Insights into mammalian adaptive evolution through genomics data

José Luis Villanueva-Cañas

TESI DOCTORAL UPF / 2015

DIRECTORA DE LA TESI

Dra. Maria del Mar Albà Soler

Evolutionary Genomics Group
Research Programme on Biomedical Informatics (GRIB)
FIMIM (Fundació Hospital del Mar Medical Research Institute)
Universitat Pompeu Fabra

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat Pompeu Fabra Barcelona

AUG

A mis abuelos

All men by nature desire to know.

Aristotle

Try to learn something about everything and everything about something.

Thomas Huxley

Acknowledgements

Uno no llega hasta aquí solo. En realidad nunca nadie llega a ningún lado sin la ayuda y soporte de los demás. Estas líneas que siguen son mi muestra de agradecimiento para todas aquellas personas que aportaron su granito de arena en el desarrollo de este trabajo.

A tu Mar, per proporcionar-me l'oportunitat de fer una mica de ciència. També per donar-me la llibertat per a cometre errors, tan necessaris per aprendre. Però sobretot per la teva integritat, com a científica i com a ciutadana.

A en Patrick Aloy, Javier de la Cruz i Lourdes Fañanas per mostrar-me el mapa i ajudar-me a triar camins.

Als membres del grup d'EG i en especial als doctorands (molts ja doctors): Steve, Alice, Macarena, Núria, Will (thanks also for the oxford comma) i en Jorge.

A l'Alfons i Miguel, *aka* "the IT Crowd", ja que sense la seva expertesa i diligència en qüestions binaries res d'això hagués estat possible.

Anne, thanks for providing me with the great opportunity to make a short internship in your group. Sheena, thanks for being an

awesome collaborator and introducing me into the lemur world. Also to all the people of the Yoder lab that took good care of me while I was overseas, especially to Erin, Peter and Jason.

To Emma and all her irish crowd (Nicole, David, Stephen, Zixia, Una, Graham and George (albeit adopted)), for giving me a human (or a bat) perspective of science and showing me what 'the craic' was. To Jens and Salla, for letting me explore their area52. To Sergio and his wonderful wife Marcia, for treating me like a family during my dublinese adventure.

To all my MSc classmates, especially to Magda and Yorgos, whom I now call friends.

Gràcies a en David, Albert, Pau, Maria, Inés, Armand, Iván i Cinta per compartir les penes i les alegries de la professió. També per les paelles.

A tots els meus companys de despatx, passats i presents, que van compartir el dia a dia amb mi, fent la rutina més agradable: Max, Adriano, Salva, Núria, Sergi, Pau, Héctor, Babita, Juan Luis, Juan Ramón, Michael, Eneritz, Nico, Imma, Emma, Abel, Amadís… i tots els que la memòria no em retorna ara mateix.

A un trozo de madera, que unas manos artesanas convirtieron en arco, y contribuyó, en gran medida, a despejar mi mente en días tristes o turbios, allá por los bosques de Dosrius.

A todos y cada uno de mis compañeros de armas de l'Associació Catalana d'Esgrima Antiga (ACEA), en especial a los instructores Oriol y Diego, que han tenido la paciencia de enseñarme la poca esgrima que sé. A Israel y David por los viajes, las risas y alguna que otra locura.

A mis viejos amigos, en especial a Aaron, Marga, Cristina, Elena, Ruth, Borja y Jaume. A Sara, por un gran año y por el futuro.

A mis abuelos y abuela, que no llegaron a ver este día. A mi iaia Carmen, que resiste el envite del tiempo. Todos ellos miembros de una dura y noble generación que lo dieron todo, sin pedir nada a cambio.

A mi familia y muy especialmente a mis padres (papá y mamá o J. Antonio y María Jesús), mi hermano Javier y mi tía Ana. Por su apoyo, siempre incondicional en cualquier proyecto (o locura) que yo haya decidido emprender.

A todo aquel que se atreva a leer más allá de estos agradecimientos.

Abstract

Although the genome sequencing revolution is still in its infancy, we must acknowledge it as the major driver of biology since the beginning of the 21st century. The availability of a large collection of complete mammalian genomes due to high-throughput sequencing technologies allows us to begin the exploration of how the evolutionary diversification of gene content reflects the ecological adaptations of different taxa. Novelty arises in evolution through the transformation or combination of existing systems and, as shown recently, also from scratch. This thesis is centered around these different mechanisms of evolutionary innovation. It includes a common methodological part in which we propose a simple method to optimize multiple alignments and examine its effect in positive selection analyses, the exploration of the origin and evolution of mammalian-specific genes, and the study of gene regulation in mammalian adaptations (e.g. hibernation) using high-throughput technologies.

Resum

Tot i que la denominada era de la genòmica es troba encara a la seva infància, ha estat un dels principals impulsors de la biologia des del començament del segle 21. L'accés a una creixent col·lecció de genomes complets de mamífers, gràcies a les tècniques de seqüenciació massiva, ens permet explorar com la diversificació evolutiva dels gens es tradueix en les diferents adaptacions ecològiques dels diferents tàxons. La innovació apareix a l'evolució a través de la transformació o la combinació de sistemes preexistents, fins i tot, nous gens poden aparèixer a partir de regions prèviament no codificants, com s'ha demostrat recentment. Aquesta tesi s'articula al voltant d'aquests mecanismes d'innovació evolutiva. Inclou una part metodològica comuna on es proposa un mètode simple per optimitzar alineaments múltiples i avaluar-ne l'efecte en anàlisis de selecció positiva, l'exploració de l'origen i evolució de gens específics de mamífers i l'estudi de la regulació gènica en una adaptació pròpia dels mamífers (hibernació) mitjançant tècniques de seqüenciació massiva.

"A curious aspect of the theory of evolution is that everybody thinks he understands it."

Jacques Monod

Preface - Prolegomena

I was, and still am, overwhelmed by the diversity and complexity of the life that surrounds us. My fascination began very early, like every other child, asking questions about everything I could lay my eyes on. I was moved by a power-source that only kids have intact: true curiosity. Mine became mature when I first heard about Darwin's (and Wallace) theory of evolution. Evolution planted in my head the seed of diversity, of change. I was amazed of how they were able to put together many different pieces of facts and converted them into new knowledge. One that would revolutionize humanity and would have a profound effect in the way we think and perceive the world.

After several years scratching the surface of evolution and genomics, and while trying to deliver a small contribution to science, I made a precious discovery: A personal one. I committed plenty of mistakes in my daily work and spent countless hours fixing subtleties, but most importantly, I was able to learn from those mistakes (and solved some problems along the way). I found out that the only possible path to true knowledge is through humility and hard work.

Everything you know (or you think you know) can change when new evidence appears or a new way to look at it comes into play; that is indeed the beauty of science. To me science is a vision of the world. A true human endeavor, which guides the irrational creature we all have inside through the path of knowledge.

José Luis Villanueva-Cañas
Barcelona, August 2015

Table of Contents

Acknowledgements ..... **VII**

Abstract ..... **XI**

Resum ..... **XIII**

Preface - Prolegomena ..... **XV**

Table of Contents ..... **XVII**

1. Mammals **1**

1.1. Distribution **3**

1.2. Origin **5**

Extant groups *8*

1.3. Characteristics and Adaptations **16**

2. Genomics **25**

2.1. Sequence Alignments **30**

3. Molecular Innovation **32**

3.1. Sequence change and positive selection **34**

3.2. Shifts in gene expression regulation **39**

3.2.1. Hibernation **41**

3.3. Origin of New Genes **47**

3.3.1. TRGs and *de Novo* Genes **50**

History and Nomenclature *50*

How do orphan genes arise? *52*

De novo genes in eukaryotes *55*

Characteristics of de novo genes *58*

4. Results **60**

4.1. Improving genome-wide scans of positive selection by using protein isoforms of similar length 61

Introduction 63

Methods 68

Results and Discussion 75

Conclusions 89

4.2. Signatures of adaptive evolution in the iberian lynx 94

Introduction 94

Signatures of positive selection 94

Lynx orphan genes 96

4.3. Comparative genomics of mammalian hibernators using gene networks. 103

Introduction 105

Materials and methods 109

Results and discussion 114

Conclusions 128

4.4. Birth of new genes and evolutionary innovation in mammals 136

Introduction 138

Results 140

Discussion 156

Methods 160

5. Discussion 172

5.1. Sequence change and positive selection 172

5.2. Gene expression regulation: Hibernation 178

5.3. Taxonomically restricted genes in mammals 182

6. Conclusions 189

7. Future research 191

8. Annex 193

9. References 196

"Homo sapiens [are] a tiny twig on an improbable branch of a contingent limb on a fortunate tree."

Stephen Jay Gould

1. Mammals

Mammals are widespread in our little blue planet, ironically called Earth. They have managed to conquer almost every ecosystem on land and even some in the sea. According to the last International Union for Conservation of Nature and Natural Resources (IUCN), there are 5,488 known species in the Mammalia class. This is a relatively low number of species for a class to have, especially when we take into account the huge diversity in morphology, size, distribution and the number of special adaptations we observe in mammals.

The class Mammalia includes the biggest animals on both land (*Loxodonta africana*) and in water (*Balaenoptera musculus*). Their range in size spans several orders of magnitude, from the Kitti's hognosed bat (*Craseonycteris thonglongyai*), weighing only 1.5 g, to the aforementioned species, which can get heavier than 120 tons.

The term Mammalia was coined by Carl Linnaeus in the tenth edition of his *Systema naturae,* published in 1758 (Linné 1758). He took the word from Latin *mamma, which means* "breast", alluding to one of the synapomorphies of the group.

Mammals are a vital economic resource for humans. We have hunted and bred many mammals throughout our history, leading to the extinction of some species along the way due to over exploitation, and strongly modifying some others as the result of

domestication. They've been an important source of meat and materials such as bone, hides, wool, and oil. We've used them as beasts of burden, assistance animals, and even in war (horses, dogs, elephants[1]…). Bats have important roles in the agricultural ecosystems; they act as insect pest control by eating thousands of insects every night. Also some species of mammals act as pollinators.

Many mammals can transmit diseases to humans or livestock, increasing their morbidity and mortality. Rabies, which can infect through many different mammalian species, or plague, transmitted via fleas carried by rodents, are popular examples. The bubonic plague (*Yersinia pestis*) had several major outbreaks including the Plague of Justinian in the 6th century and the Black Death in the 14th century that wiped out a third of the human population in Europe.

Mammals have always been at the center of research in almost all biological fields. Some species like Norway rats (*Rattus norvegicus*) and domestic mice (*Mus musculus*) are widely used as animal models in biomedical research.

Homo sapiens and our close mammalian relatives are therefore entwined through economic, social, and evolutionary bonds.

[1] One of the species/subspecies used in war in classical times is now extinct: The North African elephant (*Loxodonta africana pharaoensis)* used by Hannibal Barca.

1.1.Distribution

Mammals are found in every corner of the Earth: from the jungles to the deserts, even in the oceans. One species has even begun to colonize local space off earth off of our planet. Mammals occupy very different ecological niches and are especially diverse in tropical and subtropical regions, following the patterns of many other groups of organisms.

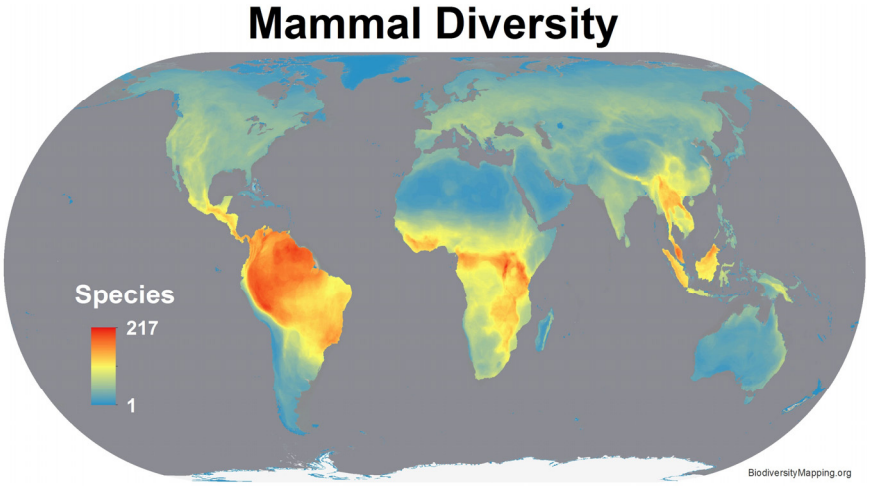


Figure 1.1a. Map of (land) mammalian diversity generated at a spatial resolution of 10x10km and using the Eckert IV equal-area projection. (Jenkins et al. 2013)

When we look at particular subgroups of mammals, we observe hotspots of diversity appearing in certain areas; however, the distribution of carnivores and rodents follows the general tropical enrichment pattern. We observe increased diversity in the islands of Sulawesi and Madagascar for small-ranged mammals, Oceania for Marsupials, South America for Chiropterans, and central Africa and the Amazon for Primates (Figure 1.1b).

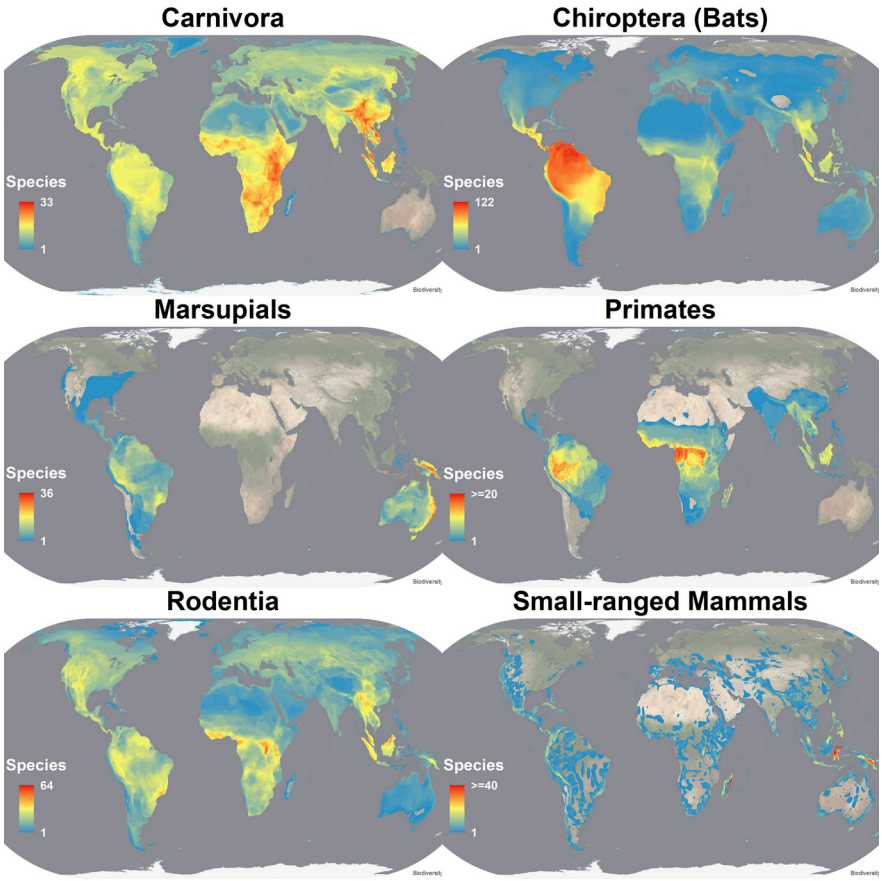


Figure 1.1b. Global maps of mammalian diversity in different groups. Modified from Jenkins et al (Jenkins et al. 2013).[2]

Since 1950 more than 914 new mammalian species have been discovered; most of them live in the tropics, where diversity was already very high (Figure 1.1a). In contrast, around 300 birds (3% of all known bird species) were discovered during the same period. Almost all of the newly discovered mammalian species belong to the most species-rich orders: Rodentia and Chiroptera.

[2] *Homo sapiens* has been removed from the Primates distribution

1.2. Origin

As skulls and teeth are the most abundant fossils, most of the taxonomy is based on them. The structure of the skull roof permits us to identify three major groups of amniotes that diverged in the Carboniferous period of the Paleozoic era: The synapsids (Greek, “fused arch”), anapsids, and diapsids (Greek, “two arches”).

Diapsids eventually led to modern, snakes, lizards, crocodilians, birds, and their ancestors. Anapsids, in the traditional meaning of the word are not a clade, but rather a paraphyletic group composed of all the early reptiles retaining the primitive skull morphology, grouped together by the absence of temporal openings. It used to contain turtles as extant species, but recent molecular studies have placed turtles within diapsids (Crawford et al. 2012).

Synapsids are characterized by lateral temporal openings in the skull. Around 360 Mya early synapsids radiated into diverse herbivorous and carnivorous forms, clustered together in a polyphyletic group formerly known as pelycosaurs. Therapsids (Greek, “beast-face”) arose from a group of synapsid carnivores and brought about morphological innovations such as an erect gait with upright limbs positioned beneath the body. The reduced stability had to be compensated with an expansion in the cerebellum and the brain to integrate the muscular coordination. Around 260 Mya, one group of therapsids, the cynodonts (Greek, “dog teeth”), diversified and gave rise to the common ancestor of all extant mammals (Figure 1.2).

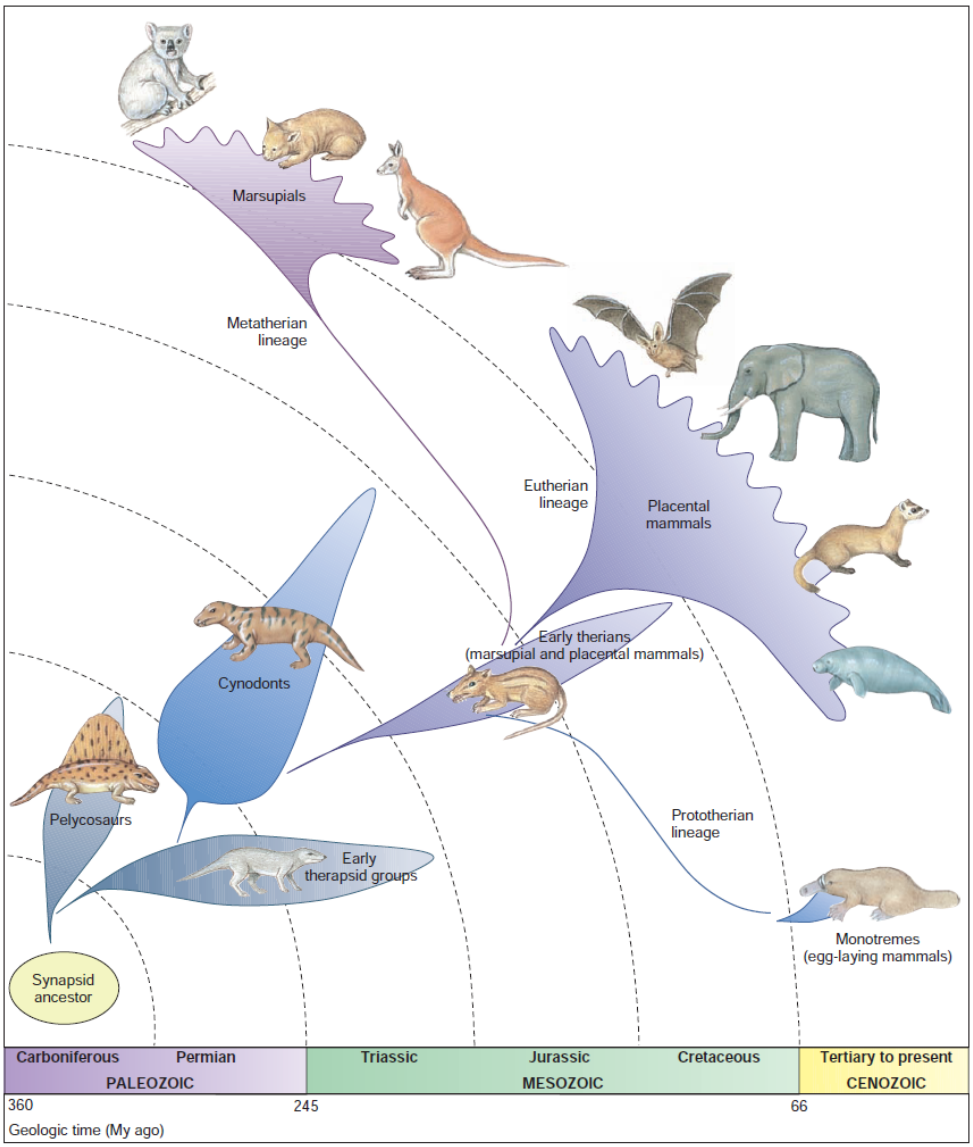


Figure 1.2. Origin of modern mammals. Early synapsids radiated extensively and evolved changes in jaws, teeth, and body form that presaged several mammalian characteristics. These trends continued in their successors, the therapsids. One lineage of therapsids, the cynodonts, gave rise in the Triassic to therians (marsupial and placental mammals). Fossil evidence, as currently interpreted, indicates that all three groups of living mammals—monotremes, marsupials, and placentals—are derived from the same therian lineage. Modified from Integrated Principles of Zoology, 14/e (Hickman et al. 2001).

Timothy Rowe defined the Mammalia class phylogenetically as the crown group of mammals (Rowe 1988). In phylogenetics, a crown group is defined as a collection of living species together with their most recent common ancestor (Figure 1.3.), as well as all of that ancestor's descendants (extant or not).

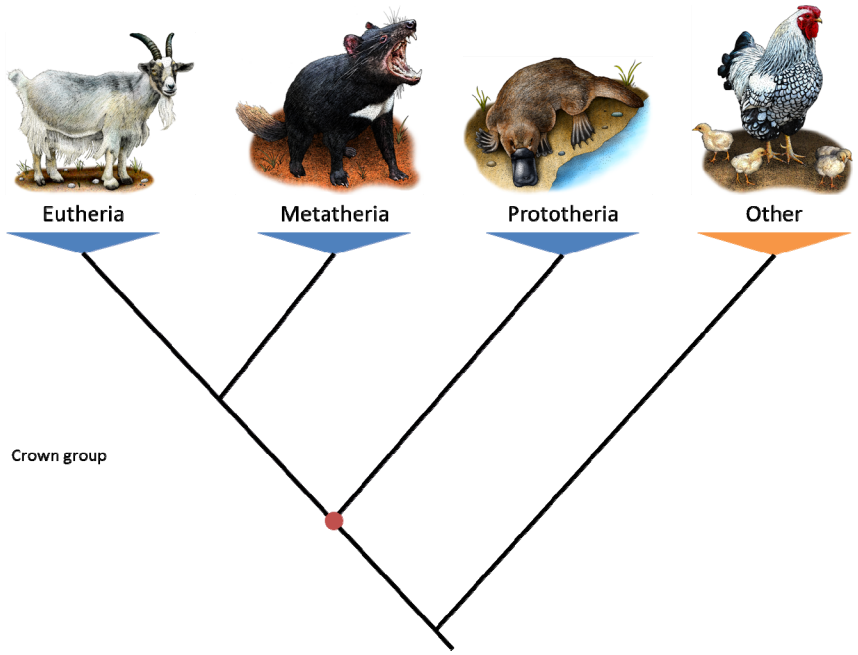


Figure 1.3. The three major extant groups of mammals (in blue), connected with their common ancestor (red dot), together with all their descendants (extant or not) form the Class Mammalia. All mammalian species belong to the same crown group (lilac). Drawings from Roger Hall.

There are three extant major groups in the Mammalia class, which are organized in two subclasses: Prototheria which includes monotremes (platypus and echidnas), and Theria which includes the infraclasses Methatheria (marsupials) and Eutheria (placental mammals).

Since the ancestor of these three extant groups lived in the Jurassic period, Rowe's definition excludes all animals from the earlier Triassic which are classified as nonmammalian synapsids, sometimes also (wrongly) called "mammal-like reptiles".

The oldest known metatherian is *Sinodelphys*, found in China's northeastern Liaoning province and dated to 125M-years ago (early Cretaceous). The fossil is nearly complete and includes tufts of fur and imprints of soft tissues (Luo et al. 2003). The oldest known fossil attributable to the Eutheria infraclass is *Juramaia sinensis*, found also in China in deposits of western Liaoning and dated 160M-years ago during the Jurassic period (Luo et al. 2011).

There are around 4,000 fossil genera described (Luo 2007) and although there were not particularly abundant in the Mesozoic, they were already quite diverse. As Luo points out, compared to the 547 known dinosaur genera, over 310 Mesozoic mammaliaform genera are now known to science. Two-thirds of them were discovered in the last 25 years (Luo 2007) including rather large animals that fed on dinosaurs (Hu et al. 2005). This new evidence has changed the classical idea proposed by George Gaylord Simpson in the 1950's in which he stated that mammals were rat-like creatures with generalized feeding and terrestrial habits, living in the shadows of the dinosaurs.

Extant groups

According to Wilson and Reeder in 2005, there were 5,416 species of mammal classified in 1,229 genera, 153 families, and 29 orders (Wilson and Reeder 2005). A detailed graphical representation of

the distribution of species along the different orders, families, and genus can be found in Figure 1.4.

Most mammalian species belong to the Eutheria group, including the most species-rich orders. The largest group in terms of raw number of species is Rodentia (which includes mice, rats, squirrels, and other gnawing mammals), followed by Chiroptera (bats) and Soricomorpha (shrews, moles and solenodons). Primates, Cetartiodactyla (including the even-toed hoofed mammals and the whales), and Carnivora (containing dogs, cats, bears, ferrets and their relatives) are next with more than 200 species in each order. While classification at the family level has been relatively stable, taxonomy at higher levels has recently been subject to substantial changes thanks to new molecular genetics analyses. For example, cetaceans are now grouped together with even-toed hoofed mammals creating the clade Cetartiodactyla (Montgelard et al. 1997); similarly, other groups have been adopted and abandoned due to molecular evidence such as the Afrotheria (Tabuce et al. 2007) and Insectivora (Wilson and Reeder 2005).

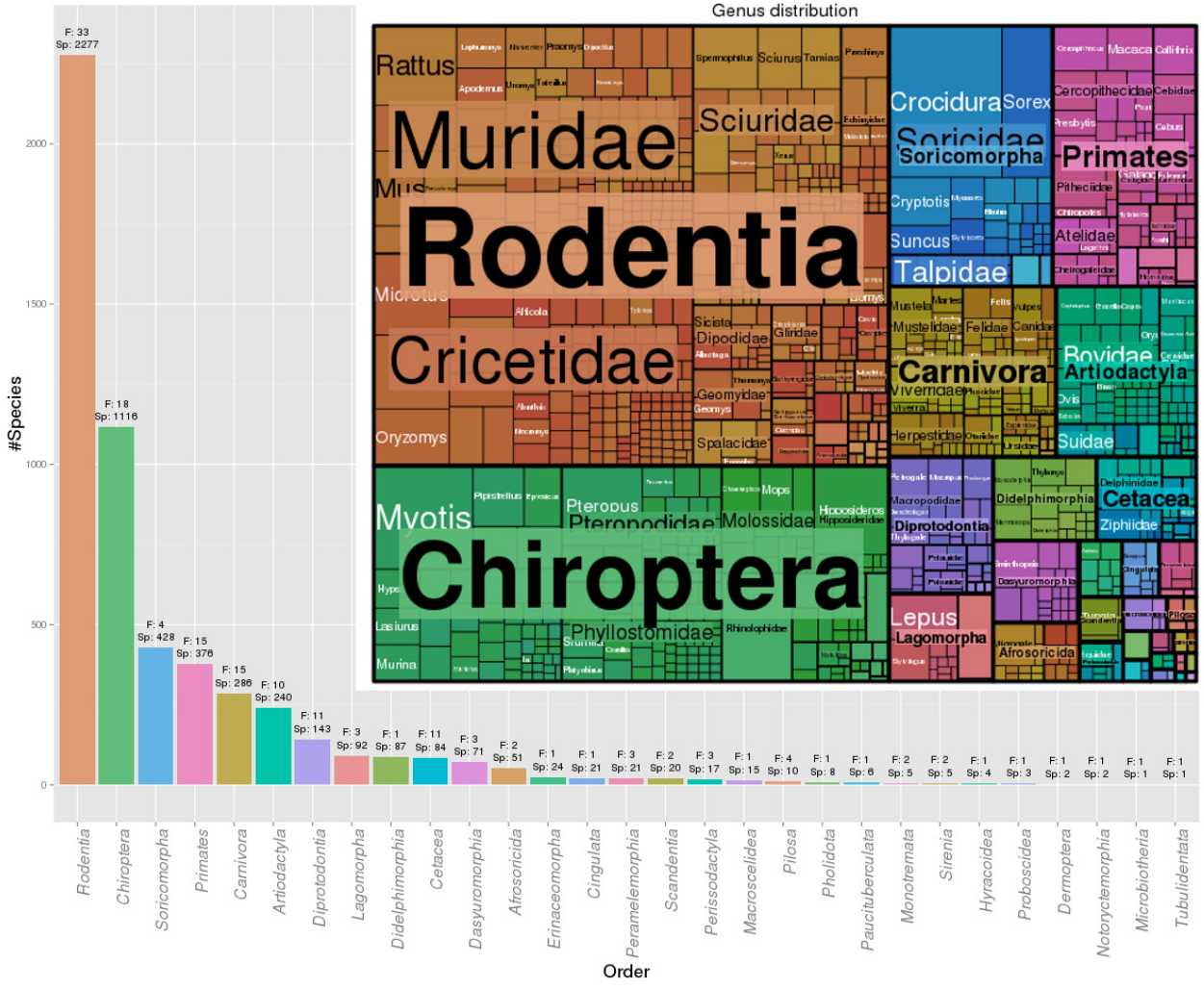


Figure 1.4. Distribution in orders, families and genus for 5,416 mammalian species. Created with data obtained from (Jones et al. 2009). Bar plot shows all mammalian orders with their total number of species (**Sp**) and families (**F**). The tree mapping of the mammalian taxonomy also includes genus information. Rectangle size shows number of species in that taxon.

Monotremata

Monotremata includes just five species which are all endemic to Australia and New Guinea: the platypus (*Ornithorhynchus anatinus*), and four species of echidna. These include the short-beaked echidna (*Tachyglossus aculeatus*) and three species of long-beaked echidna (*Zaglossus bruijni*, *Z. bartoni* and *Z. attenboroughi*). Echidnas were appropriately named after the mother of monsters in Greek mythology who was half-woman and half-snake; this is due to the echidna's strange mixture of physical features. Monotremes retain ancestral amniote characteristics (or plesiomorphic features), including egg-laying and meroblastic cleavage of the embryo, and some derived (or apomorphic) mammalian characteristics like lactation and hair. Because they have retained traits from that distant time, they give us a remarkable insight into very early mammals.

The platypus was first encountered by Europeans in 1798; only a pelt and a sketch were sent back to Great Britain as evidence of their discovery. Initially such a strange animal was considered to be a hoax constructed by a taxidermist. George Shaw, who produced the first description of the animal in 1799, stated it was impossible not to entertain doubts as to its genuine nature.

The most controversial aspect of platypus biology was whether or not they laid eggs like birds and reptiles. Almost a century passed before the egg-laying was finally confirmed in 1884. William Caldwell was sent to Sidney to study Australian mammals and *ceratodus* fish the year before. He sent a telegram to his friend professor Liversidge the 29th of August and asked him to forward it

to the British Association at Montreal. The concise telegram said: "Monotremes oviparous, ovum meroblastic", meaning that monotremes were indeed egg-laying and that their eggs were similar to those of reptiles in that only part of the egg divides as it develops (Caldwell 1887).

Monotremes are known to be the only land mammals that evolved a sense of electroreception. The electroreceptors of monotremes consist of free nerve endings located in the mucous glands of the snout. It is most highly developed in the platypus with more than 40.000 receptors, compared to the 400-2000 receptors in echidnas (Pettigrew 1999).

Although monotremes have mammary glands, they lack nipples; they nourish the newborns with milk coming from numerous abdominal pores in a depression on the mother's belly, where is lapped by the neonates. Their urinary, defecatory, and reproductive system opens into a single duct, the cloaca, like in modern reptiles.

Metatheria

Marsupials are the only living group of metatheria and comprise seven extant marsupial orders; four are Australasian and three are South American. Nearly seventy percent of the 334 extant species occur in Oceania while the remaining 100 live in America (99 in South or Central America and 1 in North America). All Australasian marsupials arose from a single marsupial migration from South America to Australia about 50 Mya, shortly after Australia had split off (Nilsson et al. 2010). Marsupial orders and their phylogenetic arrangement are shown in figure 1.5.

Marsupials are characterized by being born at a very young stage of development and by lacking a complex or true placenta to protect the embryo from their mother's immune system. Instead, they have developed a pouch that contains multiple nipples where the young are nourished and protected until they are fully developed.

The gestation period is very short, typically between 4 to 5 weeks. After gestation, the newborns climb up to the pouch using their grasping front limbs which are comparatively more developed than the rest of the body.

Marsupials have a single cloaca although they also have a separate genital tract, while most placental mammal females have separate openings for reproduction, urination, and defecation (the vagina, the urethra, and the anus).

Other common structural features with monotremes (but not placental mammals) are the presence of epipubic bones and the lack of corpus callosum between the right and left brain hemispheres.

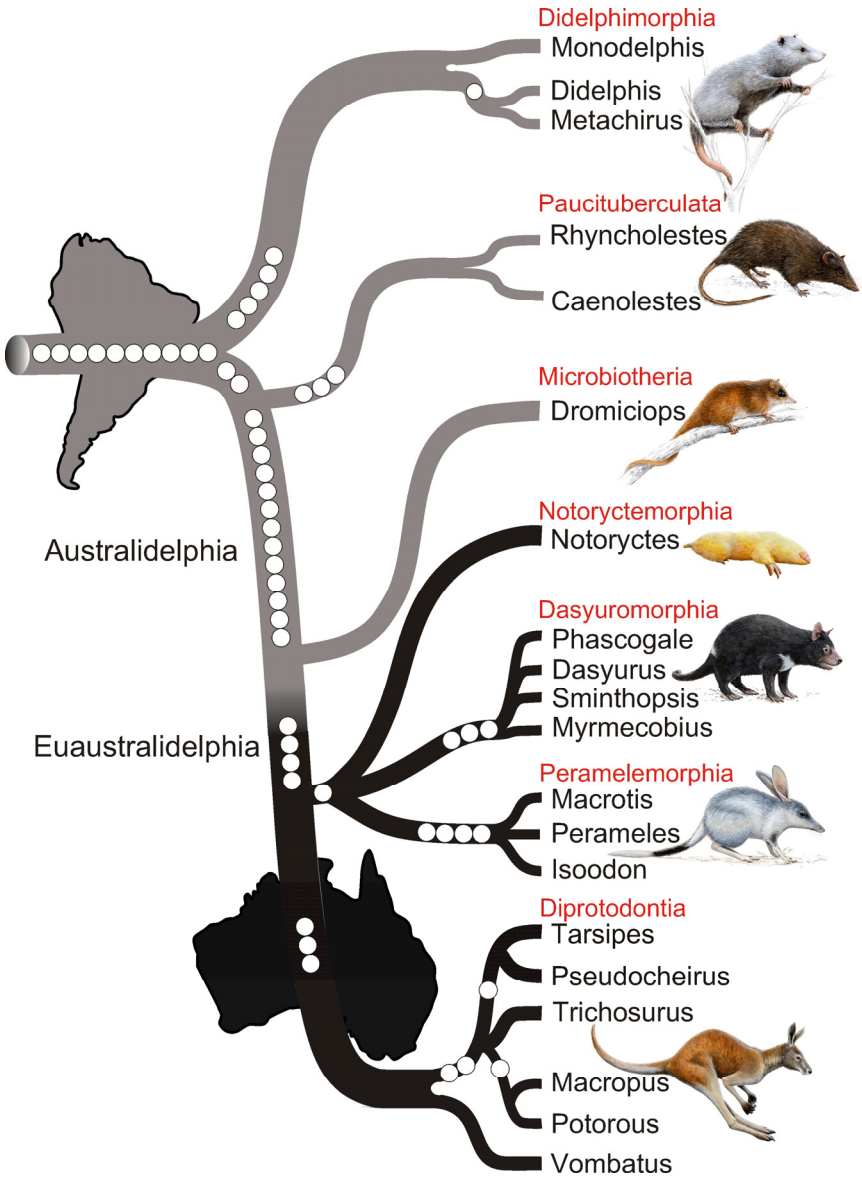


Figure 1.5. Phylogenetic tree of marsupials derived from retroposon data. The tree topology is based on presence/absence of retroposons. The names of the seven marsupial orders are shown in red. Phylogenetically informative retroposon insertions are shown as circles. Gray lines denote South American species distribution, and black lines Australasian marsupials. The cohort Australidelphia is indicated as well as the new name proposed for the four "true" Australasian orders (Euaustralidelphia). Modified from Nilsson et al 2010.

Eutheria

The term Eutheria (Greek, “True Beasts”) was introduced by Tomas Henry Huxley in 1880 and includes extant and fossil species of placental mammals. They are the most numerous group in terms of species and orders among mammals. The main defining trait of placentals is the presence of a placenta that allows the mother to carry the fetus and nourish it until it has fully developed[3]. The lack of the epipubic bone is thought to be related with this feature, allowing space for the fetus.

Other characteristics include the presence of a bony prominence on each side of the ankle known as malleolus (Latin, "small hammer") and various features of the jaw and teeth.

[3] With the notable exception of humans that are born before their bodies have developed enough to walk around

"People talk about human intelligence as the greatest adaptation in the history of the planet. It is an amazing and marvelous thing, but in evolutionary terms, it is as likely to do us in as to help us along."

Stephen Jay Gould

1.3.Characteristics and Adaptations

Mammals are homoeothermic and endothermic vertebrate animals which feed their young on milk produced by mammary glands. All have hair at some point in their lives, even if they have only a few sensory bristles on the snout like most whales. They have three middle ear bones, the malleus, incus, and stapes (more commonly referred to as the hammer, anvil, and stirrup) which function in the transmission of vibrations from the tympanic membrane (eardrum) to the inner ear. A non-exhaustive list of mammalian characteristics is summarized in table 1.1.

One of the most important defining traits of the class Mammalia is their skin structure; in particular their special modifications such as hair, glands, or horns and antlers (bovids and cervids).

Skin

In mammals, the skin is an organ of the integumentary system made up of multiple layers of ectodermal tissue, and it protects the underlying muscles, bones, ligaments, and internal organs. The skin is directly exposed to the environment and is the first line of defense

from external factors. It plays a crucial role in protecting the body against pathogens and excessive water loss.

Mammalian skin is composed of two primary layers:

1) Epidermis, which provides waterproofing and serves as the first barrier against infection. The epidermis contains no blood vessels, and cells in the deepest layers are nourished by diffusion from blood capillaries extending to the upper layers of the dermis.
2) Dermis, tightly connected to the epidermis through a basement membrane. It mostly consists of connective tissue and cushions the body from stress and strain. The dermis provides tensile strength and elasticity to the skin through an extracellular matrix composed of collagen fibrils, micro fibrils, and elastic fibers which are embedded in proteoglycans. It also harbors hair follicles, glands, and blood and lymphatic vessels.

When compared to our closest evolutionary relatives, birds and reptiles, the skin of mammals has many more glands and there is a clearer layer differentiation in the epidermis (Alibardi 2003). Mammals can also claim the thickest skin in the animal world. Steller's sea cow (*Hydrodamalis gigas*) had the thickest known skin (up to 7.5 cm) which was even used to make boats (Anderson 1995)[4].

[4] By 1768, 27 years after it had been discovered by Europeans, Steller's sea cow was extinct. The species was quickly wiped out by sailors, seal hunters and fur traders.

Glands

Mammalian skin is filled with numerous skin glands and they have the greatest variety of integumentary glands within vertebrates. Most fall into one of four different classes:

Sebaceous glands develop from the hair follicle canals and secrete a lipoidal fluid which in terrestrial mammals is water-soluble. Its main purpose seems to be to help keep the horny layer moist and pliable by admixture with absorbed sweat.

Sweat glands are a trait which is exclusive to mammals. They are tubular, highly coiled glands, spread over the body surface of most mammals. There are two types of sweat glands:

1) The **eccrine glands**, located in hairless regions like the footpads of most mammals. These secrete a watery fluid which helps in gripping the ground for walking and for grasping objects in primates. In higher primates, these glands have spread all over the body and are used to lower body temperature by sweat evaporation from the skin surface. They are either reduced or absent in rodents, rabbits, and whales.

2) **Apocrine glands** are larger than eccrine glands, with longer and more convoluted ducts. Their secretory coil is in the dermis and extends deep into the hypodermis, opening into a hair follicle or where a hair once was. In contrast to watery secretions of eccrine glands, apocrine glands produce a much more viscous, milky fluids, whitish or yellow in

color, that dry on the skin to form a film. Their activity is correlated with reproductive function and signaling, for example cycle stage and receptivity in the female or degree of maturity and relatedness of individuals within a species.

Scent glands have a very varied location and function. They are used for within species communication, warning, territorial marking or defense and can be found in interdigital regions (deers), anal regions (skunks), or the back of the head (dromedaries) among other locations.

Sebaceous glands are intimately associated with hair follicles, although some are free and open directly onto the surface. The cellular lining of the gland is discharged in the secretory process and must be renewed for further secretion. These gland cells become distended with a fatty accumulation then die, and are expelled as a greasy mixture called **sebum** into the hair follicle. Called a "polite fat" because it does not turn rancid, it serves as a dressing to keep skin and hair pliable and glossy. Most mammals have sebaceous glands over their entire body; in humans they are most numerous in the scalp and on the face.

Mammary glands, for which mammals are named, occur on all female mammals and in a rudimentary form on all male mammals. They develop by thickening the epidermis to form a milk line along each side of the abdomen in the embryo. On certain parts of these lines the mammae appear while the intervening parts of the ridge disappear. Mammary glands increase in size at maturity, becoming

considerably larger during pregnancy and subsequent nursing of young. In human females, adipose tissue begins to accumulate around the mammary glands at puberty to form the breast. In most mammals, milk is secreted from mammary glands via nipples or teats, with the exception of monotremes. Teats are also absent in male marsupials, mice, rats, and horses.

Hair

Hair is a protein filament (mostly composed by keratin) that grows from follicles found in the dermis. Thus, true hair (found only in mammals) is composed of dead keratin-packed epidermal cells.

It has a very important role in protection, temperature regulation, and sensory perception. Although hair is reduced to only a few sensory bristles in cetaceans, mammals are generally very hairy creatures. The epidermis is thinner where it is well protected by hair, but in places that are subject to much contact and use, such as palms or soles, its outer layers become thick and cornified with keratin. Hairs grow from hair follicles that are sunk in the dermis and have attached a number of structures that include the cup in which the follicle grows known as the infundibulum, arrector pili muscles, sebaceous glands, and the apocrine sweat glands. Hair follicle receptors sense the position of the hair.

There are two main kinds of hair in the pelage or fur coat: Dense and soft **underhair** which traps a layer of air for insulation, and coarse and longer **guard hair** for protection against wear and to provide coloration. In water, guard hairs become wet and adhere to each other, forming a protective blanket over the underhair.

The hair of mammals has become modified to serve many different purposes, including camouflage (with even multiple coats in arctic animals), sensory perception (vibrissae), or weapons (porcupines).

Characteristics of Class Mammalia
• Body mostly covered with hair, but reduced in some species
• Integument with sweat, scent, sebaceous, and mammary glands, overlaying a thick layer of fat
• Skull with two occipital condyles and secondary palate; turbinate bones in nasal cavity; jaw joint between squamosal and dentary bones, middle ear with three ossicles (malleus, incus, stapes); seven cervical vertebrae (except some xenarthrans [edentates] and manatees); pelvic bones fused
• Mouth with diphyodont teeth (milk, or deciduous, teeth replaced by a permanent set); teeth heterodont in most (varying in structure and function); lower jaw a single enlarged bone (dentary)
• Movable eyelids and fleshy external ears (pinnae)
• Circulatory system of a four-chambered heart (two atria and two ventricles); persistent left aortic arch, and nonnucleated, biconcave red blood corpuscles
• Respiratory system of lungs with alveoli, and larynx; secondary palate (anterior bony palate and posterior continuation of soft tissue, the soft palate) separates air and food passages; muscular diaphragm for air exchange separates thoracic and abdominal cavities; convoluted turbinate bones in the nasal cavity for warming and moistening inspired air
• Excretory system of metanephric kidneys with ureters that usually open into a bladder
• Brain highly developed, especially cerebral cortex; 12 pairs of cranial nerves; olfactory sense highly developed
• Endothermic and homeothermic

• Cloaca present only in monotremes (present but shallow in marsupials)
• Separate sexes; reproductive organs of a penis, testes (usually in a scrotum), ovaries, oviducts, and uterus; sex determination by chromosomes (males is heterogametic)
• Internal fertilization; embryos develop in a uterus with placental attachment (except in monotremes); fetal membranes (amnion, chorion, allantois)
• Young nourished by milk from mammary glands

Table 1.1. Characteristics of Class Mammalia. Modified from Integrated Principles of Zoology, 14/e (Hickman et al. 2001)

Mammals possess bigger brains than other vertebrates. On average, a mammal has a brain roughly twice as large as that of a bird of the same body size, and ten times as large as that of a reptile of the same body size (Northcutt 2002).

Size, however, is not the only difference; there are also substantial structural differences. While the hindbrain (with the exception of the cerebellum) and midbrain are fairly similar in all vertebrates, the forebrain is greatly enlarged in mammals. The cerebral cortex is the part of the brain that is most distinguishable in mammals. In non-mammalian vertebrates, the surface of the brain is lined with a comparatively simple three-layered structure called the *pallium*. In mammals, the *pallium* evolved into a complex six-layered structure.

In placental mammals, a corpus callosum also develops, further connecting the two hemispheres. The complex convolutions of the cerebral surface (gyrus) and furrows (sulci) which increase the surface area also found only in higher mammals.

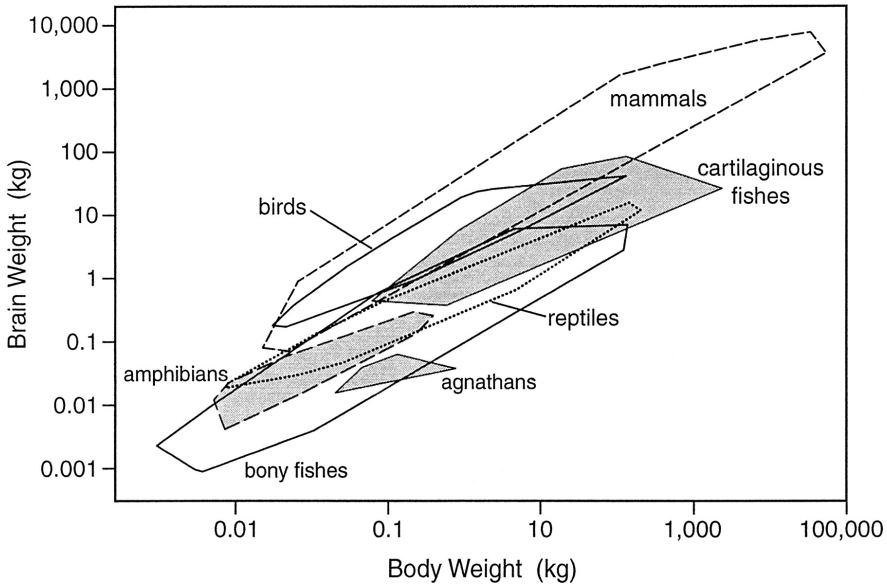


Figure 1.6. Brain weights plotted against body weights and expressed as minimal convex polygons in a double logarithmic graph for each of the major vertebrate groups. Each polygon encloses the available data for a given group. For most of these groups, there is a ten-fold range in brain size for any given body size. Furthermore, within each group, the largest brains are also generally the most complex brains. Redrawn from van Dongen (1998), taken from (Northcutt 2002).

Mammals are also considered to be highly intelligent, especially Primates, and particularly great apes. There have been many measures that have tried to estimate (with a varying degree of success) the level of cognition or intelligence by relating it to the relative size of the brain. One of such measures is the encephalization quotient (EQ), designed for mammals and defined as the ratio between brain mass and predicted brain mass for an animal of a given size (Table 1.2).

Animal taxa	Brain weight (in g)	Encephalization quotient (EQ)	Number of cortical neurons (in millions)
Whales	2,600–9,000	1.8	
False killer whale	3,650		10,500
African elephant	4,200	1.3	11,000
Humans	1,250–1,450	7.4–7.8	11,500
Bottlenose dolphin	1,350	5.3	5,800
Walrus	1,130	1.2	
Camel	762	1.2	
Ox	490	0.5	
Horse	510	0.9	1,200
Gorilla	430–570	1.5–1.8	4,300
Chimpanzee	330–430	2.2–2.5	6,200
Lion	260	0.6	
Sheep	140	0.8	
Old world monkeys	41–122	1.7–2.7	
Rhesus monkey	88	2.1	480
Gibbon	88–105	1.9–2.7	
Capuchin monkeys	26–80	2.4–4.8	
White-fronted capuchin	57	4.8	610
Dog	64	1.2	160
Fox	53	1.6	
Cat	25	1	300
Squirrel monkey	23	2.3	480
Rabbit	11	0.4	
Marmoset	7	1.7	
Opossum	7.6	0.2	27
Squirrel	7	1.1	
Hedgehog	3.3	0.3	24
Rat	2	0.4	15
Mouse	0.3	0.5	4

Table 1.2. Brain weight, encephalization quotient and number of cortical neurons in selected mammals. Modified from Roth et al (Roth and Dicke 2005)

In particular groups we can find a myriad of special adaptations such as echolocation in dolphins and some groups of bats, electroreception in monotremes (Pettigrew 1999), flight in bats, and hibernation in a wide variety of species (see section 3.1.1).

"Triumphs as well as failures of nature's past experiments appear to be contained in our genome."

Susumu Ohno

2. Genomics

The first complete genome was sequenced in 1977 by double Nobel prizewinner Fred Sanger[5]. This honor was bestowed on the ΦX174 bacteriophage, whose genome contains 5,386 nucleotides, and gave rise to the Sanger sequencing method.

Other key historical moments include the sequencing of *E. coli* (4.6Mb) in 1997 after 14 years by Blattner et al. (Blattner 1997). *S. cerevisiae* was the first genome sequence obtained by an international consortium in 1996 (Goffeau et al. 1996). Many genomes followed the consortium approach, including the first animal genome (*C. elegans*, 97Mb) two years later (C. elegans sequencing consortium 1998) and eventually bigger genomes like *D. melanogaster* (1.65 Gb) in 2000 (Adams 2000). Nevertheless, the species whose genome sequencing truly revolutionized the genomics field and allowed us to enter in the so-called 'genomic era' was our own genome, *H. sapiens* in 2001 (Lander et al. 2001).

[5] The father of the genomic era described himself as "just a chap who messed about in his lab"

A myriad of technologies encompassed by the so called High-Throughput Sequencing (HTS)[6] technologies have fostered the genomics field since the year 2000, increasing the total output and decreasing the price (Figure 1.7). Carlson correctly predicted that the doubling time of DNA sequencing technologies (measured by cost and performance) would be at least as fast as the famous Moore's law (Carlson 2003).

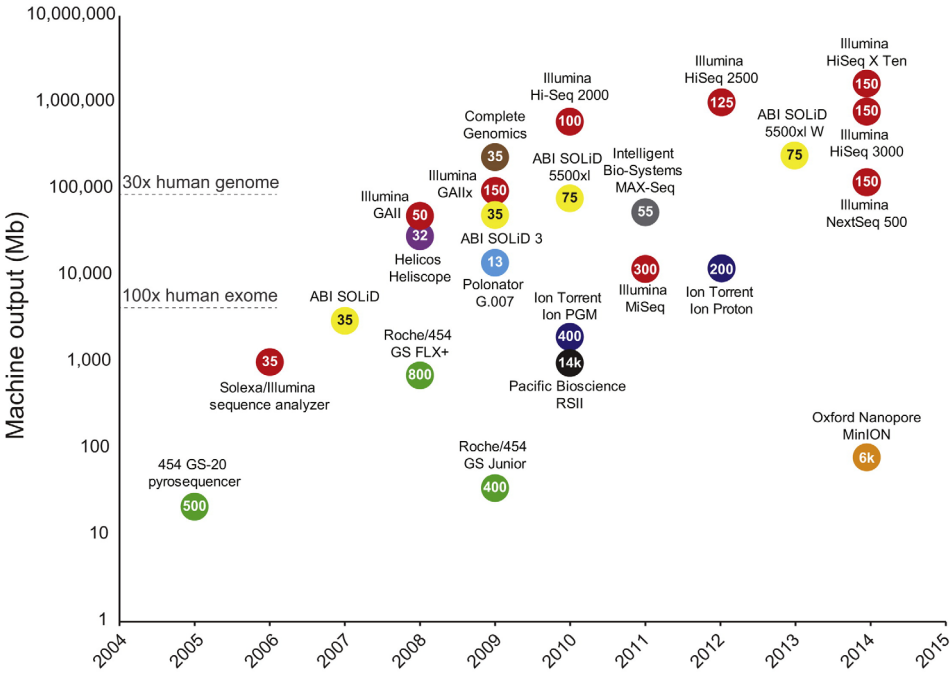


Figure 1.7. Timeline and Comparison of Commercial HTS Instruments. Commercial release dates versus machine outputs per run are shown. Numbers inside data points denote current read lengths. Sequencing platforms are color coded. Modified from Reuter et al. (Reuter et al. 2015)

[6] Also named Next-Generation Sequencing (NGS). Perhaps a name poorly chosen that will age badly as technology develops.

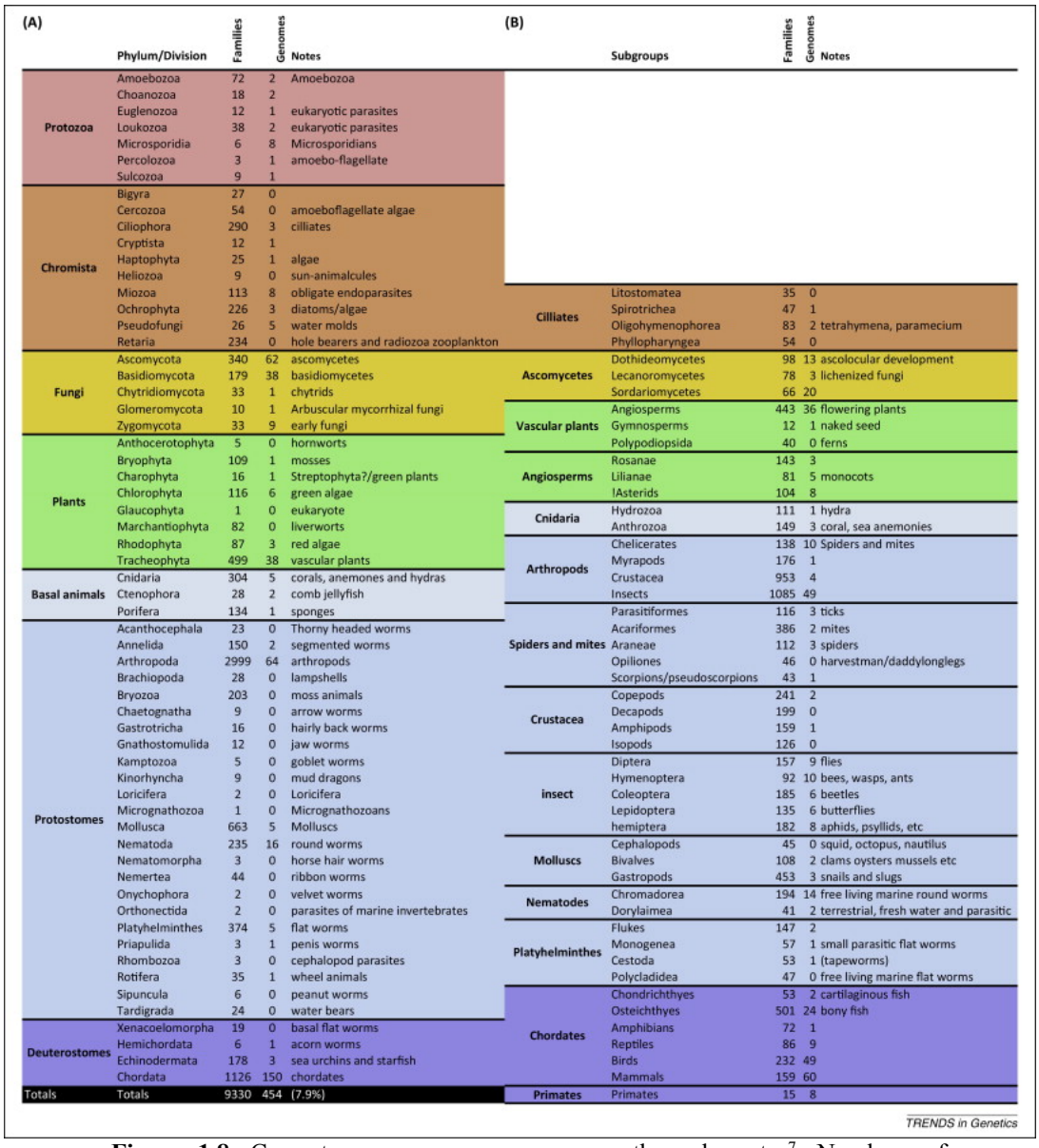
Sequencing consortiums are shrinking and some genomes (particularly small ones) will be able to be completed by a few researchers in the near future (Ekblom and Wolf 2014). The increase in the number of available genomes has boosted comparative genomics and many other fields. Genomics is considered the major driver of biology since the start of the 21st century.

When it comes to sequencing efforts, mammals are clearly overrepresented, despite their relatively low number of species and families when compared to other groups. For example at the family taxonomic category, 37.73% of the mammalian families (53.33% if we focus on primates) have at least one member sequenced, compared to only 4.51% in insects or 0.66% in gastropods. Figure 1.8 shows a snapshot of the currently sequenced eukaryotic organisms. The wide availability of mammalian genomes is clearly an advantage for comparative genomics studies.

The quality of reference genome assemblies is an important issue when interpreting differences between genomes. Coverage (read depth) is the average number of reads representing a given nucleotide in the reconstructed sequence; while a bunch of genomes like *H. sapiens* or *M. musculus* were sequenced using a hierarchical sequencing and very high coverage, other organisms remain sequenced at very low coverage (2x) and were obtained using shotgun sequencing strategies.

Another major issue corresponds to the annotation of genomes. Gene identification relies on gene prediction programs like Augustus (Stanke et al. 2006) or GeneID (Alioto et al. 2013) that

use several parameters to identify putative genes and are also capable of using external information such as EST or protein evidence to refine their gene models. Although transcriptomics data is now available for a large number of mammals, in most cases this data was not used in the annotation of their reference genome. Recently sequenced mammals incorporate this information into their pipeline, which results in improved gene annotation. Homology is also very important in the annotation of genes that are shared across species. For example, according to Ensembl v80, 17,948 out of 22,014 human protein-coding genes have a homolog in mouse (81.5%). Orthologs and paralogs are two types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication, respectively (see Koonin 2005 for a thorough review of the subject). For most of the work in this thesis we have relied on Ensembl orthology predictions. It is far more difficult to identify and correctly annotate taxonomically restricted genes (TRG or lineage specific genes) than conserved genes. Homology is an important parameter in gene discovery programs and many TRGs may be missed. Specific pipelines need to be developed for the identification of such recently evolved genes.



(A)

	Phylum/Division	Families	Genomes	Notes
Protozoa	Amoebozoa	72	2	Amoebozoa
	Choanozoa	18	2	
	Euglenozoa	12	1	eukaryotic parasites
	Loukozoa	38	2	eukaryotic parasites
	Microsporidia	6	8	Microsporidians
	Percolozoa	3	1	amoebo-flagellate
	Sulcozoa	9	1	
Chromista	Bigyra	27	0	
	Cercozoa	54	0	amoeboflagellate algae
	Ciliophora	290	3	cilliates
	Cryptista	12	1	
	Haptophyta	25	1	algae
	Heliozoa	9	0	sun-animalcules
	Miozoa	113	8	obligate endoparasites
	Ochrophyta	226	3	diatoms/algae
	Pseudofungi	26	5	water molds
	Retaria	234	0	hole bearers and radiozoa zooplankton
Fungi	Ascomycota	340	62	ascomycetes
	Basidiomycota	179	38	basidiomycetes
	Chytridiomycota	33	1	chytrids
	Glomeromycota	10	1	Arbuscular mycorrhizal fungi
	Zygomycota	33	9	early fungi
Plants	Anthocerotophyta	5	0	hornworts
	Bryophyta	109	1	mosses
	Charophyta	16	1	Streptophyta?/green plants
	Chlorophyta	116	6	green algae
	Glaucophyta	1	0	eukaryote
	Marchantiophyta	82	0	liverworts
	Rhodophyta	87	3	red algae
	Tracheophyta	499	38	vascular plants
Basal animals	Cnidaria	304	5	corals, anemones and hydras
	Ctenophora	28	2	comb jellyfish
	Porifera	134	1	sponges
Protostomes	Acanthocephala	23	0	Thorny headed worms
	Annelida	150	2	segmented worms
	Arthropoda	2999	64	arthropods
	Brachiopoda	28	0	lampshells
	Bryozoa	203	0	moss animals
	Chaetognatha	9	0	arrow worms
	Gastrotricha	16	0	hairly back worms
	Gnathostomulida	12	0	jaw worms
	Kamptozoa	5	0	goblet worms
	Kinorhyncha	9	0	mud dragons
	Loricifera	2	0	Loricifera
	Micrognathozoa	1	0	Micrognathozoans
	Mollusca	663	5	Molluscs
	Nematoda	235	16	round worms
	Nematomorpha	3	0	horse hair worms
	Nemertea	44	0	ribbon worms
	Onychophora	2	0	velvet worms
	Orthonectida	2	0	parasites of marine invertebrates
	Platyhelminthes	374	5	flat worms
	Priapulida	3	1	penis worms
	Rhombozoa	3	0	cephalopod parasites
	Rotifera	35	1	wheel animals
	Sipuncula	6	0	peanut worms
	Tardigrada	24	0	water bears
Deuterostomes	Xenacoelomorpha	19	0	basal flat worms
	Hemichordata	6	1	acorn worms
	Echinodermata	178	3	sea urchins and starfish
	Chordata	1126	150	chordates
Totals	Totals	9330	454	(7.9%)

(B)

	Subgroups	Families	Genomes	Notes
Cilliates	Litostomatea	35	0	
	Spirotrichea	47	1	
	Oligohymenophorea	83	2	tetrahymena, paramecium
	Phyllopharyngea	54	0	
Ascomycetes	Dothideomycetes	98	13	ascolocular development
	Lecanoromycetes	78	3	lichenized fungi
	Sordariomycetes	66	20	
Vascular plants	Angiosperms	443	36	flowering plants
	Gymnosperms	12	1	naked seed
	Polypodiopsida	40	0	ferns
Angiosperms	Rosanae	143	3	
	Lilianae	81	5	monocots
	lAsterids	104	8	
Cnidaria	Hydrozoa	111	1	hydra
	Anthrozoa	149	3	coral, sea anemonies
Arthropods	Chelicerates	138	10	Spiders and mites
	Myrapods	176	1	
	Crustacea	953	4	
	Insects	1085	49	
Spiders and mites	Parasitiformes	116	3	ticks
	Acariformes	386	2	mites
	Araneae	112	3	spiders
	Opiliones	46	0	harvestman/daddylonglegs
	Scorpions/pseudoscorpions	43	1	
Crustacea	Copepods	241	2	
	Decapods	199	0	
	Amphipods	159	1	
	Isopods	126	0	
insect	Diptera	157	9	flies
	Hymenoptera	92	10	bees, wasps, ants
	Coleoptera	185	6	beetles
	Lepidoptera	135	6	butterflies
	hemiptera	182	8	aphids, psyllids, etc
Molluscs	Cephalopods	45	0	squid, octopus, nautilus
	Bivalves	108	2	clams oysters mussels etc
	Gastropods	453	3	snails and slugs
Nematodes	Chromadorea	194	14	free living marine round worms
	Dorylaimea	41	2	terrestrial, fresh water and parasitic
Platyhelminthes	Flukes	147	2	
	Monogenea	57	1	small parasitic flat worms
	Cestoda	53	1	(tapeworms)
	Polycladidea	47	0	free living marine flat worms
Chordates	Chondrichthyes	53	2	cartilaginous fish
	Osteichthyes	501	24	bony fish
	Amphibians	72	1	
	Reptiles	86	9	
	Birds	232	49	
	Mammals	159	60	
Primates	Primates	15	8	

TRENDS in Genetics

Figure 1.8. Current genome sequences across the eukaryotes[7]. Numbers of eukaryotic taxonomic families represented with a reference genome assembly in NCBI. (A) Listed by phylum. (B) Breakouts for phyla with especially large numbers of taxa. The vast majority of these reference genomes are in draft status, as few large eukaryotic genomes have been finished. Modified from Richards et al (Richards 2015).

[7] Already outdated, due to the high rate of genomes published monthly.

2.1.Sequence Alignments

Multiple sequence alignment (MSA) is defined as a way of arranging multiple sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of structural, functional, or evolutionary relationships among them (Needleman and Wunsch 1970). If two or more sequences share a common ancestor, mismatches can be interpreted as point mutations and gaps in one or more lineages as insertion or deletions (InDels). The absence of substitutions or InDels in a region between two distantly related species indicates that this region may have functional or structural importance.

MSA are a fundamental and ubiquitous technique in bioinformatics; They are used for inferring phylogenies, identify homologous sequences, design PCR primers (for amplifying related genes), and assembling contigs in genome sequencing projects. Thus, alignment accuracy is crucial to a vast range of analyses and can be considered as one of the cornerstones in many biological fields.

An MSA can be obtained by the Needleman and Wunsch dynamic programming algorithm executed in n dimensions (n = number of sequences to be aligned). However, to find the global optimum for n sequences is a NP-complete problem, as the computational complexity, and the time required, increases as $O(length^n)$[8]. Therefore, MSAs are built using heuristic approaches; by far the

[8] In mathematics, big O notation describes the limiting behavior of a function when the argument tends towards a particular value or infinity, usually in terms of simpler functions.

most used approaches are based on progressive alignments (Hogeweg and Hesper 1984; Feng and Doolittle 1987). The time to perform such progressive alignments is proportional to the number of sequences: O(n) and O(n*n) for the initial pairwise alignment. This is only problematic when trying to align a very large number of sequences (tens of thousands).

In the articles presented in this thesis, two different programs have been used, depending on the objective of the study: MAFFT (Katoh et al. 2002) was used for its excellent quality/speed ratio, and PRANK+F (Löytynoja and Goldman 2008) was used when a very high-quality alignment was required for subsequent analyses.

The later uses an evolutionary framework to construct the MSA. It improves the accuracy of InDels while avoiding the alignment of non-homologous regions. PRANK+F require a phylogenetic tree in order to use the full potential from its algorithm. Although the newest versions are able to generate a tree using a fast MAFTT alignment, it is better to input the species' phylogenetic relationships if known.

A bad alignment can certainly bias or generate false positives in a wide variety of analyses, including positive selection detection (Fletcher and Yang 2010; Markova-Raina and Petrov 2011).

Trimming, which consists in removing regions with low conservation after alignment, is usually applied to get rid of badly aligned regions (Castresana 2000; Capella-Gutiérrez et al. 2009) at the cost of losing part of the sequence.

Life… finds a way.

Ian Malcolm

3. Molecular Innovation

The study of evolution becomes at some point the study of innovation; how novel adaptive traits arise and spread in biological organisms, endowing their bearers with qualitatively new abilities. Some of these major innovations become new material upon which other innovations can occur. Some well-studied examples are the epithelia in metazoans (Leys and Riesgo 2012), flowers in plants, and the development of lungs (Mess and Ferner 2010). As aforementioned we find many such innovations occurring in mammals such as homeothermy, hair or large brains. Some authors define novelty as "a feature not homologous to a feature in an ancestral taxon" (Hall 2005), yet the most common understanding of novelty is more relaxed and based on phenotypic character traits (Peterson and Müller 2013).

Most of the major evolutionary innovations are easy to distinguish at the phenotypic level, and much of the current Linnean taxonomy is, in fact, based upon them. However, our understanding of the mechanisms that underlie the genesis of novelty remains limited. The geneticist Hugo de Vries already pointed out this issue in 1905: *'natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest'* (de Vries 1905).

A new phenotype can arise through many different mechanisms, including regulatory changes, through the acquisition of new genes or the modification of existing genes. The relative contribution of each of these three mechanisms in the origination of adaptive traits is unknown, and is likely to vary among different groups of organisms (Carroll 2005).

In a recent study in threespine stickelbacks it was observed that regulatory changes were predominant in the adaptive evolution of this particular group (Jones et al. 2012). It has also been observed that mutations in genes can have pleiotropic effects, affecting fitness in an environment-dependent manner (Soskine and Tawfik 2010). This is potentially very important in the origination of morphological variation. The pleiotropic effects of an adaptive single-base pair substitution have been characterized in the bacterium *Pseudomonas fluorescens* (Knight et al. 2006). However, which portion of point mutations has small pleiotropic effects and which one has no effects remains unclear (Stern 2000).

The last major player in genetic novelty would be the origination of new genes. Traditionally, gene duplication was thought to be the only plausible mechanism of gene birth, but there is an accumulating amount of evidence that some genes can originate from scratch in previously non-coding regions (Tautz 2014; Schlötterer 2015).

It is becoming increasingly clear that genomes are highly dynamic entities; however, the genetic and molecular bases of adaptive evolution are still largely unknown.

"It is not the strongest of the species that survives, not the most intelligent that survives. It is the one that is the most adaptable to change."

Charles R. Darwin

3.1.Sequence change and positive selection

A classical issue in evolutionary biology is to understand the forces that govern the evolution of species. Mutation[9] is one of the four basic mechanisms (the others being genetic drift, selection, and migration) by which sequences evolve; generating the necessary standing variation upon which natural selection can act. Mutations can arise either from the misincorporation of nucleotides during DNA replication or from DNA lesions that remain unrepaired. Due to the redundancy of the genetic code, exhibited as the multiplicity of three-nucleotide combinations specifying an amino acid (codon), substitutions in coding sequences can be classified in two categories: synonymous substitutions and nonsynonymous substitutions.

Synonymous or silent substitutions occur when the mutated triplet codes for the same amino acid. For the sake of simplicity,

[9] Mutation and substitution are often used interchangeably but there is an important difference. Nucleotide mutations are base changes (synonymous or non-synonymous) where the mutant and wild-type forms coexist in a population. A nucleotide substitution is a fixed base change in a population or species.

synonymous substitutions are often assumed to be selectively neutral, although there is evidence that up to 40% might be under the action of selection, often because they affect splicing and/or mRNA stability (Chamary et al. 2006). Since mammalian species have relatively small population sizes, it can be assumed that synonymous substitutions are "effectively neutral" because purifying selection is less effective in smaller populations.

Nonsynonymous substitutions can truncate proteins (through a stop codon) or modify them, often producing deleterious effects. They can be further classified into conservative mutations or non-conservative mutations, depending on whether the properties of the new amino acid are equivalent to its predecessor (e.g., hydrophobicity, polarity, etc.) or not.

Deleterious mutations will not reach a high frequency in a population and will be purged by natural selection. Thus, the majority of observed substitutions are believed to be slightly deleterious or neutral, and rarely, advantageous (OHTA 1973; Akashi et al. 2012). The acquisition of new functions or the optimization of existing ones by sequence change may confer an advantage to an organism, increasing the associated allele frequency in a population through natural selection.

The search for signs of positive or adaptive selection at the molecular level is a topic of great interest but also a difficult one, as most observed changes are neutral or (slightly) deleterious. There are two families of methods used to detect positive selection: population methods, based on analyzing allele frequencies within a species, and divergence methods, based on comparing patterns of

synonymous and nonsynonymous changes in protein coding sequences. Two interesting reviews with comparisons and limitations of the different methods are Biswas et al. (Biswas and Akey 2006) and Jensen et al. (Jensen et al. 2007).

Remarkable examples of positively selected locus in humans are related with hypoxia at high altitude (Petousi and Robbins 2014) and skin pigmentation (Sturm 2009). Cetacean myoglobins, which are critical for increasing oxygen storage capacity and prolonging dive time, have been found to be under adaptive selection (Dasmeh et al. 2013). While polymorphism-based analyses are used to study selective events within populations, here we will focus on substitution-based methods, aimed at inferring species or lineage-specific adaptations.

The rate of evolution[10] is a measurement of the change of a character in an evolutionary lineage over time. In protein sequences the non-synonymous substitution rate (dN) is the number of nonsynonymous substitutions per nonsynonymous site and the synonymous substitution rate (dS) is the number of synonymous substitutions per synonymous site.

The ratio between the two can be used to infer general selection regimes. Under neutrality we expect a dN/dS of 1; while a dN/dS smaller than 1 indicates purifying selection, reflecting functional constraints. When dN/dS is higher than 1 positive selection may be acting. An important limitation is that adaptive selection is often

[10] Haldane defined a 'darwin' as a unit to measure evolutionary rates, using quantitative characteristics; one darwin is a change in a character by a factor of e in one million years.

limited to one or a few sites, which get diluted when looking at general dN/dS values.

One of the most popular packages for estimating dN/dS values in phylogenies is CodeML, which uses a maximum likelihood (ML) approach and it is included in the package PAML (Yang 2007). This method takes into account the fact that nonsynonymous mutations are thrice as likely to occur as synonymous mutations and that transitions (A to G or C to T) are more common than traversions. It also considers the possibility of recurrent changes in the same position, which are not observed in the extant species but can be inferred from ancestral reconstructions. The saturation of sequence changes over time decreases the accuracy of the substitution rate estimates at long phylogenetic distances.

In the articles presented in this thesis we have used two different models included in CodeML:

- Free-ratio models, that estimate the dN, dS, and dN/dS (sometimes called w) in a given phylogeny.
- Branch-site models, where the dN/dS ratio is allowed to vary across branches and also among sites.

The branch-site test (Zhang et al. 2005) has been designed to detect positive selection affecting a few sites in one or a few lineages. In this test the branch for which we want to detect positive selection is named foreground (labelled with '#' in the tree), and the rest of branches are considered background branches. The alternative and null models are detailed in table 3.1.

Site Class	Proportion	Background	Foreground
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 \geq 1$
2b	$(1 - p_0 - p_1)\, p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 \geq 1$

Table 3.1. Alternative hypothesis model for the branch-site test of positive selection. The null model is the same except ω2 = 1 is fixed. Taken from Yang et al. (Fletcher and Yang 2010).

To determine the most likely model, a likelihood ratio test statistic is computed as twice the log likelihood difference between both models and then the result is compared with the χ^2 distribution (one degree of freedom). Therefore, positive selection is inferred in the foreground branch if the likelihood of the alternative value is significantly higher than the one for the null model. In addition, when the null model is rejected, the test also indicates which specific amino acids (with a confidence value) are under positive selection, using a Bayes empirical Bayes (BEB) approach (Fletcher and Yang 2010).

3.2.Shifts in gene expression regulation

While some phenotypes and adaptations can be explained by protein coding sequence changes, many others are likely to be due to changes in the regulation of gene expression (Carroll 2005).

Some genes are constitutively expressed, regardless of the environmental conditions. Such genes usually control basic functions such as DNA replication or repair, or the organism's metabolism. In contrast, regulated or facultative genes are needed only occasionally and their expression is controlled or influenced by the environment or other genes.

The importance of gene expression regulation in phenotypic change was postulated in 1975 after noting that, despite big anatomic and cognitive differences between human and chimpanzee, the majority of proteins were highly conserved between the two species (King and Wilson 1975). Even gradual change, like beak size and shape in birds is likely to be caused by changes in the timing and level of expression of major developmental switch genes (Abzhanov et al. 2004; Wu et al. 2004). Other recent examples of differences between species governed by regulatory mutations are found in male wing pigmentation pattern in sibling *Drosophila* species (Prud'homme et al. 2006) or body pigmentation between *D. yakuba* and *D. santomea* (Jeong et al. 2008)

These examples indicate that regulatory mutations play an important role in inter-species phenotypic differences. Moreover, it is increasingly realized, that purifying selection conserves much more than just the protein-coding part of the genome, providing

indirect support for gene regulation as a major innovation mechanism (Bernstein et al. 2012).

Genes commonly function together; it is the concerted expression of distinct sets of genes that is often phenotypically relevant. One of the better studied pathways is the tricarboxylic acid (TCA) cycle (Krebs cycle) – present in all aerobic organisms. In this gene network, carbohydrates, lipids, and proteins are oxidized into carbon dioxide, obtaining energy in the form of adenosine triphosphate (ATP). In this example the cycle is tightly regulated by product inhibition, substrate availability and the enzyme interaction.

The 'rewiring' of gene regulatory networks can enable changes at any scale – from subtle intra-specific morphological variations (Prud'homme et al. 2007) to the creation of novelty at phylum level (Shubin et al. 2009).

Microarrays and RNAseq technologies have enabled the study of gene expression at large scale allowing us to compare different tissues and conditions. In this thesis we have studied the phenomena of hibernation, an extreme mammalian adaptation, in which gene regulation is thought to be a major player.

Swallows, for instance, have been often found in holes, quite denuded of their feathers, and the kite on its first emergence from torpidity has been seen to fly from out some such hiding-place.[11]

Aristotle

3.2.1. Hibernation

Endothermy is energetically expensive and sometimes the energy supply becomes scarce (e.g. winter, drought…). The animals usually insulate their bodies to minimize heat losses and maintain homeothermy in a constant high temperature, but this is usually not enough. Many animals have adopted migration (especially birds) as response to the lack of resources. As terrestrial locomotion is the most energetically expensive, is not surprising that very few terrestrial mammals have adopted this behavior. Those terrestrial mammals that migrate are usually large organisms like the reindeer (*Rangifer tarandus*), which migrates in Alaska, and several species which migrate in the Serengeti National Park such as the blue wildebeest (*Connochaetes taurinus*), African elephants (*Loxodonta africana*), and zebras (*Equus quagga*).

Hibernation is a strategy adopted by mammals with a large surface-to-volume ratio (small) and high mass-specific metabolic rate (Figure 3.1). Small mammals have higher heat loss than large mammals with a thicker insulation and require much more energy per Kg to maintain homeothermy. Heat loss is a function of the T_b-

[11] Aristotle deduced that swallows and other birds hibernated during winter. He considered migration an inconceivable idea because they were too small. His conclusion was accepted wisdom for over 2,000 years.

T_a differential, thermoregulatory costs at low Ta becomes prohibitory expensive for small mammals. Larger bodies can also store more relative fat, although they are also more energy-expensive to reheat. These basic energetic reasons explain why hibernation and daily torpor are strategies adopted mostly by small animals (and not by large animals[12]), where the energy savings are much higher. Hibernation is a strategy adopted by many groups of the three deepest branches of the mammalian phylogeny and has been accounted for or studied in at least 93 different mammalian species (Ruf and Geiser 2014).

It is important to differentiate between torpor and hibernation; the terms have been used interchangeably in the literature, but they are not entirely the same thing. Torpor is defined as a physiologically controlled depression of metabolic rate and activity. It starts with a regulated lowering of rate and depth of breathing and heart rate, diminishing metabolic rate as well as the Tb (Temperature of the body), which can approach ambient temperature. After a period of low, yet stable, metabolic activity (at low Tb), the animal elevates respiratory, heart, and metabolic rates to initiate rewarming. The rewarming is usually performed through non-shivering thermogenesis that occurs in brown adipose tissue (brown fat), where the uncoupling protein-1 (UCP1) allows the uncoupling of

[12] Perhaps the exception to this rule are bears, such as the black bear (*Ursus americanus*), which also exhibits a form of hibernation, although at relatively high T_b of ~30°C.

protons moving down their mitochondrial gradient from the synthesis of ATP and allowing energy to be dissipated as heat (Cannon and Nedergaard 2004).

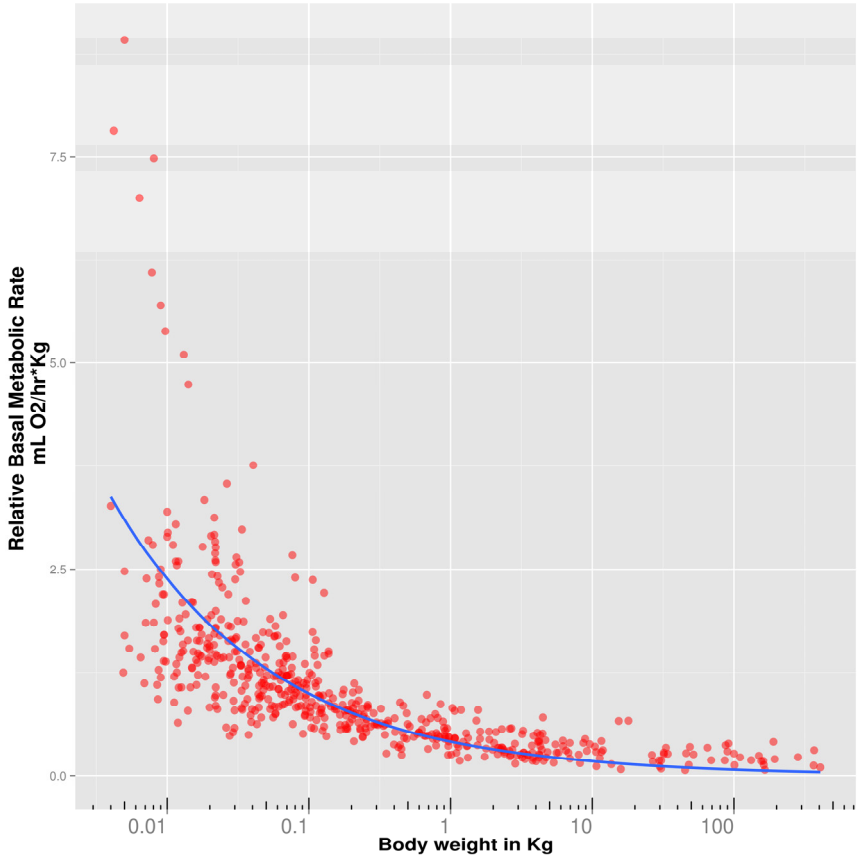


Figure 3.1. Relationship between mass-specific or relative metabolic rate and body weight for a range of 572 mammals. Smaller animals require much more energy per Kg to stay alive. Metabolic rate of a mammal varies in rough proportion (0.7) to the relative surface area rather than to body weight. Data obtained from PanTHERIA (Jones et al. 2009).

Hibernation consists in a concatenation of multi-day bouts of torpor. Torpor bouts in deep hibernators have classically been defined as those lasting more than 1 day. During deep hibernation, the

multiday periods of torpor are punctuated periodically by arousals to euthermia, known as interbout arousals (IBA), giving rise to a pattern of heterothermy (Figure 3.2).

Typically, hibernation is preceded by a hyperphagic stage where animals store all the necessary energy they will use during hibernation in the form of lipids in the white adipose tissue (WAT). The simplified basic stages of hibernation are 1) entering; 2) torpor; 3) arousing from torpor (IBA); 4) interbout euthermia (IBE).

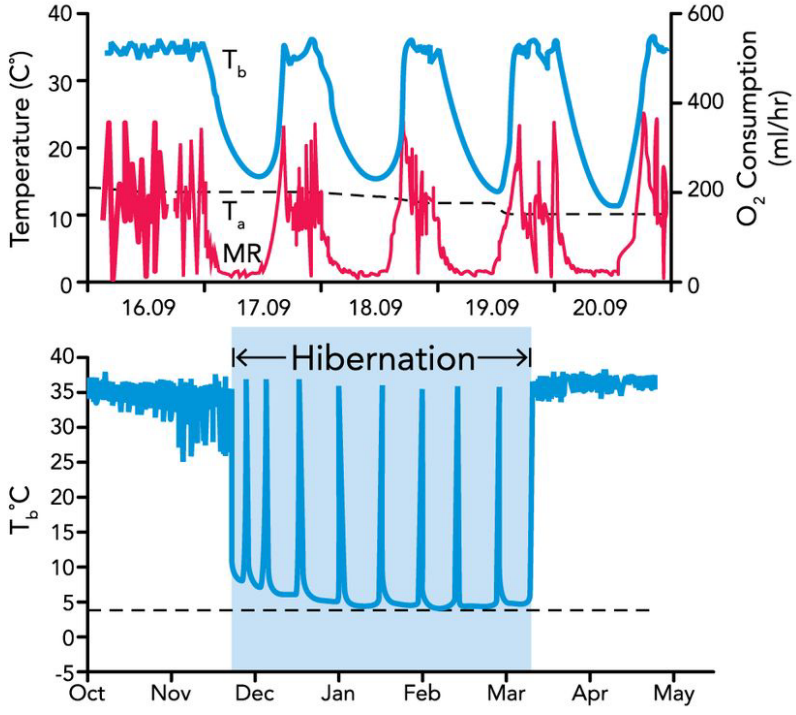


Figure 3.2. Top: daily torpor in *Glis glis* showing body temperature (Tb), ambient temperature, and metabolic rate. **Bottom:** ground squirrel hibernation: Tb across 8 months illustrates the homeothermic and heterothermic (blue shadow) periods. Modified from Breukelen et al (van Breukelen and Martin 2015).

This seasonally expressed period of heterothermy represents a huge deviation from mammalian homeostasis and can save up to 90% of the energy required to remain euthermic during resource scarcity (Wang and Lee 2011). Most of the energy spent during the torpor bouts is used in the IBA, resuming euthermia through non-shivering thermogenesis.

James E. Lovelock, best known for his Gaia theory, did some experiments in the 1950's in a period where interest in space travel was fueled by the space race. He cooled hamsters and rats to 0ºC and successfully reanimated them using microwave diathermy or a hot metal spatula applied to the chest (SMITH et al. 1954; Andjus and Lovelock 1955). Indeed, much of the knowledge about torpor physiology in mammals was produced by research on rodents. Mice and other murid rodents have been used as models for daily torpor, whereas ground squirrels are commonly used for hibernation studies (Carey et al. 2003). Well studied hibernators are the arctic squirrel and several other squirrels from the same genus (*Spermophilus*), bats like *Myotis lucifugus,* and bears. Some authors do not consider bears true hibernators, since their Tb values during torpor remain only a few degrees below normal temperature values (30-34ºC) (Tøien et al. 2011).

There is an abundance of literature about physiological changes during hibernation but not much about how it evolved. In a seminal work Srere et al. proved that the torpor length was correlated with both the concentration of mRNA and protein of the alpha2 macroglobulin gene. They were the first ones to hypothesize that the hibernation phenotype was manifested via a small number of

regulatory changes in existing mammalian genes and not through the acquisition of new genes (Srere et al. 1992).

The hibernation phenotype has a patchwork distribution in the mammalian phylogeny; we find hibernators in at least fourteen orders of mammals with representatives in all three of the major branches of the Class Mammalia: Monotremata, Marsupialia, and Placentalia (Carey et al. 2003; Andrews 2007). However, it is unknown if distantly related species use the same key pathways for activating and maintaining the hibernation phenotype and what is the contribution of sequence change and positive selection in the appearance of the hibernation phenotype.

We have formerly seen that parts many times repeated are eminently liable to vary in number and structure; consequently it is quite probable that natural selection, during the long-continued course of modification, should have seized on a certain number of the primordially similar elements, many times repeated, and have adapted them to the most diverse purposes.

Charles R. Darwin

3.3. Origin of New Genes

The third pillar in the origination of new adaptive traits is the acquisition of new genes. But, what is exactly considered a gene in 2015? Gerstein et al. offers an interesting review of what has been historically considered a gene and proposes a very relaxed definition: *A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products* (Gerstein et al. 2007). Although in our work we have primarily focused on protein coding genes, it is important to bear in mind that some new genes may have non-coding functions, including regulatory roles performed by long non-coding RNAs (lncRNAs), microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), and other short RNA.

A gene duplication is an event in which one gene gives rise to two or more genes that are identical or very similar to each other. These genes are called paralogs. Haldane pointed out gene duplication as a mechanism for the generation of new genes, in which one of the copies can be altered without negative

consequences. He also argued that multiple copies would protect the bearer against harmful mutations (Haldane 1932). Susumu Ohno refined Haldane's contributions and proposed a model (later named neofunctionalization by Force et al. 1999) where one copy retains the original (pre-duplication) function of the gene, while the second copy acquires a distinct function (Ohno 1970). Gene duplicates can arise by retrotranspositon of mRNAs back into the genome or unequal crossing over. Other genes with homology to preexisting functional elements are originated by horizontal gene transfer, gene fission/fusion and exaptation from mobile elements (Long et al. 2003; Kaessmann 2010).

Poliploidy, or whole genome duplications, can also be found in nature and are particularly common in plants (Masterson 1994). The 2R hypothesis, suggesting that the genomes of the early vertebrate lineage underwent two complete genome duplications, was first proposed by Ohno (Ohno 1970) and further confirmed using genomic data (Kasahara 2007; Smith et al. 2013).

An interesting example of adaptation linked to formation of a duplicated gene can be found in colobine monkeys of the genus *Pygathrix* (Douc langurs). With a diet based on leaves with high fiber content, they rely on bacteria to break down the cellulose. Douc langurs (like ruminants) recover nutrients by breaking and digesting the bacteria with various enzymes, including pancreatic ribonuclease (RNASE1). RNASE1B gene, which duplicated from RNASE1 gene around 4.2 Mya, acquired enhanced ribonucleolytic activity in a context of increased demands for digesting bacterial

RNA, so that nitrogen could be recycled efficiently (Zhang et al. 2002).

However, not all genes come from already existing genes. Some genes arise from previously non-coding sequences (*de novo* genes), which is a mechanism of new gene origination that was thought to be rare (Long et al. 2003). The contribution of duplicated genes to sequence and fold innovation is limited by the sequence similarity to other existing genes (Ohno 1982). In *de novo* gene evolution completely new sequences are expressed and can take new roles (see 3.3.1 TRGs and *de novo* genes). In a study in mammals, Meunier et al. found that in the origination of miRNA genes, *de novo* gene origination and duplication mechanisms showed similar frequencies (Meunier et al. 2013). In this thesis we have focused on young mammalian genes without homology to known proteins in other vertebrates. That is, genes that appeared at the root of mammals or more recently (<200Mya) and cannot be related to any existing proteins in other organisms through sequence similarity; therefore, many of these genes are likely to have originated de novo from previously non-coding parts of the genome. Many of these genes are probably related to key recent adaptive processes.

But naturalists are now beginning to look beyond this, and to see that there must be some other principle regulating the infinitely varied forms of animal life.

Alfred Russel Wallace

3.3.1. TRGs and *de Novo* Genes

History and Nomenclature

In 1996, when the complete genome of *Saccharomyces cerevisiae* was released it was found that many genes (30-35% of all ORFs) had no detectable homologues in other species. The term 'orphan gene'[13] was coined. It was hypothesized that many orphan genes would find homologues as more genomes were sequenced. It was also proposed that a irreducible fraction of them would be specific of each organism, possibly because they were fast evolving genes, making it hard to find any homologues (Dujon 1996).

For a long time it had been considered impossible or extremely unlikely, that new genes could emerge out of scratch. In his rather general essay about evolution in 1977, François Jacob states: *'The probability that a functional protein would appear de novo by random association of amino acids is practically zero[14]. In organisms as complex and integrated as those that were already living along time ago, creation of entirely new nucleotide sequences*

[13] Orphans were defined as genes without known function and without structural homologs of known function.

[14] The author does not offer a formal calculation of probability in this essay.

could not be of any importance in the production of new information' (Jacob 1977).

Much has happened since these ideas proposed by F. Jacob and Dujon. As the number of sequenced genomes increased and the taxonomic sampling became more complete the term 'orphan' lost precision. Some of the previously orphan genes could now be found in related species and were no longer 'orphan', however, homology could not still be found outside certain taxonomic rank (Genus, Family, Order, etc). Some of these taxonomically restricted genes (TRG) were found to be important for taxon-specific adaptations, such as the stinging cells in Cnidaria (Khalturin et al. 2009).

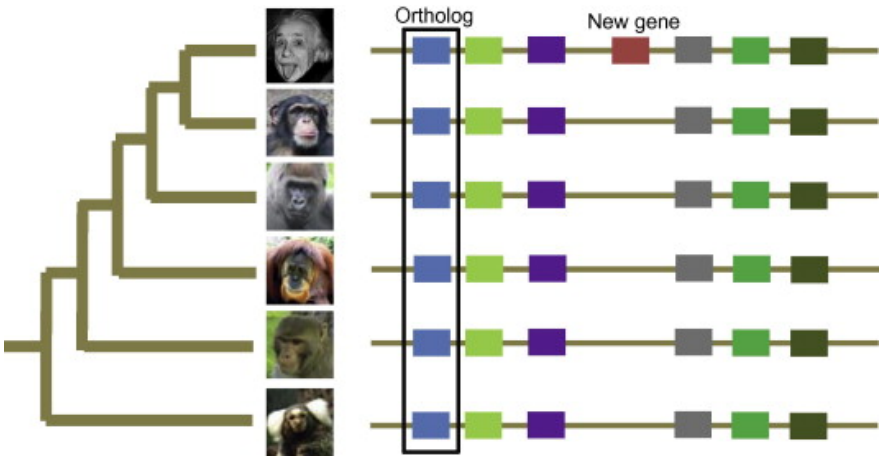


Figure 3.3. Phylogenetic tree of primate species that have high quality genome sequences. Genome data from closely related species help identify new genes based on the synteny of orthologous genes. New genes that have no homology with any other gene are likely to have originated de novo. Taken from Wu et al. (Wu and Zhang 2013).

How do orphan genes arise?

There are four different proposed mechanisms for orphan genes or TRGs:

1) **Horizontal gene transfer** (HGT) or the acquisition of a gene from other non-sequenced organisms, such as bacteria or virus.
2) **Overprinting** caused by frameshift mutations that could generate an entirely different protein with almost no change in the protein coding DNA sequence (CDS) (Okamura et al. 2006).
3) Gene duplication followed by fast evolution (**duplication-divergence model**), beyond the threshold of recognition of the parental gene through similarity (Tautz and Domazet-Lošo 2011).
4) ***De novo*** **genes** were defined as genes that originate from previously non-coding sequences.

HGT is an important mechanism in prokaryotes but is possibly negligible in eukaryotes (Wissler et al. 2013). TRG within mammals arisen through the duplication-divergence model are unlikely to be found because a small conserved motif in a sequence is enough to detect homologues by Blast. Albà and Castresana showed through simulations that Blast, when applied within eukaryotes, only misses homologues of extremely fast-evolving sequences, which are rare in mammalian genomes, as well as sequences evolving homogeneously or pseudogenes (Albà and Castresana 2007). Overprinting is relatively easy to detect if we know the chromosomic position of the gene. Therefore we think that most mammalian-specific TRGs are likely to have been

originated *de novo,* especially when very conservative filters are applied.

An enlightening model was proposed by Carvunis et al. where the birth and death of genes would form a continuum (Figure 3.3), instead of the traditional binary model (coding or not coding) where boundaries are applied in gene annotation (length, expression level, ORFs, conservation and sequence composition). In this continuum model we would have an intermediate proto-gene class, populated by potential ORFs that are not expressed, transcription without translation or different combinations of the different factors involved in gene annotation (Carvunis et al. 2012). This proto-gene class would be bidirectional, where canonical genes could become pseudogenes. This continuum would provide the reservoir from which *de novo* genes may evolve. It has also been proposed that orphan genes are often short-lived (Tautz and Domazet-Lošo 2011; Wissler et al. 2013; Palmieri et al. 2014). Whereas most *de novo* genes may become pseudogenes, others may not only become fixed (Zhao et al. 2014) but also expand to form novel gene families.

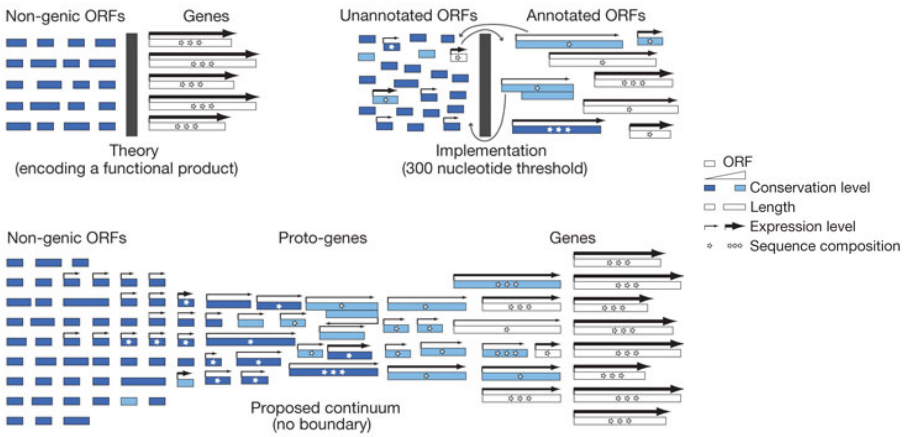


Figure 3.4. The binary model of annotation (top) and the proposed continuum (bottom)**.** Modified from Carvunis et al. 2012.

For a *de novo* gene to arise we require both transcription of a new genomic segment and an ORF that can be translated into a protein. It is not clear which is the most common order of events transcription of a previously noncoding segment of DNA followed by the elongation of an ORF (Tautz and Domazet-Lošo 2011) or the appearance of a long ORF (Figure 3.5) and the subsequent appearance of a promoter (Kaessmann 2010). The two models are further discussed by Schlötterer (Schlötterer 2015). Divergent transcription, where a promoter becomes bidirectional through mutations might also contribute to the birth of new genes (Gotea et al. 2013; Wu and Sharp 2013).

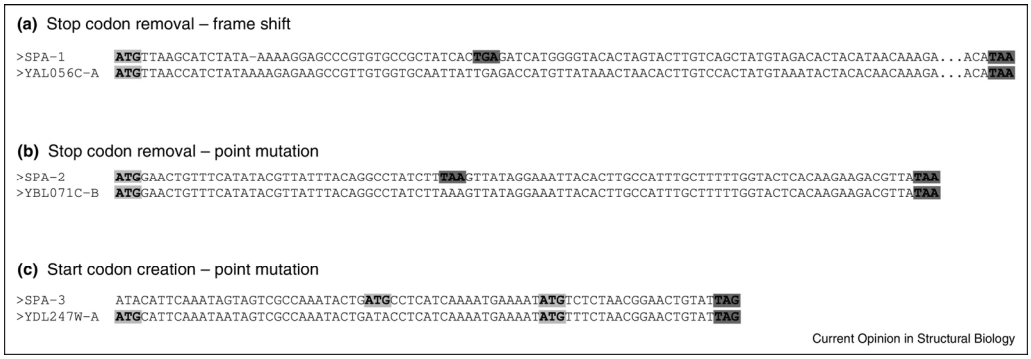


Figure 3.5. Three possible mechanisms for the proto-gene to gene transition with real examples in yeast. (a) A deletion of one or two nucleotides determines a frame shift, causing the disappearance of a stop codon. (b) A stop codon mutates into a coding triplet. (c) A point mutation creates a new start codon. YAL056C-A, YBL071C-B and YDL247W-A are *Saccharomyces cerevisiae* genes; SPA-1 to 3 are *Saccharomyces paradoxus* sequences with the following coordinates: SPA-1: AABY01000017.1:88566-88915, SPA-2: AABY01000058.1:43657-43755, SPA-3: AABY01000011.1:3135-3209. Modified from Light et al. (Light et al. 2014).

De novo genes in eukaryotes

The formation of *de novo* genes is no longer considered impossible or very rare. Despite several technical challenges (Light et al. 2014), many genes likely to have originated *de novo* have been described in recent years (Long et al. 2013; Arendsee et al. 2014; Schlötterer 2015). There are also several *bona fide* examples of protein coding genes, experimentally validated and with known function, which arose from previously non-coding sequences (*de novo*) in different species (Table 3.1).

One of the most interesting recent examples is the gene Qua-Quine Starch (QQS), found in Arabidopsis (Li et al. 2009). QQS encodes a 59 amino acid protein and its expression was confirmed by Western blot analysis. Transgenic RNA interference experiments (QQS expression was reduced) show excess leaf starch content at the end

of the illumination phase of a diurnal cycle. The modification of soybean lines to express QQS results in a marked decrease in leaf starch content and increase in leaf protein content (Li and Wurtele 2014). It is remarkable that a foreign gene can function in soybean, separated from Arabidopsis by about 100 My of evolution. In humans, the protein encoded by the NCYM gene, previously believed to be a long non-coding RNA (lncRNA), inhibits the activity of the GSK3b kinase and is associated with poor prognosis in neuroblastoma (Suenaga et al. 2014).

Orphan and *de novo* genes have been studied in a wide variety of animal groups and species. Estimates of *de novo* gene contribution in genomes vary depending on the stringency of the criteria used and the species being surveyed; in general, between 2-12% of all genes in a genome putatively originated *de novo* (Tautz and Domazet-Lošo 2011; Tautz 2014).

Toll-Riera was one of the first to perform a systematic search and characterization of orphan genes in primates, also looking for the possible mechanisms of origination (Toll-Riera et al. 2009). These genes were also identified in *Hydra* (Khalturin et al. 2008), fruit-fly (Domazet-Loso and Tautz 2003; Domazet-Loso et al. 2007), insects (Wissler et al. 2013a), *Mus musculus* (Neme and Tautz 2013), silkworm (Sun et al. 2015), and yeast (Carvunis et al. 2012).

A recent study in *Drosophila* found both segregating and fixed *de novo* genes (Zhao et al. 2014). The results indicated that at least a fraction of the genes had spread in the population by the effect of selection.

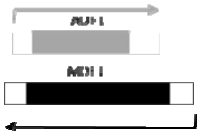
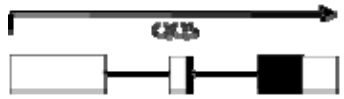
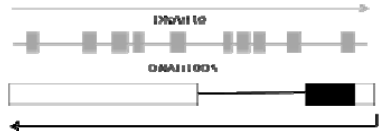
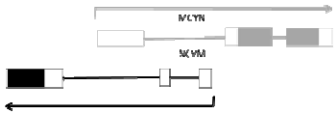
Species	Gene name Features	Sequence	Gene structure	Reference
Saccharomyces cerevisiae	MDF1(YCL058C) promotes cellular growth 152 amino acids Age <10 My	MQYHSALYVYIYVTFTTIPYKEKPDIISICFSMLSFVFDFSVRICSRTLESFSWS LISSSAFKVVSAFSLAGSCVLASRSSVGIIVSLLLFNFSTCNFVLFLSAVLIDLFF CTFLPTPTFLPTPFFFMLHLPIFSLLNALELLYLIIAGLHI		D. Li et al. 2010
Arabidopsis thaliana	QQS modulates carbon and nitrogen allocation 59 amino acids Age < 15 My	MKTNREQEIYVERSFKPNNSTIQNLMDIERFILPHTSTSGVARLKMRVISWVG LQFYNY		Li et al. 2009
Homo sapiens	PBO1 overexpressed in prostate and breast cancer 135 amino acids Age < 90 My	MRAFLRNQKYEDMHNIIHILQIRKLRHRLSNFPRLPGILAPETVLLPFCYKVFR KKEKVKRSQKATEFIDYSIEQSHHAILTPLQTHLTMKGSSMKCSSLSSEAILFT LTLQLQTLGLECCLLYLSKTIHPQII		Samusik et al. 2013
	DNAH10OS dynein, axonemal, heavy chain 10 opposite strand 163 amino acids Age < 6 My	MHSLPRSGSIRRTHSDTQATGWPPPQRIGDSPGPSPAFLSCPPSLCGGAAQTG DPVALPHGPEKWVWGGGLSPRNPHSWGIKAHGLRPPWAPRLERCMVPESE WAPWQPQLPCEPKWLGSRKSKPHRESGLRGGGPSRCAKRGTHSCGPRESGG PDTCHLPCH		Knowles and McLysaght 2009
	NCYM overexpressed in neuroblastoma 109 amino acids Age < 65 My	MQHPPCEPGNCLSLKEKKITEGSGGVCWGGETDASNPAPALTACCAAEREA NVEQGLAGRLLLCNYERRVVRRCKIAGRGRAPLGTRPLDVSSFKLKEEGRPP CLKINK		Suenaga et al. 2014
	C20orf203 overexpressed in Alzheimer 194 amino acids Age < 65 My	MFPRPVLNSRAQAILLPQPPNMLDHRQWPPRLASFPFTKTGMLSRATSVLAG LTAHLWDLGGGAGRRTSKAQRVHPQPSHQRQPPPPQHPGPYQERIWVGGEG WGEVGGLRLSKVGRRDREVRRGLRAPAGRGRAMGGMPRMGTVGDFGQAL SSLAWTSTCFQDFCLPSLPGKLPAPLISKQQFLSNSSRSLFN		C.-Y. Li et al. 2010
	TDRG1 testis-specific maximum expression during post-puberty 100 amino acids Age < 65 My	MKRREAVCAHRHFLGTGKPPHPLGRSIPVEPCPGLPAFAEVDLLSLLVPIKISS TPPSGSRLDPQIASSAFPGLGSLGGQDSSGSLVQRASCELESPYEL		Jiang et al. 2011

Table 3.1. Summary table of recently found *de novo* genes with protein evidence in a wide variety of species.

Characteristics of *de novo* genes

De novo genes are typically short and encode small proteins, although they are longer than expected by chance (Zhao et al. 2014). They are also expressed at low levels and are more tissue-specific than older genes (Palmieri et al. 2014). They are enriched in testis, which led to the suggestion that many genes were "born" in this tissue and later gained roles in other tissues – the "out of testis hypothesis" (Kaessmann 2010).

De novo genes show a positive codon usage bias when compared to non-coding sequences, indicating that many are likely to produce proteins (Toll-Riera et al. 2009). They also tend to evolve more rapidly than other coding genes (Domazet-Loso and Tautz 2003; Toll-Riera et al. 2009).

Although the functions of most *de novo* genes remain unknown, it has been shown that they can become functionally important in a relatively short span of time. Four out of five tested *de novo* genes in *D. melanogaster,* targeted with RNAi[15], resulted in a lethal phenotype (Reinhardt et al. 2013). Tinkering through non-specific protein-protein interactions might be the basis to develop a function that can be later favored by natural selection (Wu and Zhang 2013). It seems likely that regulatory functions are favored over catalytic ones (Arendsee et al. 2014).

---

[15] Also termed Short interfering RNA (siRNA): functional components of the RNAi-induced silencing complex. SiRNAs (21–23 nucleotides long) typically target and silence mRNAs by binding perfectly complementary sequences in the mRNA and causing their degradation and/or translation inhibition.

4. Results

This chapter is divided in four sections. In the first published article we develop and asses a new method to construct MSA and evaluate the impact in genome-wide positive selection studies. The second part contains relevant sections of a large sequencing consortium submitted article, where we explore orphan genes in the *Lynx pardinus* and look for signatures of positive selection in its recently sequenced genome. The third published article explores the role of gene regulation and sequence change in a mammalian adaptation: hibernation. Finally in the last submitted article we perform the largest identification of unique *de novo* gene across the entire mammalian phylogeny to date.

4.1. Improving genome-wide scans of positive selection by using protein isoforms of similar length

Villanueva-Cañas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol. Evol. 5:457–467.

Villanueva-Cañas JL, Laurie S, Albà MM. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol Evol. 2013;5(2):457-67.
doi: 10.1093/gbe/evt017

4.2.Signatures of adaptive evolution in the iberian lynx

Introduction

There are four extant lynx species: Eurasian lynx (*Lynx lynx*), Canada lynx (*L. canadiensis*), bobcat (*L. rufus*), and the Iberian lynx (*L. pardinus*), which is the only one classified as endangered by the IUCN. In 2012 we joined the Consortium to sequence and analyse the Iberian lynx genome. The main objectives were to gain insight into lynx demography and evolution, to better calibrate the population decline at a genome level, and to provide genomic resources to help in their conservation.

With these objectives in mind one Iberian male (named Candiles) was sequenced to high coverage along with the transcriptome of 11 lynx tissues. Additionally 10 Iberian and one Eurasian lynx genomes were re-sequenced.

We contributed to several stages of the sequencing and analysis and include the most relevant sections for this thesis. This includes the identification and characterization of lynx-specific genes and the search for signatures of positive selection. The full paper current status is submitted.

Signatures of positive selection

We looked for signatures of positive selection in the lynx lineage using a set of one-to-one orthologs generated in the phylogenomics analyses. We selected 8 different species: *Panthera tigris*, *Felis catus*, *Lynx lynx*, *Lynx pardinus*, *Ailuropoda melanoleuca*, *Canis*

lupus familiaris, *Homo sapiens* and *Mus musculus*. The set comprised 9,695 genes. We performed multiple sequence alignments with the software PRANK (Löytynoja and Goldman 2008) which has been shown to be particularly accurate at handling insertions and deletions, resulting in a lower number of false positives in positive selection tests (Jordan and Goldman 2011; Villanueva-Cañas et al. 2013). We conducted a branch-site test of positive selection (PS) (Yang 2007) using information from Timetree (www.timetree.org) for the input tree. This test is based on the detection of codons with an excess of non-synonymous substitutions in particular branches. It has a reasonable statistical power and low false-positive rates but it is also extremely sensitive to alignment errors (Fletcher and Yang 2010). We filtered out cases with more than one site with a probability of being under positive selection higher than 0.99 by the Bayes empirical Bayes (BEB) approach, as they typically corresponded to non-homologous stretches (Villanueva-Cañas et al. 2013). Internal branches were barely affected by this filtering, since they are more resilient to this kind of errors. We manually validated 100 lynx positive selection candidates (96 for Lynx sp. and 4 for Lynx lynx)(Extended Data Table 2).

We used Gitools (Perez-Llamas and Lopez-Bigas 2011) and annotations from Ensembl version 73 (Flicek et al. 2013) to perform an enrichment analysis in the set of positively selected genes, obtaining no significant results for either Gene Ontology terms or Kegg pathways (p-value > 0.05, False Discovery Rate correction BH (Benjamini and Hochberg 1995)). We also collected lists of

genes related to immune system or audition from NCBI genes but could see no general enrichment either.

However, inspection of the list of 100 candidates revealed that 21 of the validated genes were related to known human phenotypes, mostly diseases or syndromes recorded in OMIM. We also found structures in Protein Data Bank (PDB) for another 22 of those 100 cases. For example, the gene LYPA23A017274P1 (DARS) is an extremely conserved protein where we have a histidine (H) in the lynx lineage while all the felids present an arginine (R) and the rest of mammals a glutamine (Q).

Sensory perception is thought to be particularly important for cats (Heffner and Heffner 2003) and indeed we found two genes related to hearing in the list of positive selection candidates for the lynx branches: CACNA1D (LYPA23A015140P1) or MYO1F (LYPA23A022113P1). Mutations in these genes have been associated with deafness or hearing loss in humans (Zadro et al. 2009; Satheesh et al. 2012). In addition, two vision-related genes were also under positive selection in lynxes: OPTC (LYPA23A008195P1) and GUCY2F (LYPA23A015393P1) (Hunt et al. 2010).

Lynx orphan genes

An orphan gene is defined as a gene that lacks homologues in other lineages. Depending on which taxonomic level we are interested in, taxon-specific orphan genes (TSOGs) or species-specific orphan genes (SSOGs) can be defined (Wissler et al. 2013). Orphan genes were first discussed within the yeast genome project (Dujon 1996)

and are thought to play an important role in adaptive processes (Khalturin et al. 2008; Toll-Riera et al. 2009). In this section we describe the identification and characterization of lynx orphan genes.

We developed a pipeline to identify lynx orphan protein-coding genes. First, we discarded any proteins that had homologues in any of 23 non-mammalian eukaryotic species indicated in Figure 1, using gene protein coding annotations from Ensembl. To search for homologs we used BlastP (2.2.23+, (Altschul et al. 1990)) with an E-value threshold of 10^{-4} and the filter for low complexity regions activated. Second, we discarded any proteins for which we could indirectly trace homology to other species through a second protein in lynx. This could happen for example if the protein had evolved very rapidly after a gene duplication event (Toll-Riera et al. 2009). For these searches we used BlastP with the same parameters as previously except that we used a BLOSUM80 matrix instead of the default BLOSUM62, as we were searching for sequences that had diverged relatively recently. Third, we classified the remaining proteins as lynx-specific or mammalian-specific depending on the presence of homologues in the annotated genes from *Felis catus, Canis lupus familiaris, Ailuropoda melanoleuca, Mustela putorius furo, Homo sapiens, Mus musculus, Bos taurus, Equus ferus caballus* and *Myotis lucifugus* (Ensembl version 72). Fourth, we only selected those genes expressed in at least one tissue using a RPKM threshold of 0.3. This resulted in the identification of 323 lynx-specific genes.

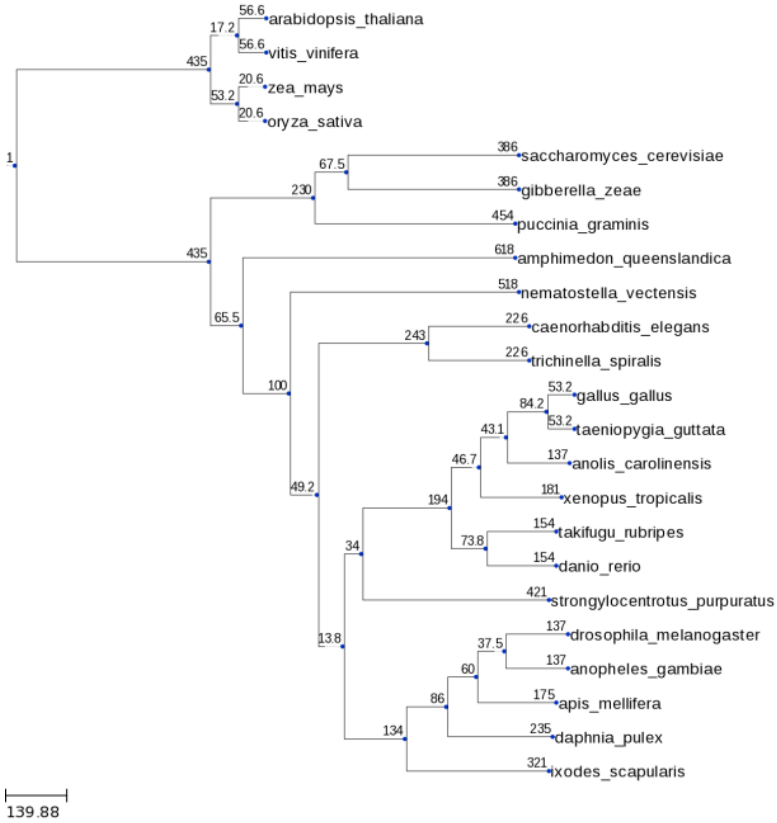


Figure 1. Tree showing the set of 23 non-mammalian species used to identify Iberian lynx orphan genes. Distances are in million years and were obtained from TimeTree.org.

The current gene catalogs are likely to be incomplete and this means that some of these 323 putatively lynx-specific genes may correspond to genes not yet annotated in other mammals. We thus employed published RNAseq data for different tissues and mammalian species (Brawand et al. 2011) to have a more comprehensive set of transcripts to compare our genes with. We run Tophat2 v2.0.8 (Trapnell et al. 2010) for pooled-tissues reads from human, mouse, chimpanzee, macaque and orangutan. Next, all long

expressed transcripts (length > 200 nucleotides) were assembled using Cufflinks (v 2.0.2) (Trapnell et al. 2010) for each species and tissue separately, not using information from gene annotations (no reference GTF file). We used Cuffmerge to obtain a comprehensive set of transcripts for each species and Cuffcompare to classify the transcripts into already known transcripts (annotated, using GTF files corresponding to Ensembl v. 60) and novel transcripts (non-annotated). The number of transcripts per species is displayed in Table 1. Finally, we run tBlastX with an E-value threshold of 10^{-6} to search for homologues of the 323 putative lynx orphan genes among these transcripts. After discarding any gene that had at least one match, the list of lynx orphan genes list was reduced to 204 (206 transcripts).

	Human	**Chimpanzee**	**Macaque**	**Orangutan**	**Mouse**
Annotated transcripts	77,597	44,067	23,414	17,156	15,876
Novel transcripts	67,573	98,979	138,240	111,858	99,315

Table 1. Number of assembled transcripts (annotated and novel ones) per species using RNAseq data.

Using the RNAseq Iberian lynx transcriptomics data we investigated if there were any biases in the tissues in which the lynx orphan genes were expressed. Brain stood out as the tissue most highly enriched in expressed orphan genes (Table 2).

		RPKM				RPKM		
Tissue	**Orphan Genes**	**Mean**	**Median**	**SD**	**All genes**	**Mean**	**Median**	**SD**
Brain	91	23.73	0.63	109.57	14,310	19.26	5.87	73.69
Heart	57	7.75	0.75	25.80	11,748	25.13	4.45	231.23
Kidney	71	11.74	0.74	41.53	13,720	20.34	6.36	128.88
Liver	37	28.73	0.90	87.57	11,023	36	4.25	935.84
Lung	54	15.96	0.95	57.21	13,453	23.81	7.05	146.28
Muscle	39	148.60	1.03	772.29	11,202	32.77	4.88	350.26
Pancreas	34	41.70	1.68	114.09	12,042	57.59	2.89	1370.74
Spleen	55	13.07	1.08	47.67	13,551	23.55	6.78	161.93
Stomach	50	38.92	1.45	170.69	12,679	34.25	6.06	694.37
Testes	74	14.92	0.96	55.04	13,975	19.89	7.75	81.20

Table 2. Summary of expression data for 206 lynx orphan transcripts. Mean, median and SD refer to RPKM values for Expressed genes (RPKM>0.3).

In an attempt to further characterize the expression of these genes in related species and given the importance of testes in the birth of new genes (Kaessmann 2010) we reconstructed the transcriptome of cat testicle. We used Tophat to map the reads to the genome (v2.0.8, cat genome version 75 ENSEMBL), Cufflinks to reconstruct the transcripts and cuffcompare to determine which genes were novel and which were already annotated in the cat (Ensembl 75, v2.1.1). We generated a fasta file from the gtf file and performed sequence similarity searches against the 204 *Lynx pardinus* orphan genes, obtaining no significant hits. This means that, at least with the current data, the 74 orphan genes expressed in lynx testes would be exclusive of the lynx lineage.

	Number of proteins	Length of proteins (aa)			Low complexity regions (LCRs)	
		Mean	*Median*	*SD*	*% of proteins with LCRs*	*Mean % covered*
ORPHAN GENES	206	113.4	94	64.83	51.5	39.32
REST OF PROTEINS	31,818	537.2	383	581.99	73.11	9.22

Table 3. Summary of properties for the set of 206 lynx-specific expressed proteins and the rest of expressed proteins (once we removed sequences with one or more 'X').

Species or lineage-specific genes tend to be short and enriched in repetitive sequences (Toll-Riera et al. 2009; Tautz and Domazet-Lošo 2011). We confirmed that lynx orphan genes are shorter than average, with a mean length of 112.3 amino acids compared to 537.2 for the rest of genes (Table 3 and Figure 2). About half of the lynx orphan genes contain low-complexity sequences as measured with the program SEG (Wootton 1994), which is less than for the rest of genes (73.11 %), but these genes contain on average nearly 40% of their sequence covered by repeats (LCRs), which is a very high portion when compared to the rest of genes (average 9.22 %) (Table 3).

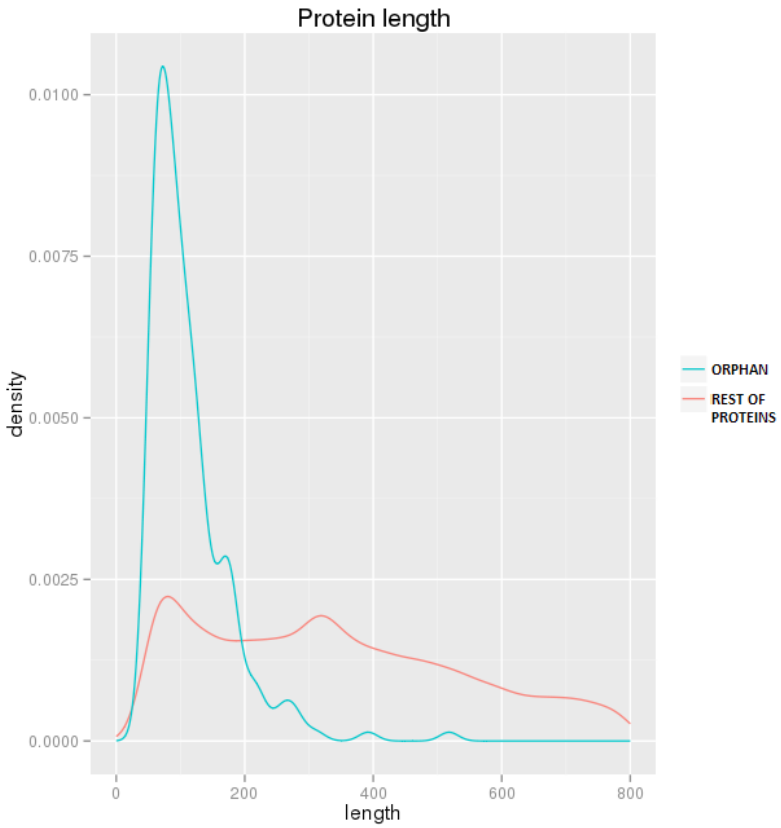


Figure 2. Distribution of lengths for the set of lynx-specific proteins compared to the rest of expressed genes.

4.3. Comparative genomics of mammalian hibernators using gene networks.

Villanueva-Cañas JL, Faherty SL, Yoder AD, Albà MM. 2014. Comparative genomics of mammalian hibernators using gene networks. Integr. Comp. Biol. 54:452–462.

Villanueva-Cañas JL, Faherty SL, Yoder AD, Albà MM. Comparative genomics of mammalian hibernators using gene networks. Integr Comp Biol. 2014 Sep;54(3):452-62. doi: 10.1093/icb/icu048.

4.4. Birth of new genes and evolutionary innovation in mammals

Villanueva-Cañas JL, Ruiz-Orera J, Albà MM. In preparation. Birth of new genes and evolutionary innovation in mammals.

ABSTRACT

The birth of new genes *de novo* from previously non-genic genomic regions is increasingly being recognized as an important mechanism of evolutionary innovation. However, these genes, which do not have homologues outside the species or taxon, remain poorly characterized. Here we used 68 complete genome sequences from different mammalian species to obtain the first global census of gene families likely to have originated in the past 200 Million years of mammalian evolution. Using a combination of gene annotations and high throughput RNA sequencing (RNA-Seq) data we identify 4,301 different mammalian-specific gene families present in more than one species, 1,863 of which probably appeared more than 65 Million years ago. The proteins encoded by these genes tend to be shorter and more positively charged than proteins that are conserved in a wide range of eukaryotic organisms. Using human tissue expression data from GTEX we observe than nearly 40% of them show maximum expression in testis. The rest of genes are significantly overrepresented in the salivary and pituitary glands. Functional enrichment analysis reveals an overrepresentation of proteins involved in the defense to pathogens, including several antimicrobial peptides. We also identify a number of proteins that are important for the development of the specialized mammalian skin, such as corneodesmosin, which shows significant molecular signatures of positive selection in the Eutheria and Euarchontoglira branches. This study highlights important composition and functional biases in *de novo* genes and provides a basis for future research on adaptive processes driven by the birth of new genes.

Introduction

Mammals are a widely diversified group that has managed to conquer almost every ecosystem available on earth. They have undergone numerous physiological and biochemical adaptations, including hair, a skin with integumentary glands, the placenta or a highly sophisticated immune system. Due to their economic and biomedical importance they have been a primary target for genome sequencing efforts (Genome 10K Community of Scientists 2009; Lindblad-Toh et al. 2011). This wealth of genomics data is very valuable to understand the molecular mechanisms behind the extraordinary range of adaptations observed in the group.

The birth of new genes can trigger important phenotypic changes. New genes can arise by the duplication of already existing genes, evolving new or modified functions or resulting in a change in the concentration of the gene (Ohno 1970; Force et al. 1999). A well-known example is RNASE1B in leaf-eating monkeys, which duplicated from RNASE1 and subsequently acquired enhanced ribonucleolytic activity for the digestion of bacterial RNA (Zhang et al. 2002). Another case involves a retrocopy of the fgf4 growth factor in dogs, which is responsible for short-legged phenotype in many breeds due to increased gene dosage (Parker et al. 2009).

New genes can also arise *de novo* from non-genic regions of the genome (Milde et al. 2009; Tautz and Domazet-Lošo 2011; Schlötterer 2015). A large fraction of the genome is transcribed, including many regions that do not contain annotated genes, providing abundant material for *de novo* gene birth (Carvunis et al.

2012; Ruiz-Orera et al. 2014). Some of these genes may acquire a novel function and be preserved by natural selection. Over time these genes are going to be present in a reduced set of species, the node connecting them in the evolutionary tree indicating the approximate time of formation of the gene (Domazet-Loso et al. 2007). Lineage-specific, or taxon-restricted, genes, which show no homology to genes outside the taxon, are believed to have originated primarily by this mechanism (Tautz and Domazet-Lošo 2011; Wissler et al. 2013; Arendsee et al. 2014). These genes are likely to play an important role in adaptive processes (Khalturin et al. 2008; Li and Wurtele 2014), yet they remain poorly characterized. They're usually short, tissue-specific and expressed at low levels (Toll-Riera et al. 2009; Palmieri et al. 2014).

In mammals, *de novo* genes have primarily been described in humans (Knowles and McLysaght 2009; Toll-Riera et al. 2009) and mouse (Murphy and McLysaght 2012; Neme and Tautz 2013). Very few have been functionally characterized. One example is the NCYM gene, present in human and chimpanzee. This gene inhibits the activity of the GSK3β kinase, resulting in the stabilization of the antisense gene product MYCN in neuroblastomas (Suenaga et al. 2014). In mouse, the gene Poldi, which arose in the past 2.5-3.5 millions, has been shown to influence sperm motility (Heinen et al. 2009).

Here we use the protein-coding genes annotated in 68 mammalian species with complete genomes, together with *de novo* transcript reconstructions using RNAseq data from a subset of 28 species, to generate a comprehensive set of mammalian-specific genes grouped

into families. We identify thousands of mammalian-specific gene families, many of which are likely to have contributed to specific adaptations at different time-points in the mammalian phylogeny. We investigate gene expression patterns in different tissues, sequence and functional biases as well as the presence of signatures of positive selection. The study provides novel insights into the evolution of *de novo* genes and provides a rich resource for future investigations on the role of these genes in adaptive processes.

Results

Construction of mammalian-specific gene families

With the aim of identifying genes originated in different mammalian lineages we selected 68 mammalian species that had their genome sequenced from a wide variety of orders (Figure 1). The nodes in the tree cover crucial evolutionary events such as the capacity to give birth to young without using a shelled egg (Theria), the appearance of the placenta (Eutheria) or synapomorphic characters of the different orders.

For each of the 68 species we obtained gene and protein annotations from Ensembl or NCBI. In a subset of 28 species we perform *de novo* transcript assembly from available RNA-Seq data to account for genes missing from the annotations. Next we identified all the annotated proteins in each species that had no annotated homologues in any of 34 non-mammalian species, including vertebrates, insects, fungi or plants (supplementary file 1, Table S2), using BlastP (E-value $<10^{-4}$).

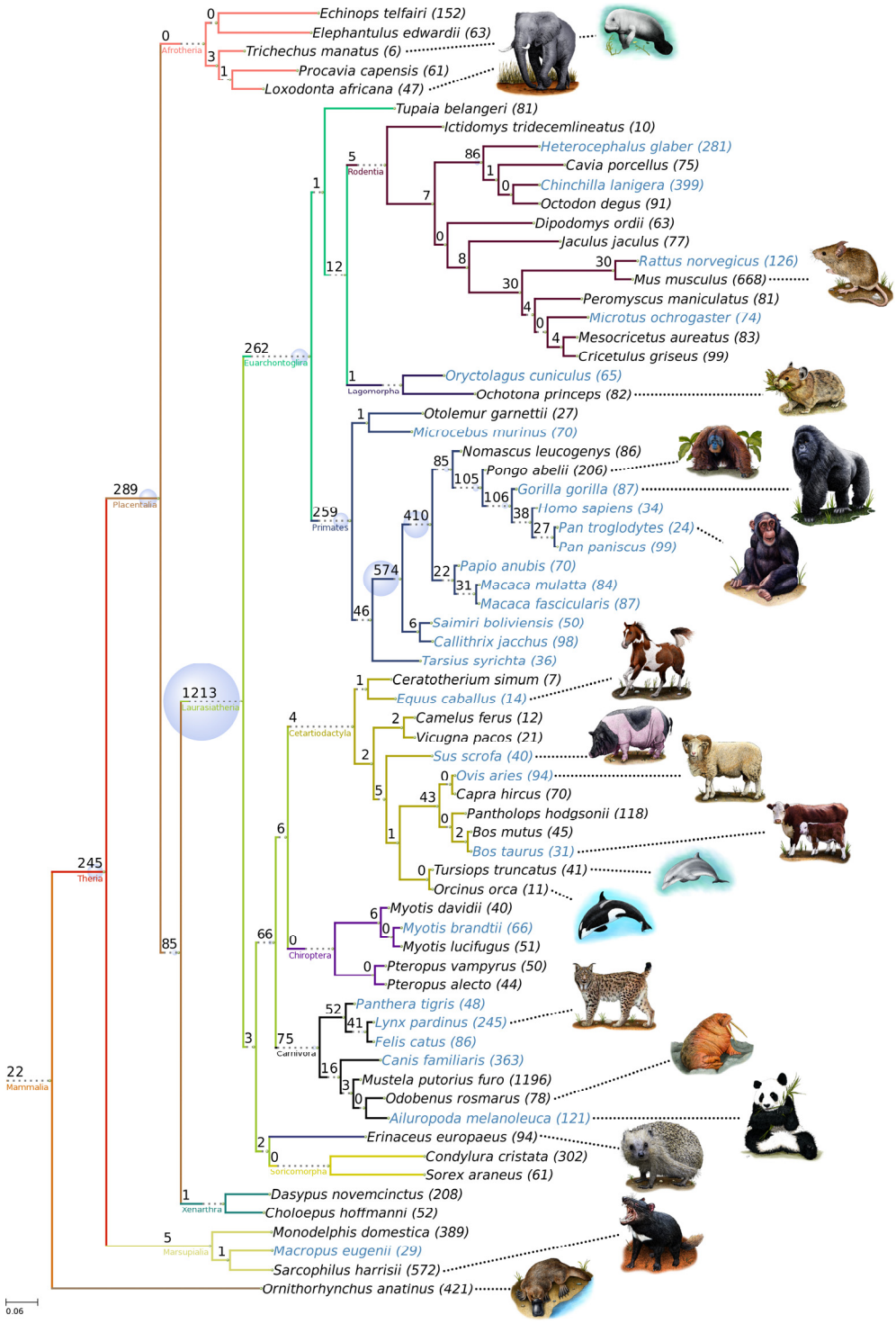


Figure 1. Tree showing the sequenced species used. Colored names indicate that RNAseq reconstructions for that species were used to improve the dating. Blue spheres and numbers represent the number of gene birth events at each node.

Branch lengths are based on Meredith (Meredith et al. 2011) with additions from other sources and represent phylogenetic distance. Branches are painted according to orders and other higher groups.

For comparison, we also collected the proteins for each species that had a significant BlastP hit in all of the 34 non-mammalian species considered and tagged them 'ancestral' proteins.

Next we constructed the set of mammalian-specific gene families, using the genes that we had already identified as restricted to mammals in the 68 species. We performed sequence similarity searches of all the proteins in one species against all the proteins in the rest of species and recursively used the lists of homologues to recover all the members of a family. The time of origin of the gene was the node connecting the two more distant species in the gene family (supplementary file 2, Figure S1). Subsequently, we refined the dating of the genes by using sequence similarity searches against the *de novo* transcript assemblies generated from RNA-Seq data, successfully reclassifying 2,220 gene birth events that were previously species-specific. Our final dataset consisted of 13,063 gene families or unique gene birth events in the past 200 Mya of mammalian history, of which 4,301 had members in several species (Figure 1).

A large number of gene families, 1,213, mapped to the Laurasiatheria branch. However, we have to consider that improved annotation of the Afrotheria and Xenartha groups may reduce this number. Gene families present in these three main groups were classified as Placentalia (289) or Theria (245), the latter also found in the marsupials *Monodelphis domestica* and/or *Macropus eugenii*.

These genes, originated in ancestral mammalian branches, may have been important for the development of shared mammalian phenotypic traits. We also identified a large number of primate-specific gene families, especially in branches predating the separation of macaques and great apes. A list of gene families with the estimated node of origin is available from supplementary file 3.
The number of species-specific genes varied greatly depending on the organism (Figure 1). Different factors, such as the distance to the closest related organism with a sequenced genome or the gene annotation protocol, may explain most of these differences. Very strict conditions when performing gene annotation may lead to the elimination of the genes that do not have homologues in other species, recovering virtually no species-specific genes, whereas too relaxed conditions, such as annotations in the absence of expression evidence, may result in the inclusion of false positives. This implies that species-specific genes are more heterogeneous as a class than gene families with members from multiple species.

Mammalian-specific genes are enriched in testis and salivary gland

Taking advantage of the extensive tissue expression data for human (GTEX, (Lonsdale et al. 2013)) and mouse (Mouse ENCODE, (Pervouchine et al. 2015)) we investigated the levels of expression of mammalian-specific and ancestral genes in different adult tissues. The majority of the mammalian-specific genes, 672 out of 834 in human and 806 out of 986 in mouse, were expressed in at least one tissue. We calculated their average expression levels in

each of the surveyed tissues, the tissue with the maximum gene expression level, and the preferentially tissue expression index (supplementary file S3 and S4).

Mammalian-specific genes collectively showed significantly higher tissue specificity than ancestral genes in both human or mouse (KS test, p-value < $2.2e^{-16}$). This finding is in line with previous knowledge about the expression patterns of young human genes (Toll-Riera et al. 2009; Wu et al. 2011). In general, a large number of genes showed maximum expression in testis: 27.4% of the ancestral genes and 45.8% of the mammalian-specific genes in humans. It is unlikely that such a large number of ancestral genes perform a function in testis or spermatogenesis, as these genes are present in a wide variety of species, including fungi and plants that are lacking these structures. The results may thus reflect deregulated gene expression in this tissue and/or high power to detect genes with very low expression level that do not play any function in the tissue (Jongeneel et al. 2005; Martínez and Reyes-Valdés 2008).

Considering genes that showed maximum expression in tissues other than testis revealed a significant enrichment of mammalian-specific genes in salivary gland and pituitary (p-value=$8.9x10^{-9}$ and p-value=0.005, respectively, Fisher test). The bulk of these genes were conserved in other primates or placental mammals. In contrast, muscle, heart or adrenal gland had significantly fewer mammalian-specific genes than expected by chance (supplementary file 1, Table S1). In conclusion, evolutionary innovation related to the birth of new genes is highly biased towards certain tissues.

Functional characterization of mammalian-specific genes

Whereas nearly 90% of genes in the ancestral class were functionally characterized, the majority of mammalian-specific genes were not. Using terms in the Description field of Ensembl gene entries for human genes we created word clouds and the term 'uncharacterized' clearly stood among mammalian-specific genes, contrasting with more diverse words from ancestral genes (Fig. 2A and 2B). We also noted that the level of functional annotation gradually decreased for younger classes of genes (Fig. 2C). This highlights our current lack of knowledge on the roles of many of these genes.

In order to gain further insight into the possible functions of mammalian-specific genes, we performed functional enrichment analysis with DAVID (Huang et al. 2009) for human and mouse genes. In both species we detected a significant enrichment in the Gene Ontology (GO) category "defense response to bacterium" (p-value $<10^{-15}$, supplementary file 1 tables S1 and S2). The terms "secreted" and "extracellular" were also significantly enriched in human and mouse, respectively (p-value $<10^{-15}$). Additionally, the mouse young genes showed a modest enrichment in keratins (p-value < 0.038).

Below we present a more detailed analysis of two functional classes of genes for which we found numerous examples and which may have been of special relevance in adaptive processes in mammals: antimicrobial peptides and skin proteins.

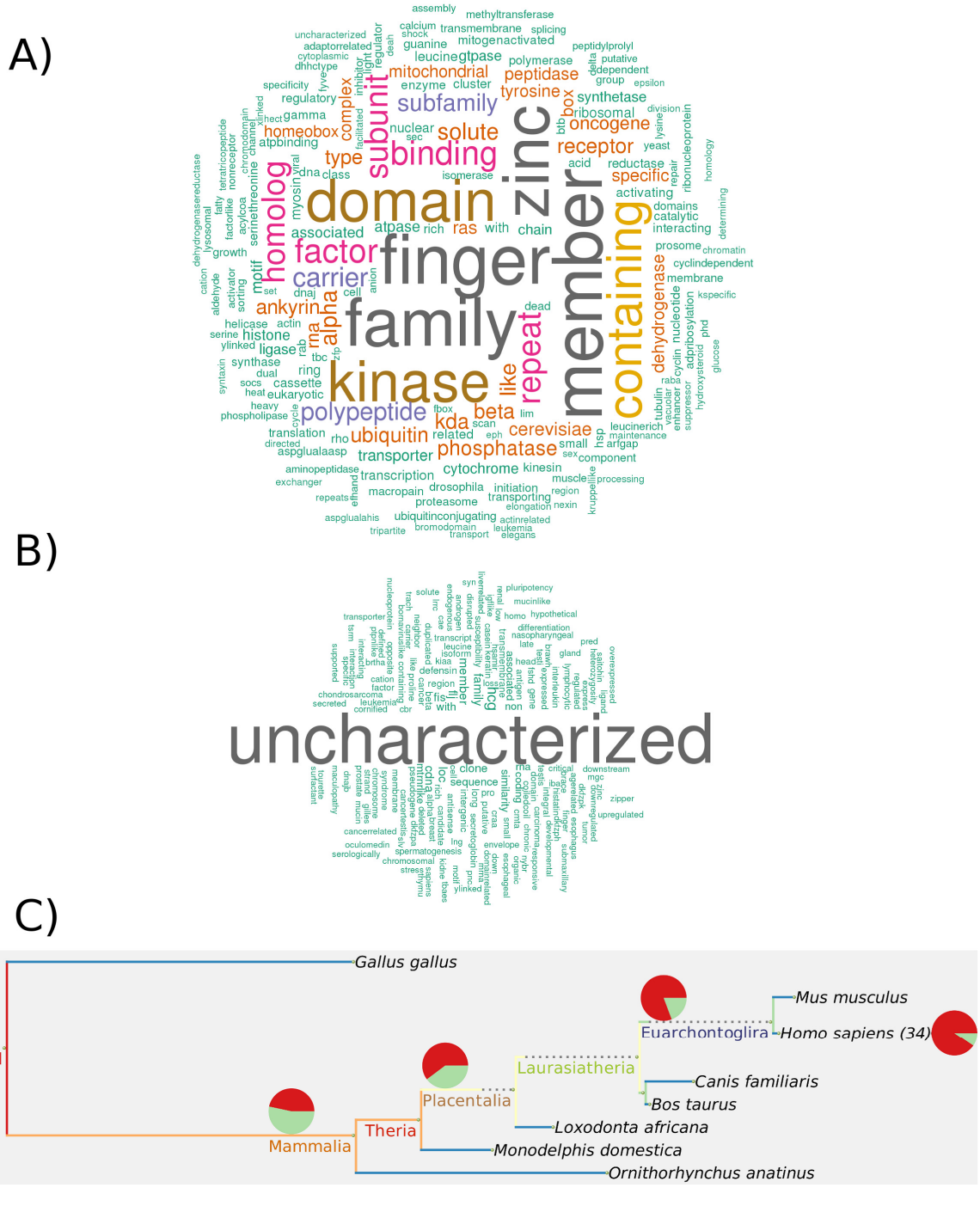


Figure 2. A) Wordcloud using the descriptions of human ancestral genes. **B)** Wordcloud using the descriptions of all young proteins identified as mammalian specific present in humans. **C)** Percentage of human genes with known function in the different conservation levels. Red: Unknown, Green: Known.

Antimicrobial peptides

Close examination of the groups “defense response to bacterium”, “secreted” and “extracellular” revealed the presence of numerous antimicrobial peptides (AMPs), including histatins (Tsai and Bobek 1998), dermcidin (Schittek et al. 2001) or lacritin (McKown et al. 2014). We also identified several alpha and beta defensins, such as DEFA6, that is only expressed in Paneth cells in the small intestine. These short proteins can acquire antimicrobial activity through a wide variety of mechanisms (Yeaman and Yount 2003). In *de novo* originated genes such activity may have been advantageous and led to their fixation in the population.

Skin proteins

The mammalian skin is adapted to have high sensitivity and allow muscles to produce a plastic deformation. The outermost layer of the skin known as *stratum corneum* (SC) is particularly dynamic in mammals as opposed to the stiff reptile skin, due to its heavy keratinization. This motivated us to investigate in more detail mammalian-specific genes with preferential tissue expression in the skin, most of which had appeared at the basis of all mammals. This includes proteins from the Late Cornified Envelope (LCE) such as LCE3B (deletion related to psoriasis (Bergboer et al. 2014)) and LCE6A. LCE are small, proline-rich precursors of the cornified envelope of the SC. Dermokine, KRTDAP are proteins that are also involved in keratinocyte differentiation.

Another mammalian-specific protein, Corneodesmosin (CDSN) is a glycoprotein expressed during the late stages of epidermal differentiation, participating in the specialized junctions called corneodesmosomes that bridge together corneocytes in the lower part of the SC. Those junctions are degraded by certain serine proteases, such as kallikreins, as corneocytes migrate toward the surface of the skin and result in desquamation. When deleted, CDSN has been shown to cause lethal epidermal barrier disruption and hair follicle degeneration in mice (Jonca et al. 2011). Corneodesmosin was exclusively expressed in the skin and was found to be under positive selection at the Eutheria and Euarchontoglira branches (see section *Signatures of natural selection*).

Sequence properties of mammalian-specific genes

Different studies have reported biases in the sequence properties of recently evolved genes, including a shorter size and lower codon usage bias (Toll-Riera et al. 2009; Carvunis et al. 2012; Neme and Tautz 2013). In order to study the sequence properties of the mammalian-specific genes we subdivided them into four conservation levels (0 to 3), roughly reflecting different periods of formation during mammalian evolution (supplementary file 2, Figure S2). The most conserved genes were those of class 3, corresponding to the Mammalia or Theria branches (due to the scarcity of Monotremata genomes a reliable distinction between the two classes was not possible). Genes in class 2 corresponded to the common Marsupialia branch or to the main Placentalia groups.

Class 1 comprised genes families for more recently evolved groups, such a Rodentia, Primates or Carnivora. Finally, class 0 only contained species-specific genes.

In agreement with previous reports we found that the proteins encoded by the mammalian-specific gene families were shorter than the proteins encoded by ancestral genes (Figure 3A). In addition the length of the protein tended to decrease as we considered younger evolutionary classes. The only exception was species-specific genes (class 0), which were slightly longer than those in class 1 (mean 187.57 vs 163.06, respectively). The later result was not observed when restricting the analysis to human or mouse genes (Table 1; supplementary file 2, Figure S3) and thus it is possibly due to spurious annotations in some organisms. Overall, the results were consistent with the hypothesis that proteins tend to become longer over time (Albà and Castresana 2005), acquiring new domains through a wide variety of mechanisms such as expansion of short repetitive elements (Toll-Riera et al. 2012) or exon shuffling (Long et al. 2003).

Mammalian-specific proteins from different conservation classes showed a consistent enrichment in proline, serine and arginine with respect to ancestral proteins (Figure 3B). This result was confirmed when we only used human or mouse proteins to perform the same analysis (supplementary file 2, Figure S4 and S5). We also found that proteins with a more recent evolutionary origin tended to show a higher isoelectric point (Table 1 and supplementary file 2, Figure S6; Kolmogorov-Smirnov (KS) test p-value < 0.05). The youngest proteins tended to be basic whereas the set of ancestral proteins

showed a bimodal distribution, with a large set of acidic proteins and other comparable set of basic proteins. These biases are consistent with the presence of AMPs among recently evolved proteins, as these proteins are particularly enriched in Arginine (R) and show a strong correlation between peptide cationicity and antimicrobial activity (Yeaman and Yount 2003).

It has been proposed that proteins with a recent evolutionary origin are enriched in low-complexity sequences and may have accumulated amino acid repeats at an increased rate (Toll-Riera et al. 2012). We did not observe a clear trend in our dataset (Table 1; supplementary file 2, Figure S7), which may be partly due to the fact that we initially discarded sequences with high percentage of such regions to avoid false positives in homology searches. However, some proteins showed patterns reminiscent of amino acid repeat expansions. One example was Corneodesmosin (CDSN), which showed a strong enrichment in glycine and serine residues (27.5% and 16%, respectively), particularly at the N-terminus of the protein. It has been hypothesized that these glycine loops could mediate intermolecular adhesions (Steinert et al. 1991). This simple interaction mechanisms and the relative simplicity of such repetitive regions could explain why some novel genes can quickly become essential if recruited into existing pathways (Reinhardt et al. 2013). We also observed a decrease in aromaticity in mammalian-specific proteins when compared to ancestral genes, although there was not a clear correlation with age (Table 1; supplementary file 2 Figure S8).

Species	Conservation level	Group	#Genes	Length	Aromaticity	IP	Simple	avg_exp	sp_t
Homo sapiens	Ancestral	A	5,362	681.28	0.079	7.31	1.14	15.79	0.65
	Young	3	60	196.30	0.064	7.49	1.25	2.79	0.90
		2	259	159.64	0.078	8.54	1.20	2.83	0.88
		1	332	115.14	0.070	8.74	1.17	0.35	0.87
		0	21	79.05	0.062	9.41	1.17	0.64	0.92
Mus musculus	Ancestral	A	5,337	667.40	0.079	7.26	1.14	23.37	0.77
	Young	3	74	196.54	0.071	6.87	1.21	26.6	0.92
		2	87	171.03	0.072	8.73	1.28	8.79	0.84
		1	77	165.62	0.054	8.73	1.25	7.23	0.83
		0	568	128.65	0.075	9.17	1.18	1.55	0.75

Table 1. Sequence properties. Mean values are shown. avg_exp is in RPKM. sp_t is the tissue-specificity score. See main text for the definition of Group.

Table 2. Positive selection. Number of branches (terminal or internal) under Positive Selection (PS) in Ancestral and mammalian-specific genes (Young). Differences between groups are significant (Fisher Test p-value < 0.05)

Terminal

	Ancestral	Young
PS	297	31
NPS	2575	176

Internal

	Ancestral	Young
PS	88	14
Total	2024	105

Table 3. Mammalian-specific genes under positive selection. Selected after manual inspection of the alignments. Positive selection in at least one branch in the mammalin tree. Ensembl Gene IDs are for the human protein.

Ensembl Gene ID	Description	Associated Gene Name
ENSG00000073803	mitogen-activated protein kinase kinase kinase 13 [Source:HGNC Symbol;Acc:6852]	MAP3K13
ENSG00000078304	protein phosphatase 2, regulatory subunit B', gamma [Source:HGNC Symbol;Acc:9311]	PPP2R5C
ENSG00000109205	odontogenic, ameloblast asssociated [Source:HGNC Symbol;Acc:26043]	ODAM
ENSG00000122188	lymphocyte transmembrane adaptor 1 [Source:HGNC Symbol;Acc:26005]	LAX1
ENSG00000125657	tumor necrosis factor (ligand) superfamily, member 9 [Source:HGNC Symbol;Acc:11939]	TNFSF9

ENSG00000125863	McKusick-Kaufman syndrome [Source:HGNC Symbol;Acc:7108]	MKKS
ENSG00000150783	testis expressed 12 [Source:HGNC Symbol;Acc:11734]	TEX12
ENSG00000153495	testis expressed 29 [Source:HGNC Symbol;Acc:20370]	TEX29
ENSG00000161180	coiled-coil domain containing 116 [Source:HGNC Symbol;Acc:26688]	CCDC116
ENSG00000167105	transmembrane protein 92 [Source:HGNC Symbol;Acc:26579]	TMEM92
ENSG00000172301	coordinator of PRMT5, differentiation stimulator [Source:HGNC Symbol;Acc:28848]	COPRS
ENSG00000173237	chromosome 11 open reading frame 86 [Source:HGNC Symbol;Acc:34442]	C11orf86
ENSG00000174437	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2 [Source:HGNC Symbol;Acc:812]	ATP2A2
ENSG00000180035	zinc finger protein 48 [Source:HGNC Symbol;Acc:13114]	ZNF48
ENSG00000184601	chromosome 14 open reading frame 180 [Source:HGNC Symbol;Acc:33795]	C14orf180
ENSG00000186118	testis expressed 38 [Source:HGNC Symbol;Acc:29589]	TEX38
ENSG00000197119	solute carrier family 25, member 29 [Source:HGNC Symbol;Acc:20116]	SLC25A29
ENSG00000204420	chromosome 6 open reading frame 25 [Source:HGNC Symbol;Acc:13937]	C6orf25
ENSG00000204539	corneodesmosin [Source:HGNC Symbol;Acc:1802]	CDSN
ENSG00000214700	chromosome 12 open reading frame 71 [Source:HGNC Symbol;Acc:34452]	C12orf71
ENSG00000239704	CMT1A duplicated region transcript 4 [Source:HGNC Symbol;Acc:14383]	CDRT4
ENSG00000240021	testis expressed 35 [Source:HGNC Symbol;Acc:25366]	TEX35

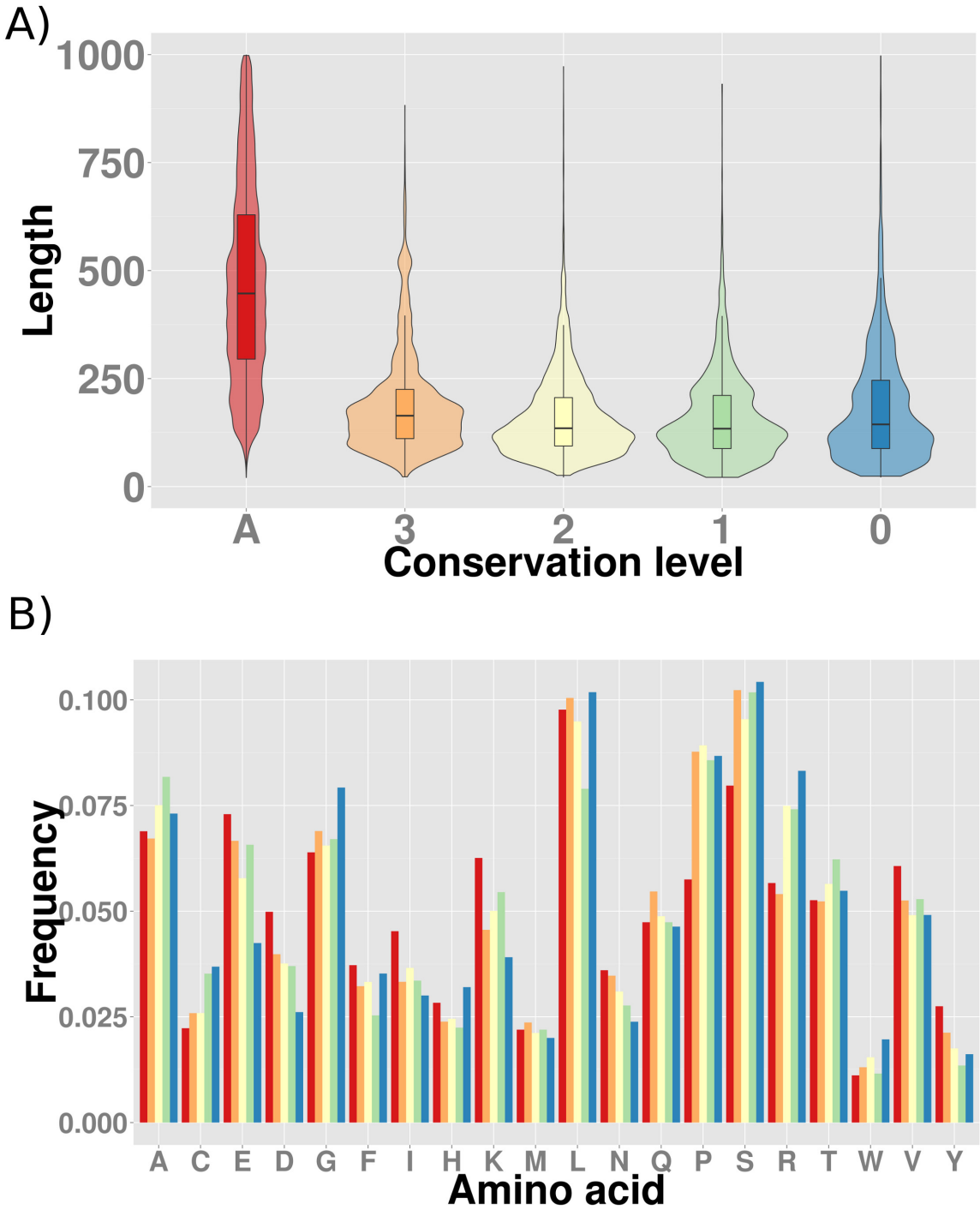


Figure 3. A) Violin plot showing length for all proteins classified into conservation classes. All distributions are significantly different among conservation classes (Kolmogorov-Smirnov, p-value<0.05). **B)** Amino acid composition (in frequency) for the concatenated sequences of all the proteins divided in conservation classes: A- red, 3- orange, 2- yellow, 1-green, 0- blue.

Signatures of natural selection

We estimated the nonsynonymous and synonymous substitution rates (dN/dS) for all young genes present in human with orthology in mouse as well as in a random subset of ancestral genes using CodeML. After applying a number of filters to remove potential alignment problems (see Methods) we obtained 457 ancestral genes and 66 young genes respectively. The mammalian-specific genes showed dN/dS values clearly lower than 1 (supplementary file 2 Figure S9), denoting that they are subject to purifying selection. However, dN/dS values were much higher in these genes than in the ancestral genes (KS test, p-value $<2.2e^{-16}$), in agreement with previous findings (Albà and Castresana 2005; Toll-Riera et al. 2008).

In order to assess if the higher dN/dS were simply due to a relaxation of purifying selection or had instead been influenced by positive selection (PS) we applied the maximum-likelihood based branch-site test (Zhang et al. 2005) to mammalian-specific and ancestral gene families, using the same set of species to build the multiple alignment (see Methods). This method allows the identification of codons likely to have experienced positive selection in a particular branch of the tree and has been specifically designed to differentiate positive selection from relaxation of purifying selection. The requirement to have additional species for a correct inference of positive selection events reduced the set of mammalian-specific genes to 33 and the set of ancestral families to 359. Nevertheless, we could observe a significant enrichment in positive selection signatures in the mammalian-specific gene

families with respect to the ancestral families (Fisher test p-value <0.05, Table 2). We detected it both at terminal and internal branches of the tree, reflecting possible adaptations both near the origin of the gene and at later periods. Table 3 shows a selection of mammalian-specific genes with signatures of positive selection. This included four proteins preferentially expressed in testis and corneodesmosin (CDSN). Another interesting gene under PS was LAX1 that encodes a transmembrane adaptor protein involved in lymphocyte signaling (Zhu et al. 2002; Hrdinka et al. 2012).

Discussion

The birth of new genes has been related to evolutionary innovations such as the development of new structures and cell types (Khalturin et al. 2009), changes in diet (Zhang et al. 2002) or increased resistance to cold (Basu et al. 2015). In mammals, syncytins, which are important components of the syncytiotrophoblast layer located at the maternofetal interface of the placenta, evolved from endogenous retrovirus sequences (Cornelis et al. 2012). *De novo* genes may also be important in the response to environmental challenges. In plants and yeast they show increased expression in response to oxidative or osmotic stresses (Donoghue et al. 2011; Carvunis et al. 2012).

Mammals are a highly diverse class of vertebrates, comprising more than 5,000 species. They underwent intense diversification in the so called Cretaceous Terrestrial Revolution (KTR), between 125 and 80 Million years ago (Meredith et al. 2011; O'Leary et al. 2013;

Yoder 2013; dos Reis et al. 2014). Here we explore the wealth of genomics data available for mammals to build the first catalog of mammalian-specific gene families. These families are composed of genes that do not show homologues outside mammals and are thus very likely to have originated in this group. The large number of genomes employed (68), and the use of transcriptomics data in addition to gene annotations, provided increased power to correctly date the time of origin of the genes when compared to previous studies (Albà and Castresana 2005; Neme and Tautz 2013). For example for a gene to be annotated as originated in the common primate branch it had to be absent from 54 other non-primate mammalian species, including 15 for which we generated *de novo* transcript assemblies to compensate for possible incomplete annotations. There were 259 such gene families, with members in several primate species. Many of these genes remain uncharacterized but it seems likely that they substantially contributed to adaptive processes in the group. In other taxons, such as the Chiroptera or Perissodactyla, we recovered very few lineage-specific gene families, probably owing to conservative criteria used in gene annotation.

The number of species and lineage-specific genes was postulated to decrease with the number of genomes sequenced (Casari et al. 1996). However, even within densely populated groups such as Primates, we observe a large number of genes that show a restricted phylogenetic distribution. High throughput RNA sequencing techniques have revealed that a larger than expected portion of the genome can be transcribed (Derrien et al. 2012). Many of the newly

discovered genes in human and mouse are annotated as long non-coding RNAs (lncRNAs) and are lineage-specific (Necsulea et al. 2014; Ruiz-Orera et al. 2015). Ribosome profiling data indicates that despite their annotation, which is based on the lack of conserved long open reading frames in the transcript, many of them are likely to translate into short peptides (Ingolia 2014; Ruiz-Orera et al. 2015). Thus, it seems possible that the number of species- and lineage-specific genes will tend to increase in the next years.

A large number of lineage-specific genes, but also of ancestral genes, are preferentially expressed in testis when different adult tissues are examined. Other authors have argued that testis-specific genes, particularly lncRNAs genes, are likely to play a role in spermatogenesis (Hezroni et al. 2015). New genes expressed in testis may have been subject to positive selection because they confer a selective advantage. An interesting example is the Poldi gene, which is specific of the Mus genus. A knockout of this gene results in reduced sperm motility and lower testis weight (Heinen et al. 2009), suggesting that it has evolved an adaptive function in testis. Another possible case is the primate-specific TDRG1 gene, which shows maximum expression in testis during post-puberty (Jiang et al. 2011). The expression in testis may also reflect pervasive transcription in this organ (Soumillon et al. 2013). Some of the genes we detect in testis may play a function in a narrow window of time in another organ, for example during development. The fact that a large number of ancestral genes, present in species such as plants or fungi, show maximum expression in testis when human and mouse tissues are analyzed supports that idea that

expression and function may be largely decoupled in this tissue. In any case, the increased power to detect expressed transcripts in testis makes it an excellent organ in which to detect new genes that have not been yet identified and study the mechanisms of *de novo* gene formation.

The analysis of annotated mammalian-specific genes has offered new insights into the functional biases of these genes. We have identified a significant overrepresentation of genes involved in the defense against pathogens, including many AMPs. In general, these proteins are short and positively charged (Yeaman and Yount 2003). These features are characteristic of mammalian-specific genes, perhaps indicating that many more AMPs await characterization. We also found a number of proteins involved in the formation of the *stratum corneum* of the mammalian skin. One such protein, Corneodesmosin, showed positive selection signatures.

Although data from the mammary gland was not available from the sources we employed, we detected milk casein proteins in other tissues. We also identified a possible precursor of some of these proteins, ODAM, to be mammalian-specific. This protein has been recently reported to be present in frog (Kawasaki et al. 2011) although we could not find it performing sequence similarity searches against the western clawed frog genome (*Xenopus tropicalis*), probably reflecting incompleteness of the genome sequence. The large number of mammalian-specific proteins with no known function, together with the incompleteness of the gene annotation for both mammalian and non-mammalian genomes,

highlights the need to prioritize annotation over the sequencing of new species.

Methods

Datasets

A total of 40 proteomes were downloaded from Ensembl v.75 (Flicek et al. 2014), 27 downloaded from NCBI (version available on March 2014) and one from a consortium in which we participate (*Lynx pardinus*). For the species downloaded from Ensembl and the *Lynx pardinus* we also obtained the cDNA files.

RNAseq data from 28 different species was downloaded from public repositories (supplementary file 1, Figure S1). For most of them we had samples for several tissues including testis, heart, muscle and brain.

Identification of mammalian genes

We run BlastP searches (e-value $<10^{-4}$) for each of the 68 proteomes, including all isoforms described in a gene, against a set of 34 non-mammalian proteomes (Supplementary file 1). All BlastP (Altschul et al. 1990) searches were run with version 2.2.28+ using a threshold of 10^{-4} and the filter for low complexity regions activated (seg = 'yes'). We discarded any gene family which could be linked to non-mammalian proteins directly or by the existence of paralogues that had hits in species other than mammals. The search

for paralogues within each species was done with BlastP using a BLOSUM80 matrix, since we are looking for similar sequences.

Novel genes are enriched in low complexity regions (LCR) (Toll-Riera et al. 2012), as such regions need to be filtered out in BlastP searches due to spurious hits, we only kept proteins with at least 20 contiguous amino acids without LCRs (Fig. S10). For the identification of low complexity regions (LCR) in protein sequences we used the software SEG (Wootton and Federhen 1993) with default parameters and calculated the length of windows without LCR with an in-house python script.

We also generated an "ancestral" dataset with conserved proteins for comparisons. For each species we selected proteins that had a BlastP match against all the 34 non-mammalian genomes used to identify mammalian-specific proteins.

Construction and dating of gene families

The first step was to run a series of BlastP searches for every set of mammalian-specific candidates in each species against the remaining 67 mammalian species. With that information we were able to assign every protein to a node. Next we run BlastP searches inside all the members of the 67 nodes, so we could create gene families and group together genes from different species through homology. We also substituted every protein in every branch for all the proteins that had a homology relation in any mammalian species. This way we could rule out some proteins connected to old genomes through other species with a more complete fragment of the protein (discard some groups) and we were also able to connect

branches that lacked annotation in a middle node, due to the loss of that gene family in a particular lineage or perhaps an incomplete annotation. The group then was assigned to the oldest node containing a species with a protein belonging to that group.

Transcript reconstructions from RNAseq

RNAseq sequencing reads from the 28 species underwent quality filtering using ConDeTri (v.2.2) (Smeds and Künstner 2011) with the following settings (-hq=30 -lq=10). Adaptors were trimmed from filtered reads if at least 5 nucleotides of the adaptor sequence matched the end of each read. In all paired-end experiments, reads below 50 nucleotides or with a single pair were not considered. We aligned the reads to the correspondent reference species genome with Tophat (v. 2.0.8) (Kim et al. 2013) with parameters –N 3, -a 5 and –m 1, and including the correspondent parameters for paired-end and stranded reads if necessary. Multiple mapping to several locations in the genome was allowed unless otherwise stated. We performed gene and transcript assembly with Cufflinks (v 2.2.0) (Trapnell et al. 2010) for each individual tissue. We only considered assembled transcripts that met the following requirements: a) the transcript was covered by at least 4 reads. b) Abundance was higher than 1% of the most abundant isoform of the gene. c) <20% of reads were mapped to multiple locations in the genome. d) The reconstructed transcripts were at least 300nt. Subsequently, we used Cuffmerge to build a simple set of assembled transcripts per each species. We use Cuffcompare to compare the coordinates of our set of assembled transcripts with gene annotation files from Ensembl

(gtf format, v.75) or NCBI (gff format, December 2014) to identify annotated transcripts and generate a novel set of reconstructed transcripts.

Sequence analyses of different conservation groups

For the amino acid analysis we concatenated all the sequences by class and calculated the relative amino acid frequency by counting the frequency of each amino acid and dividing by the total length of the sequences. The software Simple (Alba et al. 2002) was run with default parameters to calculate a relative measure of the simplicity of the sequence compositions. The remaining properties (IP, aromaticity, etc) were calculated using Biopython.

Positive selection

We grouped together genes born at conservation levels 3 (60) and 2 (259) groups (Young) and compared against the Ancestral group (5.359 genes). We used a subset of species including: *Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Canis lupus familiaris*, *Bos taurus*, *Loxodonta africana*, *Monodelphis domestica* and *Ornithorhynchus anatinus*. One to one orthologs were fetched from Ensembl v75 and we kept the genes that had an ortholog in all the detailed species and the ones that only lacked platypus.

To build the multiple sequence alignment we kept the combination of isoforms that was more similar in length using PALO (Villanueva-Cañas et al. 2013) and employed Prank+F (Löytynoja and Goldman 2005). For the input tree we downloaded the

nucleotide tree from Meredith et al (Meredith et al. 2011) and included the rat using the distances as calculated in Laurie et al (Laurie et al. 2011). In the end we obtained 22 genes out of 319 for the young group and 2,565 for ancestral that fulfilled the requests. To localize events of positive selection (PS) we applied the branch-site test included in the package PAML (Yang 2007) to all the branches in all alignments (22 for young, 359 for ancestral). We calculated differences between young and ancestral groups looking at terminal and internal branches in separate contingency tables and applying a Fisher test using the R software (R Development Core Team 2010).

Pairwise dN/dS values were calculated for all the young genes present in human with an orthology in mouse using the package PAML from alignments built using Prank+F. We removed cases with branches showing dS<0.01, as such low dS values may result in inaccurate dN/dS estimates, and branches showing dS or dN>2 indicating saturation of substitutions. Finally, we also discarded a few outlier genes showing abnormally high dN/dS values (dN/dS>10).

Expression analyses

First we computed the median for each gene and tissue available in GTEX (release 2014, Ensembl v.74) using the ~4500 available human samples, then we calculated the tissue preferential expression index (Yanai et al. 2005) and the tissue with the highest expression value (max_t). A gene was considered tissue specific with a preferential expression index higher than 0.85. For Mouse

ENCODE, which was based on Ensembl v.65 (mm9), we computed the mean for each gene and for tissues with 2 replicas we only used the data if reproducibility indexes were lower than 0.1.

Statistical Data Analyses

We used Python to code the different analysis pipelines. Analysis of data, including generation of plots and statistical tests, was done with R (R Development Core Team 2010).

Supplementary material

Supplementary files are available from our web server evolutionarygenomics.imim.es (Publications, Datasets).

Acknowledgments

We acknowledge all members of the Evolutionary Genomics Group for thoughtful discussions, Sarah Djebali for table with gene expression values from mouse, Roger Hall for the beautiful drawings in Figure 1.

References

Albà MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol. Biol. Evol. 22:598–606.

Alba MM, Laskowski RA, Hancock JM. 2002. Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics 18:672–678.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Arendsee ZW, Li L, Wurtele ES. 2014. Coming of age: orphan genes in plants. Trends Plant Sci. 19:698–708.

Basu K, Graham LA, Campbell RL, Davies PL. 2015. Flies expand the repertoire of protein structures that bind ice. Proc. Natl. Acad. Sci. U. S. A. 112:737–742.

Bergboer JGM, Dulak MG, van Vlijmen-Willems IMJJ, Jonca N, van Wijk E, Hendriks WJAJ, Zeeuwen PLJM, Schalkwijk J. 2014. Analysis of protein-protein interaction between late cornified envelope proteins and corneodesmosin. Exp. Dermatol. 23:769–771.

Carvunis A-R, Rolland T, Wapinski I, et al. 2012. Proto-genes and de novo gene birth. Nature 487:370–374.

Casari G, de Daruvar A, Sander C, Schneider R. 1996. Bioinformatics and the discovery of gene function. Trends Genet. 12:244–245.

Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Veron G, Mulot B, Dupressoir A, Heidmann T. 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc. Natl. Acad. Sci. 109:E432–E441.

Derrien T, Johnson R, Bussotti G, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22:1775–1789.

Domazet-Loso T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.

Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC Evol. Biol. 11:47.

Flicek P, Amode MR, Barrell D, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42:D749–D755.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100:659–674.

Heinen TJAJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. Curr. Biol. 19:1527–1531.

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Rep.

Hrdinka M, Otahal P, Horejsi V. 2012. The transmembrane region is responsible for targeting of adaptor protein LAX into “heavy rafts”. PLoS One 7:e36330.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4:44–57.

Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. 15:205–213.

Jiang X, Li D, Yang J, Wen J, Chen H, Xiao X, Dai Y, Yang J, Tang Y. 2011. Characterization of a Novel Human Testis-Specific Gene: Testis Developmental Related Gene 1 (TDRG1). Tohoku J. Exp. Med. 225:311–318.

Jonca N, Leclerc EA, Caubet C, Simon M, Guerrin M, Serre G. 2011. Corneodesmosomes and corneodesmosin: from the stratum corneum cohesion to the pathophysiology of genodermatoses. Eur. J. Dermatol. 21 Suppl 2:35–42.

Jongeneel CV, Delorenzi M, Iseli C, et al. 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res. 15:1007–1014.

Kawasaki K, Lafont A-G, Sire J-Y. 2011. The evolution of milk casein genes from tooth genes before the origin of mammals. Mol. Biol. Evol. 28:2053–2061.

Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TCG. 2008. A Novel Gene Family Controls Species-Specific Morphological Traits in Hydra.Patel N, editor. PLoS Biol. 6:14.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25:404–413.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. Genome Res. 19:1752–1759.

Laurie S, Toll-Riera M, Radó-Trilla N, Albà MM. 2011. Sequence shortening in the rodent ancestor. Genome Res. 22:478–485.

Li L, Wurtele ES. 2014. The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. Plant Biotechnol. J.

Lindblad-Toh K, Garber M, Zuk O, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4:865–875.

Lonsdale J, Thomas J, Salvatore M, et al. 2013. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45:580–585.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. U. S. A. 102:10557–10562.

Martínez O, Reyes-Valdés MH. 2008. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. Proc. Natl. Acad. Sci. U. S. A. 105:9709–9714.

McKown RL, Coleman Frazier E V, Zadrozny KK, Deleault AM, Raab RW, Ryan DS, Sia RK, Lee JK, Laurie GW. 2014. A cleavage-potentiated fragment of tear lacritin is bactericidal. J. Biol. Chem. 289:22172–22182.

Meredith RW, Janečka JE, Gatesy J, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524.

Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TCG. 2009. Characterization of taxonomically restricted genes in a phylum-restricted cell type. Genome Biol. 10:R8.

Murphy DN, McLysaght A. 2012. De novo origin of protein-coding genes in murine rodents. PLoS One 7:e48650.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505:635–640.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 14:117.

O'Leary MA, Bloch JI, Flynn JJ, et al. 2013. The Placental Mammal Ancestor and the Post-K-Pg Radiation of Placentals. Science (80-.). 339:662–667.

Ohno S. 1970. Evolution by gene duplication. (Unwin A, editor.). Springer-Verlag

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. Elife 3:e01311.

Parker HG, VonHoldt BM, Quignon P, et al. 2009. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325:995–998.

Pervouchine DD, Djebali S, Breschi A, et al. 2015. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat. Commun. 6:5903.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing.Austria, editor. R Found. Stat. Comput. Vienna Austria 0:{ISBN} 3–900051 – 07–0.

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 9:e1003860.

Dos Reis M, Donoghue PCJ, Yang Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol. Lett. 10:20131003.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of de novo genes in human and chimpanzee. :25.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. Elife 3:1–24.

Schittek B, Hipfel R, Sauer B, et al. 2001. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. Nat. Immunol. 2:1133–1137.

Schlötterer C. 2015. Genes from scratch – the evolutionary fate of de novo genes. Trends Genet. 31:215–219.

Smeds L, Künstner A. 2011. ConDeTri--a content dependent read trimmer for Illumina data. PLoS One 6:e26314.

Soumillon M, Necsulea A, Weier M, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 3:2179–2190.

Steinert PM, Mack JW, Korge BP, Gan SQ, Haynes SR, Steven AC. 1991. Glycine loops in proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding proteins. Int. J. Biol. Macromol. 13:130–139.

Suenaga Y, Islam SMR, Alagu J, et al. 2014. NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas.Eilers M, editor. PLoS Genet. 10:e1003996.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nat. Rev. Genet. 12:692–702.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol. Biol. Evol. 26:603–612.

Toll-riera M, Castresana J, Albà MM. 2008. Accelerated Evolution of Genes of Recent Origin. Evolution (N. Y).

Toll-Riera M, Radó-Trilla N, Martys F, Albà MM. 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol. Biol. Evol. 29:883–886.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511–515.

Tsai H, Bobek LA. 1998. Human salivary histatins: promising anti-fungal therapeutic agents. Crit. Rev. Oral Biol. Med. 9:480–497.

Villanueva-Cañas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol. Evol. 5:457–467.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol. Evol. 5:439–455.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17:149–163.

Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. PLoS Genet. 7:e1002379.

Yanai I, Benjamin H, Shmoish M, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yeaman MR, Yount NY. 2003. Mechanisms of antimicrobial peptide action and resistance. Pharmacol. Rev. 55:27–55.

Yoder AD. 2013. Evolution. Fossils versus clocks. Science 339:656–658.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zhang J, Zhang Y, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. 30:411–415.

Zhu M, Janssen E, Leung K, Zhang W. 2002. Molecular cloning of a novel gene encoding a membrane-associated adaptor protein (LAX) in lymphocyte signaling. J. Biol. Chem. 277:46151–46158.

5. Discussion

5.1. Sequence change and positive selection

Multiple sequences alignments

Natural selection is one of the main players in evolution, acting on existent phenotypic variation. If a phenotype confers a higher fitness, their genotype will be positively selected, meaning that their allele frequencies will increase and its fixation will be favored. Positive selection can be subdivided into directional selection when it favors the fixation of an advantageous allele, and balancing selection, when it acts toward maintaining the polymorphism. If a mutation generates a phenotype with low fitness, it will be eliminated by purifying selection.

We have searched for signatures of positive selection in the different studies presented here using the branch-site test of positive selection implemented in the CodeML package (Yang 2007). A good quality alignment is a prerequisite to carry out any evolutionary analysis, including the branch-site test.

Alignment quality can be difficult to evaluate, since we do not usually know the "true" alignment. The proxy used to evaluate multiple alignments is the maximization of one or more criteria (such as percentage of identity) but this may not correspond to the real evolutionary path. Besides, progressive methods use heuristics and do not guarantee to converge to a global optimum.

The most used multiple sequence alignment (MSA) programs (Clustal, Mafft, T-Coffee, Muscle) show similar accuracy but vary in computation time (Thompson et al. 2011). This is not surprising,

since all of them belong to the family of progressive alignments and their differences are mostly due to the specific implementation.

There have been attempts to amend this problem by constructing alignments with different programs and keeping the consensus (Wallace et al. 2006) or combining them (Thompson et al. 2011). Another method named head or tails (HoT) was suggested to evaluate consistency of alignment programs (Landan and Graur 2007). The HoT score is the proportion of columns shared between the "Heads" alignment, generated from the original sequences, and the "Tails" alignment, generated from the reversed sequences. It was observed that this score favors algorithms that resolved ties consistently (most algorithms break ties at random) and its use as a measure of alignment quality was discouraged (Fletcher and Yang 2010). Another possibility is the trimming of poorly aligned regions (Castresana 2000; Capella-Gutiérrez et al. 2009), but at the cost of losing some of the information.

One of the many sources of errors is the alignment of non-homologous regions, yielding unrealistic alignments from a biological point of view and seriously affecting posterior analyses such as positive selection tests, that are particularly sensitive to alignment errors (Fletcher and Yang 2010). Another disadvantage of progressive alignments is that they are based on a greedy algorithm. This means that any mistake made early on, is propagated through all the rest of the progressive alignments.

Stretches of non-homologous sequences in orthologous proteins are common. This is due to annotation mistakes or the use of different isoforms. Most MSA methods are programmed to minimize the

number of insertions or deletions (gaps) producing very compact alignments. As a result, non-homologous positions are often erroneously aligned. Although reducing the gap penalty may help ameliorate this problem, the best is to use programs that do not systematically underestimate insertions. One such program is PRANK. This program uses graph-based algorithms to improve the recognition of non-homologous regions and infer gaps along the phylogeny (Löytynoja and Goldman 2008). Several studies have shown it is more accurate than other widely used programs (Fletcher and Yang 2010; Markova-Raina and Petrov 2011; Laurie et al. 2012; Villanueva-Cañas et al. 2013).

The correct placement of insertions in a MSA is of fundamental importance for the estimation of substitution rates. Misaligned positions are interpreted as nonsynonymous substitutions by CodeML, affecting the correct inference of positive selection. Thus, it is better to have gaps in regions where we lack complete information for one or more species used in the study than misaligned positions. Although we may miss some true events of positive selection, this greatly reduces the number of false positives.

In this thesis we have developed a method to improve the quality of the MSAs not by improving the alignment program but by optimizing the set of sequences to be aligned. PALO assumes that length is a good indicator of different isoforms and, by aligning sequences of similar length, reduces the problem of dN, dS, or dN/dS overestimation (Villanueva-Cañas et al. 2013). The fraction of positively selected genes increased gradually following the order Cons (alignment with highest conservation score), PALO, Longest

(alignment with longest isoform per gene), and Random (alignment with random isoform per gene); indicating a reduction in the number of false positives in the detection of positive selection (due to the alignment of non-homologous regions) with a negligible computational cost. With the future improvement of annotations and the characterization of new isoforms in the different species this problem will grow in importance.

In general, when performing genome-wide studies, methods and algorithms that were designed for single or carefully selected genes with a well-constructed alignment should not be directly applied without carefully analyzing their flaws and limitations first.

Methodological and functional limitations

Several limitations in the inference of positive selection events are pointed by Hughes (Hughes 2007). For example, he states that codons predicted to be positively selected could be explained by a relaxation or reduced efficiency (i.e. bottleneck in the population) of purifying selection or simply by the absence of synonymous substitutions. Differentiating between a true event of positive selection and relaxation of purifying selection is not trivial and opposite interpretations could arise from the same data. Relaxation of purifying selection can be interpreted as a reduction in the importance of a gene, while positive selection involves that a gene was important in a particular adaptation.

In our work we have used the branch-site test of positive selection, which has been designed to differentiate between relaxation of purifying selection and positive selection (Yang and dos Reis 2011).

Moreover we have used filters for dN, dS, and dN/dS to discard situations in which there is saturation of substitutions (dN or dS > 2), absence of synonymous substitutions (dS<0.01) or abnormal dN/dS values (>10).

Hughes is also right when he points out that most of the search for positive selection is usually applied without a priori biological hypothesis. Many works have been centered on the "blind" identification of genes that may have undergone episodes of adaptive selection in different species or lineages using divergence data (Clark et al. 2003; Arbiza et al. 2006; Bakewell et al. 2007; Gibbs et al. 2007; Kosiol et al. 2008; Vilella et al. 2009; Carneiro et al. 2012), often with little agreement between them, even if some of them have analyzed the same species. Some of the positively selected genes are probably explained by alignment errors, inadequate filters, and the selection of non-homologous isoforms when selecting a representative isoform for each gene. Moreover experimental verification is always needed to confirm the candidate cases identified by these studies. In the work presented in this thesis we have tried to focus on biological hypotheses, such as whether hibernation-related genes were enriched in positive selection signatures with respect to other types of genes.

However, even when we tried to be as careful as possible in the different steps of the process leading to positive selection inference (alignment, filtering, multiple test corrections, hypothesis testing…), we still found that manual inspection for the alignments of the candidate genes alignments was necessary. In the analysis of Lynx genes this resulted in 100 final candidates (96 for *Lynx sp.* and

4 for *Lynx lynx*) out of 487 total genes that were initially positive in the branch-site test. Statistical significance alone cannot be used as proof of adaptive evolution and caution should be taken when making such claims. In most cases, the most reliable explanation is a methodological issue, usually the alignment.

Another limitation of positive selection studies is the fact that one single amino acid replacement can be responsible for a new phenotype, and these events are hard to detect, One example is the melanocrotin-1 receptor (MC1R) in the Florida's Gulf Coast beach mice (*Peromyscus polionotus)*: a single amino-acid change produces a new coat color, providing concealment in the beach environment (Hoekstra et al. 2006). Moreover, experimental validation of the effect of the nonsynonymous substitutions may be difficult or even impossible to be performed. Finally, as we are using a reference genome, some of the substitutions may not be fixed in the populations but correspond to polymorphic sites. In conclusion, caution should be taken when extracting conclusions from genome-wide scans for positive selection.

5.2. Gene expression regulation: Hibernation

One of the biggest endeavors in the so called postgenomic era is the integration of (functionally) relevant genomic data to help understand the physiology of complex traits and how novel adaptations arise. One of the approaches used is the analysis of large genomic data sets in the search for genes associated with structural or functional phenotypes. A common technique is the quantification of gene expression through RNAseq (Wang et al. 2009) and the posterior identification of differentially expressed (DE) genes comparing samples from different conditions.

A harsh critic to these approaches was Sidney Brenner with his famous statement *'low input, high throughput, no output'*, referring to the use of high throughput experiments to biological problem solving (Brenner 2010). RNAseq studies are in some sense reductionist since only RNA abundance measurements are taken into account. This approach ignores all modulations that happen post-transcriptionally (i.e. phosphorylation, ligand binding…) and which cannot be detected through DE. Another problem is that when we sequence a tissue we are in fact analyzing many different cell types, especially in tissues with wide cell diversity, such as the brain. This increases background noise and makes it difficult to localize real changes in expression between conditions. Single-cell sequencing is becoming more popular (and cheap) and therefore, this might become a lesser issue in the near future. Another evident drawback of DE is that biological interpretations are based on what is already known, has been correctly entered into a database, and has adequate search and retrieval facilities (Hudson et al. 2012).

Therefore, the amount of knowledge that we can extract is limited in non-model species, where most of the functions are obtained by homology to known genes in model species (mostly human or mouse).

Despite these limitations RNAseq and related techniques are being continuously improved and have proven useful in many fields, providing a comprehensive view of the shifts in gene expression that occur in different tissues and at different times. For example, we have used this technique to pinpoint to physiological pathways involved in hibernation, as discussed earlier (Villanueva-Cañas et al. 2014).

However, once the actors have been identified it is sometimes difficult to assess their causality, unless we have detailed knowledge of how molecular pathways work. Sometimes, a few genes can generate a cascade of events and the use of interaction networks can help us identify the different gene modules involved in a process.

Hibernation is a crucial adaptation for the survival of the species populations in periods of resource scarcity. Improvements in the hibernation mechanism should be subject to natural selection and we should be able to trace the underlying mutations.

A long standing question in mammalian hibernation is whether the ancestral mammal was capable of hibernation and thus the required gene set was present in all mammals, or whether the hibernation phenotype appeared several times during evolution in different lineages, perhaps through the acquisition of new genes.

In our work we have seen that proteins involved in hibernation are not being positively selected, nor do we observe similar independent amino acid changes in hibernation-related proteins among different hibernating species. Previous work in mammalian hibernators hypothesized that hibernation is activated not by a set of genes unique to hibernators, but by selective expression of genes that are present in all mammals (Srere et al. 1992). Here we made use of all the current data and tried to reconcile different studies to validate this hypothesis; whereas only a few DE genes are shared between the different hibernating species, the number of shared genes increases when we focus on interaction networks and biological processes, pointing to the common use of different physiological "modules" such as the circadian clock or the beta-oxidation pathway. Nevertheless, more research is required to determine which genes trigger the hibernation process and identify if indeed they are the same genes in the different species.

The question of whether there are universal genetic mechanisms behind the hibernation adaptation requires further scrutiny and can only be tested through additional investigations of phylogenetically distant species. Currently, experiments about hibernation have only been carried out in a small set of species (bears, several squirrels, hamsters…) and mostly in laboratory conditions. It has been suggested that physiology and behavior of hibernating animals is strongly affected by acclimation to captivity. Indeed, much of the new recent discoveries in the field of hibernation come from work on wild animals (Geiser 2013).

Following this line of research we are using RNAseq to investigate gene expression dynamics during hibernation in the only known primate hibernators—Madagascar's dwarf lemurs (genus *Cheirogaleus*). To this aim, we are comparing the varying physiological states experienced throughout the year in, both captive and wild animals.

5.3. Taxonomically restricted genes in mammals

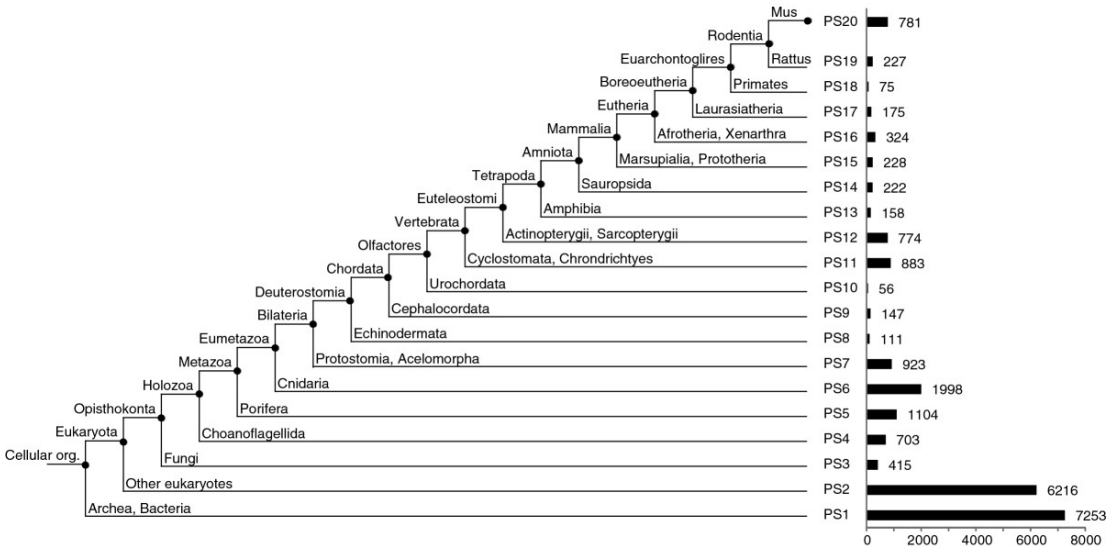
Identification and classification

The discovery of orphan genes is relatively recent (Dujon 1996). The sequencing of an ever increasing number of genomes has shown that these genes cannot be explained by lack of information from related species but that they are truly species or lineage-specific. Most of the articles in this field have been published in the last few years and our research group has been one of its pioneers in the detection and characterization of TRGs (Albà and Castresana 2007; Toll-Riera et al. 2009).

Phylostratigraphy, which is a common strategy to detect and classify Taxonomically Restricted Genes (TRGs), was initially carried out by Albà and Castresana (Albà and Castresana 2005) and further developed (and named) by Domazet-Loso (Domazet-Loso et al. 2007). It is based on sequence similarity searches (Blast) and it assigns the age of a gene from a focal species on the basis of the most distant species to which we obtain a sequence similarity match (Figure 5.1).

The main issue with this approach is that only reflects one species evolutionary history, rather than the history of the whole group, heavily relying on annotations. For example in a study performed in *Drosophila melanogaster*, the Diptera and Insecta taxonomic categories are only backed up by two species: *Anopheles gambiae* (mosquito) and *Apis mellifera* (honey bee), respectively (Domazet-Loso et al. 2007). With such a limited number of genomes, any gene present but not annotated in *A. gambiae* would be classified as Insecta specific or older if it was neither found in the bee genome.

We would be also missing Diptera or Insecta specific genes not annotated in the fruit fly genome or that were loss in its lineage. Furthermore, genes are annotated using criteria that usually include the presence of a homolog in other genomes, leading to TRG



depletion in branches close to the leaves.

Figure 5.1. Phylostratigraphy in mouse: genes are identified in one focal species (here *Mus musculus*) and their age is determined by the presence of homologues in other taxa. Modified from (Neme and Tautz 2013).

To overcome the limitations of phylostratigraphy we have developed a novel method. Although it is based also on Blast, it introduces several innovations in relation to the traditional methodology.

The first important novelty is the use of reconciled gene family's histories; using the single evolutionary histories from each of the species included in the analysis in a single tree we are able to reliably identify gene birth events in the different lineages and also compensate for incomplete annotations. However, reconcile gene

families from different species and assign them a unique common origin to a particular branch of the tree is not a trivial problem from the computational point of view. We have developed an efficient algorithm that navigates the tree and incorporates in each step all the sequence similarity searches information. It discards any genes with indirect relations to older genes and removes redundancy at every step.

The second improvement is the incorporation of RNAseq data from the different species and the reconstruction of novel transcripts. Using sequence similarity searches against our RNAseq reconstructions we were able to refine the age of the TRGs we found. This often led to the detection of an older time of birth for a gene (Villanueva et al. Submitted, chapter 4.4), reflecting the incompleteness of annotations.

Another major concern in the search for *de novo* genes is that they can sometimes be confounded with duplicated genes that evolve rapidly, beyond the threshold of detection of Blast similarity searches (Domazet-Loso and Tautz 2003). Some young genes might also arise from *Lazarus* or resurrected DNA (Graur et al. 2015): sequences that became pseudogenized and subsequently gained a new function. Although this problem only becomes apparent for genes older than mammals (Albà and Castresana 2007), we applied an additional step in our pipeline: We discarded any gene family with at least one mammalian protein with indirect sequence similarity (through their paralogs or sibling proteins) to sequences older than mammals.

Recently emerged proteins contain more low-complexity regions (LCR) in their sequences than older proteins (Toll-Riera et al. 2009). It has been argued that LCRs might play a role in the emergence of novel genes (Toll-Riera et al. 2012). However, sequences with a high percentage of repetitive regions cannot be used to perform Blast similarity searches, because they give too many spurious results, and they are filtered out. Because of this, we discarded sequences with a very high percentage of repetitive regions, lacking enough sequence length usable by Blast (<20 amino acids in a row).

Gene expression

When compared to microarrays, RNAseq experiments are more sensitive in the detection of lowly expressed genes (Wang et al. 2014), more reproducible (Zhao et al. 2014), and expression values from RNAseq correlate better with protein levels (Fu et al. 2009). This technology is therefore better suited to help in the detection and characterization of the transcriptome.

It is known that ubiquitous expression is a good indicator of housekeeping functions (Ramsköld et al. 2009). On the contrary, genes expressed in only one tissue are more likely to endure regulatory or context-dependent tasks. Many researchers have observed that young genes are more tissue-specific than older genes, pinpointing that they tend to be born in one tissue and perform a specific task (Palmieri et al. 2014; Ruiz-Orera et al. 2014; Tautz 2014). The patterns of gene expression in different mammalian tissues have been thoroughly characterized (Martínez

and Reyes-Valdés 2008; Ramsköld et al. 2009; Ardlie et al. 2015; Pervouchine et al. 2015). All the studies observed a higher diversity of expression in testis. Martinez et al. was the first one to realize that there were no abundant tissue-specific transcripts dominating the transcriptomic landscape in testis, therefore, more reads were available to allow the detection of less abundant transcripts (Martínez and Reyes-Valdés 2008).

The birth of new genes

In this thesis we have studied annotated protein coding genes that originated very recently, but we are, in fact just looking at the tip of the iceberg. The presence of such young genes points to a fundamental question: where are these genes coming from?

There is growing evidence that a large portion of a genome is being transcribed in one or more tissues at any given moment (Bernstein et al. 2012). Pervasive transcription is a steadily source of material for the generation of new genes and it is a feasible assumption that most of the transcripts expressed at low levels can be considered as selectively neutral. Consequently, regions of the genome with a high transcriptional activity would be more prone to the generation of new genes, as it was proposed for testis (Kaessmann 2010). Polev et al. suggested that transcriptional noise provides continuous scanning of the genome for those genes that can be adopted for concurrent cellular processes (Polev 2012). Although he does not mention directly *de novo* gene birth, he proposes that transcriptional noise would serve as a playground for testing existing genes in other tissues, where they could gain a novel function.

The use of deep transcriptomic sequencing has revealed the existence of many transcripts without long or conserved open reading frames (ORFs): they were termed long non-coding RNAs (lncRNAs). Most of lncRNAs are species or lineage-specific, poorly conserved and do not have a known function (Necsulea et al. 2014).

In a recent study from our group, using ribosome profiling[16] data we observed that a large fraction of the so called lncRNAs were scanned by ribosomes, suggesting that they are translated and produce short peptides in small quantities (Ruiz-Orera et al. 2014). These peptides become visible to natural selection. Interestingly lncRNAs show similar coding potential and sequence constraints than evolutionary young protein coding sequences –they are short, lineage-specific, and tissue specific– indicating that they might be a pool of proto-genes where new genes (and novel functions) can arise (Ruiz-Orera et al. 2014).

Despite this continuous flow of novel material, only a very small portion of it probably evolves under selection. This is consistent with the fact that the number of conserved protein-coding genes in a species is more or less constant, while lncRNAs are poorly conserved and very abundant. Genes that have acquired a function will tend to be conserved across species. We present many

---

[16] Ribosome profiling is a technique based on the isolation of mRNA within a ribosome through the use of nucleases that degrade unprotected mRNA regions (Ingolia et al. 2009). This technique analyzes the regions of mRNAs being converted to protein, as well as the levels of translation of each region.

examples in chapter 4.4. Many studies, including ours, have detected a bias in codon usage in young versus old genes (Toll-Riera et al. 2009; Palmieri et al. 2014). It has been suggested that the preferred usage of optimal codons might facilitate the emergence of *de novo* genes, specifically their translation. There has also been observed that some lncRNAs genes are under selective constrains. Many of the peptides encoded by lncRNAs may not yet have been found by proteomics, which shows important limitations for short proteins. Additionally, in this continuum of gene birth some RNAs may function both as coding and non-coding elements (Dinger et al. 2008). This may be a relatively widespread phenomenon and is in agreement with our model of *de novo* gene origination.

Taking all the evidence together, far from being non-coding 'junk' DNA, these transcripts with taxonomic restriction may be the source material for the birth of protein coding genes that play an important role in innovation and adaptation (Carvunis et al. 2012; Ruiz-Orera et al. 2014; Tautz 2014). It has been proven that genes can originate *de novo* (Tautz 2014; Schlötterer 2015) and some of them may even acquire essential functions (Reinhardt et al. 2013). This process, as well as the functional characterization of recently evolved genes, will be an active future of research.

6. Conclusions

1. We have shown that the use of the longest isoform in multiple sequence alignments (MSA) leads to an important overestimation of the fraction of positively selected sites, due to a higher fraction of misaligned positions in indel-rich alignments.

2. We have developed a method that selects the protein isoforms that are most similar in length for MSAs. This significantly improves the quality of the alignments generated and reduces the likelihood of wrongly identifying positively selected sites in genome-wide studies.

3. We have contributed to the assembly and annotation of the *Lynx pardinus* genome. We have identified 204 putative lynx-specific genes and detected several genes with signatures of positive selection.

4. We do not detect a significant increase in the number of hibernation-related genes with signatures of positive selection in hibernating species when compared with non-hibernating species, suggesting that species-specific changes in amino acids have not played a predominant role in this physiological adaptation.

5. There is a significant overlap in the molecular pathways involved in the hibernation adaptation in different mammalian species.

6. We have developed a novel methodology that allows the identification of *de novo* gene birth events along a phylogeny using genome and RNAseq data from all the species simultaneously.

7. We have conducted the largest search to date for taxonomically restricted genes (TRGs) in mammals. We have found that some mammalian-specific proteins are related to specific adaptations such as the formation of the skin or the defense against pathogens.

8. Young mammalian genes are short, tissue-specific, fast evolving, and positively charged when compared to older genes.

7. Future research

The promise of almost perfect assembled genomes based on Ion Torrent™ long reads (E. Myers in Brown and Morgenstern 2014) represents a huge opportunity for comparative genomics. It will allow answering many evolutionary questions that remain unanswered due to a lack of good assemblies and annotation. With this technology, the next generation of improved reference genome assemblies is almost within our reach. However, an assembly without a proper annotation has very little value, and, although the prediction of function through homology has proven very valuable we still need a lot of low-output experiments to unravel the functions and mechanisms of individual genes and pathways. This is especially relevant for orphan genes or TRGs, the vast majority of which remain functionally uncharacterized.

One of the most interesting questions is resolving the topology of the mammalian tree. Although genomic approaches have been performed extensively (Meredith et al. 2011), when contrasted with data from the fossil record, the correct relation between Xenathra, Afrotheria, and Laurasiatheria is currently unknown, with several competing hypotheses (Teeling and Hedges 2013). It is known that several genes tell different evolutionary stories and convergent evolution or gene duplication might affect such analyses. *De novo* gene origination is thought to be a rare event, thus genes that are lineage-specific should be key in resolving the tree topology, as it is unlikely that they appeared independently in several lineages. Better annotation of these genomes will allow the identification of many more TRGs.

Most of the studies that have been performed so far were done at the species level, looking for *de novo* genes that are already (or almost) fixed in the population. An area of research that will prove key to fully understand the dynamics of *de novo* gene birth and death will be population genetics, where we can explore all the regions that started to be transcribed (and/or translated) only recently and quantify their frequencies within populations. This will allow us to observe the whole spectrum of proto-genes, as described by Carvunis et al., and detect which ones are selected and spread throughout a population, perhaps due to a slight advantage being positively selected.

8. Annex

List of papers and other contributions associated with this thesis

Journal Articles

Villanueva-Cañas JL, Ruiz-Orera J., Albà MM. In preparation. Birth of new genes and evolutionary innovation in mammals.

Salla Vartia, **Villanueva-Cañas JL** et al. Submitted. A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding.

Lynx consortium, including **Villanueva-Cañas JL.** Submitted. Extreme genomic erosion in the highly endangered Iberian lynx.

Faherty SL*, **Villanueva-Cañas JL***, Peter H. Klopfer, Albà MM, Yoder AD. Submitted. Primate hibernators demonstrate universality in genetic regulation of mammalian hibernation.

Villanueva-Cañas JL, Faherty SL, Yoder AD, Albà MM. 2014. Comparative genomics of mammalian hibernators using gene networks. Integr. Comp. Biol. 54:452–462.

Villanueva-Cañas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol. Evol. 5:457–467.

Oral communications

Ciència i ficció. L'exploració creativa dels mons reals i dels irreals. Barcelona, 2015.

Title: *Hibernació humana: ciencia i ficció*

XII symposium on bioinformatics. Sevilla, 2014

Title: *Exploring hibernation in mammals through transcriptomics: the case of the hibernating lemur*

XIV Jornada de Biologia Evolutiva. Barcelona, 2014

Title: *Exploring hibernation in mammals through transcriptomics: the case of the hibernating lemur*

UCD Earth Institute, Dublin 2013

Title: *Insights into mammalian adaptation through genome-wide analysis*

XIII Jornada de Biologia Evolutiva. Barcelona, 2013

Title: *Incorporating protein isoform information into genome-wide studies: Impact on positive selection*

Poster presentations

Annual meeting of the Society for Molecular Biology and Evolution. Vienna, 2015

Title: *Characterization of* de novo *gene evolution across the mammalian phylogeny*

Annual meeting of the Society for Molecular Biology and Evolution. Dublin, 2012
Title: *Not another alignment program paper: How to improve multiple alignments by selecting the appropriate isoform combination.*

Tercer Congreso de la Sociedad Española de Biología Evolutiva (SESBE), Madrid, 2011.
Title: *Which protein isoform to use when making multiple alignments of orthologous genes? An empirical approach*

Other

Dia de la Ciència a les Escoles, 2014.
Title: *Històries de la genòmica*

9. References

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science 305:1462–1465.

Adams MD. 2000. The Genome Sequence of Drosophila melanogaster. Science (80-.). 287:2185–2195.

Akashi H, Osada N, Ohta T. 2012. Weak selection and protein evolution. Genetics 192:15–31.

Albà MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol. Biol. 7:53.

Alibardi L. 2003. Adaptation to the land: The skin of reptiles in comparison to that of amphibians and endotherm amniotes. J. Exp. Zool. B. Mol. Dev. Evol. 298:12–41.

Alioto T, Picardi E, Guigó R, Pesole G. 2013. ASPic-GeneID: a lightweight pipeline for gene prediction and alternative isoforms detection. Biomed Res. Int. 2013:502827.

Anderson PK. 1995. COMPETITION, PREDATION, AND THE EVOLUTION AND EXTINCTION OF STELLER'S SEA COW, HYDRODAMALIS GIGAS. Mar. Mammal Sci. 11:391–394.

Andjus RK, Lovelock JE. 1955. Reanimation of rats from body temperatures between 0 and 1° C by microwave diathermy. J. Physiol. 128:541–546.

Andrews MT. 2007. Advances in molecular biology of hibernation in mammals. Bioessays 29:431–440.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput. Biol. 2:e38.

Ardlie KG, Deluca DS, Segre A V., et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.). 348:648–660.

Arendsee ZW, Li L, Wurtele ES. 2014. Coming of age: orphan genes in plants. Trends Plant Sci. 19:698–708.

Bakewell M a, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. U. S. A. 104:7489–7494.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet. 22:437–446.

Blattner FR. 1997. The Complete Genome Sequence of Escherichia coli K-12. Science (80-.). 277:1453–1462.

Brenner S. 2010. Sequences and consequences. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365:207–212.

Van Breukelen F, Martin SL. 2015. The Hibernation Continuum: Physiological and Molecular Aspects of Metabolic Plasticity in Mammals. Physiology (Bethesda). 30:273–281.

Brown D, Morgenstern B eds. 2014. Algorithms in Bioinformatics. Berlin, Heidelberg: Springer Berlin Heidelberg

C. elegans sequencing consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium [published erratum appears in Science 1999 Jan 1;283(5398):35]. Science (80-.). 282:2012–2018.

Caldwell WH. 1887. The Embryology of Monotremata and Marsupialia. Part I. Philos. Trans. R. Soc. B Biol. Sci. 178:463–486.

Cannon B, Nedergaard J. 2004. Brown adipose tissue: function and physiological significance. Physiol. Rev. 84:277–359.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carey H V, Andrews MT, Martin SL. 2003. Mammalian hibernation: cellular and molecular responses to depressed metabolism and low temperature. Physiol. Rev. 83:1153–1181.

Carlson R. 2003. The pace and proliferation of biological technologies. Biosecur. Bioterror. 1:203–214.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar J a, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (Oryctolagus cuniculus) genome. Mol. Biol. Evol.

Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3:e245.

Carvunis A-R, Rolland T, Wapinski I, et al. 2012. Proto-genes and de novo gene birth. Nature advance on:1–5.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chamary J V, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7:98–108.

Clark AG, Glanowski S, Nielsen R, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783–786.

Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI. 2013. Positively selected sites in cetacean myoglobins contribute to protein stability. PLoS Comput. Biol. 9:e1002929.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput. Biol. 4:e1000176.

Domazet-Loso T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 13:2213–2219.

Dujon B. 1996. The yeast genome project: what did we learn? Trends Genet. 12:263–270.

Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7:n/a – n/a.

Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. 27:2257–2267.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Fu X, Fu N, Guo S, et al. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics 10:161.

Geiser F. 2013. Hibernation. Curr. Biol. 23:R188–R193.

Gerstein MB, Bruce C, Rozowsky JS, et al. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res. 17:669–681.

Gibbs R a, Rogers J, Katze MG, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

Goffeau A, Barrell BG, Bussey H, et al. 1996. Life with 6000 genes. Science 274:546, 563–567.

Gotea V, Petrykowska HM, Elnitski L. 2013. Bidirectional Promoters as Important Drivers for the Emergence of Species-Specific Transcripts. PLoS One 8.

Haldane JBS. 1932. The Causes of Evolution.

Hall BK. 2005. Consideration of the neural crest and its skeletal derivatives in the context of novelty/innovation. J. Exp. Zool. B. Mol. Dev. Evol. 304:548–557.

Hickman CP, Roberts LS, Larson A. 2001. Integrated Principles of Zoology.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. Science 313:101–104.

Hogeweg P, Hesper B. 1984. The alignment of sets of sequences and the construction of phyletic trees: An integrated method. J. Mol. Evol. 20:175–186.

Hu Y, Meng J, Wang Y, Li C. 2005. Large Mesozoic mammals fed on young dinosaurs. Nature 433:149–152.

Hudson NJ, Dalrymple BP, Reverter A. 2012. Beyond differential expression: the quest for causal mutations and effector molecules. BMC Genomics 13:356.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity (Edinb). 99:364–373.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223.

Jacob F. 1977. Evolution and tinkering. Science 196:1161–1166.

Jenkins CN, Pimm SL, Joppa LN. 2013. Global patterns of terrestrial vertebrate diversity and conservation. Proc. Natl. Acad. Sci. U. S. A. 110:E2602–E2610.

Jensen JD, Wong A, Aquadro CF. 2007. Approaches for identifying targets of positive selection. Trends Genet. 23:568–577.

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. 2008. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell 132:783–793.

Jiang X, Li D, Yang J, Wen J, Chen H, Xiao X, Dai Y, Yang J, Tang Y. 2011. Characterization of a novel human testis-specific gene: testis developmental related gene 1 (TDRG1). Tohoku J. Exp. Med. 225:311–318.

Jones FC, Grabherr MG, Chan YF, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

Jones KE, Bielby J, Cardillo M, et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals.Michener WK, editor. Ecology 90:2648–2648.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Kasahara M. 2007. The 2R hypothesis: an update. Curr. Opin. Immunol. 19:547–552.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TCG. 2008. A novel gene family controls species-specific morphological traits in Hydra. PLoS Biol. 6:e278.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25:404–413.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Knight CG, Zitzmann N, Prabhakar S, Antrobus R, Dwek R, Hebestreit H, Rainey PB. 2006. Unraveling adaptive evolution: how a single point mutation affects the protein coregulation network. Nat. Genet. 38:1015–1022.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. Genome Res. 19:1752–1759.

Koonin E V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

Kosiol C, Vinar TT, Da Fonseca RR, et al. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4:e1000144.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol. Biol. Evol. 24:1380–1383.

Lander ES, Linton LM, Birren B, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Leys SP, Riesgo A. 2012. Epithelia, an evolutionary novelty of metazoans. J. Exp. Zool. B. Mol. Dev. Evol. 318:438–447.

Li C-Y, Zhang Y, Wang Z, et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput. Biol. 6:e1000734.

Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. 20:408–420.

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. Plant J. 58:485–498.

Li L, Wurtele ES. 2014. The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. Plant Biotechnol. J.

Light S, Basile W, Elofsson A. 2014. Orphans and new gene origination, a structural and evolutionary perspective. Curr. Opin. Struct. Biol. 26C:73–83.

Linné C von. 1758. Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Holmiae : Impensis Direct. Laurentii Salvii,

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4:865–875.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. Annu. Rev. Genet. 47:307–333.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635.

Luo Z-X, Ji Q, Wible JR, Yuan C-X. 2003. An Early Cretaceous tribosphenic mammal and metatherian evolution. Science 302:1934–1940.

Luo Z-X, Yuan C-X, Meng Q-J, Ji Q. 2011. A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476:442–445.

Luo Z-X. 2007. Transformation and diversification in early mammal evolution. Nature 450:1011–1019.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 21:863–874.

Martínez O, Reyes-Valdés MH. 2008. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. Proc. Natl. Acad. Sci. U. S. A. 105:9709–9714.

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science 264:421–424.

Meredith RW, Janečka JE, Gatesy J, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524.

Mess AM, Ferner KJ. 2010. Evolution and development of gas exchange structures in Mammalia: the placenta and the lung. Respir. Physiol. Neurobiol. 173 Suppl:S74–S82.

Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. Genome Res. 23:34–45.

Montgelard C, Catzeflis FM, Douzery E. 1997. Phylogenetic relationships of artiodactyls and cetaceans as deduced from the comparison of cytochrome b and 12S rRNA mitochondrial sequences. Mol. Biol. Evol. 14:550–559.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505:635–640.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 14:117.

Nilsson MA, Churakov G, Sommer M, Tran N Van, Zemann A, Brosius J, Schmitz J. 2010. Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS Biol. 8:e1000436.

Northcutt RG. 2002. Understanding vertebrate brain evolution. Integr. Comp. Biol. 42:743–756.

Ohno S. 1970. Evolution by gene duplication. (Unwin A, editor.). Springer-Verlag

Ohno S. 1982. Evolution Is Condemned to Rely upon Variations of the Same Theme: The One Ancestral Sequence for Genes and Spacers. Perspect. Biol. Med. 25:559–572.

OHTA T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature 246:96–98.

Okamura K, Feuk L, Marquès-Bonet T, Navarro A, Scherer SW. 2006. Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics 88:690–697.

Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of Drosophila orphan genes. Elife 3:e01311–e01311.

Pervouchine DD, Djebali S, Breschi A, et al. 2015. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat. Commun. 6:5903.

Peterson T, Müller GB. 2013. What is evolutionary novelty? Process versus character based definitions. J. Exp. Zool. B. Mol. Dev. Evol. 320:345–350.

Petousi N, Robbins PA. 2014. Human adaptation to the hypoxia of high altitude: the Tibetan paradigm from the pregenomic to the postgenomic era. J. Appl. Physiol. 116:875–884.

Pettigrew JD. 1999. Electroreception in monotremes. J. Exp. Biol. 202:1447–1454.

Polev D. 2012. Transcriptional noise as a driver of gene evolution. J. Theor. Biol. 293:27–33.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc. Natl. Acad. Sci. U. S. A. 104 Suppl :8605–8612.

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh S-D, True JR, Carroll SB. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440:1050–1053.

Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5:e1000598.

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLoS Genet. 9.

Reuter JA, Spacek D V., Snyder MP. 2015. High-Throughput Sequencing Technologies. Mol. Cell 58:586–597.

Richards S. 2015. It's more than stamp collecting: how genome sequencing can unify biological research. Trends Genet.

Roth G, Dicke U. 2005. Evolution of the brain and intelligence. Trends Cogn. Sci. 9:250–257.

Rowe T. 1988. Definition, diagnosis, and origin of Mammalia. J. Vertebr. Paleontol. 8:241–264.

Ruf T, Geiser F. 2014. Daily torpor and hibernation in birds and mammals. Biol. Rev. Camb. Philos. Soc.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. Elife 3:1–24.

Samusik N, Krukovskaya L, Meln I, Shilov E, Kozlov AP. 2013. PBOV1 is a human de novo gene with tumor-specific

expression that is associated with a positive clinical outcome of cancer.Prokunina-Olsson L, editor. PLoS One 8:e56162.

Schlötterer C. 2015. Genes from scratch – the evolutionary fate of de novo genes. Trends Genet. 31:215–219.

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. Nature 457:818–823.

SMITH AU, LOVELOCK JE, PARKES AS. 1954. Resuscitation of hamsters after supercooling or partial crystallization at body temperature below 0 degrees C. Nature 173:1136–1137.

Smith JJ, Kuraku S, Holt C, et al. 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat. Genet. 45:415–421, 421e1–e2.

Soskine M, Tawfik DS. 2010. Mutational effects and the evolution of new protein functions. Nat. Rev. Genet. 11:572–582.

Srere HK, Wang LC, Martin SL. 1992. Central role for differential gene expression in mammalian hibernation. Proc. Natl. Acad. Sci. U. S. A. 89:7119–7123.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.

Stern DL. 2000. Evolutionary developmental biology and the problem of variation. Evolution 54:1079–1091.

Sturm RA. 2009. Molecular genetics of human pigmentation diversity. Hum. Mol. Genet. 18:R9–R17.

Suenaga Y, Islam SMR, Alagu J, et al. 2014. NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas.Eilers M, editor. PLoS Genet. 10:e1003996.

Sun W, Zhao X-W, Zhang Z. 2015. Identification and evolution of the orphan genes in the domestic silkworm, Bombyx mori. FEBS Lett.

Tabuce R, Marivaux L, Adaci M, Bensalah M, Hartenberger J-L, Mahboubi M, Mebrouk F, Tafforeau P, Jaeger J-J. 2007. Early Tertiary mammals from North Africa reinforce the molecular Afrotheria clade. Proc. Biol. Sci. 274:1159–1166.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nat. Rev. Genet. 12:692–702.

Tautz D. 2014. The discovery of de novo gene evolution. Perspect. Biol. Med. 57:149–161.

Teeling EC, Hedges SB. 2013. Making the impossible possible: rooting the tree of placental mammals. Mol. Biol. Evol. 30:1999–2000.

Thompson JD, Linard B, Lecompte O, Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6:e18093.

Tøien Ø, Blake J, Edgar DM, Grahn DA, Heller HC, Barnes BM. 2011. Hibernation in black bears: independence of metabolic suppression from body temperature. Science 331:906–909.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol. Biol. Evol. 26:603–612.

Toll-Riera M, Radó-Trilla N, Martys F, Albà MM. 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol. Biol. Evol. 29:883–886.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.

Villanueva-Cañas JL, Faherty SL, Yoder AD, Albà MM. 2014. Comparative genomics of mammalian hibernators using gene networks. Integr. Comp. Biol. 54:452–462.

De Vries H. 1905. Species and Varieties: Their Origin by Mutation. The Open Court Publishing Company

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 34:1692–1699.

Wang LCH, Lee TF. 2011. Torpor and Hibernation in Mammals: Metabolic , Physiological , and Biochemical Adaptations. In: Compr Physiol. p. 507–532.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57–63.

Wilson DE, Reeder DM. 2005. Mammal Species of the World.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013a. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol. Evol. 5:439–455.

Wu D-D, Zhang Y-P. 2013. Evolution and function of de novo originated genes. Mol. Phylogenet. Evol. 67:541–545.

Wu P, Jiang T-X, Suksaweang S, Widelitz RB, Chuong C-M. 2004. Molecular shaping of the beak. Science 305:1465–1466.

Wu X, Sharp PA. 2013. Divergent transcription: A driving force for new gene origination? Cell 155.

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. 28:1217–1228.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zhang J, Zhang Y, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. 30:411–415.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. Science (80-.). 343:769–772.

La impressió d'aquesta tesi ha estat posible gràcies a l'ajut per a la finalització de tesis doctorals de la Fundació IMIM

UAG