UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Animal Medicine, Production and Health (MAPS)

Ph.D Course in *Veterinary Science*
Series: XXXIV

Identification of the Molecular Mechanisms and Signatures associated with Longevity and Cancer Resistance in Mammals

Coordinator: Ch.mo Prof. Mattia Cecchinato
Supervisor: Ch.mo Prof. Cristian Taccioli

Ph.D Student: Chiara Vischioni

Declaration.

I declare that the present thesis has not been previously submitted as an exercise for degree at University of Padova, or any other University, and I further declare that the work embodied is my own.

© 2022 – Chiara Vischioni
All rights reserved.

Abstract.

Cancer is a rooted evolutionarily disease, born with the development of the multicellularity, and inherently caused by mutations occurring at somatic level or inherited through the germline. Yet, there is a whole world behind this simple academic definition. Some authors argue that it is not just a disease, but it rather represents a force able to drive the biological systems, acting itself as evolutionary mechanism able to selectively shape the adaptation of a species. Surprisingly, at phylogenetic level, susceptibility to cancer greatly varies from one species to another. Indeed, it is known that within the same species body size and lifespan are strongly correlated with the probability of developing cancer, whereas, across different ones, this association disappears, being replaced by what it is recognized as Peto's Paradox biological dilemma: theoretically, over time, cells acquire and accumulate mutations that, in some cases, can lead to the development of a tumorigenic event. Since every cell in the body has the same potential to become cancerous, larger and longer-living species should proportionally have a higher risk of cancer. However, Peto teaches us that some of them have evolved cancer suppression strategies able to parallelly coexist alongside their grater size and longevity. In this framework, oncology and comparative genomics are the only tools able to answer those question wondering why some species are more resistant to cancer compared to others, despite their phenotypic constraints such as size and high longevity. Understanding how Nature has solved the problem of cancer suppression during evolution could, therefore, be translated into cancer prevention strategies for human and veterinary research. To date, mechanisms proposed for the resolution of Peto's paradox include the reduction in the number of oncogenes copies, or, conversely, the increase in the number of suppressor genes. In particular, Copy Number Variations (CNVs), are regions of DNA found deleted and/or duplicated within the genome, which may reflect a phenotypic variation, causing, in some cases, disease. Therefore, investigating the copy number composition of genes in the genome of long-living and/or big size animals showing a low cancer rate could shed light on new molecular targets related to ageing and cancer-resistance that are still unknown. Specifically, **Chapter II** describes VarNuCopy, the first online tool that I developed during the course of my Ph.D, that collects and compares CNVs from the genome of 233 organisms (mammalian and non-mammalian), correlating, for a selected subset, the copy number with some phenotypic traits of the species. **Chapter III**, exploiting VarNuCopy data, identifies

for the first time the microRNAs family as a new biomarker able to discriminate the cancer predisposition of a species. Finally, **Chapter IV** explains how and why the single-cell organism *S. cerevisiae* can be considered as a key model in the study of ageing processes and cancer-related pathways, reporting also my personal research experience carried out during the nine months of my Ph.D spent abroad.

Riassunto.

Il cancro è una malattia che evolve seguendo le regole della selezione naturale, nata con lo sviluppo della multicellularità, e intrinsecamente causata da mutazioni che si verificano sia a livello somatico che ereditate attraverso la linea germinale. Eppure, aldilà di questa semplice definizione accademica, dietro lo sviluppo delle malattie oncologiche, si cela un mondo molto più ampio, per la maggior parte ancora sconosciuto. Alcuni autori sostengono che il cancro non sia solo una malattia, ma che rappresenti una forza evolutiva in grado di modellare selettivamente l'adattamento di una specie. Sorprendentemente, a livello filogenetico, la suscettibilità al cancro varia notevolmente. Infatti, è noto che all'interno della stessa specie le dimensioni del corpo e la durata della vita siano fortemente correlate alla probabilità di sviluppare un tumore, mentre, tra specie diverse, questa associazione scompare, lasciando il posto a quello che viene definito come il paradosso di Peto: teoricamente, poiché ogni cellula del corpo ha la stessa probabilità di diventare cancerosa, le specie dotate di una massa maggiore e quelle più longeve dovrebbero proporzionalmente avere un rischio maggiore di tumorigenesi. Tuttavia, Peto ci insegna che alcune di esse hanno evoluto strategie di soppressione in grado di coesistere con la loro grande dimensione e l'elevata longevità. In questo contesto, l'oncologia e la genomica comparativa sono gli unici strumenti in grado di rispondere a quelle domande sul perché, nonostante i loro vincoli fenotipici come dimensioni ed elevata longevità, alcune specie siano più resistenti al cancro rispetto ad altre. Nel corso del tempo, le cellule acquisiscono e accumulano mutazioni che, in alcuni casi, possono portare allo sviluppo di tumore. Capire in che modo la Natura abbia risolto il problema della soppressione del cancro durante l'evoluzione, potrebbe quindi essere tradotto in strategie di prevenzione nell'ambito della ricerca umana e veterinaria. Ad oggi, tra i meccanismi proposti per la risoluzione del paradosso di Peto si trovano la riduzione del numero di copie degli oncogeni o, al contrario, l'aumento del numero di geni soppressori. In particolare, le *Copy Number Variations* (CNVs), sono regioni di DNA delete e/o duplicate all'interno del genoma, e portano ad una variazione fenotipica, causando, in alcuni casi, malattia. Pertanto, indagare la composizione in *copy number* nel genoma di animali longevi e/o di taglia grande, ma che mostrano un basso tasso di incidenza di neoplasia, potrebbe far luce su nuovi target molecolari legati all'invecchiamento ad oggi ancora sconosciuti. In particolare, il **Capitolo II** descrive VarNuCopy, il database che ho sviluppato durante il corso del mio dottorato, e che raccoglie e confronta le CNVs del

genoma di 233 organismi (mammiferi e non), correlando, per un sottoinsieme selezionato, il numero di copie con alcuni tratti fenotipici della specie. Il **Capitolo III**, sfruttando i dati di VarNuCopy, riporta per la prima volta la famiglia dei microRNA come un nuovo target molecolare in grado di discriminare per la predisposizione al cancro di una specie. Infine, il **Capitolo IV** spiega come e perché il lievito unicellulare *S. cerevisiae* possa essere considerato un modello chiave nello studio dei processi di invecchiamento e del cancro, riportando anche la mia esperienza di ricerca personale svolta durante i nove mesi trascorsi all'estero.

To my parents and sister,
who have always encouraged all my adventures,
for always being there even when I didn't know I needed it.

List of Papers.

This manuscript is based on the following papers:

I. De Chiara, M., Barré, B. P., Persson, K., Irizar, A., **Vischioni, C.**, Khaiwal, S., ... & Liti, G. (2022). Domestication reprogrammed the budding yeast life cycle. *Nature Ecology & Evolution*, 1-13.

II. **Vischioni, C.**, Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2022). Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research. *Big Data Research*, *27*, 100298.

III. **Vischioni, C.**, Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2021). miRNAs Copy Number Variations repertoire as hallmark indicator of cancer species predisposition. *bioRxiv*.

IV. Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., Taccioli, C., & **Vischioni, C**. (2020, September). VarCopy: a Visual Exploratory Data Analysis Platform for Copy Number Variation Studies. In *2020 24th International Conference Information Visualisation (IV)* (pp. 391-396). IEEE.

Additional papers.

The following works have been published during my doctoral studies, but they are not directly linked with the aim of this thesis:

I. Bergamini, C. M., **Vischioni, C.**, Aguiari, G., Grandi, C., Terrazzan, A., Volinia, S., ... & Taccioli, C. (2021). Inhibition of the lncRNA Coded within Transglutaminase 2 Gene Impacts Several Relevant Networks in MCF-7 Breast Cancer Cells. *Non-coding RNA*, *7*(3), 49.

II. Modi, A., Vergata, C., Zilli, C., **Vischioni, C.**, Vai, S., Tagliazucchi, G. M., ... & Taccioli, C. (2021). Successful extraction of insect DNA from recent copal inclusions: limits and perspectives. *Scientific reports*, *11*(1), 1-8.

III. **Vischioni, C.**, Giaccone, V., Catellani, P., Alberghini, L., Scapin, R. M., & Taccioli, C. (2021). GBRAP: a tool to retrieve, parse and analyze GenBank files of viral and bacterial species. *bioRxiv*.

IV. Ghidoni, G., Martoglia, R., Taccioli, C., & **Vischioni, C.** (2020, September). InstaCircos: a Web Application for Fast and Interactive Circular Visualization of Large Genomic Data (Work in Progress). In *2020 24th International Conference Information Visualisation (IV)* (pp. 385-390). IEEE.

Contents.

Chapter I:

Introduction.

1. General Overview

Aging is defined as the process of accumulation of structural, molecular, cellular, and functional changes affecting a cell or an organism over the passage of time. According to the World Health Organization (https://www.who.int/), in the next thirty years, the number of persons aged older than 80 is expected to triple, potentially impacting the entire economic sectors of multiple nations. Indeed, in the light of this imminent demographic change, all countries must be prepared to face multiple challenges in order to ensure that their health and social systems will be ready to withstand the high pressure of this event. In this framework, global efforts are required to implement existing knowledge about the prevention and the treatment of age-related diseases, such as cardiovascular disease, Parkinson, Alzheimer, and cancer. Indeed, it is possible that life expectancy will keep increasing until a stable plateau defined by individual intrinsic genetics is reached. At this point cardiovascular failure or cancer will be the principal hazard of human life (Kirkwood, 2017). Nowadays, one of the goals of aging research is to promote the investigation on the study of age-related diseases, moving towards the development of technologies capable to delay the process of aging and extend human health and life. Because this process is defined as a progressive deterioration of physiological functions accompanied by an increase in vulnerability and mortality, the primary focus of aging research must be based on two fundamental concepts: (*i*) health benefits, with the aim to preserve the health of the individual by postponing the onset of disease, and (*ii*) life extension, obtained slowing down, tackle, and therapeutically curing aging in order to enable people to live longer. Although in the last two centuries it has been subject of numerous studies, currently, within the scientific community, it is not possible to identify a common consensus on what the nature of the aging process is. In particular, the issues on which different scientists are still debating are how aging is related to mortality rate, functional decline and damage accumulation, and whether it is biologically programmed (Cohen et al., 2020a). Lack of consensus on an aging biology paradigm clearly emerged after the 'Biology of Aging Symposium' held in Montreal in 2019, which gathered 44 expert speakers of the field (Cohen et al., 2020b). As pointed out by A.A. Cohen, the main discussion dynamically revolved around the question *'Do we know what ageing is?'*. Surprisingly, the debate highlighted lack of agreement and common vision both on the core question and its basic principles. On the other hand, all the scientists agreed on the heterogeneity and the multifactorial nature of the process, emphasizing the need to reach a common consensus on key issues. Personally, what I found interesting for the purpose of this manuscript, is the diversity in the answers to the additional question "*describe your understanding of what causes aging, and what it is, or is not, at a mechanistic level*", that are partly summarized in Table 1. In my opinion, all the definitions given by the audience raise important questions regarding the topic, but they may be imbalanced according to the different scientific background and expertise of the speaker. As a young scientist who has just approached the aging research field for a few years, I cannot myself identify a global and fulfilling definition of aging, besides the usual and academic one. Probably, my personal point of view is a combination between multiple perspectives. I agree on the idea of the imperfect optimization in maintaining a balance between evolutionary constraints, environmental constraints, and cellular dynamical equilibrium proposed by Vanhaelen Q.; on the other hand, I also think that with time, a gradual break-down of cellular components caused by multiple factors such as DNA damage, mutations, genetic/epigenetic pre-programmed senescence occurs, leading to a dysregulation of the system, which fails to balance internal and external damages (as hinted by Gourbunova V., and Anglas U.)

	Aging definition
L. Ferrucci	"Aging is the progressive shrinking of the biological mechanisms that surveil and repair cellular and intercellular damage and miscommunication. Since the manifestations of aging are observable and stereotyped, the mechanisms of aging should be discovered using these manifestations as gold standard."
J.F. Lemaître	"As an evolutionary biologist, I see aging as the decrease in the age-specific contribution to fitness. In fact, I prefer the term senescence to describe this evolutionary process. […] In that case, aging could represent the effect of time on organisms from birth to death while senescence should represent the decline in survival and reproductive probabilities with increasing age (themselves underpinned by a decline in physiological functions). […] At a mechanistic level, aging (or I should say senescence) correspond to any deterioration of cellular/physiological traits that will ultimately impact fitness."
Q. Vanhaelen	"Mechanistically, aging is a disruption of the homeostasis established between cellular processes. […] Aging is the result of this imperfect optimization to maintain a balance between evolutionary constraints, cellular dynamical equilibrium, and environmental constraints […]. Aging is not a programmed process but a consequence of this search for an equilibrium."
J. Van Raamsdonk	"Aging is the progressive decline of function with increasing chronological age due to internal factors that are not dependent on environment, which leads to an increased probability of death. Aging is caused by a genetically programmed switch that downregulates cellular pathways involved in homeostasis, stress response, repair etc. […]"
V. Gorbunova	"Repair machineries fail to keep up with internal and external damage; systems become dysregulated and eventually collapse."
G. Pawelec	"Ageing is caused by a breakdown in repair mechanisms due to a shift of resource allocation after reproduction, modulated according to the environmental niche of the organism."
D. Frasca	"Aging is the result of the ability of an individual to adapt to the progressive accumulation of stress stimuli that accumulate throughout life at different rates in all tissues and organs. […]"

V.N. Gladyshev	“In its essence, aging is the accumulation of deleterious changes. More specifically, aging is the increase in deleterious changes (the deleteriome) as a by-product of life (metabolism) under ecological/ evolutionary constraints, with its rate adjusted by genetic (a major contributor), environmental and stochastic processes (these two primarily contribute to variation within species). Aging starts very early in life and may be tracked by a combination of clocks and biomarkers as well as by functional assays and mortality (the latter is not always, e.g. not in all life stages, not in all species that age).”
Anatoli I. Yashin	“Aging results from the imperfect design of our bodies to deal with existing environment. […] It might be a reason for the evolution to develop mortal organisms with different lifespans and different rates of aging. This reason could deal with the need to maintain sustainability of the Earth's ecological system in the condition of limited energy supply (sun energy) […].
M. Ivanchenko	“Aging is a continuous dynamical process, starting as development, then adaptation, and at the later stage accumulation of damage, garbage, adaptation controversies and failures. It is a product of genetic background and environmental factors, and as a result highly heterogeneous among individuals.”
U. Anglas	“The gradual break-down of cellular components, leading to the eventually death of the organism, associated with time. This is caused by DNA damage, mutations, genetic/epigenetic pre-programmed senescence, protein aggregates and environmental stress.”
V. Legault	“I view aging as the gradual exhaustion of the system (organism), arising from the repeated responses to external and internal perturbations. The system might find alternative stable states along the way but will eventually collapse at a certain point.”
F. Dufour	“Aging is not programmed. Aging is the result of complex interactions between the genome and physiological and environmental changes, with feedback loops that lead to physical and mental impairment. Organisms are not adapted to naturally live indefinitely.”

Table 1: Subset of the different definition of Aging, presented at the “Biology of Aging Symposium” (2019)
(**Source**: adapted from Table1 & Supplementary Material1 - *Cohen et al., 2020a*)

While writing this thesis, I was wondering how other researchers with different interests and backgrounds would think in this respect. Recently, during a routine discussion, I brought the topic on the table, asking some colleagues to choose a definition of ageing among the ones proposed by the experts panel in 2019. I was curious about how they might define the process from their (scientific) point of view, to verify if there were any repetitive common words or concepts among their quotes. For the moment, this is just a personal and informal questioning, but I think that inquiring on external opinions, especially given by people who are not familiar with the field, can be extremely interesting in order to recognize similar identity patterns. Indeed, the words used to describe the phenomena can be used as proxies of how the ageing paradigm is seen by the general public and the society, and most importantly, to define a universal concept of aging to guide future research (Gladyshev, 2016).

1.1. Aging theories

Currently, several different theories exist, all of them covering different aspects of aging (Moldakozhayev et al., 2021). The most important and relevant, are summarized below:

1) **programmed theory**: aging as an altruistic advantageous genetic program that has evolved to benefit future generations, freeing up resources consumed by the older ones in order to be used by future organisms (Longo et al., 2005);
2) **evolutionary theory**: mutations can accumulate in the genomes over evolutionary time scales when the forces of natural selection decline as a function of age. Although they may show a beneficial or neutral effect during early life stages, they can turn to be detrimental when accumulating during evolution, causing ageing (Medawar, 1952; Williams, 1957);
3) **free radical theory**: reactive oxygen species (ROS) are a by-product of cellular metabolic processes, being one of the primary cellular damage sources (Harman, 1956). Accumulation of these damages leads to aging and senescence.
4) **disposable soma theory**: the inability of organisms to support both maintenance and reproduction, lead to damage accumulation (Kirkwood, 1977). Recently, this theory has been replaced by the **hyperfunction theory of ageing**, according to which the excessive gene functions replaced the concepts of molecular damages caused by the unbalanced allocation of the available resources (Blagosklonny, 2008).

Aging theories are very different, each of them describing a particular aspect of the whole process. A new hypothesis supported by the recent reinterpretation of the Second Law of Thermodynamics (Lambert, 2007), argues that these changes are the result of modifications in the entropy levels. Indeed, energy state levels capable of maintaining high fidelity during the molecular processes tend to be evolutionarily selected until, during aging, their alteration renders those molecules inactive or malfunctioning (Hayflick, 2007; Fariselli and Taccioli et al., 2021). The only biological feature able to overcome the evolutionary scale of time is the information encoded within DNA molecules, but even this one is not entirely immune to mutation or change (Hayflick, 2000). For this reason, in 2016, Gladyshev V.N., developed a principle capable of enclosing in a single point of view all the existing perspectives: the **deleteriome model**. According to his vision, the *deleteriome* is described as the accumulation of all the molecular damage of an organism, due to the imperfect activity of all biological systems. Therefore, the biological age mirrors the set of cumulative, deleterious age-related changes, that can be measured through the biomarkers of aging. Biological molecules and processes are imperfect, leading to unwanted and disordering damaging consequences that produce different traits which are not selected during evolution (Gladyshev, 2013). In 2012, the same author, proposed a new fascinating metaphor, describing lifespan as the time needed for the water of a river to flow from the mountain to the ocean (Gladyshev, 2012). This time can be extended by building a dam for example, but this would not provide any explanation regarding the reason why the water continues to flow down the river, which is well known to be gravity. Similarly, any interventions that might regulate the aging process, such as lifespan extension, do not provide any answer related to what causes aging itself. In this context, the measures of the *deleteriome* are the best markers of aging. The origin of these aging damages has not been well defined yet, and we are still missing a satisfactory explanation about the reason why the cells are not able to clear them thanks to their protective machineries.

1.2. The aging hallmarks

As previously mentioned, in the last decades, a quasi-consensus opinion on the heterogeneity and the multifactorial aspect of aging has clearly emerged (Taffett, 2003). Figure 1 briefly summarizes how various authors integrated the different types of cumulative aging damage into (*i*) seven major damage types, (Matise, 2018), (*ii*) nine synthetic and twenty analytic hallmarks (Otin et al., 2013; Lemoine, 2021), (*iii*) and seven pillars (Kennedy et al., 2014). The seven major damage types suggested by the authors to be the main targets of anti-ageing research are extracellular aggregates, death-resistant cells, extracellular matrix stiffening, intracellular aggregates, mitochondrial mutations, cancerous cells, and cell loss, or tissue atrophy (Matise, 2018). The nine hallmarks include instability of the genome, telomere exhaustion, epigenetic changes, proteostasis loss, alteration of nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell shortage, and aberrated intercellular communication (Lemoine, 2021; López-Ótin et al., 2013). The seven pillars concept are represented by macromolecular damage, changes in the epigenome, chronic inflammation, loss of adaptation to stress conditions, proteostasis loss, stem cell shortage, and metabolism alteration (Kennedy et al., 2014).

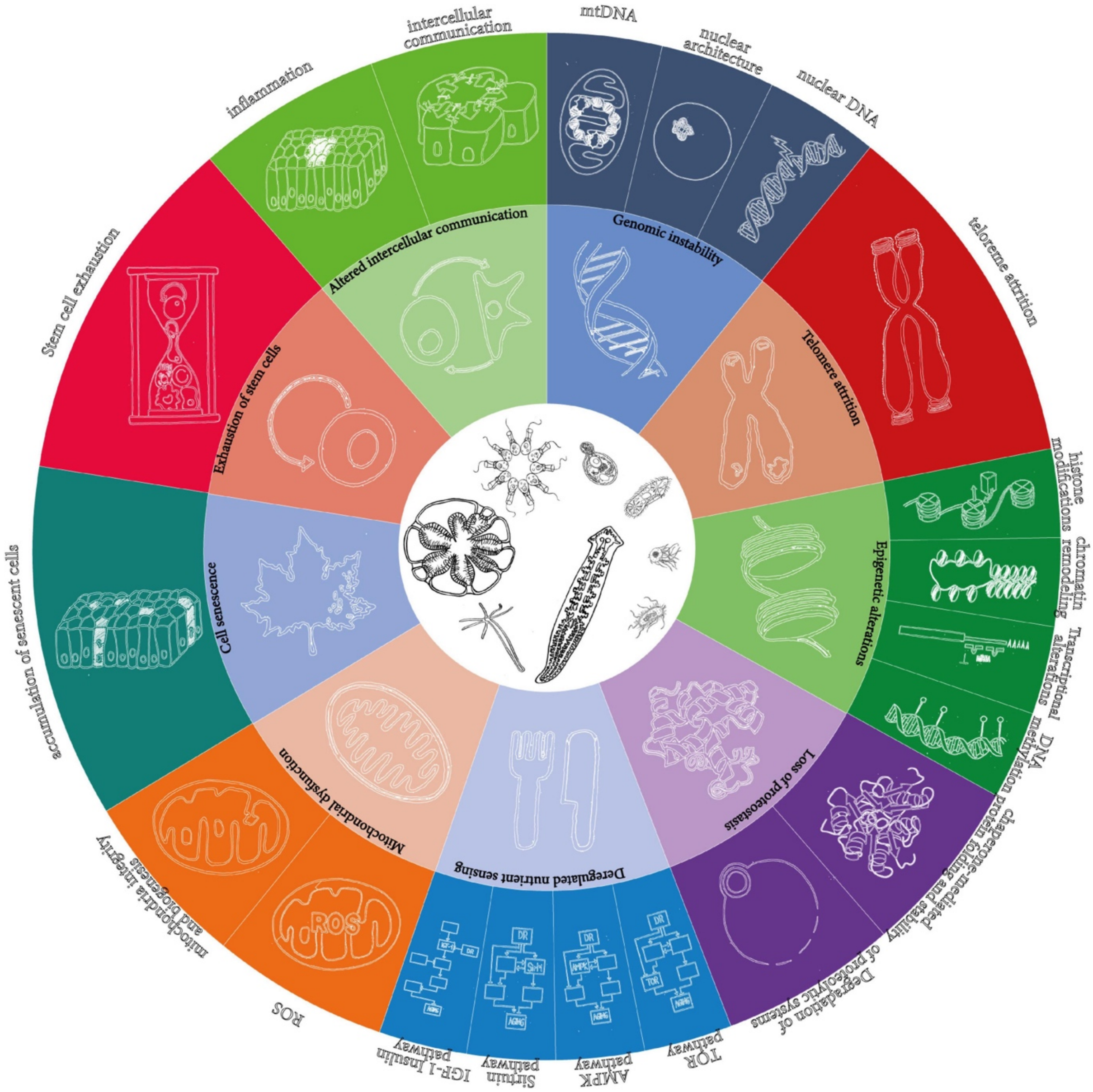


Figure 1: The hallmarks of aging
(**Source**: *Lemoine., 2021; López-Otín et al., 2013*).

Different forms of cellular damage are capable of causing variable effects on the organism's fitness. Natural selection can interact during evolution by removing those traits that present a more severe form of damage, in turn promoting the survival of those armed with more effective protective systems. On the contrary, if the damage exists in a lesser strength, this can rapidly accumulate during the life cycle, becoming impossible to compensate by the cellular protective systems. Moreover, an elevated number of protective mechanisms would not be sustainable in terms of fitness costs. For all these reasons, the cell is not able to repair the whole accumulation of damages it may encounters during its life, and therefore it must somehow select the most severe ones to be treated with specific mechanisms. In this regard, it is believed that cells are able to prevail over the overload of damage through its dilution at the time of cellular division, optimizing the balance between damage generation and dilution (Gladyshev, 2012).

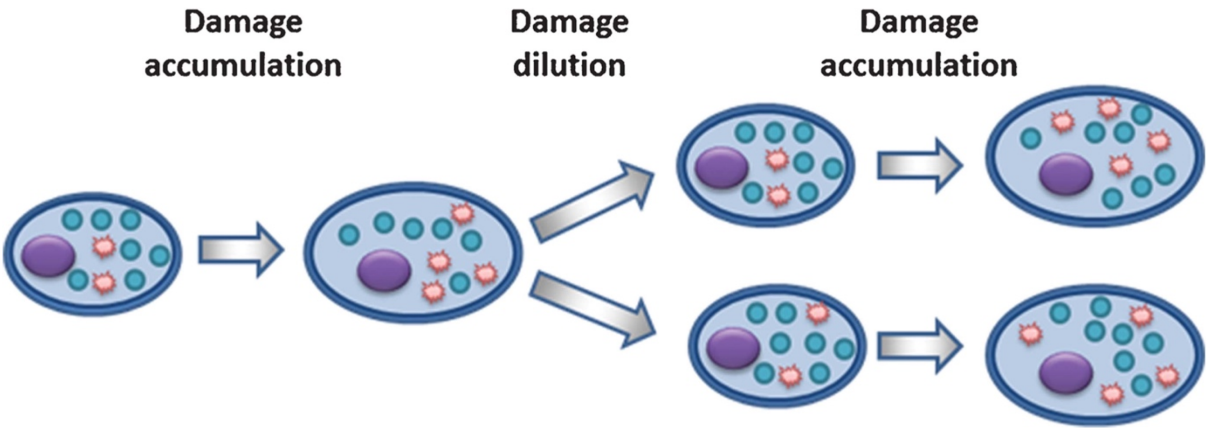


Figure 2: Damage dilution theory. Mild damage (red dots) is diluted during cell divisions progression (**Source**: *Gladyshev, 2012*).

According to this theory, damage dilution is the simplest life-sustaining biological strategy for the cell, which, in this way, no longer needs to maintain expensive protective systems. Many differentiated cells accumulate damage, albeit by dividing symmetrically. In fact, this can accumulate faster than it is diluted, causing clonal senescence and eventually death. In this context, mammalian cells possess control systems capable of activating apoptotic processes that limit the proliferation of one cell compared to another, thus limiting the potential development of cancer. These systems are capable of killing newly emerged cancer cells until they themselves fail, as a result of damages accumulation. Therefore, cancer can be seen as a condition that removes these check-point mechanisms, manifesting as mutations, which directly or indirectly disrupt apoptotic and related protective systems.

1.3. Damage accumulation and Cancer

Regarding the damage accumulation and its dilution, cancer is directly linked to aging: the damage accumulated during the aging process is removed by the protective systems, which in turn age and undergo mutations. Some cells can re-set their metabolism evading the cell cycle check-point systems, thus becoming cancerous. Cancer and aged cells represent the two different side of the same coin, characterized by shared or highly divergent pathways and molecular mechanisms (Table 2) (Aunan, et al., 2017).

	Aging	Cancer
Genomic Instability	Increased	Increased
Telomere Attrition	Shortened telomers	Shortened telomers, but telomerase activation
Epigenetic Alteration	• Global hypomethylation • Complex non-coding downregulation (miRNAs)	• Hyper- of tumor suppressors • Hypo- of oncogenes • Complex miRNA deregulation
Deregulated Nutrient Sensing	Inhibition of mTOR signaling that increases lifespan	Inhibition of mTOR signaling is antineoplastic
Cellular Senescence	Increased	• Prevalent in premalignant tumors • Evaded in fully malignant tumors
Proteostasis	Impairment of: • Chaperoning • Proteasome activity • Autophagy-lysosome activity	Augmented in: • Chaperoning • Proteasome activity • Autophagy-lysosome activity
Stem Cells	Exhausted	Potential nidus for tumorigenesis

Table 2: shared/divergent hallmarks between Aging and Cancer
(**Source**: adapted from Table1 - *Aunan et al., 2017*)

Mirroring those of aging, cancer hallmarks summarize the disease, and include proliferation, evasion of the suppression mechanisms, resistance to cell death, replicative immortality, angiogenesis, invasion, and metastasis (Hanahan and Weinberg, 2011). Additionally, genome instability and inflammation are placed at the base of all these processes (Hanahan and Weinberg, 2011). Specifically, genomic instability can lead to uncontrolled cell proliferation, giving rise to cancer phenomena due to deleterious mutations accumulation (Aunan et al., 2017). Indeed, during their life-span human cells undergo through billions of rounds of DNA replications, during which the risk of introducing mutational events is highly elevated. Being not particularly harmful, most of them can be repaired by the DNA repair complexes; however, during time, a certain level of DNA damage is still accumulated (Moskalev et al., 2013). Human cancers are usually characterized by genome instability and high mutational rates, while normal tissues control cell division through their life cycle. Cancer cells, by deregulating these pathways, sustain the proliferative signaling in alternative and independent ways. In this context, they can avoid those programs that arrest cell proliferation, such as the ones performed through suppressor genes. For example, TP53 gene in a stressed intracellular environment, is able to stop the progression of the cell cycle. Alternatively, if the damage is irreparable, it can trigger apoptosis through the activation of programmed cell death, thus acting as a natural barrier against the cancer development.

2. Intersection between Aging, DNA repair and Cancer

"Age is the single largest risk factor for an enormous number of diseases. So, if you can essentially postpone aging, then you can have beneficial effects on a whole wide range of disease."

Cynthia Kenyon

The existence of a direct connection between aging and the possible development of cancer has been already demonstrated (Frank, 2007). However, what is still unclear, is the complex mechanism by which the interaction of aging-related changes may be associated with cancer progression. Surely, cancer cells adopt a kind of invasive-selfish behavior, through which they modify the normal functioning of tissues (and sometimes organs), causing, in the worst case, the death of the individual itself (Nunney et al., 2015). Nevertheless, it is also true, that other mechanisms normally develop over the course of an organism's life that can inhibit the process of tumor development, whereas others may even hamper it (de Magalhães, 2013). As repeatedly emphasized throughout the introduction of this thesis, aging, represents the main risk factor for the development of many cancers. In fact, as time progressed, there is an implicit risk of accumulation of DNA mutations and damage. In the same way, the risk of developing errors in the newly synthesized DNA increases, as the number of cell divisions rises over the lifetime of the organism. Therefore, cells naturally become damaged and/or mutated during DNA replication, but many of these molecular errors can be repaired. For all these reasons, it is known that ageing is somehow closely related to the number of unresolved mutations due to the continuous proliferation of the cell cycle (Nunney et al., 2015). On the other hand, many dark sides regarding the intrinsic biological link between aging and tumorigenesis have not been revealed yet. To date, 70 years after its formulation, the multistage model of carcinogenesis proposed by Nordling in 1953, is still the most used to describe the increasing in cancer incidence rate in relation to the age of an individual (Jacobs et al., 2012; Nordling, 1953). According to this, a cell becomes malignant once it has accumulated a sufficient number of sequential mutations, which obviously increases with age (Anisimov, 2003). The species that we commonly identify as "cancer-resistant", have somehow developed a complex multistep trajectory, whereby it is more complicated for a somatic mutation to propagate its deleterious effects (Nunney, 2016). For these reasons, understanding which are the mechanisms preventing cancer development in different species could lay the foundation for a new therapeutic perspective in humans biomedical research (Caulin and Maley., 2011).

2.1. Peto’s paradox and Cancer

Although its evolutionary origin remains a mystery, many authors suggest that cancer should be as ancient as the appearance of multi-cellular organisms (Domazet-Lošo et al., 2014). Tumor incidence is not the same for all metazoans. For example, this value is low for elephants, blind and naked mole rats, while, on the contrary, it is extremely high for mice and dogs. In this context, the question that has been haunting researchers for decades is always the same: *how to explain this difference?* At the very beginning, the multistage process of tumorigenesis is triggered by a single oncogenic mutation that initiates the disease. From a theoretical point of view, since the probability of accumulating mutations increases during cell division progression, it is reasonable to argue that cancer should affect predominantly those organisms that are large, with a higher number of cells, and long-lived, with a greater number of cell divisions (Leroi, Koufopanou, and Burt, 2003). Indeed, somatic mutations that give rise to cancer events can appear whenever a cell enters its division cycle. In theory, we can assume that the risk of developing some form of tumor increases as a function of the number of cell divisions, where large, long-lived organisms have a greater chance of accumulating cancer-causing mutations (Caulin et al., 2015; Gaughran et al., 2016). However, throughout the evolution of multicellularity, longer-lived animals have been equipped with powerful tumor suppression mechanisms (Wolf, 2021). In fact, numerous studies have shown that some animals have solved the problem of cancer, by subjectively excelling at different molecular mechanisms of protection, being able to limit malignant tumor growth, whereas ensuring longer life spans and larger body size (Boddy et al., 2020; Seluanov et al., 2018; Tollis et al., 2017). This phenomenon is referred to Peto's Paradox, a biological enigma raised for the first time by the English statistician and epidemiologist Richard Peto in 1975 (Peto., 1975). According to this evolutionary conundrum, species that are very large and long-living have evolved additional suppression mechanisms that make them less prone to cancer development compared to their close relatives (Caulin and Maley, 2011). Recently, different mathematical models have been developed, showing that animals such as elephants and whales could not survive if they carried the equivalent risk of cancer by cell division as humans do (Caulin and Maley, 2011; Gaughran et al., 2016).

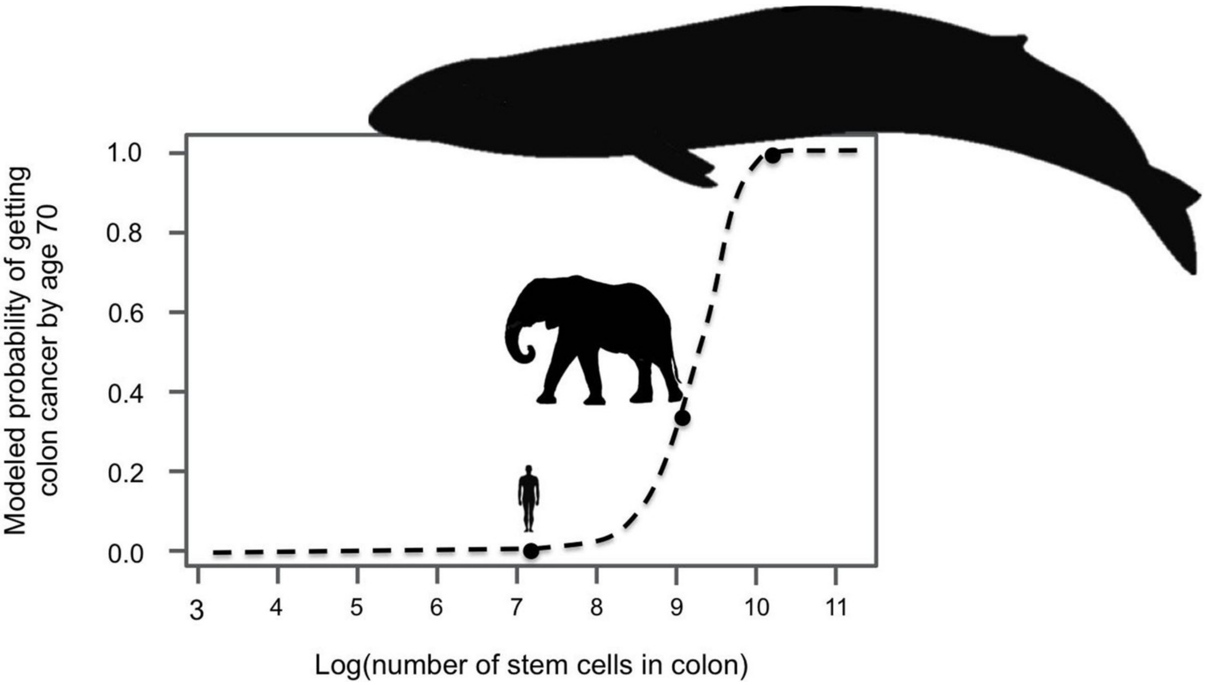


Figure 3: Cancer prediction model for large-bodied animals.
(**Source**: *Gaughran et al., 2016* – adapted from *Caulin and Maley., 2011*).

Obviously (and fortunately), elephants, whales, and other giant animals continue to populate our planet. Moreover, there is not a single evidence highlighting higher cancer rates for these species, confirming, instead, the hypothesis of the *super-human* cancer suppression development, which acted on big size and high longevity during the evolution of the anti-tumoral defenses. Furthermore, because size is one of the phenotypic traits that has evolved independently many times among the tree of life (Keane et al., 2015), it is reasonable to hypothesize that different species use multiple strategies to resolve this paradox, likely due to family subjective evolutionary pressures and trajectories (Callier, 2019). Parallelly, along the tree-of life, the so called *'trade-off'* evolutionary phenomena, can compensate the presence of tumor suppression mechanisms against the other survival processes, such as the reproductive success (Boddy et al., 2015). Indeed, the defense machineries, DNA repair or cell cycle control for example, can be very costly in terms of fitness, and therefore they must be balanced in order to ensure the evolution of the species (Tollis et al, 2017). Nowadays, different researchers are benefitting from the genome of those animals that represent size and longevity outlier, in order to study and unravel the knot hiding this biological puzzle. However, a single answer does not exist. At least, not for the moment I would add (Vischioni, this thesis, 2022). What is certain is that different longevity pathways are conserved in multiple eukaryotic species, and that, in this context, comparative studies are the ideal target for the identification of the genetic mechanisms underlying longevity and cancer suppression.

3. Alternative animal models for cancer research

Turning away from a single model organism to a set of different species, could be the key to understanding and developing a new knowledge about unknown molecular mechanisms. Furthermore, limiting the investigation to only a few model species, undermines the efforts to gain a broad and general understanding of the mechanisms behind the biological processes that determine aging. *Homo sapiens*, per se, is a long living species. Therefore, short living models are not sufficient to study the entire dynamics of its ageing processes. Indeed, why an organism that live for just three or four years, such as the mouse or the rat, should have developed additional resources to sustain up to the age of 80? None. Recently, researcher's attention has shifted towards the study of those animals that are beyond the usual standard models in biomedicine, and that can equal, if not exceed, the human longevity rate. From an evolutionary perspective, during chronological aging, a high mortality risk associated with a lower reproductive success represents a general decrease in organisms' fitness (Reichard, 2017). Mortality and fitness loss are generally found throughout the entire animal kingdom, but it seems that some species can age more slowly than others. In this context, the term "senescence" denotes a generic definition used to describe the physiological deterioration which led towards a mortality increase and/or a fertility decline with age. This phenomenon is often considered inevitable, and if this is true, all mortality rates among species should increase with age, correlating with the physiological deterioration itself. However, since the 1990s, researchers have identified some species whose mortality rates appear to be against this tendency (Finch, 1994), exhibiting a level of senescence which was later called "negligible". Shortly, there are some species which that do not show increased mortality rates with age, such as lobsters (*Homarus* spp.), the quahog bivalve, the *Sebastes* spp. fish, and Testudinidae (turtles) for example (Finch 1994). Generally, studying organisms with longer lifespans could help to better understand the mechanisms underlying their successful aging and life expectancy optimization, while minimizing the physical deterioration, potentially paving the way for new therapeutic interventions in age-related diseases, or providing insights into successful strategies for healthy aging (Reichard, 2016). The basic idea is that the acquisition of larger bodies and higher lifespans, have themselves favored the need to develop powerful mechanisms of tumor suppression. Parallelly, in this optics, Evolution would have been able to balance the costs and the benefits of such defensive machineries, maximizing the reproductive success of the species (DeGregori, 2011). From one side, this evolutionary "trade-off" increases the cancer-vulnerability in long living species, whereas, from the other, it pushes to the development of compensatory strategies against the disease onset (Aktipis et al., 2013; Nunney et al., 2015). To date, animals such as the mouse (*Mus musculus*) and the rat (*Rattus norvegicus*), have been the most widely used conventional models in the field of aging research. However, presenting a lifespan of only few months, they are not always able to recapitulate the wide range of variability across the whole animal kingdom (Holtze et al., 2021). Indeed, until now, mice are primarily used to confirm targets identified in large screenings involving invertebrate models. For example, invertebrates such as the yeast *Saccharomyces cerevisiae* (**Chapter IV**), have been instrumental in the discovery of aging-related genes and pathways, that are conserved throughout the eukaryotic domain (Guarente and Kenyon, 2000). In recent decades, many laboratories have specialized in the study of unconventional model organisms, which by their phenotypic and genotypic characteristics differ from traditional ones, and which have specific peculiarities, often displaing unusual biological features, as illustrated in Figure 4 (Shepard and Kissil., 2020).

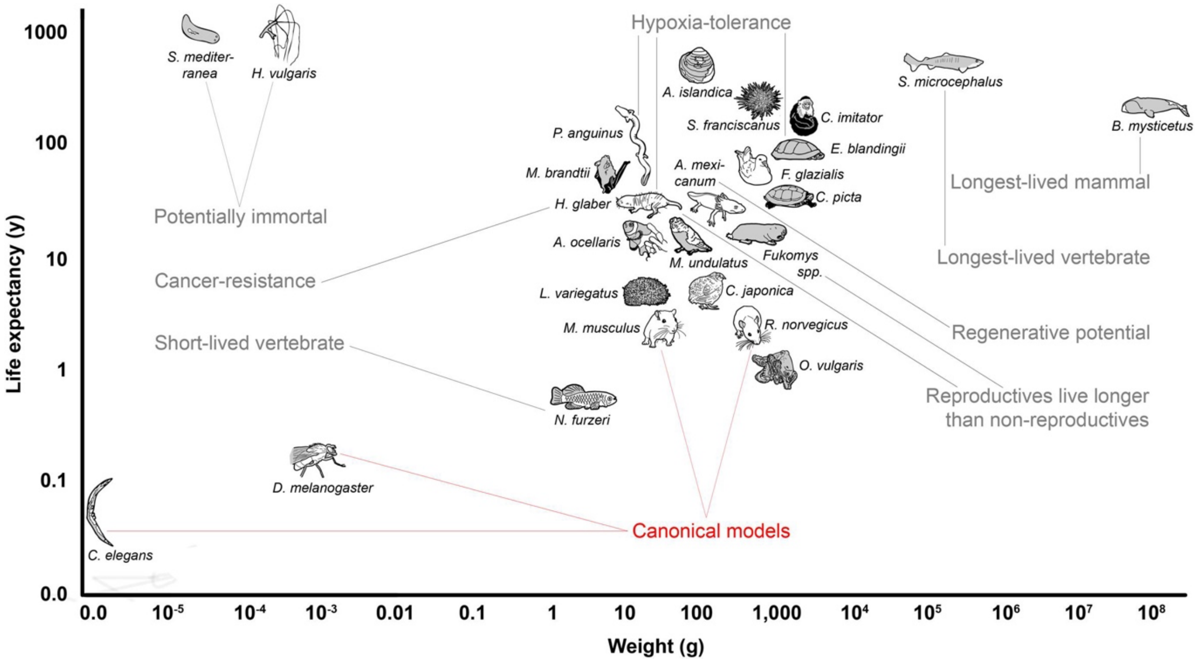


Figure 4: Canonical and Alternative animal models of aging and comparative research (**Source**: *Holtze et al., 2021*).

All these species carrying exceptional life expectancy could help discovering strategies, pathways, and/or gene variants that, although possibly absent in the human genome, could shed light on how to delay aging, while inhibiting related diseases (Brunet, 2020). Not only that. Precisely, because of their uniqueness, non-canonical organisms offer a whole new pull of species-specific cancer resistance molecular targets to be further explored *ex-novo* (Holtze at al., 2021). Indeed, according to Cohen and co-authors (2020a), each aging mechanism is often typical for a single organism, whereas the upstream regulatory pathways would be conserved across species (Cohen et al., 2020b). Indeed, recent analyses based on whole-genome sequencing, transcriptomics, omics, and metabolomics suggested the presence of anti-aging mechanisms that might contribute to their extreme longevity, while failing the functional and physiological decline (Ma & Gladyshev 2017; Tian et al. 2017), allowing them to eventually become senescent, but to age much slower compared to humans, for example (de Magalhães., 2015). Finally, a comparative analysis of the so-called longevity outliers will allow the generation of new transversal information applicable to all the biological systems, having great impact on understanding the unique biological proprieties of multiple cancer resistance mechanisms (Brunet, 2020).

3.1. *Loxodonta africana*, elephant

The African savannah elephant (*Loxodonta africana*) is one of the representatives of the cancer-resistant species category. In 2015, researchers began to unravel the secret underlying how this organism is able to overcome Peto's paradox, maintaining such a large size, but simultaneously, living so long (Abegleen et al., 2015). Indeed, these giants organisms show an high resistance to cancer thanks to the amplification of the number of copies of TP53 gene (Sulak et al., 2016), which is also associated with the alteration of other tumor suppressors copy number (Vazquez and Lynch, 2021). TP53 gene functions as *"guardian"* of the cellular cycle, detecting the molecular damages, and triggering cell death. Therefore, thanks to the extra TP53s, the elephant genome appears more stable, and able to get rid of the deleterious mutations that could possibly lead to irreparable DNA damage, thus promoting tumor development. While in the human genome TP53 is present in singular copy (as well as in many other mammals), the one of the elephant encodes 20 copies of the same one, including pseudogenes. Having a higher number of copies of TP53 ensures greater efficiency in activating apoptotic pathway during uncontrolled cellular proliferation. Moreover, in more recent years, Lynch and collaborators have also discovered that the genome of *Loxodonta africana* codes for 11 copies of a gene called LIF. One of them, LIF6, if over-expressed and activated by the cellular damage, can act as another powerful apoptotic factor, even in the absence of TP53 itself (Vazquez et al., 2018).

3.2. *Heterocephalus glaber*, naked mole rat

Although several organisms are intriguing for many different reasons, the *Heterocephalus glaber* has recently attracted scientist's research interests in the field of cancer comparative genomics. Despite its small size, the naked mole rat (NMR), with a lifespan of more than 30 years, is, to date, the longest-living member of the rodent family. Nonetheless, being about the same size as a mouse, it suffers from a very low level of cancer incidence (Buffenstein, 2005; Lewis et al., 2016; Miyawaki et al., 2016; Seluanov et al., 2018; Shepard and Kissil, 2020; Tian et al., 2013), especially compared to the 90% rate found in the common rat (Lipman et al., 2004). Indeed, numerous studies have shown that this organism is able to resist to a whole range of age-related diseases such as, for example, sarcopenia (O' Connor et al., 2002; Stoll et al., 2016), and neurodegeneration (Edrey et al., 2013). Just a few months ago, Rochelle Buffenstein, one of the pioneers of cancer genomics and ageing research, relaunched the paradigm according to which NMR is, and will be, a fundamental key character in order to understand how evolution has adapted the biology of the species to survive in extreme and adverse conditions (Buffenstein et al., 2021). The first report showing unusual values of cancer mortality for this animal dates to 2008 (Buffenstein, 2008). Since this time, many other studies have been performed, which seem to reflect and confirm a common low susceptibility to the cancer insurgence. In this context, in 2009, Seluanov and co-authors were able to demonstrate the existence of a particular cellular phenomenon called early contact inhibition (Seluanov et al., 2009). Few years later, this hypothesis was confirmed through the investigation of the extracellular matrix environment, that, in the naked mole rat, possess peculiar characteristics able to prevent the onset of cancer. Indeed, the cellular division of naked mole rat cells would stop as soon as the environment becomes too crowded, mediated by ultra-high molecular mass hyaluronan (Tian et al., 2015, 2013). In other words, regulatory feedbacks can order the cell to abort division cycles (Tian, 2016; Rankin and Frankel, 2016) ensuring the reestablishment of the equilibrium condition. However, the contact inhibition phenomenon is not the only mechanism involved into its resistance towards tumorigenesis. To give a general overview, Figure 5 summarizes all the processes by which the naked mole rat can delay aging and prevent cancer (Shepard and Kissil., 2020). Among the most significant ones, we retrieve the overexpression of alpha-2 macroglobulin (Thieme et al., 2015), a more efficient repair system of the DNA damage (MacRae et al., 2015), and efficient pathways of apoptosis and autophagy against the damaged cells (Evdokimov et al., 2018).

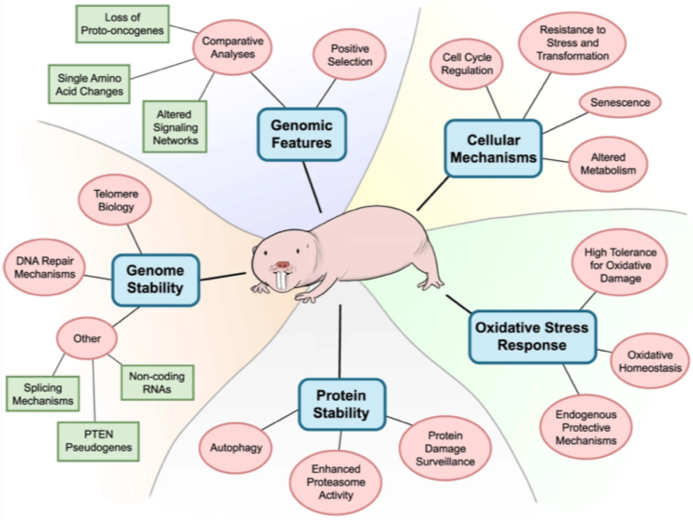


Figure 5: NMR's cancer resistance mechanisms overview (**Source**: *Shepard and Kissil., 2021*).

3.3. *Spalax spp.*, blind mole rat

The super-species *Spalax ehrenbergi* possess anti-aging properties that render it particularly attractive in cancer research (Lagunas-Rangel., 2018). Even with its high longevity (~20 years), in more than 40 years of investigations, there have been identify approximately zero cases of neoplasia, or any other phenotypic changes associated with aging for this animal (Fang et al., 2014; Manov et al., 2013;). Despite living in a stressful environment, the blind mole rat has developed a surprising resistance to cancer, allowing it to overcome the damages caused by the oxidative stress and the genomic instability due to the constant hypoxia and the rapid re-oxygenation coming from its natural habitat (Manov et al., 2013). Moreover, analyzing the transcriptomes of Spalax, compared to the one of the others shorter-living rodents, some authors found a specific blind mole rat signature of over-expressed genes involved in DNA repair, metabolism, and recombination pathways, and cell cycle mechanisms (Malik et al., 2016). In the previous section, I described how *Heterocephalus glaber* resolved tumor initiation through the "early contact inhibition" strategy (Seluanov et al., 2009). In contrast, Spalax, adopts another type of process that was described in 2012 as "concerted cell death" (Gorbunova et al., 2012). This represents a specific mechanism mediated by a combination of necrotic and apoptotic processes, involving also p53 protein among others. Specifically, genomic analyses revealed that duplications of genes involved in the interferon signaling pathway such as IFNβ1, would be involved in the regulation of cell death and inflammation, modulating necrosis and inflammatory responses (Fang et al., 2014). Thus, Spalax is protected from the persistence of damaged cells, which could eventually give rise to carcinogenic events. Furthermore, the blind mole rat possesses specific amino-acid changes in the DNA binding domain of p53 protein. Normally, p53 monitor cellular adaptations against a variety of stress conditions, including DNA damage and hypoxia, resulting in cell cycle arrest and apoptosis. However, this substitution creates a bias against apoptosis, for which the genes associated to the promotion of the cell cycle arrest are favored compared to the ones related to the apoptotic targets (Ashur-Fabian et al., 2004; Avivi et al., 2007).

3.4. *Balena mysticetus*, bowhead whale (and other cetacea)

Despite the extreme conditions of food deprivation and cold temperatures, Arctic and Antarctic are the natural habitat of some of the longest-living species on Earth. In such a challenging environment characterized by a predominance of frost and ice, high salinities, and limited resources, some animals have evolved physiologically and behaviorally to support the species survival and reproduction (Blix, 2016). Generally, these are primarily studied in order to understand their adaptations to temperature, diet, and metabolism, which, in turn, are closely related factors to the extension of life expectancy (Keane et al., 2015). For example, whales are among the longest-living mammals on Earth, which exhibit some age-associated physiological aspects together with an extremely low cancer incidence rate (Lagunas-Rangel, 2021). Species such as the bowhead whale (BWH) (*Balaena mysticetus*), with a record of 211 years (George et al., 1999; George and Bockstoce, 2008), represents the longest-lived mammal on our planet. Actually, many members of the *Mysticeti* cetacean order have a lifespan exceeding 100 years, fully satisfying and respecting Peto's paradox criteria. Their high life expectancy, coupled with their very low tumorigenic events, make them an ideal study model for human age-related diseases, such as cancer. Even if the genome and the transcriptome of some whale species have already been sequenced and are currently publicly available, the molecular mechanisms underlying their longevity and resistance to age-related diseases have not yet been fully elucidated. According to Omotoso and co-author (2021), such a low risk of neoplasia development would be the result of the combination between the evolution of large body size and very efficient genes suppression networks, which, interacting together in a balanced harmony, provide higher protection against cancer (Caulin et al., 2015; Caulin and Maley, 2011; Nunney et al., 2015; Wensink, 2016). Remarkably, cetacean genomes present multiple duplication events, occurring in those genes related to cell cycle control and cancer protection. Surprisingly, comparative analyses found that the bowhead whale genome does not exhibit TP53 gene duplication, as the elephant does. However, in 2021, Tejada-Martinez demonstrated that the "large body" phenotype significantly correlates with the number of different tumor suppression pathways, involving genes such as EEF1A1, H2AFX, HSPD1, MAPK9, GSTP1, PTPN11, which are known to be key players in aging as well as in senescence processes (Tejada-Martinez et al., 2021). This founding would confirm the hypothesis that the duplication in suppressor genes among the cetacean clade underwent through positive selection during the course of evolution. Moreover, sequencing of the BHW genome and comparative analysis revealed several SNPs and amino acids substitutions in those genes associated to ecological adaptation and to the disease resistance, such as PCNA, LAMTOR1, PSMD4, UCHL3, ARPP19, STOML2, HSBP1, DLD, SMS, and ST13 (Keane et al., 2015). Additionally, the longevity associated EIF2 and PABP genes have been found significantly expensed in her genome (Doherty and de Magalhães, 2016). All together, these insights provide evidence that the evolution of cancer resistance and longevity in these species have been modeled under the molecular adaptation of DNA repair genes, and other genes related to DNA replication, cell cycle, cellular damage and survival, as shown in Figure 6.

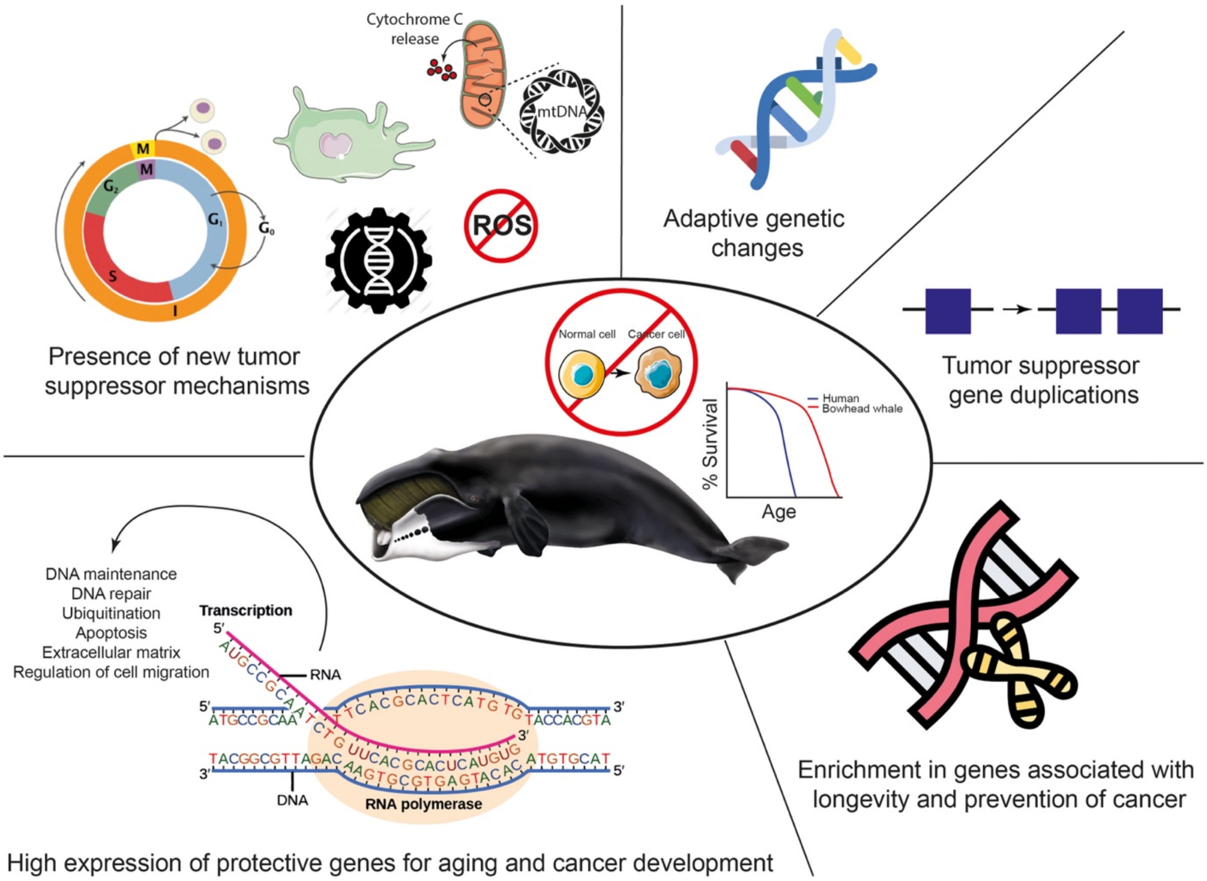

Cytochrome C release
mtDNA
M
M
G2
G1
G0
S
I
ROS
Presence of new tumor suppressor mechanisms
Adaptive genetic changes
Normal cell
Cancer cell
Human
Bowhead whale
% Survival
Age
Tumor suppressor gene duplications
DNA maintenance
DNA repair
Ubiquitination
Apoptosis
Extracellular matrix
Regulation of cell migration
Transcription
RNA
DNA
RNA polymerase
Enrichment in genes associated with longevity and prevention of cancer
High expression of protective genes for aging and cancer development

Figure 6: BHW's cancer resistance and longevity mechanisms overview (**Source**: *Lagunas-Rangel., 2021*).

Finally, the LINEs transposable elements family appears to be highly active within the bowhead whale genome (Keane et al., 2015; Tollis et al., 2019). The percentage of active TEs, which in BWH is around 30%, normally increases as a function of the organism age, becoming potentially highly mutagenic in older individuals (Anwar et al., 2017; Bravo et al., 2020). However, those from cetaceans appear to possess an extremely slow mutation rate, that would prevent the cancer cells progression (Tollis et al., 2019).

4. Copy Number Variations (CNVs)

Peto's paradox evidence implies that the evolution of species with high cancer resistance rates should overlap and coexist directly with the maintenance of large body sizes and long-lived organisms. However, scientists still question the genetic mechanisms underlying the interspecies diversity of lifespan and cancer rates. In this perspective, comparative genomics and new sequencing studies may help in understanding the mechanisms responsible for the extreme longevity of some animals. Currently, a unanimous solution that could totally explain the Peto’s enigma does not exist yet. Among the multiple hypotheses, the increased copy number of tumor suppressor (TS) genes as protective defense against cancer is one of the most widely acclaimed. An alteration of the CNV landscape of TS could, in fact, limit the uncontrolled cell proliferation in the presence of molecular damage (Domazet-Lošo and Tautz, 2010). In the last 10-15 years, there have been many efforts in sequencing the genome of an increasing number of long-lived species, in order to understand whether there is a common mechanism of positive selection regarding the conservation of suppressor genes among different species, and clarify which targets are involved in maintaining the paradox (Árnason et al., 2018; Fang et al., 2014; Gorbunova et al., 2014; Howe et al., 2021; Keane et al., 2015; Kim et al., 2011; Lewis et al., 2016; Seim et al., 2013; Zepeda Mendoza et al., 2018). Indeed, since mutations in suppressor and oncogenic genes (OG) are directly related to cancer (Stratton et al., 2009), many researchers have begun to investigate individual cases of gene duplication involved in disease development (Abegglen et al., 2015; Sulak et al., 2016; Vazquez et al., 2018), believing that this phenomenon is a key player in addressing cancer resistance, especially for those species with a high theoretical risk of neoplasia. The first discovery in this field, which can be seen as one of the founders of this research line, was undoubtedly the amplification of TP53 in the elephant genome proposed in 2015 and 2016 (Abegglen et al., 2015, Sulak et al., 2016), and subsequently confirmed in 2018 (Vazquez et al., 2018). It is already well known that many ageing-related pathways have evolved thanks to the duplication of certain genes (Ritter et al., 2013), and that even complex human traits, such as disease susceptibility or drug response (Gamazon and Stranger., 2015), depend on CNVs of specific markers. Many authors, nowadays, argue that studying CNVs in a comparative perspective is the key to understand the maintenance of particular traits, where gene duplication itself is one of the main actors of the evolution of the “long living” and the “cancer resistant” phenotypes (Caulin et al., 2015; Doherty and de Magalhães, 2016).

According to Caulin and Maley (2011), multiple copies of tumor suppressor genes and a reduced presence of proto-oncogenes are two among the possible reasons why organisms with high longevity rates can outcompete cancer onset, despite their size and lifespan (Caulin and Maley, 2011).
Specifically:

1) to become cancerous, a cell needs to develop an oncogenic mutation. Reducing the number of proto-oncogenes, therefore, would also decrease the probability of triggering a tumorigenic event (Davenport et al., 2002). At the same time, however, each oncogene has his own function and the lack of this genomic elements would be deleterious in term of fitness and survival;
2) a higher number of suppressor genes could prevent the onset of cancer, since more mutagenic events would be needed to produce a malignant phenotype (Nunney, 1999).

By definition, copy number variation (CNV) represents a biological event in which sequences in a genome are duplicated, creating an important source of genetic variation. They are often involved in diseases and important molecular pathways such as cancer, metabolic and neurodegenerative disorders. Therefore, elucidating the role of genes driving copy number alteration, and investigating how this may protect or promote cancer is an important new promising field of research (Zack et al., 2013). In this context, the exploration of cancer development, its maintenance processes, and the mechanisms that prevent its occurrence in different species offers enlightened potential knowledge to be translated into the field of biomedical research in order to uncover the mechanisms of cancer susceptibility for humans, and possibly propose new strategies for studying and developing treatments for the disease (Nunney et al., 2015).

In the following paragraphs, I will provide a brief summary of each chapter of my thesis. At the beginning of every chapter, I have included an introductory description of the ideas, concepts, and main findings contained within. For the study that has been published (**Chapter II**), I have included the article in the format of the journal.

There are two limitations affecting this kind of scientific investigations, which led me to the two main findings of this thesis. First, to date, there are few available sources of data collecting the true prevalence of cancer incidence in different mammals (Vittecoq et al., 2013). For this reason, I developed *VarNuCopy*, a user-frendly open source database that collects CNVs data from the genome of 233 organisms, allowing researchers, not only to obtain the CNVs landscape for each animal, but to correlate it with some phenotypic traits such as cancer rate and lifespan, by means of statistical models (**Chapter II**).

Secondly, many of these studies have focused on genes that are chosen *a priori* based on a sub-selection of those already known to be involved in tumor development, maintenance, and progression. In particular, I analyzed the entire genes copy number landscape of 9 representative species, discovering that some of the most important human onco-miRNAs and miRNAs suppressor are able to discriminate cancer prone and cancer resistant organism categories (**Chapter III**). This highlights the potential of microRNAs as tumor/anti-tumors regulators, not only strengthening their value as future anticancer molecular targets, but also paving the way of a new trajectory in trying to solve the puzzling Peto's Paradox dilemma.

In the framework of comparative genomics and the study of alternative organisms in cancer and ageing research, **Chapter IV** underlies the potentiality of using the yeast *S. cerevisiae* as a model system in these fields. In particular, I highlight the general biology and the main features of the species, coupling it with the description of its two ageing paradigms, which were also the topics and the aim of the project I carried out during the Ph.D period spent abroad.

Finally, in my concluding remarks (**Chapter V** and **Chapter VI**), I briefly return to the main concepts listed and discussed throughout the entire manuscript, highlighting the current limitations of cancer comparative genomics research, with an eye on the possible future directions and developments.

References:

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., ... & Schiffman, J. D. (2015). Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. Jama, 314(17), 1850-1860.

Aktipis, C., Boddy, A. M., Gatenby, R. A., Brown, J. S., & Maley, C. C. (2013). Life history trade-offs in cancer evolution. Nature Reviews Cancer, 13(12), 883-892.

Anisimov, V. N. (2003). The relationship between aging and carcinogenesis: a critical appraisal. Critical reviews in oncology/hematology, 45(3), 277-304.

Anwar, S. L., Wulaningsih, W., & Lehmann, U. (2017). Transposable elements in human cancer: causes and consequences of deregulation. International journal of molecular sciences, 18(5), 974.

Árnason, Ú., Lammers, F., Kumar, V., Nilsson, M. A., & Janke, A. (2018). Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. Science advances, 4(4), eaap9873.

Ashur-Fabian, O., Avivi, A., Trakhtenbrot, L., Adamsky, K., Cohen, M., Kajakaro, G., ... & Rechavi, G. (2004). Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. Proceedings of the National Academy of Sciences, 101(33), 12236-12241.

Aunan, J.R., Cho, W.C., Søreide, K., 2017. The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. Aging Dis. 8, 628. https://doi.org/10.14336/AD.2017.0103.

Avivi, A., Ashur-Fabian, O., Joel, A., Trakhtenbrot, L., Adamsky, K., Goldstein, I., Amariglio, N., Rechavi, G., Nevo, E., 2007. P53 in blind subterranean mole rats – loss-of-function versus gain-of-function activities on newly cloned Spalax target genes. Oncogene 26, 2507–2512. https://doi.org/10.1038/sj.onc.1210045.

Aunan, J. R., Cho, W. C., & Soreide, K. (2017). The biology of aging and cancer: a brief overview of shared and divergent molecular hallmarks. Aging Dis 8: 628–642.

Blagosklonny, M. V. (2008). Aging: Ros or tor. Cell cycle, 7(21), 3344-3354.

Blix, A. S. (2016). Adaptations to polar life in mammals and birds. Journal of Experimental Biology, 219(8), 1093-1105.

Boddy, A. M., Abegglen, L. M., Pessier, A. P., Aktipis, A., Schiffman, J. D., Maley, C. C., & Witte, C. (2020). Lifetime cancer prevalence and life history traits in mammals. Evolution, medicine, and public health, 2020(1), 187-195.

Bravo, J. I., Nozownik, S., Danthi, P. S., & Benayoun, B. A. (2020). Transposable elements, circular RNAs and mitochondrial transcription in age-related genomic regulation. Development, 147(11), dev175786.

Brunet, A. (2020). Old and new models for the study of human ageing. Nature Reviews Molecular Cell Biology, 21(9), 491-493.

Buffenstein, R. (2008). Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. Journal of Comparative Physiology B, 178(4), 439-445.

Buffenstein, R. (2005). The naked mole-rat: a new long-living model for human aging research. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 60(11), 1369-1377.

Buffenstein, R., Amoroso, V., Andziak, B., Avdieiev, S., Azpurua, J., Barker, A. J., ... & Smith, E. S. J. (2022). The naked truth: a comprehensive clarification and classification of current 'myths' in naked mole-rat biology. Biological Reviews, 97(1), 115-140.

Callier, V. (2019). Core concept: solving Peto's paradox to better understand cancer. Proceedings of the National Academy of Sciences, 116(6), 1825-1828.

Caulin, A. F., Graham, T. A., Wang, L. S., & Maley, C. C. (2015). Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1673), 20140222.

Caulin, A. F., & Maley, C. C. (2011). Peto's Paradox: evolution's prescription for cancer prevention. Trends in ecology & evolution, 26(4), 175-182.

Cohen, A. A., Kennedy, B. K., Anglas, U., Bronikowski, A. M., Deelen, J., Dufour, F., ... & Fülöp, T. (2020). Lack of consensus on an aging biology paradigm? A global survey reveals an agreement to disagree, and the need for an interdisciplinary framework. Mechanisms of ageing and development, 191, 111316.

Cohen, A. A., Legault, V., & Fülöp, T. (2020). What if there's no such thing as "aging"?. Mechanisms of Ageing and Development, 192, 111344.

Davenport, M. P., Ward, R. L., & Hawkins, N. J. (2002). The null oncogene hypothesis and protection from cancer. Journal of medical genetics, 39(1), 12-14.

De Magalhães, J. P. (2013). How ageing processes influence cancer. Nature Reviews Cancer, 13(5), 357-365.

de Magalhães, J. P. (2015). The big, the bad and the ugly: extreme animals as inspiration for biomedical research. EMBO reports, 16(7), 771-776.

DeGregori, J. (2011). Evolved tumor suppression: why are we so good at not getting cancer?. Cancer research, 71(11), 3739-3744..

Doherty, A., & de Magalhães, J. P. (2016). Has gene duplication impacted the evolution of Eutherian longevity?. Aging Cell, 15(5), 978-980.

Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature, 468(7325), 815-818.

Edrey, Y. H., Medina, D. X., Gaczynska, M., Osmulski, P. A., Oddo, S., Caccamo, A., & Buffenstein, R. (2013). Amyloid beta and the longest-lived rodent: the naked mole-rat as a model for natural protection from Alzheimer's disease. Neurobiology of aging, 34(10), 2352-2360.

Evdokimov, A., Kutuzov, M., Petruseva, I., Lukjanchikova, N., Kashina, E., Kolova, E., ... & Lavrik, O. (2018). Naked mole rat cells display more efficient excision repair than mouse cells. Aging (Albany NY), 10(6), 1454.

Fang, X., Nevo, E., Han, L., Levanon, E. Y., Zhao, J., Avivi, A., ... & Wang, J. (2014). Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. Nature communications, 5(1), 1-11.

Fariselli, P., Taccioli, C., Pagani, L., & Maritan, A. (2021). DNA sequence symmetries from randomness: the origin of the Chargaff’s second parity rule. Briefings in bioinformatics, 22(2), 2172-2181.

Finch, C. E. (1994). Longevity, senescence, and the genome. University of Chicago Press.

Gamazon, E. R., & Stranger, B. E. (2015). The impact of human copy number variation on gene expression. Briefings in functional genomics, 14(5), 352-357..

Gaughran, S. J., Pless, E., & Stearns, S. C. (2016). Evolutionary biology: how elephants beat cancer. Elife, 5, e21864.

George, J. C., Bada, J., Zeh, J., Scott, L., Brown, S. E., O'Hara, T., & Suydam, R. (1999). Age and growth estimates of bowhead whales (Balaena mysticetus) via aspartic acid racemization. Canadian Journal of Zoology, 77(4), 571-580.

Bockstoce, J. R. (2008). Two historical weapon fragments as an aid to estimating the longevity and movements of bowhead whales. Polar Biology, 31(6), 751-754.

Gladyshev, V. N. (2016). Aging: progressive decline in fitness due to the rising deleteriome adjusted by genetic, environmental, and stochastic processes. Aging cell, 15(4), 594-602.

Gladyshev, V. N. (2013). The origin of aging: imperfectness-driven non-random damage defines the aging process and control of lifespan. Trends in genetics, 29(9), 506-512.

Gladyshev, V. N. (2012). On the cause of aging and control of lifespan: heterogeneity leads to inevitable damage accumulation, causing aging; control of damage composition and rate of accumulation define lifespan. Bioessays, 34(11), 925-929.

Gorbunova, V., Hine, C., Tian, X., Ablaeva, J., Gudkov, A. V., Nevo, E., & Seluanov, A. (2012). Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism. Proceedings of the National Academy of Sciences, 109(47), 19392-19396.

Gorbunova, V., Seluanov, A., Zhang, Z., Gladyshev, V. N., & Vijg, J. (2014). Comparative genetics of longevity and cancer: insights from long-lived rodents. Nature Reviews Genetics, 15(8), 531-540.

Guarente, L., & Kenyon, C. (2000). Genetic pathways that regulate ageing in model organisms. Nature, 408(6809), 255-262.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

Harraan, D. (1955). Aging: a theory based on free radical and radiation chemistry.

Hayflick, L. (2000). The illusion of cell immortality. British journal of cancer, 83(7), 841-846.

Hayflick, L. (2007). Entropy explains aging, genetic determinism explains longevity, and undefined terminology explains misunderstanding both. PLoS genetics, 3(12), e220.

Holtze, S., Gorshkova, E., Braude, S., Cellerino, A., Dammann, P., Hildebrandt, T. B., ... & Sahm, A. (2021). Alternative animal models of aging research. Frontiers in Molecular Biosciences, 8, 311.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., ... & Flicek, P. (2021). Ensembl 2021. Nucleic acids research, 49(D1), D884-D891.

Jacobs, K. B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., ... & Stolzenberg-Solomon, R. Z. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nature genetics, 44(6), 651-658.

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., ... & de Magalhães, J. P. (2015). Insights into the evolution of longevity from the bowhead whale genome. Cell reports, 10(1), 112-122.

Kennedy, B. K., Berger, S. L., Brunet, A., Campisi, J., Cuervo, A. M., Epel, E. S., ... & Sierra, F. (2014). Geroscience: linking aging to chronic disease. Cell, 159(4), 709-713.

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., ... & Gladyshev, V. N. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature, 479(7372), 223-227.

Kirkwood, T. B. (1977). Evolution of ageing. Nature, 270(5635), 301-304.

Kirkwood, T. B. (2017). Why and how are we living longer?. Experimental physiology, 102(9), 1067-1074.

Lagunas-Rangel, F. A. (2018). Cancer-free aging: insights from Spalax ehrenbergi superspecies. Ageing Research Reviews, 47, 18-23.

Lagunas-Rangel, F. A. (2021). Deciphering the whale's secrets to have a long life. Experimental Gerontology, 151, 111425.

Lambert, F. L. (2007). Entropy and the second law of thermodynamics.

Lemoine, M. (2021). The evolution of the hallmarks of aging. Frontiers in Genetics, 1511.

Leroi, A. M., Koufopanou, V., & Burt, A. (2003). Cancer selection. Nature Reviews Cancer, 3(3), 226-231.

Lewis, K. N., Soifer, I., Melamud, E., Roy, M., McIsaac, R. S., Hibbs, M., & Buffenstein, R. (2016). Unraveling the message: insights into comparative genomics of the naked mole-rat. Mammalian Genome, 27(7), 259-278.

Lipman, R., Galecki, A., Burke, D. T., & Miller, R. A. (2004). Genetic loci that influence cause of death in a heterogeneous mouse stock. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 59(10), B977-B983.

Longo, V. D., Mitteldorf, J., & Skulachev, V. P. (2005). Programmed and altruistic ageing. Nature Reviews Genetics, 6(11), 866-872.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. Cell, 153(6), 1194-1217.

Ma, S., & Gladyshev, V. N. (2017, October). Molecular signatures of longevity: insights from cross-species comparative studies. In Seminars in cell & developmental biology (Vol. 70, pp. 190-203). Academic Press.

MacRae, S. L., Zhang, Q., Lemetre, C., Seim, I., Calder, R. B., Hoeijmakers, J., ... & Zhang, Z. D. (2015). Comparative analysis of genome maintenance genes in naked mole rat, mouse, and human. Aging Cell, 14(2), 288-291.

Malik, A., Domankevich, V., Lijuan, H., Xiaodong, F., Korol, A., Avivi, A., & Shams, I. (2016). Genome maintenance and bioenergetics of the long-lived hypoxia-tolerant and cancer-resistant blind mole rat, Spalax: a cross-species analysis of brain transcriptome. Scientific reports, 6(1), 1-14.

Manov, I., Hirsh, M., Iancu, T. C., Malik, A., Sotnichenko, N., Band, M., ... & Shams, I. (2013). Pronounced cancer resistance in a subterranean rodent, the blind mole-rat, Spalax: in vivo and in vitro evidence. BMC biology, 11(1), 1-18.

Matise, M. (2018). Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime.

Medawar, P. B. (1952). An unsolved problem of biology.

Miyawaki, S., Kawamura, Y., Oiwa, Y., Shimizu, A., Hachiya, T., Bono, H., ... & Miura, K. (2016). Tumour resistance in induced pluripotent stem cells derived from naked mole-rats. Nature communications, 7(1), 1-9.

Moldakozhayev, A., Tskhay, A., & Gladyshev, V. N. (2021). Applying deductive reasoning and the principles of particle physics to aging research. Aging (Albany NY), 13(18), 22611.

Moskalev, A. A., Shaposhnikov, M. V., Plyusnina, E. N., Zhavoronkov, A., Budovsky, A., Yanai, H., & Fraifeld, V. E. (2013). The role of DNA damage and repair in aging through the prism of Koch-like criteria. Ageing research reviews, 12(2), 661-684.

Nordling, C. (1953). A new theory on the cancer-inducing mechanism. British journal of cancer, 7(1), 68.

Nunney, L. (2016). Commentary: The multistage model of carcinogenesis, Peto's paradox and evolution. International Journal of Epidemiology, 45(3), 649-653.

Nunney, L. (1999). Lineage selection and the evolution of multistage carcinogenesis. Proceedings of the Royal Society of London. Series B: Biological Sciences, 266(1418), 493-498.

Nunney, L., Maley, C. C., Breen, M., Hochberg, M. E., & Schiffman, J. D. (2015). Peto's paradox and the promise of comparative oncology. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1673), 20140177.

O'Connor, T. P., Lee, A., Jarvis, J. U., & Buffenstein, R. (2002). Prolonged longevity in naked mole-rats: age-related changes in metabolism, body composition and gastrointestinal function. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, 133(3), 835-842.

Omotoso, O., Gladyshev, V. N., & Zhou, X. (2021). Lifespan extension in long-lived vertebrates rooted in ecological adaptation. Frontiers in Cell and Developmental Biology, 9.

Rankin, K. S., & Frankel, D. (2016). Hyaluronan in cancer–from the naked mole rat to nanoparticle therapy. Soft matter, 12(17), 3841-3848.

Reichard, M. (2016). Evolutionary ecology of aging: time to reconcile field and laboratory research. Ecology and Evolution, 6(9), 2988-3000.

Reichard, M. (2017, October). Evolutionary perspectives on ageing. In Seminars in cell & developmental biology (Vol. 70, pp. 99-107). Academic Press.

Ritter, A. D., Shen, Y., Bass, J. F., Jeyaraj, S., Deplancke, B., Mukhopadhyay, A., ... & Walhout, A. J. (2013). Complex expression dynamics and robustness in C. elegans insulin networks. Genome research, 23(6), 954-965.

Seim, I., Fang, X., Xiong, Z., Lobanov, A. V., Huang, Z., Ma, S., ... & Gladyshev, V. N. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt’s bat Myotis brandtii. Nature communications, 4(1), 1-8.

Seluanov, A., Gladyshev, V. N., Vijg, J., & Gorbunova, V. (2018). Mechanisms of cancer resistance in long-lived mammals. Nature Reviews Cancer, 18(7), 433-441.

Seluanov, A., Hine, C., Azpurua, J., Feigenson, M., Bozzella, M., Mao, Z., ... & Gorbunova, V. (2009). Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. Proceedings of the National Academy of Sciences, 106(46), 19352-19357.

Shepard, A., & Kissil, J. L. (2020). The use of non-traditional models in the study of cancer resistance—the case of the naked mole rat. Oncogene, 39(28), 5083-5097.

Stoll, E. A., Karapavlovic, N., Rosa, H., Woodmass, M., Rygiel, K., White, K., ... & Faulkes, C. G. (2016). Naked mole-rats maintain healthy skeletal muscle and Complex IV mitochondrial enzyme function into old age. Aging (Albany NY), 8(12), 3468.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719-724.

Sulak, M., Fong, L., Mika, K., Chigurupati, S., Yon, L., Mongan, N. P., ... & Lynch, V. J. (2016). TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. elife, 5, e11994.

Taffett, G. E. (2003). Physiology of aging. In *Geriatric Medicine* (pp. 27-35). Springer, New York, NY.

Tejada-Martinez, D., de Magalhães, J. P., & Opazo, J. C. (2021). Positive selection and gene duplications in tumour suppressor genes reveal clues about how cetaceans resist cancer. Proceedings of the Royal Society B, 288(1945), 20202592.

Thieme, R., Kurz, S., Kolb, M., Debebe, T., Holtze, S., Morhart, M., ... & Birkenmeier, G. (2015). Analysis of alpha-2 macroglobulin from the long-lived and cancer-resistant naked mole-rat and human plasma. PLoS One, 10(6), e0130470.

Tian, X., Azpurua, J., Ke, Z., Augereau, A., Zhang, Z. D., Vijg, J., ... & Seluanov, A. (2015). INK4 locus of the tumor-resistant rodent, the naked mole rat, expresses a functional p15/p16 hybrid isoform. Proceedings of the National Academy of Sciences, 112(4), 1053-1058.

Tian, X. (2016). Identification of longevity and cancer resistance mechanisms in long-lived rodent species. University of Rochester.

Tollis, M., Boddy, A. M., & Maley, C. C. (2017). Peto's Paradox: how has evolution solved the problem of cancer prevention?. BMC biology, 15(1), 1-5.

Tollis, M., Robbins, J., Webb, A. E., Kuderna, L. F., Caulin, A. F., Garcia, J. D., ... & Maley, C. C. (2019). Return to the sea, get huge, beat cancer: an analysis of cetacean genomes including an assembly for the humpback whale (Megaptera novaeangliae). Molecular biology and evolution, 36(8), 1746-1763.

Vazquez, J. M., & Lynch, V. J. (2021). Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. Elife, 10, e65041.

Vazquez, J. M., Sulak, M., Chigurupati, S., & Lynch, V. J. (2018). A zombie LIF gene in elephants is upregulated by TP53 to induce apoptosis in response to DNA damage. Cell Reports, 24(7), 1765-1776.

Vittecoq, M., Roche, B., Daoust, S. P., Ducasse, H., Missé, D., Abadie, J., ... & Thomas, F. (2013). Cancer: a missing link in ecosystem functioning?. Trends in ecology & evolution, 28(11), 628-635.

Wensink, M. J. (2016). Size, longevity and cancer: age structure. Proceedings of the Royal Society B: Biological Sciences, 283(1838), 20161510.

Williams, G. C. (2001). Pleiotropy, Natural Selection, and the Evolution of Senescence: Evolution 11, 398-411 (1957). Science of Aging Knowledge Environment, 2001(1), cp13-cp13.

Wolf, A. M. (2021). The tumor suppression theory of aging. Mechanisms of Ageing and Development, 200, 111583.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., ... & Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. Nature genetics, 45(10), 1134-1140.

Zepeda Mendoza, M. L., Xiong, Z., Escalera-Zamudio, M., Runge, A. K., Thézé, J., Streicker, D., ... & Gilbert, M. P. (2018). Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. Nature ecology & evolution, 2(4), 659-668.

Chapter II:

Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research.

Chiara Vischioni[1], Fabio Bove[2], Federica Mandreoli[2], Riccardo Martoglia[2], Valentino Pisi[2], Cristian Taccioli[1]

[1] MAPS - University of Padova, Italy

[2] FIM - University of Modena and Reggio Emilia, Italy

It has been suggested that gene duplication is one of the main factors driving genetic diversity. Indeed, it is not only responsible for the adaptation towards different environmental fluctuations, but it also somehow manages to shape the genome of certain species, contributing to the emergence of alternative traits and pathways of developmental programs (Magadum et al. 2013). In this context, for example, an altered number of TS genes can play a key role in forging the route leading to cancer resistance in large size and long-lived organisms (Vazquez and Lynch, 2021). In particular, CNVs are gains or losses in copies of genes contained in cell DNA, which can be associated with phenotypic variations, including disease (Feuk et al., 2006). Moreover, the variation in CNVs has recently been correlated with longevity and cancer resistance. According to the hypothesis that positively selected CNVs tend to recur during cancer progression (Beroukhim et al., 2010; Bignell et al., 2010), but also during the evolution, during the course of my Ph.D, I developed **VarNuCopy**, the first database of genome Copy Number Variation across the animal kingdom. In the actual version, it includes the variation landscape of genes copies among the genome of 233 organisms, combining, for some of them, CNV, longevity, mass, metabolism, gene names and other important biological and genomic parameters. The database is built under two main parallel sections: the first one provides a general analysis, describing the genes copy number panel in a particular species, or, inversely, searching for a specific gene presence in multiple organisms. The second analysis is more cancer-oriented: together with my collaborators, I created some models able to correlate the copy-number variation of genes, and in particular of tumour suppressors and oncogenes of 24 organisms, with the phenotypical characteristics described above. In few words, the platform allows the exploration of complex data-sets to assess copy number variation of any given gene throughout the referenced species. Moreover, by performing a basic query in the Genes Exploration section, researchers investigating particular gene copies gain/loss can easily retrieve and analyze the difference in copy number in an inter-species comparison approach. Finally, VarNuCopy Descriptive Analysis Models (DAMs), which are exclusively built using both the species cancer incidence and the genomic copy number variation, can help scientists to statistically discriminate between cancer–prone and cancer–resistant organisms. The final aim of this unique tool is, indeed, to easily compare patterns of copy number changes, in order to identify new oncogenes or tumour suppressors targets, related to species longevity, weight, and metabolism. According to our bioinformatics analysis, VarNuCopy is able to confirm that a gene can be (possibly) involved in biological processes of cancer onset. As far as I know, this is the only tool of its kind, that is able to compare the CNVs landscape of multiple species, both for model and non-model organisms in biomedicine, pinpointing towards the discovery of specific genes and pathways that might play a role in cancer resistance.

The following pages contain a reproduction of the paper published in Big Data Research:
Vischioni, C., Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2022). Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research. *Big Data Research*, *27*, 100298.
https://doi.org/10.1016/j.bdr.2021.100298
Reprint was made with the permission of the publisher, Copyright Elsevier (2022).

Abstract

The study of *Copy Number Variations* (CNVs) is recently emerging as a hot topic for biomedical cancer research. While different data sources, websites, and tools concerning genomic CNVs have been made publicly available, CNV data is still a largely unexplored source of biological information, due to the limitations of currently available analysis tools. To this respect, we propose a novel platform, named VarNuCopy, that overcomes such limitations by pursuing the core principles of Exploratory Data Analysis (EDA) in the context of Copy Number Variation (CNV) data. The platform has been made publicly available as a web application, and is, to our best knowledge, the first tool enabling visual, interactive exploration and analysis of the CNV landscape of multiple species. Through novel client and server-side optimizations inspired by scalable data science, VarNuCopy implements a comprehensive and efficient data exploration solution that empowers researchers to easily recognize complex trends and patterns within a huge amount of data concerning CNVs, and to identify new target genes that might function as tumor suppressor and oncogenes.

Keywords: Copy Number Variations (CNVs); Interactive visualization; Exploratory Data Analysis (EDA); Scalable data science; Data analysis models.

1. Introduction

The study of the species showing peculiar properties in terms of cancer resistance and high rates of longevity is recently emerging as a hot topic for biomedical research, producing fundamental insights into the mechanisms which could protect an organism against the development of tumorigenesis (Serrano., 2015). A promising area of research is the study of *Copy Number Variations* (CNVs), which are defined as the number of gene copies within a genome, that might be also related to genome instability (Feuk et al., 2006) and phenotype alteration. For this reason, in the last years, different data sources, websites, and tools covering genomic CNVs have been made publicly available (Bragin et al., 2014; Chen et al., 2009; Howe et al., 2021; Qiu et al., 2012; Zarrei et al., 2019). However, despite their potentialities in providing useful insights for the identification of oncogenes and tumor suppressors, these data still represent a largely unexplored source of biological information, due to the poor analysis functionalities offered by the currently available tools. Generally speaking, databases of biological knowledge have become essential resources that are used daily by biologists around the world, and different biogenetics databases are nowadays available to support comparative genomics, i.e., the branch of science in which computational tools are used to compare the genome sequences of multiple species, distinguishing different pattern of similarity and/or variation. In this context, we developed VarNuCopy, a new tool able to compare and correlate the CNV landscape of different organisms, allowing the identification of genomic regions of high instability. In VarNuCopy, we address the problem of providing access to large and rich biogenetics databases for biomedical research by starting from the methodological aspect, where the issue is that of database usability, and propose to adopt the *Exploratory Data Analysis* (EDA) principles (Behrens., 1997) to provide visual and interactive ways to explore, summarize, and analyze data in a simple and user-friendly manner. The contributions of the VarNuCopy platform, whose current version has been made publicly available as a web application at http://isgroup.mat.unimore.it:8083 (as done in the past for other genomic exploration tools) (Lomonaco et al., 2014), are the following:

a. the platform is based on a unique and rich database that combines CNV data among different species with other vital parameters;
b. the information is retrieved from public on-line genomic libraries;
c. exploratory data analysis is made possible through a variety of tools, allowing researchers to: (*i*) freely combine and use different *Data*

Analysis Models (DAMs), newly introduced analytical tools useful to generate visual and interactive reports and plots, (*ii*) access additional related data available in reference online sources, (*iii*) submit sophisticated research questions by means of custom queries over the database schema;

d. the client technologies are coupled with server-side optimizations inspired by scalable data science that together enable the necessary real-time interaction experience on large amounts of data.

In this way, VarNuCopy implements a comprehensive and efficient solution of data exploration that empowers researchers to easily recognize complex trends and patterns within a huge amount of data concerning CNVs. By building their own data exploration paths according to their deductions, researchers can hypothesize new possible target genes based on previous predictions, highlighting which are the ones possibly linked to cancer resistant species and/or long-living organisms. Following the idea that copy number variation of important genes can theoretically protect a species from cancer insurgence, the platform allows the exploration of complex data-sets to assess copy number variation of any given gene throughout the referenced species. Moreover, researchers investigating particular gene copies gain/loss can easily retrieve and analyze the difference in copy number in an inter-species comparison approach to underlie the higher/lower need for that gene. Finally, VarNuCopy DAMs models, which are exclusively built using both the species cancer incidence and the genomic copy number variation, can help scientists to statistically discriminate between cancer–prone and cancer–resistant organisms and eventually to discover new possible genetic mechanisms involved in oncogenesis.

This paper is based on the preliminary results presented in (Bove et al., 2020) and significantly extends the previous work by providing new detailed descriptions of the platform functionalities (including up-dated database and analysis tools), comprehensive background and related work discussions and novel insights on technological, ar- chitectural and performance evaluation aspects, as well as on the potential of the platform from a bioinformatic research point of view.

The paper is structured in the following way: Section 2 discusses the background/methodological claim of EDA as a mean for biogenetics database usability. Section 3 describes the database on which the platform is built on, whereas the platform itself is described in Sections 4 (high-level overview), 5 (EDA functionalities) and 6 (implementation strategies). A performance evaluation is presented in Section 7, while Section 8 analyzes and compares related works to our proposal. Finally, Section 9 concludes the paper and discusses future works.

2. EDA as a mean for biogenetics database usability

In this background section, we recall the principles, applications and issues of EDA and graphical analysis, and discuss the methodological claim of this work, i.e., how EDA can effectively support biogenetics research.

EDA and Graphical analysis. Researchers often struggle to develop hypotheses despite data availability abundance. In recent years, EDA and data visualization techniques (Tufte., 2001) have been suggested as effective steps for pattern and hypothesis generation in a data science process (Di Blas et al., 2017; Ma et al., 2017; Schutt & O'Neil., 2013). EDA is a well-established statistical tradition that provides conceptual and computational tools to discover patterns in a data science context (Behrens., 1997). EDA is typically characterized by an emphasis on: (i) a substantive understanding of data; (ii) graphic representations of data; (iii) tentative model building in an iterative/inter- active process; (iv) flexibility regarding which is the best method to apply. The final aim is to discover patterns within data. In 1977, Tukey proposed the data analyst as the one able to listen to the data in different ways, until a plausible "story" becomes clear. The works (Hoaglin et al., 1991; Tukey., 1977; Velleman et al., 1992) are among the most important ones in the classical EDA tradition. EDA data science method is equivalent to the data-driven abductive approach (Ma et al., 2017), in which it is possible to obtain plausible information starting from the phenomenon observation; conversely, deduction is the process to refine the hypothesis with new supportive evidence, while induction approach extrapolates the hypothesis based on a general law or theory. Graphical analysis plays a central role in the EDA context: by increasing the number of algebraic summaries, graphics can simultaneously show numerous data values, thus avoiding missing important patterns. Indeed, in recent years, researchers have proposed that data visualization should be applied in each data science process (Fox & Hendler., 2011). The incorporation of modern visualization tools into the analytics process enables scientists to easily understand large scientific datasets, and to produce, in an easy-to-use way, quick methods to explore new hypotheses.

Data management issues in EDA: scalable data science. In order to achieve these goals, the scientific community has to overcome a wide range of analysis problems

(Fox & Hendler., 2011). First of all, it is not always obvious which is the best and the most effective method to visualize specific data at each analysis stage. For this reason, from one side, many visualization techniques have been developed, including histograms, scatter and box plots, mind maps (Buzan., 2015), and conceptual maps (Novak & Cañas., 2006), but, on the other side, the development of new approaches is still required. Moreover, in order to quickly and easily provide scientists the requested data, it is important to merge particular kinds of statistics with proper ways of visualization (Card et al., 1999). Generally speaking, one of the biggest challenges related to EDA is data management (Di Blas et al., 2017; Buoncristiano et al., 2015). On one hand, the amount of data that has to be analyzed often requires the linking of an extremely large number of complex and difficult-to-model data sources. On the other, EDA tools require instant responses to each user interaction, where even a single click on the user interface can trigger very complex processing and manipulations in the back-end. In order to utilize high-performance computing resources, most of the existing workflow systems submit jobs to external resource management systems (such as HTCondor) (Thain et al., 2015). Recently, new big data management techniques have been proposed to integrate the workflow and resource management facilities (Tang et al., 2013). These include the use of materialization and materialized views, to improve the performance of query workloads in the underlying database engines (Du et al., 2017; Elghandour & Aboulnaga., 2012). These techniques can allow the creation of standalone analysis platforms able to provide sufficient capacity and performance needed in the EDA workflows. Following the scalable data science trend, many analytic platforms are becoming available, empowering the domain sciences, healthcare, humanities, governance, journalism, among others, ready to be studied at scales and granularities which were impossible before: from cloud data mining solutions (Talia., 2019) to data streams visualization (Breve et al., 2020), from large social network analysis (Cao et al., 2015; Birkholz et al., 2012) to the medical field, where platforms such as (McPadden et al., 2019) are proposed as new ways to real-time access health care research data.

Biogenetics Database Usability through EDA. In particular, from a biogenetics database point of view, in the last few decades the availability of (biological) data has been constantly increasing day by day, making data science, analysis and visualization a fundamental instrument to deal with such a rise in information sources (Schmidt., 2020). Databases play an important role in exploring, understanding, and finding new insights among different datasets, allowing humans to benefit from a better data representation (Tufte., 2001). In this context, performance and functionality alone are not enough to support data analysis itself, but they also need to be associated with the usability of the platform. Therefore, in recent years, since usability plays a pivotal role in human-computer interaction, issues related to that topic have gained more and more importance within the database community (Li & Jagadish., 2012). Regarding this, the methodological claim of this work is that EDA can become a very effective way to support biogenetics research and, more specifically, biogenetics database usability. The platform we propose, VarNuCopy, follows both the principal criteria of the so-called "database usability": innovative query interface design and database personalization. Moreover, it can be also described by all the multiple components of the usability description proposed in (Nielsen., 1994) and discussed later by Catarci (Catarci., 2000): (i) learnability, rapidly and easy to learn, (ii) efficiency, efficient to use, (iii) memorability, easy to remember, (iv) low error rate, (v) satisfaction, system pleasant to use, and (vi) human-centered system design.

3. The Dataset

VarNuCopy allows researchers to access a unique dataset that combines different kinds of information regarding genes, organisms, and families with the number of copies of each gene in a species genome (CNVs). Data originates from different sources: the CNV landscape for each species was obtained from Ensembl comparative genomics resources (Yates et al., 2019); the list of oncogenes and tumor suppressors was collected from NCBI Genome (http://www.ncbi.nlm.nih.gov/genome), Tumor Suppressor Gene (TS-Gene) (Zhao et al., 2016), and Oncogene Database (Liu et al., 2017), whereas gene families data were downloaded from Gene Ontology (http://geneontology.org). The dataset is stored in a relational database that has the following schema (see also Figure 1 for the ER diagram):

```
Gene (ID_gene, gene_name, is_oncogene, is_tumor_suppressor)
Species  (ID_species,  species_name,  is_long_lived,  is_cancer_prone,
          metabolism, avg_weight, necropsies,class)
Family (ID_family, family_name)
Classification (ID_gene, ID_family)
     FK: ID_gene REFERENCES Gene
     FK: ID_family REFERENCES Family
CopyNumber (ID_gene, ID_species, qty)
     FK: ID_gene REFERENCES Gene
     FK: ID_species REFERENCES Species
```

It contains the following number of tuples per table: 21,067 tuples for the Gene table; 9,996 tuples for the Family table; 233 tuples for the Species table; 3,473,365 tuples for the CopyNumber table; 835,089 tuples for the Classification table. The Gene table contains two boolean attributes, is_oncogene and is_tumor_suppressor, that indicate whether the instance is known to be an oncogene or a tumor suppressor, respectively. This information is available only for a small portion of the genes (12% of the total). The Species table, instead, reports a list of phenotypic parameters that are someway related to the process of cancer insurgence (Myers et al., http://animaldiversity.org): metabolism contains the species metabolic rate measured in Watt, avg_weight is the average weight of the organism measured in Kg, longevity represents the maximum longevity of the species, necropsies represents the number of necropsies performed on a single organism to obtain the percentage of cancer incidence, class refers to the phylogenetic classification of the species. Finally, is_cancer_prone (boolean) distinguishes species between cancer-prone and cancer resistant, whereas is_long_lived (boolean) assesses whether the instance is a long-living species or a short-living one. Globally, this information is currently available for 11% of the species.

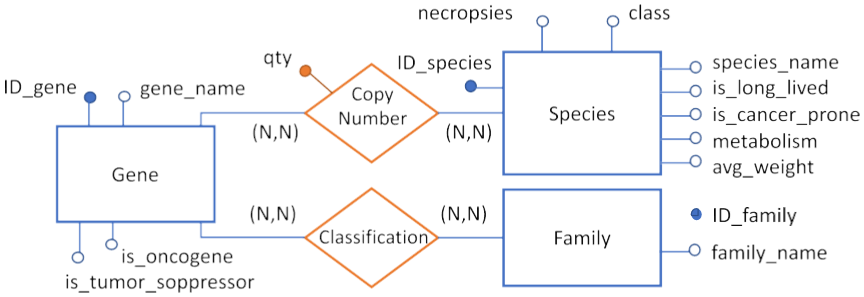


Figure 1: VarNuCopy Database schema: ER Diagram

4. Platform overview

VarNuCopy is a web-based platform built with the aim to compare and correlate the CNVs of different organisms, in order to understand if they have any role in tumors onset and, ultimately, to identify new potential genes that might be function as tumor suppressor and oncogenes. To this end, VarNuCopy implements a comprehensive approach of data exploration over the rich dataset presented in Section 3. The main challenges we addressed cover four different aspects (Di Blas et al., 2017): (1) the user point of view, (2) the user experience during exploration, (3) the possible outcomes, and (4) the relevance of the data output. The platform is geared towards researchers having different background and exploratory goals, spanning from the ones with no knowledge on CNVs domain, and therefore with very generic exploratory goals, to those who want to explore the dataset through sophisticated research questions. Therefore, an exploratory session can start with an in-depth or with a generic information request and can proceed in an on-going conversation between the user and the system, sustained by search forms and interactive plots showing query analysis results. These graphical representations can be *Descriptive Analysis Models* (DAMs), that show CNVs trends of groups of species known to have properties in terms of longevity and cancer rate, or contained in target reports, in which a summary of CNV statistics about target genes or species is shown. The system supports users in progressively gathering the information of interest by highlighting the CNVs statistical properties that can be exploited to identify interesting subsets of the query analysis results that are worth of further explorations. As shown in Figure 2, our web application is organized in four areas and seven different data exploration components:

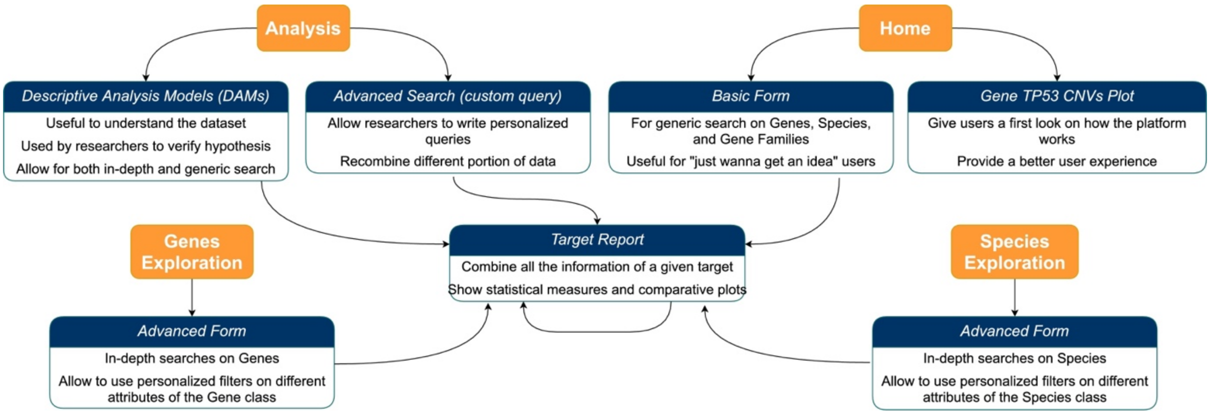

Analysis
Home
Descriptive Analysis Models (DAMs)
Useful to understand the dataset
Used by researchers to verify hypothesis
Allow for both in-depth and generic search
Advanced Search (custom query)
Allow researchers to write personalized queries
Recombine different portion of data
Basic Form
For generic search on Genes, Species, and Gene Families
Useful for "just wanna get an idea" users
Gene TP53 CNVs Plot
Give users a first look on how the platform works
Provide a better user experience
Genes Exploration
Target Report
Combine all the information of a given target
Show statistical measures and comparative plots
Species Exploration
Advanced Form
In-depth searches on Genes
Allow to use personalized filters on different attributes of the Gene class
Advanced Form
In-depth searches on Species
Allow to use personalized filters on different attributes of the Species class

Figure 2: Platform schema: interactions between the different areas of the platform (shown in orange), and their components (in blue and italics).

- The “**Home**” area welcomes researchers and proposes a basic overview of the platform; moreover, it allows simple search functionalities on genes, species and families.
- The “**Species Exploration**” and “**Genes Exploration**” areas provide advanced search queries on species and genes, respectively.
- The “**Analysis**” area proposes several DAMs and an advanced query form useful for ad-hoc searching and filtering.
- Target reports implement the different models, measures, and statistical tests, allowing the user to interact with them.

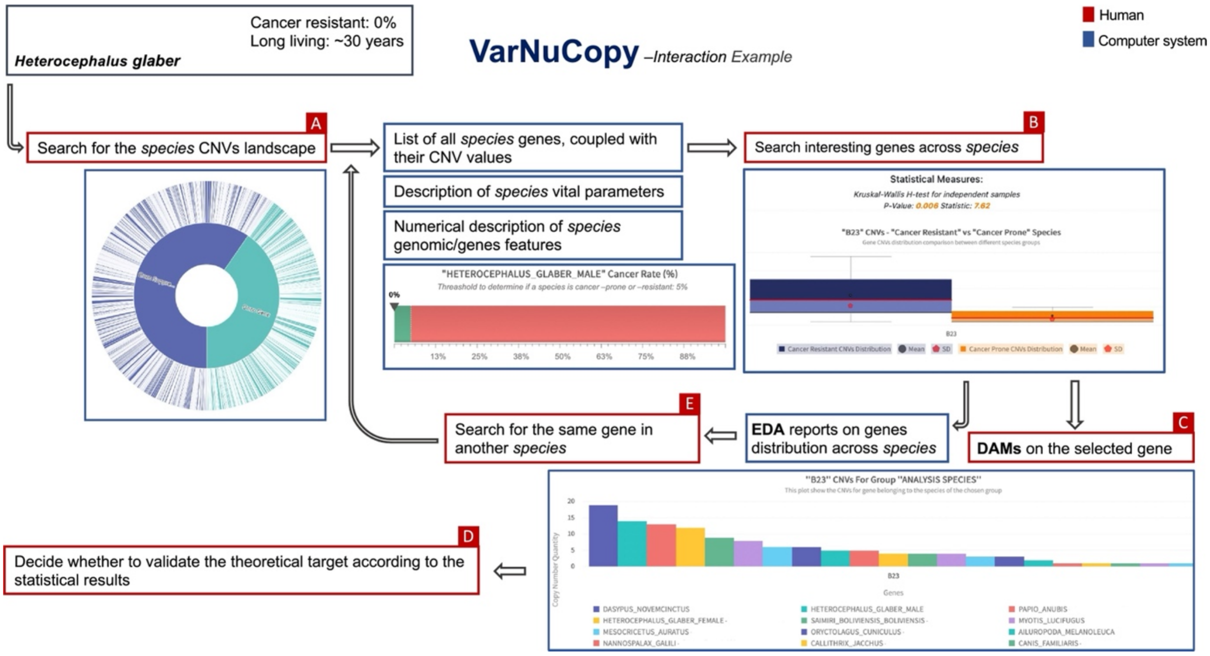

Cancer resistant: 0%
Long living: ~30 years
Heterocephalus glaber
VarNuCopy –Interaction Example
Human
Computer system
A
Search for the species CNVs landscape
List of all species genes, coupled with their CNV values
Description of species vital parameters
Numerical description of species genomic/genes features
"HETEROCEPHALUS_GLABER_MALE" Cancer Rate (%)
Threashold to determine if a species is cancer –prone or –resistant: 5%
0%
13%
25%
38%
50%
63%
75%
88%
B
Search interesting genes across species
Statistical Measures:
Kruskal-Wallis H-test for independent samples
P-Value: 0.006 Statistic: 7.62
"B23" CNVs - "Cancer Resistant" vs "Cancer Prone" Species
Gene CNVs distribution comparison between different species groups
B23
Cancer Resistant CNVs Distribution
Mean
SD
Cancer Prone CNVs Distribution
Mean
SD
E
Search for the same gene in another species
EDA reports on genes distribution across species
C
DAMs on the selected gene
D
Decide whether to validate the theoretical target according to the statistical results
"B23" CNVs For Group "ANALYSIS SPECIES"
This plot show the CNVs for gene belonging to the species of the chosen group
Copy Number Quantity
20
15
10
5
0
B23
Genes
DASYPUS_NOVEMCINCTUS
HETEROCEPHALUS_GLABER_FEMALE
MESOCRICETUS_AURATUS
NANNOSPALAX_GALILI
HETEROCEPHALUS_GLABER_MALE
SAIMIRI_BOLIVIENSIS_BOLIVIENSIS
ORYCTOLAGUS_CUNICULUS
CALLITHRIX_JACCHUS
PAPIO_ANUBIS
MYOTIS_LUCIFUGUS
AILUROPODA_MELANOLEUCA
CANIS_FAMILIARIS

Figure 3: Descriptive Analysis Models (DAMs) UI - Interaction example starting the analysis from the *Heterocephalus glaber* species.

Following the "exploratory" philosophy, where the EDA process is flexible and the result is uncertain (Behrens., 1997), the users can incrementally build their own data exploration path according to their deductions. In the following, we describe a possible EDA session showcasing how VarNuCopy can support a biologist in the challenging task of hypothesis generation and testing. The example focuses on the *Heterocephalus glaber* species, whose common name is naked mole rat, and the test is aimed at searching its CNV trends to verify if this organism has particular molecular targets that can underlie its high longevity rate and its low incidence tumor percentage, compared to the other species. Such a session, whose main exploratory interactions are shown in Figure 3, starts with a simple query on the target species as follows:

A) knowing that the naked mole rat is a well-defined cancer resistant species (Tian et al., 2013), the user starts from the "Home" area and enters the species name to obtain its entire genomic copy number landscape. This is depicted in the corresponding target report as a pie plot where the CNV ratio between tumor suppressors and oncogenes are made evident through two different colors, blue for the former and green for the latter, and where each gene bar width is proportional to its CNV. The same report also includes the description of the vital parameters of the species (lifespan, weight, metabolism, and cancer incidence). Moreover, it shows relevant statistical measures concerning its genome, such as the total number of genes, and the fraction of tumor suppressor/oncogenes in respect to the entire genome composition (as shown by the bar plot in Figure 3).

B) The user is interested in a specific gene, B23, which presents a copy number expansion in the naked mole rat genome. By clicking on its gene bar in the pie plot, the researcher gets the gene CNVs across different species and can easily compare their distribution within cancer -prone and -resistant organisms through a box plot. The result is shown in a new target report, from which the user can start to perform DAM analysis.

C) For example, in the figure, it is depicted the "Common Genes in Species Group" model on B23 gene, which provides the bar plot of its copy number distribution in decreasing order of species.

D) These statistical measures allow the biologist to make a new hypothesis: as B23 is a new molecular target that can potentially discriminate between the 2 groups (p-value < 0.05), it is an interesting gene to be further investigated and experimentally validated.

E) The user can now try to verify if the same gene expansion happens in some other species. Starting from the previous DAM and clicking on one of the species represented in the bar plot, it is possible to iterate the EDA process on this species for new investigations.

In order to efficiently support the above EDA functionalities, VarNuCopy implements data caching and parallel threading solutions enabling rapid and effective response from the platform, also against expensive requests. The design of such mechanisms, inspired by the scalable data science trend, will be described in Section 6.

5. Exploratory Data Analysis functionalities

EDA functionalities constitute the core of the platform analytical power. Data analytic requests can be:

- **triggered** in different ways: (a) by means of the query forms; or (b) by interacting in a recursive way with one of the plots from previous analyses (i.e., from a DAM or target report)

- **targeted** on different objects: (a) specific database *entities* (i.e., genes, species, families, and groups) as prescribed by each DAM; or (b) custom subsets of the dataset for even greater flexibility.

5.1. Descriptive Analysis Models (DAMs)

Descriptive Analysis Models (DAMs) are novel analytical tools developed to support effective exploratory data analysis over CNVs properties related to both genes and species. VarNuCopy includes different DAMs for different analytical needs, which are dynamically generated by selecting one of the available models applied to a specific target group. Four default target groups are provided: cancer resistant and cancer-prone species, long-living and non- long-living species. Finally, through response formatting, the output of the analysis is generated, and the results are sorted on the basis of different ranking strategies. Figure 4 shows these schema elements.

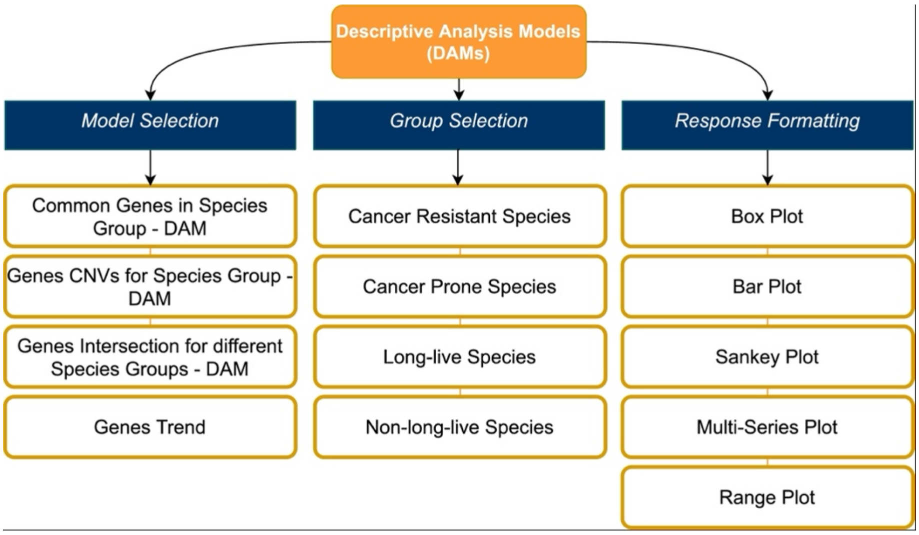


Figure 4: Descriptive Analysis Models (DAMs) schema: a *model* is applied to a *group*, while the resulting report is created as *response.*

The user interfaces of the four different DAMs coupled with the response output are instead shown in Figure 5:

- **Common Genes Model**: bar plot of the CNVs of genes of the species of the selected group (e.g., cancer -prone or long-living species) where results are ranked in decreasing CNV order. This is essential to highlight the most promising genes; for instance, the example introduced in Section 4 allowed the selection of B23 gene;

- **Genes CNVs for Species Group Model**: box plot of CNVs statistical measures such as mean, standard deviation, max and min. It is useful to know which are the genes with highest CNV in one species of the selected group and how their statistical measures fluctuate. Results are sorted by decreasing maximum number of gene copies;
- **Genes Intersection Model**: this model requires two different target groups and shows in a Sankey plot the maximum number of copies of the genes in the species belonging to one of the two. The thickness of the lines connecting each gene with each group allows researchers to easily identify those genes that, showing remarkable differences between the two different groups, prove to be interesting for further investigation;
- **Genes Trend**: differently from the other models, this model does not require the selection of a group but works on both cancer-prone and cancer resistant species data. It provides a multi-series plot showing the diffusion of the different genes across the available species, using an ad-hoc ranking criterion (see Section 5.4 for details).

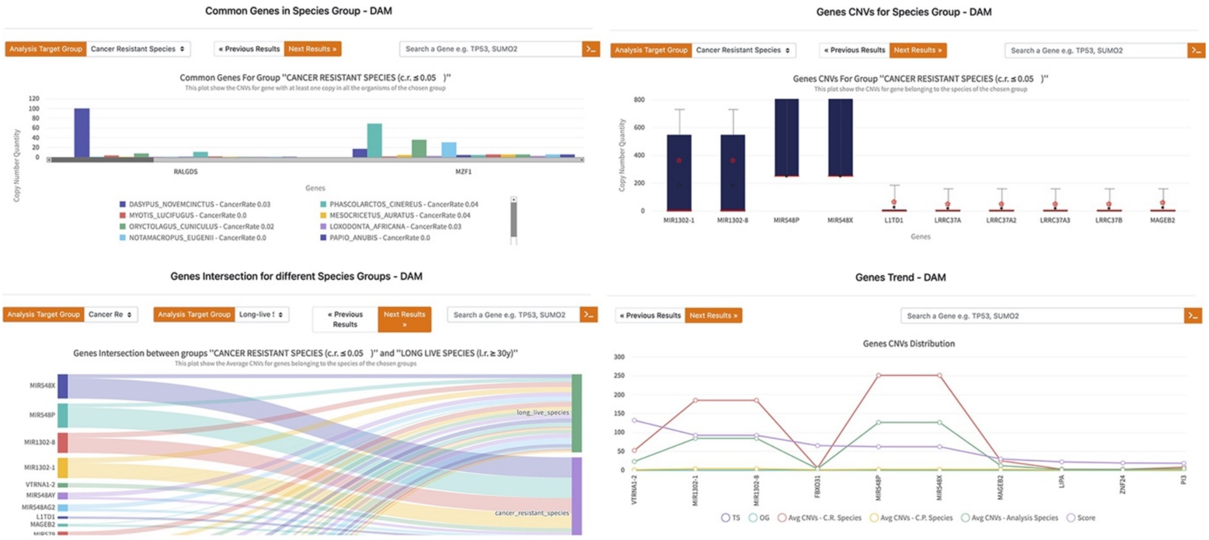

Common Genes in Species Group - DAM
Analysis Target Group
Cancer Resistant Species
« Previous Results
Next Results »
Search a Gene e.g. TP53, SUMO2
Common Genes For Group "CANCER RESISTANT SPECIES (c.r. ≤ 0.05)"
This plot show the CNVs for gene with at least one copy in all the organisms of the chosen group
Copy Number Quantity
120
100
80
60
40
20
0
RALGDS
MZF1
Genes
DASYPUS_NOVEMCINCTUS - CancerRate 0.03
MYOTIS_LUCIFUGUS - CancerRate 0.0
ORYCTOLAGUS_CUNICULUS - CancerRate 0.02
NOTAMACROPUS_EUGENII - CancerRate 0.0
PHASCOLARCTOS_CINEREUS - CancerRate 0.04
MESOCRICETUS_AURATUS - CancerRate 0.04
LOXODONTA_AFRICANA - CancerRate 0.03
PAPIO_ANUBIS - CancerRate 0.0
Genes CNVs for Species Group - DAM
Analysis Target Group
Cancer Resistant Species
« Previous Results
Next Results »
Search a Gene e.g. TP53, SUMO2
Genes CNVs For Group "CANCER RESISTANT SPECIES (c.r. ≤ 0.05)"
This plot show the CNVs for gene belonging to the species of the chosen group
Copy Number Quantity
800
600
400
200
0
MIR1302-1
MIR1302-8
MIR548P
MIR548X
L1TD1
LRRC37A
LRRC37A2
LRRC37A3
LRRC37B
MAGEB2
Genes
Genes Intersection for different Species Groups - DAM
Analysis Target Group
Cancer Re
Analysis Target Group
Long-live !
« Previous Results
Next Results »
Search a Gene e.g. TP53, SUMO2
Genes Intersection between groups "CANCER RESISTANT SPECIES (c.r. ≤ 0.05)" and "LONG LIVE SPECIES (l.r. ≥ 30y)"
This plot show the Average CNVs for genes belonging to the species of the chosen groups
MIR548X
MIR548P
MIR1302-8
MIR1302-1
VTRNA1-2
MIR548AY
MIR548AG2
L1TD1
MAGEB2
MIR570
long_live_species
cancer_resistant_species
Genes Trend - DAM
« Previous Results
Next Results »
Search a Gene e.g. TP53, SUMO2
Genes CNVs Distribution
300
250
200
150
100
50
0
VTRNA1-2
MIR1302-1
MIR1302-8
FBXO31
MIR548P
MIR548X
MAGEB2
LIPA
ZNF24
PI3
TS
OG
Avg CNVs - C.R. Species
Avg CNVs - C.P. Species
Avg CNVs - Analysis Species
Score

Figure 5: User interfaces of the four main DAMs: Common genes in Species Group (top left), Genes CNVs for Species Group (top right), Genes Intersection for different Species Groups (bottom left) and Genes Trend (bottom right).

5.2. Target reports

A target report provides detailed information about a selected gene or species. Reports contain interactive graphs constructed from CNV data related to the specific target in the dataset, and additional information provided by dynamically generated links to external knowledge sources, such as Gene Cards and Gene Ontology for genes (https://www.genecards.org/; http://geneontology.org/), and ADW (http://animaldiversity.org/) for species. Specifically, additional resources for target genes include gene families and function descriptions, while mean copies of the genes and variance in cancer-prone and cancer resistant species, associated with the Kruskal-Wallis H-test p-value, are computed from the databases. The H-test is used to determine if there are significant differences between two groups of species, where a *p-value* lower than 0.05 can be used to discriminate between cancer-prone and cancer resistant. Conversely, for a subset of 24 species, the platform returns cancer rate, longevity, metabolism, average weight and number of copies of oncogenes and tumor suppressors within the genome. Most of this information is presented in graphical form to give a quick glance of the results. For instance, an interactive box plot is used to represent the CNV statistics described above for gene B23 (as shown in Figure 3, step B).

5.3. Custom queries

Custom queries aim to support advanced research questions and can be specified in a dedicated free text form. Unlike the basic forms, where the queries concern genes or species and are suggested by auto-complete mechanisms, custom queries give the user access to the whole database by means of a simple easy-to-use syntax. When prompting for a custom query, the system helps the user by showing a popup window summarizing the database schema and the different available attributes.
The syntax allows the user to:

- request any attribute (column) of any table in the database (**show** clause) (the syntax is simply a list of the attributes required in the output, including possible attributes involved in filtering clauses);
- filter the retrieved information according to conditions over the values of the selected columns (**filter** clause, to limit the results to the ones satisfying the condition(s), and **exclude** clause, to exclude from the results the ones satisfying the condition(s)). Conditions are simply expressed in the form `<attribute> <comparator> <value>`, where possible comparators are `=, >=, <=, >, <`. String values require quotes.

For instance, the following query (Q_A) retrieves the genes and the related CNVs of the *Homo Sapiens* species:

```
show:(species_name, gene_name, qty)
filter:(species_name = HOMO SAPIENS)
```

whereas the following query (Q_B) retrieves the species where the copy number of the gene TP53 is less than 5:

```
show: (species_name, gene_name, qty)
filter: (gene_name=''TP53'') filter: (qty<5)
```

5.4. Ranking analysis

Data analytics requests often output a very large number of results, either genes or species. It is therefore of utmost importance to implement ranking strategies that reward the genes or species that are the most representatives in the selected group. For most plots, ranking is in CNV decreasing order. Moreover, VarNuCopy introduces a new ranking mechanism, named "genes trend", which relies on the CNV property, and which is used in the corresponding DAM. Its aim is to identify other genes within the genome of mammalian organisms with a CNV behavior similar to TP53 gene in *Loxodonta africana* genome, another well-known long-living species. Indeed, this gene presents a high amplification in the number of copies which has been described as a protective mechanism against the cancer insurgence for the elephant (Sulak et al., 2016). Since the amplification of TP53 gene copies is not present in any other analyzed species, we consider elephant TP53 CNV as an "anomaly" compared to the other organisms. By definition, anomalies represent data patterns that have different data characteristics from normal instances, and the ability to detect anomalies has significant relevance, providing critical and actionable information that is worth of further investigation. To this end, the ranking formula we propose assigns high raking scores to genes having an elevated number of copies in cancer-resistant species, but a low distribution of gene copies in the rest of the population. The ranking formula for the gene *g* is the following:

$$rank\ (g) = {(MaxCNVg_{CancerResistance} - AvgCNVg_{AllSpecies})}\Big/{MaxCNVg_{CancerProne}}$$

where MaxCNV stands for the maximum value of number of copies of a gene in the cancer resistance and cancer prone group respectively, whereas AvgCNV represents the CNVs mean value, considering all the species. The higher the rank(g) value is, the more the gene is interesting in terms of copy number variation across species, and in particular within the two groups (cancer-prone vs -resistant organisms). For example, considering again *Loxodonta africana* genome, among the top ranked genes, besides TP53, we can find two other important hits, MT1G and BEX2, which seem to be promising for further experimental validation. Indeed, from a biological point of view, we know that MT1G can indirectly increase the stability of TP53 and lead to cell cycle arrest and apoptosis by inhibiting the expression of MDM2 gene TP53 ubiquitination factor (Wang et al., 2019), while BEX2 gene encodes a protein detected in some types of cancer, such as breast tumors (Naderi et al., 2010).

6. Implementation strategies

VarNuCopy has been implemented with advanced server- and client-side technologies and optimizations, also inspired by scalable data science, enabling the needed real-time interaction experience on a large amount of data. The following sections describe the exploited architectural optimizations (Section 6.1) and how data requests are managed (Section 6.2).

6.1. Detailed architecture and optimizations

The platform is implemented as a web application written in Python and based on several up-to-date technologies for an efficient client-server web application development and data management (Figure 6).
Specifically, it is composed of:

- a **client-side** part (top of figure, green color), offering an analysis interface for the final user, and an administrator interface for platform and data maintenance. The front-end is implemented in Django (http://www.djangoproject.com/) and FusionCharts (http://www.fusioncharts.com/).
- a **server-side** part (bottom of figure, blue color) offering data management functionalities, exploiting Pandas (http://pandas.pydata.org/), Scikit-Learn (http://scikit-learn.org/), and PostgreSQL (http://www.postgresql.org/). Pandas was used for data caching, manipulation, and analysis, providing the facilities for the management of data frames and time series, while PostgresSQL was chosen for database storage. A relational database ensures good write performances (in the future we plan to allow users to store user-defined data) and is able to deliver shorter time-to-insight than big data management frameworks due to its capability of analyzing non-huge data sets in real or near-real-time. Moreover, it best fits the well-defined and stable database structure the dataset presents.

The advanced and interactive analysis functions offered by the platform would be nearly useless without efficient processing strategies allowing immediate response time even for complex requests. The platform forms require the querying, retrieval and presentation of large amounts of data (as highlighted in Figure 6 for the detailed architecture and the operation flow). In particular, the user interacts with the platform (upper left part of figure) and triggers the generation and visualization of new results in form of reports (upper right part); this requires a number of processing operations that are performed at client-side and/or at server-side, depending on the specific request. For efficient processing, VarNuCopy implementation exploits server-side caching and threading as well as client-side caching and chunking optimizations. Before considering typical examples of requests submitted to our platform and discussing what kinds of operations and optimizations are performed in order to efficiently answer them, we will start by giving general details on the specific optimizations:

- **Server-side caching**. As a commonly used technique in scalable data analysis research (Du et al., 2017; Elghandour & Aboulnaga., 2012), we store intermediate materialized results and create highly modular caches ("server-side caches" in Figure 6) containing the data views that are most often required in the analyses. To this end, we assumed the platform workload would, for the majority of times, be triggered by DAM generation requests and DAM interactions, and selected the intermediate query results to be materialized according to the model in the DAM schema (Figure 4). VarNuCopy stores cached data in three main repositories:
  – a "Species" cache containing the information about the different species as available from the `Species` table (see Section 3)
  – a "Gene" cache containing the copy numbers information with their gene names (`Gene` and `CopyNumber` tables). This information is available in two versions: as it is, or joined with the involved species
  – a "Family" cache containing the families (`Family` table) already joined with the `Classification` table (and ready to be joined with the involved genes).

  Server-side caches are implemented by means of Pandas dataframes, where information is already available in memory, indexed, joined (if needed) and ready to be further manipulated. Since data ordering is also critical to the production of the final reports, the cached data is also kept ordered w.r.t. the most used sorting options.
- **Server-side threading**. Further response latency reduction is achieved by threading, enabling the parallelization of the processes needed for the construction of the form responses (parallelized tasks are shown in Figure 6 as vertically aligned boxes and will be discussed in detail in Section 6.2 for different use cases).

- **Client-side caching and chunking**. User requests inevitably produce a high number of results, typically close to 20,000 records and beyond. Trying to visualize this data without any pre-processing would lead to an inefficient plot rendering, resulting in a bad user experience and compromising the platform functionality. The solution to this problem has been the division of the results of each DAM in multiple chunks (this process is done client-side) and the exploitation of client-side result caching. Each chunk includes the information already organized in such a way to be readily visualized.

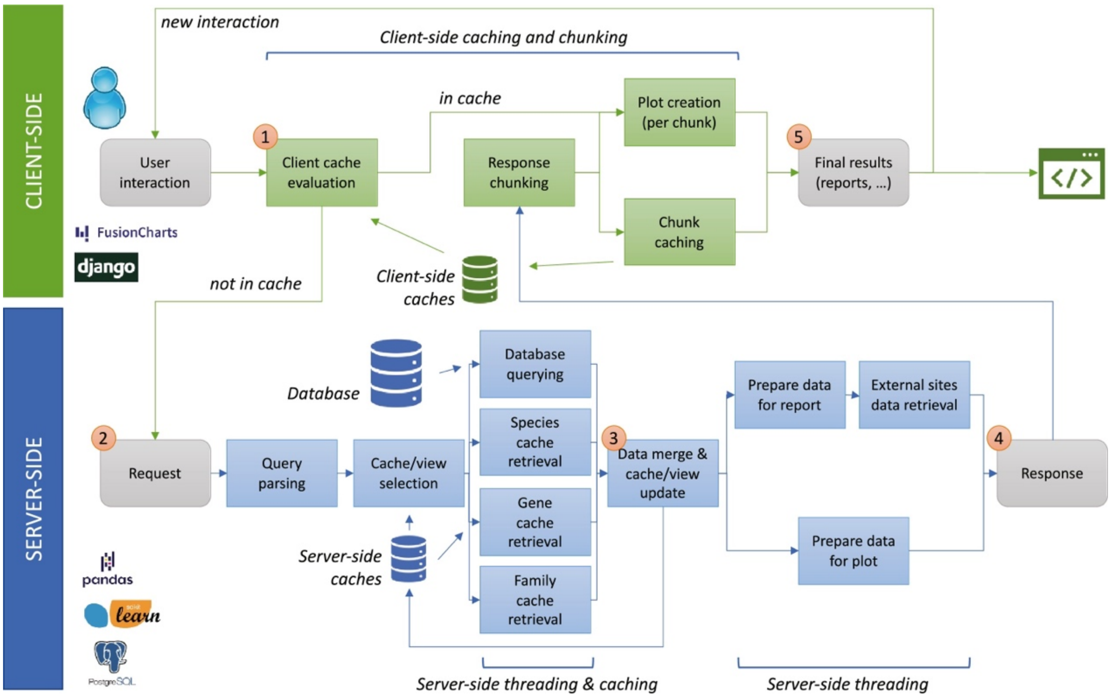


Figure 6: VarNuCopy architecture schema detailing client-side (upper part) and server-side (lower part) operations.

6.2. Data requests processing

In this section we describe how the three kinds of data requests are dealt with.

Use case 1: DAM request. The user requests an analysis on one of the available DAMs (for example, the Genes Copy Number Quantity for TP53 gene in cancer-prone species). The following are the performed steps (please also consider Figure 6, in which the numbers within the circles reference the step numbers):

1) after the parameters have been selected (in this case, gene TP53 and the cancer-prone group), the client performs "*Client cache evaluation*" in order to determine if the data needed to produce the results are already in its cache. If the results

are in the client-side caches, the execution continues in the client alone with “Plot creation” (see step 4) and concludes with step 5;

2) if they are not (which is the standard case), the full client-server interaction is needed: a request is sent to the server, which parses the query, then “cache/view selection” is performed in order to determine how the views stored in the server-side caches can help to obtain the final results. In general, data requests are dealt by rewriting the query according to the materialized views associated to the cache repositories of interest (query rewriting under views approach) (Halevy., 2001). In our case, a thread will access the “Gene” cache to retrieve the materialized join between the `Gene`, `CopyNumber` and `Species` tables, then a selection will be performed to restrict the result to the gene and species of interest. In the case where more than one cache is needed, different threads will access each of them;
3) after the data are merged and the caches are updated, other threaded operations are performed in order to efficiently send the response back to the client (lower right part of figure). One thread deals with the construction of the structures necessary for formatting report table data and retrieving additional resources: it computes statistical measures and retrieves data from external sites. The second one is dedicated to preparing the data for the plot (prepare the correct data structures to pass to the client, sort their contents, etc.);
4) the response is sent back to the client: first of all “Response chunking” is performed on the results, dividing the large data of the response in smaller chunks. Then, while the first chunk is visualized in a plot (“Plot creation” step), another thread (“Chunk caching”) performs the storage of the other chunks in the client-side cache: the other cached results will be in case subsequently used without the need to send new requests to the server;
5) Finally, the plot and the additional information are shown to the user.

Use case 2: custom query. In the case of an advanced search (custom query), query parsing, cache and data merging proceeds in the following way (then, processing proceeds as discussed in steps 3-5 for Use Case 1):

1) the `show` clause list is scanned, and the different schema attributes are identified based on their names (there is no ambiguity since the schema attributes names are unique);
2) the involved tables are identified (i.e., minimum subset of tables containing all the required attributes, multiple instances of the same table are not allowed)

and caches are accessed in order to retrieve their join. With respect to Use case 1, in the case of custom queries server-side parallelism for step 2 is even higher: a thread that retrieves the cached data is created for each needed materialized view in order to make the results available for the subsequent phases in the shortest amount of time;

3) filtering conditions in the `filter/exclude` clauses are applied to the results.

For instance, for queries Q_A and Q_B (Section 5.3), the required information contained in the `Gene`, `CopyNumber` and `Species` tables is extracted, already in joined form, from the "Gene" cache, then filtering conditions are applied.

Use case 3: plot interaction. Finally, let us consider the iterative nature of the analyses performed on the platform, where a simple interaction by the user on a report will trigger new requests to the platform (note the "new interaction" loop in the upper part of the figure). For instance, when the user receives the report, the data used for the plot (showing the genes CNV) has already been chunked and cached in the client-side cache. Therefore, when the user interacts with the plot in order to focus on a single gene (e.g., TP53), the answer is almost instantaneous: being the data already in the cache, the updated plot with the highlighted gene and its information can be directly created and shown without passing further requests to the server.

7. Performance evaluation

In this section we present a selection of the tests we performed to quantify the performance of the platform. In particular, we measured the overall response time for typical requests with and without server-side caching and threading optimizations (Section 7.1). Then, we deepened the analysis with additional tests designed to evaluate the benefits provided by server-side caching (Section 7.2). Finally, we consider user and client interactions benefits given by client-side optimizations (Section 7.3). All tests were performed on a server with AMD Ryzen 1920X CPU, 32 GB RAM, 256 GB SSD and 2 TB HDDs.

7.1. Overall response time

Figure 7 shows two response time comparisons. The first one (left side, blue bars) is relative to the mean response time measured for standard interaction/querying on the genes and species exploration forms (similarly to the Use case 1 discussed in previous section), with and without threading. As depicted, threading optimizations almost produce a 4x speedup, making all interactions almost real-time. The second comparison (on the right side, orange bars) quantifies the advantages offered by server-side caching, in particular in the complex case of custom queries (as in the Use case 2 discussed in previous section). A large number of queries are avoided on the database, leading to a sharp reduction in time response. In particular, the required time depends on the number of tables involved in the user query. For this test, we considered the worst case (all tables involved): the measured speedup reaches 10x, going from 25 seconds (no caching) to less than 3 seconds (caching active).

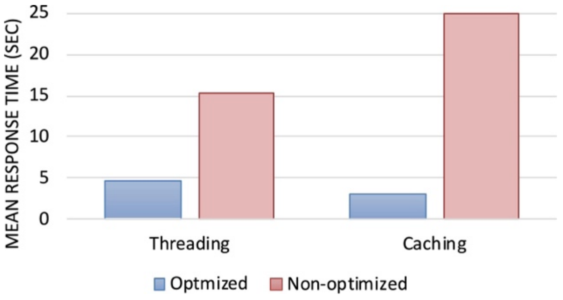


Figure 7: Response time tests (mean time in seconds).

7.2. Server-side caching optimization benefits

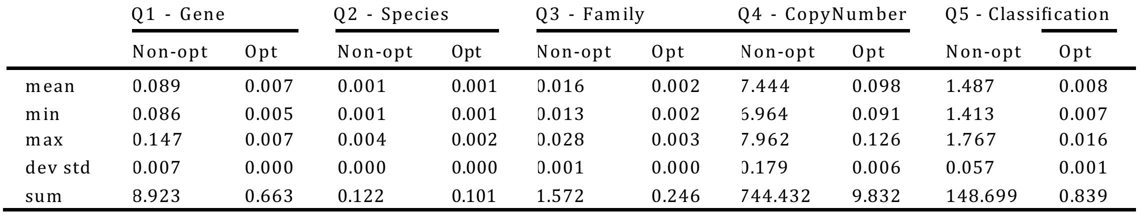
We designed a benchmark composed of 8 different queries, representative of different data retrieval operations performed in response to custom querying or DAM user interactions. We considered only the time required to fetch the data and compare optimized (server-side caching on, time required to retrieve data from the caches) and non-optimized time (time required to fetch the data directly from the relational database and put it in dataframe format for further elaboration) on 100 executions for each query. Table 1 below lists the complete query specifications.

	Involved tables (Join)					Filtering	Aggregation	
	Gene	Species	Family	CopyNumber	Classification		On	Aggr funct
Q1	V							
Q2		V						
Q3			V					
Q4				V				
Q5					V			
Q6	V	V		V		tumor suppressor		
Q7	v	V		V				
	qty (desc)							
Q8	V	V		V			gene name	avg(qty)
	avg(qty) desc							

	Attributes in output	Custom query syntax
Q1	*	show: (ID_gene, gene_name, is_oncogene, is_tumor_suppressor)
Q2	*	show: (ID_species, species_name, is_long_lived, is_cancer_prone, metabolism, avg_weight,
Q3	*	show: (ID_family, family_name)
Q4	*	show: (ID_gene, ID_species, qty)
Q5	*	show: (ID_gene, ID_family)
Q6	species_name, gene_name, qty	show: (species_name, gene_name, qty) filter: (is_tumor_suppressor=TRUE)
Q7	species_name, gene_name, qty, cancer_rate	n/a
Q8	gene_name, avg(qty)	n/a

Table 1: Test queries specifications: involved tables, filtering, aggregation (aggregation column and applied aggregate function), sorting operations (top), attributes required in output (* stands for all attributes in tables), and custom query syntax (bottom).

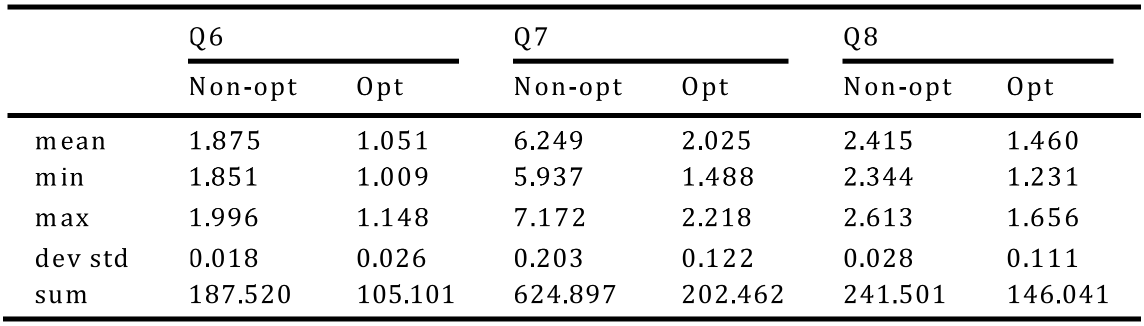
Table 2, instead, shows the complete statistics (mean, minimum, maximum, standard deviation and sum) for the 100 executions of the first five queries (Q1-Q5): such queries are designed to plainly retrieve the information on the genes (Q1), species (Q2), families (Q3), copy numbers (Q4) and classifications (Q5) available in the platform data, without specific elaborations (joins, selections, etc.). By looking at the mean response time, we can see that the caching strategies enable consistent time savings on delivering the required data (of 13.46x, 1.21x, 6.38x, 75x and 177x for Q1-Q5, respectively).



	Q1 - Gene		Q2 - Species		Q3 - Family		Q4 - CopyNumber		Q5 - Classification	
	Non-opt	Opt	Non-opt	Opt	Non-opt	Opt	Non-opt	Opt	Non-opt	Opt
mean	0.089	0.007	0.001	0.001	0.016	0.002	7.444	0.098	1.487	0.008
min	0.086	0.005	0.001	0.001	0.013	0.002	6.964	0.091	1.413	0.007
max	0.147	0.007	0.004	0.002	0.028	0.003	7.962	0.126	1.767	0.016
dev std	0.007	0.000	0.000	0.000	0.001	0.000	0.179	0.006	0.057	0.001
sum	8.923	0.663	0.122	0.101	1.572	0.246	744.432	9.832	148.699	0.839

Table 2: Response time statistics for server-side caching tests (queries Q1-Q5)

Finally, Table 3 shows the response time statistics for queries Q6-Q8, which represent more elaborate requests (involving multiple tables with joins and sorts): species having tumor suppressor genes with relative CNVs (Q6), genes, CNVs and cancer rate information (ordered by decreasing CNVs) for each species (Q7), average copy number for each gene (results ordered by decreasing average CNV) (Q8). Note that queries Q7 and Q8 also involve aggregations and sorting and are therefore representative of elaborations which go beyond custom query requests but are constantly automatically performed by the system in response to user interactions.



	Q6		Q7		Q8	
	Non-opt	Opt	Non-opt	Opt	Non-opt	Opt
mean	1.875	1.051	6.249	2.025	2.415	1.460
min	1.851	1.009	5.937	1.488	2.344	1.231
max	1.996	1.148	7.172	2.218	2.613	1.656
dev std	0.018	0.026	0.203	0.122	0.028	0.111
sum	187.520	105.101	624.897	202.462	241.501	146.041

Table 3: Response time statistics for server-side caching tests (queries Q6-Q8)

Also in this case, VarNuCopy caching strategies are able to deliver good average speedups (1.78x, 3.08x and 1.65x, respectively for Q6-Q8). Indeed, Figure 8 shows an overall mean response time comparison between all queries.

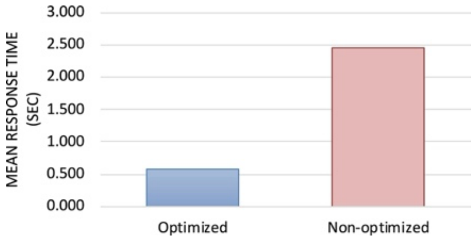


Figure 8: Mean response time comparison between all queries

As a concluding remark, the queries have different selectivity (as they are designed to retrieve small and large parts of the dataset) and the data volume retrieved by the 100 executions of the queries is 279.46 MB (Q1), 5.98 MB (Q2), 101.68 MB (Q3), 11.11 GB (Q4), 20.04 GB (Q5), 35.15 GB (Q6), 56.84 GB (Q7) and 155.13 MB (Q8).

7.3. Client-side optimization benefits

We concluded our tests with two experiments designed to evaluate the benefits of the client-side optimizations (caching, chunking) and the interactive nature of the reports. We considered a situation similar to Use case 3 (Section 6.2) where a user interacts on a received report plot: to search for a gene (TP53) within the DAM showing the genome of the *Loxodonta africana* species. The non-optimized scenario requires server access (17 ms) also to perform the gene search; moreover, result rendering is also slightly faster in the optimized scenario, since the data in the chunks of the client cache are already organized for visualization. The total time in the optimized scenario is nearly halved (33.5 ms instead of 60 ms): this would allow the server to manage more than 3300 of such user requests per minute, instead of less than 1800 in the non-optimized case. Finally, in our last test we quantified the benefits given by the interactive nature of the reports returned by the platform. More specifically, as described in the previous test, we considered the situation where the user searches for a genome of a species, focusing then on a specific gene: we measured the overall time (including user interaction time) required in the optimized case (with client-side optimizations and interactive reports) and non-optimized (where reports do not allow to trigger new searches and therefore a new search for the gene has to be submitted from scratch).

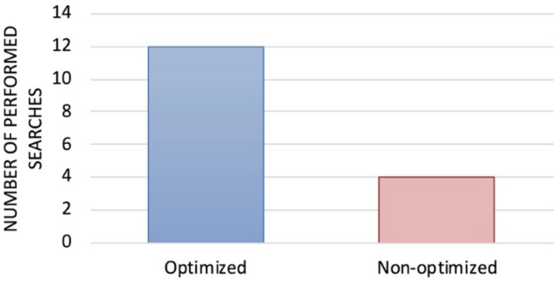


Figure 9: User interaction test: number of searches that can be performed in 1 minute time window (including user interaction time)

Figure 9 shows that in this case user productivity is nearly tripled in the optimized version, allowing a much larger number of analyses in less time.

8. Comparison with related works

From a pragmatic point of view, one of the hardest challenges that scientists have to face in the NGS (Next Generation Sequencing) era, is the handling and the organization of a huge volume of information. Since the availability of genomic data is constantly increasing in time, refined computational methods are needed to manage them, especially through the construction and the use of (bio)informatics databases. Visualization techniques are essential, even if they are currently mostly applied during the report stage, at the very end of the data science workflow (Schmidt., 2020). In this context, CNVs have been studied in several research works, highlighting their importance in many biological processes (Feuk et al., 2006; Stratton et al., 2009). Nowadays, a unique and comprehensive catalogue of animal gene copies across the genome of different organisms doesn’t exist; nonetheless, an extended literature regarding CNVs events, especially related to human (Bragin et al., 2014) or human related diseases is available (Qiu et al., 2012). In particular, CNVD website collects and organizes information regarding copy number variations events related to human diseases manually mined from more than 6000 research papers, with the aim of providing a platform to study the role of CNVs in many diseases based on literature research (Qiu et al., 2012). In 2009, Chen and collaborators, developed CNVVdb (Chen et al., 2009), a database of putative copy number variations across 16 vertebrate genomes, built using the pairwise alignment of sequences. However, both CNVD and CNVVdb, besides the low number of analyzed species, do not contain, for example, any comparative statistical analysis based on the number of gene copies, or advanced graphical representations of the results. VarNuCopy is the first tool that can compare the CNVs landscape of multiple species, both for model and non-model organisms in biomedicine. Using a comparative approach, we developed this new web-platform which includes the variation landscape of genes copies among the genome of 233 organisms, combining, for 24 of them, the gene presence in multiple or deleted copies, together with cancer related parameters, such as longevity, mass, and metabolism. To our knowledge, VarNuCopy is the first tool allowing a multi-species gene copy number comparison. Thanks to the EDA approach and the DAMs analysis, our platform easily performs and retrieves statistical information on the number of gene copies among different target groups (cancer-prone and -resistant organism, long- and short-living species), guiding the users to the discovery of known and/or unknown molecular mechanisms related to the cancer predisposition of a species. Building a personalized research path, following the exploratory analysis paradigm, it is possible to combine simple genomic features of a species with advanced queries based exclusively on both cancer incidence and genomic CNVs. Indeed, the final aim of the platform is to deeply investigate the genes copy number relationship among different species, and eventually retrieve new biomarkers involved in the oncogenic processes. Among related biogenetic research tools, Ensembl "gene gain/ loss" feature (Herrero et al., 2016) is probably one of the available tools which, in some ways, is most similar to the platform we propose. In fact, it can map the number of copies of each gene in multiple species, providing the number of extant homologues. Other tools such as EnsemblCNV (Howe et al., 2021) include human-readable and machine-readable information about the functions of genes from different species; they can be used together only through APIs and, despite being a powerful resource in the field, they present a certain number of shortcomings:

I. typically, slow and time-expensive research speed;
II. "heavy" graphical style, often difficult to read, which can make the user interaction less friendly and less immediate;
III. lack of proper data visualization techniques.

VarNuCopy is aimed to go beyond the current offerings of publicly available repositories in terms of data/feature availability, efficiency, and ease of use:

i. the client offers ad-hoc visualization facilities associated with convenient plots representing the data distribution. Reports are presented as single panels or multiple dashboards, but, in any case, are designed to be as easy-to-understand as possible, allowing scientists to look at data from multiple directions and studying different aspects to understand their structure;

ii. the scalable data science implementations include ad-hoc data structures for a faster and less expensive management of information, allowing real-time results and interaction;

iii. the EDA tools enable recursive searches directly from the output charts and visualizations, and personalized and advanced searches through which the user can specify complex search patterns on the database;

iv. the peculiar database and analysis features include, for a subset of mammals, the possibility to show and analyze their vital parameters, with the aim to correlate the gene number of copies with their cancer incidence rate. This is an exclusive novelty in the field of biological databases.

9. Conclusions and future works

One of the goals of information visualization when dealing with large and rich datasets is to give the observer sets of query results, which may themselves be large, as quickly as possible, thus effectively highlighting the most relevant ones. VarNuCopy leverages statistical measures related to the genome CNVs to quantify gene relevance and supports different types of plots and ranking criteria for their effective visualization. For example, the DAM model "Genes CNVs for Species Group" is associated with the representation of two different box plots coupled with stat and *p-value* measures, to give the user the opportunity to straightforwardly check if the CNV distribution is able to significantly discriminate the gene within two different categories (e.g. cancer-prone vs cancer resistant species). As pinpointed through the paper, the user who wishes to find interesting patterns can explore the platform personalizing his research process that can involve repeated cycles of visualization and interactive commands. In other words, the human and the computer system are solving a problem as two different actors of the same play: the former looks for patterns thinking which presentation may be most informative, while the latter is crunching through large quantities of data to develop the best informative visual representation.

In the future, we plan to expand VarNuCopy capabilities by:

- extending the biogenetic database with further related publicly available data sources concerning, for instance, biological pathways and perturbations;

- incorporating machine learning algorithms for gene classification;
- studying data mining approaches for the discovery of such parameters that can cooperate in the cancer onset.

Declaration of competing interest

The authors declare that they have no known competing finan- cial interests or personal relationships that could have appeared to influence the work reported in this paper.

References:

(Paper)

Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., Taccioli, C., & Vischioni, C. (2020, September). VarCopy: a Visual Exploratory Data Analysis Platform for Copy Number Variation Studies. In 2020 24th International Conference Information Visualisation (IV) (pp. 391-396). IEEE.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2(2), 131.

Birkholz, J. M., Bakhshi, R., Harige, R., Steen, M. V., & Groenewegen, P. (2012, December). Scalable analysis for large social networks: the data-aware mean-field approach. In International Conference on Social Informatics (pp. 406-419). Springer, Berlin, Heidelberg.

Bragin, E., Chatzimichali, E. A., Wright, C. F., Hurles, M. E., Firth, H. V., Bevan, A. P., & Swaminathan, G. J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic acids research, 42(D1), D993-D1000.

Breve, B., Caruccio, L., Cirillo, S., Deufemia, V., & Polese, G. (2020). Visualizing Dependencies during Incremental Discovery Processes. In EDBT/ICDT Workshops.

Buoncristiano, M., Mecca, G., Quintarelli, E., Roveri, M., Santoro, D., & Tanca, L. (2015). Database challenges for exploratory computing. ACM SIGMOD Record, 44(2), 17-22.

Buzan, T. (2005). Mind map: The ultimate thinking tool. London: Thorsons.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. Computers, Environment and Urban Systems, 51, 70-82.

Card, M. (1999). Readings in information visualization: using vision to think. Morgan Kaufmann.

Catarci, T. (2000). What happened when database researchers met usability. Information Systems, 25(3), 177-212.

Chen, F. C., Chen, Y. Z., & Chuang, T. J. (2009). CNVVdb: a database of copy number variations across vertebrate genomes. Bioinformatics, 25(11), 1419-1421.

Di Blas, N., Mazuran, M., Paolini, P., Quintarelli, E., & Tanca, L. (2017). Exploratory computing: a comprehensive approach to data sensemaking. International Journal of Data Science and Analytics, 3(1), 61-77.

Du, J., Miller, R. J., Glavic, B., & Tan, W. (2017). DeepSea: Progressive Workload-Aware Partitioning of Materialized Views in Scalable Data Analytics. In EDBT (pp. 198-209).

Elghandour, I., & Aboulnaga, A. (2012, May). ReStore: reusing results of MapReduce jobs in pig. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 701-704).

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85-97.

Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. Science, 331(6018), 705-708.

Halevy, A. Y. (2001). Answering queries using views: A survey. The VLDB Journal, 10(4), 270-294.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., ... & Flicek, P. (2016). Ensembl comparative genomics resources. Database, 2016.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1991). Fundamentals of exploratory analysis of variance (Vol. 261). John Wiley & Sons.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., ... & Flicek, P. (2021). Ensembl 2021. Nucleic acids research, 49(D1), D884-D891.

Hoaglin, D. C. (1992). Perspectives on Contemporary statistics (Vol. 21). Mathematical Assn of Amer.

Li, F., & Jagadish, H. V. (2012). Usability, Databases, and HCI. IEEE Data Eng. Bull., 35(3), 37-45.

Lomonaco, V., Martoglia, R., Mandreoli, F., Anderlucci, L., Emmett, W., Bicciato, S., & Taccioli, C. (2014). UCbase 2.0: ultraconserved sequences database (2014 update). Database, 2014.

Ma, X., Hummer, D., Golden, J. J., Fox, P. A., Hazen, R. M., Morrison, S. M., ... & Meyer, M. B. (2017). Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. ISPRS International Journal of Geo-Information, 6(11), 368.

Myers, R. Espinosa, C.S. Parr, T. Jones, G.S. Hammond, T.A. Dewey, ADW - Animal Diversity Web, http://animaldiversity.org/.

McPadden, J., Durant, T. J., Bunch, D. R., Coppi, A., Price, N., Rodgerson, K., ... & Schulz, W. L. (2019). Health care and precision medicine research: analysis of a scalable data science platform. Journal of medical Internet research, 21(4), e13043.

Naderi, A., Liu, J., & Hughes-Davies, L. (2010). BEX2 has a functional interplay with c-Jun/JNK and p65/RelA in breast cancer. Molecular cancer, 9(1), 1-18.

Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.

Novak, J. D., & Cañas, A. J. (2008). The theory underlying concept maps and how to construct and use them.

Liu, Y., Sun, J., & Zhao, M. (2017). ONGene: a literature-based database for human oncogenes. Journal of Genetics and Genomics, 44(2), 119-121.

Qiu, F., Xu, Y., Li, K., Li, Z., Liu, Y., DuanMu, H., ... & Li, X. (2012). CNVD: text mining-based copy number variation in disease database. Human mutation, 33(11), E2375-E2381.

Schmidt, J. (2020, February). Usage of Visualization Techniques in Data Science Workflows. In VISIGRAPP (3: IVAPP) (pp. 309-316).

Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. Sebastopol, CA: O'Reilly.

Serrano, M. (2016). Unraveling the links between cancer and aging. Carcinogenesis, 37(2), 107-107.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719-724.

Sulak, M., Fong, L., Mika, K., Chigurupati, S., Yon, L., Mongan, N. P., ... & Lynch, V. J. (2016). TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. elife, 5, e11994.

Talia, D. (2019). A view of programming scalable data analysis: from clouds to exascale. Journal of Cloud Computing, 8(1), 1-16.

Tang, W., Wilkening, J., Desai, N., Gerlach, W., Wilke, A., & Meyer, F. (2013, October). A scalable data analysis platform for metagenomics. In 2013 IEEE International Conference on Big Data (pp. 21-26). IEEE.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the Condor experience. Concurrency and computation: practice and experience, 17(2-4), 323-356.

Tian, X., Azpurua, J., Hine, C., Vaidya, A., Myakishev-Rempel, M., Ablaeva, J., ... & Seluanov, A. (2013). High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. Nature, 499(7458), 346-349.

Zhao, M., Kim, P., Mitra, R., Zhao, J., & Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic acids research, 44(D1), D1023-D1031.

Tufte, E. (2001). The visual display of quantitative information.

Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).

Wang, Y., Wang, G., Tan, X., Ke, K., Zhao, B., Cheng, N., ... & Liu, J. (2019). MT1G serves as a tumor suppressor in hepatocellular carcinoma by interacting with p53. Oncogenesis, 8(12), 1-11.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., ... & Flicek, P. (2020). Ensembl 2020. Nucleic acids research, 48(D1), D682-D688.

Zarrei, M., Burton, C. L., Engchuan, W., Young, E. J., Higginbotham, E. J., MacDonald, J. R., ... & Scherer, S. W. (2019). A large data resource of genomic copy number variation across neurodevelopmental disorders. NPJ genomic medicine, 4(1), 1-13.

References:

(Chapter II preface)

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., ... & Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. Nature, 463(7283), 899-905.

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., ... & Stratton, M. R. (2010). Signatures of mutation and selection in the cancer genome. Nature, 463(7283), 893-898.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85-97.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. Journal of genetics, 92(1), 155-161.

Vazquez, J. M., & Lynch, V. J. (2021). Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. Elife, 10, e65041.

Vischioni, C., Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2022). Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research. Big Data Research, 27, 100298.

Chapter III:

miRNAs Copy Number Variations repertoire as hallmark indicator of cancer species predisposition.

Chiara Vischioni[1], Fabio Bove[2], Federica Mandreoli[2], Riccardo Martoglia[2], Valentino Pisi[2], Cristian Taccioli[1]

[1] Department of Animal Medicine, Production and Health, University of Padova, Italy.

[2] FIM, University of Modena and Reggio Emilia, Modena, Italy.

Genes containing CNVs are related to several molecular pathways, and are involved in diseases such as cancer, metabolic and neurodegenerative disorders. As already mentioned in the previous chapters, it is already well known that an altered number of tumor suppressors/oncogenes can protect/expose a species to a lower/higher cancer incidence rate, but previous works have often focused just on the comparative analysis of candidate genes which were selected *a priori*. Leveraging on the new online tool that I have recently developed (and presented in **Chapter II**), I analyzed the entire genomic CNVs spectra of multiple mammals, in order to identify which are the best molecular targets able to statistically discriminate between cancer -prone and -resistant species. Contrary to what is usually done, I did not sub-select only for those genes that are already known to be involved in tumor development, maintenance and progression, such as tumor suppressors or oncogenes, but I kept the entire species genomic CNVs landscape, including noncoding RNAs (ncRNAs), that are often discarded from this kind of analysis. Indeed, ncRNAs, and microRNAs (miRNAs) in particular, may represent non-canonical genomic features that might contribute to genome stability. For the first time, I proposed miRNAs gene family as one of the most important key players in determining the cancer predisposition of a species, suggesting that an altered microRNAs copy number patterns may help lower its cancer risk. In this context, I have classified miRNAs as oncogenes or tumor suppressors along the discussion, justifying their specific behavior using different examples described in literature. Given the double side behavior of microRNAs genes, which can act both as oncogenes or TSGs, I have firstly performed the statistical analysis to identify the target, and secondly verified its possible functioning as onco-miRNA, or onco-suppressor. This allowed me to detect, for example, amplifications of microRNAs known as oncogenes in cancer-resistant species, or conversely, suppressors microRNAs in cancer-prone species. Moreover, this kind of approach allows me to include as significant results those microRNAs not usually known as cancer related genes, but also to straighten the hypothesis that, indeed, microRNAs genes function as hallmarks and discriminants of a species cancer predisposition.

This paper is currently being prepared for submission.
The following pages are reproduced as an update of the first version of the article, which has been posted on BioRxiv: Vischioni, C., Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2021). miRNAs Copy Number Variations repertoire as hallmark indicator of cancer species predisposition. *bioRxiv*.

The supplemental material for this paper is included as Appendix A.

Abstract

Aging is one of the hallmarks of multiple human diseases, including cancer. However, the molecular mechanisms associated with high longevity and low cancer incidence percentages characterizing long-living organisms have not been fully understood yet. In this context, we hypothesized that variations in the number of copies (CNVs) of specific genes may protect some species from cancer onset. Based on the statistical comparison of gene copy numbers within the genomes of cancer -prone and -resistant organisms, we identified novel gene targets linked to the tumor predisposition of a species, such as CD52, SAT1 and SUMO protein family members. Furthermore, for the first time, we were able to discover that, considering the entire genome copy number landscape of a species, microRNAs (miRNAs) are among the most significant gene families enriched for cancer progression and predisposition. We identified through bioinformatics analysis, several alterations in miRNAs copy number patterns, represented by duplication of miR-221, miR-222, miR-21, miR-372, miR-30b, miR-30d and miR-31 among others. Therefore, our analysis provides the first evidence that an altered copy number miRNAs signature is able to statistically discriminate species more susceptible to cancer than those that are tumor resistant, helping researchers to discover new possible therapeutic targets involved in tumor predisposition.

Keywords: CNVs; cancer resistance; miRNAs

1. Introduction

Aging is one of the hallmarks of cancer insurgence, being considered also one of its possible related risk factors (Serrano., 2016). Therefore, it is probable that, in order to maintain high longevity rate, some species have developed intrinsic molecular mechanisms that protect them from cancer onset or development (Tian et al., 2017). Along this assumption, as two binary parallel lines, those organisms that live longer should theoretically possess a higher risk of cancer occurrence. Nevertheless, considering different species, according to Peto's Paradox theory (Peto et al., 1975), the body size of an organism and/or its lifespan expectation are not directly correlated with the species percentage of cancer incidence. After more than 40 years of research, the solution to this puzzling paradox is still an open challenge to be solved. For example, despite its small size, the naked mole rat, being able to live more than 30 years, is, to date, the longest-living member of the rodent family. Several studies highlighted that, besides the delayed aging, this species also shows the capacity to resist spontaneous and experimentally induced tumorigenesis (Buffenstein, 2008; Kim et al., 2011; Liang et al., 2010; Seluanov et al., 2009). Conversely, in mice, the cancer-related mortality can reach 90%, coupled with a species maximum life expectancy of four-five years (Lipman et al., 2004). The long-living *Myotis lucifugus* bat species, has been recently identified as prospective organism for comparative cancer research (Boddy, Harrison, et al., 2020). Given their extended life-span (Wilkinson & Adams, 2019), it has been suggested that bats develop very low cancers events, as confirmed from different pathological studies performed in different area of the world (Tollis, Schiffman, et al., 2017; Wang et al., 2011). The elephant has been pinpointed as another cancer resistant species (Abegglen et al., 2015), with a cancer incidence rate value considerably low compared to the human one for example (~5% vs approximately 22%) (Ferlay et al., 2015). Interestingly, various authors recently reported that the genome of the African elephant encodes multiple copies of the TP53 gene, also known as the "guardian of the genome stability". This amplification could be at the basis of its anti-cancer and longevity mechanisms, leading, for example, to increased levels of apoptotic events in response to DNA damage (Abegglen et al., 2015; Sulak et al., 2016). Indeed, according to Caulin and Maley (2011), the genome of large long-living organisms can reveal an altered number of tumor suppressors and oncogenes (in multiple or reduced copies), that might represent a possible mechanism underlying their capacity of exceeding the threshold of cancer onset, despite their phenotypic predisposition such as size and longevity (Caulin & Maley, 2011). Copy Number

Variations (CNVs) are duplications or deletions of genomic regions which can be associated with phenotypic alterations, including tumorigenic diseases (Feuk et al., 2006). In particular, a variation in the gene copy numbers can activate or inactivate tumor suppressors and oncogenes, leading to the development of cancer (Stratton et al., 2009). Within this framework, long-living animals have to rely on compensatory mechanisms to suppress and prevent cancer progression, that can be straightened by different molecular and genomics mechanisms such as altered gene copy numbers that increase the number of tumor suppressors paralogues or reduce copies of oncogenes (Tollis, Boddy, et al., 2017; Tollis et al., 2020). As previously mentioned, mammals have evolved lifespan and cancer incidence rates which vary among species (Boddy, Abegglen, et al., 2020), however mechanisms underlying these differences are still unclear. In order to test the hypothesis that genomic CNVs are related to the cancer incidence rate of a species, we compared the entire genomic copy number landscape of 9 different mammals (5 cancer resistant and 4 cancer prone organisms), to retrieve among their genomes, which target genes are able to significantly discriminate between these two groups.

2. Material and Methods

2.1. Data collection

In a comparative biology framework, taking advantage of the advancement of the Next Generation Sequences (NGS) technologies, it is now possible to investigate and speculate about new factors that can control longevity and cancer susceptibility. According to the hypothesis that positively selected CNVs tend to recur during cancer progression (Beroukhim et al., 2010; Bignell et al., 2010), but also during the evolution (Iskov et al., 2012), we have recently developed VarNuCopy database, a tool unique of its kind, which collects the CNVs landscape for multiple organisms, with the aim to compare patterns of copy number changes across the genome of different species (Vischioni et al., 2022). We used a homemade script written in Perl 5.14 and Python 3 in order to download the CNV data from Ensembl comparative genomics resources (http://www.ensembl.org) (Howe et al., 2021), an ideal system to perform and support vertebrate comparative genomic analyses, given the consistency of gene annotation across the genomes of different vertebrate species. In this context, we leveraged Ensembl's “gene gain/loss tree” feature, which maps the number of copies of extant homologous gene for each species, as a taxonomic tree view (Herrero et al., 2016). Remarkably, these data are estimated through CAFE, a computational tool useful to study gene family evolution, which, when calculating CNVs data, takes into account a priori the species phylogenetic tree (De Bie et al., 2006; Herrero et al., 2016). The Perl API script provided by the Ensembl website was used to access the genomic databases and used to download all the available CNVs data. We encoded a new Python script in order to format the CNVs data counts as a readable tab delimited matrix, useful to perform the subsequent analysis.

2.2. Statistical comparison

Using a comparative approach, we analyzed the variation landscape of genes copies among the genome of 9 organisms sub-set in two categories: "cancer resistant" (*Heterocephalus glaber* [Hg], *Nannospalax galili* [Ng], *Dasypus novemcinctus* [Dn], *Loxodonta africana* [La], *Myotis lucifugus* [Ml]), and "cancer prone" (*Mus musculus* [Mm], *Rattus norvegicus* [Rn], *Canis familiaris* [Cf], and *Homo sapiens* [Hs]) species. We classified as "cancer resistant" those species that, based on the literature review, are known to possess a low cancer incidence rate. Conversely, "cancer prone" organisms, were referred to those for which the percentage of tumors found in a certain number of necropsies is known to be high. To determine whether microRNAs CNVs independently contribute to the variation in cancer incidence percentages among our species, we applied a linear regression model through PGLS R package (Orme., 2013), in order to check for potential bias due to species phylogeny or population structure (Figure 1D; Supplementary Data S1). The phylogenetic tree included in the analysis was derived from VertLife (Upham et al., 2019) and created through the Interactive Tree of Life web-tool (Figure 1C) (Letunic & Bork, 2019), while cancer rate data were collected from different recently published literature (Boddy, Abegglen, et al., 2020; Boddy, Harrison, et al., 2020; Lagunas-Rangel, 2018; Seluanov et al., 2018; Sulak et al., 2016; Wang et al., 2011). We performed a statistical comparison between the CNVs of the two different species groups, cancer prone and resistant organisms, with the aim to identify new possible gene targets able to discriminate between the two categories. Thus, a statistical unpaired 2-group Wilcoxon test was performed using R.3.1.1 (https://www.r-project.org/), to compare their entire CNVs spectra. Data processing, plots, and statistical tests were performed with R.3.1.1 (www.cran.r-project.org) and RStudio 1.4.1717 (https://www.rstudio.com/). Figures were made using the ggplot2 R package, in association with different R Shiny apps such as BoxPlotR, PlotsOfData, and ClustVis (Metsalu & Vilo, 2015; Postma & Goedhart, 2019; Spitzer et al., 2014).

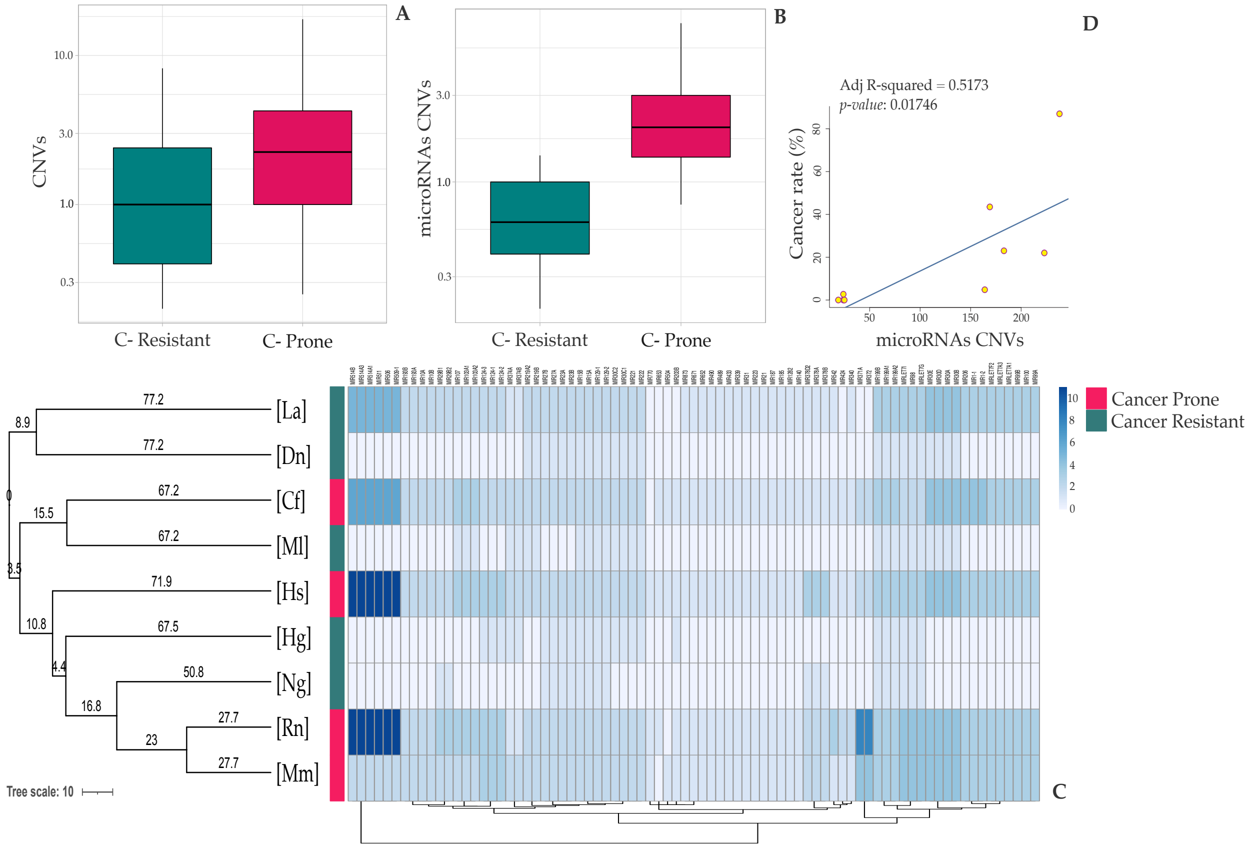


Figure 1: **CNVs landscapes comparisons**: **A,** Boxplot of the distribution of significant genes CNVs in cancer prone vs cancer resistant species. **B,** Boxplot of the distribution of significant microRNAs CNVs in cancer prone vs cancer resistant species. Cancer resistant species are highlighted in green, cancer prone species in red. In the boxplots, the Y-axis scale has been changed to log one. The boxplots are built considering the average number of copies of each gene in the two different target groups. **C**, heatmap representing the MicroRNAs CNVs repertoire within the 9 analyzed species. [Hg]: Heterocephalus glaber; [Ng]: Nannospalax galili; [Dn]: Dasypus novemcinctus; [La]: Loxodonta Africana; [Ml]: Myotis lucifugus; [Mm]: Mus musculus; [Rr]: Rattus norvegicus; [Cf]: Canis familiaris; [Hs]: Homo sapiens. [Hg], [Ng], [Dn], [La] and [Ml] have been previously described as cancer resistant species. [Mm], [Rr], [Cf] and [Hs] are known to be "cancer prone" species. Phylogeny was inferred from VertLife (Upham et al., 2019), and created through the Interactive Tree of Life web-tool (Letunic & Bork, 2019). **D.**, PGLS correlating Cancer incidence rate ~ Number of significant microRNAs copies across the 9 species included in the analysis. The blue line represent a positive correlation between the two variables (Adjusted $R^2 = 0.5173$; p-value = 0.01746).

2.3. Pathway analysis

To determine if CNVs are enriched in specific gene families, we used Gene SeT AnaLysis Toolkit, a tool for the interpretation of lists of interesting genes which is commonly used to extract biological insights from targets of interest (Liao et al., 2019). The set of significant genes were tested for pathway associations using the hyper-geometric test for over-representation analysis (ORA) (Khatri et al., 2012) (Supplementary Table3 [S3]). We selected different pathway enrichment categories (KEGG: https://www.genome.jp/; Wikipathway: https://www.wikipathways.org; Reactome: https://reactome.org/; PANTHER: http://www.pantherdb.org/), considering over-represented those molecular networks with FDR significance level lower than 0.05, after a correction with Benjamini–Hochberg method. In this context, the ORA analysis was the preferred option among the others (e.g. gene set enrichment or network topology-based analysis) in order to obtain biological information underlying the significantly enriched genes, resulting in a reduction in the complexity of the data interpretation (Khatri et al., 2012).

3. Results

A two-group comparison was performed using a Wilcoxon rank sum test, in order to identify an existing distinction in terms of distribution in the number of gene copies between cancer-prone and cancer-resistant species. A list of the most significant hits (p-value < 0.05), including known tumor suppressors and oncogenes, is reported in Table 1 (See SupplementaryTable2 [S2] for the extended version). Our analysis, which exclusively considered the variation in number of gene copies within different species, was able to identify those genes involved in biological processes related to cancer development and maintenance.

Gene	*p*-value	Known_TS	Known_OG	References
CD52	0.007	NO	NO	Wang et al., 2020
SAT1	0.007	NO	NO	Thakur et al., 2019
MIR424	0.009	YES	NO	Xu et al., 2013
MIR372	0.010	NO	YES	Sun & Gao, 2018
DMD	0.014	YES	NO	Jones et al., 2021
EIF5	0.017	NO	NO	Spilka et al., 2013
MIR107	0.022	YES	YES	Turco et al., 2020
MIR124-1, MIR124-2, MIR124-3	0.022	YES	NO	Wang et al., 2014
SUMO2, SUMO3, SUMO4	0.024	NO	NO	Schneeweis et al., 2021
MIR506	0.029	YES	YES	Wen et al., 2015
MIR509-1	0.029	NO	NO	Zhai et al., 2012
MIR511	0.029	YES	NO	Squadrito et al., 2012
MIR514A1, MIR514A3, MIR514B	0.029	NO	NO	Ren et al., 2018
MIR378A	0.030	YES	NO	Chen et al., 2016
S100A16	0.030	NO	NO	Zhu et al., 2016
MBD1, MBD2, MBD3	0.031	NO	YES (MDB1)	Miremadi et al., 2007
FGFBP1	0.032	NO	NO	Zhang et al., 2019
FOXJ1	0.032	NO	NO	Xian et al., 2018
MIR1-1, MIR1-2	0.032	YES	NO	Nohata et al., 2011
MIR206	0.032	YES	NO	Zhang et al., 2013
MIR340	0.032	YES	NO	Wu et al., 2011
MIR542	0.032	NO	NO	Kureel et al., 2014
NUPR1	0.032	YES	NO	Cano et al., 2011
SELENOW	0.032	NO	NO	Yim et al., 2019
JUND	0.034	NO	YES	Elliott et al., 2019

Table 1*: *Genomic CNVs landscape comparison. Subset of 25 significant hits resulting from the unpaired 2-group Wilcoxon test (p-value < 0.05). The statistical comparison was made in order to identify those genes able to discriminate between cancer -prone and -resistant species groups, relying exclusively on the genomic copy number values. Some of these genes are already known to be tumor suppressor and/or oncogenes, whereas the others can

play pivotal roles in tumorigenesis events, and, for this reason, can be considered as targets to be further investigated and validated in the context of cancer development.

3.1. Best candidate cancer-related genes

The distribution of the average number of each gene copy plotted in Figure 1A highlights a difference between the two species categories, which appears even higher if we only refer to the microRNAs CNVs landscape (Figure 1B). Among the most significant genes presenting an altered number of copies we found CD52 (p-value = 0.007), SAT1 (p-value = 0.007), DMD (p-value = 0.014), EIF5 (p-value = 0.017), SUMO2, SUMO3, SUMO4 (p-value = 0.024), S100A16 (p-value = 0.030), MBD1, MBD2, MBD3 (p-value = 0.031), FGFBP1 (p-value = 0.032), FOXJ1 (p-value = 0.032), NUPR1 (p-value = 0.032), SELENOW (p-value = 0.032) and JUND (p-value = 0.034). Some of these, such as DMD, MDB1, NUPR1 and JUND have been already well described as tumor suppressors or oncogenes (Cano et al., 2011; Elliott et al., 2019; Jones et al., 2021; Miremadi et al., 2007), whereas the others do not officially belong to any of these two categories and they have been proposed as key regulators in biological processes such as cell proliferation, migration, and cancer development and progression (Müller et al., 2004; Peters et al., 2018; Schneeweis et al., 2021; Spilka et al., 2013; Thakur et al., 2019; J. Wang et al., 2020; Xian et al., 2018; Yim et al., 2019; Zhu et al., 2016). A Principal Component Analysis (PCA) of CNVs values of the 9 species reported in Figure 2A-B, showed a clear dichotomy between cancer -prone and -resistant groups, supporting the hypothesis that an altered landscape of CNV is able to discriminate between the two categories. To confirm these results, we performed another unsupervised clustering analysis using Euclidean distance (Figure 2C).

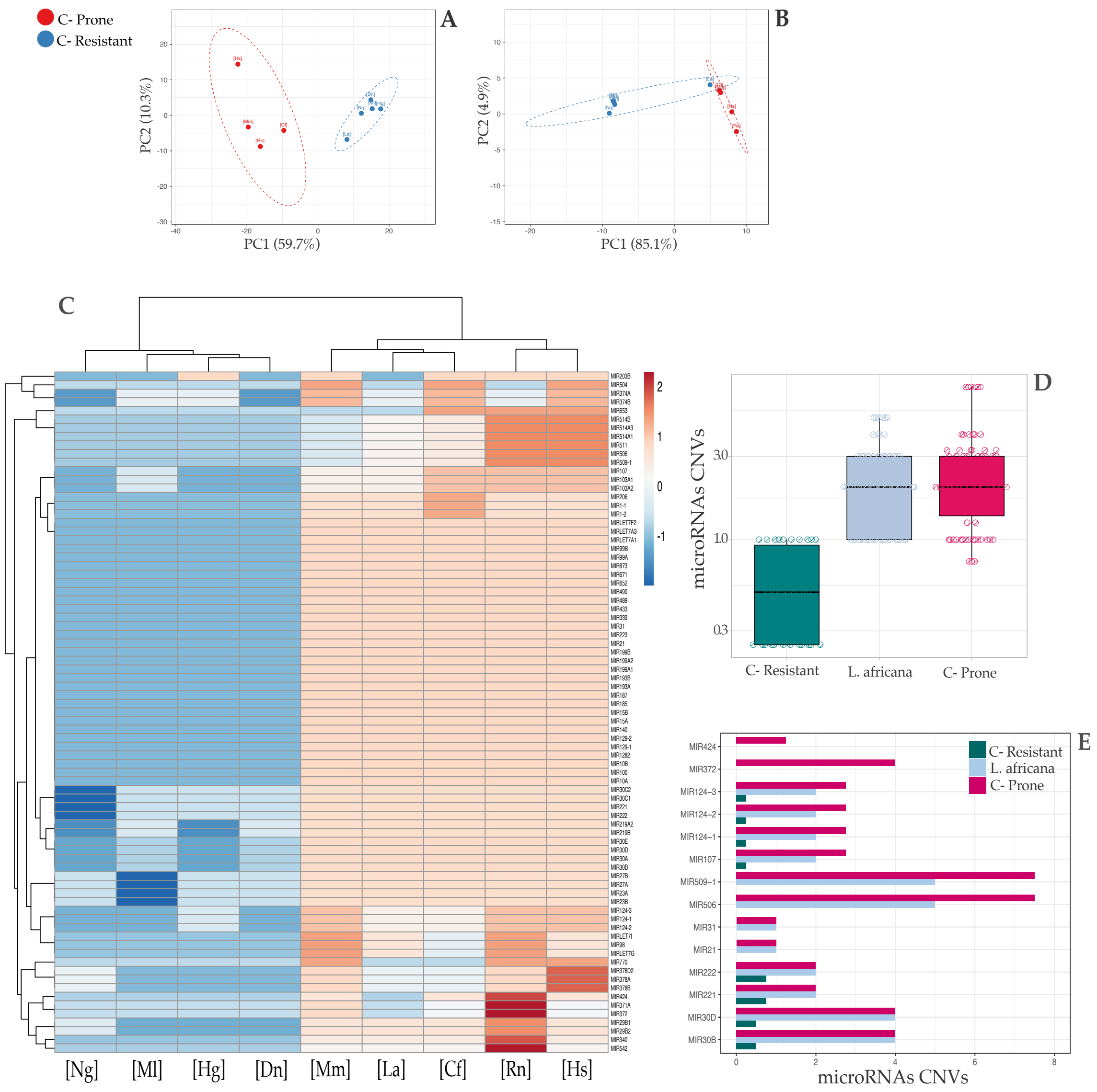


Figure 2: **A**, PCA based on the CNVs of all the significant genes. **B**, PCA based on the CNVs of the significant microRNAs subset. Both plots show a dichotomy between cancer resistant (blue) and cancer prone species (red). **C**, Heatmap of the significant microRNAs, clustered with Euclidean distance and complete linkage**. D,E** Bar- and Box- plot of significant microRNAs CNVs in cancer prone species, cancer resistant species, and Loxodonta Africana. The microRNAs repertoire of Loxodonta africana seems to reflect the cancer prone miRNAs copy number alteration landscape, rather than the one typical of the cancer resistant organisms. In the boxplots, the Y-axis scale has been changed to log one. The boxplots are built considering the average number of copies of each gene in the two different target groups.

As depicted in the heatmap, each cluster has a distinct set of copy number values, and the main branches representing cancer -prone and -resistant organisms perfectly distinguish the two groups. No additional information was given to the algorithm (other than copy numbers), which was able to discriminate between the two groups. In particular, we applied the Euclidean distances, with both 'complete' and 'ward' methods (criteria that direct how the subclusters are merged) (Supplementary materials). Remarkably, also using this method, *Loxodonta africana* microRNAs CNV landscape seems to have a different pattern as compared to the other cancer resistant species (Figures 2C, and Supplementary material), confirming, our idea of identifying the elephant as outlier species of the cancer -resistant group (See Discussion paragraph).

3.2. Cancer related MicroRNAs pathways are the most enriched biological families.

Our analysis shows an enrichment of onco-miRNAs amplifications in the cancer -prone species group. According to our results, important tumor-related miRNAs are able to discriminate between the two organism categories. In particular, miR-424 (p-value = 0.009), miR-372 (p-value = 0.010), miR-107 (p-value = 0.022), miR-124 (p-value = 0.022), miR-506 (p-value = 0.029), miR-511 (p-value = 0.029), miR-378A (p-value = 0.030), miR-1 (p-value = 0.032), miR-206 (p-value = 0.032), and miR-340 (p-value = 0.032), are the most significant microRNA hits, which possess a suppressor and/or oncogenic role (Figure 1C). Given the high diversity among our species, we used the generalized least squares (PGLS) phylogenetic method (Orme., 2013) in order to assess whether copy number and cancer incidence rates evolved in a dependent manner along the tree, or if their relationship might be the consequence of common ancestry, resulting in similar patterns of miRNAs copy number alteration (Supplementary Data S1). Indeed, as shown in Figure 1C-D, the PGLS comparative method was used to establish the association between the cancer incidence rate and the number of the total significant microRNAs taking into account the genetic structure of the population, which outputted a significant correlation between the two traits independently of the shared evolutionary history of the species (adjusted $R^2 = 0.5173$; p-value = 0.01746).

3.3. ORA analysis confirms a significant enrichment in the miRNAs gene family.

To confirm the hypothesis that microRNAs CNVs can represent one of the most important gene family that can potentially discriminate for cancer predisposition, we also performed an Over-Representation Analysis (ORA) (Liao et al., 2019) on the total list of significant genes, in order to underlie functional enriched candidates potentially related to cancer (Table 2). The most enriched pathways outputted by ORA analysis were: "MicroRNAs in cancer", "miRNAs involved in DNA damage response", “Metastatic brain tumor”, “miRNA targets in ECM and membrane receptors'', “let-7 inhibition of ES cell reprogramming”, and “miRNAs involvement in the immune response in sepsis” (Kanehisa, 2019; Martens et al., 2021). These results indicate that the most deregulate genes were miRNAs involved in cancer initiation, chronic inflammation, and immune response. Remarkably, performing the ORA analysis applying PANTHER algorithm (Thomas., 2003), we also found a significant enrichment in the “Cadherin signaling network”, which is a well-known molecular pathway de-scribed as a key player in cancer (Kourtidis et al., 2017).

	Description	FDR (BH)	Genes
KEGG	MicroRNAs in cancer	0	MIR103A1; MIR103A2; MIR107; MIR124-1; MIR124-2; MIR124-3; MIR1-1; MIR1-2; MIR206; MIR100; MIR10A; MIR10B; MIR129-1; MIR129-2; MIR15A; MIR15B; MIR193B; MIR199A1; MIR199A2; MIR199B; MIR203B; MIR21; MIR223; MIR31; MIR99A; MIRLET7A1; MIRLET7A3; MIRLET7F2; MIR29B1; MIR29B2; MIRLET7G; MIRLET7I; MIR221; MIR222; MIR23A; MIR23B; MIR27A; MIR27B; MIR30C1; MIR30C2; MIR30A; MIR30B; MIR30D; MIR30E.
	Taste transduction	3.16E-10	TAS2R10; TAS2R13; TAS2R14; TAS2R19; TAS2R20; TAS2R3; TAS2R30; TAS2R31; TAS2R42; TAS2R43; TAS2R45; TAS2R46; TAS2R50; TAS2R7; TAS2R8; TAS2R9
	Progesterone-mediated oocyte maturation	2.43E-04	SPDYE1; SPDYE11; SPDYE16; SPDYE17; SPDYE2; SPDYE2B; SPDYE3; SPDYE4; SPDYE5; SPDYE6; INS
	Oocyte meiosis	2.73E-04	PPP3R2; SPDYE1; SPDYE11; SPDYE16; SPDYE17; SPDYE2; SPDYE2B; SPDYE3; SPDYE4; SPDYE5; SPDYE6; INS
PANTHER	Cadherin signaling pathway	0.040196	PCDHB14; PCDHB7; PCDHGB1; PCDHB16; PCDHB6; PCDHGB4; PCDHGA6; PCDHGB6; PCDHGB7
Wikipathway	miRNAs involved in DNA damage response	3.76E-09	MIR371A; MIR372; MIR542; MIR100; MIR15B; MIRLET7A1; MIR374B; MIR221; MIR222; MIR23A; MIR23B; MIR27A; MIR27B
	Alzheimers Disease	5.31E-05	MIR124-1; MIR124-2; MIR124-3; MIR10A; MIR129-1; MIR129-2; MIR199B; MIR21; MIR433; MIR671; MIR873; PPP3R2; MIR29B1; MIR30C2; MIR219A2
	Metastatic brain tumor	0.00230738	MIRLET7A1; MIRLET7A3; MIRLET7F2; MIR29B1; MIR29B2; MIRLET7G
	miRNA targets in ECM and membrane receptors	0.00230738	MIR107; MIR15B; MIR30C1; MIR30C2; MIR30B; MIR30D; MIR30E
	MicroRNAs in cardiomyocyte hypertrophy	0.00276823	MIR103A1; MIR103A2; MIR140; MIR15B; MIR185; MIR199A1; MIR199A2; MIR23A; MIR27B; MIR30E
	Cell Differentiation - Index	0.012506063	MIR1-1; MIR206; MIR199A1; MIR199A2; MIR221; MIR222
	let-7 inhibition of ES cell reprogramming	0.01250606	MIRLET7A1; MIRLET7F2; MIRLET7G; MIRLET7I
	miRNAs involvement in the immune response in sepsis	0.01427985	MIR187; MIR199A1; MIR199A2; MIR203B; MIR223; MIR29B1; MIRLET7I

	Cell Differentiation - Index expanded	0.02382508	MIR1-1; MIR206; MIR199A1; MIR199A2; MIR221; MIR222
	Role of Osx and miRNAs in tooth development	0.03346077	MIRLET7A1; MIRLET7F2; MIR29B1; MIRLET7G; MIRLET7I

Table 2: Pathway analysis. Gene Over-Representation Analysis (ORA) using KEGG, PANTHER, and Wikipathway. The enrichment test used Benjamini-Hochberg's FDR correction (FDR < 0.05). CNVs data were previously analyzed by an unpaired 2-group Wilcoxon test (p-value < 0.05). Significant genes altered in their number of copies within the entire genomic landscape were used to perform the ORA analysis, which highlighted a significant enrichment in MicroRNAs and cancer related pathways.

4. Discussion

Being theoretically more susceptible to cancer, big and long living species need additional cancer defense molecular mechanisms. On the other hand, short living and small size organisms might not need them because of their lower intrinsic predisposition to cancer due to their short lifespan rate. CNVs can therefore be considered one of the multiple protection ways against tumor insurgence that can explain Peto's paradox. In fact, we hypothesized that, all cancer resistant organisms implemented a series of molecular mechanisms aimed to preserve themselves from their cancer predisposition, which in turn depends on and derives from their own specific evolutionary history. We believe that CNVs that increase the onco-suppressive capacity of specific genes, can be one of the excellent defenses against tumor diseases in species at risk. Indeed, some authors have recently suggested that one of the most effective cancer resistance strategies is represented by an augmentation in the number of copies of tumor suppressor genes (Vazquez & Lynch, 2021). Parallelly, at a macromolecular level, a reduced cancer resistance rate could be caused by a selective loss of the same suppressor genes (Glenfield & Innan, 2021). For instance, CD52 gene (higher number of copies in the cancer prone group), a membrane glycoprotein expressed on the surface of mature lymphocytes, monocytes and dendritic cells, resulted as one of the most significant hits of our analysis (p-value = 0.007). Recently, Wang and co-authors (J. Wang et al., 2020) identified CD52 as a key role player in tumor immunity, affecting tumor behavior by regulating the associated tumor microenvironment. With the same significant p-value of 0.007, we also identified SAT1 gene (higher number of copies in the cancer prone group) as one of the possible targets to be further investigated in the context of tumor onset. In particular, this gene can regulate and drive brain tumor aggressiveness, promoting molecular pathways in response to DNA damage and regulation of cell cycle (Thakur et al., 2019). Another significant gene resulting from our analysis was represented by the SUMO protein family members (higher number of copies in the cancer resistant group). During cell cycle progression, many tumor suppressors and oncogenes are regulated via SUMOylation (Eifler & Vertegaal, 2015), a biological process that, if deregulated, can lead to genome instability and altered cell proliferation (Müller et al., 2004). In this context, it is evident that some tumors could be dependent on the functional SUMO pathway, but whether it is required for tumor growth remains to be established. For this reason, SUMO2, SUMO3, and SUMO4 can be potentially exploited in further anti-cancer mechanisms investigations (p-value = 0.024 in the present study), in order to shed light on the regulatory mechanisms underlying the activity of SUMO machinery in an oncogenic framework. Among the most significant hits, we also retrieved some genes that are already known to be tumor suppressor or oncogenes (DMD and JUND respectively). Indeed, mutation or deregulated expression of Duchenne Muscular Dystrophy gene (DMD), is often linked to the development and progression of some major cancer types (Jones et al 2021), such as sarcomas, carcinomas, melanomas, lymphomas and brain tumors (Gallia et al., 2018; Ruggieri et al., 2019), being a well-known tumor suppressor in different types of human cancers. On the other hand, JUND (member of the AP-1 family), that is related to MYC signaling pathway, regulates cell cycle and proliferation, and its overexpression is linked to many types of cancer cell (PCA i.e.) (Elliott et al., 2019).

Notably, our results show that miRNAs are the most enriched gene family in discriminating between cancer-prone and cancer-resistant species. The specific role of these miRNAs is not yet fully understood, but we speculate that some of them possess important regulatory functions aimed at defending some species (big size and long lifespan organisms) from cancer, while, at the same time, they are capable of exposing others to tumorigenesis (small size and short lifespan mammals). MicroRNAs (miRNAs) are small post-transcriptional molecular regulators, which are able to modify gene expression levels increasing the amount of mRNA degradation or inhibiting protein translation (Schmiedel et al., 2015). Since each single miRNA can regulate the expression of dozens of genes, many authors were able to correlate their activity with cell development, homeostasis, and disease (Martinez-Sanchez & Murphy, 2013), including cancer (Iorio et al., 2007; Jansson & Lund, 2012). Indeed, some tumorigenic events are caused by a malfunction in the heterogeneous regulatory activity of microRNAs inside the eukaryotic cells. Depending on the specific tissue and on the relationship with the immune system, they can behave both as tumor suppressors and as oncogenes (Svoronos et al., 2016). Furthermore, epigenetic factors and species genetic predisposition can drive their double side behavior, although some of them are evolutionarily conserved within vertebrate taxonomic families (Bartel, 2018). Several miRNAs have been already previously described in literature as oncogenes and tumor suppressors. For example, miR-424 is known to be a human tumor suppressor that can inhibit cell growth enhancing apoptosis or suppressing cell migration (Xu et al., 2013). MiR-372, instead, can participate in WNT cancer molecular pathway (Fan et al., 2018), whereas the overexpression of miR-107, mediating p53 regulation of hypoxic signaling, can suppress tumor angiogenesis and growth in mice (Yamakuchi et al., 2010). MiR-1 is another example of tumor suppressor microRNA that has been previously found significantly down-regulated in squamous carcinoma cells (Nohata et al., 2011). MiR-30b and miR-30d are considered suppressors in those tumors that do not involve immune cells, whereas they have been found upregulated in melanoma (Gaziel-Sovran et al., 2011). Similarly, for the first time, our analysis revealed several miRNAs candidates that might be involved in a mammalian species cancer predisposition (Figure 1C).

Interestingly, all the miRNAs we have found show many more copies in the cancer prone group compared to the cancer free species, and most of them are well known as oncogenes (miR-221, miR222, and miR-372, etc.). MiR-372, for instance, is not present in cancer free species, whereas it shows multiple copies in almost all those ones belonging to the cancer prone group. This microRNA can play a pivotal role in the initiation of breast cancer and may activate WNT and E2F1 pathway during the epithelial-mesenchymal transition process (Fan et al., 2018; Sun & Gao, 2018). We also found an amplification in the cancer prone category for miR-221 and miR-222. Extensive literature has described these two RNAs as oncogenes, being deregulated in primary brain tumors and in Acute Lymphoid Leukemia among other malignancies (Ciafrè et al., 2005; Di Leva et al., 2010). According to our results, surprisingly, cancer prone species showed an amplification of miR-15 tumor suppressor, which is known to be able to regulate cancer proliferation initiation by targeting BCL2 gene (Cimmino et al., 2005). Our hypothesis is that this apparent paradox may underlie a defensive role of this microRNA in those species that are, a priori, susceptible to tumor insurgence. Indeed, according to the so-called "gene dosage hypothesis", gains or losses of specific gene copies can have a dramatic impact on the fitness of a species, leading to altered phenotypes due to the change in the expression levels of the affected genes (Tang & Amon, 2013). On the other hand, oncogenes amplification or tumor suppressors deletions are not always detrimental, but can recapitulate tumorigenic events, being drivers or modulators of the disease (Gordon et al., 2012). As mentioned before, in fact, differences in ecology and evolutionary history are believed to give rise to significant differences between short and long living animals (Kirkwood, 2017), and consequently in cancer prone and resistant species. In 2020, Tollis and co-authors (Tollis et al., 2020) showed that, mammalian lifespan can be correlated to both suppressor genes and of oncogenes CNV, a phenomenon that they themselves called "paradoxical". Interestingly, our analysis also leans in the same direction, suggesting that where high copy numbers of oncogenes might shorten lifespan, they must somehow be counterbalanced by higher copy numbers of TSGs.

In this framework, compared to the other mammals, elephant miRNAs amplification signature resembles the organisms of cancer prone group (Figure 1C-D). In fact, it showed an alteration in the copy numbers of known oncogenes, such as miR-221 and miR-222, together with miR-30b/d and miR-31. In our opinion, Loxodonta africana, should be placed in a new category of organisms, which share both oncogenic and cancer free characteristics, being also clustered as outlier species of the cancer resistant group (Figure 2B). During the evolution, elephants may have selected molecular defenses, such as the amplification of TP53 and pseudogenes (Abegglen et al., 2015; Sulak et al., 2016), with the aim to defend their cells from the tumorigenic action of a high percentage of onco-miRNAs copy number amplification and high longevity. Consequently, an additional amplification in the number of tumor suppressor microRNAs would have not been sustainable/useful in term of fitness and/or survival. The hypothesis is that species with bigger sizes and longer lifespans have an expanded number of TSGs, which is even higher than the one of their oncogenic counterparts. In this way the direct elimination of oncogenes which implies elevated costs in terms of growth and cellular functions maintenance can be avoided, thus reducing the cancer incidence risk. In support of this, recently, Vazquez and Lynch (2021) (Vazquez & Lynch., 2021) reported that, within the Afrotheria order, the tumor suppressor genes found in an altered number of copies was relatively lower compared to what might be expected. This finding can mirror the trade-off mechanism that natural selection has developed during evolution in order to compensate for the multi copies effect which can lead to an increased risk of cancer, due to the unbalanced number of copies of the same genes. Indeed, long-living species might possess mechanisms which are capable of maintaining the equilibrium between proliferation and tumor control. Their regulatory networks can create positive feedback in which the amplification of tumor suppressor families functions as a buffer against the oncogene co-expansion, or vice-versa (Tollis et al., 2020). On the other hand, cancer-prone organisms included in our analysis, do not develop these gene defenses because they have a lower lifespan, which does not make them particularly exposed to a severe lack of fitness due to cancer progression (except in the case of Homo sapiens that has reached a high lifespan only recently, thanks to the advance of medicine treatments and health care).

5. Limitations and Perspectives

Gene duplication is a fundamental process that can lead to the emergence of new phenotypic traits. Analyzing patterns of gene copy number alteration across the genome of large and long-living organisms, may reveal new insights about those mechanisms underlying cancer resistance in mammals (Abegglen et al., 2015; Tejada-Martinez et al., 2021; Tollis et al., 2020). Here, we have developed a simple way to test the hypothesis that CNVs confer protection or increase vulnerability to cancer among species. Using the absolute number of copies of each gene by species, we were able to identify, for the first time, an alteration in miRNA CNVs, that are overrepresented and enriched in molecular pathways related to cancer. Further analyses will help to validate these findings by better defining the correlation between miRNAs and their targets. In our opinion, studying microRNAs that are related to human malignancies from a comparative genomics perspective, can provide additional clues about their role, and potentially point towards novel targets involved in tumorigenic diseases. Focusing on patterns of miRNAs copy number changes may, for the first time, give new insights into the conserved molecular pathways influencing cancer incidence across species, and may lead to the discovery of novel therapeutic approaches.

Supplementary Materials:

Table S1: Species description. Phenotype characteristics of the 9 species included in our analysis**.**

Table S2-S3: Cancer Prone vs Cancer Resistant: a two-group statistical comparison. List of the significant hits resulting from the unpaired 2-group wilcoxon test (p-value <0.05) applied on the total genomic CNVs landscape of the selected species.

Table S4: Pathway analysis – extended version. Gene Over-Representation Analysis (ORA). The enrichment test used Benjamini-Hochberg's FDR correction (FDR < 0.05). CNVs data were previously analyzed by an unpaired 2-group wilcoxon test (p-value < 0.05).

Supplementary Data:

S1: PGLS modelling results: Cancer incidence rate ~ miRNAs CNVs.

S2-S4: Clustering heatmaps. Heatmaps of the significant genes, clustered with Euclidean distance, and ward and complete linkage methods.

Author Contributions: Conceptualization, C.V. and C.T.; methodology, C.V. and C.T.; F.B., F.M., R.M., V.P., established the bioinformatical resource used to generate the data; C.V., F.B., F.M., R.M., V.P., C.T., performed the analysis and the investigation; data curation, C.V., F.B., V.P.; data interpretation, C.V.; writing—original draft preparation, C.V., C.T.; writing—review and editing, C.V., C.T., F.M., R.M.; supervision, C.T., F.M., R.M..; funding acquisition, C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the UNIVERSITY OF PADOVA through M.A.P.S. department under the program BIRD213010.

Data Availability Statement: All data necessary for confirming the conclusions of the article are present within the article, figures, tables, and its supplementary material.

Acknowledgments: We thank Dr. Nicoletta Bianchi for critical reading of our manuscript and for her precious suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References:

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., Kiso, W. K., Schmitt, D. L., Waddell, P. J., Bhaskara, S., Jensen, S. T., Maley, C. C., & Schiffman, J. D. (2015). Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. JAMA, 314(17), 1850.

Bartel, D. P. (2018). Metazoan MicroRNAs. Cell, 173(1), 20–51.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., ... & Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. Nature, 463(7283), 899-905.

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., ... & Stratton, M. R. (2010). Signatures of mutation and selection in the cancer genome. Nature, 463(7283), 893-898.

Boddy, A. M., Abegglen, L. M., Pessier, A. P., Aktipis, A., Schiffman, J. D., Maley, C. C., & Witte, C. (2020). Lifetime cancer prevalence and life history traits in mammals. Evolution, Medicine, and Public Health, 2020(1), 187–195.

Boddy, A. M., Harrison, T. M., & Abegglen, L. M. (2020). Comparative Oncology: New Insights into an Ancient Disease. IScience, 23(8), 101373.

Buffenstein, R. (2008). Negligible senescence in the longest living rodent, the naked mole-rat: Insights from a successfully aging species. Journal of Comparative Physiology B, 178(4), 439–445.

Cano, C. E., Hamidi, T., Sandi, M. J., & Iovanna, J. L. (2011). Nupr1: The Swiss-knife of cancer. Journal of Cellular Physiology, 226(6), 1439–1443.

Caulin, A. F., & Maley, C. C. (2011). Peto's Paradox: Evolution's prescription for cancer prevention. Trends in Ecology & Evolution, 26(4), 175–182.

Chen, Q., Zhou, W., Han, T., Du, S., Li, Z., Zhang, Z., Shan, G., & Kong, C. (2016). MiR-378 suppresses prostate cancer cell growth through downregulation of MAPK1 in vitro and in vivo. Tumor Biology, 37(2), 2095–2103.

Ciafrè, S. A., Galardi, S., Mangiola, A., Ferracin, M., Liu, C.-G., Sabatino, G., Negrini, M., Maira, G., Croce, C. M., & Farace, M. G. (2005). Extensive modulation of a set of microRNAs in primary glioblastoma. Biochemical and Biophysical Research Communications, 334(4), 1351–1358.

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., Wojcik, S. E., Aqeilan, R. I., Zupo, S., Dono, M., Rassenti, L., Alder, H., Volinia, S., Liu, C. -g., Kipps, T. J., Negrini, M., & Croce,

C. M. (2005). MiR-15 and miR-16 induce apoptosis by targeting BCL2. Proceedings of the National Academy of Sciences, 102(39), 13944–13949.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. Bioinformatics, 22(10), 1269–1271.

Di Leva, G., Gasparini, P., Piovan, C., Ngankeu, A., Garofalo, M., Taccioli, C., Iorio, M. V., Li, M., Volinia, S., Alder, H., Nakamura, T., Nuovo, G., Liu, Y., Nephew, K. P., & Croce, C. M. (2010). MicroRNA Cluster 221-222 and Estrogen Receptor α Interactions in Breast Cancer. JNCI: Journal of the National Cancer Institute, 102(10), 706–721.

Eifler, K., & Vertegaal, A. C. O. (2015). SUMOylation-Mediated Regulation of Cell Cycle Progression and Cancer. Trends in Biochemical Sciences, 40(12), 779–793.

Elliott, B., Millena, A. C., Matyunina, L., Zhang, M., Zou, J., Wang, G., Zhang, Q., Bowen, N., Eaton, V., Webb, G., Thompson, S., McDonald, J., & Khan, S. (2019). Essential role of JunD in cell proliferation is mediated via MYC signaling in prostate cancer cells. Cancer Letters, 448, 155–167.

Fan, X., Huang, X., Li, Z., & Ma, X. (2018). MicroRNA-372-3p promotes the epithelial-mesenchymal transition in breast carcinoma by activating the Wnt pathway. Age (years), 50(27), 14.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., & Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. International Journal of Cancer, 136(5), E359–E386.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97.

Gallia, G. L., Zhang, M., Ning, Y., Haffner, M. C., Batista, D., Binder, Z. A., Bishop, J. A., Hann, C. L., Hruban, R. H., Ishii, M., Klein, A. P., Reh, D. D., Rooper, L. M., Salmasi, V., Tamargo, R. J., Wang, Q., Williamson, T., Zhao, T., Zou, Y., … Bettegowda, C. (2018). Genomic analysis identifies frequent deletions of Dystrophin in olfactory neuroblastoma. Nature Communications, 9(1), 5410.

Gaziel-Sovran, A., Segura, M. F., Di Micco, R., Collins, M. K., Hanniford, D., Vega-Saenz de Miera, E., Rakus, J. F., Dankert, J. F., Shang, S., Kerbel, R. S., Bhardwaj, N., Shao, Y., Darvishian, F., Zavadil, J., Erlebacher, A., Mahal, L. K., Osman, I., & Hernando, E. (2011). MiR-30b/30d Regulation of GalNAc Transferases Enhances Invasion and Immunosuppression during Metastasis. Cancer Cell, 20(1), 104–118.

Glenfield, C., & Innan, H. (2021). Gene Duplication and Gene Fusion Are Important Drivers of Tumourigenesis during Cancer Evolution. Genes, 12(9), 1376.

Gordon, D. J., Resio, B., & Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. Nature Reviews Genetics, 13(3), 189–203.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. 2016, 17.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. Nucleic Acids Research, 49(D1), D884–D891.

Iorio, M. V., Visone, R., Di Leva, G., Donati, V., Petrocca, F., Casalini, P., Taccioli, C., Volinia, S., Liu, C.-G., Alder, H., Calin, G. A., Ménard, S., & Croce, C. M. (2007). MicroRNA Signatures in Human Ovarian Cancer. Cancer Research, 67(18), 8699–8707.

Iskow, R. C., Gokcumen, O., & Lee, C. (2012). Exploring the role of copy number variants in human adaptation. Trends in Genetics, 28(6), 245-257.

Jansson, M. D., & Lund, A. H. (2012). MicroRNA and cancer. Molecular Oncology, 6(6), 590–610.

Jones, L., Naidoo, M., Machado, L. R., & Anthony, K. (2021). The Duchenne muscular dystrophy gene and cancer. Cellular Oncology, 44(1), 19–32.

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Science, 28(11), 1947–1951.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology, 8(2), e1002375.

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., Marino, S. M., Sun, X., Turanov, A. A., Yang, P., Yim, S. H., Zhao, X., Kasaikina, M. V., Stoletzki, N., Peng, C., Polak, P., Xiong, Z., Kiezun, A., Zhu, Y., … Gladyshev, V. N. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature, 479(7372), 223–227.

Kirkwood, T. B. L. (2017). Why and how are we living longer?: Why and how are we living longer? Experimental Physiology, 102(9), 1067–1074.

Kourtidis, A., Lu, R., Pence, L. J., & Anastasiadis, P. Z. (2017). A central role for cadherin signaling in cancer. Experimental Cell Research, 358(1), 78–85.

Kureel, J., Dixit, M., Tyagi, A. M., Mansoori, M. N., Srivastava, K., Raghuvanshi, A., Maurya, R., Trivedi, R., Goel, A., & Singh, D. (2014). MiR-542-3p suppresses osteoblast cell proliferation and differentiation, targets BMP-7 signaling and inhibits bone formation. Cell Death & Disease, 5(2), e1050–e1050.

Lagunas-Rangel, F. A. (2018). Cancer-free aging: Insights from Spalax ehrenbergi superspecies. Ageing Research Reviews, 47, 18–23.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Research, 47(W1), W256–W259.

Liang, S., Mele, J., Wu, Y., Buffenstein, R., & Hornsby, P. J. (2010). Resistance to experimental tumorigenesis in cells of a long-lived mammal, the naked mole-rat (Heterocephalus glaber): Oncogene resistance in naked mole-rat cells. Aging Cell, 9(4), 626–635.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research, 47(W1), W199–W205.

Lipman, R., Galecki, A., Burke, D. T., & Miller, R. A. (2004). Genetic Loci That Influence Cause of Death in a Heterogeneous Mouse Stock. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 59(10), B977–B983.

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A. Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., Dupuis, L. J., Winckers, L. A., Coort, S. L., Willighagen, E. L., Evelo, C. T., Pico, A. R., & Kutmon, M. (2021). WikiPathways: Connecting communities. Nucleic Acids Research, 49(D1), D613–D621.

Martinez-Sanchez, A., & Murphy, C. (2013). MicroRNA Target Identification—Experimental Approaches. Biology, 2(1), 189–205.

Metsalu, T., & Vilo, J. (2015). ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Research, 43(W1), W566–W570.

Miremadi, A., Oestergaard, M. Z., Pharoah, P. D. P., & Caldas, C. (2007). Cancer genetics of epigenetic genes. Human Molecular Genetics, 16(R1), R28–R49.

Müller, S., Ledl, A., & Schmidt, D. (2004). SUMO: A regulator of gene expression and genome integrity. Oncogene, 23(11), 1998–2008.

Nohata, N., Sone, Y., Hanazawa, T., Fuse, M., Kikkawa, N., Yoshino, H., Chiyomaru, T., Kawakami, K., Enokida, H., Nakagawa, M., Shozu, M., Okamoto, Y., & Seki, N. (2011). MiR-1 as a tumor suppressive microRNA targeting TAGLN2 in head and neck squamous cell carcinoma. Oncotarget, 2(1–2), 29–42.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2013). The caper package: comparative analysis of phylogenetics and evolution in R. R package version, 5(2), 1-36.

Peters, K. M., Carlson, B. A., Gladyshev, V. N., & Tsuji, P. A. (2018). Selenoproteins in colon cancer. Free Radical Biology and Medicine, 127, 14–25.

Peto, R., Roe, F. J., Lee, P. N., Levy, L., & Clack, J. (1975). Cancer and ageing in mice and men. British Journal of Cancer, 32(4), 411–426.

Postma, M., & Goedhart, J. (2019). PlotsOfData—A web app for visualizing data together with their summaries. PLOS Biology, 17(3), e3000202.

Ren, L.-L., Yan, T.-T., Shen, C.-Q., Tang, J.-Y., Kong, X., Wang, Y.-C., Chen, J., Liu, Q., He, J., Zhong, M., Chen, H.-Y., Hong, J., & Fang, J.-Y. (2018). The distinct role of strand-specific miR-514b-3p and miR-514b-5p in colorectal cancer metastasis. Cell Death & Disease, 9(6), 687.

Ruggieri, S., De Giorgis, M., Annese, T., Tamma, R., Notarangelo, A., Marzullo, A., Senetta, R., Cassoni, P., Notarangelo, M., Ribatti, D., & Nico, B. (2019). Dp71 Expression in Human Glioblastoma. International Journal of Molecular Sciences, 20(21), 5429.

Schmiedel, J. M., Klemm, S. L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D. S., & van Oudenaarden, A. (2015). MicroRNA control of protein expression noise. Science, 348(6230), 128–132.

Schneeweis, C., Hassan, Z., Schick, M., Keller, U., & Schneider, G. (2021). The SUMO pathway in pancreatic cancer: Insights and inhibition. British Journal of Cancer, 124(3), 531–538.

Seluanov, A., Gladyshev, V. N., Vijg, J., & Gorbunova, V. (2018). Mechanisms of cancer resistance in long-lived mammals. Nature Reviews Cancer, 18(7), 433–441.

Seluanov, A., Hine, C., Azpurua, J., Feigenson, M., Bozzella, M., Mao, Z., Catania, K. C., & Gorbunova, V. (2009). Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. Proceedings of the National Academy of Sciences, 106(46), 19352–19357.

Serrano, M. (2016). Unraveling the links between cancer and aging. Carcinogenesis, 37(2), 107–107.

Spilka, R., Ernst, C., Mehta, A. K., & Haybaeck, J. (2013). Eukaryotic translation initiation factors in cancer development and progression. Cancer Letters, 340(1), 9–21.

Spitzer, M., Wildenhain, J., Rappsilber, J., & Tyers, M. (2014). BoxPlotR: A web tool for generation of box plots. Nature Methods, 11(2), 121–122.

Squadrito, M. L., Pucci, F., Magri, L., Moi, D., Gilfillan, G. D., Ranghetti, A., Casazza, A., Mazzone, M., Lyle, R., Naldini, L., & De Palma, M. (2012). MiR-511-3p Modulates Genetic Programs of Tumor-Associated Macrophages. Cell Reports, 1(2), 141–154.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719–724.

Sulak, M., Fong, L., Mika, K., Chigurupati, S., Yon, L., Mongan, N. P., Emes, R. D., & Lynch, V. J. (2016). TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. ELife, 5, e11994.

Sun, H., & Gao, D. (2018). Propofol suppresses growth, migration and invasion of A549 cells by down-regulation of miR-372. BMC Cancer, 18(1), 1252.

Svoronos, A. A., Engelman, D. M., & Slack, F. J. (2016). OncomiR or Tumor Suppressor? The Duplicity of MicroRNAs in Cancer. Cancer Research, 76(13), 3666–3670.

Tang, Y.-C., & Amon, A. (2013). Gene Copy-Number Alterations: A Cost-Benefit Analysis. Cell, 152(3), 394–405.

Tejada-Martinez, D., de Magalhães, J. P., & Opazo, J. C. (2021). Positive selection and gene duplications in tumour suppressor genes reveal clues about how cetaceans resist cancer. Proceedings of the Royal Society B, 288(1945), 20202592.

Thakur, V. S., Aguila, B., Brett-Morris, A., Creighton, C. J., & Welford, S. M. (2019). Spermidine/spermine N1-acetyltransferase 1 is a gene-specific transcriptional regulator that drives brain tumor aggressiveness. Oncogene, 38(41), 6794–6800.

Thomas, P. D. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Research, 13(9), 2129–2141.

Tian, X., Seluanov, A., & Gorbunova, V. (2017). Molecular Mechanisms Determining Lifespan in Short- and Long-Lived Species. Trends in Endocrinology & Metabolism, 28(10), 722–734.

Tollis, M., Boddy, A. M., & Maley, C. C. (2017). Peto's Paradox: How has evolution solved the problem of cancer prevention? BMC Biology, 15(1), 60.

Tollis, M., Schiffman, J. D., & Boddy, A. M. (2017). Evolution of cancer suppression as revealed by mammalian comparative genomics. Current Opinion in Genetics & Development, 42, 40–47.

Tollis, M., Schneider-Utaka, A. K., & Maley, C. C. (2020). The Evolution of Human Cancer Gene Duplications across Mammals. Molecular Biology and Evolution, 37(10), 2875–2886.

Turco, C., Donzelli, S., & Fontemaggi, G. (2020). miR-15/107 microRNA Gene Group: Characteristics and Functional Implications in Cancer. Frontiers in Cell and Developmental Biology, 8, 427.

Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biology, 17(12), e3000494.

Vazquez, J. M., & Lynch, V. J. (2021). Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. ELife, 10, e65041.

Vischioni, C., Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2022). Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research. Big Data Research, 27, 100298.

Wang, J., Zhang, G., Sui, Y., Yang, Z., Chu, Y., Tang, H., Guo, B., Zhang, C., & Wu, C. (2020). CD52 Is a Prognostic Biomarker and Associated With Tumor Microenvironment in Breast Cancer. Frontiers in Genetics, 11, 578002.

Wang, L.-F., Walker, P. J., & Poon, L. L. M. (2011). Mass extinctions, biodiversity and mitochondrial function: Are bats 'special' as reservoirs for emerging viruses? Current Opinion in Virology, 1(6), 649–657.

Wang, P., Chen, L., Zhang, J., Chen, H., Fan, J., Wang, K., Luo, J., Chen, Z., Meng, Z., & Liu, L. (2014). Methylation-mediated silencing of the miR-124 genes facilitates pancreatic cancer progression and metastasis by targeting Rac1. Oncogene, 33(4), 514–524.

Wen, S.-Y., Lin, Y., Yu, Y.-Q., Cao, S.-J., Zhang, R., Yang, X.-M., Li, J., Zhang, Y.-L., Wang, Y.-H., Ma, M.-Z., Sun, W.-W., Lou, X.-L., Wang, J.-H., Teng, Y.-C., & Zhang, Z.-G. (2015). MiR-506 acts as a tumor suppressor by directly targeting the hedgehog pathway transcription factor Gli3 in human cervical cancer. Oncogene, 34(6), 717–725.

Wilkinson, G. S., & Adams, D. M. (2019). Recurrent evolution of extreme longevity in bats. Biology Letters, 15(4), 20180860.

Wu, Z., Wu, Q., Wang, C., Wang, X., Huang, J., Zhao, J., Mao, S., Zhang, G., Xu, X., & Zhang, N. (2011). MiR-340 inhibition of breast cancer cell migration and invasion through targeting of oncoprotein c-Met. Cancer, 117(13), 2842–2852.

Xian, S., Shang, D., Kong, G., & Tian, Y. (2018). FOXJ1 promotes bladder cancer cell growth and regulates Warburg effect. Biochemical and Biophysical Research Communications, 495(1), 988–994.

Xu, J., Li, Y., Wang, F., Wang, X., Cheng, B., Ye, F., Xie, X., Zhou, C., & Lu, W. (2013). Suppressed miR-424 expression via upregulation of target gene Chk1 contributes to the progression of cervical cancer. Oncogene, 32(8), 976–987.

Yamakuchi, M., Lotterman, C. D., Bao, C., Hruban, R. H., Karim, B., Mendell, J. T., Huso, D., & Lowenstein, C. J. (2010). P53-induced microRNA-107 inhibits HIF-1 and tumor angiogenesis. Proceedings of the National Academy of Sciences, 107(14), 6334–6339.

Yim, S. H., Clish, C. B., & Gladyshev, V. N. (2019). Selenium Deficiency Is Associated with Pro-longevity Mechanisms. Cell Reports, 27(9), 2785-2797.e3.

Zhai, Q., Zhou, L., Zhao, C., Wan, J., Yu, Z., Guo, X., Qin, J., Chen, J., & Lu, R. (2012). Identification of miR-508-3p and miR-509-3p that are associated with cell invasion and migration and involved in the apoptosis of renal cell carcinoma. Biochemical and Biophysical Research Communications, 419(4), 621–626.

Zhang, L., Liu, X., Jin, H., Guo, X., Xia, L., Chen, Z., Bai, M., Liu, J., Shang, X., Wu, K., Pan, Y., & Fan, D. (2013). MiR-206 inhibits gastric cancer proliferation in part by repressing CyclinD2. Cancer Letters, 332(1), 94–101.

Zhang, Z., Liu, M., Hu, Q., Xu, W., Liu, W., Sun, Q., ... & Qin, Y. (2019). FGFBP1, a downstream target of the FBW7/c-Myc axis, promotes cell proliferation and migration in pancreatic cancer. American journal of cancer research, 9(12), 2650.

Zhu, W., Xue, Y., Liang, C., Zhang, R., Zhang, Z., Li, H., Su, D., Liang, X., Zhang, Y., Huang, Q., Liu, M., Li, L., Li, D., Zhao, A. Z., & Liu, Y. (2016). S100A16 promotes cell proliferation and metastasis via AKT and ERK cell signaling pathways in human prostate cancer. Tumor Biology, 37(9), 12241–12250.

Chapter IV:

Saccharomyces cerevisiae: a budding model for cancer and ageing research.

Even though the quest of eternal youth has always fascinated mankind, the interest in aging research only boomed around the beginning of the 1990. At that time, Johnson and Kenyon were among the first scientists to identify genes that appear to control lifespan, challenging former beliefs according to which longevity was predetermined and immutable. Specifically, a mutation in age-1 gene was found to increase the mean lifespan of *C. elegans* by about the 65%, whereas the deletion of the gene daf-2 granted adult hermaphrodites worms as much as a twice longer longevity compared to their wild-type counterpart (Johnson, 1990; Kenyon et al., 1993). This guided the current currents of thought towards the idea of aging as a process not genetically immutable. From this point and forward, hundreds and hundreds of genes regulating lifespan have therefore been identified, in multiple species and different conditions, leading to the fundamental conclusion that the mechanisms underlying aging are highly conserved from yeast to humans (Kenyon, 2010). The budding yeast *Saccharomyces cerevisiae* is a single-cell organism from the fungi kingdom, which has extensively been exploited by humans for thousands of years in food and beverage making processes. Although unicellular, yeasts usually live included into a colony organization. Yeasts mostly reproduce by asexual mitotic growth through a process called budding, during which the "mother cell" produces a protrusion, commonly referred to as a “bud”, that eventually forms a genetically identical newborn daughter cell. However, *S. cerevisiae* is also able to enter sexual reproduction when subjected to stress or nutrient depletion, initiating both meiosis and sporulation at the same time to form dormant gametes embedded into a protective spore wall (Figure 1).

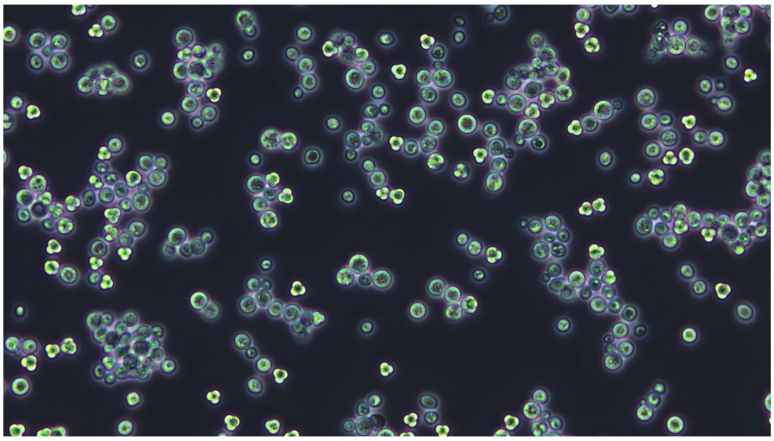


Figure 1: Heterogenous population of single, budding, and tetrad cells of *Saccharomyces spp*. under the microscope.

The co-evolution of gametogenesis and sporulation is thought to result from a selective advantage driven in wild habitats as it leads to an increased genetic diversity to better cope with changing environments, while it offers an enhanced protection against external stresses such as heat, desiccation, or digestive tract enzymes (Coluccio et al., 2008; de Chiara et al., 2022). In the last decades, *S. cerevisiae* has become a pivotal eukaryotic model system which led to the discovery of major biological processes in eukaryotes, owing to the remarkable properties of this simple organism (Botstein and Fink, 2011). Firstly, with a doubling time of less than 2 hours, it can be easily cultured in laboratory conditions, allowing rapid production and maintenance of multiple strains at very low costs. Secondly, *S. cerevisiae* was the first fully sequenced eukaryotic organism (Goffeau et al., 1996), and its small genome (12Mb) split in 16 chromosomes is the best annotated one as of today (www.yeastgenome.org). Taking advantage of its proficient homologous recombination repair machinery, *S. cerevisiae* can be easily genetically engineered with gene deletions or insertions, controlled gene expression, or recombinant DNA (Ito et al., 1983). Thanks to these features, researchers can tweak yeast genomes in almost any possible way and within a very short time-lapse. This enabled the construction of multiple genome-wide yeast collection mutants, which allowed to unravel gene functionality at an unprecedented level (Ho et al., 2009; Huh et al., 2003; Giaever et al., 2002; Winzeler et al., 1999). Thirdly, being eukaryote, yeasts share most cellular, molecular, and metabolic processes in common with multicellular eukaryotes. In addition, numerous genetic pathways are conserved from yeasts to mammals and about 17% of *S. cerevisiae* genes belong to orthologous gene families associated with human diseases (Heinicke et al., 2007). For all these reasons, *S. cerevisiae* is at the forefront of the new emerging fields of functional genomics and systems biology, which aim at developing a more holistic understanding of the cellular machinery and how it eventually defines any given living entity.

Maybe more unexpectedly, *S. cerevisiae* is also used as a prime model in ageing research to study two distinct paradigms: the replicative and the chronological lifespans. The replicative lifespan (RLS), refers to the total number of divisions a yeast cell can undergo before entering senescence (~25, depending on the strain background) (Stumpferl et al., 2012; Zhang et al, 2012; Kaeberlein, 2010; Mortimer and Johnston, 1959), whereas the chronological lifespan (CLS) is defined as the duration yeast cells can survive in a non-proliferative state (up to several weeks, depending on the strain background and the environment) (Longo et al., 2012; Fabrizio & Longo, 2003; MacLean, Harris, and Piper, 2001). In particular, RLS helps to investigate the ageing processes affecting proliferative tissues, such as lymphocytes or fibroblasts

(Wasko and Kaeberlein, 2014; Longo et al., 2012; Steinkraus et al., 2008), while CLS is used to untangle the mechanisms underlying maintenance of post-mitotic tissues, such as mature neurons or muscular cells (Ruetenik and Barrientos, 2015; Longo and Fabrizio, 2012; Longo et al., 2012; Longo et al., 1996;) (Figure 2).

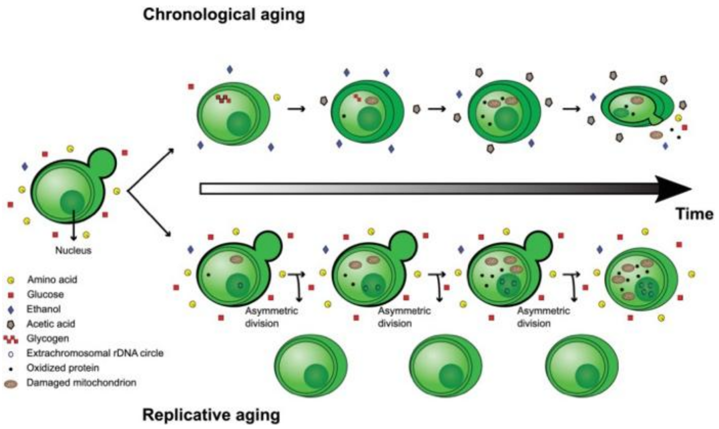


Figure 2: **The two ageing paradigms of** ***S. cerevisiae***. Chronological Life Span (CLS) represents the time cells can sustain under non-proliferative conditions; Replicative Life Span (RLS) refers to the number of divisions a mother cell can undergo before senescence.
(**Source**: *Wauters, Britton, & Verstrepen 2021*).

In nutrient rich conditions, *S. cerevisiae* asymmetrically divides by mitosis leading to a different partitioning of cellular components between mother and daughter cells. The latter specifically inherits damage free material through a tightly regulated filtering process occurring at the bud neck, whereas the former retains altered constituents such as carbonylated proteins, deficient mitochondria, impaired vacuoles, and episomal DNA. This leads to the rejuvenation of newborn daughters at each division and fully resets their replicative potential, thus preventing the senescence and extinction of the cellular lineage (Denoth Lippuner et al., 2014). However, because nutrients are scarce in natural habitats and because yeasts divide exponentially and readily saturate their environment, yeast cells have mostly evolved in non-proliferative conditions, a situation in which they tend to exit the cell cycle and to sustain in a reversible and resilient quiescent state (de Virgilio, 2012; Gray et al., 2004). Interestingly though, it was observed that not all cells are able to enter quiescence upon nutrient exhaustion and that stationary phase cultures are a complex mixture of heterogeneous cell types with very variable life expectancies (Allen et al., 2006). During aging, both models exhibit signs of physiological decline as observed in multicellular organisms. Cells undergoing RLS exhibit a gradually slower growth rate and experience a drop in fertility and mating efficiencies (Lee et al., 2012), while chronologically ageing cells progressively lose their replicative potential (Ashrafi et al., 1999). However, the common denominator between the RLS and CLS paradigms lies on the innate damage limit a cell can tolerate before becoming irreversibly impaired. Nowadays, there is a wide understanding of the genetic factors controlling both aging models. Indeed, hundreds of genes regulating lifespan have been described and linked to cellular processes involved in aging, such as:

- mitochondrial functioning,
- amino acid homeostasis,
- glycogen accumulation,
- apoptosis,
- regulation of the cell cycle,
- TOR and protein kinase A (PKA) signaling,
- autophagy,

(Campos et al., 2018; Garay et al., 2014; Gresham et al., 2011; Fabrizio et al., 2010; Alvers et al., 2009; Powers et al., 2006). Additionally, aging is accompanied with a raise in genomic instability that occurs in both yeast aging systems, akin to what is observed in higher eukaryotes. Indeed, it was found that the rate of Loss of Heterozygosity (LOH), which results from the semi-conservative repair of double strand breaks by homologous recombination, exponentially scales in old cells during replicative lifespan, fostered by increasing replication stress and mitochondrial dysfunction (Lindstrom and Gottschling, 2011; Veatch et al., 2009; McMurray and Gottschling, 2003). During CLS, DNA damage mainly occurs in the form of base substitutions, indels, gross chromosomal rearrangements (Wei et al., 2009; Fabrizio et al., 2004; Maclean et al., 2003; Longo et al., 1999), and also as LOH (Qin et al., 2008). Moreover, induction of replicative stress through deletion of genes involved in DNA repair shortens CLS by preventing cells from entering into quiescence (Weinberger et al., 2013; Weinberger et al., 2007; Laschober et al., 2010).

1. My experience with *S. cerevisiae* ageing research

Starting in September 2020, I spent 9 months as a visiting Ph.D student in the "Population Genomics and Complex Traits" team led by Dr. Gianni Liti, at the Institute for Research on Cancer and Aging (IRCAN), in Nice, France. The team uses both experimental approaches and bioinformatics to discover and untangle the architecture and the evolution of *S. cerevisiae* genome, and to understand how genetic diversity drives phenotypic variation. In 2018, they contributed to the development of the 1011 Yeast Genomes Project (Peter et al., 2018), which provides genomics data from over a thousand natural yeast isolates and thus constitutes the most comprehensive study of yeast population genomics to date. Besides genome evolution, downstream applications of this resource will substantially help to dissect the architecture of complex traits as it enables the usage of Genome-Wide Association studies in yeasts for the first time. Specifically, during my stay, I have been working on two projects related to genetic instability in chronologically ageing cells, and to the identification and validation of natural genetic drivers of CLS, thus having the opportunity to combine my bioinformatics background to the exploration and the learning of advanced experimental approaches already used in the team. During chronological aging, cell survival can be measured in different ways (Figure 3B). Generally, counting the colony forming units (CFU) able to grow on solid selective medium over time is the gold standard approach (Mirisola et al., 2014; Piper, 2011). However, because this method is poorly accurate, error prone, and low-throughput, scientists started to develop alternative procedures to substantially increase the scalability of yeast CLS assays and to allow systematic screening of large collections. One of these methods relies on the usage of fluorescent viability markers, such as Propidium Iodide (PI), which is unable to enter living cells, while it specifically stains dead cells upon membrane disruption. This approach can readily be coupled with high-throughput flow cytometry and enables to automatically obtain a direct estimation of cell viability at any given time point (Barré et al., 2020). Specifically, this is the technology I have learnt and used in the following projects I was involved with.

As previously mentioned, genome instability is one of the hallmarks of cancer and ageing. Earlier studies in *S. cerevisiae* have provided direct evidence of increased genomic instability in replicative old cells, and more particularly of LOH, which however has not been studied in CLS so far (McMurray & Gottschling., 2003). Accordingly, my task was to use yeast genetics and genomics tools to determine the occurrence of LOH in chronologically aged cells. I have applied a simple genetic system described in Mozzachiodi et al., (2021), in order to measure the LOH rate in five different diploid yeast backgrounds. In a few words, I used engineered yeast strains carrying a replacement of the LYS2 gene with the URA3 one, and measured how often this URA3 marker was lost due to LOH events (Figure 3A). To do this, I used a so-called "URA3-LYS2" system, based on the counterselection of clones able to grow in a solid media containing 5-fluoroorotic Acid Monohydrate (5-FOA). In normal conditions, 5-FOA is specifically recognized as a substrate of the enzyme encoded by URA3 gene and is metabolized into a toxic compound for the cells. Hence, only cells becoming URA3-/URA3-, by substitution of URA3 with LYS2 resulting from a LOH, can grow on media supplemented with 5-FOA. Thanks to this simple approach, it is possible to count, select and identify URA3- cells, and therefore to quantify the LOH rate of yeast cells at different ages (Figure 3C).

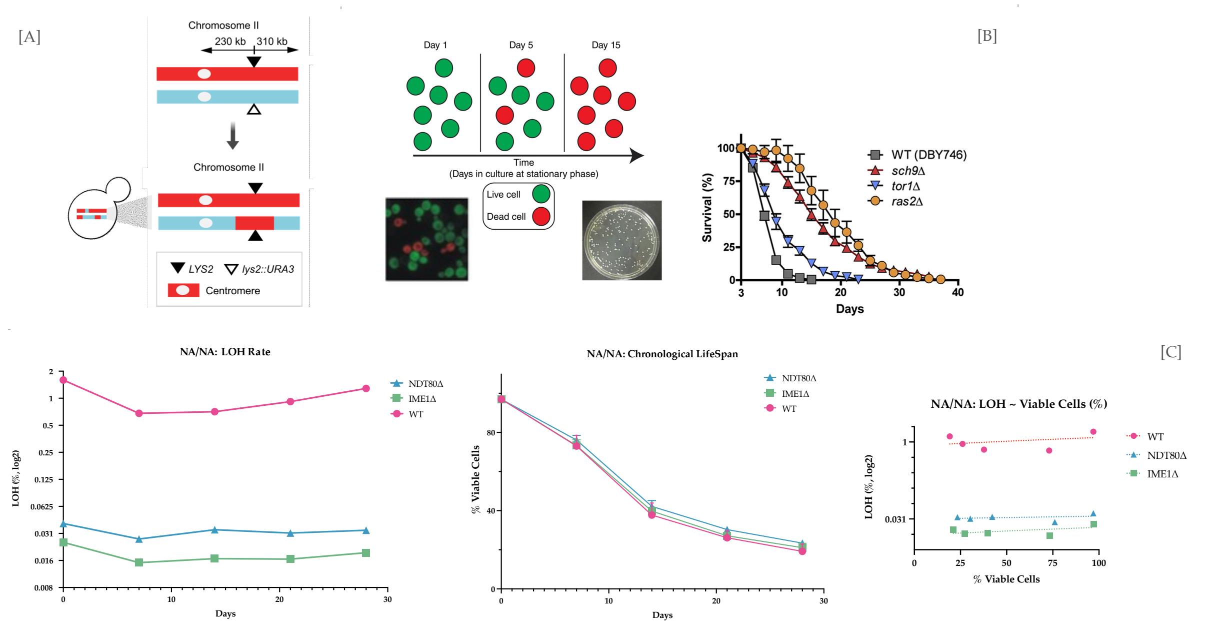


Figure 3: **CLS-LOH experiment: methods and preliminary results**.
A, URA3-LYS2 system used for measuring LOH rate
(**Source**: adapted from *Mozzachiodi & Tattini et al., 2021*).
B, Experimental procedure of CLS assay
(**Source**: adapted from *Chadwick et al., 2016; Kwolek-Mirek & Zadrag-Tecza., 2014; Longo et al., 2012).*
C, Preliminary results of my CLS-LOH assay, performed in the NA/NA yeast strain background, using three different genetic constructions (wild-type, NDT80 gene knock-out, and IME1 gene knock-out). From the left to the right: row LOH rate, NA/NA Cell Survival, and LOH rate ~ Cell Viability.

I hypothesized that yeast capacity to maintain genomic integrity during ageing should gradually decrease and may underlie or, at least, correlate with the variation in chronological lifespan of the different strains tested. In this context, the rate of LOH events, measured as described below (Equation 1), is expected to represent the level of genomic instability occurring during chronological lifespan, which should correlate with the loss of cell viability.

$$LOH(\%) = {}^{Number\ of\ cells\ in\ 5 - FoA}/_{Number\ of\ cells\ in\ YPD} \times 100$$

Equation 1: LOH rate calculation

Parallelly, in addition, I participated in a population genomics project whose scope was to understand how human domestication remodeled key features of the yeast life cycle by studying a large collection of 1011 natural yeast isolates (de Chiara and Barré et al., 2022). Relying on a gene candidate approach and on Genome-Wide Association studies (GWAS), the final objective was to identify and validate the genetic variants driving this domestication syndrome. In this context, I had the opportunity to learn innovative techniques of molecular biology such as cloning and genome editing. Then, I leveraged these techniques to validate by genetic engineering multiple candidate polymorphisms falling in different categories: single nucleotide variants, Copy Number Variations, gene presence/absence, and losses of function. More specifically, I focused on the validation of CLS candidates, for which we mapped potential causative variants in the genes WHI2 and HPF1 (Figure 4A). We found loss of function mutations in the transcription factor WHI2, which were associated with short living isolates. I substituted the impaired WHI2 allele with a functional one, which rescued the short lifespan of the strains tested (Figure 4B). Therefore, I could validate the effect of WHI2 loss-of-function on CLS as predicted in silico. Moreover, one of the strongest hits mapped as a genetic determinant of the CLS variation, was the presence/absence of a sequence we called HPF1-*like*, which was also associated with shorter lifespan when present. This gene corresponds to a functionally unknown open reading frame (ORF) that shares identity with the gene HPF1 that encodes for a cell wall protein. Accordingly, to test the detrimental incidence of HPF1-*like* on CLS, I fully deleted this gene in five different strain backgrounds. In three out of five strains, I observed a significant lifespan extension after removing HPF1-*like*. Likewise, I was able to relate the presence of HPF1-*like* to chronological life span shortening as predicted, even though the different results observed across genetic backgrounds suggest the existence of epistatic interactions that may buffer the effect of this gene (Figure 4C).

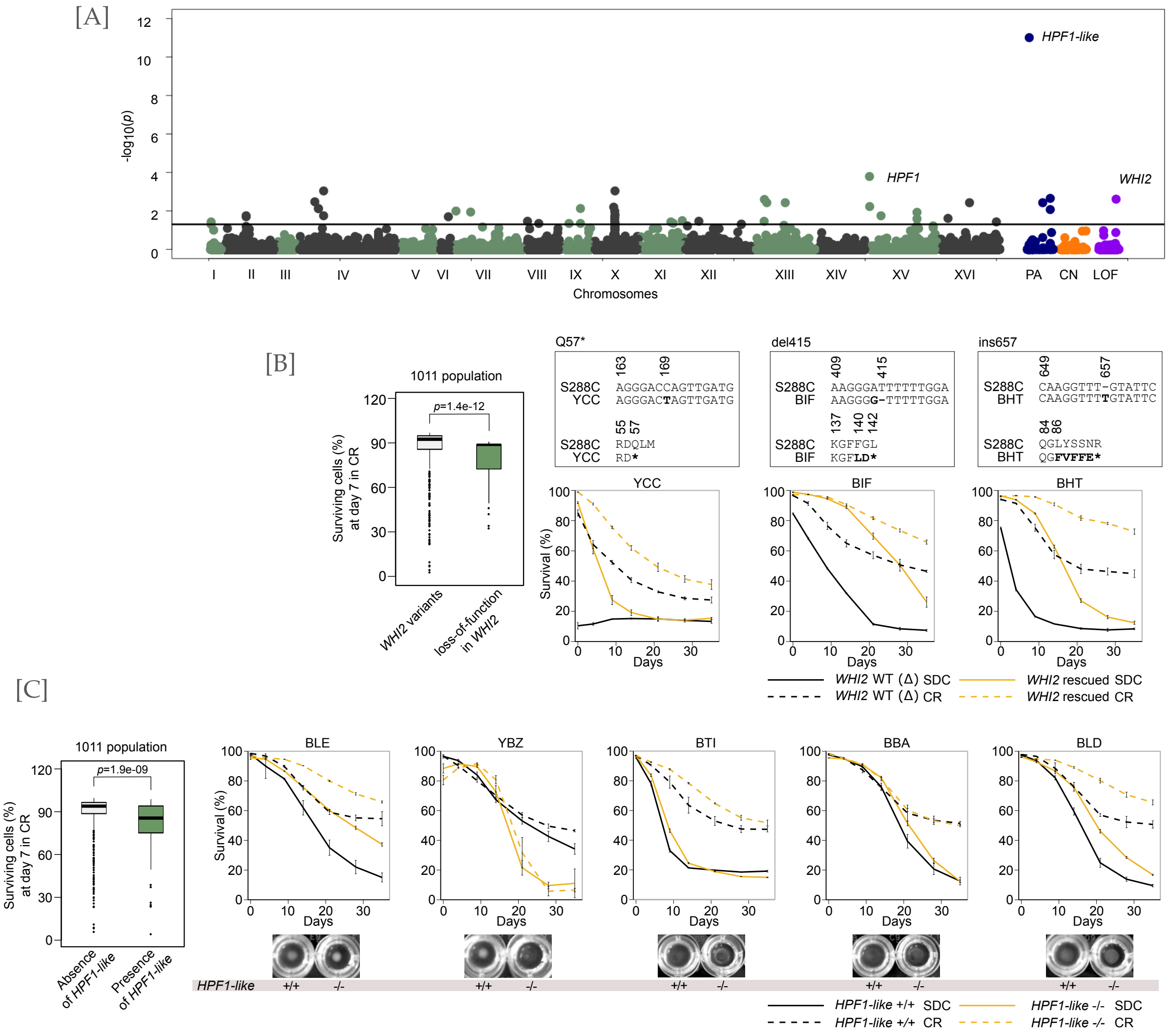


Figure 4: **Genetics variants impairing yeast CLS**. **A**, GWAS-Manhattan plot showing different CLS hits (experimental condition: Caloric Restriction, Day7). **B**, WHI2 Loss-of-function variants depicted in the upper panel have been associated with yeasts lifespan shortening. The replacement of the functional allele extended the lifespan rate in 3 out of the 4 tested strains. **C**, HPF1-*like* presence/absence variant have been associated to a yeast life span shortening. Knock-out of HPF1-*like* extended the lifespan rate in 3 out of 5 tested strains.
(**Source**: adapted from *de Chiara & Barré et al., 2022*)

References:

Allen, C., Büttner, S., Aragon, A. D., Thomas, J. A., Meirelles, O., Jaetao, J. E., ... & Werner-Washburne, M. (2006). Isolation of quiescent and nonquiescent cells from yeast stationary-phase cultures. The Journal of cell biology, 174(1), 89-100.

Alvers, A.L., Fishwick, L.K., Wood, M.S., Hu, D., Chung, H.S., Dunn Jr, W.A. and Aris, J.P., 2009. Autophagy and amino acid homeostasis are required for chronological longevity in Saccharomyces cerevisiae. Aging cell, 8(4), pp.353-369.

Ashrafi, K., Sinclair, D., Gordon, J.I. and Guarente, L., 1999. Passage through stationary phase advances replicative aging in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences, 96(16), pp.9100-9105.

Barré, B. P., Hallin, J., Yue, J. X., Persson, K., Mikhalev, E., Irizar, A., ... & Liti, G. (2020). Intragenic repeat expansion in the cell wall protein gene HPF1 controls yeast chronological aging. Genome research, 30(5), 697-710.

Botstein, D., & Fink, G. R. (2011). Yeast: an experimental organism for 21st Century biology. Genetics, 189(3), 695-704.

Campos, S. E., Avelar-Rivas, J. A., Garay, E., Juárez-Reyes, A., & DeLuna, A. (2018). Genomewide mechanisms of chronological longevity by dietary restriction in budding yeast. Aging cell, 17(3), e12749.

Coluccio, A. E., Rodriguez, R. K., Kernan, M. J., & Neiman, A. M. (2008). The yeast spore wall enables spores to survive passage through the digestive tract of Drosophila. PLoS One, 3(8), e2873.

De Chiara, M., Barré, B. P., Persson, K., Irizar, A., Vischioni, C., Khaiwal, S., ... & Liti, G. (2022). Domestication reprogrammed the budding yeast life cycle. Nature Ecology & Evolution, 6(4), 448-460.

De Virgilio, C. (2012). The essence of yeast quiescence. FEMS microbiology reviews, 36(2), 306-339.

Denoth Lippuner, A., Julou, T., & Barral, Y. (2014). Budding yeast as a model organism to study the effects of age. FEMS microbiology reviews, 38(2), 300-325.

Fabrizio, P., & Longo, V. D. (2003). The chronological life span of Saccharomyces cerevisiae. Aging cell, 2(2), 73-81.

Fabrizio, P., Battistella, L., Vardavas, R., Gattazzo, C., Liou, L. L., Diaspro, A., ... & Longo, V. D. (2004). Superoxide is a mediator of an altruistic aging program in Saccharomyces cerevisiae. The Journal of cell biology, 166(7), 1055-1067.

Fabrizio, P., & Longo, V. D. (2008). Chronological aging-induced apoptosis in yeast. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 1783(7), 1280-1285.

Fabrizio, P., Hoon, S., Shamalnasab, M., Galbani, A., Wei, M., Giaever, G., ... & Longo, V. D. (2010). Genome-wide screen in Saccharomyces cerevisiae identifies vacuolar protein sorting, autophagy, biosynthetic, and tRNA methylation genes involved in life span regulation. PLoS genetics, 6(7), e1001024.

Garay, E., Campos, S. E., Gonzalez de la Cruz, J., Gaspar, A. P., Jinich, A., & DeLuna, A. (2014). High-resolution profiling of stationary-phase survival reveals yeast longevity factors and their genetic interactions. PLoS genetics, 10(2), e1004168.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., ... & Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. nature, 418(6896), 387-391.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., ... & Oliver, S. G. (1996). Life with 6000 genes. Science, 274(5287), 546-567.

Gray, J. V., Petsko, G. A., Johnston, G. C., Ringe, D., Singer, R. A., & Werner-Washburne, M. (2004). "Sleeping beauty": quiescence in Saccharomyces cerevisiae. Microbiology and molecular biology reviews, 68(2), 187-206.

Gresham, D., Boer, V. M., Caudy, A., Ziv, N., Brandt, N. J., Storey, J. D., & Botstein, D. (2011). System-level analysis of genes and functions affecting survival during nutrient starvation in Saccharomyces cerevisiae. Genetics, 187(1), 299-317.

Heinicke, S., Livstone, M. S., Lu, C., Oughtred, R., Kang, F., Angiuoli, S. V., ... & Dolinski, K. (2007). The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. PloS one, 2(8), e766.

Ho, C. H., Magtanong, L., Barker, S. L., Gresham, D., Nishimura, S., Natarajan, P., ... & Boone, C. (2009). A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. Nature biotechnology, 27(4), 369-377.

Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., & O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. Nature, 425(6959), 686-691.

Ito, H., Fukuda, Y. A. S. U. K. I., Murata, K., & Kimura, A. (1983). Transformation of intact yeast cells treated with alkali cations. Journal of bacteriology, 153(1), 163-168.

Johnson, T. E. (1990). Increased life-span of age-1 mutants in Caenorhabditis elegans and lower Gompertz rate of aging. Science, 249(4971), 908-912.

Kaeberlein, M. (2010). Lessons on longevity from budding yeast. Nature, 464(7288), 513-519.

Kenyon, C., Chang, J., Gensch, E., Rudner, A., & Tabtiang, R. (1993). A C. elegans mutant that lives twice as long as wild type. Nature, 366(6454), 461-464.

Kenyon, C. J. (2010). The genetics of ageing. Nature, 464(7288), 504-512.

Laschober, G. T., Ruli, D., Hofer, E., Muck, C., Carmona-Gutierrez, D., Ring, J., ... & Jansen-Dürr, P. (2010). Identification of evolutionarily conserved genetic regulators of cellular aging. Aging cell, 9(6), 1084-1097.

Lee, S. S., Vizcarra, I. A., Huberts, D. H., Lee, L. P., & Heinemann, M. (2012). Whole lifespan microscopic observation of budding yeast aging through a microfluidic dissection platform. Proceedings of the National Academy of Sciences, 109(13), 4916-4920.

Lindstrom, D. L., Leverich, C. K., Henderson, K. A., & Gottschling, D. E. (2011). Replicative age induces mitotic recombination in the ribosomal RNA gene cluster of Saccharomyces cerevisiae. PLoS genetics, 7(3), e1002015.

Longo, V.D., Liou, L.L., Valentine, J.S. and Gralla, E.B., 1999. Mitochondrial superoxide decreases yeast survival in stationary phase. Archives of biochemistry and biophysics, 365(1), pp.131-142.

Longo, V. D., Shadel, G. S., Kaeberlein, M., & Kennedy, B. (2012). Replicative and chronological aging in Saccharomyces cerevisiae. Cell metabolism, 16(1), 18-31.

MacLean, M., Harris, N., & Piper, P. W. (2001). Chronological lifespan of stationary phase yeast cells; a model for investigating the factors that might influence the ageing of postmitotic tissues in higher organisms. Yeast, 18(6), 499-509.

Maclean, M. J., Aamodt, R., Harris, N., Alseth, I., Seeberg, E., Bjørås, M., & Piper, P. W. (2003). Base excision repair activities required for yeast to attain a full chronological life span. Aging cell, 2(2), 93-104.

McMurray, M.A. and Gottschling, D.E., 2003. An age-induced switch to a hyper-recombinational state. Science, 301(5641), pp.1908-1911.

Mirisola, M. G., Braun, R. J., & Petranovic, D. (2014). Approaches to study yeast cell aging and death. FEMS yeast research, 14(1), 109-118.

Mortimer, R. K., & Johnston, J. R. (1959). Life span of individual yeast cells. Nature, 183(4677), 1751-1752.

Mozzachiodi, S., Tattini, L., Llored, A., Irizar, A., Škofljanc, N., D'angiolo, M., ... & Liti, G. (2021). Aborting meiosis allows recombination in sterile diploid yeast hybrids. Nature communications, 12(1), 1-13.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A. and Cruaud, C., 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature, 556(7701), pp.339-344.

Piper, P. W. (2011). Maximising the yeast chronological lifespan. Aging research in yeast, 145-159.

Powers, R. W., Kaeberlein, M., Caldwell, S. D., Kennedy, B. K., & Fields, S. (2006). Extension of chronological life span in yeast by decreased TOR pathway signaling. Genes & development, 20(2), 174-184.

Qin, H., Lu, M. and Goldfarb, D.S., 2008. Genomic instability is associated with natural life span variation in Saccharomyces cerevisiae. PloS one, 3(7), p.e2670.

Ruetenik, A., & Barrientos, A. (2015). Dietary restriction, mitochondrial function and aging: from yeast to humans. Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1847(11), 1434-1447.

Steinkraus, K. A., Kaeberlein, M., & Kennedy, B. K. (2008). Replicative aging in yeast: the means to the end. Annual review of cell and developmental biology, 24, 29-54.

Stumpferl, S. W., Brand, S. E., Jiang, J. C., Korona, B., Tiwari, A., Dai, J., ... & Jazwinski, S. M. (2012). Natural genetic variation in yeast longevity. Genome research, 22(10), 1963-1973.

Veatch, J.R., McMurray, M.A., Nelson, Z.W. and Gottschling, D.E., 2009. Mitochondrial dysfunction leads to nuclear genome instability via an iron-sulfur cluster defect. Cell, 137(7), pp.1247-1258.

Wasko, B. M., & Kaeberlein, M. (2014). Yeast replicative aging: a paradigm for defining conserved longevity interventions. FEMS yeast research, 14(1), 148-159.

Wauters, R., Britton, S. J., & Verstrepen, K. J. (2021). Old yeasts, young beer—The industrial relevance of yeast chronological life span. Yeast, 38(6), 339-351.

Wei, M., Fabrizio, P., Madia, F., Hu, J., Ge, H., Li, L. M., & Longo, V. D. (2009). Tor1/Sch9-regulated carbon source substitution is as effective as calorie restriction in life span extension. PLoS genetics, 5(5), e1000467.

Weinberger, M., Feng, L., Paul, A., Smith Jr, D.L., Hontz, R.D., Smith, J.S., Vujcic, M., Singh, K.K., Huberman, J.A. and Burhans, W.C., 2007. DNA replication stress is a determinant of chronological lifespan in budding yeast. PloS one, 2(8), p.e748.

Weinberger, M., Sampaio-Marques, B., Ludovico, P., & Burhans, W. C. (2013). DNA replication stress-induced loss of reproductive capacity in S. cerevisiae and its inhibition by caloric restriction. Cell Cycle, 12(8), 1189-1200.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., ... & Davis, R. W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. science, 285(5429), 901-906.

Zhang, Y., Luo, C., Zou, K., Xie, Z., Brandman, O., Ouyang, Q., & Li, H. (2012). Single cell analysis of yeast replicative aging using a new generation of microfluidic device. PloS one, 7(11), e48275.

Chapter V:

Discussion & Future Perspectives.

"It's easy to make things worse [with a single mutation], but it's much more difficult to make them better."

Vadim Gladyshev

Aging is considered one of the risk factors of cancer insurgence due to the mutational burden derived from cell division and DNA replication. By definition, the cancer process is triggered by a cell that begins to behave abnormally, thus starting to divide in an uncontrolled manner probably due to a somatic genomic alteration. Previous theories (Burnet., 1974; Kirkwood & Holliday., 1979) have predicted an inverse correlation between the rate of somatic mutation and the species lifespan (Cagan et al., 2022). Indeed, as an organism ages, its cells might constantly accumulate DNA damage in time. Theoretically, each cell has the same probability to develop cancer, converging towards the idea that organisms possessing more cells and higher life expectancy should be more prone to the disease development. Therefore, animal of extreme large size such as whales or elephants, which possess thousands more cells than humans, should have a much higher risk of developing cancer, and consequently a relatively shorter life span. As extensively described in the introductory chapter, however, the scenario is not always as theoretically thought. The non-correlation between cancer occurrence and body size among species has been coined as Peto's paradox, which states that, although the association between the size/longevity of a species and its cancer incidence should be positively correlated, in nature, this relationship is not always respected (Peto et al., 1975). According to this, since each cell division has an identical probability of generating these errors and bring to a malignant transformation, all those species representing positive longevity outliers in relation to their body-size, should present greater cancer rates than their smaller counterpart, mainly because of the higher number of cell divisions (Caulin & Maley, 2011). Indeed, the disparity with this Peto's null hypothesis, in which each cell is assumed to have an identical probability of malignant transformation, is particularly important for understanding additional and more efficient mechanisms serving as defense weapons against cancer onset. Since these rates are expected to raise the risk of cancer occurrence, special mechanisms must necessarily exist to make large and long-lived species different from all the others. Scientists often use logic in order to show relationships between the parts of an idea and the whole thought, but it is not always easy to find logic confirmations in Natural laws. As Albert Einstein would say, logic would bring us straight from A to B, while imagination would take us everywhere. Here, we are not speaking about pure imagination, but mainly referring to some unknown strategies whereby some animals have increased their size and life expectancy, overcoming the threshold of the disease onset, or increasing its suppression. Indeed, there are some animals that, in addition of being huge and make up by an enormous number of cells, not only survive to reach what we would call "old age", but they also show very low cancer incidence rates. What is currently known and accepted is that natural selection on large size and/or extended longevity must intrinsically walk parallel and inseparably from the anti-cancer defenses evolution. In this context, what we can learn from solving this enigma, could therefore greatly contribute to increase the understanding of natural anticancer mechanisms, which could potentially be exploited in the future (bio) medical research. In this context, comparative genomics is just one of the multiple methods and biological tool required to investigate the mechanisms underlying Peto's paradox. Few months ago, Vincze and collaborators (Vincze et al., 2022), published on the prestigious *Nature* journal one of the most thorough review and examination of the paradox across species to date, finally concluding and confirming the independence of the cancer risk from body size and life expectancy. Interestingly, according to the authors, the solution to the paradox lies in the co-evolution of potent cancer resistance strategies and tumor suppressive mechanisms, which appear to have evolved along with the high longevity of very large and long-living animals.

In the framework of this work, I have examined a precise and determined angle by which Peto's paradox can be resolved. On the one hand, this includes the duplication of tumor suppressor genes, whereas, more generally, on the other, it considers the overall alteration of the CNVs landscape of cancer-related and non-cancer-related genes as biomarkers of the species tumor predisposition. Gene duplication, in addition to being one of the fundamental strategies for the emergence of new genetic traits (Ohta, 1989), is also one of the main contributors to the adaptation to unfavorable conditions and environments, as well as responsible for the evolution and the maintenance of growth, development, and cell regulatory pathways (Magadum et al. 2013). Therefore, in order to shed light on what are the ***mechanisms underlying cancer resistance and high longevity of certain mammals***, it is fundamental and promising to investigate gene copy number alteration in these outliers organisms using a genomic approach, and, more specifically, performing comparative oncology analysis across species from an evolutionary and candidate gene perspectives (Abegglen et al., 2015; Caulin et al., 2015; Caulin and Maley, 2011; Fang et al., 2014; Gorbunova et al., 2014; Keane et al., 2015; Kim et al., 2011; Seim et al., 2013; Tollis et al., 2017; Vazquez et al., 2018; Vazquez and Lynch, 2021).

In **Chapter II**, I have presented and described VarNuCopy, the new database I have developed during the course of my doctoral studies, and which represents the first online database of CNVs across the animal kingdom (Vischioni et al., 2022), enabling visual, interactive exploration, and analysis of gene copies alterations landscape among multiple species. From the very beginning, the goal of my project was to build a useful tool freely available for the scientific community, which would have filled the the lack of a user-friendly instrument able to investigate the relationship between genes number of copies among species, having the ultimate aim of identifying new molecular markers involved in the processes of tumorigenesis. In fact, nowadays, VarNuCopy is the first tool allowing a multi-species gene copy number comparison, both for model and non-model organisms in biomedicine. More specifically, it exploits CNVs data to highlight genes relevance, supporting different types of statistical measures, graphs, and classification criteria for their effective visualization. Therefore, users can explore the platform performing customize research processes by repeating multiple cycles of visualization and interactive commands. With the help of my collaborators, I built our database following an Exploratory Data Analysis approach, thanks to which the platform is able to perform and easily return the statistical information related to the genes number of copies, thus comparing them among the different target groups. This type of interaction, coupled with the outputs of the DAMs reports, guides the users towards the discovery of new molecular mechanisms, or the deepening of those already known to be related to the cancer susceptibility of a species. The data underlying VarNuCopy have been obtained through a homemade script written in Perl 5.14 and Python 3, specifically coded to download CNVs values from Ensembl resources (http://www.ensembl.org). This allowed us to retrieve gene copy numbers across species without performing multiple pairwise and various genome alignments, thus using less computational resources. Moreover, these data are collected from CAFE algorithm estimations (De Bie et al., 2006), which has the advantage of including a priori the species phylogenetic tree and the related lineage information when outputting the CNVs data. As of today, the current version of VarNuCopy is freely available at http://isgroup.mat.unimore.it:8083/, but we plan to expand the functionalities of the database including other biological features such as networks and pathways information to the analysis and categorization, implementing new genes classification algorithms such as the machine learning ones, and exploring new data mining and data discovery approaches.

In **Chapter III** I have described the first results obtained thanks to the use of this new resource, which allowed me to identify the alteration of the microRNAs CNVs pattern as a hallmark of the cancer predisposition of a species. In particular, I developed a simple way to test the hypothesis that CNVs confer protection or increase vulnerability to cancer. In this context, we were able to identify, for the first time, an over-representation of the microRNAs CNV signature as enriched in cancer-related pathways. Although we have found statistical support in our hypothesis, this test did not pass the significance criteria after correction for multiple testing. However, we performed a generalized least squares (PGLS) phylogenetic method in order to establish the association between copy number and cancer incidence rates independently of the shared evolutionary history, the ancestry phenomena, or the species population structure. As a result, we obtained a significant and strong correlation between the two traits (adjusted $R^2 = 0.5173$; p-value = 0.01746), confirming our hypothesis that additional data, such as a higher number of species, may be sufficient to straighten the statistical power of the different tests. Moreover, another caveat of the work can be represented by the fact that, sometimes, genomic data from online repositories can be incomplete, thus carrying assembly errors which can be translated into a misidentification of real deletions or expansion events. Because of the poor quality of many non-model-organism genomes, noncoding RNAs has been almost completely uncovered in the context of Peto's paradox research. Conversely, thanks to the nature of our data, we were able to overcome these limitations, avoiding obtaining biased and unreliable CNVs data which can result from the commonly used BLAST alignment approaches. The Discussion paragraph included in **Chapter III** focused on giving examples from the literature of the identified miRNAs being involved in cancer. Here, we have classified miRNAs as oncogenes or tumor suppressors, justifying their specific behavior using different examples described in literature. We have chosen not to use an *a priori* classification in order to avoid failures in discovering new possible targets. Given the double side nature of microRNAs genes, which, as pointed above can act both as oncogenes or TSGs, we preferred to perform firstly the statistical analysis to identify the target, and secondly verify its possible functioning as onco-miRNA, or onco-suppressor. We think that this kind of approach allowed us not only to include as significant results those microRNAs not usually known as cancer genes, but also to straighten our hypothesis that, indeed, microRNAs genes function as hallmarks and discriminants of a species cancer predisposition.

An interesting future direction of this work would include an evolutionary analysis of duplicated genes relative to each other. Throughout the whole manuscript, I categorized the identified molecular targets in three different groups, represented by tumor suppressors, oncogenes, and other non-cancer-related genes. However, the dividing line between these classes is far from being defined. Indeed, as it also happens in the case of the microRNAs family, many genes can act as either suppressors or oncogenes, depending, for example, on their expression condition or the variation signatures they are subjected with time. A recent paper published by João Pedro de Magalhães (de Magalhães, 2021), has highlighted that, nowadays, relying on literature and public databases, such as PubMed (https://pubmed.ncbi.nlm.nih.gov/), any human gene have been studied and possibly associated in the context of cancer. Surprisingly, according to his work, the 87.7% of human genes reported in scientific literature, can be also found in at least one paper mentioning cancer, underlying that, most of the genes has been already studied in the context of this disease. Indeed, I do agree with the author in thinking that the real challenge of our genomics era is to identify which are the *real* key players acting in the disease, and to determine which are the promising therapeutic targets to direct the efforts and the resources of the scientific research.

Human genes orthologs and homologs were used for my analyses, but, in doing this, we made the assumption that they perform the same function in the other species. However, we lack the experimental evidence that miRNAs and genes we have identified share the same molecular role in all mammalian species. In this perspective, all our targets need to be tested and validated in order to confirm their involvement in tumorigenic events and maintenance. Regarding this point, during my visiting period at IRCAN in Nice, besides the results I have already described along **Chapter IV**, I have also performed other experimental and computational works relevant for the final aim of this Ph.D thesis. Specifically, leveraging the CRISPR/Cas9 genome editing system (Hsu, Lander, and Zhang., 2014), I have started to set up a yeast-based genetic model to investigate the effect of copy number variation on candidate cancer-related genes, using chronologically aged cells. Briefly, in this system, Cas9 enzyme is used to cut the specific DNA region we are interested in, being leaded by a small guide RNA; once the sequence is targeted and cut, a synthetic cassette will, indeed, function as repair template in order to add, delete, or change the original DNA sequence, giving, as a result, the new edited fragment. In this framework, I have engineered some yeast strains with an additional copy of some of those genes found to be significant in the previously described analyses (**Chapter II** and **Chapter III**), such as HPA2, SCP1, and SMT3 yeasts orthologs/homologs genes (*p*-value of 0.007, 0.005 and 0.01 respectively). My final aim was to investigate the effect of the extra-copy on the yeast ageing processes, verifying if this may impair CLS, and finally measure its impact on the strain genomic instability. Unfortunately, mainly due to timing constraints, I was unable to complete these experiments. However, this is still a promising open project that I would like to finalize, and through which possibly validate my bioinformatics results.

References:

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., ... & Schiffman, J. D. (2015). Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. Jama, 314(17), 1850-1860.

Burnet, S. F. M. (1974). Intrinsic mutagenesis: a genetic approach to ageing. Lancaster: MTP.

Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T. H., Sanders, M. A., ... & Martincorena, I. (2022). Somatic mutation rates scale with lifespan across mammals. Nature, 1-8.

Caulin, A. F., Graham, T. A., Wang, L. S., & Maley, C. C. (2015). Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1673), 20140222.

Caulin, A. F., & Maley, C. C. (2011). Peto's Paradox: evolution's prescription for cancer prevention. Trends in ecology & evolution, 26(4), 175-182.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics, 22(10), 1269-1271.

de Magalhães, J. P. (2021). Every gene can (and possibly will) be associated with cancer. Trends in Genetics.

Fang, X., Nevo, E., Han, L., Levanon, E. Y., Zhao, J., Avivi, A., ... & Wang, J. (2014). Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. Nature communications, 5(1), 1-11.

Gorbunova, V., Seluanov, A., Zhang, Z., Gladyshev, V. N., & Vijg, J. (2014). Comparative genetics of longevity and cancer: insights from long-lived rodents. Nature Reviews Genetics, 15(8), 531-540.

Holtze, S., Gorshkova, E., Braude, S., Cellerino, A., Dammann, P., Hildebrandt, T. B., ... & Sahm, A. (2021). Alternative animal models of aging research. Frontiers in Molecular Biosciences, 8, 311.

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell, 157(6), 1262-1278.

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., ... & de Magalhães, J. P. (2015). Insights into the evolution of longevity from the bowhead whale genome. Cell reports, 10(1), 112-122.

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., ... & Gladyshev, V. N. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature, 479(7372), 223-227.

Kirkwood, T. B., & Holliday, R. (1979). The evolution of ageing and longevity. Proceedings of the Royal Society of London. Series B. Biological Sciences, 205(1161), 531-546.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. Journal of genetics, 92(1), 155-161.

Ohta, T. (1989). Role of gene duplication in evolution. Genome, 31(1), 304-310.

Peto, R., Roe, F. J., Lee, P. N., Levy, L., & Clack, J. (1975). Cancer and ageing in mice and men. British journal of cancer, 32(4), 411-426.

Seim, I., Fang, X., Xiong, Z., Lobanov, A. V., Huang, Z., Ma, S., ... & Gladyshev, V. N. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii. Nature communications, 4(1), 1-8.

Tollis, M., Schiffman, J. D., & Boddy, A. M. (2017). Evolution of cancer suppression as revealed by mammalian comparative genomics. Current opinion in genetics & development, 42, 40-47.

Vazquez, J. M., & Lynch, V. J. (2021). Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. Elife, 10, e65041.

Vazquez, J. M., Sulak, M., Chigurupati, S., & Lynch, V. J. (2018). A zombie LIF gene in elephants is upregulated by TP53 to induce apoptosis in response to DNA damage. Cell Reports, 24(7), 1765-1776.

Vincze, O., Colchero, F., Lemaître, J. F., Conde, D. A., Pavard, S., Bieuville, M., ... & Giraudeau, M. (2022). Cancer risk across mammals. Nature, 601(7892), 263-267.

Vischioni, C., Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2022). Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research. Big Data Research, 27, 100298.

Chapter VI:

Conclusion.

Age is one of the major risk factors for the development of cancer. With time, the interrelated connections between the mechanisms underlying aging and the ones responsible of tumorigenesis cause progressive deleterious changes affecting cells and organisms (Piano and Titorenko, 2015). Therefore, longer-living species must necessarily interface with the risk of fitness decrease, primarily due to changes in entropy levels or DNA molecular damages. In this perspective, comparative cancer research based on genetic and genomic investigations is a fundamental resource which can lead to the entanglement of the mechanisms behind cancer risk and development (Wong et al., 2019). Given the high heterogeneity level of tumor incidence values across species, to date, aging and cancer are studied using both standard and non-standard models, ranging from yeast and worms microorganisms (Kaphai et al., 2017; Longo et al., 2012), to bigger species such as mice or naked mole rats (Yuan et al., 2011; Ruby et al., 2018), elephants or whales (Vazquez and Lynch., et al., 2021; Abegglen et al., 2015; Keane et al., 2015). By exploiting the conserved features along the phylogenetic tree, multiple studies are shedding light on the protective mechanisms present in some species, with the final aim of improving our general understanding of cancer biology (Tollis, Schiffman, et al. 2017; Seluanov et al., 2018). Focusing on the resolution of Peto’s paradox, throughout this entire manuscript, I have examined different and unexplored area of comparative cancer genomics, coupling the evolutionary bioinformatic investigation of sequencing data with the development of new analytical tools. Indeed, in **Chapter II**, I have presented and described VarNuCopy, a new on-line resource representing the first database of CNVs across the animal kingdom (Vischioni et al., 2022). In **Chapter III** I have described for the first time the alteration of the microRNAs CNVs pattern as a hallmark of the cancer predisposition of a species. Finally, since most of the vertebrates aging hallmarks are also conserved in yeasts, **Chapter IV**, describes a new genetic system that I started to develop, which is useful to investigate the effect of copy number variation on candidate cancer-related genes and genomic instability. Altogether, I think that, nowadays, the current challenge is to develop and optimize new experimental design and strategies able to test the theoretical hypotheses deriving from this type of comparative studies, and to transfer the new acquisitive knowledge into the human (Holtze et al., 2021), and veterinary biomedical research. Indeed, whenever a potential cancer suppression mechanism is discovered in a species, there is the real possibility of identifying a new molecular target or therapeutic approach for the prevention of the disease. In conclusion, I personally believe that the bioinformatics results highlighted in this thesis can be seen as a first starting point for a new wave of research in CNVs studies, opening the door to a new era of experimental validation in the context of comparative cancer genomics and evolution.

References:

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., ... & Schiffman, J. D. (2015). Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. Jama, 314(17), 1850-1860.

Holtze, S., Gorshkova, E., Braude, S., Cellerino, A., Dammann, P., Hildebrandt, T.B., Hoeflich, A., Hoffmann, S., Koch, P., Terzibasi Tozzini, E., Skulachev, M., Skulachev, V.P., Sahm, A., 2021. Alternative Animal Models of Aging Research. Front. Mol. Biosci. 8, 660959.

Kapahi, P., Kaeberlein, M., & Hansen, M. (2017). Dietary restriction and lifespan: Lessons from invertebrate models. Ageing research reviews, 39, 3-14.

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., ... & de Magalhães, J. P. (2015). Insights into the evolution of longevity from the bowhead whale genome. Cell reports, 10(1), 112-122.

Longo, V. D., Shadel, G. S., Kaeberlein, M., & Kennedy, B. (2012). Replicative and chronological aging in Saccharomyces cerevisiae. Cell metabolism, 16(1), 18-31.

Piano, A., & Titorenko, V. I. (2015). The intricate interplay between mechanisms underlying aging and cancer. Aging and disease, 6(1), 56.

Ruby, J. G., Smith, M., & Buffenstein, R. (2018). Naked mole-rat mortality rates defy Gompertzian laws by not increasing with age. elife, 7, e31157.

Seluanov, A., Gladyshev, V. N., Vijg, J., & Gorbunova, V. (2018). Mechanisms of cancer resistance in long-lived mammals. Nature Reviews Cancer, 18(7), 433-441.

Tollis, M., Schiffman, J. D., & Boddy, A. M. (2017). Evolution of cancer suppression as revealed by mammalian comparative genomics. Current opinion in genetics & development, 42, 40-47.

Vazquez, J. M., & Lynch, V. J. (2021). Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. Elife, 10, e65041.

Vischioni, C., Bove, F., Mandreoli, F., Martoglia, R., Pisi, V., & Taccioli, C. (2022). Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research. Big Data Research, 27, 100298.

Wong, K., van der Weyden, L., Schott, C. R., Foote, A., Constantino-Casas, F., Smith, S., ... & Adams, D. J. (2019). Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. Nature communications, 10(1), 1-14.

Yuan, R., Peters, L. L., & Paigen, B. (2011). Mice as a mammalian model for research on the genetics of aging. ILAR journal, 52(1), 4-15.

Appendix A:

Supplementary material for Chapter III.

The following pages contain the supplementary material associated with chapter III.

Table S1: Species description. Phenotype characteristics the 9 species included in our analysis: name, mass (kg), maximum lifespan (years), metabolism (W), and “Cancer YES/NO” labelling.

Species	Common_name	Mass_kg	Lifespan_max_yr	Metabolism	Cancer
Heterocephalus_glaber	Naked_mole_rat	0.035	32	0.136	NO
Nannospalax_galili	Blind_mole_rat	0.325	21	0.585	NO
Dasypus_novemcinctus	Armadillo	3.95	22.34	4.55	NO
Loxodonta_africana	Elephant	4500	80	NA	NO
Myotis lucifugus	Little_brown_bat	0.01	34	0.051	NO
Mus_musculus	Mouse	0.02	6	0.271	YES
Rattus_norvegicus	Rat	0.28	5	1404	YES
Canis_familiaris	Dog	40	24	17.25	YES
Homo_sapiens	Human	70	80	76825	YES

[a] https://genomics.senescence.info/ (Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., ... & de Magalhães, J. P. (2018). Human ageing genomic resources: new and updated databases. Nucleic acids research, 46(D1), D1083-D1090.)

Table S2: Cancer Prone vs Cancer Resistant: a two-group statistical comparison. List of the significant hits resulting from the unpaired 2-group wilcoxon test (p-value <0.05) applied on the total genomic CNVs landscape of the selected species. Our analysis, which exclusively considered the variation in the number of gene copies within different species, was able to identify genes involved in biological processes related to cancer development and maintenance.

Cancer_rate (%)	0	0	2.7	4.81	0	43.5	87	23	22	
ID	[Hg]	[Ng]	[Dn]	[La]	[Ml]	[Mm]	[Rn]	[Cf]	[Hs]	p-value
CD52	0	0	0	0	0	1	1	1	1	0.0072
CSN1S1	0	0	0	0	0	1	1	1	1	0.0072
EEF1AKMT4	0	0	0	0	0	1	1	1	1	0.0072
GP2	0	0	0	0	0	1	1	1	1	0.0072
IQCM	0	0	0	0	0	1	1	1	1	0.0072
PCDHB14	0	0	0	0	0	1	1	1	1	0.0072
PCDHB7	0	0	0	0	0	1	1	1	1	0.0072
RNF224	0	0	0	0	0	1	1	1	1	0.0072
SAT1	2	2	2	2	2	3	3	3	3	0.0072
SATL1	2	2	2	2	2	3	3	3	3	0.0072
SMIM31	0	0	0	0	0	1	1	1	1	0.0072
ZNF169	0	0	0	0	0	1	1	1	1	0.0072
ABCG2	0	0	0	0	0	1	2	1	1	0.0093
MIR424	0	0	0	0	0	1	2	1	1	0.0093
SCARNA21	0	0	0	0	0	3	2	2	2	0.0093
DPPA3	0	0	0	0	0	1	8	2	1	0.0104
MIR371A	0	0	0	0	0	4	8	2	2	0.0104
MIR372	0	0	0	0	0	4	8	2	2	0.0104
CNGB1	1	1	0	1	1	2	2	2	2	0.0108
CNGB3	1	1	0	1	1	2	2	2	2	0.0108
GAR1	1	1	0	1	1	2	2	2	2	0.0108
GJA9	3	3	2	3	3	4	4	4	4	0.0108
GJD2	3	3	2	3	3	4	4	4	4	0.0108
GJD3	3	3	2	3	3	4	4	4	4	0.0108
GJD4	3	3	2	3	3	4	4	4	4	0.0108
GJE1	3	3	2	3	3	4	4	4	4	0.0108
NANOS1	1	2	2	2	2	3	3	3	3	0.0108
NANOS2	1	2	2	2	2	3	3	3	3	0.0108
NANOS3	1	2	2	2	2	3	3	3	3	0.0108
NOS3	1	2	2	2	2	3	3	3	3	0.0108
DMD	2	2	3	3	6	1	1	1	1	0.0142
SCARNA20	0	1	0	1	0	3	3	5	3	0.0155
SPAG11A	0	1	0	1	1	3	2	2	2	0.0155
SPAG11B	0	1	0	1	1	3	2	2	2	0.0155
EIF5	6	3	3	2	2	1	1	0	1	0.0170
EID1	2	1	1	1	1	3	3	2	3	0.0184
EID2	2	1	1	1	1	3	3	2	3	0.0184
EID2B	2	1	1	1	1	3	3	2	3	0.0184
PCDHGB1	0	0	0	0	1	2	2	1	2	0.0184
HIST1H4D	2	3	2	1	1	5	4	9	6	0.0189
HIST1H4J	2	3	2	1	1	5	4	9	6	0.0189
HIST2H4A	2	3	2	1	1	5	4	9	6	0.0189
HIST4H4	2	3	2	1	1	5	4	9	6	0.0189
RETN	1	2	1	1	1	4	4	3	2	0.0201
RETNLB	1	2	1	1	1	4	4	3	2	0.0201
SCARNA16	0	0	0	1	0	2	5	2	1	0.0201
MIR103A1	0	0	0	2	1	2	3	3	3	0.0219
MIR103A2	0	0	0	2	1	2	3	3	3	0.0219
MIR107	0	0	0	2	1	2	3	3	3	0.0219
MIR124-1	1	0	0	2	0	3	3	2	3	0.0219
MIR124-2	1	0	0	2	0	3	3	2	3	0.0219
MIR124-3	1	0	0	2	0	3	3	2	3	0.0219
RPLP1	4	4	4	3	7	2	3	1	1	0.0237
SUMO2	15	8	8	4	6	3	4	3	3	0.0237
SUMO3	15	8	8	4	6	3	4	3	3	0.0237
SUMO4	15	8	8	4	6	3	4	3	3	0.0237
ZFP41	1	2	1	1	0	5	2	7	22	0.0243
ZNF100	1	2	1	1	0	5	2	7	22	0.0243
ZNF114	1	2	1	1	0	5	2	7	22	0.0243
ZNF253	1	2	1	1	0	5	2	7	22	0.0243
ZNF257	1	2	1	1	0	5	2	7	22	0.0243
ZNF430	1	2	1	1	0	5	2	7	22	0.0243
ZNF431	1	2	1	1	0	5	2	7	22	0.0243
ZNF479	1	2	1	1	0	5	2	7	22	0.0243
CRYGA	1	0	0	0	0	1	2	2	1	0.0260
KRTAP24-1	4	3	3	3	3	5	4	5	4	0.0260
KRTAP3-1	4	3	3	3	3	5	4	5	4	0.0260
KRTAP3-2	4	3	3	3	3	5	4	5	4	0.0260
KRTAP3-3	4	3	3	3	3	5	4	5	4	0.0260
SPACA7	0	1	0	0	0	2	2	1	1	0.0260
TGIF2LX	0	0	0	1	0	2	1	1	2	0.0260
TGIF2LY	0	0	0	1	0	2	1	1	2	0.0260
CLEC2D	0	0	1	0	0	5	7	1	1	0.0267
CNTNAP3	0	1	0	0	0	1	1	2	3	0.0267
CNTNAP3B	0	1	0	0	0	1	1	2	3	0.0267
CNTNAP3C	0	1	0	0	0	1	1	2	3	0.0267
RHOXF1	0	0	0	0	1	2	1	3	1	0.0267
SCARNA14	0	0	0	1	0	3	4	1	1	0.0267
FAM237A	1	0	0	0	1	2	1	2	2	0.0282
FAM237B	1	0	0	0	1	2	1	2	2	0.0282
MIR506	0	0	0	5	0	2	11	6	11	0.0289
MIR509-1	0	0	0	5	0	2	11	6	11	0.0289
MIR511	0	0	0	5	0	2	11	6	11	0.0289
MIR514A1	0	0	0	5	0	2	11	6	11	0.0289
MIR514A3	0	0	0	5	0	2	11	6	11	0.0289
MIR514B	0	0	0	5	0	2	11	6	11	0.0289
SCARNA4	0	0	0	2	0	8	4	1	3	0.0297

DEFB105A	1	0	0	0	1	2	1	3	2	0.0304
DEFB105B	1	0	0	0	1	2	1	3	2	0.0304
DEFB130A	1	2	1	0	2	4	4	2	4	0.0304
DEFB130B	1	2	1	0	2	4	4	2	4	0.0304
DEFB4A	1	2	1	0	2	4	4	2	4	0.0304
DEFB4B	1	2	1	0	2	4	4	2	4	0.0304
FLG	1	2	3	3	2	4	4	3	4	0.0304
FLG2	1	2	3	3	2	4	4	3	4	0.0304
HRNR	1	2	3	3	2	4	4	3	4	0.0304
MIR378A	0	1	0	1	0	2	2	1	3	0.0304
MIR378B	0	1	0	1	0	2	2	1	3	0.0304
MIR378D2	0	1	0	1	0	2	2	1	3	0.0304
RPTN	1	2	3	3	2	4	4	3	4	0.0304
S100A16	1	2	3	3	2	4	4	3	4	0.0304
MBD1	3	3	4	3	5	5	5	6	9	0.0312
MBD2	3	3	4	3	5	5	5	6	9	0.0312
MBD3	3	3	4	3	5	5	5	6	9	0.0312
MBD3L1	3	3	4	3	5	5	5	6	9	0.0312
MBD3L2	3	3	4	3	5	5	5	6	9	0.0312
MBD3L2B	3	3	4	3	5	5	5	6	9	0.0312
MBD3L3	3	3	4	3	5	5	5	6	9	0.0312
MBD3L4	3	3	4	3	5	5	5	6	9	0.0312
MBD3L5	3	3	4	3	5	5	5	6	9	0.0312
PCM1	3	3	4	3	5	5	5	6	9	0.0312
C4orf3	0	0	0	1	0	1	1	2	1	0.0318
ERVFRD-1	0	1	0	0	0	3	1	1	1	0.0318
FGFBP1	1	1	2	1	1	2	2	2	3	0.0318
FGFBP2	1	1	2	1	1	2	2	2	3	0.0318
FGFBP3	1	1	2	1	1	2	2	2	3	0.0318
FOXJ1	1	2	2	2	2	1	0	1	1	0.0318
FRG2	0	0	1	0	0	1	1	1	3	0.0318
FRG2B	0	0	1	0	0	1	1	1	3	0.0318
FRG2C	0	0	1	0	0	1	1	1	3	0.0318
MIR1-1	0	0	0	3	0	3	3	4	3	0.0318
MIR1-2	0	0	0	3	0	3	3	4	3	0.0318
MIR206	0	0	0	3	0	3	3	4	3	0.0318
MIR340	0	0	0	1	0	1	2	1	1	0.0318
MIR542	0	0	0	1	0	1	3	1	1	0.0318
NUPR1	1	2	1	1	1	2	3	2	2	0.0318
NUPR2	1	2	1	1	1	2	3	2	2	0.0318
SELENOW	0	1	0	0	0	1	2	1	1	0.0318
SPINK14	0	0	1	0	0	1	3	1	1	0.0318
SYNE1	2	2	2	1	2	1	0	1	1	0.0318
CCDC8	6	8	8	5	6	13	9	8	14	0.0334
MOAP1	6	8	8	5	6	13	9	8	14	0.0334
PNMA1	6	8	8	5	6	13	9	8	14	0.0334
PNMA2	6	8	8	5	6	13	9	8	14	0.0334
PNMA3	6	8	8	5	6	13	9	8	14	0.0334
PNMA5	6	8	8	5	6	13	9	8	14	0.0334
PNMA6A	6	8	8	5	6	13	9	8	14	0.0334
PNMA6E	6	8	8	5	6	13	9	8	14	0.0334
PNMA6F	6	8	8	5	6	13	9	8	14	0.0334
PNMA8A	6	8	8	5	6	13	9	8	14	0.0334
PNMA8B	6	8	8	5	6	13	9	8	14	0.0334
PNMA8C	6	8	8	5	6	13	9	8	14	0.0334
ZCCHC12	6	8	8	5	6	13	9	8	14	0.0334
ZCCHC18	6	8	8	5	6	13	9	8	14	0.0334
ACOXL	1	0	0	0	0	1	1	1	1	0.0339
ADRA2C	1	0	0	0	0	1	1	1	1	0.0339
AGBL4	2	2	2	1	2	1	1	1	1	0.0339
AKAP14	0	0	1	0	0	1	1	1	1	0.0339
C2CD4C	0	0	1	0	0	1	1	1	1	0.0339
C3orf22	0	0	1	0	0	1	1	1	1	0.0339
C6orf201	0	0	0	0	1	1	1	1	1	0.0339
CCDC179	0	1	0	0	0	1	1	1	1	0.0339
CCDC185	0	0	1	0	0	1	1	1	1	0.0339
CNBD1	0	0	0	1	0	1	1	1	1	0.0339
COL27A1	0	0	0	0	1	1	1	1	1	0.0339
EXOC6	3	2	3	3	3	2	2	2	2	0.0339
EXOC6B	3	2	3	3	3	2	2	2	2	0.0339
FAM227B	0	0	1	0	0	1	1	1	1	0.0339
FMR1NB	0	0	0	1	0	1	1	1	1	0.0339
FOXL1	0	1	0	0	0	1	1	1	1	0.0339
GALP	0	1	0	0	0	1	1	1	1	0.0339
GAS1	0	0	1	0	0	1	1	1	1	0.0339
GFRA4	0	1	0	0	0	1	1	1	1	0.0339
GIMAP8	0	0	0	0	1	1	1	1	1	0.0339
GPR160	0	0	0	0	1	1	1	1	1	0.0339
GPR88	0	1	0	0	0	1	1	1	1	0.0339
HNRNPA0	2	2	2	1	2	1	1	1	1	0.0339
HSBP1L1	1	0	0	0	0	1	1	1	1	0.0339
IGIP	0	1	0	0	0	1	1	1	1	0.0339
IGKC	0	0	1	0	0	1	1	1	1	0.0339
IGSF23	0	0	0	0	1	1	1	1	1	0.0339
JUND	0	1	0	0	0	1	1	1	1	0.0339
KLF16	0	1	0	0	0	1	1	1	1	0.0339
KPRP	0	0	1	0	0	1	1	1	1	0.0339
KRT9	0	1	0	0	0	1	1	1	1	0.0339
LMLN2	0	0	1	0	0	1	1	1	1	0.0339
LRIT2	1	0	0	0	0	1	1	1	1	0.0339
LRWD1	1	1	2	1	1	2	2	2	2	0.0339
LSMEM2	0	0	0	1	0	1	1	1	1	0.0339
MAGEE2	0	0	0	1	0	1	1	1	1	0.0339
MAP4K4	2	2	1	2	2	1	1	1	1	0.0339
MIR100	0	0	0	3	0	3	3	3	3	0.0339

MIR10A	0	0	0	2	0	2	2	2	2	0.0339
MIR10B	0	0	0	2	0	2	2	2	2	0.0339
MIR1282	0	0	0	1	0	1	1	1	1	0.0339
MIR129-1	1	1	1	2	1	2	2	2	2	0.0339
MIR129-2	1	1	1	2	1	2	2	2	2	0.0339
MIR140	0	0	0	1	0	1	1	1	1	0.0339
MIR15A	1	1	1	2	1	2	2	2	2	0.0339
MIR15B	1	1	1	2	1	2	2	2	2	0.0339
MIR185	0	0	0	1	0	1	1	1	1	0.0339
MIR187	0	0	0	1	0	1	1	1	1	0.0339
MIR193A	0	0	0	2	0	2	2	2	2	0.0339
MIR193B	0	0	0	2	0	2	2	2	2	0.0339
MIR199A1	1	1	1	3	1	3	3	3	3	0.0339
MIR199A2	1	1	1	3	1	3	3	3	3	0.0339
MIR199B	1	1	1	3	1	3	3	3	3	0.0339
MIR203B	1	0	0	0	0	1	1	1	1	0.0339
MIR21	0	0	0	1	0	1	1	1	1	0.0339
MIR223	0	0	0	1	0	1	1	1	1	0.0339
MIR31	0	0	0	1	0	1	1	1	1	0.0339
MIR339	0	0	0	1	0	1	1	1	1	0.0339
MIR433	0	0	0	1	0	1	1	1	1	0.0339
MIR489	0	0	0	1	0	1	1	1	1	0.0339
MIR490	0	0	0	1	0	1	1	1	1	0.0339
MIR652	0	0	0	1	0	1	1	1	1	0.0339
MIR671	0	0	0	1	0	1	1	1	1	0.0339
MIR873	0	0	0	1	0	1	1	1	1	0.0339
MIR99A	0	0	0	3	0	3	3	3	3	0.0339
MIR99B	0	0	0	3	0	3	3	3	3	0.0339
MIRLET7A1	0	0	0	3	0	3	3	3	3	0.0339
MIRLET7A3	0	0	0	3	0	3	3	3	3	0.0339
MIRLET7F2	0	0	0	3	0	3	3	3	3	0.0339
NHLRC4	2	3	2	2	2	3	3	3	3	0.0339
NLRP10	0	0	0	0	1	1	1	1	1	0.0339
NPB	0	2	0	0	0	2	2	2	2	0.0339
NPW	0	2	0	0	0	2	2	2	2	0.0339
OPTC	1	1	2	1	1	2	2	2	2	0.0339
PBX4	0	1	0	0	0	1	1	1	1	0.0339
PCDHB16	0	0	0	0	1	1	1	1	1	0.0339
PCDHB6	1	0	0	0	0	1	1	1	1	0.0339
PHF3	2	2	2	1	2	1	1	1	1	0.0339
PMAIP1	0	1	0	0	0	1	1	1	1	0.0339
PMM1	3	3	3	2	3	2	2	2	2	0.0339
PMM2	3	3	3	2	3	2	2	2	2	0.0339
PPM1N	0	0	0	0	1	1	1	1	1	0.0339
PRR35	2	3	2	2	2	3	3	3	3	0.0339
PRR9	0	0	0	0	1	1	1	1	1	0.0339
RBM17	2	2	2	1	2	1	1	1	1	0.0339
RNF14	2	1	2	2	2	1	1	1	1	0.0339
RNF225	0	1	0	0	0	1	1	1	1	0.0339
SCFD2	1	2	2	2	2	1	1	1	1	0.0339
SP8	0	1	0	0	0	1	1	1	1	0.0339
SRRM5	0	0	1	0	0	1	1	1	1	0.0339
TEX11	0	0	0	0	1	1	1	1	1	0.0339
UFSP1	0	0	1	0	0	1	1	1	1	0.0339
UMOD	2	2	2	2	1	1	1	1	1	0.0339
UNCX	0	1	0	0	0	1	1	1	1	0.0339
UTS2B	0	0	0	1	0	1	1	1	1	0.0339
WFDC13	0	1	0	0	0	1	1	1	1	0.0339
ZBED9	0	0	1	0	0	1	1	1	1	0.0339
ZFP57	0	0	0	1	0	1	1	1	1	0.0339
HNRNPH1	4	3	3	3	4	2	2	3	2	0.0362
HNRNPH2	4	3	3	3	4	2	2	3	2	0.0362
MIR374A	1	0	0	1	1	2	1	2	2	0.0362
MIR374B	1	0	0	1	1	2	1	2	2	0.0362
PCDHGB4	1	0	1	0	1	2	2	1	2	0.0362
PCDHGB5	1	0	1	0	1	2	2	1	2	0.0362
ZNF383	1	1	1	0	0	1	2	2	2	0.0362
C10orf95	0	0	0	0	0	1	0	1	1	0.0369
C2orf92	0	0	0	0	0	1	0	1	1	0.0369
C6orf226	0	0	0	0	0	1	1	0	1	0.0369
CCDC192	0	0	0	0	0	1	0	1	1	0.0369
CFB	1	1	1	1	1	2	2	1	2	0.0369
CTXND2	0	0	0	0	0	1	0	1	1	0.0369
DEFB115	0	0	0	0	0	1	1	0	1	0.0369
DPEP2NB	0	0	0	0	0	1	0	1	1	0.0369
FDCSP	0	0	0	0	0	0	1	1	1	0.0369
FREM3	0	0	0	0	0	1	1	0	1	0.0369
HBE1	0	0	0	0	0	1	1	0	1	0.0369
HIGD2A	1	1	1	1	1	1	2	2	2	0.0369
HIGD2B	1	1	1	1	1	1	2	2	2	0.0369
IER2	2	2	2	2	2	3	2	3	3	0.0369
IER5	2	2	2	2	2	3	2	3	3	0.0369
IER5L	2	2	2	2	2	3	2	3	3	0.0369
KDM5C	1	1	1	1	1	2	2	1	2	0.0369
KDM5D	1	1	1	1	1	2	2	1	2	0.0369
LRPPRC	0	0	0	0	0	1	0	1	1	0.0369
LTB4R	1	1	1	1	1	1	2	2	2	0.0369
LTB4R2	1	1	1	1	1	1	2	2	2	0.0369
MFSD10	0	0	0	0	0	1	0	1	1	0.0369
MIR504	0	0	0	0	0	1	0	1	1	0.0369
MIR653	0	0	0	0	0	0	1	1	1	0.0369
MIR770	0	0	0	0	0	1	1	0	1	0.0369
PATE4	0	0	0	0	0	1	1	0	1	0.0369
PCDHGA6	0	0	0	0	0	1	1	0	1	0.0369
PCDHGA9	0	0	0	0	0	1	1	0	1	0.0369

PCDHGB6	0	0	0	0	0	1	1	0	1	0.0369
PCDHGB7	0	0	0	0	0	1	1	0	1	0.0369
PPP3R2	0	0	0	0	0	1	1	0	1	0.0369
PRR18	0	0	0	0	0	1	1	0	1	0.0369
SMIM36	0	0	0	0	0	1	0	1	1	0.0369
SMIM38	0	0	0	0	0	1	1	0	1	0.0369
SMIM41	0	0	0	0	0	1	1	0	1	0.0369
SRY	0	0	0	0	0	1	1	0	1	0.0369
TEX54	0	0	0	0	0	1	1	0	1	0.0369
TIMM10B	1	1	1	1	1	2	2	1	2	0.0369
TMEM249	0	0	0	0	0	1	0	1	1	0.0369
TMEM41A	1	1	1	1	1	0	0	0	1	0.0369
WFDC12	0	0	0	0	0	1	1	0	1	0.0369
ZBTB42	0	0	0	0	0	1	1	0	1	0.0369
ZNF142	0	0	0	0	0	1	1	0	1	0.0369
ZNF235	1	1	1	1	1	0	0	0	1	0.0369
EIF4A1	4	3	2	2	2	1	1	2	1	0.0371
FAM110A	3	4	3	3	2	4	4	5	4	0.0371
FAM110B	3	4	3	3	2	4	4	5	4	0.0371
FAM110C	3	4	3	3	2	4	4	5	4	0.0371
FAM110D	3	4	3	3	2	4	4	5	4	0.0371
MIR29B1	0	1	0	2	0	2	3	2	2	0.0371
MIR29B2	0	1	0	2	0	2	3	2	2	0.0371
PPP1R14A	3	4	2	4	4	5	5	5	4	0.0371
PPP1R14B	3	4	2	4	4	5	5	5	4	0.0371
PPP1R14C	3	4	2	4	4	5	5	5	4	0.0371
PPP1R14D	3	4	2	4	4	5	5	5	4	0.0371
MIR98	1	1	1	3	1	4	4	2	3	0.0389
MIRLET7G	1	1	1	3	1	4	4	2	3	0.0389
MIRLET7I	1	1	1	3	1	4	4	2	3	0.0389
SPDYE1	1	2	3	1	2	3	3	3	16	0.0389
SPDYE11	1	2	3	1	2	3	3	3	16	0.0389
SPDYE16	1	2	3	1	2	3	3	3	16	0.0389
SPDYE17	1	2	3	1	2	3	3	3	16	0.0389
SPDYE2	1	2	3	1	2	3	3	3	16	0.0389
SPDYE21P	1	2	3	1	2	3	3	3	16	0.0389
SPDYE2B	1	2	3	1	2	3	3	3	16	0.0389
SPDYE3	1	2	3	1	2	3	3	3	16	0.0389
SPDYE4	1	2	3	1	2	3	3	3	16	0.0389
SPDYE5	1	2	3	1	2	3	3	3	16	0.0389
SPDYE6	1	2	3	1	2	3	3	3	16	0.0389
SPDYE8P	1	2	3	1	2	3	3	3	16	0.0389
SPDYE9P	1	2	3	1	2	3	3	3	16	0.0389
DUSP18	3	1	11	2	2	1	1	0	1	0.0398
HLA-DRB1	3	2	2	2	5	1	0	0	2	0.0398
HLA-DRB5	3	2	2	2	5	1	0	0	2	0.0398
C15orf65	2	1	1	1	0	2	2	2	2	0.0400
C9orf116	2	1	1	1	0	2	2	2	2	0.0400
CRIPT	4	4	4	3	6	3	3	3	3	0.0400
DNAJC19	1	2	6	2	2	1	1	1	1	0.0400
GALR1	1	1	0	2	1	2	2	2	2	0.0400
GALR3	1	1	0	2	1	2	2	2	2	0.0400
INS	1	1	0	1	1	2	2	1	2	0.0400
INS-IGF2	1	1	0	1	1	2	2	1	2	0.0400
MIR221	1	0	1	2	1	2	2	2	2	0.0400
MIR222	1	0	1	2	1	2	2	2	2	0.0400
MIR23A	1	1	1	2	0	2	2	2	2	0.0400
MIR23B	1	1	1	2	0	2	2	2	2	0.0400
MIR27A	1	1	1	2	0	2	2	2	2	0.0400
MIR27B	1	1	1	2	0	2	2	2	2	0.0400
MIR30C1	1	0	1	2	1	2	2	2	2	0.0400
MIR30C2	1	0	1	2	1	2	2	2	2	0.0400
NACA	5	2	1	2	2	1	1	1	1	0.0400
POLR1C	4	4	4	3	6	3	3	3	3	0.0400
POLR2C	4	4	4	3	6	3	3	3	3	0.0400
PPP2R2A	11	5	5	4	5	4	4	4	4	0.0400
PPP2R2B	11	5	5	4	5	4	4	4	4	0.0400
PPP2R2C	11	5	5	4	5	4	4	4	4	0.0400
PPP2R2D	11	5	5	4	5	4	4	4	4	0.0400
RB1CC1	3	2	1	2	2	1	1	1	1	0.0400
RNF113A	1	1	0	1	2	2	2	2	2	0.0400
RNF113B	1	1	0	1	2	2	2	2	2	0.0400
SUB1	1	3	2	2	2	1	1	1	1	0.0400
ZNF112	1	1	1	0	1	2	2	1	2	0.0400
C19orf33	0	0	0	0	0	1	2	0	2	0.0404
C20orf141	0	0	0	0	0	1	0	1	2	0.0404
GBP2	0	0	0	0	0	2	1	0	1	0.0404
GOLGA2	1	1	1	1	1	2	2	1	19	0.0404
GOLGA6A	1	1	1	1	1	2	2	1	19	0.0404
GOLGA6B	1	1	1	1	1	2	2	1	19	0.0404
GOLGA6C	1	1	1	1	1	2	2	1	19	0.0404
GOLGA6D	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8A	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8B	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8F	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8G	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8H	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8J	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8K	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8M	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8N	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8O	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8Q	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8R	1	1	1	1	1	2	2	1	19	0.0404
GOLGA8S	1	1	1	1	1	2	2	1	19	0.0404

GOLGA8T	1	1	1	1	1	2	2	1	19	0.0404
HUS1	1	1	1	1	1	2	3	1	2	0.0404
HUS1B	1	1	1	1	1	2	3	1	2	0.0404
IGKV1OR2-108	0	0	0	0	0	1	2	0	1	0.0404
SEC61G	1	1	1	1	1	2	5	2	1	0.0404
SERPINA1	1	1	1	1	1	6	2	1	2	0.0404
SERPINA2	1	1	1	1	1	6	2	1	2	0.0404
SH2D1B	0	0	0	0	0	2	2	0	1	0.0404
SMIM40	0	0	0	0	0	2	0	1	2	0.0404
IGKV6-21	0	0	0	0	0	5	2	0	3	0.0416
IGKV6D-21	0	0	0	0	0	5	2	0	3	0.0416
IGKV6D-41	0	0	0	0	0	5	2	0	3	0.0416
RBMY1A1	0	0	0	0	0	10	1	0	6	0.0416
RBMY1B	0	0	0	0	0	10	1	0	6	0.0416
RBMY1D	0	0	0	0	0	10	1	0	6	0.0416
RBMY1E	0	0	0	0	0	10	1	0	6	0.0416
RBMY1F	0	0	0	0	0	10	1	0	6	0.0416
RBMY1J	0	0	0	0	0	10	1	0	6	0.0416
SEMG1	0	0	0	0	0	3	6	0	2	0.0416
SEMG2	0	0	0	0	0	3	6	0	2	0.0416
TRAV18	0	0	0	0	0	9	11	0	1	0.0416
ACKR4	4	2	3	3	2	4	4	4	4	0.0421
CUL2	4	2	2	1	4	1	1	1	1	0.0421
CXCR1	4	2	3	3	2	4	4	4	4	0.0421
CXCR2	4	2	3	3	2	4	4	4	4	0.0421
CXCR5	4	2	3	3	2	4	4	4	4	0.0421
MIR219A2	0	0	1	2	1	2	2	2	2	0.0421
MIR219B	0	0	1	2	1	2	2	2	2	0.0421
MIR30A	0	0	1	4	1	4	4	4	4	0.0421
MIR30B	0	0	1	4	1	4	4	4	4	0.0421
MIR30D	0	0	1	4	1	4	4	4	4	0.0421
MIR30E	0	0	1	4	1	4	4	4	4	0.0421
THOC7	3	3	2	1	2	1	1	1	1	0.0421
TRAPPC13	3	1	3	2	2	1	1	1	1	0.0421
AP3S1	9	5	4	4	3	3	3	3	3	0.0431
AP3S2	9	5	4	4	3	3	3	3	3	0.0431
AP4S1	9	5	4	4	3	3	3	3	3	0.0431
C1D	2	4	2	1	3	1	1	1	1	0.0431
CSNK1A1	5	4	2	1	2	1	1	1	1	0.0431
GLUL	6	2	3	1	2	1	1	1	1	0.0431
HNRNPA2B1	2	4	3	1	2	1	1	1	1	0.0431
LPA	1	1	1	1	0	3	2	1	2	0.0431
NUCKS1	3	6	7	2	3	2	2	2	2	0.0431
PRSS56	1	1	1	1	0	3	2	1	2	0.0431
RAD51AP1	3	6	7	2	3	2	2	2	2	0.0431
RWDD1	8	3	2	2	1	1	1	1	1	0.0431
TAF13	4	2	3	1	2	1	1	1	1	0.0431
TCAF1	2	2	1	2	2	3	4	2	3	0.0431
TCAF2	2	2	1	2	2	3	4	2	3	0.0431
TCAF2C	2	2	1	2	2	3	4	2	3	0.0431
PRELID1	11	5	6	4	8	4	4	4	4	0.0442
PRELID2	11	5	6	4	8	4	4	4	4	0.0442
PRELID3A	11	5	6	4	8	4	4	4	4	0.0442
PRELID3B	11	5	6	4	8	4	4	4	4	0.0442
SSX1	0	1	1	1	1	14	4	1	8	0.0442
SSX2	0	1	1	1	1	14	4	1	8	0.0442
SSX2B	0	1	1	1	1	14	4	1	8	0.0442
SSX3	0	1	1	1	1	14	4	1	8	0.0442
SSX4	0	1	1	1	1	14	4	1	8	0.0442
SSX4B	0	1	1	1	1	14	4	1	8	0.0442
SSX5	0	1	1	1	1	14	4	1	8	0.0442
SSX7	0	1	1	1	1	14	4	1	8	0.0442
KRTAP4-12	0	2	1	0	0	7	3	1	3	0.0444
KRTAP4-6	0	2	1	0	0	7	3	1	3	0.0444
KRTAP4-7	0	2	1	0	0	7	3	1	3	0.0444
ANKRD62	0	2	0	0	1	7	10	8	1	0.0453
ZNF195	3	5	0	0	0	21	7	13	3	0.0453
ZNF429	3	5	0	0	0	21	7	13	3	0.0453
TAS2R10	2	14	3	4	6	25	24	6	15	0.0491
TAS2R13	2	14	3	4	6	25	24	6	15	0.0491
TAS2R14	2	14	3	4	6	25	24	6	15	0.0491
TAS2R19	2	14	3	4	6	25	24	6	15	0.0491
TAS2R20	2	14	3	4	6	25	24	6	15	0.0491
TAS2R3	2	14	3	4	6	25	24	6	15	0.0491
TAS2R30	2	14	3	4	6	25	24	6	15	0.0491
TAS2R31	2	14	3	4	6	25	24	6	15	0.0491
TAS2R42	2	14	3	4	6	25	24	6	15	0.0491
TAS2R43	2	14	3	4	6	25	24	6	15	0.0491
TAS2R45	2	14	3	4	6	25	24	6	15	0.0491
TAS2R46	2	14	3	4	6	25	24	6	15	0.0491
TAS2R50	2	14	3	4	6	25	24	6	15	0.0491
TAS2R7	2	14	3	4	6	25	24	6	15	0.0491
TAS2R8	2	14	3	4	6	25	24	6	15	0.0491
TAS2R9	2	14	3	4	6	25	24	6	15	0.0491
CENPJ	1	2	2	1	1	4	2	2	4	0.0495
CMTM1	0	0	1	0	1	2	1	1	2	0.0495
CMTM2	0	0	0	1	1	2	2	1	1	0.0495
KRTAP15-1	0	1	1	0	0	2	1	2	1	0.0495
SCARNA2	0	1	0	1	0	2	2	1	1	0.0495
SCART1	0	0	1	0	1	2	2	1	1	0.0495
TCP10	1	2	2	1	1	4	2	2	4	0.0495
TCP10L	1	2	2	1	1	4	2	2	4	0.0495
TCP10L2	1	2	2	1	1	4	2	2	4	0.0495
ZNF420	1	0	1	0	0	2	1	1	2	0.0495

Table S3: Cancer Prone vs Cancer Resistant: a two-group statistical comparison. List of the significant microRNAs resulting from the unpaired 2-group wilcoxon test (p-value <0.05) applied on the total genomic CNVs landscape of the selected species.

ID	[Hg]	[Ng]	[Dn]	[La]	[Ml]	[Mm]	[Rn]	[Cf]	[Hs]	p-value
Cancer-	Resistant	Resistant	Resistant	Resistant	Resistant	Prone	Prone	Prone	Prone	
MIR424	0	0	0	0	0	1	2	1	1	0.0093
MIR371A	0	0	0	0	0	4	8	2	2	0.0104
MIR372	0	0	0	0	0	4	8	2	2	0.0104
MIR103A1	0	0	0	2	1	2	3	3	3	0.0219
MIR103A2	0	0	0	2	1	2	3	3	3	0.0219
MIR107	0	0	0	2	1	2	3	3	3	0.0219
MIR124-1	1	0	0	2	0	3	3	2	3	0.0219
MIR124-2	1	0	0	2	0	3	3	2	3	0.0219
MIR124-3	1	0	0	2	0	3	3	2	3	0.0219
MIR506	0	0	0	5	0	2	11	6	11	0.0289
MIR509-1	0	0	0	5	0	2	11	6	11	0.0289
MIR511	0	0	0	5	0	2	11	6	11	0.0289
MIR514A1	0	0	0	5	0	2	11	6	11	0.0289
MIR514A3	0	0	0	5	0	2	11	6	11	0.0289
MIR514B	0	0	0	5	0	2	11	6	11	0.0289
MIR378A	0	1	0	1	0	2	2	1	3	0.0304
MIR378B	0	1	0	1	0	2	2	1	3	0.0304
MIR378D2	0	1	0	1	0	2	2	1	3	0.0304
MIR1-1	0	0	0	3	0	3	3	4	3	0.0318
MIR1-2	0	0	0	3	0	3	3	4	3	0.0318
MIR206	0	0	0	3	0	3	3	4	3	0.0318
MIR340	0	0	0	1	0	1	2	1	1	0.0318
MIR542	0	0	0	1	0	1	3	1	1	0.0318
MIR100	0	0	0	3	0	3	3	3	3	0.0339
MIR10A	0	0	0	2	0	2	2	2	2	0.0339
MIR10B	0	0	0	2	0	2	2	2	2	0.0339
MIR1282	0	0	0	1	0	1	1	1	1	0.0339
MIR129-1	1	1	1	2	1	2	2	2	2	0.0339
MIR129-2	1	1	1	2	1	2	2	2	2	0.0339
MIR140	0	0	0	1	0	1	1	1	1	0.0339
MIR15A	1	1	1	2	1	2	2	2	2	0.0339
MIR15B	1	1	1	2	1	2	2	2	2	0.0339
MIR185	0	0	0	1	0	1	1	1	1	0.0339
MIR187	0	0	0	1	0	1	1	1	1	0.0339
MIR193A	0	0	0	2	0	2	2	2	2	0.0339
MIR193B	0	0	0	2	0	2	2	2	2	0.0339
MIR199A1	1	1	1	3	1	3	3	3	3	0.0339
MIR199A2	1	1	1	3	1	3	3	3	3	0.0339
MIR199B	1	1	1	3	1	3	3	3	3	0.0339
MIR203B	1	0	0	0	0	1	1	1	1	0.0339
MIR21	0	0	0	1	0	1	1	1	1	0.0339
MIR223	0	0	0	1	0	1	1	1	1	0.0339
MIR31	0	0	0	1	0	1	1	1	1	0.0339
MIR339	0	0	0	1	0	1	1	1	1	0.0339
MIR433	0	0	0	1	0	1	1	1	1	0.0339
MIR489	0	0	0	1	0	1	1	1	1	0.0339
MIR490	0	0	0	1	0	1	1	1	1	0.0339
MIR652	0	0	0	1	0	1	1	1	1	0.0339
MIR671	0	0	0	1	0	1	1	1	1	0.0339
MIR873	0	0	0	1	0	1	1	1	1	0.0339
MIR99A	0	0	0	3	0	3	3	3	3	0.0339
MIR99B	0	0	0	3	0	3	3	3	3	0.0339
MIRLET7A1	0	0	0	3	0	3	3	3	3	0.0339
MIRLET7A3	0	0	0	3	0	3	3	3	3	0.0339
MIRLET7F2	0	0	0	3	0	3	3	3	3	0.0339
MIR374A	1	0	0	1	1	2	1	2	2	0.0362
MIR374B	1	0	0	1	1	2	1	2	2	0.0362
MIR504	0	0	0	0	0	1	0	1	1	0.0369
MIR653	0	0	0	0	0	0	1	1	1	0.0369
MIR770	0	0	0	0	0	1	1	0	1	0.0369
MIR29B1	0	1	0	2	0	2	3	2	2	0.0371
MIR29B2	0	1	0	2	0	2	3	2	2	0.0371
MIR98	1	1	1	3	1	4	4	2	3	0.0389
MIRLET7G	1	1	1	3	1	4	4	2	3	0.0389
MIRLET7I	1	1	1	3	1	4	4	2	3	0.0389
MIR221	1	0	1	2	1	2	2	2	2	0.0400
MIR222	1	0	1	2	1	2	2	2	2	0.0400
MIR23A	1	1	1	2	0	2	2	2	2	0.0400
MIR23B	1	1	1	2	0	2	2	2	2	0.0400
MIR27A	1	1	1	2	0	2	2	2	2	0.0400
MIR27B	1	1	1	2	0	2	2	2	2	0.0400
MIR30C1	1	0	1	2	1	2	2	2	2	0.0400
MIR30C2	1	0	1	2	1	2	2	2	2	0.0400
MIR219A2	0	0	1	2	1	2	2	2	2	0.0421
MIR219B	0	0	1	2	1	2	2	2	2	0.0421
MIR30A	0	0	1	4	1	4	4	4	4	0.0421
MIR30B	0	0	1	4	1	4	4	4	4	0.0421
MIR30D	0	0	1	4	1	4	4	4	4	0.0421
MIR30E	0	0	1	4	1	4	4	4	4	0.0421

Table S3: Pathway analysis – extended version. Gene Over-Representation Analysis (ORA) using both Gene Ontology (biological processes) and Pathway analysis (KEGG, PANTHER, Reactome and Wikipathway) as enrichment categories. ORA was performed by the WebGesTalt functional enrichment analysis tool available at http://www.webgestalt.org. The enrichment test used Benjamini-Hochberg's FDR correction (FDR < 0.05). CNVs data were previously analyzed by an unpaired 2-group wilcoxon test (p-value < 0.05).

	Enrichment_category_1	Enrichment_category_2	Reference_list	FDR_method	Significance_level
A	gene_ontology	biological_proc	genome	BH	FDR<0.05
B	Pathway	Kegg	genome	BH	FDR<0.05
C	Pathway	Panther	genome	BH	FDR<0.05
D	Pathway	Reactome	genome	BH	FDR<0.05
E	Pathway	Wikipathway	genome	BH	FDR<0.05

	geneSet	description	enrichmentRatio	pValue	FDR	Genes
A	GO:0001580	detection of chemical stimulus involved in sensory perception of bitter taste	2.422E+13	2.220E-16	2.019E-12	TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	GO:0050913	sensory perception of bitter taste	2.197E+01	8.882E-16	4.037E-12	TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	GO:0050912	detection of chemical stimulus involved in sensory perception of taste	2.099E+01	1.887E-15	5.719E-12	TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	GO:0050909	sensory perception of taste	1.453E+01	5.317E-13	1.208E-09	TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	GO:0006346	methylation-dependent chromatin silencing	2.841E+01	1.365E-10	2.481E-07	MBD1;MBD2;MBD3;MBD3L1;MBD3L2;MBD3L3;MBD3L4;MBD3L5
	GO:0007030	Golgi organization	7.442E+00	1.679E-09	2.544E-06	SYNE1;GOLGA2;GOLGA6A;GOLGA6B;GOLGA6C;GOLGA6D;GOLGA8A;GOLGA8B;GOLGA8H;GOLGA8J;GOLGA8M;GOLGA8N;GOLGA8O;GOLGA8R;CSNK1A1
	GO:0051225	spindle assembly	7.762E+00	1.253E-08	1.628E-05	GOLGA2;GOLGA6A;GOLGA6B;GOLGA6C;GOLGA6D;GOLGA8A;GOLGA8B;GOLGA8H;GOLGA8J;GOLGA8M;GOLGA8N;GOLGA8O;GOLGA8R
	GO:0007051	spindle organization	5.012E+00	2.128E-06	2.418E-03	GOLGA2;GOLGA6A;GOLGA6B;GOLGA6C;GOLGA6D;GOLGA8A;GOLGA8B;GOLGA8H;GOLGA8J;GOLGA8M;GOLGA8N;GOLGA8O;GOLGA8R
	GO:0006342	chromatin silencing	7.687E+00	2.488E-06	2.513E-03	HIST4H4;MBD1;MBD2;MBD3;MBD3L1;MBD3L2;MBD3L3;MBD3L4;MBD3L5
B	geneSet	description	enrichmentRatio	pValue	FDR	Genes
	hsa05206	MicroRNAs in cancer	7.795E+00	0.000E+00	0.000E+00	MIR103A1;MIR103A2;MIR107;MIR124-1;MIR124-2;MIR124-3;MIR1-1;MIR1-2;MIR206;MIR100;MIR10A;MIR10B;MIR129-1;MIR129-2;MIR15A;MIR15B;MIR193B;MIR199A1;MIR199A2;MIR199B;MIR203B;MIR21;MIR223;MIR31;MIR99A;MIRLET7A1;MIRLET7A3;MIRLET7F2;MIR29B1;MIR29B2;MIRLET7G;MIRLET7I;MIR221;MIR222;MIR23A;MIR23B;MIR27A;MIR27B;MIR30C1;MIR30C2;MIR30A;MIR30B;MIR30D;MIR30E
	hsa04742	Taste transduction	1.021E+01	1.940E-12	3.163E-10	TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R42;TAS2R43;TAS2R45;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	hsa04914	Progesterone-mediated oocyte maturation	5.886E+00	2.238E-06	2.432E-04	SPDYE1;SPDYE11;SPDYE16;SPDYE17;SPDYE2;SPDYE2B;SPDYE3;SPDYE4;SPDYE5;SPDYE6;INS
	hsa04114	Oocyte meiosis	5.126E+00	3.346E-06	2.727E-04	PPP3R2;SPDYE1;SPDYE11;SPDYE16;SPDYE17;SPDYE2;SPDYE2B;SPDYE3;SPDYE4;SPDYE5;SPDYE6;INS
C	geneSet	description	enrichmentRatio	pValue	FDR	Genes
	P00012	Cadherin signaling pathway	3.808E+00	3.557E-04	4.020E-02	PCDHB14;PCDHB7;PCDHGB1;PCDHB16;PCDHB6;PCDHGB4;PCDHGA6;PCDHGB6;PCDHGB7
D	geneSet	description	enrichmentRatio	pValue	FDR	Genes
	R-HSA-420499	Class C/3 (Metabotropic glutamate/pheromone receptors)	2.65E+01	0.00E+00	0.00E+00	TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	R-HSA-500792	GPCR ligand binding	4.08E+00	7.15E-10	6.17E-07	ADRA2C;NPB;NPW;UTS2B;LTB4R;LTB4R2;GALR1;GALR3;ACKR4;CXCR1;CXCR2;CXCR5;TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	R-HSA-418594	G alpha (i) signalling events	4.17E+00	5.26E-09	3.03E-06	CNGB1;ADRA2C;NPB;NPW;GALR1;GALR3;CXCR1;CXCR2;CXCR5;TAS2R10;TAS2R13;TAS2R14;TAS2R19;TAS2R20;TAS2R3;TAS2R30;TAS2R31;TAS2R43;TAS2R46;TAS2R50;TAS2R7;TAS2R8;TAS2R9
	R-HSA-1461957	Beta defensins	1.20E+01	1.58E-06	6.83E-04	DEFB105A;DEFB105B;DEFB130A;DEFB130B;DEFB4A;DEFB4B;DEFB115
	R-HSA-1461973	Defensins	9.66E+00	6.98E-06	2.41E-03	DEFB105A;DEFB105B;DEFB130A;DEFB130B;DEFB4A;DEFB4B;DEFB115
	R-HSA-6803157	Antimicrobial peptides	5.92E+00	5.94E-05	1.71E-02	DEFB105A;DEFB105B;DEFB130A;DEFB130B;DEFB4A;DEFB4B;DEFB115;SEMG1
	R-HSA-3214842	HDMs demethylate histones	8.45E+00	7.08E-05	1.75E-02	HIST1H4D;HIST1H4J;HIST2H4A;HIST4H4;KDM5C;KDM5D
	R-HSA-6803204	TP53 Regulates Transcription of Genes Involved in Cytochrome C Release	1.44E+01	1.47E-04	3.18E-02	PMAIP1;PRELID1;PRELID3A;ZNF420
E	geneSet	description	enrichmentRatio	pValue	FDR	userId
	WP1545	miRNAs involved in DNA damage response	1.323E+01	7.077E-12	3.758E-09	MIR371A;MIR372;MIR542;MIR100;MIR15B;MIRLET7A1;MIR374B;MIR221;MIR222;MIR23A;MIR23B;MIR27A;MIR27B
	WP2059	Alzheimers Disease	5.088E+00	2.000E-07	5.311E-05	MIR124-1;MIR124-2;MIR124-3;MIR10A;MIR129-1;MIR129-2;MIR199B;MIR21;MIR433;MIR671;MIR873;PPP3R2;MIR29B1;MIR30C2;MIR219A2
	WP2249	Metastatic brain tumor	1.090E+01	1.365E-05	2.307E-03	MIRLET7A1;MIRLET7A3;MIRLET7F2;MIR29B1;MIR29B2;MIRLET7G
	WP2911	miRNA targets in ECM and membrane receptors	8.282E+00	1.738E-05	2.307E-03	MIR107;MIR15B;MIR30C1;MIR30C2;MIR30B;MIR30D;MIR30E
	WP1544	MicroRNAs in cardiomyocyte hypertrophy	5.037E+00	2.607E-05	2.768E-03	MIR103A1;MIR103A2;MIR140;MIR15B;MIR185;MIR199A1;MIR199A2;MIR23A;MIR27B;MIR30E
	WP2029	Cell Differentiation - Index	7.268E+00	1.516E-04	1.251E-02	MIR1-1;MIR206;MIR199A1;MIR199A2;MIR221;MIR222
	WP3299	let-7 inhibition of ES cell reprogramming	1.357E+01	1.649E-04	1.251E-02	MIRLET7A1;MIRLET7F2;MIRLET7G;MIRLET7I
	WP4329	miRNAs involvement in the immune response in sepsis	5.653E+00	2.151E-04	1.428E-02	MIR187;MIR199A1;MIR199A2;MIR203B;MIR223;MIR29B1;MIRLET7I
	WP2023	Cell Differentiation - Index expanded	6.105E+00	4.038E-04	2.383E-02	MIR1-1;MIR206;MIR199A1;MIR199A2;MIR221;MIR222
	WP3971	Role of Osx and miRNAs in tooth development	7.066E+00	6.301E-04	3.346E-02	MIRLET7A1;MIRLET7F2;MIR29B1;MIRLET7G;MIRLET7I

Supplementary Data S1:

Statistical results of the PGLS model correlating Cancer incidence rate ~ Number of significant microRNAs copies across the 9 species included in the analysis, which has been applied in order to check for potential bias due to species phylogeny or population structure.

```
Model <- pgls(cancer_rate ~ CNV, lambda="ML", data = comp.data)
summary(Model)


pgls(formula = cancer_rate ~ CNV, data = comp.data, lambda =
"ML")
kappa[Fix]: 1.000
lambda[ ML]: 0.460
delta[Fix]: 1.000
```

Coefficients:

- Residual standard error: 2.235 on 7 degrees of freedom
- Multiple R-squared: 0.5776,
- Adjusted R-squared: 0.5173
- F-statistic: 9.574 on 1 and 7 DF
- p-value: **0.01746**

Supplementary Data S2-S4.

The heatmap plots shown below highlight the relationships between individual data points, and the corresponding relationships within clusters.

Each group has a distinct set of copy number values, and the main branches representing cancer- prone and resistant organisms perfectly distinguish the species. No additional information was given to the algorithm other than copy number. The model was able to discriminate between the two groups only using CNVs data. Euclidean distances, with both 'complete' and 'ward' methods have been applied.

Figure S2: Heatmap of all the significant genes, clustered with Euclidean distance and ward linkage.

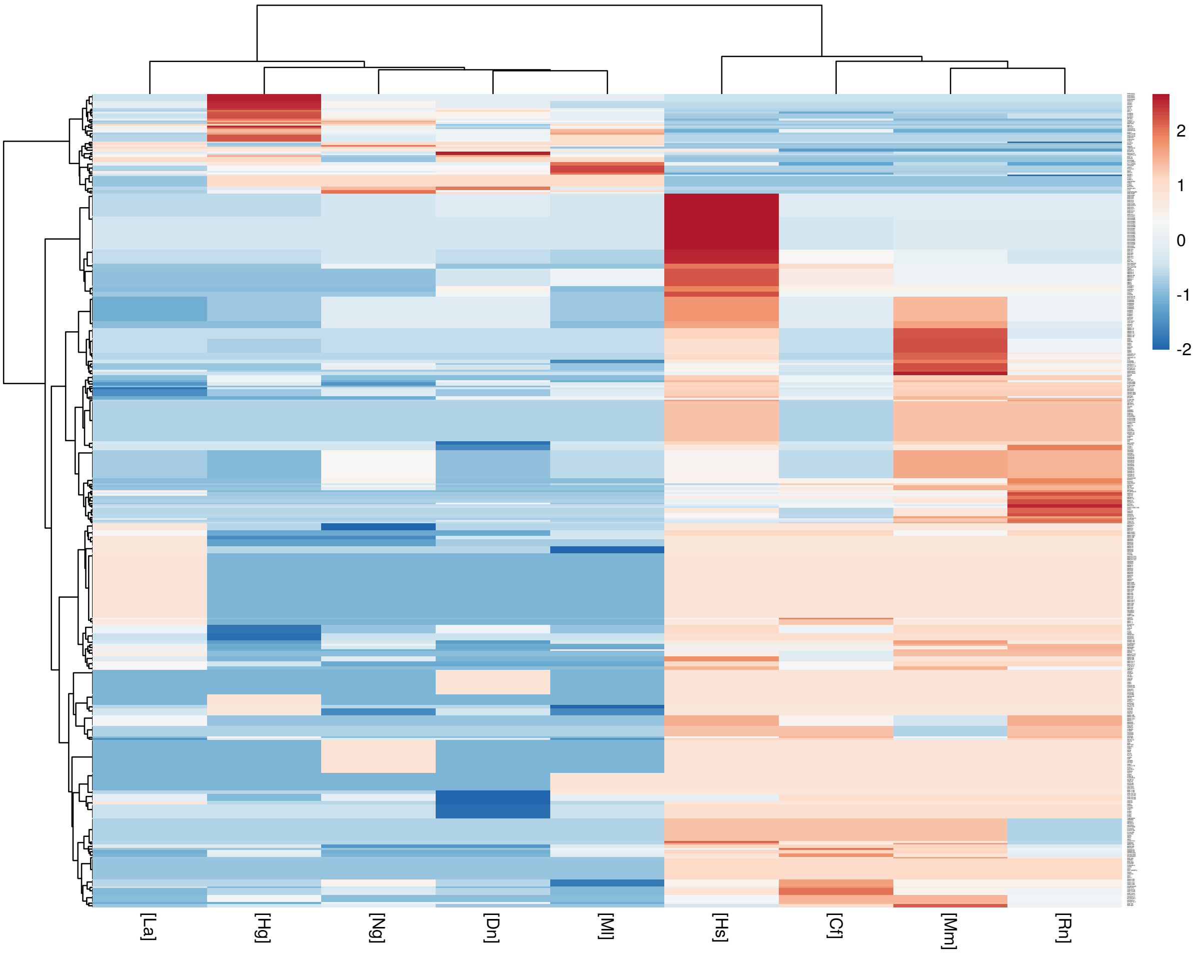


Figure S3: Heatmap of all the significant genes, clustered with Euclidean distance and complete linkage.

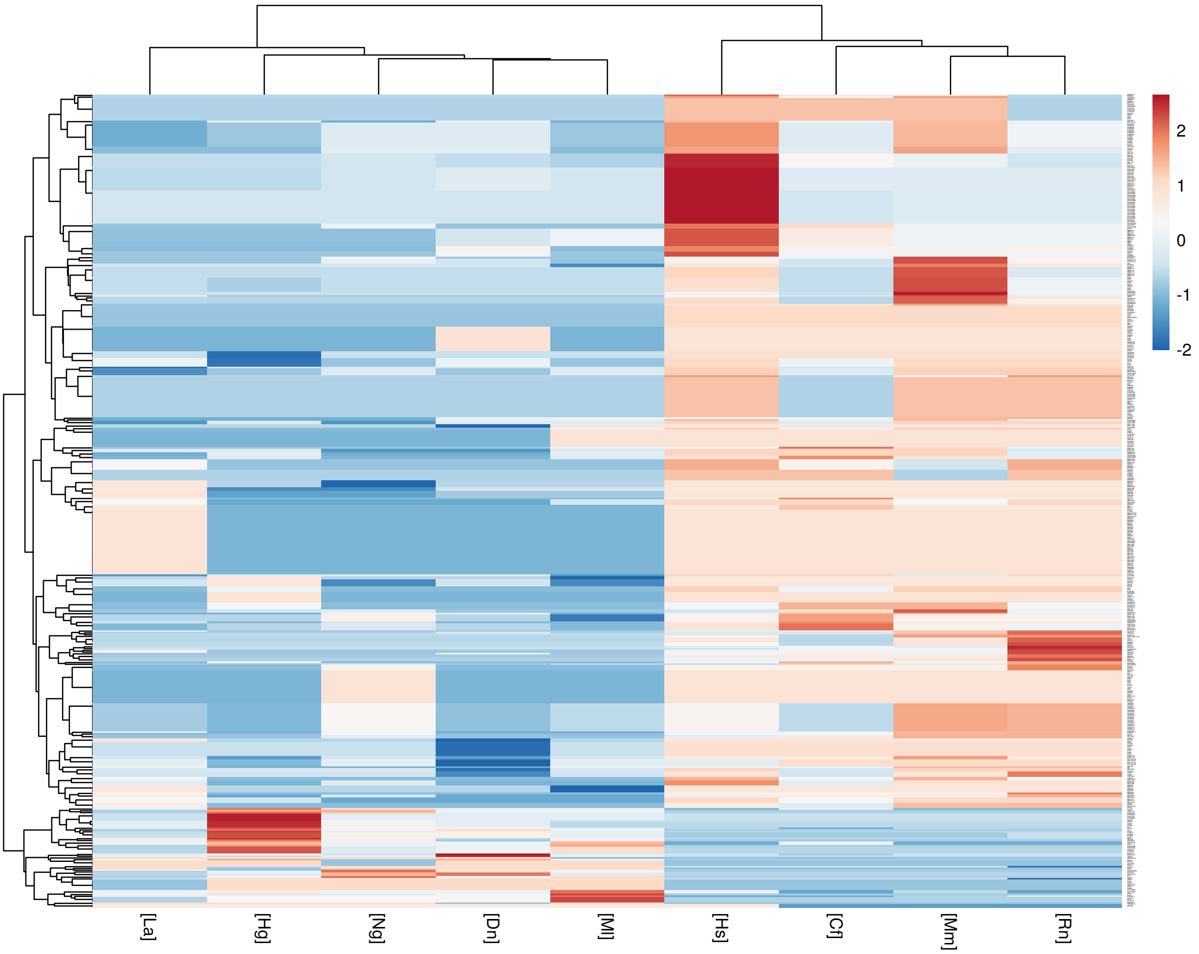
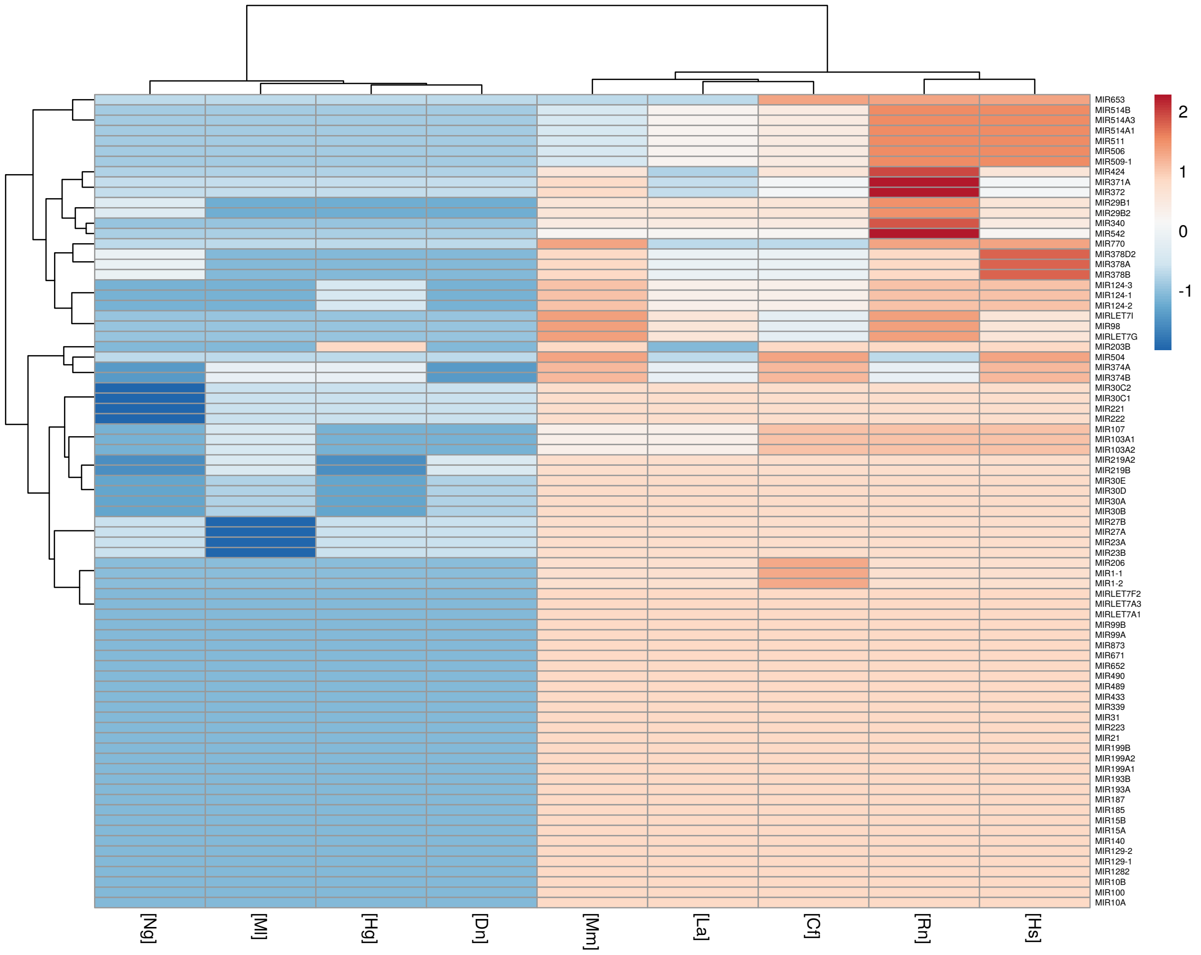


Figure S4: Heatmap of the significant MicroRNAs, clustered with Euclidean distance and ward linkage.



“Per aspera, sic itur ad astra.”