168–171. ↪ *page 8*

[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. ↪ *pages 14, 15, 18, and 83*

[Bartolomé *et al.* 2002] Carolina Bartolomé, Xulio Maside, and Brian Charlesworth, On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Molecular Biology and Evolution* **19**(6) (2002), 926–37. ↪ *page 69*

[Bassi 2007] Sebastian Bassi, *Python for Bioinformatics*. CHAPMAN & HALL/CRC Mathematical and Computational Biology Series, 2007. ↪ *page 103*

[Becher & Heiber 2011] Verónica Becher and Pablo Ariel Heiber, On extending de Bruijn sequences. *Information Processing Letters* **111**(18) (2011), 930–932. ↪ *page 59*

[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 33 and 39*

[Blair Hedges & Kumar 2003] S Blair Hedges and Sudhir Kumar, Genomic clocks and evolutionary timescales. *Trends in Genetics* **19**(4) (2003), 200–6. ↪ *page 37*

[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 55*

[Borcard *et al.* 2011] Daniel Borcard, Francois Gillet, and Pierre Legendre, *Numerical Ecology with R*. Springer, 2011. ↪ *page 18*

[Brudno *et al.* 2003] Michael Brudno, Chuong B Do, Gregory M Cooper, Michael F Kim, Eugene Davydov, Eric D Green, Arend Sidow, and Serafim Batzoglou, LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**(4) (2003), 721–31. ↪ *page 101*

[de Bruijn 1946] N G de Bruijn, A combinatorial problem. *Koninklijke Netherlands: Academe Van Wetenschappen* **49** (1946), 758–764. ↪ *page 59*

[Burrows & Wheeler 1994] Michael Burrows and David J Wheeler, A block-sorting lossless data compression algorithm. *Digital SRC Research Report* **124** (1994). ↪ *page 23*

[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. ↪ *pages 14 and 18*

[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 90*

[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 20 and 37*

[Caswell 1976] Hal Caswell, Community Structure: A Neutral Model Analysis. *Ecological Monographs* **46**(3) (1976), p. 327. ↪ *page 3*

[Cavalier-Smith 2005] Thomas Cavalier-Smith, Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Annals of Botany* **95**(1) (2005), 147–75. ↪ *page 74*

[Cavalier-Smith 2006] Thomas Cavalier-Smith, Rooting the tree of life by transition analyses. *Biology Direct* **1**(1) (2006), p. 19. ↪ *page 5*

[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 61*

[de Chardin 1955] Pierre Teilhard de Chardin, *The Phenomenon of Man.* William Collins Sons & Co. Ltd., London, 1955. ↪ *page 3*

[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. ↪ *pages 14, 15, 18, 83, and 86*

[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 16, 17, 39, 83, and 85*

[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 39*

[Darwin 1872] Charles R Darwin, *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life.* John Murray, London, 6th edition, 1872. ↪ *pages 1 and 11*

[Dawkin 1976] Richard Dawkin, *The selfish gene.* Oxford University Press, USA, 1976. ↪ *page 12*

[Delport *et al.* 2010] Wayne Delport, Art F Y Poon, Simon D W Frost, and Sergei L Kosakovsky Pond, Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics (Oxford, England)* **26**(19) (2010), 2455–7. ↪ *page 20*

[Dereeper *et al.* 2008] A Dereeper, V Guignon, Guillaume Blanc, S Audic, S Buffet, F Chevenet, J-F Dufayard, S Guindon, V Lefort, M Lescot, J-M Claverie, and O Gascuel, Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**(Web Server issue) (2008), W465–9. ↪ *page 20*

[Doolittle & Sapienza 1980] W Ford Doolittle and C Sapienza, Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**(5757) (1980), 601–3. ↪ *page 12*

[Doolittle *et al.* 2011] W Ford Doolittle, Julius Lukeš, John M Archibald, Patrick J Keeling, and Michael W Gray, Comment on "Does constructive neutral evolution play an important role in the origin of cellular complexity?" DOI 10.1002/bies.201100010. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **33**(6) (2011), 427–9. ↪ *page 5*

[Dopazo] Joaquín Dopazo, Formulating and testing hypotheses in functional genomics. *Artificial Intelligence in Medicine* **45**(2-3) , 97–107. ↪ *page 39*

[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 54*

[Edgar 2004] Robert C Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**(5) (2004), 1792–7. ↪ *pages 19 and 35*

[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 30*

[Endler 1986] J A Endler, *Natural selection in the wild.* Princeton University Press, 1986. ↪ *page 85*

[Etienne 2005] Rampal S Etienne, A new sampling formula for neutral biodiversity. *Ecology Letters* **8**(3) (2005), 253–260. ↪ *pages 4, 18, 31, and 90*

[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 32 and 90*

[Ewens 1972] Waren J Ewens, The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**(1) (1972), 87–112. ↪ *pages 3 and 29*

[Fisher 1930] Ronald A Fisher, *The genetical theory of natural selection.* Oxford University Press, 1930. ↪ *page 2*

[Fisher *et al.* 1943] Ronald A Fisher, A Steven 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. ↪ *pages 2 and 11*

[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 10, 19, 25, 28, 30, 35, and 58*

[Fourment & Gillings 2008] Mathieu Fourment and Michael R Gillings, A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics* **9** (2008), p. 82. ↪ *page 103*

[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. ↪ *page 90*

[Frank 2012] Steven A Frank, Wright's adaptive landscape versus Fisher ' s fundamental theorem. In *The Adaptive Landscape in Evolutionary Biology*, ed. Erik I Svensson and Ryan Calsbeek, 1–34. Oxford University Press, 2012. ↪ *page 2*

[Friz 1968] Carl T Friz, The biochemical composition of the free-living Amoebae Chaos chaos, Amoeba dubia and Amoeba proteus. *Comparative Biochemistry and Physiology* **26**(1) (1968), 81–90. ↪ *page 5*

[Fujita *et al.* 2011] Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R Dreszer, Belinda M Giardine, Rachel A Harte, Jennifer Hillman-Jackson, et al., The UCSC Genome Browser database: update 2011. *Nucleic Acids Research* **39**(Database issue) (2011), D876–82. ↪ *page 19*

[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 D 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*

[Gause 1934] GF Gause, *The struggle for existence.* Williams & Wilkins Company, Baltimore, 1934. ↪ *page 3*

[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 54*

[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 54*

[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 54*

[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 47*

[Gojobori 1983] T. Gojobori, Codon substitution in evolution and the "saturation" of synonymous changes. *Genetics* **105**(4) (1983), p. 1011. ↪ *page 19*

[Gonçalves *et al.* 2011] Luís G Gonçalves, Nuno Borges, François Serra, Pedro L Fernandes, Hernán Dopazo, and Helena Santos, Evolution of the biosynthesis of di-myo-inositol phosphate, a marker of adaptation to hot marine environments. *Environmental Microbiology* **In Press** (2011). ↪ *page 103*

[Granlund 2000] Torbjörn Granlund, GMP: The GNU Multiple Precision Arithmetic Library, http://gmplib.org/, 2000. ↪ *page 90*

[Graur & Li 2000] Dan Graur and Wen-Hsiung Li, *Fundamentals of Molecular Evolution.* Sinauer Associates, Inc., Sunderland, Massachusetts, USA, second edi edition, 2000. ↪ *page 37*

[Gray *et al.* 2010] Michael W Gray, Julius Lukes, John M Archibald, Patrick J Keeling, and W Ford Doolittle, Cell biology. Irremediable complexity? *Science (New York, N.Y.)* **330**(6006) (2010), 920–1. ↪ *page 5*

[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 5*

[Gregory 2005] T Ryan Gregory, Synergy between sequence and size in large-scale genomics. *Nature Reviews. Genetics* **6**(9) (2005), 699–708. ↪ *pages 5, 7, and 8*

[Gregory 2012] T Ryan Gregory, Animal Genome Size Database, http://www.genomesize.com, 2012. ↪ *page 5*

[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 102*

[Haegeman & Weitz 2012] Bart Haegeman and Joshua S Weitz, A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* **13**(1) (2012), p. 196. ↪ *page 73*

[Haldane 1957] John Burdon Sanderson Haldane, The cost of natural selection. *Journal of Genetics* **55**(3) (1957), 511–524. ↪ *page 2*

[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 18 and 90*

[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 85*

[Hershberg & Petrov 2008] Ruth Hershberg and Dmitri a Petrov, Selection on codon bias. *Annual Review of Genetics* **42**(iv) (2008), 287–99. ↪ *page 15*

[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 9*

[Hu *et al.* 2011] Tina T Hu, Pedro Pattyn, Erica G Bakker, Jun Cao, Jan-Fang Cheng, Richard M Clark, Noah Fahlgren, Jeffrey a Fawcett, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Jesse D Hollister, Stephan

Ossowski, Robert P Ottilar, Asaf a Salamov, et al., The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics* **43**(5) (2011), 476–81. ↪ *pages 52, 58, and 62*

[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 39*

[Hubbell 2001] Stephen P Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001. ↪ *pages 3, 4, 11, 18, 31, 67, 72, 89, and 91*

[Hubbell *et al.* 2005] Stephen P Hubbell, R Condit, and R B Foster, Barro Colorado Forest Census Plot Data, `https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci`, 2005. ↪ *pages 91 and 92*

[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. ↪ *pages 20, 37, 96, and 102*

[Huerta-Cepas *et al.* 2012] Jaime Huerta-Cepas, Marina Marcet-Houben, and Toni Gabaldón, Scalable resolution in a growing eukaryotic Tree Of Life. *Submitted* (2012). ↪ *page 67*

[Hutchinson 1959] G Evelyn Hutchinson, Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist* (1959). ↪ *page 3*

[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 39*

[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 12, 19, 32, 37, and 90*

[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 19 and 90*

[Jones *et al.* 2001] Eric Jones, Travis E. Oliphant, and Pearu Peterson, Scipy: Open source scientific tools for Python, `http://www.scipy.org/`, 2001. ↪ *page 90*

[Jurka *et al.* 2005] Jerzy Jurka, Vladimir 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 25*

[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. ↪ *pages 10, 13, and 28*

[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 20*

[Keller 1999] Laurent Keller, Intragenomic Conflict. In *Levels of selection in evolution*, chapter 7, 121–152. Monographs in Behavior and Ecology, 1999. ↪ *page 10*

[Kiefer 1953] Jack Carl Kiefer, Sequential Minimax Search for a Maximum. *Proceedings of the American Mathematical Society* **4**(3) (1953), p. 502. ↪ *page 90*

[Kimura 1968] Motoo Kimura, Evolutionary Rate at the Molecular Level. *Nature* **217**(5129) (1968), 624–626. ↪ *page 2*

[Kimura 1985] Motoo Kimura, *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, UK, 1985. ↪ *pages 4, 11, 14, and 16*

[King & Jukes 1969] Jack Lester King and Thomas H Jukes, Non-Darwinian evolution. *Science (New York, N.Y.)* **164**(3881) (1969), 788–98. ↪ *page 2*

[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 28 and 35*

[Knight *et al.* 2007] Rob Knight, Peter Maxwell, Amanda Birmingham, Jason Carnes, J Gregory Caporaso, Brett C Easton, Michael Eaton, Micah Hamady, Helen Lindsay, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Michael Robeson, Raymond Sammut, Sandra Smit, et al., PyCogent: a toolkit for making sense from sequence. *Genome Biology* **8**(8) (2007), p. R171. ↪ *page 96*

[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 16, 17, 39, 77, 83, 85, and 86*

[Kraft 1988] Dieter Kraft, A software package for sequential quadratic programming. *Forschungsbericht / Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt* **88**(28) (1988), p. 33. ↪ *page 90*

[Lander *et al.* 2001] Eric S Lander, L M Linton, Bruce W Birren, C Nusbaum, Michael C Zody, Jennifer 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. ↪ *page 9*

[Lavagnino *et al.* 2012] Nicolás Lavagnino, François Serra, Leonardo Arbiza, Hernán Dopazo, and Esteban Hasson, Evolutionary Genomics of Genes Involved in Olfactory Behavior in the Drosophila melanogaster Species Group. *Evolutionary Bioinformatics* **8** (2012), 89–104. ↪ *page 103*

[Le Rouzic *et al.* 2007a] 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) (2007a), 19375–80. ↪ *page 12*

[Le Rouzic *et al.* 2007b] Arnaud Le Rouzic, Stéphane Dupas, and Pierre Capy, Genome ecosystem and transposable elements species. *Gene* **390**(1-2) (2007b), 214–20. ↪ *pages 12 and 13*

[Leigh 2007] E G Leigh, Neutral theory: a historical perspective. *Journal of Evolutionary Biology* **20**(6) (2007), 2075–91. ↪ *page 2*

[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 12*

[Lewontin 1974] RC Lewontin, *The genetic basis of evolutionary change.* New York: London, Columbia University Press, 1974. ↪ *page 85*

[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. ↪ *page 8*

[Lüke *et al.* 2011] Lena Lüke, Alberto Vicens, François Serra, Juan José Luque-Larena, Hernán Dopazo, Eduardo R. S. Roldan, and Montserrat Gomendio, Sexual Selection Halts the Relaxation of Protamine 2 among Rodents. *PloS One* **6**(12) (2011), p. e29247. ↪ *page 103*

[Lynch 2000] Michael Lynch, The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290**(5494) (2000), 1151–1155. ↪ *page 62*

[Lynch 2007] Michael Lynch, *The Origins of Genome Architecture.* Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 2007. ↪ *page 76*

[Lynch & Conery 2003] Michael Lynch and John S Conery, The origins of genome complexity. *Science (New York, N.Y.)* **302**(5649) (2003), 1401–4. ↪ *page 72*

[MacArthur & Wilson 1967] Robert H MacArthur and Edward O Wilson, *The Theory of Island Biogeography.* Princeton University Press, 1967. ↪ *pages 3 and 11*

[Maddison & Schulz 2007] D. R. Maddison and K.-S. Schulz, The Tree of Life Web Project, `http://tolweb.org`, 2007. ↪ *page 6*

[Magurran 2004] Anne E Magurran, *Measuring Biological Diversity.* Blackwell Science Ltd, 2004. ↪ *pages 11 and 72*

[Martelli 2007] Alex Martelli, GMPY Multiprecision arithmetic for Python, `http://code.google.com/p/gmpy/`, 2007. ↪ *page 90*

[Martín-Trillo *et al.* 2011] Mar Martín-Trillo, Eduardo González Grandío, François Serra, Fabien Marcel, María Luisa Rodríguez-Buey, Gregor Schmitz, Klaus Theres, Abdelhafid Bendahmane, Hernán Dopazo, and Pilar Cubas, Role of tomato BRANCHED1-like genes in the control of shoot branching. *The Plant Journal : for Cell and Molecular Biology* **67**(4) (2011), 701–14. ↪ *page 103*

[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 20*

[Mayr 1942] Ernst Mayr, *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press, 1942. ↪ *page 13*

[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 35*

[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 11 and 67*

[McGrath & Katz 2004] Casey L McGrath and Laura a Katz, Genome diversity in microbial eukaryotes. *Trends in Ecology & Evolution* **19**(1) (2004), 32–8. ↪ *page 5*

[McShea 1996] Daniel W. McShea, Perspective: Metazoan complexity and evolution: ss there a trend? *Evolution* **50**(2) (1996), p. 477. ↪ *page 5*

[Mikkelsen *et al.* 2005] Tarjei S Mikkelsen, LaDeana W Hillier, Evan E Eichler, Michael C Zody, David B Jaffe, Shiaw-Pyng Yang, Wolfgang Enard, Ines Hellmann, Kerstin Lindblad-Toh, Tasha K Altheide, Nicoletta Archidiacono, Peer Bork, Jonathan Butler, Jean L Chang, Ze Cheng, et al., Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055) (2005), 69–87. ↪ *page 83*

[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 5*

[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 15*

[Moreira & Warnes 2004] W Moreira and G R Warnes, Rpy, a robust Python interface to the R Programming Language, `http://rpy.sourceforge.net/index.html`, 2004. ↪ *pages 19 and 103*

[Motomura 1932] I Motomura, A statistical treatment of associations. *Japanese Journal of Zoology* **44** (1932), 379–83. ↪ *page 2*

[Mustonen & Lässig 2009] Ville Mustonen and Michael Lässig, From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics* **25**(3) (2009), 111–9. ↪ *pages 74, 84, and 87*

[Nachman & Crowell 2000] Michael W Nachman and S L Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1) (2000), 297–304. ↪ *page 27*

[Nelder & Mead 1965] J A Nelder and R Mead, A Simplex Method for Function Minimization. *The computer journal* **7**(4) (1965), 308–313. ↪ *page 90*

[Néron *et al.* 2009] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal, Mobyle: a new full web bioinformatics framework. *Bioinformatics (Oxford, England)* **25**(22) (2009), 3005–11. ↪ *page 20*

[Nielsen 2001] Rasmus Nielsen, Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**(Pt 6) (2001), 641–7. ↪ *page 16*

[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. ↪ *pages 14, 15, 18, and 83*

[Nies 2009] André Nies, *Computability and Randomness.* Oxford University Press, 2009. ↪ *page 61*

[Nocedal & Wright 2000] Jorge Nocedal and Stephen J Wright, *Numerical Optimization.* Springer, 2000. ↪ *page 90*

[Notredame 2010] Cedric Notredame, Computing multiple sequence/structure alignments with the T-coffee package. *Current Protocols in Bioinformatics* **Chapter 3** (2010), Unit 3.8.1–25. ↪ *page 20*

[Ohta 1992] Tomoko Ohta, The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23** (1992), 263–86. ↪ *page 2*

[Orgel & Crick 1980] L E Orgel and F H Crick, Selfish DNA: the ultimate parasite. *Nature* **284**(5757) (1980), 604–7. ↪ *page 12*

[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 27*

[Oster & Alberch 1982] George Oster and Pere Alberch, Evolution and bifurcation of developmental programs. *Evolution* **36**(3) (1982), 444–59. ↪ *page 17*

[Petit & Barbadilla 2009] N Petit and Antonio Barbadilla, Selection efficiency and effective population size in Drosophila species. *Journal of Evolutionary Biology* **22**(3) (2009), 515–26. ↪ *page 80*

[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 15*

[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 96*

[Posada 2008] David Posada, jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**(7) (2008), 1253–6. ↪ *pages 20 and 101*

[Posada & Crandall 1998] David Posada and Keith A Crandall, MODELTEST: testing the model of DNA substitution. *Bioinformatics (Oxford, England)* **14**(9) (1998), 817–8. ↪ *page 101*

[Preston 1948] FW Preston, The Commonness, And Rarity, of Species. *Ecology* **29**(3) (1948), p. 254. ↪ *pages 2 and 11*

[Pritham 2009] Ellen J Pritham, Transposable elements and factors influencing their success in eukaryotes. *The Journal of Heredity* **100**(5) (2009), 648–55. ↪ *page 10*

[Proost *et al.* 2011] Sebastian Proost, Pedro Pattyn, Tom Gerats, and Yves Van de Peer, Journey through the past: 150 million years of plant genome evolution. *The Plant Journal : for Cell and Molecular Biology* **66**(1) (2011), 58–65. ↪ *page 52*

[Ridley 2004] Mark Ridley, *Evolution*. Blackwell Science Ltd, third edition, 2004. ↪ *page 2*

[Rizzon *et al.* 2002] Carène Rizzon, Gabriel Marais, Manolo Gouy, and Christian Biémont, Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. *Genome Research* **12**(3) (2002), 400–7. ↪ *page 69*

[Roger & Simpson 2009] Andrew J Roger and Alastair G B Simpson, Evolution: revisiting the root of the eukaryote tree. *Current biology* **19**(4) (2009), R165–7. ↪ *page 67*

[Ronquist & Huelsenbeck 2003] Fredrik Ronquist and John P Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)* **19**(12) (2003), 1572–4. ↪ *page 102*

[Rosenzweig 1995] ML Rosenzweig, *Species diversity in space and time.* Cambridge University Press, Cambridge, UK, 1995. ↪ *page 71*

[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), 340–8. ↪ *pages 11 and 72*

[Ryabko 1980] B Ya Ryabko, Data Compression by Means of a 'Book Stack'. *Problems Information Transmission* **16**(4) (1980), 16–21. ↪ *page 23*

[Sánchez *et al.* 2011] Rubén Sánchez, François Serra, Joaquín Tárraga, Ignacio Medina, José Carbonell, Luis Pulido, Alejandro de María, Salvador Capella-Gutiérrez, 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. ↪ *pages 37 and 100*

[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 25*

[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 19*

[Schmid & Tautz 1997] K J Schmid and D Tautz, A screen for fast evolving genes from Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **94**(18) (1997), 9746–50. ↪ *page 37*

[Serra *et al.* 2011] François Serra, Leonardo Arbiza, Joaquín Dopazo, and Hernán Dopazo, Natural selection on functional modules, a genome-wide analysis. *PLoS Computational Biology* **7**(3) (2011), p. e1001093. ↪ *page 103*

[Serra *et al.* 2012a] François Serra, Verónica Becher, and Hernán Dopazo, Neutral Theory Predicts the Relative Abundance and Diversity of Genetic Elements in a Broad Array of Eukaryotic Genomes. *Submitted* (2012a). ↪ *page 103*

[Serra *et al.* 2012b] François Serra, Hernán Dopazo, ERS Roldan, and M Gomendio, Sexual Selection and the Evolution of Protamines among Primates. *Submitted* (2012b). ↪ *page 103*

[Shannon 1948] C E Shannon, *A mathematical theory of communication*, vol. 5. The Bell System Technical Journal, 1948. ↪ *pages 23 and 32*

[Shapiro & Alm 2008] B Jesse Shapiro and Eric J Alm, Comparing patterns of natural selection across species using selective signatures. *PLoS Genetics* **4**(2) (2008), e23+. ↪ *pages 16, 17, and 86*

[Smit *et al.* 2010] Arian F. A. Smit, R Hubley, and P Green, RepeatMasker Open-3.0, `http://www.repeatmasker.org`, 2010. ↪ *page 25*

[Smith & Smith 1996] J M Smith and N H Smith, Synonymous nucleotide divergence: what is "saturation"? *Genetics* **142**(3) (1996), 1033–6. ↪ *page 19*

[Speijer 2011] Dave Speijer, Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **33**(5) (2011), 344–9. ↪ *page 5*

[Stoltzfus 1999] Arlin Stoltzfus, On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* **49**(2) (1999), 169–81. ↪ *page 5*

[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 16*

[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 20*

[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 8*

[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 20*

[Tárraga *et al.* 2007] Joaquín Tárraga, Ignacio Medina, Leonardo Arbiza, Jaime Huerta-Cepas, Toni Gabaldón, Joaquín Dopazo, and Hernán Dopazo, Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. *Nucleic Acids Research* **35**(Web Server issue) (2007), W38–42. ↪ *page 20*

[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 30*

[Team 2011] R Development Core Team, R: A Language and Environment for Statistical Computing, `http://www.r-project.org`, 2011. ↪ *pages 18 and 103*

[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 13*

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

[Van Rossum & Drake 2003] Guido Van Rossum and F L Drake, *Python language reference manual.* Network Theory Ltd., 2003. ↪ *pages 19, 25, and 103*

[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 5*

[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* **25**(7) (2009), 317–23. ↪ *page 4*

[Venter *et al.* 2001] J Craig Venter, Mark 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 35*

[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 12 and 18*

[Watterson 1974] GA Watterson, The sampling theory of selectively neutral alleles. *Advances in Applied Probability* **6**(3) (1974), 463–88. ↪ *page 3*

[Weber & Helentjaris 1989] D Weber and T Helentjaris, Mapping RFLP loci in maize using B-A translocations. *Genetics* **121**(3) (1989), 583–90. ↪ *page 54*

[Wernegreen 2002] Jennifer J Wernegreen, Genome evolution in bacterial endosymbionts of insects. *Nature Reviews. Genetics* **3**(11) (2002), 850–61. ↪ *page 47*

[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. ↪ *pages 10 and 28*

[Wilks 1938] Samuel 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 32*

[Williams 1966] George C Williams, *Adaptation and Natural Selection*. Princeton University Press, Princeton, New Jersey, 1966. ↪ *page 1*

[Wolfe 2001] Kenneth H Wolfe, Yesterday's polyploids and the mystery of diploidization. *Nature Reviews. Genetics* **2**(5) (2001), 333–41. ↪ *pages 54 and 58*

[Wright 1931] Stephen J Wright, Evolution in Mendelian Populations. *Genetics* **16**(2) (1931), 97–159. ↪ *page 11*

[Wright 1932] Stephen J Wright, The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of The Sixth International Congress on Genetics* **I** (1932), 356–66. ↪ *page 2*

[Yang 2007] Ziheng Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8) (2007), 1586–91. ↪ *pages 20, 37, and 43*

[Yang 2009] Ziheng Yang, User Guide PAML : Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **3**(September) (2009). ↪ *page 19*

[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 15*

[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*

[Zhu *et al.* 1994] C Zhu, Richard H Byrd, and Peihuang Lu, L-BFGS-B: a limited memory FORTRAN code for solving bound constrained optimization problems. *ACM Transactions on Mathematical Software* (1994), 1–17. ↪ *page 90*

[Zmasek 2012] Christian M Zmasek, Archaeopteryx, `http://www.phylosoft.org/archaeopteryx`, 2012. ↪ *page 102*

Glossary

biotype A transcript classification including protein coding, pseudogene, and non-coding RNAs. 10, 13, 28, 30

Burrows-Wheeler *-transform* Also called block-sorting compression, this is an algorithm used in data compression techniques such as bzip2. It is based on the concept of sorting all possible rotations of a given string, sorting the results in lexicographic order and finally taking the last character of each rotated string. 23, 24

C-value Refers to the amount of DNA contained within a haploid nucleus. The unit of measurement is picogram (pg). 5, 7, 8

class In computer science, a class is the description of the characteristics defining an object. Basically the class is what is written in the program while the object is the result of the execution of a class.. 89, 96, 127

de Bruijn *∼-sequence* A mathematically-defined string of characters with a perfect equal frequency of sub-sequences, i.e.: every possible combination of logarithmic length appears exactly once as a sequence of consecutive symbols. 59

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. 11, 13

LINE A long interspersed element sequence - typically used for non-long terminal repeat retrotransposons. 12, 25, 29, 54, 55, 67

LTR Long Terminal Repeat A kind of retrotransposon with direct repeats of 300-500bp of DNA at each end of the element. These sequences resemble the integrated proviruses of retroviruses. 12, 25, 29, 54, 55, 67

object In computer science, an object is a symbolic container with its own being defined in a class. An object can incorporate data and methods relating to something of the real world manipulated in a computer program.. 89, 91, 96, 127

patch *ecological* ~ A homogeneous area with a given shape and spatial configuration differing from the rest of the ecosystem. It is the lowest unit of a landscape. 13

retroposon A mobile DNA sequence that can move to new locations through an RNA intermediate. 12, 128

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). 10, 12, 65, 67

satellite A kind of tandem repeats, larger than minisatellites (10-60 bp) and microsatellites (2-6 bp).. 12, 13, 25, 29, 54, 55, 61, 67, 69

script A program written for a software environment that automates the execution of tasks that could alternatively be executed in sequence by a human operator. 18, 89, 103

seed *-sequence* of a gene or a protein, is the sequence used as a starting point in the search for homologous sequences within a given set of entries. Extending this concept to the genomic level, gives rise to the concepts of *seed-genomes* or *seed-species*. ***Note:*** in a phylome, it is expected to observe an over-representation of proteins from the seed-species.
~ *-species*, in the case of ortholog retrieval, a *seed species* is the equivalent of a *seed sequence*. 19, 35, 131

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. 12

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. 12, 25, 29, 54, 55, 61, 67, 70, 153

superfamily *Transposable elements'*-, The fourth level in the classification of transposable elements according to `http://www.bioinformatics.org/wikiposon/doku.php?id=main`. This level classifies elements based on the structure of the internal sequence. 13, 26, 28, 29

tandem repeat Repetitive sequence of DNA, comprising a pattern of two or more sequentially repeated nucleotides.. 67, 128

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. 10, 12, 13, 25, 65

trophic Involving the feeding habits or food relationship of different organisms in a food chain. 11

A. Overview on the detection of selective pressures at the genomic level

In this appendix, I review the different steps conducted to detect selective pressures in protein-coding genes, and detail some of the improvements in protocols implemented in this thesis.

The classical pipeline can be divided into five main steps, itemized as:

- **Definition of a set of species**: firstly a seed species or sequence needs to be defined. Using this seed as a basis, a set of sequences needs to be selected. These sequences may not be too distant from the 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 four sequences [Yang 2009].

- **Homologous sequences retrieval**: once the set of species has been selected, the next step consists in retrieving homologous sequences; the most popular options being *Ensembl* [Flicek *et al.* 2011], the database resources of the NCBI [Sayers *et al.* 2011], or the UCSC database [Fujita *et al.* 2011].

- **Alignment**: the importance of this step is often underestimated. However, when measuring selective pressure and, in particular, for the detection of positive selection, misalignment generate a high proportion of false positives. The 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 remove abnormally-divergent columns (sites) from the alignment. It is vital to pay special attention to generating accurate alignments, since misaligned regions can be misinterpreted as fast-evolving regions, thereby strongly biasing the detection of positive selection. Moreover, since this step is computationally the fastest, the cost for ensuring accurate alignments is generally low.

- **Phylogenetic reconstruction**: this step is unnecessary if the species tree is already known (as is generally the case). In any case, I review how to construct accurate phylogenetic trees, paying special attention to the use of model testing [Posada 2008, Abascal *et al.* 2005].

- **Identification of selective pressures and testing of evolutionary hypotheses**: this is the final step of the analysis. It requires most of the computation time and, here, proved a real challenge to automate. The programs used to calculate selective pressures and fit evolutionary models in this study were SLR [Massingham & Goldman 2005] and CodeML from the PAML package [Yang 2007].

Among these steps, the most critical are certainly the first ones i.e. the definition of the target groups, sourcing homologous sequences, and their subsequent alignment. However, the essentially technical aspect of these steps is often underestimated. In the sections below, I review the classical methodologies and propose some solutions in order to improve their accuracy.

A.1. The selection of homologous sequences

The first step in an analysis of selective pressures acting on a gene is the definition of a set of homologous genes. Two parameters have to be taken into account here: the number of sequences to be compared and their degree of similarity. Based on the recommendations of Ziheng Yang in PAML's "FAQs" [Yang 2007], I mention these points:

- The number of sequences: preferentially a minimum of four to five sequences with optimal sequence divergence.

- The optimal divergence: The sum of dS over all branches in a tree is greater than 0.5. The maximum divergence that can be handled to reconstruct an accurate phylogenetic tree is assumed to be limited by saturation of synonymous sites. However, since the emergence of maximum likelihood in phylogenetic analysis, synonymous saturation is only considered to be problematic when dS estimation gives values higher than 2 o 3 or when third codon position divergence is higher than 30-40%. As a result, other limitations may take precedence, such as the accuracy of the alignment, or the heterogeneity of nucleotide frequencies (which would bias the substitution process in some species relative to others) [Yang 1998b].

Thus, accurately aligning a set of at least 4-5 sequences, and summing $dS > 0.5$, is generally sufficient to start building a phylogenetic tree and estimating selective pressures over it.

A.2. The alignment

The first aspect to consider when aligning coding sequences is their intrinsic structure; the three nucleotides constituting a codon should not be separated in the alignment process. This consideration plus the fact that alignments based on amino acid sequences are generally more accurate, leads to the conclusion that the best way to align coding-sequences is to use their protein-based translation.

The variety of software for aligning sequences might seem bewildering. Software needs to e carefully chosen, especially since inaccurate alignments present a high possibility of giving false positive results when testing for evolutionary hypotheses. Even if some multi-alignment tools seem to stand out in terms of accuracy and computation time (e.g. Muscle [Edgar 2004], MAFFT [Katoh *et al.* 2005], DIALIGN-TX [Subramanian *et al.* 2008], ProbCons [Do *et al.* 2005] or T-COFFEE [Notredame 2010]), results may vary.

Some studies have attempted to classify these tools in terms of accuracy, e.g. [Plyusnin & Holm 2012] using the BAliBASE [Bahr *et al.* 2001] dataset. However, even though some alignment tools in general seem to bebetter than others (ProbCons or T-Coffee), their accuracy may be highly dependent on the data used [Edgar & Batzoglou 2006].

In the protocol for identifying selective pressures among genes, though crucial, the alignment step, is usually the fastest in terms of computational time. Thus, consideration should be given to employing the solution of [Edgar & Batzoglou 2006], whereby the consensus solution of several alignments derived from different algorithms is used (the example proposed in the review is T-COFFEE, ProbCons and MUSCLE). This consensus alignment can be obtained for example, with the M-COFFEE tool. This strategy would allow the deletion of those sites in the consensus alignment that vary depending on the alignment software used.

Finally, once an accurate alignment has been obtained, removal of columns (or rows, if possible) in the alignment that represent unrealistic variation is necessary. This step can be carried out using TrimAl [Capella-Gutiérrez *et al.* 2009].

A.3. Model testing and phylogenetic inference

In the pipeline for the detection of selective pressures at a molecular level, the phylogenetic relationship between sequences must be known. This step may be skipped if the phylogenetic tree is already known with confidence. Otherwise, the phylogeny must be constructed.

As with the alignment step, the reconstruction of a phylogeny is generally more accurate when dealing with amino acid sequences. The only exception is if sequences are too similar to present differences in their amino acid sequences.

The distances between sequences are calculated according to a model. In the case of nucleotide models, there is a range from the simplest Jukes and Cantor (JC) model [Jukes & Cantor 1969] (which assumes equal transition and transversion rates as well as equal base equilibrium frequencies), to the most complex General Time-Reversible (GTR) model [Rodríguez *et al.* 1990] (which takes into accounts different rates for each possible nucleotide substitution). Several evolutionary models are also available for amino acid sequences, which are based on empirical data.

Whether nucleotide or amino acid sequences are being considered, the model that best describes the substitutions observed between sequences must be chosen. Currently, the most accepted methodologies consist of comparing the likelihood of models through LRT (if the models are nested) or their Akaike (or AIC) score [Akaike 1974] otherwise (e.g. amino acid models).

Although this methodology can successfully identify the most likely model explaining the successive substitutions that differentiate sequences, the cost in terms of computational time is high. Some approximations, however, can be used in order to accelerate the process, e.g. using fixed topology build by neighbor-joining (NJ) is believed to generate good approximations in determining the best model for use [Posada & Crandall 2001].

Once the evolutionary model with the best fit has been identified, the phylogeny can be constructed. It is not necessary to discuss this topic more extensively here, but if the chosen model fails to generate a phylogeny with strongly-supported nodes, “multi-forked” can be used to go on with the protocol. Using a consensus solution is safer than employing one with poor statistical support.

A.4. Testing for evolutionary scenarios in protein-coding genes

Once formulated, an evolutionary hypothesis can tested. The classical methodology for testing evolutionary hypothesis involves the use of programs like CodeML from the PAML package [Yang 2007] or the SLR program [Massingham & Goldman 2005] (among others). However, these tools are designed to be used for studying one or very few trees. First because of the preparation of the data, the configuration file and the successive tests needed in order to find with confidence the optimal values of each of the parameters. Second, because the interpretation and summary of the results is also complicated, given the quantity of information usually produced.

This section is dedicated to reviewing the most important points to take into account when analyzing different evolutionary scenarios, together with examples of the most classical tests for protein-coding regions.

A.4.1. Codon substitution model

The detection of selective pressures in the protein-coding regions of genomes basically consists of counting the number of synonymous and non-synonymous changes between pairs of sequences. More precisely, the distance between two codons is calculated according to a Markov-chain model, using the proportions of each codon and a substitution-rate matrix, $Q = \{q_{ij}\}$. q_{ij}, representing the rates of changes from codon i to codon j. The most commonly used model is a simplification of that of Goldman and Yang [Goldman & Yang 1994]:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 codon positions.} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by 1 synonymous transversion.} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by 1 synonymous transition.} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by 1 non-synonymous transversion.} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by 1 non-synonymous transition.} \end{cases} \tag{A.1}$$

With π_j representing the equilibrium frequency of codon j, κ the ratio of transitions over transversions (see Figure A.1), and ω the ratio of non-synonymous over synonymous mutations.

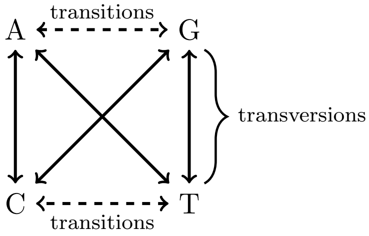


Figure A.1.: Transition and transversion.
Schematic diagram defining the different kinds of substitutions between each nucleotide.

The different codon models correspond to the different assumptions made on the distribution of the equilibrium codon frequencies, π_j. Most common codon models assume either ***1)*** that each codon has the same frequency ("F1×4" in CodeML – 1 degree of freedom) ***2)*** codon frequencies are estimated based on the observed frequencies of nucleotides ("F3×4" in CodeML – 3 degrees of freedom), or ***3)*** codon frequencies are different for each codon ("F61" in CodeML – 59 or 60 degrees of freedom).

To relate the model to real data over time, a transition-probability for any time (t) needs to be defined, and for all possible i and j codons ($p_{ij}(t) = Pr\{X(t) = j|X(0) = i\}$). The matrix of this transition-probability ($P(t)$) can be calculated as:

$$P(t) = \{p_{ij}(t)\} = e^{Qt} \tag{A.2}$$

Equation A.1 provides the major parameters that affect the number of changes between two codons (plus the time Equation A.2). For real data, within a phylogenetic tree, this estimation has to be done for each codon, and at each internal node. Nowadays solutions for this problem mainly center on maximum likelihood methods, using Felsenstein's pruning algorithm [Felsenstein 1981].

A.4.2. Overview of major/classical evolutionary models

In the previous section, the parameters to be considered in order to compute distances between sequences are described. However, the quantity of parameters that need to be estimated for each codon and at each internal node may lead to a model that is "over-fit". In order to estimate the importance of the optimization of any given parameter, a likelihood-ratio test can be performed. For example, in the previous section, the most frequently used codon models (F1×4, F3×4 and F61) were mentioned. In order to decide which one of these models would best describe the changes observed between codons, their individual likelihoods for a given dataset can be computed.

Another example lies in the assumption of different selective pressures, either over branches or over sites. Selective pressures (related to the ω ratio) can be optimized according to different hypotheses of heterogeneity. Below I focus on three main groups of models; site, branch and branch-site models.

Site models

Site models assume that all branches of a given phylogenetic tree evolve at the same rate ω, but allow different selective pressures to occur along the alignment. These models are useful in order to quickly see which parts of a sequence alignment are under strong selective constraints. Several strategies are available in order to obtain the pattern of ω values among sites, two of which are the most commonly used. **First**, using a "site wise based likelihood-ratio" methodology implemented in the SLR program [Massingham & Goldman 2005] which computes a LRT at each site i: $\Lambda_i = 2 \times (l_i(1) - l_i(\hat{\omega_i}))$, with $l_i(1)$ as the likelihood of the null model assuming that $\omega = 1$, and $l_i(\hat{\omega_i})$ the likelihood of the alternative model letting the estimate of ω_i vary. **Second**, site models implemented in CodeML [Yang 2007] which are based on the definition of a prior, that segregates different categories of sites based on the distribution of the random variation of ω values (e.g. categories of sites with $\omega > 1$), followed by assignment of a value of probability for each site to belong to one of these categories.

Branch models

These models are orthogonal to site models in the sense that the rate ω is not allowed to vary along the alignment, but it can be estimated independently for each branch of the phylogeny. The simplest (and most unrealistic) branch model, the "M0" model, assumes that all branches evolve at the same rate (as for site models), and calculate a unique value of ω for the entire phylogeny. At the other end of the scale, the most complex model, the "free-ratio" model estimates a different value of ω for each branch of the phylogeny. Although, both the M0 and free-ratio models might be either unrealistic or classic cases of over-fitting models, they are useful as respective null or alternative models when testing a given evolutionary scenarios.

Overall, the assumptions made by branch models are dangerously unacceptable, as the high values of ω detected are usually due to the counterbalanced effect of conserved and accelerated sites. However, branch models are useful for detecting differences between the evolutionary rates of sequences, either taken individually or grouped (e.g. in clades). A classical example of the use of these models was in the detection of different selective pressures occurring between colobine and hominid lysozyme protein [Yang 1998a].

Branch-site models

As the name suggests, these models represent a compromise between the previous two groups of models. More precisely, phylogenetic information is used to contrast differences in rates at a given site. This model is usually used to detect, or contrast, sites under a characteristic selective pressure; the contrast being made between two parts of a phylogenetic tree, generally referred to as the foreground and background branches. Foreground branches have an extra class of sites that allows ω to be higher than 1. In a classic branch-site model (*branch-site A* in CodeML), foreground branches are fitted to a model similar to the M2 model, while background branches are fitted to a model similar to the M1 model. These models are more realistic then site or branch models, and are the most reliable for tests of positive selection.

A.4.3. Testing for the best evolutionary model

Beyond the descriptive use of the models described above, models can be compared through LRT in order to evaluate, with statistical significance, the importance of the optimization of a given parameter. This methodology, for example, was extensively used on each of the groups of models described above in the context of the test for positive selection. Here, some examples are listed of the most popular tests that can be done by comparing the fit of different evolutionary models by means of their likelihood.

Test of positive and relaxed selection at sites

As seen above, a site analysis in CodeML basically consists of classifying all the sites in an alignment into the categories of sites defined by the model. For the determination of positively-selected sites, models can be classified into two categories: neutral models and positive-selection models; the difference between them being that positive-selection models have an extra category of sites that allows ω to be higher than 1. A LRT can be conducted between nested models (usually with 1 degree of freedom, which corresponds to the estimation of the extra class of sites). If the positive-selection model is optimal, then sites that belong to this extra category (with $\omega > 1$) are considered to be truly under positive selection.

A classical positive-selection test is done by comparing model the M1a (neutral, with two categories of sites: $0 < \omega_0 < 1$ and $\omega_1 = 1$) with model M2a (positive selection, with the same categories as M1a plus $\omega_2 > 1$).

Test of different selective regimes in branches

These models are perhaps currently considered the most unrealistic for testing for positive selection. However, they still are useful for comparing and differentiating selective regimes in branches. For example, the hypothesis that a given clade is undergoing an accelerated mutation rate compared to the rest of the phylogeny, can be tested in this way.

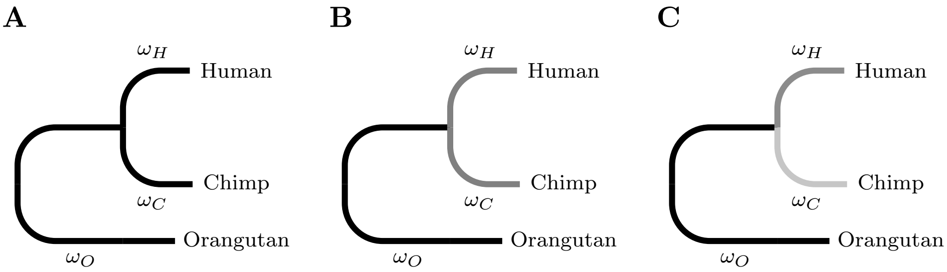


Figure A.2.: A simple example of allowing / disallowing different ω rates in a tree.
A simple phylogeny is represented here, where the colors of the branches represents the different estimations of ω. In (**A**), the ω value of all the branches is the same ($\omega_O = \omega_C = \omega_H$), in (**B**) the ω_O of the Orangutan, is different from the rest of the tree ($\omega_O \neq \omega_C = \omega_H$), and in (**C**) each branch has a different value of ω ($\omega_O \neq \omega_C \neq \omega_H$). Figure adapted from [Yang 2006].

Figure A.2 is an example of application of this methodology to a simple phylogeny representing three sequences (human, chimp and orangutan). (**A**), (**B**) and

(**C**) in the figure, represent different branch models. (**A**) is the simplest branch model, where a single rate ω is estimated for all branches. In (**B**), two values of ω are estimated (thus, there is one extra parameter compared to (**A**)). In (**C**) three values of ω have been calculated (this time, there are two more parameters compared to (A) and on more compared to (B)). A comparison of the models (**A**) and (**B**) using LRT could determine if the evolutionary rate calculated for the orangutan sequence is significantly different from the rest of the tree. In the same way, the comparison between (**B**) and (**C**) tests if each branch is evolving at a different rate.

Test of positive and relaxed selection in sites of a given set of branches

This test is perhaps the most sensitive for detecting positive selection in protein-coding genes. For the branch tests described above, a protein is considered as being under positive selection only if the average ω over all sites is higher than 1. However, this can be misleading if only a few sites are evolving fast in the context of global purifying selection for the rest of the protein. In the same way, in the case of site tests, the ω value of a given site is averaged over all branches. Both branch and site tests have successfully been used to detect positive selection in protein-coding genes; though, in most of the cases significant accelerations of dN relative to dS would only affect some sites in a given lineage. This is the main reason that prompted Yang and Nielsen [Yang & Nielsen 2002] to implement a new test for positive selection that can detect positive selection at only a few sites in a particular lineage. Originally, the test consisted of comparing the branch-site model A (bsA) to the model M1a (this test is often referred to as "test I"). From the specification of the bsA model (see Table A.1), the only difference with model M1a is the presence of sites evolving at rate ω_2 in foreground branches.

Site class	Proportion	Background ω	Foreground ω
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

Table A.1.: Assumed ω ratios in branch-site model A.
In a phylogenetic tree where branches are divided into two categories, "background" and "foreground", sites are allowed to evolve at ω rates higher than 1 only in foreground branches. p_0 and p_1 are the proportion of sites evolving at rates ω_0 and ω_1, respectively.

Even if, theoretically, the test between bsA and M1a (test I) seems robust, simulation studies have found high proportions of false positive results [Zhang 2004], resulting in the elaboration of a new test [Zhang *et al.* 2005]. This new branch-site test (referred to as "test II") consists of comparing two branch-site models; one with $\omega_2 > 1$ (bsA) and the other with $\omega_2 = 1$ (branch-site A1, bsA1).

The null model is bsA1. A gene would be considered to be under positive selection if the bsA model fits the dataset better than the bsA1 model.

As the first branch-site test (test I) is considered to generate an unacceptably high proportion of false positive results, it is often used to detect relaxed selection in genes (see Figure A.3). Thus, a protein-coding gene would be considered to be under positive selection if it passes test I, and considered to be under relaxed selection if it passes test II but not test I.

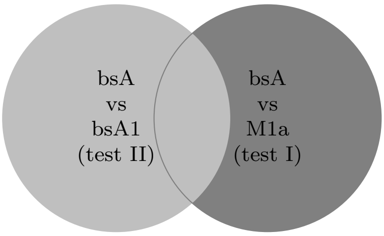


Figure A.3.: Branch-site tests for the detection of positive selection and relaxation.
Circles here represent positive results of the LRT between branch-site model A (bsA) and branch-site model A1 (bsA1) ("Test II") on the left, and bsA versus site model M1a on the right ("Test I"). The color represents the set of genes that should be considered as under positive selection (light gray on the left) or under relaxation (dark gray on the right).

Appendix A – 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 132*

[Akaike 1974] Hirotsugu Akaike, A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6) (1974), 716–723. ↪ *page 134*

[Bahr *et al.* 2001] Anne Bahr, Julie D Thompson, J-C Thierry, and Olivier Poch, BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research* **29**(1) (2001), 323–6. ↪ *page 133*

[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 131 and 133*

[Do *et al.* 2005] Chuong B Do, Mahathi S P Mahabhashyam, Michael Brudno, and Serafim Batzoglou, ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research* **15**(2) (2005), 330–40. ↪ *page 133*

[Edgar 2004] Robert C Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**(5) (2004), 1792–7. ↪ *pages 131 and 133*

[Edgar & Batzoglou 2006] Robert C Edgar and Serafim Batzoglou, Multiple sequence alignment. *Current Opinion in Structural Biology* **16**(3) (2006), 368–73. ↪ *page 133*

[Felsenstein 1981] J Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**(6) (1981), 368–76. ↪ *page 136*

[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. ↪ *page 131*

[Fujita *et al.* 2011] Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R Dreszer, Belinda M Giardine, Rachel A Harte, Jennifer Hillman-Jackson, et al., The

UCSC Genome Browser database: update 2011. *Nucleic Acids Research* **39**(Database issue) (2011), D876–82. ↪ *page 131*

[Gojobori 1983] T. Gojobori, Codon substitution in evolution and the "saturation" of synonymous changes. *Genetics* **105**(4) (1983), p. 1011. ↪ *page 131*

[Goldman & Yang 1994] Nick Goldman and Ziheng Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**(5) (1994), 725–36. ↪ *page 135*

[Jukes & Cantor 1969] T H Jukes and CR Cantor, Mammalian Protein molecules. In *Evolution of Protein Molecules*, 21–132. Academic Press, New York, in munro (edition, 1969. ↪ *page 134*

[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. ↪ *pages 131 and 133*

[Massingham & Goldman 2005] Tim Massingham and Nick Goldman, Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**(3) (2005), 1753–62. ↪ *pages 132, 134, and 136*

[Notredame 2010] Cedric Notredame, Computing multiple sequence/structure alignments with the T-coffee package. *Current Protocols in Bioinformatics* **Chapter 3** (2010), Unit 3.8.1–25. ↪ *pages 131 and 133*

[Plyusnin & Holm 2012] Ilya Plyusnin and Liisa Holm, Comprehensive comparison of graph based multiple protein sequence alignment strategies. *BMC Bioinformatics* **13**(1) (2012), p. 64. ↪ *page 133*

[Posada 2008] David Posada, jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**(7) (2008), 1253–6. ↪ *page 132*

[Posada & Crandall 2001] D. Posada and K. A. Crandall, Selecting Models of Nucleotide Substitution: An Application to Human Immunodeficiency Virus 1 (HIV-1). *Molecular Biology and Evolution* **18**(6) (2001), 897–906. ↪ *page 134*

[Rodríguez *et al.* 1990] F. Rodríguez, J.L. Oliver, A. Marín, and J.R. Medina, The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* **142**(4) (1990), 485–501. ↪ *page 134*

[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 131*

[Smith & Smith 1996] J M Smith and N H Smith, Synonymous nucleotide divergence: what is "saturation"? *Genetics* **142**(3) (1996), 1033–6. ↪ *page 131*

[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. ↪ *pages 131 and 133*

[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 131*

[Yang 1998a] Ziheng Yang, Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**(5) (1998a), 568–73. ↪ *page 137*

[Yang 1998b] Ziheng Yang, On the Best Evolutionary Rate for Phylogenetic Analysis. *Systematic Biology* **47**(1) (1998b), 125–133. ↪ *page 132*

[Yang 2006] Ziheng Yang, *Computational Molecular Evolution*, vol. 1. Oxford University Press, Oxford, England, 2006. ↪ *page 138*

[Yang 2007] Ziheng Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8) (2007), 1586–91. ↪ *pages 132, 134, and 136*

[Yang 2009] Ziheng Yang, User Guide PAML : Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **3**(September) (2009). ↪ *page 131*

[Yang & Nielsen 2002] Ziheng Yang and Rasmus Nielsen, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**(6) (2002), 908–17. ↪ *page 139*

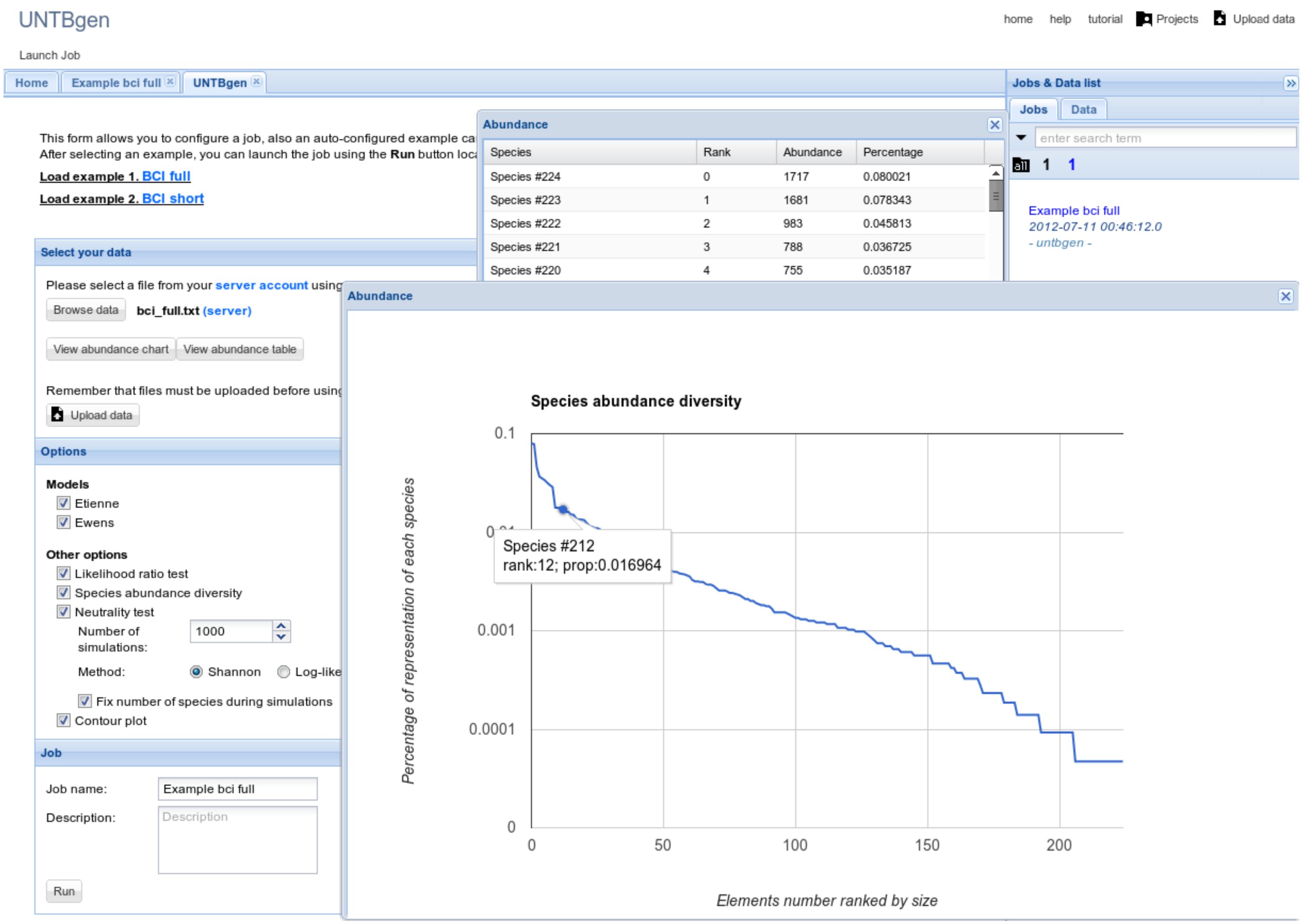
[Zhang 2004] Jianzhi Zhang, Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution* **21**(7) (2004), p. 1332. ↪ *page 139*

[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 139*

B. Screen-shot UNTBGen main form with uploaded data

In the following page is shown a screen-shot of the main form of the UNTBGen web server. BCI-full example dataset is loaded, and both descriptive tools, the RSA chart and the table of abundances are opened.

B. Screen-shot UNTBGen main form with uploaded data



This form allows you to configure a job, also an auto-configured example ca
After selecting an example, you can launch the job using the **Run** button loca

Load example 1. BCI full

Load example 2. BCI short

Select your data

Please select a file from your server account using

Browse data **bci_full.txt** (server)

View abundance chart | View abundance table

Remember that files must be uploaded before using

Upload data

Options

Models

- ☑ Etienne
- ☑ Ewens

Other options

- ☑ Likelihood ratio test
- ☑ Species abundance diversity
- ☑ Neutrality test

Number of simulations: 1000

Method: ● Shannon ○ Log-like

- ☑ Fix number of species during simulations
- ☑ Contour plot

Job

Job name: Example bci full

Description: Description

Run

Example bci full
2012-07-11 00:46:12.0
- untbgen -

Abundance

Species	Rank	Abundance	Percentage
Species #224	0	1717	0.080021
Species #223	1	1681	0.078343
Species #222	2	983	0.045813
Species #221	3	788	0.036725
Species #220	4	755	0.035187

Abundance

Species abundance diversity
0.1
0.01
0.001
0.0001
0
Percentage of representation of each species
0
50
100
150
200
Elements number ranked by size
Species #212
rank:12; prop:0.016964

Resumen en castellano

Introducción

En el estudio de las modificaciones genéticas que conducen a las poblaciones a adaptarse a su ambiente, es importante distinguir de forma inequívoca, los cambios que incrementan la eficacia biológica de aquellos que son neutros o levemente deletéreos en un genotipo.

Aunque la neutralidad fue claramente señalada en el mismo "Origen" de Charles Darwin, su relevancia dentro del proceso evolutivo fue desestimada hasta el descubrimiento de la enorme variación poblacional detectada en los primeros datos moleculares. Estos descubrimientos condujeron al desarrollo de los modelos neutros de evolución molecular. Gracias a su desarrollo y aceptación, los modelos neutros permitieron formular y probar estadísticamente hipótesis sobre la evolución adaptativa de secuencias biológicas.

En ecología el desarrollo de modelos neutros capaces de entender el patrón observado de diversidad y abundancia de especies es más reciente, y su uso se extendió sólo recientemente, demostrando un ajuste significativo a la casi totalidad de los ecosistemas analizados.

Tanto en ecología como en biología molecular, la definición explícita de los modelos neutros lleva a valorar los cambios adaptativos producidos por selección natural. Más allá de su valor descriptivo, el modelo neutro es también una poderosa herramienta estadística.

Esta tesis analiza tres aproximaciones biológicas diferentes en genomas completos. Cada uno de ellos plantea de forma explícita un modelo neutro y sus desviaciones respectivas. El planteamiento de este común denominador a lo largo de los tres capítulos principales le dan valor explicativo a esta tesis. Sin él, las conclusiones serian irrelevantes. Estas aproximaciones son:

- **Contenido de información**: este es el análisis más simple que se puede hacer de un genoma cómo que entidad contenedora de información biológica. Bajo esta perspectiva, se tiene exclusivamente en cuenta el conjunto de nucleótidos A, C, G y T, considerando estos constituyentes del genoma como unidades fundamentales e independientes. En este estudio consideramos los genomas como una secuencia simple con un determinado contenido informativo. Nuestro primer objetivo es medir el contenido de información de genomas completos. Para ello utilizamos un amplio rango de especies representativo de gran parte de la diversidad de la vida. En el marco de esta tesis, la descripción del contenido de información en genomas se inscribe como el estudio más elemental y seguramente el menos biológico de la descripción del sustrato genómico.

- **Ecológico**: el estudio de la abundancia y diversidad de familias de elementos genéticos (mayoritariamente elementos transponibles) en genomas eucariotas, si bien está definido desde la genética de poblaciones nunca ha pasado de la modelización de unas pocas de estas familias en un mismo genoma. No obstante los ecólogos ya implementaron modelos estadísticos y mecanísticos capaces de predecir con precisión los patrones y procesos subyacentes a la composición de los ecosistemas. En este contexto, puede parecer natural aplicar estos modelos ecológicos sobre las diferentes familias de elementos genéticos que componen nuestro genoma. De hecho, la ecología del genoma no es un concepto nuevo asociado a esta tesis. Esta segunda aproximación a los genomas completos tiene como fin la elaboración de un modelo general de ecología de genomas para todos los elementos que pueblan los cromosomas eucariotas; y la posterior verificación de este modelo en genomas y cromosomas de un gran número de organismos.

- **Sistémico**: esta es la última aproximación bajo la cual analizamos los genomas. Para ello cambiamos el nivel de análisis, pasando de genes individuales a grupos de genes con características funcionales similares. Concretamente, nos centramos en proteínas que actúan conjuntamente para completar una ruta bioquímica o asociadas a una determinada función. En este contexto varios estudios genómicos realizados gen a gen, ya intentaron discernir enriquecimientos funcionales entre los genes sujetos a selección positiva. Sin embargo esta metodología es propensa a perder poder estadístico cuando es aplicada a nivel genómico. En consecuencia, estos estudios no consiguieron resaltar patrones estadísticamente significativos. Nuestro objetivo será implementar una nueva metodología capaz de detectar la huella de la selección natural en diferentes categorías funcionales.

El último capítulo de esta tesis constituye la parte mas técnica derivada de cada uno de los tres estudios ya mencionados. Esta parte consiste en el desarrollo de varias herramientas bioinformáticas específicas, entre la cuales se hayan algunas con potencial interés para la comunidad científica. En tres apartados se presenta las siguientes herramientas: Ecolopy, diseñada para el estudio de ecosistemas genómicos; ETE-Evol, para la manipulación de árboles filogenéticos y pruebas de hipótesis evolutivas en regiones codificantes; y Phylemon, un servidor web que propone un amplio abanico de herramientas, todas enmarcadas en los campos de la filogenia, la filogenómica y el test de hipótesis evolutivas.

Material y métodos

Análisis de la complejidad de genomas

Para el estudio de la complejidad del genoma en términos de cantidad de información, definimos la tasa de complejidad (CR – por sus siglas en inglés *complexity ratio*). El CR se elabora a partir de tres funciones diferentes que transforman una secuencia de ADN en un valor numérico. Cada una de estas funciones se basa en algoritmos clásicamente utilizados en informática para la compresión de datos. El primero es una transformación de *Burrows-Wheeler*, que ordena los caracteres de una secuencia. Cada carácter se ordena acorde a los caracteres que le siguen en la secuencia. El segundo consiste en aplicar sobre la secuencia previamente ordenada, el algoritmo *Move-To-Front*, que valora numéricamente el desordenamiento que supone cada carácter para la secuencia. Finalmente calculamos la entropía de Shannon correspondiente a los valores derivados del *Move-To-Front*, resultando en un valor de complejidad (CV – por *complexity value*) que podemos normalizar en base al largo de la secuencia para obtener el CR.

Este análisis fue aplicado sobre 54 genomas cubriendo un amplio rango de formas de vida, desde virus a mamíferos pasando por bacterias, plantas y aves, e incluyendo especies con genomas poliploides, organismos de vida en condiciones extremas, parásitos intracelulares, organismos con expansiones génicas, reducciones genómicas, organismos con genoma de ARN, de una sola hebra, e incluso con un genoma sintético.

Estudio de la distribución de abundancia de elementos genéticos en eucariotas

En el contexto de este estudio, usamos 31 genomas de especies eucariotas. El primer paso consistió en identificar los diferentes elementos genéticos y clasificarlos en familias. En este estudio hemos analizado elementos repetidos derivados de la base de datos *RepBase*, y biotipos (o tipos de transcritos, como los codificantes, los ARN-t o los micro-ARN, entre otros) derivados de la bases de datos *Ensembl* de anotación de genomas.

Para poner a prueba la distribución aleatoria de elementos genéticos en cromosomas y genomas hemos simulado su distribución equiprobable y su distribución a través de un proceso neutro donde cada uno de estos elementos está sujeto al principio de equivalencia ecológica.

Para ello hemos desarrollado herramientas estadísticas, inspiradas en las usadas por los ecólogos, a fin de poner a prueba la teoría neutra unificada de biodiversidad (UNTB – por *Unified Neutral Theory of Biodiversity*) originalmente planteada en ecología por Stephen Hubbell y aplicada en esta tesis para los 548 cromosomas.

Presiones selectivas a nivel genómico

El primer paso en el estudio de presiones selectivas en grupos de genes relacionados funcionalmente, es la definición de las especies a analizar y la selección de genes ortólogos. Para este trabajo decidimos concentrarnos en 2 grupos de organismos modelos, los mamíferos con humano, chimpancé, ratón y rata; y *Drosophila* con *D. melanogaster*, *D. sechellia*, *D. simulans*, *D. yakuba* y *D. erecta*). Después de la aplicación de diferentes filtros, identificamos 12.453 y 9.240 grupos de genes ortólogos, respectivamente.

Las presiones selectivas en genes codificantes fueron medidas mediante el valor de ω (dN sobre dS). Además de este cálculo, estimamos el conjunto de genes bajo selección positiva ($\omega > 1$) usando herramientas clásicamente utilizadas en evolución molecular computacional.

En este estudio desarrollamos el "*Gene Set Selection Analysis*" (GSSA), una propuesta estadística cuyo propósito es detectar desviaciones significativas en la distribución genómica de una variable evolutiva como los valores ω, dS y dN. El GSSA consiste en la aplicación de cinco pasos sucesivos: ***1)*** ordenar los genes en función de una variable evolutiva, ***2)*** anotar los genes con categorías funcionales, ***3)*** cortar la lista de genes en dos particiones, ***4)*** aplicar un test de Fisher entre los genes anotados con una función dada, y los pertenecientes a una de las particiones definidas, y ***5)*** corregir el conjunto de los p-valores de los resultados por test múltiples. Los resultados del GSSA para los diferentes grupos de genes funcionalmente relacionados son de tres tipos: ***1)*** los no significativos (NS – por *no-significant*), con ausencia de desviaciones significativas respecto a lo observado en el genoma, y ***2)*** los que muestran una desviación significativa hacia valores altos (SH – por *significantly high*), o ***3)*** hacia valores bajos (SL – por *significantly low*) en comparación con lo observado en cada genoma. En función de esta observación describimos procesos evolutivos para cada una de las especies y grupos de especies

Resultados y discusión

Estructura cuasi-aleatoria del ADN

Nuestro primer resultado consiste en la observación de una correlación extraordinaria entre el valor de complejidad (CV) y el tamaño de la secuencia analizada en 54 especies pertenecientes a los 20 grupos sistemáticos estudiados. Esta correlación es observada en cromosomas y genomas completos (con una pendiente de 0'924 para cromosomas y de 0'967 para genomas). Las tasas de complejidad de información son en su mayoría cercanos al máximo (CR $>$ 0'95). Entre los genomas de menor entropía encontramos, por una parte los poliploides recientes como el maíz, el sorgo o *Danio rerio*, y por otra parte los genomas con una composición de nucleótidos fuertemente sesgada como *Plasmodium* o *Dictyostelium* con concentraciones en A + T superiores al 75%.

En cuanto a los valores de complejidad en familias de elementos genéticos, observamos que los elementos de mayor complejidad son los genes, y en particular los exones. Por otra parte, en cuanto a los elementos repetidos (donde esperábamos encontrar valores muy bajos de complejidad en información), resultó que las únicas categorías de elementos con valores de complejidad bajo fueron los SINEs y los satélites. Estos resultados apuntan a que los genomas presentan un nivel alto de variabilidad en las regiones con alta concentración de elementos repetidos.

Finalmente, para poder apreciar las diferencias en CR que observamos entre poliploides y no poliploides, simulamos eventos de mutación y de transposición sobre secuencias aleatorias representando genomas y cromosomas poliploides, y también sobre algunos cromosomas de maíz y de sorgo. El resultado de esta simulación, fue que en ambos casos, a través de mutaciones o de transposiciones, el máximo de complejidad se recobró al cabo de un numero de generaciones elevado. En el caso de las mutaciones, usando una tasa de mutación intermedia entre las observaciones para plantas y mamíferos, 30 millones de generaciones fueron suficientes para recobrar un valor de CR $>$ 0'95 para el conjunto de secuencias analizadas.

En este contexto, interpretamos que la evolución del genoma sigue un patrón de sucesivas caídas y crecimientos en cuanto a su CR. Durante este proceso, los estadíos en los que los genomas poliploides recientes sufren mutaciones y reordenamientos, podrían ser propicios para dar origen a nuevas secuencias funcionales, proporcionando así la materia prima de la divergencia entre especies, y el crecimiento de la complejidad biológica.

La conclusión mas destacable es sin duda que, sea cual sea el genoma o el cromosoma analizado, la estructura del ADN esta fuertemente atraída hacia el estado de complejidad máxima. Dejando de lado las excepciones previamente citadas (como los poliploides recientes), observamos que la totalidad de los genomas analizados, desde virus hasta mamíferos, presentan un valor de CR muy cercano a 1.

Generalizando nuestras observaciones podemos formular las siguientes hipótesis:

- Los genomas presentan una estructura combinatoria cuasi-aleatoria independientemente del grado de complejidad biológico de los organismos.
- Los genomas poliploides recientes, tienden a recobrar una máxima complejidad a través de procesos de mutación o translocación, después de un número elevado de generaciones.
- Puesto que la estructura combinatoria del ADN es cuasi-aleatoria, la complejidad del genoma sólo puede aumentar mediante amplificación, y posterior divergencia durante el proceso evolutivo.

Nuestras hipótesis podrían verse falseadas si se encontrasen:

- Poliploides recientes con una estructura de ADN cuasi-aleatoria.
- No poliploides que muestren una estructura de ADN no aleatoria (CR baja).

Diversidad y abundancia de los elementos genéticos en genomas eucariotas

En primer lugar comparamos la cantidad de los elementos genéticos presentes en cada uno de los 548 cromosomas con lo esperado por una distribución equiprobable de dichos elementos. Sólo un 4% de los elementos genéticos fueron efectivamente observados en las proporciones esperadas por azar.

Seguimos el estudio con dos metodologías descriptivas clásicas utilizadas en ecología, las curvas de abundancia relativa de especies (RSA – por *relative species abundances*) que tienen como característica principal representar las especies únicamente acorde a sus abundancias relativas, y la relación entre el número de especies y el tamaño de la área de distribución. Para apoyar la analogía con los estudios de ecología, nos referimos a las diferente familias de elementos genéticos como especies genéticas (GS – por *genetic species*).

Sorprendentemente las RSA correspondientes a la distribución de GSs, además de presentar una forma muy similar a la observada en ecosistemas, se ajustaban muy bien a lo esperado por azar. El ajuste observado, para el 86% de los cromosomas estudiados, solo puede explicarse por un proceso balanceado de sobre- y sub-abundancias de cantidad de elementos pertenecientes a diferentes familias en los cromosomas. Esta primera evidencia de ajuste a un patrón ajeno a parámetros biológicos, fue confirmada por la correlación significativa que observamos entre el número de GSs y el tamaño del cromosoma en cuestión.

Ambos resultados evidencian que un proceso aleatorio diferente al proceso de distribución equiprobable rechazado, ajusta la abundancia y diversidad de elementos genéticos en los genomas eucariotas.

El ajuste de los datos al modelo neutro propuesto por la UNTB no pudo ser rechazado para ninguno de los cromosomas analizados (siempre que se aplican las correcciones por múltiples pruebas estadísticas). Este último resultado supone ciertamente la aceptación del principio de equivalencia de cada una de las GSs para explicar el patrón general de abundancias y diversidad en genomas eucariotas.

Si bien el ajuste de un modelo neutro no implica necesariamente la existencia de un proceso neutro responsable del patrón observado, la amplitud del ajuste plantea una pregunta: ¿por qué no somos capaces de detectar el diferencial de la selección natural previamente descrita para la diversidad de familias genéticas en los genomas? Independientemente de la respuesta, el modelo propuesto en este capítulo sirve como hipótesis nula en el estudio de mecanismos alternativos capaces de explicar la abundancia y diversidad de especies genéticas en genomas eucariotas.

Búsqueda de patrones evolutivos en grupos de genes funcionalmente relacionados

En este capítulo nos dedicamos al estudio de las presiones selectivas en grupos de genes funcionalmente relacionados, adoptando así una escala sistémica para el análisis de los genomas de mamíferos y de *Drosophila*.

Tras la aplicación del GSSA (definido en material y métodos), encontramos muy pocas funciones con sesgos significativos en sus valores dS. Sin embargo, para el resto de la variables, encontramos un gran número de resultados positivos, tanto significativamente bajos (SL) como significativamente altos (SH). En gran parte las categorías funcionales que encontramos significativamente aceleradas (ω SH) coinciden con las tendencias descritas en estudios previos basados en metodologías clásicas de agrupación de genes seleccionados positivamente (PSGs – por *positively selected genes*). Entre los resultados más destacados podemos mencionar que los módulos funcionales relacionadas con la percepción sensorial presentan valores de ω alto en primates; o relacionados con inmunidad, también significativamente acelerados, en roedores. En *Drosophila*, encontramos también muchas funciones o rutas metabólicas relacionadas con la percepción sensorial, diferentes metabolismos o proteólisis presentando valores de ω significativamente altos. De forma general, en mamíferos y en moscas, las rutas metabólicas y funciones moleculares relacionadas con el desarrollo y con la transcripción/traducción resultaron estar muy conservadas (caracterizadas como SL).

Dada la aparente relación entre nuestros resultados y las tendencias encontradas en estudios sobre grupos de PSGs, decidimos relacionar nuestras categorías funcionales con los PSGs, y dividir así nuestros resultados en dos subconjuntos, las categorías funcionales con o sin PSGs.

Resumiendo el resultado obtenido, los PSGs se distribuyen en módulos funcionales bajo diferentes escenarios evolutivos (con ω SH, SL e incluso NS), sin embargo, su distribución es significativamente sesgada hacia los grupos funcionales cambiando a tasas elevadas de ω en roedores y moscas. Por otra parte, en primates, los PSGs parecen distribuirse de manera uniforme entre los módulos funcionales, independientemente de las presiones selectivas observadas en el conjunto de genes asociados a la función.

Esta observación sugiere que los PSGs podrían estar implicados en procesos más complejos que el de su participación directa en los cambios adaptativos de los fenotipos.

A través de la estrategia de evaluación presentada en este capítulo, conseguimos aumentar el poder estadístico en el contexto del análisis de la evolución de genomas y sugerimos que los PSGs podrían cumplir funciones adicionales a la de contribuir a los cambios adaptativos en la evolución de los fenotipos.

Herramientas y programas

En este último capítulo, se hace referencia a dos herramientas que fueron desarrolladas, en un primer lugar, para responder a necesidades específicas relacionadas con el trabajo presentado en esta tesis y adaptadas, en segundo lugar, para prestar servicio al resto de la comunidad científica. También se presenta el servidor web Phylemon, un recopilatorio de herramientas enmarcadas en la filogenética, la filogenómica y los test de hipótesis evolutivas.

La primera de estas herramientas es Ecolopy, un programa diseñado para estudiar la distribución y abundancias de especies en ecosistemas, y probar estadísticamente su neutralidad mediante modelos enmarcados en la UNTB. Como característica adicional Ecolopy, ofrece la posibilidad de tratar con valores de abundancia muy grandes, como pueden ser los derivados de censos de elementos genéticos. Además del programa de libre acceso a partir del cual se puede llamar las diferentes funciones implementadas, Ecolopy se puede usar a través de un servidor web que integra los principales componentes necesarios para llevar a cabo un test de neutralidad en ecosistemas o genomas.

El segundo programa es una extensión de un paquete de programas llamado ETE, diseñado para tratar con arboles filogenéticos. Esta extensión, ETE-Evol, permite formular y probar un amplio abanico de hipótesis evolutivas, usando internamente programas como CodeML o SLR. ETE-Evol representa sobretodo un avance en el contexto de los estudios genómicos ya que permite enlazar directamente diferentes modelos evolutivos y prueba estadística (por ejemplo el test de selección positiva) a arboles filogenéticos. También resulta útil para el estudio de genes específicos, ya que propone soluciones para representar gráficamente los resultados del cómputo de diferentes modelos evolutivos.

Finalmente, se presenta la segunda versión del servidor web Phylemon. Phylemon nace naturalmente respondiendo a la necesidad de investigadores no-bioinformáticos llamados a usar herramientas de uso complejo y asociadas a cómputos pesados; y a investigadores bioinformáticos, intentando alentar el uso de sus herramientas para llegar a un público más amplio de investigadores y estudiantes. Las herramientas propuestas en Phylemon se dividen en la siguientes secciones ***1)*** Alineamiento: para alinear secuencias, ***2)*** Filogenia: para la construcción de árboles filogenéticos a partir de secuencias alineadas, ***3)*** Pruebas evolutivas: desde las pruebas de ajuste a modelos de substitución de nucleótidos o amino-ácidos hasta pruebas mas complejos como los de selección positiva, ***4)*** "*Pipeliner*": una utilidad que permite conectar gráficamente muchas de las herramientas que propone Phylemon formando así un encadenamiento de pasos necesarios, por ejemplo, para pasar de un grupo de secuencias homólogas a la representación de sus relaciones filogenéticas y ***5)*** Utilidades: sección bajo la que se agrupan herramientas accesorias cubriendo un rango de funciones, desde limpiar alineamientos hasta calcular distancias entre árboles filogenéticos.

Conclusiones

1. A lo largo de toda la diversidad de la vida, desde virus hasta mamíferos, el contenido de información de los genomas muestra valores constantes cercanos al máximo. Sólo los cambios drásticos en el incremento del tamaño del genoma, como pueden ser eventos de poliploidización o sesgos muy evidentes en el contenido de nucleotídicos, son capaces de disminuir el contenido de información del genoma.

2. Este ajuste universal de los genomas a la máxima complejidad, sugiere que los aumentos en complejidad biológica son la consecuencia de eventos anteriores de expansiones del genoma (mediante duplicación o polyploidización).

3. Del mismo modo que para la distribución de especies en ecosistemas, los genomas eucariotas presentan una distribución heterogénea de familias o "especies" genéticas: unas pocas son muy abundantes, otras relativamente frecuentes y la mayoría raras.

4. Al igual que la relación especie-área en ecología, en los genomas eucariotas se observa que el número de especies genéticas es proporcional al tamaño de los cromosomas donde se encuentran.

5. La distribución y abundancia de las familias de elementos genéticos en genomas eucariotas, ya sea funcional o repetitivo, sigue lo esperado por un modelo neutro similar al desarrollado en la teoría UNTB.

6. A través del desarrollo y puesta a prueba del GSSA identificamos las principales categorías funcionales candidatas a ser dianas de la selección natural tanto positiva como purificadora durante la evolución de linajes de especies de mamíferos y de *Drosophila.* Dado que el GSSA no está limitado por la presencia obligatoria de genes seleccionados positivamente, la lista de funciones biológicas detectadas como dianas de la selección natural es mayor a las descritas anteriormente.

7. Los genes bajo selección positiva se distribuyen en categorías funcionales con evidencias significativas de mayor, menor o igual tasa de evolución (ω) que la observada en genomas. Sin embargo se observa un sesgo significativo hacia categorías cambiando a altas tasas de dN/dS en roedores y *Drosophila.* En el caso de primates, los genes seleccionados positivamente se distribuyen de forma uniforme, sugiriendo que, en este caso los tamaños poblacionales afectan la eficacia de la selección natural como se sugiere en la teoría de mutaciones levemente deletéreas.

8. Dada esta observación sugerimos que el papel de los genes bajo selección positiva no consiste solamente en brindar cambios adaptativos a los fenotipos, sino que posiblemente sirvan para compensar mutaciones deletéreas en una red de genes relacionados funcionalmente.

9. El trabajo llevado a cabo a lo largo de está tesis condujo al desarrollo de tres herramientas bioinformáticas implementadas con la perspectiva de facilitar y extender futuras investigaciones de la comunidad científica. Estas herramientas se enmarcan en los campos de la ecología (de genomas), la filogenia, la filogenómica y la formulación y prueba de hipótesis evolutivas.