Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Programa de Doctorado en Biomedicina y Oncología Molecular

Comparative analysis of the Degradome of long-lived metazoans

Doctoral Thesis

José María González Pérez-Silva

October, 2020

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

AUTORIZACIÓN PARA LA PRESENTACIÓN DE TESIS DOCTORAL

Año Académico: 2019/2020

1.- Datos personales del autor de la Tesis		
Apellidos: González Pérez-Silva	Nombre: José María	
DNI/Pasaporte/NIE: 71675501A	Teléfono: 630344262	Correo electrónico: ereboperezsilva@gmail.com

2.- Datos académicos	
Programa de Doctorado cursado: Biomedicina y oncología molecular	
Órgano responsable: Universidad de Oviedo	
Departamento/Instituto en el que presenta la Tesis Doctoral: Instituto Universitario de Oncología del Principado de Asturias (IUOPA)	
Título definitivo de la Tesis	
Español/Otro Idioma: Análisis comparativo del degradoma de metazoos longevos	Inglés: Comparative analysis of the Degradome of long lived metazoans
Rama de conocimiento: Ciencias de la salud	

3.- Autorización del Director/es y Tutor de la tesis	
D: Víctor Quesada Fernández	DNI/Pasaporte/NIE: 53507141X
Departamento/Instituto: Departamento de bioquímica y biología molecular / IUOPA	
D/Dª:	DNI/Pasaporte/NIE:
Departamento/Instituto/Institución:	
Autorización del Tutor de la tesis	
D: José María Pérez Freije	DNI/Pasaporte/NIE: 09371813A
Departamento/Instituto: Departamento de bioquímica y biología molecular / IUOPA	

Autoriza la presentación de la tesis doctoral en cumplimiento de lo establecido en el Art. 32 del Reglamento de los Estudios de Doctorado, aprobado por el Consejo de Gobierno, en su sesión del día 20 de julio de 2018 (BOPA del 9 de agosto de 2018)

En Oviedo, a 27 de abril de 2020

Director de la Tesis

QUESADA FERNANDEZ VICTOR - 53507141X

Firmado digitalmente por QUESADA FERNANDEZ VICTC 53507141X
Fecha: 2020.04.28 11:03:36 +02'00'

Fdo.: Víctor Quesada Fernández

Tutor de la Tesis

PEREZ FREIJE JOSE MARIA - 09371813A

Firmado digitalmente por PEREZ FREIJE JOSE MA 09371813A
Fecha: 2020.04.28 11:57:(+02'00'

Fdo.: José María Pérez Freije

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN BIOMEDICINA Y ONCOLOGÍA MOLECULAR

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RESOLUCIÓN DE PRESENTACIÓN DE TESIS DOCTORAL

Año Académico: 2019/2020

1.- Datos personales del autor de la Tesis		
Apellidos: González Pérez-Silva	Nombre: José María	
DNI/Pasaporte/NIE: 71675501A	Teléfono: 630344262	Correo electrónico: ereboperezsilva@gmail.com

2.- Datos académicos	
Programa de Doctorado cursado: Biomedicina y oncología molecular	
Órgano responsable: Universidad de Oviedo	
Departamento/Instituto en el que presenta la Tesis Doctoral: Instituto Universitario de Oncología del Principado de Asturias (IUOPA)	
Título definitivo de la Tesis	
Español/Otro Idioma: Análisis comparativo del degradoma de metazoos longevos	Inglés: Comparative analysis of the Degradome of long lived metazoans
Rama de conocimiento: Ciencias de la salud	
Señale si procede: ☑ Mención Internacional ☑ Idioma de presentación de la Tesis distinto al español ☐ Presentación como compendio de publicaciones	

3.- Autorización del Presidente de la Comisión Académica	
D/Dª: Luis Menéndez Antolín	DNI/Pasaporte/NIE: 11410713E
Departamento/Instituto: Instituto Universitario de Oncología del Principado de Asturias (IUOPA)	

Resolución: La Comisión Académica del Programa de Doctorado en Biomedicina y Oncología Molecular del Instituto Universitario de Oncología del Principado de Asturias (IUOPA), en su reunión telemática de fecha 27/04/2020 , acordó la presentación de la tesis doctoral a la Comisión de Doctorado, previa comprobación de que la tesis presentada y la documentación que la acompaña cumplen con la normativa vigente, según lo establecido en el Art.32.8 del Reglamento de los Estudios de Doctorado, aprobado por el Consejo de Gobierno, en su sesión del día 20 de julio de 2018 (BOPA del 9 de agosto de 2018)
Además, informa:

	Favorable	Desfavorable
• Mención Internacional	■	☐
• Idioma	■	☐
• Presentación como compendio de publicaciones	☐	☐

FOR-MAT-VOA-012 (Reg. 2018)

Justificación
Se autoriza la presentación de la tesis doctoral del doctorando D. José María González Pérez-Silva para su depósito y posterior defensa.

Oviedo, 28 de Abril de 2020

Presidente de la Comisión Académica del Programa de Doctorado

Luis Menéndez Antolín
Firmado digitalmente por Luis Menéndez Antolín
Fecha: 2020-04-29 12:18:53

Fdo.: Luis Menéndez Antolín

Contra la presente Resolución, podrá interponer recurso de alzada ante el Rectorado, en el plazo de un mes, a partir del día siguiente al de la presente notificación, de conformidad con el art. 122 de la Ley 39/2015, de 1 de octubre, de Procedimiento Administrativo Común de las Administraciones Públicas

SR. DIRECTOR DEL CENTRO INTERNACIONAL DE POSTGRADO

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Análisis comparativo del degradoma de metazoos longevos	Inglés: Comparative analysis of the Degradome of long lived metazoans

2.- Autor	
Nombre: **José María González Pérez-Silva**	DNI/Pasaporte/NIE: 71675501A
Programa de Doctorado: Oncología y medicina molecular	
Órgano responsable: Universidad de Oviedo	

RESUMEN (en español)

El envejecimiento se puede definir como el declive de la eficacia biológica de un organismo en el tiempo. Este proceso afecta a casi todos los organismos multicelulares, aunque de diversas formas. Las consecuencias del envejecimiento son pleiotrópicas, e incluyen un elevado riesgo de sufrir enfermedades. Se han identificado 9 claves generales (hallmarks) del envejecimiento en animales. La intervención en estos procesos puede llevar a una mejora de la duración y calidad de vida. De entre estos, 4 destacan por ser primarios (inestabilidad genómica, reducción de telómeros, alteraciones epigenéticas y pérdida de proteostasis). Dado que el resto derivan de estos, son el mejor punto para comenzar a comprender el envejecimiento, al mismo tiempo que aportan dianas terapéuticas muy convenientes, ya que su atenuación supone una atenuación de las hallmarks secundarias.

Paralelamente, definimos el degradoma como el conjunto de proteasas de un organismo dado. Este sistema cumple diversas funciones basadas en la rotura regulada de proteínas, siendo de gran relevancia en procesos tan importantes como el envejecimiento y el cáncer. Dada la complejidad bioquímica del degradoma, nuestra primera aproximación para comprender su papel se basa en el estudio de su evolución. Así, la caracterización del degradoma en especies longevas debería mostrar evidencias de selección en aquellos componentes del sistema que estén involucrados en el envejecimiento.

En este contexto, la presente Tesis se propone analizar y comparar los genomas de dos especies longevas, el cachalote y la extinta Tortuga gigante de Pinta (Islas Galápagos), con el fin de extrapolar adaptaciones que puedan contribuir de alguna forma a entender mejor el proceso del envejecimiento en la especie humana. De esta forma se encontraron diversos tipos de alteraciones presumiblemente relevantes, como expansiones génicas, silenciamiento de genes, o alteraciones en sus motivos funcionales. Comparando estas alteraciones entre distintas especies más o menos cercanas a la especie analizada, se puede determinar su historia evolutiva, y gracias a los efectos conocidos de deficiencias en determinadas proteasas (degradomopatías), podemos extrapolar un efecto en determinadas rutas o sistemas, traduciendo así una alteración genética en un posible efecto biológico.

En el genoma del cachalote (*Physeter macrocephalus)* se anotaron un total de 546 proteasas implicadas en sistemas tan distintos como el inmune, la regulación de la presión sanguínea, el mantenimiento de la piel, o el sistema digestivo. En general, una parte significativa de estas mutaciones se corresponden con proteasas con conocidos efectos en el control de la inflamación, sugiriendo un mayor control sobre esta, y por ende un efecto ligeramente mitigado de sus efectos adversos (*inflammaging*). Por otra parte, se anotaron en torno a 600 proteasas en el genoma de la tortuga gigante de las Galápagos (*Chelonoidis abingdonii*), cuyo ensamblaje también se discute en la presente tesis. Entre las diversas adaptaciones identificadas en estas, destacan nuevamente aquellas con una clara vinculación al sistema inmune, sugiriendo en este caso un favorecimiento del sistema innato frente al adquirido. Además, fueron también significativas diversas adaptaciones relacionadas con la sensibilidad a nutrientes (especialmente a la glucosa), o con el desarrollo neurológico. Finalmente, el análisis automático del ensamblaje del genoma, aportó información sobre expansión de familias génicas y sobre presión selectiva de genes. En ambos casos, un número significativos de los genes destacados, guardaba alguna relación con el sistema ERAD, que a su vez guarda estrecha relación con la comunicación intercelular, otro de los *hallmarks* del envejecimiento.

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RESUMEN (en Inglés)

Ageing can be defined as the decay of biological fitness of an organism in time. This process affects to almost every multicellular organism, but in different ways. Similarly, the results of ageing are pleiotropic, including a higher sensibility to diseases. In this sense, 9 hallmarks of this process were identified in animals. Its intervention leads to an enhanced vitality and quality of life. Among these, 4 are highlighted as primary hallmarks (genome instability, telomere attrition, epigenetics alterations, and loss of proteostasis). Since the rest derive from these, they are the best point to start understanding ageing, while they also provide convenient therapeutic targets, since their attenuation mean san attenuation of secondary hallmarks.

Parallelly, we define the degradome as the set of proteases in a given organism. This system has several functions based on the controlled break of proteins, playing an important part of many biological processes such as ageing or cancer. Given the enormous biochemical complexity of the degradome, our first approach to unveil its role is based in the study of its evolution. Hence, degradome characterization in long lived species should yield evidence of selection in those components involved in ageing.

In this context, the present Thesis pretends to analyse and compare the genomes of two long lived species, the sperm whale and the extinct giant tortoise of Pinta (Galápagos Islands), aiming to deduce the adaptation that may contribute in any way to better understand the ageing process in humans. Thus, many different types of presumably relevant adaptations were found, such as gene families expansion, gene silencing, or alteration of their functional motives. By comparing these alterations among more or less closely related species, one can determine the evolutive the evolutive history of said adaptions, and thanks to the known pathological effects of these (degradomepathies), we can extrapolate an effect in certain routes or systems, translating, this way, a genetical alteration into a possible biological effect.

In the genome of sperm (*Physeter macrocephalus)* we annotated a total of 546 proteases, implicated in different systems, such as the immune, blood pressure regulation, skin maintenance, or the digestive system. In general, a significant part of these relate to proteases with known effects in the regulation of inflammation, suggesting a greater control over it, which in turn would mitigate some of its antagonistic effects (*inflammaging*). On the other hand, around 600 proteases were annotated in the genome of the Galápagos giant tortoise (*Chelonoidis abingdonii*), whose assembly is also discussed in the present thesis. Among the several identified adaptations, those with a clear bond with the immune system stand out again, suggesting, in this case, a favouring of the innate immune system versus the adaptative one. Additionally, adaptations linked to nutrient sensing (specially those related to glucose) were also significant, similarly with alterations related to neurological development. Finally, the automatic analysis of the genome assembly, yielded information regarding gene family expansions and footprints of selective pressure in a significant number of genes. In both cases a considerable number of results had some relation with the ERAD system, which, in turn, is tightly linked with another hallmark of ageing, the alteration of intercellular communications.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN BIOMEDICINA Y ONCOLOGÍA MOLECULAR

Esta tesis es la culminación de varios años de trabajo, pero, más aún, representa la síntesis de una etapa de mi vida, y como tal, además de recoger en ella los resultados y conclusiones de los proyectos en que he participado, es necesario dedicar también unas frases a todas las personas que, de una u otra forma, han contribuido a que este barco llegue a buen puerto.

Por ello, y en primer lugar, quiero agradecer a Carlos la oportunidad que me dio de empezar todo esto una tarde extremadamente cálida, hace ya muchos años. Gracias, Carlos, por apostar por mí en una situación en que muchos otros no lo habrían hecho, por abrirme las puertas al que, desde siempre, había sido mi sueño.

Gracias a Víctor, por introducirme en el mundo de la bioinformática, por su dirección y tutela cuando necesité guía, y por dejarme espacio para crecer independientemente cuando la situación lo permitía. Gracias por tu paciencia, Víctor, sé que no siempre fue fácil.

Gracias a Gloria, a Josemari, y a Xose, porque sin ellos el laboratorio no sería lo que es, por cuidar de todos nosotros en las buenas y en las malas, por su preocupación y dedicación constante. Gracias por vuestra implicación y dedicación. Gracias también a Yaiza, que con su trabajo hace el de los demás más fácil. Gracias por intentar animarnos y ayudarnos siempre.

Gracias a todos los compañeros de laboratorio, con los que tantas horas he compartido. Gracias por toda la ayuda prestada y por tantos buenos ratos, por los cafés, las conversaciones, y por tantas tardes de sesiones músicales.

Thanks to Dr. Kevin Howe for letting me visit his lab for three wonderful months. Thank you for your help and willingness. Thanks also to Fergal and his team (especially to Leanne), for your patience, your advice and guidance, and for making me feel welcomed and cozy*. You helped me to learn a different way of doing science and for that I'm (again) grateful.*

Y por supuesto, gracias a la Universidad de Oviedo, al Instituto Universitario Oncológico del Principado de Asturias, y al ministerio de Salud por su continuado apoyo financiero y sus ayudas durate la realización de esta tesis.

Pero no todo iba a ser trabajar, así que tengo que dedicar unas palabras de agradecimiento especial a un grupo que pese a haber nacido entre las paredes del labo, ha trascendido con creces a lo largo de los años. Gracias a las personas que pasaron de no atreverse a compartir coche en un viaje de 200 kilómetros a venirse conmigo al otro lado del globo. No son palabras vacías si os digo que sin vosotros esto no habría sido posible. No solo por el constante apoyo y los buenos momentos, sino por compartir los malos (y los muy malos). Por todas las fiestas, las "autoinvitaciones", los viajes y las escapadas, mil gracias.

Gracias también a mi "cohorte de Sandramandra", gracias por haber estado ahí desde tiempos inmemoriales. Por las innumerables horas de aventuras imaginarias en tierras desconocidas, por todas las visitas (casi improvisadas) a mi país adoptivo, las reuniones (algo más planificadas) en el país bávaro, y los (de momento inexistentes) viajes a las tierras de la lengua incomprensible. Y, por supuesto, por las "tertulias científicas" durante las noches de vermú. Gracias por todo.

A mi "equipo de IT", gracias. Gracias por que con vuestros chistes y provocaciones me disteis la determinación que necesitaba para avanzar en un campo que me era desconocido, y con vuestros consejos y eterna disposición, me ayudasteis a progresar en el mismo. Gracias por estar "aquí", por las noches de peli, por las sesiones de turismo de "aperitivo", por los campings y por todos los buenos momentos.

No puedo terminar sin darle las gracias a mi familia, por su constante e incondicional apoyo. A mis padres, sin quienes no habría llegado tan lejos. Gracias por llevarme a la universidad, por ayudarme a avanzar curso a curso, y por seguir a pie del cañón día a día. También a mi hermana, compañera en muchos viajes (no solo de carretera), a mis tíos y tías, por vuestro interés en mi trabajo y vuestros ánimos, a mis primos por todos los buenos ratos de esparcimiento, y, de una forma muy especial, gracias a mis abuelas, por todo vuestro apoyo y amor.

Finalmente, me gustaría dedicarle la presente tesis a mis abuelos Paco y Pepe, que de una forma u otra me han inspirado a tomar este camino y han hecho que me sienta orgulloso de haberlo seguido. Gracias por ayudarme a estar aquí hoy.

Nothing in biology makes sense
except in the light of evolution.

THEODOSIUS DOBZHANSKY

Contents

List of Figures

List of Tables

List of Abbreviations

AJAX	Asynchronous JavaScript And XML
ATP	Adenosine TriPhosphate
BATI	Blast Annotate Tune and Iterate
BLAST	Basic Local Alignment Search Tool
CGI	Common Gateway Interface
CNV	Copy number variations
DNA	Deoxyribonucleic Acid
ERAD	Endolasmatic-Reticulum-Associated protein Degradation
HTML	HyperText Markup Language
JSON	JavaScrip Object Notation
KKS	Kinnin-Kallikreyn System
LSF	Load Sharing Facility
NK	Natural Killers
ROS	Reactive Oxygen Species

Introduction

In 1958, Theodosius Dobzhansky published an essay discussing the relation between homoeostasis and senility, and how natural selection affects both [Dobzhansky, 1958]. On it, Dobzhansky reflected on the concept of homoeostasis, defined by Walter Cannon some years before as the "wisdom of the body" [Cannon, 1934], or the ability to, through controlled changes in the organism, adapt to changes in the eviroment, provided that they are within the "normal" changes the organism had evolved to withstand. He also linked this concept with ageing, by defining the later as the reduction of this adaptability or plasticity agains "normal" enviromental changes. Specifically, he proposed that "*the homeostatic buffering against enviromental shocks is weakened during the postreproductive phase*".

In his work "The Causes of Evolution", Haldane considered this topic without any clear conclusion. To him, natural selection "may either favor or hinder the prolongation of life during the postreproductive phase" [Haldane, 1933]. To Dobzhansky, the fact that homoeostatic mechanisms "tend to deteriorate during the autumm of life", while the same function most efficiently during youth and maturity, is indication of how these are fashioned by natural selection. Of course, we now understand in much more detail how far the link between homoeostasis and ageing goes, knowing specific cell paths, cellular systems, and other homoeostasis mechanisms for which *heritable* malfuction (or even slightly under-optimal function) is a cause for a hastened ageing process.

After decades of tentative work on these determinants of ageing, in 2013, a seminal work finally established a precise framework to study how genetic determinants and their interaction with the environment regulates this process [López-Otín et al., 2013]. The framework is based on nine *hallmarks of ageing*.

1.1 Hallmarks of ageing

Ageing is quasi-universal among multicellular organisms, yet, it is probably one of the least-understood of the natural processes of our biology [Kirkwood, 2005]. It can be described as a progressive loss of biological fitness, or the progressive decline in functional integrity and homoeostasis, culminating in death [Singh et al., 2019]. As such, it is expected to be caused by a continuous accumulation of damage with points of inflexion, at which the attempts of the body of fighting this creeping decay adds to the deleterious nature of ageing.

The characterization of the hallmarks of ageing, much like what happened with the hallmarks of cancer previously published [Hanahan and Weinberg, 2011], has helped conceptualize the underlying nature of ageing whilst setting a frame in its study.

For its categorization, each hallmark must satisfy 3 key principles:

1. It should manifest under normal ageing.
2. Its experimental aggravation should accelerate the normal ageing process.
3. Its experimental betterment should delay the normal ageing process, increasing healthspan and lifespan.

Following this canon, nine hallmarks were defined. These were genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication. These nine hallmarks were grouped in 3 categories, depending on their impact on the process of ageing. The first category, *primary hallmarks*, is composed of those considered to be the causes of cellular damage, including genomic instability, telomere attrition, epigenetic alterations, and loss of proteostasis. The second category, (*antagonistic hallmarks*), was defined as the group of compensatory or antagonistic responses to the damage produced by the primary hallmarks. While these responses are able to initially mitigate the damaging effects produced by the primary hallmarks, after becoming chronic they will turn deleterious as well. This group includes deregulated nutrient sensing, mitochondrial dysfunction, and cellular senescence. Lastly, *integrative hallmarks*, which arise from the consequences of the previous groups, are directly responsible for the functional decline associated with ageing. This category consists of two hallmarks, stem cell exhaustion, and altered intercellular communications [López-Otín et al., 2013].

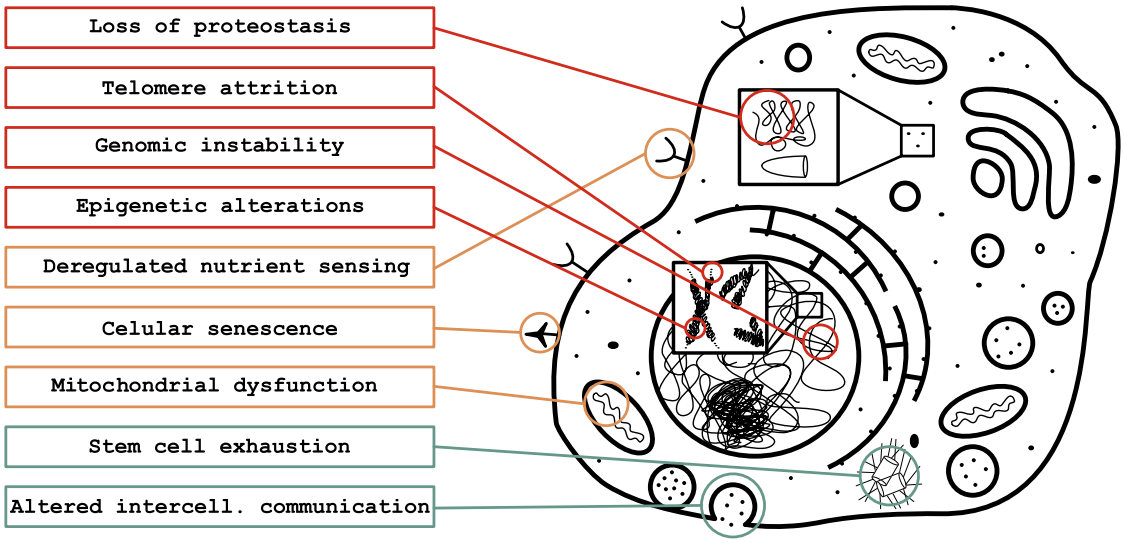

Fig. 1.1: The nine hallmarks of ageing, grouped by category using the different colours. In **red** are displayed the primary hallmarks, such as genomic instability, telomere attrition, epigenetic alterations, and loss of proteostasis. In **orange** we can see the secondary or antagonistic hallmarks, *i.e.* deregulated nutrient sensing, mitochondrial dysfunction, and cellular senescence. Finally, those in **green** are the integrative hallmarks, namely, stem exhaustion, and altered intercellular communication.

1.1.1 Primary hallmarks

Genomic instability refers to the gradual accumulation of genetic damage throughout life [Moskalev et al., 2013]. As we age, the integrity and stability of our DNA is being continuously challenged by both exogenous and endogenous threats. Physical, chemical and biological agents such as ultraviolet light, mutagenic substances, or even viral replication are some examples of exogenous damage that DNA must withstand. In addition, DNA faces internal sources of instability, such as replication errors, or the damage produce by some by-products of our own metabolism, like ROS [Hoeijmakers, 2009]. Overall, the genome is constantly suffering highly diverse lesions, including point mutations, translocations, chromosomal gains and losses, telomere shortening, and gene disruption caused by the integration of viruses and transposons. For that reason, organisms have evolved a complex network of DNA repair mechanisms capable of tackling most of this problems [Lord and Ashworth, 2012]. Excessive DNA damage, or insufficient DNA repair favours the ageing process. On the other hand, incorrect handling of this repair process is in itself one of the sources of further damage to the DNA. Thus, many lesions are originated after mismatch repair, non-homologous

end-joining, translesion synthesis, or base excision repair [Agathangelou et al., 2018]. Hence, these aquired complex systems and characteristics are subject to heredity and hence to the forces of natural selection, adding weight to the hypothesis of heritable longevity [Lord and Ashworth, 2012].

The aforementioned damage to the DNA is seemingly random in location, but there are some specific chromosome regions that are specially susceptible to age-related damage, such as the telomeres. Telomeres, a genomic region made up of repeated sequences at the end of each chromosomal arm, are among such regions, making *telomere attrition* another primary hallmark. This reduction in length that accompanies ageing originates from the natural incapability of the DNA polymerase to completely replicate the terminal ends of linear DNA molecules [Turner et al., 2019]. There is a specific DNA polymerase capable of replicating this "caps", known as telomerase, but it is not expressed by most of mammalian cell types, leading to the progressive reduction of this DNA segment over time [Shay, 2016]. Researchers have observed not only a natural reduction of telomere size during normal ageing, but also that the pathological dysfunction of telomerase in experimental models greatly accelerates the process of ageing.

As we advance in the field of epigenetics, we increasingly recognise the importance of its physiological and pathological role in a lot of different biological processes, including ageing. For this reason, we include *epigenomic alterations* as one of the primary hallmarks of ageing. These modifications affect all cells and tissues throughout life, via DNA methylation, post-translational modification of histones, and chromatin remodelling, and can be linked to accelerated ageing processes [Pal and Tyler, 2016]. For instance, it has been experimentally observed that specific epigenetic alterations can mimic progeroid syndromes in model organisms [Osorio et al., 2010]. Because of this, understanding and manipulating the epigenome holds promise for improving age-related pathologies, hence extending lifespan and, more importantly, healthspan.

The last primary hallmark of ageing, *loss of proteostasis*, relates to the maintenance of a correct protein homoeostasis in the cell, meaning, getting rid of defective proteins, or correcting them. The function of most proteins depends on their tridimensional structure, hence a misfolded protein could either not have function at all or have a different, unregulated one. Specifically, we use the term "proteostasis" to reffer collectively to all the paths that deal with unfolded or misfolded proteins, namely the autophagy path, proteasome and lysosome degradation paths, and refolding via chaperones [Klaips et al., 2018]. These paths can be regulated based on the action of the heat-shock protein family [Hartl et al., 2011, Koga et al., 2011, Mizushima et al., 2008]. In a normal situation, the different pathways works coordinatedly towards restoring the structure of

misfolded peptides or to remove and degrade them completely, avoiding the accumulation of damaged (and damaging) proteins [Powers et al., 2009]. Numerous studies have shown that ageing tends to alter proteostasis, decreasing its effectiveness [Hipp et al., 2019]. When this happens, it has been observed that chronic accumulation of misfolded proteins may lead to age-related pathologies, such as Alzheimer and Parkinson's disease, or cataracts [Powers et al., 2009]. Finally, when inducing perturbations in this system, age-associated pathologies as well as hastened ageing has been observed. In contrast, by improving proteostasis we can achieve the opposite effect and delay ageing [Zhang and Cuervo, 2008].

1.1.2 Antagonistic hallmarks

The *deregulation of nutrient sensing* hallmark relates to the system that detects and corrects the levels of different nutrients, energy, and other elements that alter body homoeostasis. This complex hallmark stems from the relationship between anabolic signalling and accelerated ageing [Fontana et al., 2010]. As an example of this, a pharmacological manipulation that mimics low availability of nutrients has been observed to extend longevity in mice [Harrison et al., 2009]. Also, consistent with this, dietary restriction has shown to increase healthy lifespan in several species, including some primates [Mattison et al., 2017].

Ageing also affects the respiratory chain, whose efficacy diminishes with age, increasing electron leakage and reducing ATP generation [Green et al., 2011]. We know this hallmark as *mitochondrial dysfunction*. In turn, these problems tend to increase the concentration of ROS, which, as mentioned before, also plays an important role in accelerating ageing [Harman, 1965]. While this association between defects in the respiratory chain and hastened ageing has being known for a long time, major details remain a research challenge. In fact, the fulfilment of the third principle of hallmarks (amelioration leading to higher lifespan) is still under discussion.

Cellular senescence is defined as a stable arrest of the cell cycle coupled with phenotypic changes [Campisi and D'Adda Di Fagagna, 2007, Kuilman et al., 2010]. Several ageing-related stimuli are believed to trigger this process, including telomere shortening, non-telomeric DNA damage, and depression of the INK4/ARF locus [Collado et al., 2007]. The accumulation of senescence cells in a tissue can be indirectly measured through surrogate markers, such as DNA damage or the presence of metabolites such as senescence-associated β-galactosidase [Dimri et al., 1995]. It has been observed that these senescent cells are not equally accumulated across tissues in old age. In

fact, some organs, like heart or kidney, do not show changes in senescence associated to age [Hoenicke and Zender, 2012, Kang et al., 2011, Xue et al., 2007]. It has been widely assumed that senescent cells contribute to ageing, however, cellular senescence should be considered a response from the body to ageing, by marking cells for their deletion. Sadly, as we age, due to all the other problems associated, the mechanisms that should get rid of these "marked" cells function suboptimally, and the deletion does not happen as fast as it should (or at all) [Calcinotto et al., 2019]. Therefore, it is the process of ageing which promotes the accumulation of senescent cells that otherwise should have been killed and removed. Thus, the malfunction of the turnover system that replaces cells in tissues lowers the regenerative capacity of the progenitor cells and leads to their exhaustion.

1.1.3 Integrative hallmarks

As we previously stated, these hallmarks are the direct result of the previous ones. Its chronification is the final cause for most of the phenotypic changes associated with ageing. And so, as a direct response to cellular senescence, the first of these hallamarks is *stem cell exhaustion*. It can be defined as the decline in the regenerative potential of tissues. For instance, it has been shown that haematopoietic stem cells of aged mice decrease in cell-cycle activity compared to those of young mice [De Haan and Lazare, 2018, Shaw et al., 2010]. Of course, this directly correlates with a higher accumulation of DNA damage and over-expression of cell-cycle arrest proteins such as $p16^{INK4a}$ [Janzen et al., 2006, Rossi et al., 2007, Stenvinkel et al., 2017]. While a deficit in proliferation is detrimental for the long-term maintenance of the organism, its excess is equally damaging, since it translates into an early exhaustion of the proliferating capacity of the stem cells of the tissue. This effect has been observed in experiments with *Drosophila melanogaster* intestinal stem cells, where excessive proliferation lead to exhaustion and premature ageing [Rera et al., 2011, Wang et al., 2014]. As an integrative hallmark, stem-cell exhaustion can be modulated by interventions on primary and antagonistic hallmarks. For instance, it has been shown that inducing *INK4a* (related to cellular senescence), or decreasing IGF-1 (related to deregulated nutrient sensing), we can help preserve the quiescence of stem cells, delaying stem cell exhaustion in the organism [Chakkalakal et al., 2012, Tümpel and Rudolph, 2019].

Besides cell-autonomous alterations, ageing also involves changes at the level of cell-cell integration, either by endocrine, neuronal or neuroendocrine paths [Laplante and Sabatini, 2012, Rando and Chang, 2012, Zhang et al., 2013]. Thus, *altered intercellular*

communication, the last hallmark of ageing, has ubiquitous effects in several different signalling paths, including a deregulation in neurohormonal signalling (such as renin-angiotensin, adrenergic or insulin-IGF1 signalling) [Bocheva et al., 2019]. This provokes an increase in inflammatory reactions, a decrease in immunosurveillance against both pathogens and premalignant cells, and also changes in the composition of the peri- and extracellular environment. One of the most interesting effects of an altered intercellular communication is a smouldering proinflammatory phenotype associated with ageing called *inflammageing* [Franceschi and Campisi, 2014]. Inflammaging has different causes, such as an accumulation of proinflammatory tissue damage, a failure of an immune system to effectively clear pathogens or dysfunctional cells, propensity to secrete proinflammatory cytokines by senescent cells, or an autophagic response [Franceschi et al., 2017, Salminen et al., 2012].

1.2 The Degradome

As clearly reflected by the hallmarks of ageing, ageing is an extremely complex trait, with multiple intertwined systems, and a network of causes and consequences that extend to almost any remarkable cellular pathway. A useful step in tackling these complexities is to separate a simpler system that nevertheless recapitulates some of the characteristics of the whole system. In this regard, we have extensively worked on *proteases*, *i. e.*, proteins capable of degrading other proteins by means of peptide bond hydrolysis in an essentially irreversible way [Pérez-Silva et al., 2016]. Proteases influence diverse biological features of the organisms, such as the immune system, digestion process, skin regeneration, cell cycle progression, tissue remodelling, neuronal outgrowth, haemostasis, wound healing, angiogenesis, apoptosis or metastasis, [López-Otín and Bond, 2008, Quirós et al., 2015, Reinhard et al., 2015, Voskoboinik et al., 2015].

Due to these ubiquitous functions, the proper function of the proteolytic system is key to maintain homoeostasis in the organism, so it must be tightly regulated in terms of both activation and specificity. Failings in their regulation underlie very diverse pathological conditions, such as progeria, cancer or even mental illness [Turk et al., 2012]. Thus, this set of genes experienced a remarkable evolutive expansion as an adaptation to regulate the large set of substrates depending on its correct functioning.

The large number of protease genes [Pérez-Silva et al., 2016] and their interdependence led to the definition of the *degradome* as the complete set of proteases in a given organism [López-Otín and Overall, 2002]. The set of techniques that study and characterise proteases in this context is known as *degradomics*. Given the large amount of

information we often work with when delving in degradomics, bioinformatic techniques are important tools, as they allow researchers to focus on the biological meaning of the data and not its processing. Our experience and the mining of literature are usually a vital complement to the manual curation of annotations after automatic predictions and other analysis. All together aimed towards finding links between protease alterations and pathological consequences, including hereditary diseases, sometimes referred as *degradomopathies* [Quesada et al., 2009]

1.3 Molecular strategies against ageing and model organisms

With information on hallmarks of ageing and the influence of proteases on them, we can look for clues on how longevity affects the genome of long lived animals. The abundant data on protease function and biochemistry can then be leveraged to discern some of the molecular mechanisms that underlie extended life-span. As controls, we can use evolutionary comparison with both close and distant relatives of diverse longevity.

1.3.1 *Physeter macrocephalus* and other Cetacea

The sperm whale is a marine mammal, part of the Cetacea infraorder. As such, they are part of the order Artiodactyla[1] (also known as even-toed ungulates), placing them as relatives to Camelidae (suborder Tylopoda), the pig and close family (suborder Suina), the ruminants (suborder Ruminantia), and hippopotamuses (with whom Cetacea share the Whippomorpha suborder; figure 1.4) [Agnarsson and May-Collado, 2008]. Cetacea and Sirenia (the order of manatees or sea cows) are the only mammals that live their entire lives inside the water. Interestingly, they are not closely related to those, nor to Superfamily Pinnipedia (seals and the like), the only other mammal adapted to semiaquatic life [Arnason et al., 2007, Tabuce et al., 2008]. This suggests that aquatic adaptation developed independently several times in Mammalia, an example of convergent evolution. On its own, Cetacea is further divided into two parvorders, Odontoceti, to which the sperm whale belongs, and Mysticeti [Mancia, 2018].

Mysticeti or "Baleen Whales" feed on plankton, by using a filter-like system based on extremely thin keratin bristle-like structures called baleen (as in its "common" name

[1]Sometimes referred as Cetartiodactyla, a combination of "Cetacean" and "Artiodactylia", arguing that despite the clear evidence supporting this clade, the enormous morphological differences between cetaceans and the rest justify this distinction.

[Fordyce and Marx, 2018]. The fifteen members composing this parvorder are quite diverse, ranging in size from the 6 m and 3,000 kg of the pygmy right whale to the 31 m and 190,000 kg of the blue whale [Agnarsson and May-Collado, 2008, Wada et al., 2003]. They also diverge in other features, such as life expectancy and diving capacity. Of special interest in the present work is the Bowhead whale (*B. mysticetus*), some of whose specimens have been shown to live over 400 years, the largest life span recorded among Mammalia [Keane et al., 2015]. In addition, the availability of genomic data on Minke whale (*B. acutorostrata*) allows genetic comparisons, as both whales are representatives of the Mysticeti parvorder [Yim et al., 2014].

Odontoceti or toothed whales, are those presenting teeth. These cetaceans feed on meat from various species, from aquatic mammals to fishes, including, in the case of the sperm whale, giant squids [Best, 1979]. As found in the parvorder of the Mysticeti, there is great divergence between members of this group as well, with species ranging from 1.5 m and 50 kg to more than 20 m and 50,000 kg [Warren et al., 2017]. Odontoceti is a much larger parvorder than Mysticeti, with more than 70 different species [Agnarsson and May-Collado, 2008]. As a reflection of this diversity, Odontoceti are further divided in *Physeteroidea*, the superfamily comprising the sperm whales and relatives; *Delphinoidea*, including all salt-water dolphins and relatives (*e.g.*, the killer whale, or the *Monodontidae* family, composed of narwhals and beluga whales); *Ziphioidea*, the superfamily of the beaked whales (*e.g.*, Cuvier's whale); and 3 other superfamilies that include the fresh-water dolphins, *Inioidea*, *Platanistoidea*, and *Lipotoidea* (figure 1.5) [Agnarsson and May-Collado, 2008]. In terms of longevity this parvorder is also diverse, featuring species with life-spans that range from 20 to 100 years, the maximum life expectancy for a sperm whale [Whitehead, 2003].

Sperm whales are believed to posses great cognitive capacities. Not only do they have the largest brain on earth, but also present complex social behaviours. They take care as a group of calves and elderly, protecting wounded or weaker individuals from predators, and they migrate as organised groups [Best, 1979]. One of their most salient features is their echolocation apparatus, which in fact originated the name *sperm whale*. The peculiar shape of the cranium of these animals allows the allocation of a waxy substance inside it, the spermaceti[2] [Alam et al., 2016,R. Clarke, 1970]. The spermaceti, whose function is still partly unclear, is though to provide the internal needed resonation that allows these whales to use a sophisticated system of communication based on ultrasounds (*clicks*) [Møhl, 2001]. This is key to their cognitive development and their

[2]Called 'spermaceti' after wrongly assuming the function that it played in sperm whale biology.

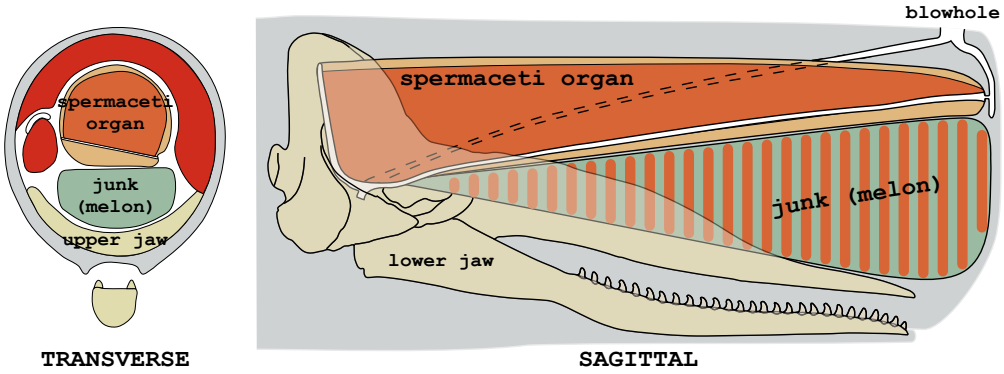

Fig. 1.2: Outline of the spermaceti organ and the craneum architecture of the sperm whale. Both the spermaceti and the sponge-like melon are thought to participe in the complex resonance system described. Adapted from *WikiCommons*, Creative Commons license.

social behaviour, since these clicks are supposed to be used for a variety of reasons. Thus, not only do sperm whales use echo location in the dark depths, but also, according to some researchers, they are able to perform some 'low-level' communication between them using this system, to the point of assigning specific names to members of the family [Schulz et al., 2011]. It was believed that another use of these clicks was to "stun" prey during hunting, but it has been proven that while useful in the chase and to buzz them, there is no stunning involved [Fais et al., 2016].

As a member of one of the few groups of mammals solely living in aqueous environments (Cetacea), sperm whales have undergone a process of adaptation to underwater distinctive features. First and foremost, given the nature of water, gravity is much more forgiving with organisms, allowing them to reach larger sizes, which grant these creatures a series of attributes. Thus, a link between longevity and size was already hinted by Aristotle in the *IV* century BC [Speakman, 2005]. This link is particularly apparent in the case of Bowhead whales who, as we mentioned before, can live to up to 4 centuries. In addition to that, as dictated by Peto's paradox, the observation of a moderate frequency of cancer occurrences in this gargantuan animals suggests extremely tight cancer protection mechanisms. Namely, by virtue of having a very large body, cetaceans are made of a larger number of cells than other mammals, which means more cells potentially able to develop cancer-like mutations. Therefore, this increase in cell number must be compensated by a much lower probability for tumorigenesis in each cell, accomplished via different systems [Tollis et al., 2017a]. Secondly, changes in pressure from shallow to deep waters demand different ways of regulation of blood homoeostasis (also known as *haemostasis*) [Stewart, 2009]. This is especially important

in the case of sperm whales, which, along with the Cuvier's whale, is one of the deepest divers known among mammals [Schorr et al., 2014].

On top of all of these features, sperm whales develop a distinctive skin structure, presenting a significant thicker *stratum corneum* than that of other Cetacea. In addition to the tissue structure of the skin, it differs macroscopically from other cetaceans in which it is more wrinkled (sometimes it is compared to a raisin). Also, sperm whales shed skin more regularly than other species [Sokolov, 1982]. Given the fact that skin is the first line of defence against the environment, these special traits may provide some advantage to the sperm whale in terms of physical defence, both against the attacks of their prey and the medium itself. Thus, it has been proposed a relationship between the diving peculiarities of the sperm whale and those of its skin, maybe also weighting in its hunting habits [Strauss, 1969].

1.3.2 Lonesome George and other giant tortoises

On September of 1835, *HMS Beagle* arrived to the shores of the Galápagos Archipelago to perform cartography mappings. For some weeks, on-board naturalist and geologist Charles Darwin also observed and recollected data on the island biodiversity, which, eventually, along with other findings, helped him enact his theory of evolution by means of natural selection. Among these observations, he noticed that the tortoises from different islands displayed conspicuous differences. Even so, tortoises from the same island presented very diverse shells depending on the altitude of their habitat [Darwin, 1845]. Ever since, partly due its implication in the development of this theory, the Galápagos giant tortoises have become well known everywhere, and have greatly contributed to the well-deserved notoriety of the archipelago. The most famous Galápagos tortoise was Lonesome George, the last member of his species, *Chelonoidis abingdonii*, who tragically died in 2012, when believed to be 101-102 years old, becoming an important symbol for conservation efforts, both in the Galápagos Islands and around the world.

At some point it was considered that all Galápagos tortoises were, in fact, one single species, *Chelonoidis nigra*, composed of several subspecies (*e.g.*, Lonesome George would have belonged to the subspecies *C. nigra abingdonii*) [Caccone et al., 1999]. Nowadays, studies support the idea of several differentiated species, sharing the same genus (*Chelonoidis*), but each independent [Le et al., 2006].

As a group, these species are all part of the class Reptilia (also referred to as Sauropsida). Reptilia is a complex clade, since we must contemplate that birds are part of it in order to considerate it a monophyletic group, or to assume that it's paraphyletic (leav-

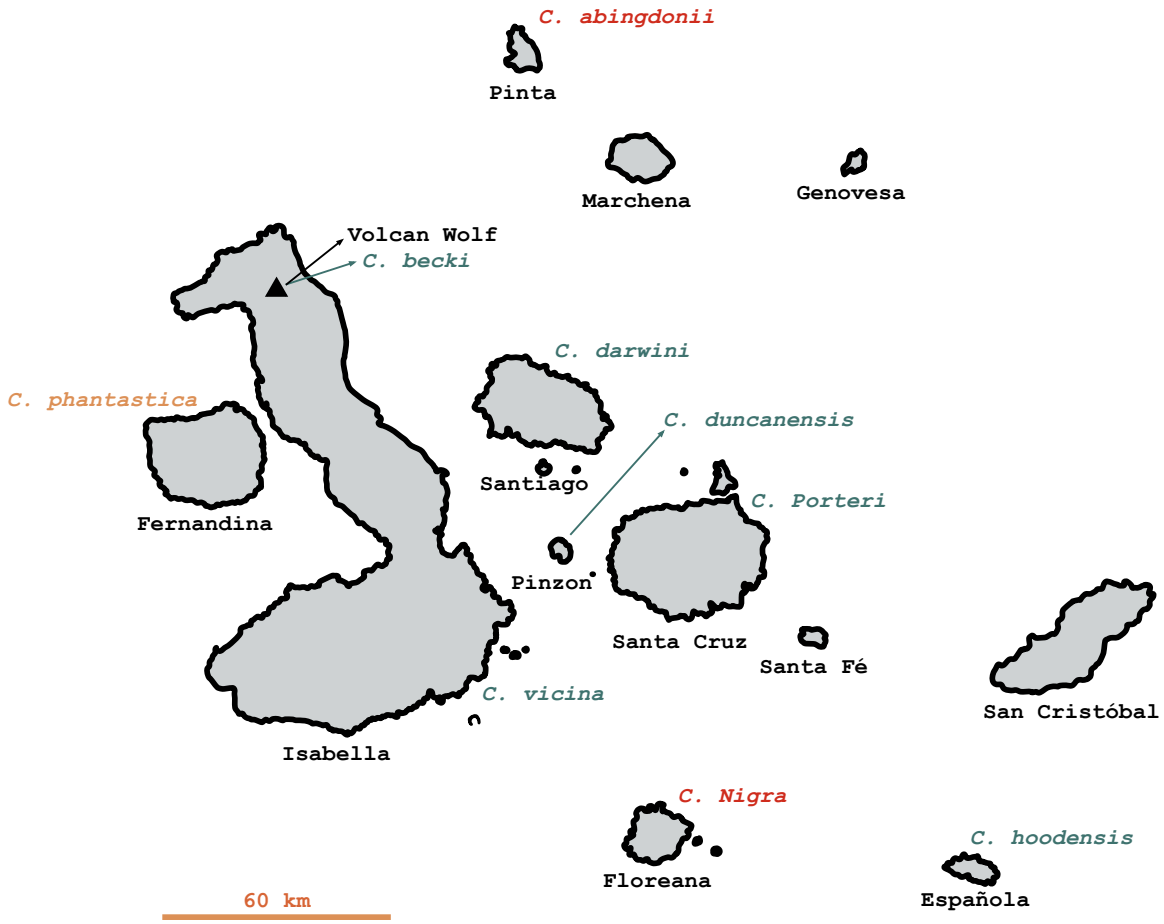

Fig. 1.3: Representation of the Galápagos Island (names in **black**) and the Galápagos giant tortoises that inhabits them, indicating in **red** those that are currently extinct, in **orange** those in doubt, and in **green** the rest. Adapted from *WikiCommons*, Creative Commons license.

ing crocodiles on a side). The traditional approach went by considering 2 subdivision of this class, Diapsida, and Anapsida, which refers to the number of openings in their skulls (Diapsida having two of them, and Anapsida having none). This classification,based on directly observable morphological features, has been revised and the current approach consist on 2 subdivisions as well, this time being Eureptilia, and Parareptilia [Benton, 2014]. While Parareptilia is extinct, Eureptilia contains itself another subdivision, called Diapsida, in which diapsides have being merged with anapsides, that has being rename as infraclass Neodiapsida. Inside this infraclass, Testudinata constitutes an order on its own, while Lepidosauromorpha (Tuataras, lizards, and snakes) and Archosauromorpha (crocodiles and birds, along with dinosaurs), maintain their status of

infraclasses (figure 1.5) [Mannen and Li, 1999, Modesto and Anderson, 2004].

Turtles[3] comprises more than 300 species [Rhodin et al., 2017]. This group is subdivided into two suborders, Pleurodira and Cryptodira, depending on whether their neck retracts sidewards or backwards, respectively. The extant members of Pleurodira are largely restricted to fresh-water ecosystems [Ferreira et al., 2018]. On the other hand, Cryptodira thrive in land, aquatic and mixed habitats. They present different adaptations according to feeding, both herbivorous and carnivorous, not to mention sizes, ranging from some centimetres to more than 2 metres.

Another remarkable feature of turtles is their longevity. Allegedly, one of the oldest turtles to ever live, *Adwaita* an Aldabra giant tortoise (*A. gigantea*), reached more than 250 years. Sadly no records are kept to support this claim. Among those with recorded or tested age, *Tu'i Malila* (deceased in 1956) and *Jonathan* (still alive) are both 188 years old and are thus considered the longest lived terrestrial animals. Notably, while *Jonathan* is also an Aldabra tortoise, *Tu'i Malila* was a member of *Geochelone radiata*, one of the land relatives of Galápagos Giant tortoises (of a much smaller size). Next in the list would be *Harriet*, a Galápagos giant tortoise (*C. porteri*), who died in 2006 at the age of 176 years. Another small tortoise, *Timothy*, lived 160 years, from around 1844 to 2004. Timothy belonged to *Testudo gracea*, also related to the land family of Galápagos giant tortoises. In this sense, it has been discussed that the "early" dead of Lonesome George had some pathological component to it, but this doesn't make his age negligible.

All these long-lived metazoans must show genomic footprints of the evolutionary processes that lent them a combination of molecular characteristics that solve the problems associated with longevity. With the intention of shedding some light on the subject of ageing, and to contribute to the collective efforts in solving this impending problem, the present Doctoral Thesis aims to study some of nature's solutions from an evolutionary point of view, paying special attention to the study of the degradome and its impact on longevity.

[3]Disclaimer: "Turtle" can refer to the order as a whole or, when differentiating between them, to the aquatic Testudines. In this sense, we use "tortoises" to refer to the land members of the order, while those able of both walking and swimming will be referred at as "terrapins". Whenever talking about the specific varieties we will try to stick to the more specific names possible.

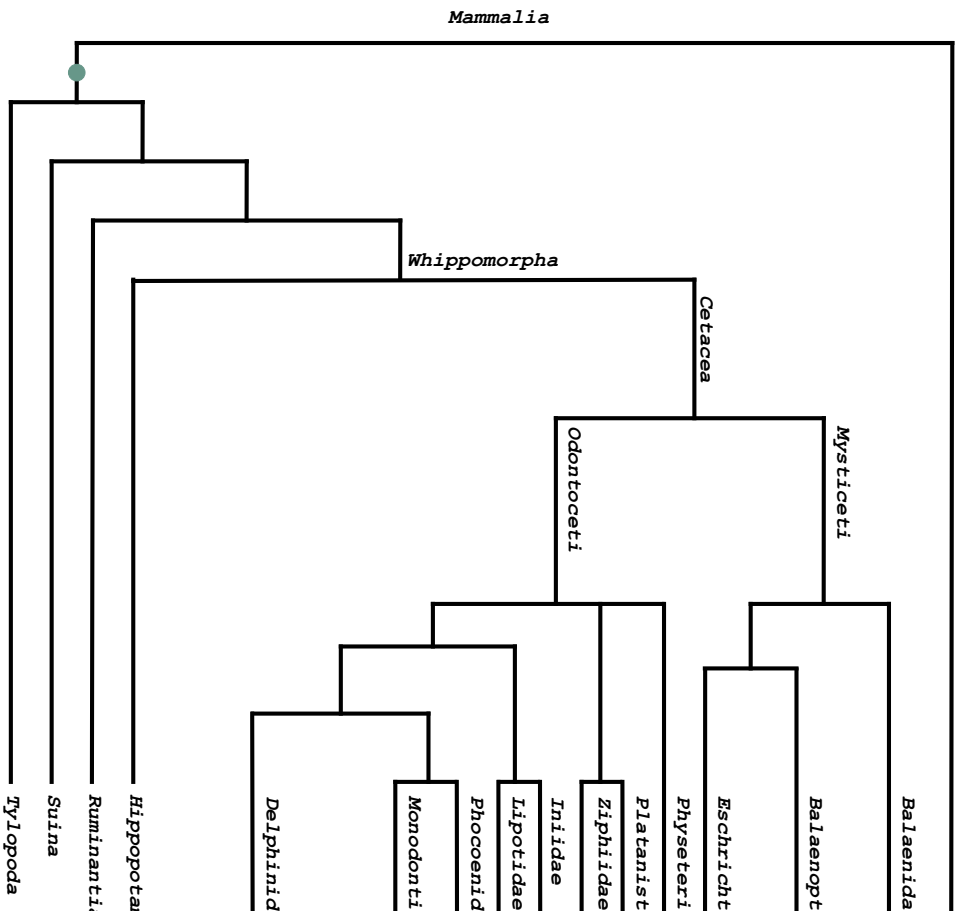

Fig. 1.4: Cetacea taxonomy tree, including the rest of families composing the order Artiodactyla (denoted by the green point). Marked in orange those specifically referred to in the text and hence used in the comparison studies. Marked with an asterisk are those which annotation is part of the work presented in this thesis, although punctual annotations for specific genes in the rest were also performed. Based in the relations shown by "Time tree" [Kumar et al., 2017].

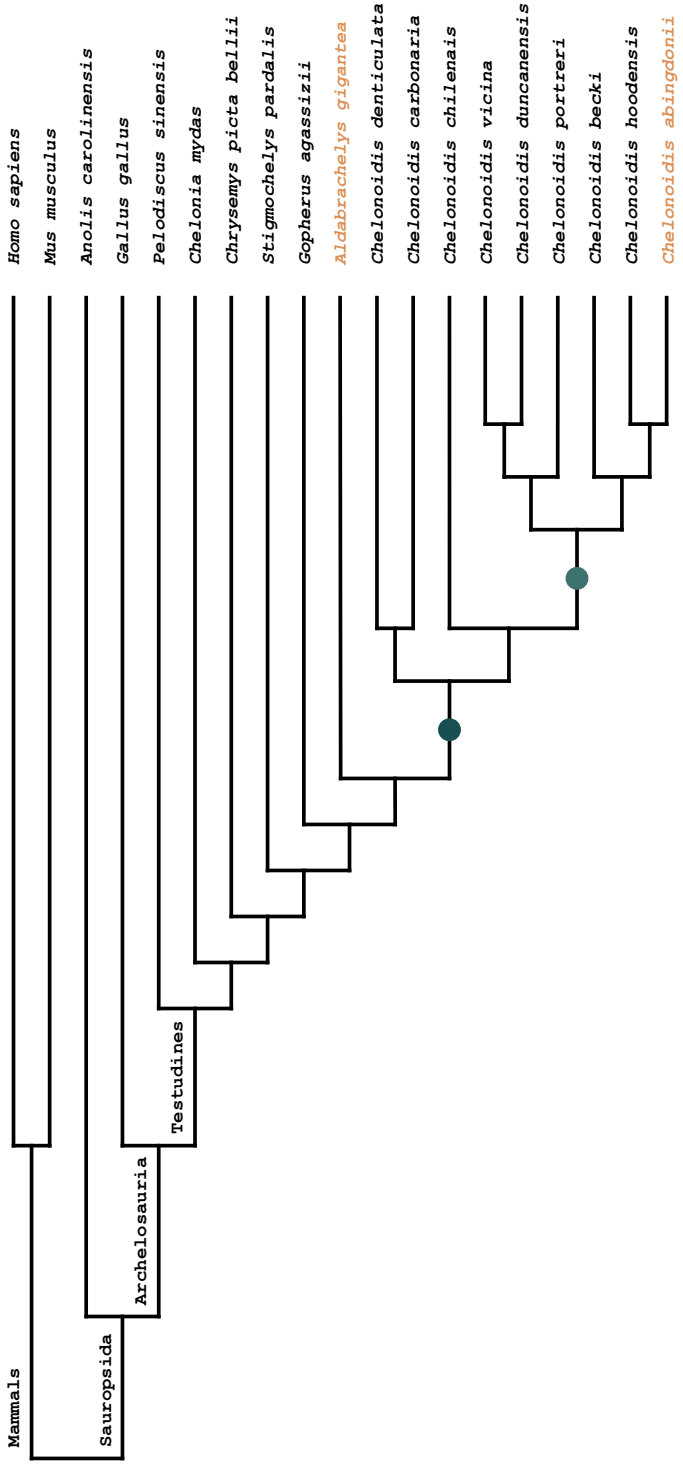

Fig. 1.5: Testudines taxonomi tree, including the rest of species used in the comparative studies as representatives of the main families comprising the clade of Sauropsida, and *H. sapiens* as the out-group an and reference. Marked in **orange** are those annotated as part of the present thesis, although punctual annotations for specific genes in the rest were also performed. Closely related tortoises both continental and from the Galápagos archipelago are indicated with the **darker green** point, while the radiation of island tortoises is marked using the **lighter green**. Based on the relations shown by "Time tree" [Kumar et al., 2017].

Objectives

Taking into consideration both the previous work in the field and the experience of our laboratory in degradomics, in this thesis we aimed to annotate, analyse and compare proteases and other genes in the genomes of long-lived metazoans.

Thus, the specific objectives for this Doctoral Thesis were:

1. Study and analysis of *Physeter macrocephalus*'s Degradome.
    - 1.1 Manual annotation of *P. macrocephalus*'s Degradome.
    - 1.2 Comparison of this gene set with that of the other annotated marine mammals (i.e. Bowhead whale, Minke whale, Bottlenose dolphin, and Killer whale).
    - 1.3 Study of the differences and similarities in an evolutionary context.
2. Study and analysis of *C. abingdonii*'s Degradome.
    - 2.1 Manual annotation of *C. abingdonii*'s Degradome.
    - 2.2 Comparison of this gene set with that of the other annotated tortoises (i.e. other Giant Galapagos tortoises, continental relatives, and Aldabra tortoise).
    - 2.3 Study of the differences and similarities in an evolutionary context.
3. Study and analysis of other genomic families of interest in *C. abingdonii*.
    - 3.1 Automatic annotation of the complete genome of *C. abingdonii*.
    - 3.2 Study of the evolutionary traits of the different families.
    - 3.3 Study of the genomic variation of gene families of interest among related species.

Materials and methods

3.1 Molecular biology methods

3.1.1 Data collection

Sperm whale

Tissue samples were obtained from an adult female specimen found in the northern Gulf of Mexico. DNA was then extracted with the DNAeasy kit from Quiagen, according to the manufacturer's protocol. RNA samples were collected from skin tissues (n=5) also from other specimens at the Gulf of Mexico using Trizol reagent (Invitrogen) according to the manufacturer specifications. RNA quality was assessed by electrophoresis using an Agilent 2100 Bioanalyzer (Santa Rosa, CA). Additional samples from four different individuals were extracted from "Voyage of the *Odyssey* samples" using a high-salt procedure [Godard et al., 2003]. The gender of each specimen was determined based on amplification by PCR of the *SRY* gene [Richard et al., 1994].

Galápagos giant tortoise

Both DNA and RNA samples from *C. abingdonii* were recovered from a frozen sample of blood belonging to Lonesome George. In parallel, we obtained samples of a granulome in an individual of *A. gigantea* from which we extracted DNA and RNA. Samples for PCR and Sanger sequencing were obtained from an array of samples provided by Yale University.

3.1.2 Genome sequencing

Sperm whale

Library collections for genome assembly consisted of paired-end (200 bp), and mate-pair libraries (insert sizes: 3, 8, and 40 kbp). All libraries were sequenced using paired 100 bp reads on an *Illumina HiSeq2000* ultrasequencer. Additional samples were sequenced to medium depth (20-30X) by sequencing paired-end short insert libraries (300bp) to 125bp length on an *Illumina Hi-Seq X10* instrument. All RNAseq data are available through the NCBI SRA, under BioProject number PRJNA177694.

Galápagos giant tortoises

Library collections built for sequencing on the *Illumina Hi-Seq 2000* platform, from a 180 bp-insert paired-end library, a 5 kb-insert mate-pair library and a 20 kb-insert mate-pair library. Additionally, reads from 18 PacBio SMRT^{TM} cells were used to extend the contigs. For the DNA sample from Aldabra tortoise we used Illumina technology and a 180 bp paired-end library to obtain whole-genome data. All the reads are available under BioProject number PRJNA416050.

3.1.3 RNA sequencing

Sperm whale

RNAseq paired-end data (100 bp length) were generated from Illumina TruSeq cDNA (stranded) libraries using the Hi-Seq 2000 instrument. All RNAseq data are available through the NCBI SRA, under BioProject number PRJNA177694.

Galápagos giant tortoise and other tortoises

RNAseq paired-end data were generated from Illumina TruSeq libraries using the HiSeq-2000 instrument. All the reads are available under BioProject number PRJNA416050.

3.1.4 Gene selection

Whenever manual annotation of genes was performed, the first step would be to select the set of genes to annotate. In the cases in which the target set to annotate is the Degradome, an already curated database is prepared. In other cases, genes were chosen after extensive literature mining using our experience in the field of ageing. In the case

of Lonesome George, more than 3,000 genes were chosen by this method. This number includes the more than 600 genes that comprise the Degradome. Unless otherwise noted, the sequences of all starting gene sets are taken from the human genome.

3.2 Bioinformatics methods

3.2.1 Genome assembly

Sperm whale

The combined sequence reads were assembled with the AllPaths software [Butler et al., 2008] using default parameter settings. This draft assembly was gap-filled with a version of Image [Tsai et al., 2010] that was modified for large genomes, and cleaned of contaminating contigs by performing a MegaBLAST [Zhang et al., 2000] of the contigs against bacterial and vertebrate genome databases. Contigs that displayed the best alignment over 50% of their length with a different species were removed. Using a genome size estimate of 2.8 Gbp, the total raw sequence depth of Illumina reads was $>$ 90X. The final sperm whale genome assembly was repeat-masked using WindowMasker [Morgulis et al., 2006].

Galápagos giant tortoise

The assembly of these libraries was performed with the AllPaths algorithm [Butler et al., 2008] to yield a draft genome of 64,657 contigs with an *N50* of 74 kb (Table 4.1). Then, we scaffolded these contigs using Sspace (v3.0) [Boetzer and Pirovano, 2014] employing the long-insert mate-pair libraries. Finally, we filled the gaps using PBJelly (v15.8.24) [English et al., 2012] and the reads obtained from 18 PacBio cells. The final assembly (*CheloAbing 1.0*) was 2.3 Gb long. Over this final assembly, we soft-masked repeated regions with RepeatMasker [Smit et al.,] using a database of chordate repeated elements (provided by the software) as reference. We then aligned the resulting reads to the *C. abingdonii* assembly with BWA (v0.7.5a) [Li and Durbin, 2009]. Similarly, raw genomic reads from *C. abingdonii* were aligned to *CheloAbing 1.0* for manual curation purposes.

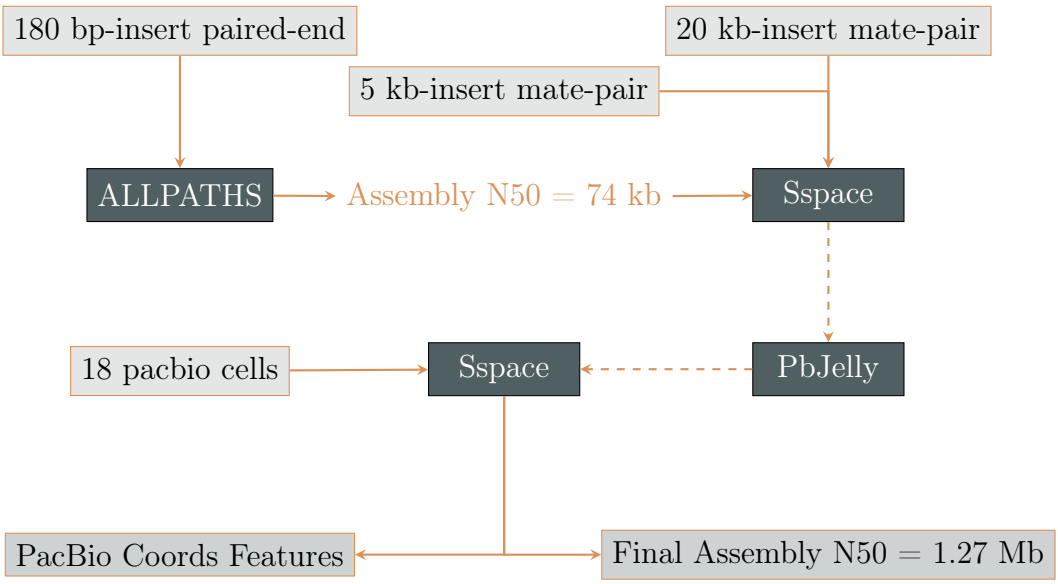

Fig. 3.1: Assembly process in Lonesome George's project.

3.2.2 RNA mapping and assembly

Galápagos giant tortoises

We aligned RNA-Seq data from *C. abingdonii* blood and *A. gigantea* granuloma to the assembled genome using TopHat (v2.0.14) [Trapnell et al., 2009].

3.2.3 Genome completeness assessment

The relative completeness in terms of expected gene content of the assembled genomes and their annotated gene sets was assessed using the Benchmarking Universal Single-Copy Ortholog (BUSCO) assessment tool [Seppey et al., 2019].

Sperm whale

In the case of the sperm whale, we used the laurasiatheria_odb9 lineage dataset that contains 6253 BUSCOs (v3.0.0). The dependencies used were Augustus v3.2.3, and HMMER v3.1b1 [Eddy, 2011].

Galápagos giant tortoise

In this case the program was run from an all-dependencies included Ubuntu virtual machine (available in https://busco-archive.ezlab.org/). For this assessment, we used

the vertebrata_odb9 lineage dataset. We performed this analysis *de novo* in CH38 human's assembly, to compare between them. Additionally, by using the published data of the Mojave dessert tortoise, we increase the scope of the comparison.

3.2.4 Genome automatic annotation

Galápagos giant tortoise

We performed de novo annotation on the genome assembly of *C. abingdonii* using *MAKER2*, a multi-threaded, parallelized computational tool designed to produce accurate annotations for novel genomes based on a machine-learning approach [Campbell et al., 2014]. We fed the algorithm with both the *CheloAbing 1.0* assembly and the RNA-Seq data, as well as reference genome sequences from human and *P. sinensis*. We also provided multifasta files of the complete annotated set of human and *P. sinensis* proteins. With this input, MAKER2 completed two runs in a Microsoft Azure virtual machine. Finally, predicted genes were assigned a putative function as part of the MAKER2 pipeline.

EBL-EBI annotation pipeline

In the automatic annotation process of the narwhal and the beluga whale, during my short internship in the European Bioinformatic Institute, I used a different approach. In its first step, this method takes advantage of the data repository inside ENSEMBL. Thus, the pipeline automatically searches for all the evidence used in the annotation process. The algorithm accepts a unique accession number for the assembly. By using the data associated to said number in the SQL databases, it finds and uses any RNASeq information available.

From this point on, a program which works on top of a LSF (a job scheduler), manages a swarm of scripts, each of which takes care of one part of the annotation process. They work on general tasks, such as the masking of the genome or the generation of an index, but also more annotation-related tasks, such as model generation, or comparison among related organisms. Finally, by using the main annotation softwares, such as Geneblast or Augustus, and comparing the results between all of them and also with the RNASeq-generated models, a final set is generated. This set is then manually revised to check for anomalies that may indicate a poorly performed annotation. In addition, the researcher is tasked with controlling the correct performance of the pipeline by checking guiHive, a program designed to graphically show the steps of the annotation, current

state, several inputs, outputs and options of each step, and the warnings and errors produced in the annotation.

3.2.5 Manual genome annotation

Manual annotation is largely based on the search for orthologs of our genes of interest in the genome of the species we want to annotate. The most commonly used tool for this task is *BLAST* [Altschul et al., 1990], an alignment algorithm designed to compare different sequences of nucleotides or amino acids. This alone would yield a low-quality set of annotated proteins, since, given the own bias of the aligner, there will be some errors in the sequence, specially in the exon-intron junction sequences. In order to correct the predicted alignment and assure the proper genetic structure, further steps are required. These steps of manual curation can be tedious and error-prone. To make the task easier and safer, we performed all manual annotations using the BATI algorithm, developed in this laboratory.

The main ideas behind this pipeline are to perform all alignments automatically, to provide a graphical environment to easily correct said alignments, and to summarize all the results in a comprehensive format that allows the user to effortless point out duplicates or new genes. This is achieved through four independent programs, that, once initialized, can be simultaneously used by several researchers working in the same set without conflicting each other's work. These scripts are written in Perl v5 and can be obtained from our group web site (http://degradome.uniovi.es/downloads.html).

1. `tbex` The first script to be executed. It prepares all required files and runs all the `tblastn` comparisons.
2. `bgmix` Summarizes all the hits from the different `tblastn` results in one single file.
3. `bsniffer` Generates a file per model gene, containing the `tblastn` result in a more readable format for users to choose.
4. `genetuner` Provides a graphical environment in which we can adjust intron-exon junctions and add or remove sequence stretches.
5. `bgmix` Non-redundantly summarizes all the `tblastn` hits, highlighting those belonging to an annotated gene. This allows the user to quickly find further copies of the genes under study.

`tbex` has two functions. The first one consists of creating a file containing the necessary data for the pipeline (*i.e.* the genome of the organism we want to annotate,

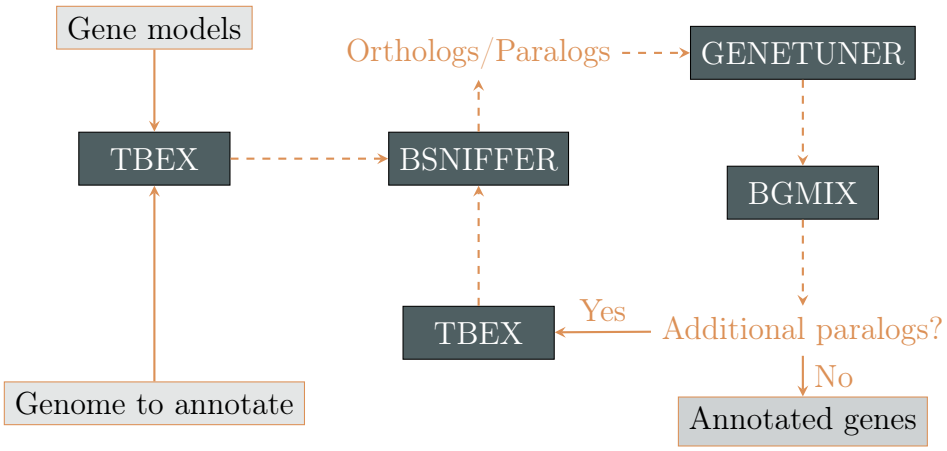

Fig. 3.2: Information flux in the execution process of BATI.

the protein sequence of the genes we are interested in annotate, and, optionally, the cDNA of said sequences, all in fasta format). The genome must be indexed. If necessary, the script itself will run `formatdb` on it. Once this is complete, the script will launch an instance of BLAST for each protein sequence we have given it. Specifically, the flavour of BLAST used is `tblastn` (as `blastall -p tblastn`, which will search the protein sequence given in a translated version of the nucleotidic genomic sequence. This flavour was chosen because evolutive pressures on genes are more evident on the protein sequence.

The program `bsniffer` generates a file per gene in which the results are reorganized in a more readable way. Additionally, the script calculates a score based on how complete, large and good match a hit is. Considering all of this, one can choose the best combination of hits, only being restricted by the contigs and not by the `tblasn` combinations. The program will also mark the hits that have already been used in building a gene to prevent their reuse. Once the best choice has been made, a predicted model is created and the next gene can be analysed in the same way.

The step run by `genetuner` provides a graphical interface that allows the user to move around the annotated exons of the gene, and shows the genome to be annotated and its three translated frames in the chosen strand. Besides the clearly distinguishable sequences of exons and introns, if the cDNA for the model genes was provided, this will also be available to double-check similarities. The aim of this step is to use the interactive view of the chosen alignment and to polish it as much as possible. This can be accomplished by correcting the exon-intron junctions, adding or deleting exons, or even splitting or mixing different exons (usually because of a *frameshift*). Some of these cases may be tackled by paying attention to the conserved splicing points, but in many

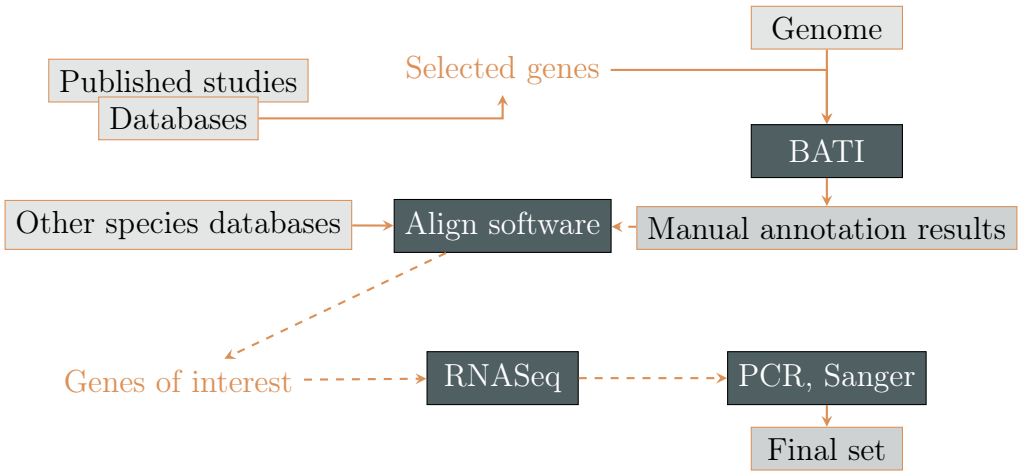

Fig. 3.3: Complete manual annotation process, including all the steps for the initial data to the final selection of corroborated genes of interest.

other cases it may be helpful to check external information sources such as published works regarding a specific gene, different databases, or related (and already annotated) organisms. This step is crucial, as it allows users to improve the annotation in ways that are not accessible during automatic annotation. The program also allows users to note down comments on particular regions. This is very important, since a wrong exon-intron union may be later interpreted as a spurious mutation. The graphical interactive interface includes several ways to move around, and several way to edit the current selection. Additionally, it allows the writing of warning when needed and the ability to increase or decrease sensitivity locally, as well as to perform different kinds of searches. This is the step where the researcher experience is most valuable.

`bgmix` summarizes all the hits in one file, indicating to which model protein each hit is more similar, and also if it has been used to build a gene already. Because of its usefulness, it is usually run twice, first after the `tbex` step, to have a general idea of the best match to each hit, and once everything is annotated, to check for unused hits, since those can build up a whole gene which can be a duplication or even a new one. If this is the case, the protocol is to duplicate the protein sequence file (and, if available, the cDNA) of the duplicated gene, and re-run the whole pipeline. Once this result of `bgmix` has no more obvious hits, we can consider the set of genes of interest annotated.

By using the data sets mentioned in subsection 3.1.4 the whole annotation process, including the pertinent comparisons and the final validation using Sanger is summarized in figure 3.3.

When manually annotating a genome, especially in the case of *de novo* genomes,

one must consider that some of the results may be artefacts produced by the *de novo* assembly. There are multiple causes for this error to appear, *e.g.* absence or reduced coverage of a specific regions of the genome leading us to think that genes may have been lost, or a lot of heterozygous positions concentrated in a region, which may provoke the assembler to assume that they are different regions, hence misidentifying a duplication. For this reason, hypotheses that arise from annotations must always be corroborated by other studies in order to be fully reliable. Some of these additional tests can range from checking the quality of the specific region we are interested on, or studying RNA-Seq data (if available), to performing PCR amplification and Sanger sequencing of the interesting region.

3.2.6 Expansion of gene families

We performed several pairwise alignments of the predicted proteins from the automatic annotation to the UniProt [Bateman, 2019] databases of human and *P. sinensis* proteins using BLAST (v2.6.011) [Altschul et al., 1990]. Using in-house perl scripts (available in a public repository (https://github.com/vqf/LG), we grouped these sequences into one-to-one, one-to-many, and many-to-many orthologous relationships. Only alignments with a coverage of at least 80% of the longer protein, and with more than 60% of identity were considered for the analysis. Finally, we searched for family expansions specifically present in *C. abingdonii*, by examining the aforementioned groups of orthologs. The results were manually curated. This way, we constructed extended orthology sets that may contain more than one sequence per species.

3.2.7 Positive selection

To search for signatures of selection affecting the predicted set of genes, we used BLAST and in-house perl scripts to pairwise align all available protein sequences from human (*H. sapiens*), mouse (*M. musculus*), dog (*Canis lupus familiaris*), gecko (*Gekko japonicus*), green anole lizard (*A. carolinensis*), python snake (*Python bivittatus*), common garter snake (*Thamnophis sirtalis*), Habu viper (*Trimeresurus mucrosquamatus*), budgerigar (*Melopsittacus undulatus*), zebra finch (*Taeniopygia guttata*), flycatcher (*Ficedula albicollis*), duck (*Anas platyrhynchos*), turkey (*Meleagris gallopavo*), chicken (*G. gallus*), Chinese softshell turtle (*P. sinensis*), green sea turtle (*C. mydas*) and painted turtle (*C. p. bellii*). We focused only on those genes with a one-to-one ortholog status in every species, and missing in no more than 3 species (excluding *C. abingdonii*),

as described in previous studies [Keane et al., 2015]. We then aligned each group separately with PRANK v.150803 using the codon model and analysed the alignments with `codeml` from the PAML package [Yang, 2007].

To search for genes with signatures of positive selection affecting *C. abingdonii* genes specifically, we executed two different branch models, M0, with a single $\omega 0$ value (where ω represents the ratio of non-synonymous to synonymous substitutions) for all the branches (nested), and M2a, with a foreground $\omega 2$ value exclusive for *C. abingdonii* and a background $\omega 1$ value for all the other branches. Genes with a high $\omega 2$ value (>1) and a low $\omega 1$ value ($\omega 1<0.2$ and $\omega 1 \approx \omega 0$) in *C. abingdonii*, but not in *P. sinensis* (Table B.1) were then considered as candidates to be under positive selection. As a control, M2a was repeated using *P. sinensis* as the foreground branch, and no overlapping genes were found in the result. Then, we used the M8 branch model to assess the individual importance of every site in these positively selected genes, obtaining a list of sites possibly under selection. The equivalent sites were examined in the Aldabra tortoise through alignments, to evaluate which of these important residues were altered (Table B.2).

3.3 Code development

3.3.1 Database management CGI

The different scripts, coded in Perl (version $>$ 5.20), and using a CGI protocol for executing them via web requests, orchestrate the interaction with the user, the creation/edition of the database, and the editing/creating of the final HTML that will display the website. In parallel, a couple of simple HTML files create the UIfor the editing of the database and the CGI request for the main set of programs. Briefly, the steps of the process would be as follow:

1. A JSON-build database, containing all the information about the degradome, degradomopathies, protease family, and information related to our laboratory (*e.g.* members, news, software, ...) is taken as input by one script. This JSON is then displayed as a interactive HTML-coded table.
2. In this table one can made editions, additions or even deletions. For each of these modifications, there is an appropriate button to execute the pertinent script and perform the desired change.
3. The invoked script then creates a copy of the database as it is (to be kept as a

security copy), then applies the required changes to the database, and finally calls for the last script.

4. This will take the new altered database as an input and build the different parts of the website taking into consideration the changes everywhere (*e.g.* if you add a protease, every line that mentions the number of proteases will change so that the number displayed is increased by 1).
5. Ideally, the user that made the modification should now check that everything is in order, since by repeating this process with a new modification will overwrite the saved copy. Ultimately, if even after checking some mistake is spotted and it is too late to recover the data not all will be lost, since an automatic process in the server keeps weekly security copies.

It is noteworthy that regular queries to the website work in a similar way as step 4, since the information in the JSON file is queried using the AJAX technology through jQuery, hence being dynamically fetched by request. If the browser lacks JavaScript or if this is disabled, the website will redirect the user to a static table with all the information.

Website Public Interface

The website is built using responsive-by-design technology, which allows the browser to "rearrange" the different HTML-containers in order to adapt it to the available display. Also, by interacting with the displayed information, the user may ask for more details of specific parts. By requesting this information interactively and using the aforementioned technologies waiting times for the loading of the different tables is reduced.

Results

4.1 Sperm whale degradome

We have annotated the complete set of proteases (or degradome) of the sperm whale using human proteins as model sequences, and we have performed several comparisons between them and those of other cetaceans and human. Overall, most of the predicted losses and gains of protease genes mirror those already described in minke [Yim et al., 2014] and bowhead whales [Keane et al., 2015]. Nevertheless, several events stand out as independent or specific, providing interesting hypotheses about the evolution of sperm whales in the context of cetacean evolutionary history (Figure A.1). In addition to the usual selective pressure on the immune and reproductive system of mammals, the unique aquatic environment of cetaceans has prompted numerous changes affecting protease genes involved in blood homoeostasis, digestion and skin maintenance.

4.1.1 Immunity and inflammation

In the context of immunology, we have found a conserved premature stop codon at the coding sequence of cetacean *MMP12*. Interestingly, besides the conserved premature stop codon (in p.W109), sperm whale present an additional, specific one in p.G93 (figure 4.1). This metalloprotease is preferentially expressed in macrophages, where it can modulate inflammatory responses. Its role is essential in multiple pathological settings, including lung emphysema [Ishii et al., 2014] and atherosclerotic disease [Proietta et al., 2014]. Notably, the loss of expression of *MMP12* in lung adenocarcinoma cells inhibits their growth and invasion capacities [Lv et al., 2015]. Similarly, *TPSAB1*, one of the major proteases present in mast cells, has been inactivated in all analysed cetaceans via conserved premature stop codon in p.D276 (figure 4.1). Additional premature stop

codons, both of them independently arisen, can be found in the sequences of sperm whale and minke whale, but they appear in non-conserved areas of the gene. This protease is secreted in the response to degranulation, and has been implicated as a mediator in the pathogenesis of asthma and other allergic disorders [Abdelmotelb et al., 2014].

In addition, we have found that *CASP12*, a modulator of the activity of inflammatory caspases, presents several premature stop codons as a result of a shared frameshift (starting in p.L101 and is likely a pseudogene (Figure 4.1; [McIlwain et al., 2015]). The annotations of this gene in the other cetaceans, suggests a complex evolutionary history. Specifically, Delphinoideos and Physeteroideos seem to share the same frameshift (starting in p.L30) which must have arisen via convergent evolution, since their closer relatives Monodontiddeos lack this alteration. Similarly, a different position presents another premature stop codon (p.R221*), shared among all cetacea but Delphinoidea (Figure A.3). While this caspase is conserved and functional in most terrestrial mammals, it is lost in most humans through a different mechanism involving the loss of the canonical stop codon. It has been shown that human displaying this protease in its active form are more sensitive to infection and sepsis [Saleh et al., 2004].

Finally, we have found interesting hallmarks of the evolutionary history of *MASP2* in cetaceans. This serine protease binds specific carbohydrates in the surface of invading microorganisms and activates the alternative complement pathway of innate immunity. Thus, MASP2 cleaves factors C4 and C2 to initiate the proteolytic cascade that leads to the formation of the membrane-attack complex. Consistent with this important role, a missense mutation in humans MASP2 (D105G), which abolishes its activity, correlates with severe immunological deficiencies [Stengaard-Pedersen et al., 2003]. However, the genome of the sperm whale shows a truncated version of this important gene because of a premature stop codon, which is upheld by the RNA-Seq analysis. This truncated ortholog also features a specific D105A change. Interestingly, dolphins also display a truncated version of this protein, in this case caused by a premature stop codon resulting from a frameshift at p.G454, again suggesting convergent evolution (Figure, A.4). Therefore, these data suggest that *MASP2* has been independently lost in at least two Odontoceti, but is present and functional in killer whales. Further studies will be necessary to identify the putative compensatory mutations that permit the loss of *MASP2* without any obvious disadvantage in certain odontocetes.

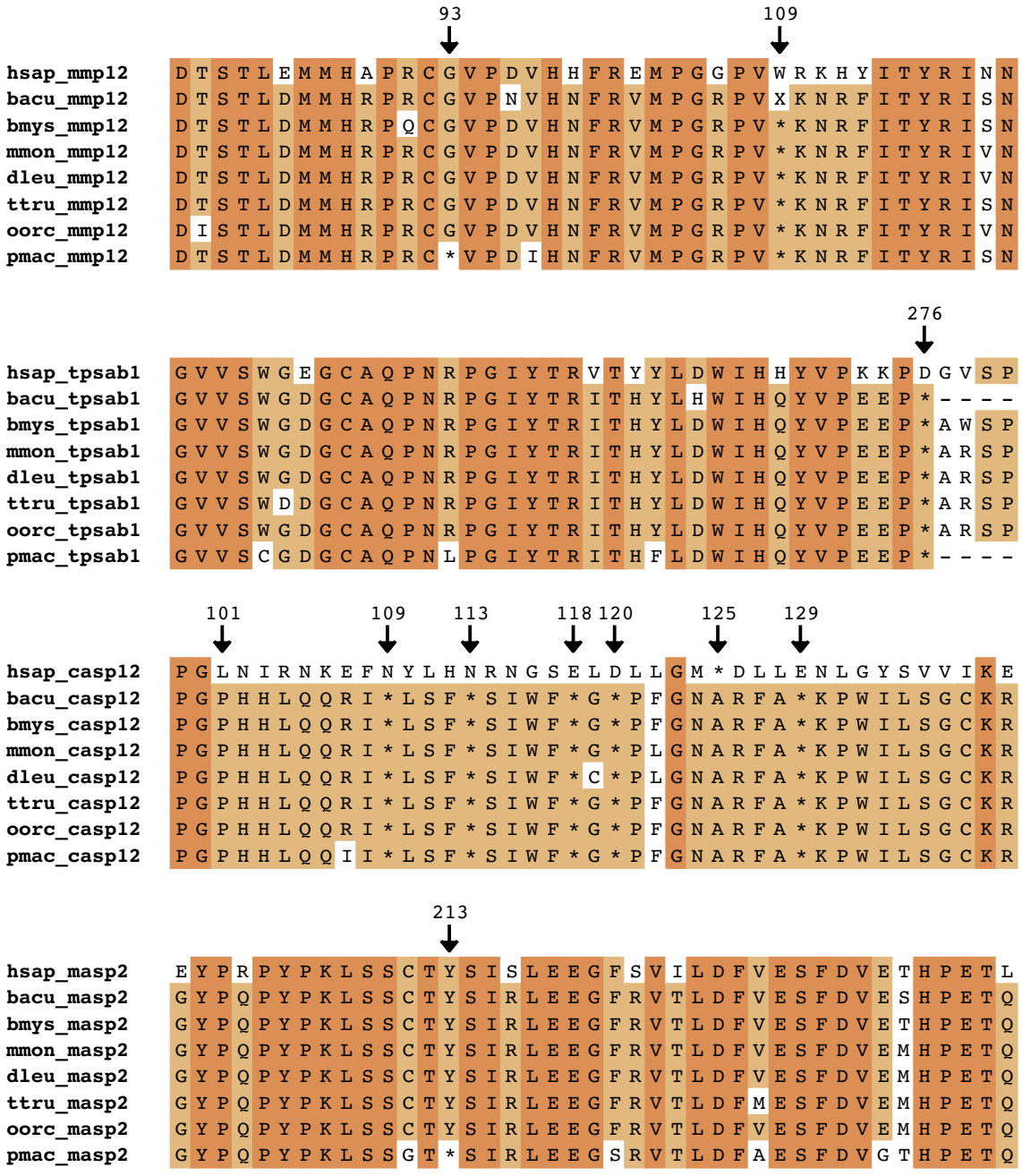

Fig. 4.1: Alignments of proteases related to immunology in cetaceans. Arrows mark the positions mentioned in the text, and the different intensity of colour reflects the degree of conservation within the alignment. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

4.1.2 Coagulation and blood pressure

We have confirmed the lack of both *F12* and *KLKB1*, two serine proteases which participate in the kinin-kallikrein system [Irmscher et al., 2018]. According to our annotation, *F12* was lost in a common ancestor to all cetaceans, presenting 2 conserved premature stop codons (p.Y391* and p.E521*). An alignment of the annotated sequences also shows evidence of more recent events, like an additional premature stop codon shared by all non-Physeteroideos Odontoceti in p.E514*, an additional premature stop codon shared by Delphinoideos in p.W406* (Figures 4.2, and A.5). On the other hand, *KLKB1*, seems to be in different states of pseudogenization in Misticeti and appears to be completely absent in some Odontoceti. However, it must be noted that complete gene losses can be mimicked by assembly artefacts. The kinin-kallikrein system is important in inflammation, blood pressure control, coagulation and pain [Verweij et al., 2013].

Furthermore, the related serine proteases *F7*, *TMPRSS11F* and *TMPRSS11B* have been shown to constitute targets of selection in Mysticeti, probably related to their role in coagulation [Keane et al., 2015]. In our data set, *F7* appears to be a functional gene in sperm whales, whereas both *TMPRSS11F* and *TMPRSS11B* have been lost through premature stop codons. Specifically, *TMPRSS11B* seems to be in different states of pseudogenization in the analysed cetaceans. It presents a common premature stop codon shared by all Delphinoidea (p.R402*), whereas in sperm, bowhead, and Minke whales a different point mutation has caused the gain of a premature, non-conserved stop codon (Figures 4.2, and A.6). *TMPRSS11F*, on the other hand, presents a frameshift (starting in p.R411), conserved in all cetaceans, which causes several premature stop codons (Figure 4.2).

4.1.3 Skin homoeostasis

The cysteine-protease *CAPN12* has been lost through different, non-overlapping premature stop codons and frameshifts in sperm whale, dolphin, and bowhead and minke whales. In the case of the sperm whale specifically, the truncation is produced by one frameshift starting at p.F131 (Figure 4.3). This protease is preferentially expressed at the cortex of the hair follicle [Dear et al., 2000].

Likewise, the serine-protease KLK8 appears to be absent in all analysed cetaceans, except for sperm whales, through two different mechanisms, each specific to one parvorder. While mysticetes share a common premature stop codon at p.88, Delphinoidea

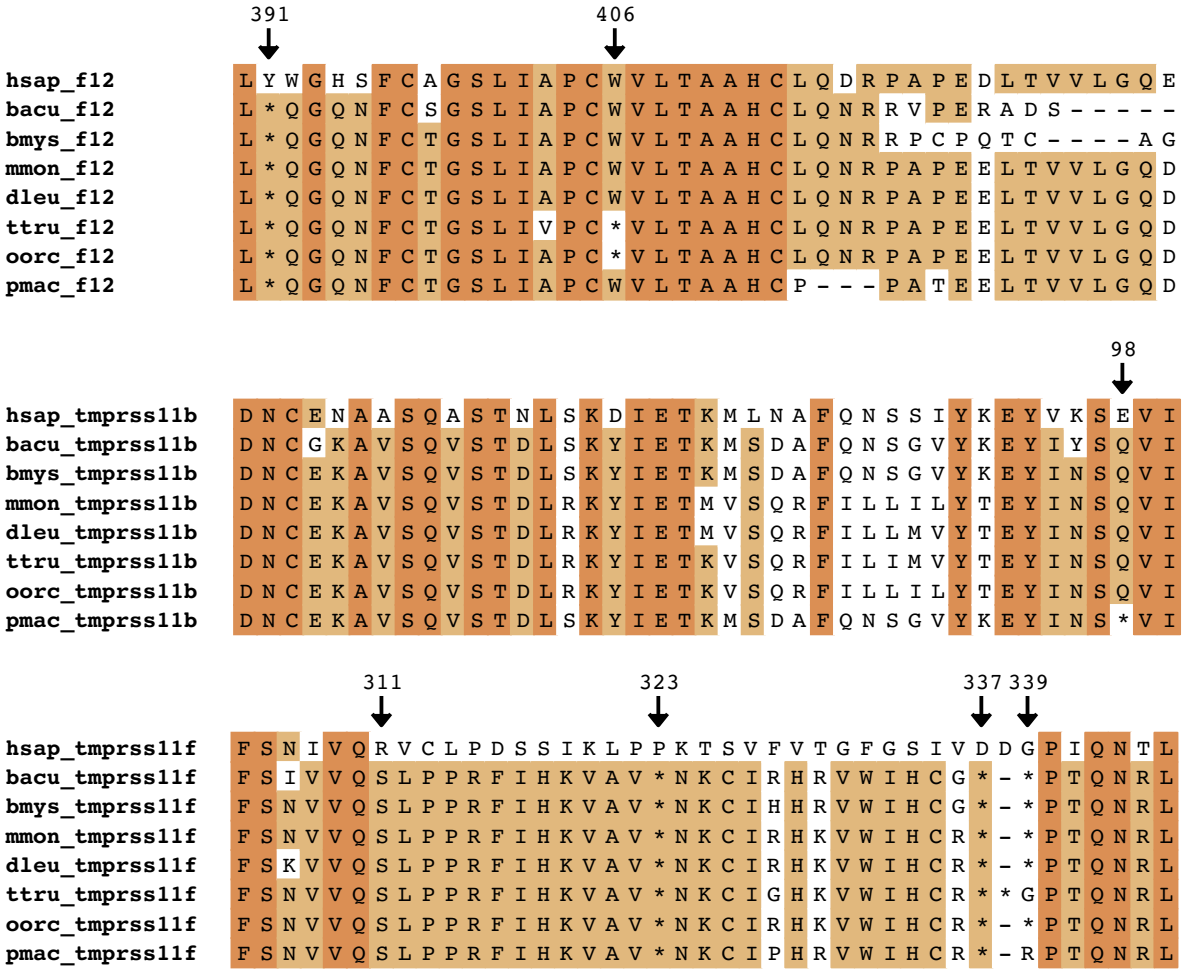

Fig. 4.2: Alignments of proteases related to blood homoeostasis (and coagulation) in cetaceans. Arrows mark the positions mentioned in the text, and the different intensity of colour reflects the degree of conservation within the alignment. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

displays a different premature stop codon at p.170. In sperm whales, this KLK8 displays a complete open-reading frame, but its catalytic site is mutated (starting at p.G255A) to a theoretically inactive form (Figure 4.3). Therefore, sperm whale's *KLK8* is expected to be a functional gene producing a non-functional protease. Nevertheless, inactive proteases can be important, since they can still bind and sequester substrates.

Finally, another kallikrein, *KLK7*, seems to have been specifically lost in a common ancestor of mysticetes, but not in the odontocetes analysed (including sperm whales). Both *KLK7* and *KLK8* have been related to skin homoeostasis [Kishibe et al., 2007],

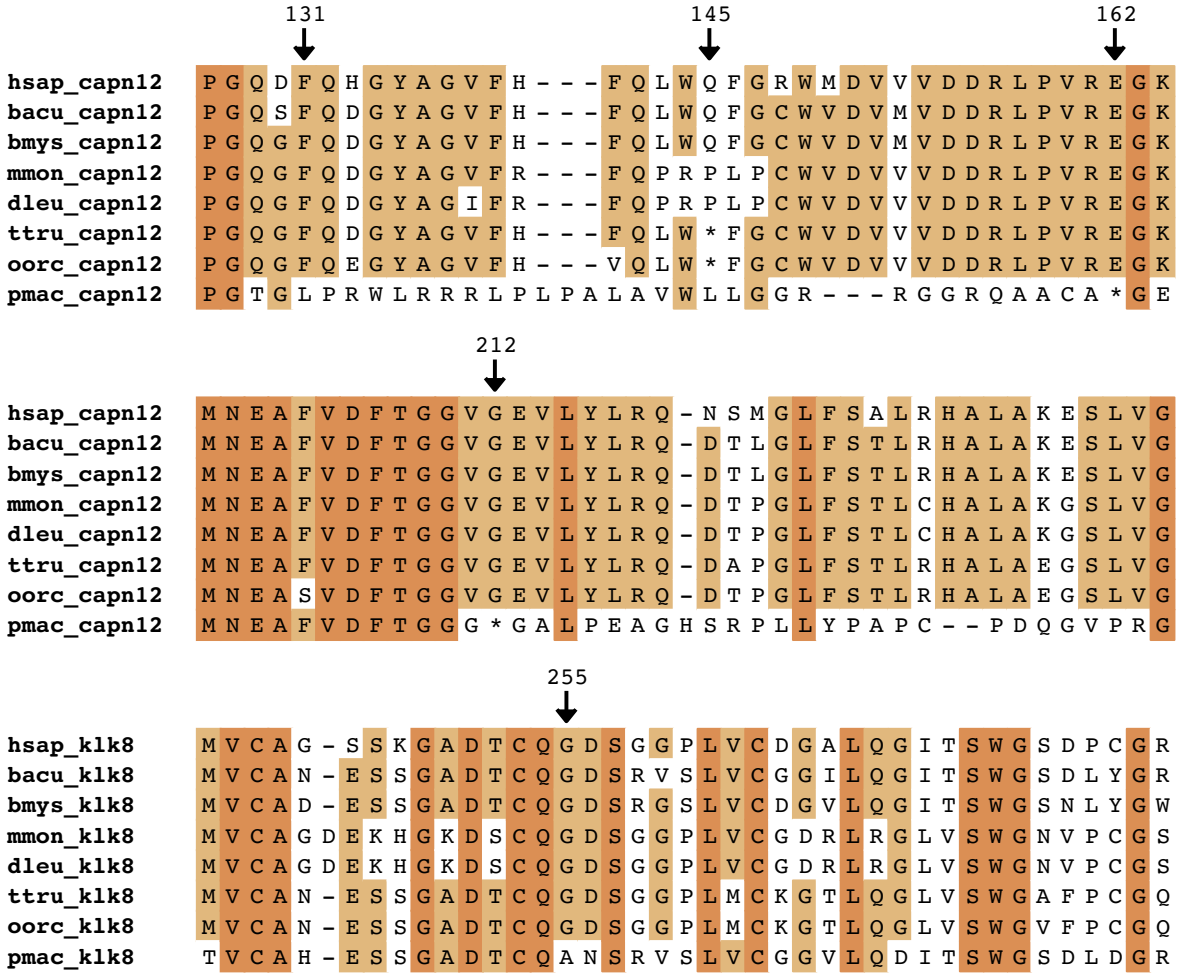

Fig. 4.3: Alignments of proteases related to skin homoeostasis in cetaceans. Arrows mark the positions mentioned in the text, and the different intensity of colour reflects the degree of conservation within the alignment. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

and some experiments suggest that *KLK8* might be directly involved in the terminal differentiation and desquamation of the *stratum corneum*, the utmost layer of the skin in mammals [Kuwae et al., 2002].

4.1.4 Digestive system

The evolution of the digestive system of these mammals has included the loss of several metallocarboxypeptidases from the M14 family. Thus, *CPA2*, *CPA3* and *CPO* were lost in all cetaceans examined. Surprisingly, *CPA3* may have been lost independently

in sperm whales and in an ancestor of baleen whales. Dolphin *CPA3* also features an independent premature stop codon. Furthermore, sperm-whale *CPB1*, also from the M14 family, features two premature stop codons not present in Mysticeti. Notably, those stop codons are not present in the corresponding ortholog from dolphins. As expected, odontocetes retain functional orthologs of *KLK4* and *MMP20*, two proteases involved in dentition which are lost in mysticetes [Keane et al., 2015].

4.1.5 Sperm whale-specific traits

Several important events in the sperm-whale degradome seem to be specific for this mammal, and probably merit further study. Thus, *MMP7* (matrilysin-1), a metalloprotease with the ability to cleave scaffolding proteins in the extracellular matrix, contains a premature stop codon (p.W149*) in sperm-whales (figure 4.4) [Grindel et al., 2018]. The second member of the matrilysin sub-family, *MMP26* or *matrilysin-2*, is primate-specific [Uria and Lopez-Otin, 2000]. If confirmed, this would be the first case of spontaneous lack of matrilysins in a mammalian species.

Finally, the cysteine-protease *CASP3* has been specifically duplicated in sperm whales in a retrotranscription-involving event (Figure 4.5). The resulting single-exon duplicate contains a complete, uninterrupted open reading frame and features few amino acid changes compared to the original form. This protease is involved in one of the proteolytic cascades leading to apoptosis [McIlwain et al., 2015]. In addition to its putative role in cancer progression, *CASP3* has also been involved in brain physiology as the predominant caspase that cleaves amyloid-beta 4A precursor protein.

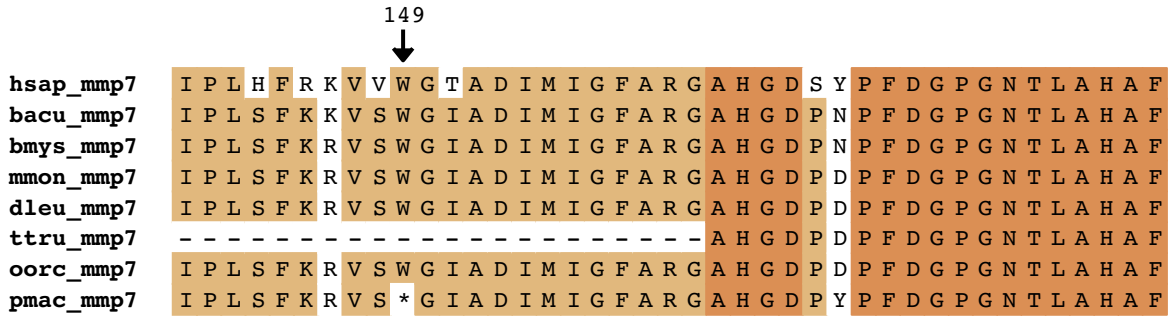

Fig. 4.4: Alignment of protease MMP7 in cetaceans. The arrows marks positions mentioned in the text. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `hsap`, *H. sapiens*.

4.2 Galápagos giant tortoise genome analysis

4.2.1 Genome assembly

We used Illumina paired reads, mate pairs and PacBio reads to assemble the genome of Lonesome George in 10,623 scaffolds with an N50 of 1.27 Mb, with the largest scaffold

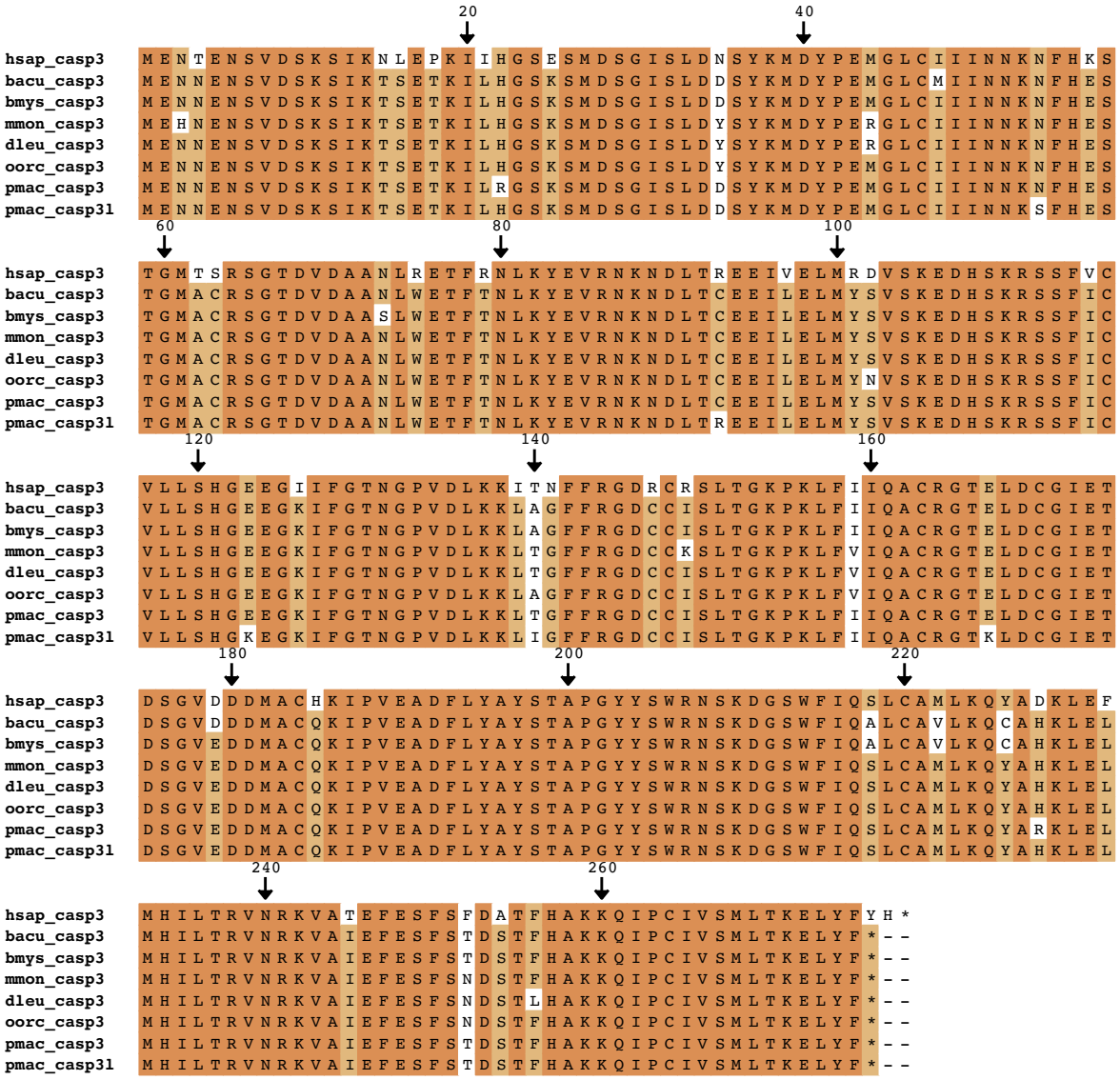

Fig. 4.5: Alignment of protease CASP3 in cetaceans, including the sequence of sperm whale's specific duplication. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

being longer than 10 Mb (Table 4.1).

Table 4.1: *C. abingdonii* genome statistics.

	Scaffolds with GAPS	Scaffolds without GAPs	Contigs with Ns
Seqs	10,623	10,623	10,623
Min	886 bp	886 bp	130 bp
Median	6,499 bp	4,304 bp	16,213 bp
Mean	216,581 bp	204,233 bp	33,555 bp
Max	10,495,589 bp	10,223,643 bp	1,220,627 bp
Total	2,300,749,194 bp	2,169,570,871 bp	2,169,570,871 bp
N50	1,277,207 bp	1,227,724 bp	74,527 bp
N90	337,476 bp	331,274 bp	18,547 bp
N95	174,219 bp	172,536 bp	10,818 bp
Non-gapped Ns		23,455 bp	

The final assembly (*CheloAbing 1.0*) is 2.3 Gb long, a size consistent with other turtle assemblies such as *C. p. bellii* (2.59 Gb) [Bradley Shaffer et al., 2013], and *C. mydasa* and *P. sinensis* (both around 2.2 Gb) [Wang et al., 2013].

According to the masking procedure, 28.5% of the genome represents repetitive elements. Of these, the larger group is that of retroelements, which add up to 17.86% of the assembly (Table 4.2).

Table 4.2: Repeated elements in the genomes of *C. abingdonii* and *C. p. bellii*, showing number of elements (`n`), total length (`bp`), and percentage of genome covered (`%`).

	C. abingdonii			*C. p. bellii*	
Repeat type	n	bp	%	bp	%
SINEs	215,691	33,047,802	1.40	44,277,662	1.87
PLEs	156,168	35,941,354	1.56	-	-
LINEs	572,190	231,609,979	10.07	248,848,977	10.52
LRTs	243,123	146,157,459	6.35	123,030,173	5.20
DNA Transposons	1,044,675	227,423,152	9.88	18,952,044	9.80
Unclassified	84,016	17,050,133	0.74	20,076,666	0.80
Small RNAs	42,858	8,412,400	0.37	8,527,612	0.36
Simple repeats	8,199	1,475,310	0.06	11,395,517	0.48
Low complexity	192	64,291	∼0.00	2,244,030	0.09

Additionally, since PacBio sequencing features frequent errors, we compiled regions covered only by these reads and added them to the header of the assembly.

A file containing this information is available in the assembly entry at the NCBI database (www.ncbi.nlm.nih.gov/nuccore/PKMU00000000.1/). Point variants belonging to these regions are not reported unless validated by other means.

In parallel, we assessed the suitability of *CheloAbing 1.0* for gene annotation. From this analysis, we located 96.4% of the common core of conserved genes, out of which 0.6% were duplicated genes. While the percentage of fragmented genes (5.7%) was relatively high, these results are compatible with a moderately good quality of assembly, with only 1.9% totally missing. In other species considered for comparison, the percentage of missing genes was always higher, while maintaining a similar completeness percentage (Table 4.3).

Table 4.3: Comparative BUSCO analysis of *C. abingdonii*, published data from *G. agassizii*, and *de novo* analysis of *H. sapiens* (CH38).

	C. abingdonii		*G. agassizii*		*H. sapiens* (CH38)	
	n	%	n	%	n	%
Completed	2,391	92.4	2,387	92.4	2,355	91.1
Duplicated	16	0.6	22	0.9	72	2.8
Fragmented	147	5.7	138	5.3	112	4.3
Missing	48	1.9	61	2.3	119	4.6
Total	2,586	100	3,023	100	2,586	100

4.2.2 Automatic annotation of *CheloAbing 1.0*

As a result of the automated annotation process, we obtained 27,208 predicted genes with putative functions assigned. A multifasta file containing all protein sequences is available in a public repository (https://github.com/vqf/LG).

With these data, we constructed extended orthology sets that may contain more than one sequence per species. These orthology sets capture most of the known protein families, although some of these families appear split according to sequence similarity. Almost all of these splits occur both in the human-to-*P. sinensis* and in the human-to-*C. abingdonii* comparisons. Since assembly errors may mimic gene losses, we decided to only test these sets for *C. abingdonii*-specific expansions. The interrogation of these sets suggests the existence of several high copy-number gene families in tortoises and turtles but absent in humans. Most of these families show homology to viral retrotranscriptases, consistent with the hypothesized expansion of CR1 retrotransposons in ancestral amniotes followed by dominance of L1 LINEs in therians [Suh et al., 2014].

After manual curation of the results, we found 12 examples of gene families displaying extra copies in *C. abingdonii* compared to turtles (Table 4.4). Each of those genes was also identified from the aligned reads in the Aldabra giant tortoise genome, and 10 of these amplifications were also identified in the genome of Agassiz's desert giant tortoise [Tollis et al., 2017b]. Interestingly, a functional annotation clustering analysis of these 12 genes found a significant enrichment in the "extracellular exosome" GO category (8 genes; *ATP6V1A*, *EEF1A1*, *EEF2*, *RPL11*, *RPS25*, *STXBP1*, *TPT1*, and *VCP*; $p = 0.0021$ after Benjamini correction).

Table 4.4: Tortoise-specific gene expansions, including *H. sapiens* as a reference and *P. sinensis*, representing turtles.

	H. sapiens	*C. abingdonii*	*G. agassizii*	*P. sinensis*
OTX2	1	2	2	1
ATP6V1A	1	2	2	1
LAMTOR4	1	2	2	1
STXBP1, STXBP2	2	4	6	2
TPT1	1	2	1	1
VCP	1	2	2	1
RPL11	1	2	2	1
EEF2	1	2	2	1
GJD2	1	2	2	1
RPS25	1	2	1	1
EEF1A1, EEF1A2	2	6	5	3
POLR2L	1	2	2	1

Next, we checked for signatures of positive selection. This analysis pointed to multiple biochemical pathways which may have been affected by selection. Two of the top three genes with the strongest evidence for positive selection were tubulins (*TUBE1* and *TBG1*), suggesting that changes in cytoskeletal dynamics have been important in the evolution of giant tortoises. Consistent with the role of this pathway in the biology of the cell, the alterations found in the selected residues (p.E169D and p.I186V in TUBE1) are conservative and probably do not affect the main role of this protein in microtubule formation. In addition, two genes with evidence for positive selection, *BAG2* (*NEF*) and *UBE2J1* (*Ubc6/7*) are involved in endoplasmic-reticulum-associated protein degradation (ERAD).

Notably, one of the genes duplicated in giant tortoises (*VCP*, also known as *p97*) also plays a central role in this pathway. In this regard, the list of positively selected genes also features *TDO2*, whose product is involved in the regulation of tryptophan-mediated proteostasis. Interestingly, the inhibition of *TDO2* has been shown to protect against age-related neurodegeneration [Breda et al., 2016].

In addition, two positively selected genes, *AHSG* and *FGF19*, are listed in a panel of four proteins whose expression levels correlate with successful ageing in humans [Sanchis-Gomar et al., 2015]. Notably, one of the selected alterations in FGF19 (p.S116A) is expected to affect the receptor-interaction site of this protein. These factors intervene in the regulation of glucose and lipid metabolism [Kir et al., 2011, Pal et al., 2012], another hallmark of ageing, suggesting that the adaptation to the challenges that longevity poses on this system may have been important in the evolution of giant tortoises. Finally, three genes with evidence for positive selection in these organisms (*MVK*, *IRAK1BP1* and *IL1R2*) play important roles in the modulation of the immune system, which in turn participates in the phenotypes of altered intercellular communication associated with ageing [López-Otín et al., 2013].

4.3 Galápagos giant tortoise degradome

We manually annotated more than 600 protease genes in *CheloAbing 1.0*, using our human degradome database as a reference to predict each ortholog and any additional paralogsi (Figure A.2).

4.3.1 Immunology

We identified some features in the genomes of *C. abingdonii* and *A. gigantea* that support the enhanced role of innate immune defence in these organisms. Specifically, we found some putatively deleterious changes in genes involved in B-lymphocyte maturation. As an illustrative example, *MEP1A*, a metalloprotease involved in the activation of interleukin-6 (IL6), shows premature stop codons, the first at p.Q320*, due to the presence of two separate frameshift mutations detected only in Galáapagos giant tortoises (Figure 4.6), as assessed by Sanger sequencing validation. Consistent with the role of IL6 in the different aspects of the immunolgy of B cells [Eto et al., 2011], disruptions in MEP1A have been associated with altered homoeostasis of monocytes and natural killer (NK) cells in mice [Sun et al., 2009].

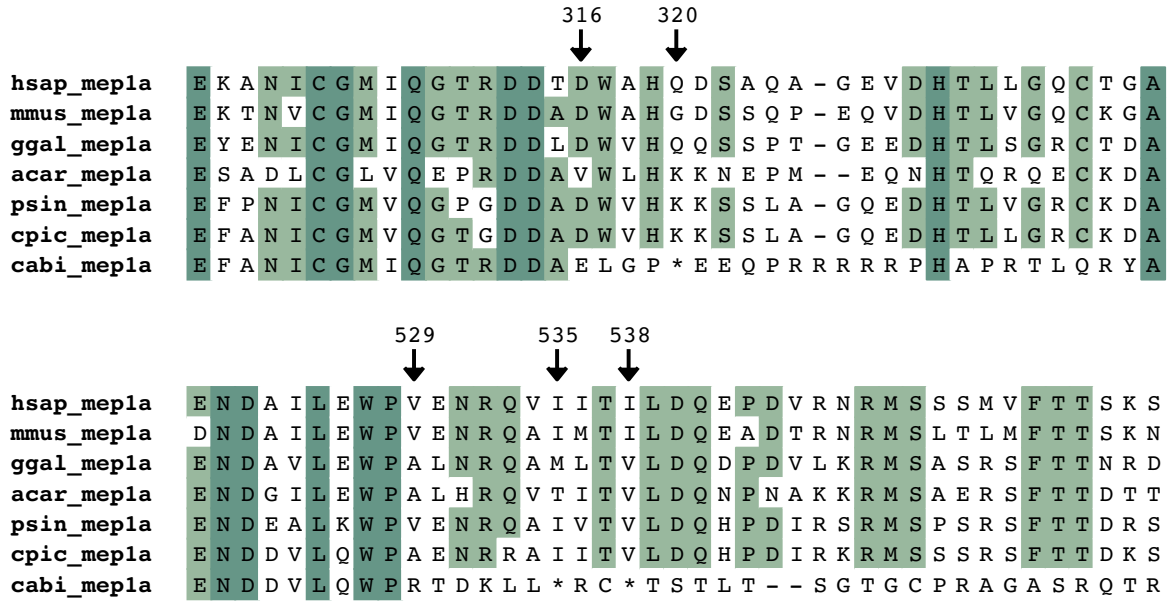

Fig. 4.6: Alignment of protease MEP1A in testudines, showing the two frameshifts that cause the truncation of its function. Arrows mark positions of interest, such as the start of frameshifts or the resulting premature stop codons. `hsap`, *H. sapiens*; `mmus`, *M. musculus*; `acar`, *A. carolinensis*; `ggal`, *G. gallus*; `psin`, *P. sinensis*; `cpic`, *C. picta belli*; and `cabi`, *C. abingdonii*.

Moreover, we found a family expansion involving the granzyme (GZM) serine proteases (Figure A.7). This well-conserved family of proteolytic enzymes is expressed in 3 clusters, the chymase, the met-ase and the GZM A/K loci, the latter being conserved among Craniata [Akula et al., 2015]. The chymase locus, encoding *CMA1*, *CTSG*, *GZMA*, *GZMB*, and *GZMH*, is greatly expanded, with 1 extra copy of *CMA1*, 6 extra copies of *CTSG*, 4 extra copies of *GZMB*, and 1 extra copy of *GZMH* (Table 4.5). In addition, several other copies appear to have been pseudogenised and hence are not included in the previous counting. These duplications detected in the genome of *C. abingdonii* are also present in all the other Galapagos giant tortoises tested and in *A. gigantea*, as assessed by Sanger sequencing of these genomic regions. We also detected a similar amplification in the genome of *G. agassizii*. Although some of these copies are pseudogenes, this copy-number variation evidences the importance of innate immunological pathways in tortoises. These serine proteases are key components of the CTL and NK cell secretory granules, playing important roles in defence against both pathogens and cancer [Akula et al., 2015].

Table 4.5: Percentages of identity and coverage between members of the chimase locus in Galápagos giant tortoises.

	Nucleotide		Protein	
Gene	coverage (%)	identity (%)	coverage (%)	prot_identity (%)
CMA1L_1	71,10	62,10	73,10	46,20
CTSGL_1	40,60	30,00	45,30	18,60
CTSGL_2	29,90	22,30	32,00	15,00
CTSGL_3	73,10	55,10	84,10	42,50
CTSGL_4	64,80	47,70	76,60	36,30
CTSGL_5	70,10	53,20	82,90	41,60
CTSGL_6	70,40	52,20	81,90	37,50
GZMBL_1	59,50	44,60	61,90	26,40
GZMBL_2	59,50	44,00	66,10	24,60
GZMBL_3	50,70	39,00	71,20	22,10
GZMBL_4	57,50	42,50	59,60	21,90
GZMHL_1	76,00	59,80	96,20	53,00

Immune regulators of inflammation

We found that *CASP12* was apparently functional in all Testudines, without changes in essential residues.

4.3.2 Coagulation

We next analysed different genes involved in the coagulation pathway, as both coagulation and blood homoeostasis can be greatly impacted by environmental changes and species adaptation to new habitats [Keane et al., 2015]. We found interesting variations in some members of the coagulation factor family, such as factor VII, factor X, and factor XI (figure 4.7). These variants, common to all species of Galápagos giant tortoises tested, *A. gigantea*, and *G. agassizii*, lead to putative F7 (p.F64L), F10 (p.E91K), and F11 (p.E315K) deficiencies. This could result in altered functions of these proteases. In humans, mutations affecting these residues are associated with pathologies such as factor VII, X, and XI deficiency [Al-Hilali et al., 2007, Quélin et al., 2006].

Additionally, *KLKB1*, encoding a serine protease that participates in the kinin-kallikrein system [Wong and Takei, 2013], was absent both in *C. abingdonii* and in *P. sinensis*. This loss suggests diverse blood pressure and coagulation control mechanisms, compared to mammals [Khan et al., 2007]. Likewise, plasminogen (*PLG*) displays point

variants in fibrin-binding sites, like p.R134L, also present in *A. gigantea* and *P. sinensis*, and p.R136K, found in all studied turtles (Figure 4.7). Finally, the serine protease

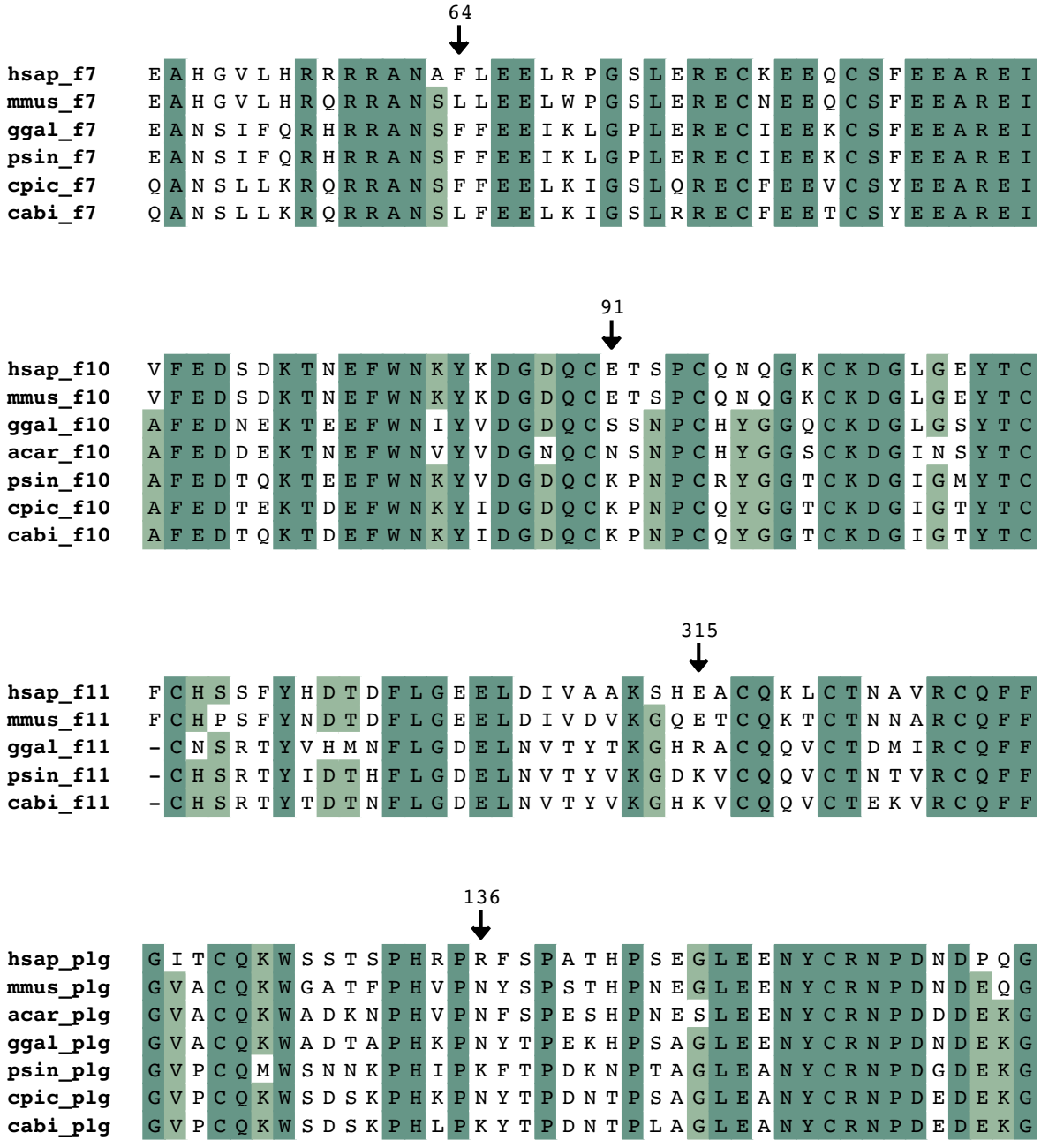

Fig. 4.7: Alignments of proteases related to blood homoeostasis (and coagulation) in testudines and related organisms. Arrows show positions mentioned in the main text. `hsap`, *H. sapiens*; `mmus`, *M. musculus*; `acar`, *A. carolinensis*; `ggal`, *G. gallus*; `psin`, *P. sinensis*; `cpic`, *C. picta belli*; and `cabi`, *C. abingdonii*.

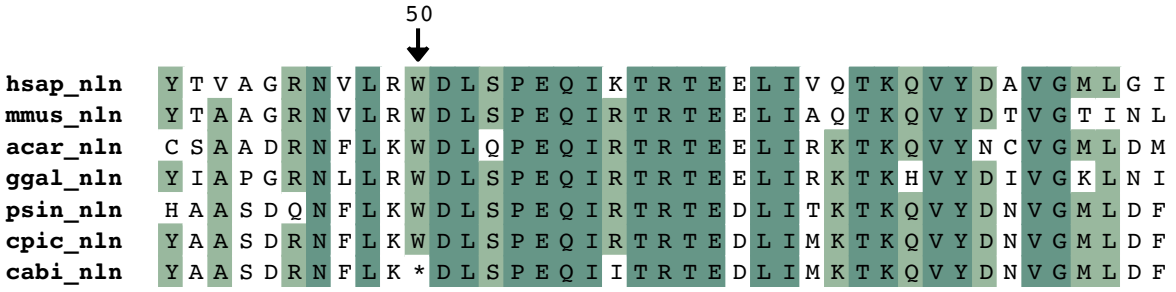

Fig. 4.8: Alignment of protease NLN in testudines and relatives. Arrows mark the premature stop codon gained in giant tortoises. `hsap`, *H. sapiens*; `mmus`, *M. musculus*; `acar`, *A. carolinensis*; `ggal`, *G. gallus*; `psin`, *P. sinensis*; `cpic`, *C. picta belli*; and `cabi`, *C. abingdonii*.

PROC that inhibits the generation of plasmin, was absent only in *C. abingdonii*.

4.3.3 Metabolism and diet

Glucose is one of the most important carbohydrates due to its pivotal position as a source of energy, as well as a precursor of vitamins and different polymers essential for cells. Among the multiple enzymes involved in these metabolic pathways, the mitochondrial metalloprotease neurolysin (*NLN*) is a key component in multiple glusose-related processes. In *C. abingdonii*, this gene is truncated due to a premature stop codon (p.W50* (Figure 4.8), present in all tested species of Galápagos giant tortoises and in their continental outgroups, but not in Aldabra tortoise (as validated through Sanger sequencing). The genome of *G. agassizii* shows a different premature stop codon, which suggests a case of convergent evolution.

Additionally, *C. abingdonii* also displayed a duplication of *CTRB1* (Figure 4.9), a pancreatic serine protease involved in digestion. While this duplication is shared by *A. gigantea*, it is not duplicated in other Sauria.

4.3.4 Development features

Neurological alterations

Motopsin (*PRSS12*), a serin protease linked to nonsyndromic mental retardation [Mitsui et al., 2013], appeared to be truncated exclusively in *C. abingdonii*, as validated by Sanger sequencing, due to a premature stop codon at p.691 (Figure 4.10). Similarly, the aspartyl protease *PSEN1* was found to present a point variant at position p.R352E common to all Sauropsida (Figure 4.10). This variant was validated by Sanger

sequencing in giant tortoises, both from Galápagos and from Aldabra, and their respective outgroups. In humans, a mutation in this same residue (p.R352C) has been linked to the early development of Alzheimer's disease [Jiang et al., 2015, Ryazantseva et al., 2016].

In addition, neurology-related protease *XPNPEP1* presents a premature stop codon (confirmed by RNA-Seq analysis) at the beginning of the protein (p.D16*). This variant, present in all giant tortoises (including all of the Galápagos tortoises, and Aldabra), and their continental relatives, and validated through Sanger sequencing, likely results in the loss of the function of the protease. The absence of *XPNPEP1* in humans is associated

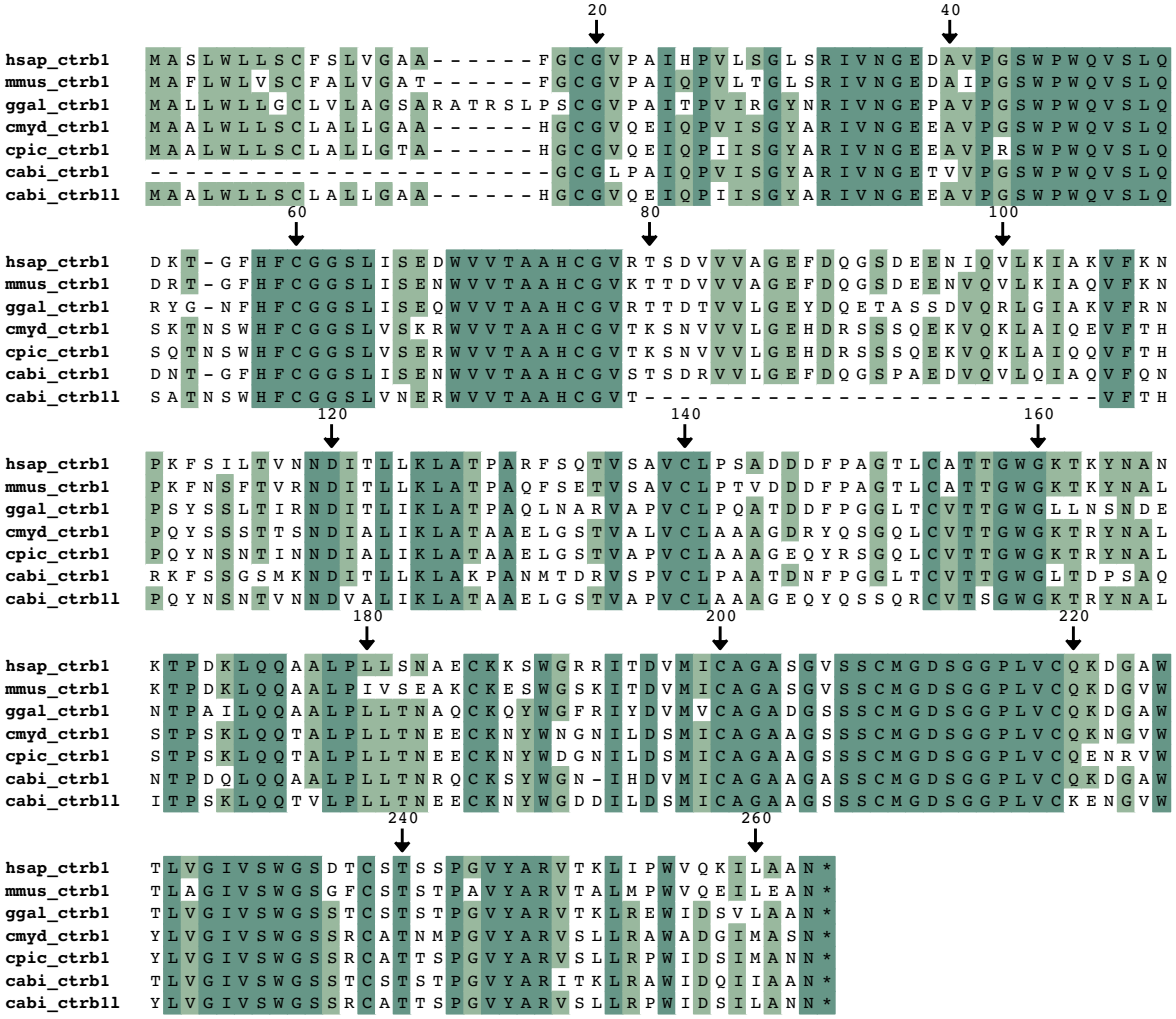

Fig. 4.9: Alignment of protease CTRB1 in testudines and other related organisms. Arrows show positions used as reference. `hsap`, *H. sapiens*; `mmus`, *M. musculus*; `ggal`, *G. gallus*; `cmyd`, *C. mydas*; `cpic`, *C. picta belli*; and `cabi`, *C. abingdonii*.

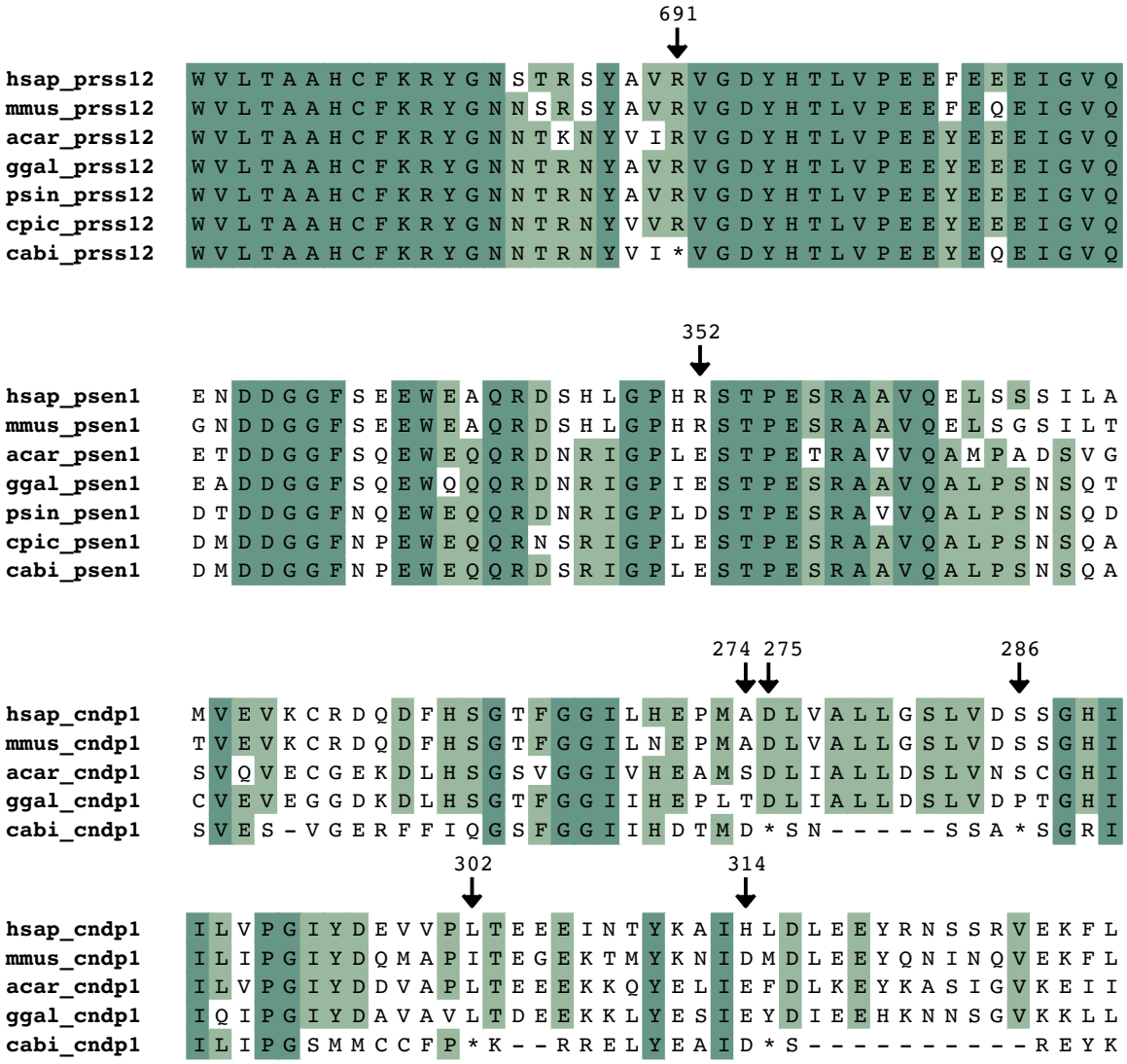

Fig. 4.10: Alignments of proteases involved in development in testudines and related organisms. Arrows mark positions of interest mentioned in the main text. `hsap`, *H. sapiens*; `mmus`, *M. musculus*; `acar`, *A. carolinensis*; `ggal`, *G. gallus*; `psin`, *P. sinensis*; `cpic`, *C. picta belli*; and `cabi`, *C. abingdonii*.

with several neurological dysfunctions, such as microcephaly or neurodevelopmental retardation [Yoon et al., 2012]. Similarly, *CNDP1* presents a truncating frameshift starting at residue p.27 (Figure 4.10) in all Galápagos tortoises, the Aldabra tortoise and their continental outgroups. In mammals, the loss of this protease causes carnosinemia, a recessive deficiency that leads to mental retardation, developmental delay, and various neurophaties [Bellia et al., 2014, Hu et al., 2007].

Dentition

Testudines lost the ability to grow teeth approximately 150-200 million years ago, becoming the oldest extant edentulous lineage of tetrapods [Davit-Béal et al., 2009]. Previous studies in birds and edentulous Mysticeti whales proved that tooth loss is closely associated with the pseudogenization and subsequent loss of tooth-specific genes, including proteases *KLK4* and *MMP20* [Keane et al., 2015, Meredith et al., 2014]. Western painted turtle's genome analysis indicated an homologous loss in the same tooth-specific genes [Bradley Shaffer et al., 2013]. As expected, we did not find any of these genes in *C. abingdonii* nor in *A. gigantea*, indicating their likely absence or pseudogenization in Testudines. These results confirm a pattern of multiple pseudogenizations associated with tooth loss, followed consistently and independently in multiple vertebrates.

Discussion

Ageing, the progressive deterioration of homoeostatic capabilities of the body, is a multifactor, quasi-universal, and poorly understood process. In an attempt at preliminarily characterising the evolutionary strategies to counteract ageing, we have annotated, analysed and comparatively studied the degradomes of sperm whales and the iconic tortoise Lonesome George. As a result of this, we have identified multiple genomic alterations affecting protease genes, some of which may be associated with ageing-related pathways or systems.

Overall, most of the predicted losses and gains of protease genes in the sperm whale mirror those described in the previously annotated genomes of minke [Yim et al., 2014] and bowhead whales [Keane et al., 2015]. Similarly, many of the findings in the Galápagos Giant tortoise are well conserved across Giant Tortoises and even across Archelosauria or Sauropsida. Nevertheless, several events stand out as independent or specific, providing interesting hypotheses about the evolutionary history of these species in the context of ageing as well as specific features of adaptation to the environment. As a *driver of evolution*, the immune system has been targeted by selection in many instances in both species, each in a different way, proving once more the important role that this system plays in the history of a species. We have found several instances of mutations related to four of the hallmarks of ageing, namely, *altered intercellular communication*, *stem cell exhaustion*, *derregulated nutrient sensing*, and *loss of proteostasis*. Altered intercellular communication, in particular, may play a role in explaining Peto's paradox in both species. Finally, some results may be related to specific adaptations to the distinct history and environment of the sperm whale and the Galápagos giant tortoise of Pinta Island.

When annotating a genome aiming to find differential features related to the biology

of an organism, a key step is to compare the results with those of other species, both closely related and outgroups. In the case of sperm whale (*P. macrocephalus*), the organisms of choice were: humans as an outgroup; bowhead and Minke whales; the bottlenose dolphin; and the killer whale. As they became available, genomic data from the narwhal and the beluga whale were included. Overall, most of the predicted losses and gains of protease genes mirror those described in the previously annotated genomes of minke [Yim et al., 2014] and bowhead whales [Keane et al., 2015]. This comparison served as a preliminary screening for our results. Nevertheless, several events stand out as independent or specific, providing interesting hypotheses about the evolution of sperm whales in the context of cetacean evolutionary history. In addition to the usual selective pressure on the immune and reproductive system of mammals, the unique aquatic environment of cetaceans has prompted numerous changes affecting protease genes involved in blood homoeostasis, skin maintenance, and digestion.

The strong selective pressure on the immune system has imposed multiple variations in mammalian proteases in several species [Keane et al., 2015, Puente et al., 2006, Worley et al., 2014]. Additionally, the underwater environment cetaceans live in poses challenges unique to this species of cetaceans, thus prompting novel immune strategies. Also, the immune system plays an important role in cancer development. This role may be relevant in massive mammalian organisms, given the much higher number of cells they possess compared to smaller animals. As stated in *Peto's paradox*, a similar propensity of each cell to become tumoural would lead to cancer incidences orders of magnitude higher in large mammals, which is not observed [Caulin and Maley, 2011].

Therefore, conspicuous genomic changes affecting cancer-related genes in large mammals might offer interesting candidates in ageing and cancer research. For instance, given the contribution of MMP12 to metastasis [Lv et al., 2015], the premature stop codon in cetacean orthologs may play a role in the biology of cancer in these organisms. Considering the level of conservation among the studied cetaceans of this premature stop codon, this could be an early adaptive event in cetaceans evolutionary history.

In addition, a putative loss of function in matrix metalloprotease 7 (*MMP7*), and a duplication affecting cystein protease *CASP3* may also merit further research. Since *MMP7*'s main role consist in degrading the extracellular matrix, which allows the cancerous cells to spread, this loss may impact the process of metastasis in sperm whales. *CASP3*, on the other hand, is known to be implicated in the proteolytic cascades associated with the apoptosis process, of great importance as an antitumoral measure.

As already mentioned, the putative loss of *MMP7* would mean the total absence of any member of the matrysilin subfamily in sperm whales, the only known case in

all mammals. Of course, the *MMP* family has a large number of members with partially overlapping functions, and therefore the loss of one protease can be overcome by the activity of other paralogs. However, mice deficient in *MMP7* have been shown to respond differently to challenges like re-epithelialization [Swee et al., 2008]. Interestingly, high levels of expression of this metalloprotease have also been shown to promote metastasis [Li et al., 2014, Koskensalo et al., 2011], which suggests a putative sperm whale-specific mechanism to counteract the problems underlined in Peto's paradox. Indeed, a duplication in a pro-autophagic protease, such as *CASP3*, could lead to an enhanced capability to enter an autophagic process, which would be a useful response mechanism against the development of tumour cells.

Adaptations related to inflammation, a process whose correct regulation is tightly linked to ageing, suggest that this response may be comparatively mild in cetaceans. Specifically, *CASP12* and *TPSAB1* apparent loss of function, both point to this mitigated reaction to pathogens via down-regulating the activity of inflammatory caspases, and altering mast-cells degranulation process, respectively [Abdelmotelb et al., 2014, McIlwain et al., 2015]. Interestingly, *CASP12* seems to have been lost in all cetaceans. Despite its complex evolutionary history, the frameshift starting at p.L101 is possibly the result of a single event, early in cetacean radiation, given its high degree of conservation in this clade, suggesting a possible secondary role in the adaptation to the aquatic habitat. In addition, the putative loss of function of *TPSAB1*, which also seems to have occurred via single, early event in the history of cetaceans, could have repercussions on the degranulation process occurring in mast cells during the first immunological response [Wilcock et al., 2019]. Finally, *MASP2* is specifically truncated in sperm whales, not only by gaining a premature stop codon, but by mimicking a deleterious point mutation that causes pathologies in humans. It is important to keep in mind that, according to Dobzhansky, which constitutes a pathological mutation to one species, can be a concerted, adaptive response in another [Dobzhansky, 1958]. This suggests that MASP2 may have been lost in a stepwise mechanism. First, a point mutation may have altered its function in concert with other events that rendered this change innocuous. Once the biology of this system was adapted to the loss of MASP2 activity, a second truncating event would have inactivated the gene. A different, independent truncation with the same expected result occurs in the bottlenose dolphin. Intriguingly, both truncated proteins are expected to contain lectin-binding domains, but not the serine-protease domain, which suggests a possible compensating mechanism through binding of a different protease to these domains. These events also provide a remarkable example of convergent evolution, and support the idea that loss of the

serine-protease domain of MASP2 is favoured in some cetaceans. When considered in terms of its function, these results could lead to less aggressive immune innate systems, with a diminished capacity to activate the complement path.

Taken together, these data reinforce the important role of the immune system, particularly the inflammatory response, in the evolution of cetaceans. This is highlighted by the independent losses of important protease genes participating in this system in sperm whales and other cetaceans, specially given the putative participation of *MASP2* mutations in speciation events through a complementation mechanism [Kondrashov et al., 2002]. The apparent general trend towards a milder inflammatory response may also be relevant in the study of the lower tumorigenic potential of cells in large mammals, and of course it could also yield some light into the longevity of this order.

One of the most conspicuous traits of the aquatic environment is the hydrostatic pressure and lack of net weight experienced by cetaceans. These conditions, so different from those encountered in land, must prompt compensatory mechanisms in the control of blood pressure and coagulation to avoid haemostatic accidents.

Related to this, we reported several losses that may impact blood homoeostasis. Specifically, two serine proteases, *F12* and *KLKB1*, that relate to the Kinin-Kallikreyn System, appear to have been lost in all cetaceans. Interestingly, in the case of *KLKB1*, it appears that there are several different stages of pseudogenization, including complete absences (such is the case of sperm whale), partial absences (with only two-four exons identified), and various numbers of premature stop codons in those presenting some exons. It's worth mentioning that the automatic annotation of the Monodontidae family members yielded no results for this gene. In all species presenting some exons, at least one stop codon is conserved. This suggest that despite the different stages, the process that leaded to the loss of this gene could have started at the same time in an early stage of cetacean speciation. As mentioned, both F12 and KLKB1 function as part of the KKS, a complex network of proteins and peptides involved in several biological processes, associated with exocrine glands and plasma. In these processes, they act as potent vasodilators, increase vascular permeability, produce pain, increase lymph flow, and (in high doses) cause the accumulation of polymorphonuclear leukocytes [Bader, 2011]. In fact, a genome association analysis with human populations has uncovered variants of these serine proteases putatively related to increased levels of vasoactive peptides [Verweij et al., 2013]. Therefore, their absence may be related to the extreme differences found in the aquatic versus terrestrial environment. Moreover, the KKS also impacts the acute phase of the inflammatory process. This suggests that the loss of *F12* and *KLKB1* may affect the inflammatory response along with the aforementioned

variants.

Another interesting aspect of blood homoeostasis, coagulation, seems to have undergone some alterations as well. In this sense, we reported the truncation of *TMPRSS11B* and *TMPRSS11F* (also known as *HATL5* and *HATL4* respectively) via different mechanisms. In addition, factor VII (*F7*), which also participates in coagulation and was reported as truncated in Mysticeti [Keane et al., 2015], appears to be functional in sperm whale. This adds another dimension to the convoluted ways in which blood homoeostasis has evolved to adapt to a new environment, thus reflecting the intrinsic complexity of this system. It must be noted that the sperm whale shows a natural disposition to much deeper dives than its relatives, which may constrict changes in the coagulation system.

Together, all these changes suggest that the mammalian potential for clotting and blood pressure are excessive in an aquatic environment, and these systems had to be modulated through processes that may have included the loss of proteases implicated in related proteolytic cascades.

Living underwater also sets the skin of mammals as a target of evolutionary pressure. Several events in the degradome of cetaceans might be related to this adaptive process. For instance, Hair-follicle cortex-related cystein protease, *CAPN12* seems to have been lost through several independent events in *Physeteroidea*, *Delphinoidea* (including *Monodontidae*, and *Mysticeti*. The annotated sequences are compatible with progressive gene losses, at different points of history. This case of convergent evolution suggests that the loss of this gene was favoured by selection. In addition, the loss of *KLK8*, specifically linked to the maintenance of the most external layer of the skin (*stratum corneum*). Possibly relevant as well is the apparent functionality of *KLK7*, lost in Mysticeti. The loss of the catalytic site in the case of *KLK8* may have more repercussions than simply a diminished function, since the putative inactive enzyme may yet be able to sequester substrates in the cell, playing a potentially antagonistic role to other physiological cell functions. Its independent loss in the rest of Cetacea suggests, again, that the suppression of its role may be important in the adaptation to the environment, to which the skin acts as first barrier and defence.

This complex pattern of convergent evolution suggests that skin-related proteases have played important roles in aquatic adaptation, in a process possibly influenced by the specific and somewhat contradictory requirements of heat insulation, buoyancy and diving. Interestingly, sperm whales presents several peculiarities in the skin that may be related to the reported differences in adaptation process. Not only does the sperm whale presents one of the relatively thinnest *stratum corneum*, but it also presents an

extremely wrinkled skin (sometimes referred as raisin-like) when compared to other Cetacea [Sokolov, 1982]. Besides this, sperm whales shed more frequently than its relatives and, as mentioned before, is one of the deepest divers among the aquatic mammals, only behind Cuvier's whale.

Regarding nutrition, different feeding strategies and diets may underlie similarly notable genetic adaptations. Thus, multiple metalloproteases have been lost at different points during cetacean evolution, probably due to their diverse diets (fish and bigger animals, versus krill). In short, *CPA2*, *CPA3*, *CPO*, and *CPB1*, take part in the digestive process, severing the peptidic link as part of the digestion of proteins. As such, the specific loss of *CPB1* in sperm whales and the diverse premature stop codons of *CPA3* in other cetaceans could be related to the peculiar diet of this enormous creature, the only known predator of giant squids. Unsurprisingly, when compared with the annotation degradome of Mysticeti, we found that those proteases linked to the dentition process and enamel maintenance, *KLK4* and *MMP20*, were apparently functional. Hence, as expected, we could validate that the phenotypic differences that most clearly divide Cetacea, the presence or absence of teeth, is likely to have a distinct genetic base in the presence or absence of these proteases.

These results suggest that protease gene losses have been important in the evolution of the digestive system of cetaceans. At least in some cases, the genetic causes for these losses have been independent even between Odontoceti. This cases of convergent evolution suggest that those events were highly favoured at the trophic level where cetaceans thrive.

Hence, the manual annotation of the degradome of sperm whales yields some insight into some of the biological peculiarities of this fascinating organism. We have set forth hypotheses on the genetic basis of the mechanics underlining the immune response of *P. macrocephalus*, reinforcing the key role of this system in the evolution of Metazoa. In turn, this may offer information on the impact of the aquatic environment in mammalian immunological systems, and the impact of inflammation on the organism. We have also reported several events that seem to be pivotal in the evolutionary history of this order, as suggested by convergent evolution acting on genes related to blood or skin homoeostasis. Finally, we noted some hypothesis that would be related to Peto's paradox in these enormous animals.

Lonesome George was the iconic last member of *C. abingdonii*. Like other giant tortoises, George lived a long life. Therefore, its genome is expected to hold clues to a different and independent solution to the problems associated with ageing. With this hypothesis, we undertook the sequencing and annotation of this genome.

From the automatic annotation of `CheloAbing 1.0`, we found twelve expanded gene families, eight of which belong to the "extracellular exosome" GO category. This suggests that *C. abingdonii* this family has been subjected to selective pressure during the evolutionary history of this species. The correct activity of this pathway, directly impacts intercellular communication, one of the Hallmarks of ageing, in which exosomes play a crucial role. As such, exosomes take part in many biological processes and signalling pathways related to immunity, cancer, and ageing [Baixauli et al., 2014,Becker et al., 2016,Prattichizzo et al., 2017]. In addition, this expansion be related to gigantism (and its associated cancer protection, as predicted by Peto's paradox) in giant tortoises.

Consistent with this, an additional analysis suggested that several genes involved in ERAD, *BAG2*(*NEF*) and *UBE2J1* (*Ubc6/7*) may have been subjected to positive selection. Not only this strengthened the results from previous analysis by highlighting the apparent importance of these related pathways in giant tortoises, but one of the expanded genes, *VCP*, is also a central part of this route. In this sense, ERAD is important in the unfolded-protein response, and consequently, in the correct proteostasis of the cell, whose deregulation constitutes a hallmark of ageing [López-Otín et al., 2013, Scheper and Hoozemans, 2015].

Interestingly, two genes showing putative positive selection (*TUBE1* and *TBG1*) are related to tubulin assembly during cell cycle progression [Chinen et al., 2015]. This suggests that certain alterations affecting these genes might be related to the increase in the number of cellular divisions associated to gigantism in tortoises. Taken together, these results suggest altered inner- and intercellular communication strategies underlying the biology of giant tortoises.

Amongst the targets of selection we found two genes previously linked to successful ageing in humans. Specifically, expression levels of *AHSG* and *FGF19* have been linked to successful ageing in humans [Sanchis-Gomar et al., 2015]. The proteins encoded by these genes are involved in glucose and lipids metabolism, meaning that their alteration could impact the regulation of nutrient sensing, one of the secondary hallmarks of ageing. Finally, this analysis also singled out three genes, *MVK*, *IRAK1BP1* and *IL1R2*, all of them with important roles in the modulation of the immune system. In this regard, it is important to notice that, in addition to its role in proteostasis, ERAD is a target of viral infection, as multiple viruses depend on this process for successful delivery to the cytoplasm [Morito and Nagata, 2015].

Taken together, this hypothesis-free analysis highlights proteostasis, metabolism regulation and immune response as key processes during the evolution of giant tortoises, and provide starting points for future work on this subject.

On the other hand, manual annotation of Lonesome George's degradome uncovered several point mutations, truncations, and copy-number-variations, specific of George, Giant tortoises or testudines, that may offer interesting information. From these results, it seems that most reptiles and specifically giant tortoises have drifted towards a situation in which the innate immune response outweighs the adaptive one. Although several immune mediators can have dual functions both in innate and adaptive immune responses, it is thought that the innate branch of the immune system in vertebrates evolved earlier than the adaptive route [Zimmerman et al., 2010]. All multicellular organisms have some form of innate immune response, which acts as an initial step in the defence against pathogens. Among vertebrates, Reptilia are the only ectothermic amniotes, and therefore the study of their immune system could provide new important insights into its evolution under different circumstances.

On this topic, the truncation of *MEP1A*, a protease responsible for the maturation of B-lymphocytes, is expected to impact the immune response based on specific antibodies. In addition, we have found CNVs affecting granzyme serine proteases, a set of enzymes linked to the citotoxicity mediated by Natural Killers, one of the cellular types tasked with the innate immunological response [Voskoboinik et al., 2015]. These alterations are exclusive of the Galápagos giant tortoises in the case of *MEP1A*, and shared only with *A. gigantea* in the case of the granzyme expansion, probably playing an important role in its adaptations to size or longevity. Other analysed species showed an expansion in the granzyme family but not so extensive as the one present in giant tortoises.

Taking all of this into account, an unbalance between the two types of defence is apparent. Of course, we should consider that this does not mean a total abandonment of the adaptive system.

Regarding proteolytical systems regulating blood homoeostasis, several proteases appear to be truncated, including *F7*, *F10*, and *F11*. Truncations affecting these proteases usually cause Factors *VII*, *X* and *XI* deficiencies in humans, although, which points to a Dobzhansky anomaly in Galápagos tortoises. Interestingly, due to several point mutations in catalytic sites (even if not specific to giant tortoises), the function of PLG seems possibly lost as well. Since the main function of its product is to aid in the dissolution of blood clots [Wu et al., 2019], its truncation adds weight to our hypothesis that blood homoeostasis systems have been under selective pressure in turtles. Finally, *PROC*, which inhibits the generation of plasmin, is apparently absent. The absence of this gene may contribute to severe the consequences of the other alterations in the coagulation system in turtles, since its deficiency is often link to thrombosis diseases in humans [Cheng et al., 2016].

On the subject of metabolism and diet, giant tortoises seem to have (expectedly) lost the proteases associated with dentition (*KLK4* and *MMP20*) as the rest of the clade, besides this, we report some interesting alterations linked to the hallmark of deregulated nutrient sensing. Specifically, alterations relate with glucose tolerance and intake.

First, we found that neurolysin (*NLN*) presents a premature stop codon, quite early in the sequence, that would abolish the function of the protein. Interestingly, while Aldabra giant tortoise does not present a similar truncation, Mohave's dessert tortoise (*G. agassizii*) does, suggesting a case of convergent evolution. A second alteration, a duplication in pancreatic serine protease *CTRB1* is expected to impact the process of digestion.

The development of the nervous system is a complex and intricate process in which a lot of different genes and pathways take part. Because of this complexity, it is a metabolically expensive system to invest in, from an evolutionary point of view. For this reason, nervous system development and derivative capabilities, are part of a trade-off. In this regard, George's exclusive truncation of *PRSS12*, giant tortoise truncations of *XPNPEP1* and *CNDP1*, and point variant affecting *PSEN1*, could be related to an underdevelopment of the neural and cognitives functions when compared with other related animals [Bellia et al., 2014, Jiang et al., 2015, Mitsui et al., 2013, Yoon et al., 2012].

In summary, the manual annotation of the degradome of sperm whales has yielded some insight on the biological peculiarities of this fascinating organism. We have set forth hypotheses on the genetic basis of the mechanics underlining the immune response of *P. macrocephalus*, reinforcing the key role of this system in the evolution of Metazoa. In turn, this may offer information on the impact of the aquatic environment in mammalian immunological systems, and the impact of inflammation on the organism. We have also reported several events that seem to be pivotal in the evolutionary history of this order, as suggested by convergent evolution acting on genes related to blood or skin homoeostasis. Finally, we noted some results that might be related to the Peto paradox in these enormous animals. In addition, the automatic and manual annotation of the genome of Lonesome George has provided information that may help in understanding the genomic basis of particular features of giant tortoises. Considering alterations to immune-related genes, these species seem to show a different balance between innate and adaptive responses compared to mammals. Of course, this does not mean a total abandonment of the adaptive system over the innate, but rather a more important role of the innate response. Altogether, this analysis highlights proteostasis, metabolism

regulation, cell division, and immune response as key and potentially age-related processes during the evolution of giant tortoises, and provide starting points for future work on these subjects.

Therefore, the present Thesis investigates the results of the manual and automatic annotation of *P. macrocephalus* and *C. abingdoni* degradomes. By searching for links between these degradomes and the hallmarks of ageing, we point out multiple genes and pathways affected by natural selection, such as *loss of proteostasis*, *altered intercellular communication*, and *deregulated nutrient sensing*. Hopefully, by unravelling the evolutionary history of these hallmarks in long-lived metazoans, we inch closer to greater understanding of the ageing process, and to make P. Medawar's reference to ageing as "*An unsolved problem in biology*" obsolete.

Conclusions

1. The complete degradome of the sperm whale, *Physeter macrocephalus*, contains 546 protease genes, including 58 classified as pseudogenes.

2. We found multiple events of pseudogenization and truncation in sperm whales affecting proteases, including *MMP12*, *TPSAB1*, and *MASP2*, with known roles in the immune system.

3. We report nonsense variants in genes linked to cancer development in *P. macrocephalus*, such as *CASP3* and *MMP7*, that are expected to eliminate their products.

4. We have assembled the genome of Lonesome George, last member of *Chelonoidis abingdonii*, one of the Galápagos giant tortoises. The automatic annotation of this assembly yielded 27,208 predicted genes.

5. The complete degradome of Lonesome George is composed of 515 genes, 63 of which are classified as pseudogenes.

6. Through hypothesis-driven manual annotation of the degradome of Lonesome George, we have found multiple specific variants predicted to affect the activity of proteases with known roles in *altered intercellular communication*, an integrative hallmark of ageing.

Conclusiones

1. El degradoma completo del cachalote, *Physeter macrocephalus*, incluye 546 genes de proteasas, 58 de ellos clasificados como pseudogenes.

2. Encontramos diversos eventos de pseudogenización y truncantes que afectan a proteasas, incluyendo *MMP12*, *TPSAB1*, y *MASP2*, con funciones conocidas en el sistema inmune.

3. Reportamos variantes truncantes en genes vinculados con el desarrollo del cáncer en *P. macrocephalus*, como *CASP3* y *MMP7*, que se espera que eliminen el producto génico.

4. Hemos ensamblado el genoma de Solitario George, el último miembro de la especie *Chelonoidis abingdonii*, una de las especies de tortugas gigantes de las Islas Galápagos. La anotación automática de este ensamblado predijo la existencia de 27.208 genes.

5. El degradoma completo de Solitario George se compone de 515 genes, incluyendo 63 que han sido clasificados como pseudogenes.

6. Mediante una anotación manual y dirigida por hipótesis, del genoma de George, hemos encontrado múltiples variantes que predecimos que afecten a la actividad de proteasas con funciones conocidas en la *alteración de las comunicaciones intercelulares*, una de las "marcas distintivas" integrativas del envejecimiento.

Bibliography

[Abdelmotelb et al., 2014] Abdelmotelb, A. M., Rose-Zerilli, M. J., Barton, S. J., Holgate, S. T., Walls, A. F., and Holloway, J. W. (2014). Alpha-tryptase gene variation is associated with levels of circulating IgE and lung function in asthma. *Clinical and Experimental Allergy*, 44(6):822–830.

[Agathangelou et al., 2018] Agathangelou, K., Apostolou, Z., and Garinis, G. A. (2018). Nuclear DNA damage and ageing. In *Subcellular Biochemistry*, volume 90, pages 309–322. Springer New York.

[Agnarsson and May-Collado, 2008] Agnarsson, I. and May-Collado, L. J. (2008). The phylogeny of Cetartiodactyla: The importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. *Molecular Phylogenetics and Evolution*, 48(3):964–985.

[Akula et al., 2015] Akula, S., Thorpe, M., Boinapally, V., and Hellman, L. (2015). Granule associated serine proteases of hematopoietic cells-an analysis of their appearance and diversification during vertebrate evolution. *PLoS ONE*, 10(11):1–26.

[Al-Hilali et al., 2007] Al-Hilali, A., Wulff, K., Abdel-Razeq, H., Saud, K. A., Al-Gaili, F., and Herrmann, F. H. (2007). Analysis of the novel factor X gene mutation Glu51Lys in two families with factor X-Riyadh anomaly. *Thrombosis and haemostasis*, 97(4):542–5.

[Alam et al., 2016] Alam, P., Amini, S., Tadayon, M., Miserez, A., and Chinsamy, A. (2016). Properties and architecture of the sperm whale skull amphitheatre. *Zoology*, 119(1):42–51.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[Arnason et al., 2007] Arnason, U., Gullberg, A., Janke, A., and Kullberg, M. (2007). Mitogenomic analyses of caniform relationships. *Molecular Phylogenetics and Evolution*, 45(3):863–874.

[Bader, 2011] Bader, M. (2011). *Kinins*. Walter de Gruyter GmbH and Co. KG.

[Baixauli et al., 2014] Baixauli, F., López-Otín, C., and Mittelbrunn, M. (2014). Exosomes and autophagy: Coordinated mechanisms for the maintenance of cellular fitness. *Frontiers in Immunology*, 5(AUG):1–7.

[Bateman, 2019] Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515.

[Becker et al., 2016] Becker, A., Thakur, B. K., Weiss, J. M., Kim, H. S., Peinado, H., and Lyden, D. (2016). Extracellular Vesicles in Cancer: Cell-to-Cell Mediators of Metastasis. *Cancer Cell*, 30(6):836–848.

[Bellia et al., 2014] Bellia, F., Vecchio, G., and Rizzarelli, E. (2014). Carnosinases, their substrates and diseases. *Molecules*, 19(2):2299–2329.

[Benton, 2014] Benton, M. J. M. J. (2014). *Vertebrate palaeontology*.

[Best, 1979] Best, P. B. (1979). Social Organization in Sperm Whales, Physeter macrocephalus. In *Behavior of Marine Animals*, pages 227–289. Springer US.

[Bocheva et al., 2019] Bocheva, G., Slominski, R. M., and Slominski, A. T. (2019). Neuroendocrine aspects of skin aging. *International Journal of Molecular Sciences*, 20(11):1–19.

[Boetzer and Pirovano, 2014] Boetzer, M. and Pirovano, W. (2014). SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, 15(1):1–9.

[Bradley Shaffer et al., 2013] Bradley Shaffer, H., Minx, P., Warren, D. E., Shedlock, A. M., Thomson, R. C., Valenzuela, N., Abramyan, J., Amemiya, C. T., Badenhorst, D., Biggar, K. K., Borchert, G. M., Botka, C. W., Bowden, R. M., Braun, E. L.,

Bronikowski, A. M., Bruneau, B. G., Buck, L. T., Capel, B., Castoe, T. A., Czerwinski, M., Delehaunty, K. D., Edwards, S. V., Fronick, C. C., Fujita, M. K., Fulton, L., Graves, T. A., Green, R. E., Haerty, W., Hariharan, R., Hernandez, O., Hillier, L. D. W., Holloway, A. K., Janes, D., Janzen, F. J., Kandoth, C., Kong, L., de Koning, A. P., Li, Y., Literman, R., McGaugh, S. E., Mork, L., O'Laughlin, M., Paitz, R. T., Pollock, D. D., Ponting, C. P., Radhakrishnan, S., Raney, B. J., Richman, J. M., St John, J., Schwartz, T., Sethuraman, A., Spinks, P. Q., Storey, K. B., Thane, N., Vinar, T., Zimmerman, L. M., Warren, W. C., Mardis, E. R., and Wilson, R. K. (2013). The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology*, 14(3).

[Breda et al., 2016] Breda, C., Sathyasaikumar, K. V., Idrissi, S. S., Notarangelo, F. M., Estranero, J. G., Moore, G. G., Green, E. W., Kyriacou, C. P., Schwarcz, R., and Giorgini, F. (2016). Tryptophan-2,3-dioxygenase (TDO) inhibition ameliorates neurodegeneration by modulation of kynurenine pathway metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, 113(19):5435–5440.

[Butler et al., 2008] Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research*, 18(5):810–820.

[Caccone et al., 1999] Caccone, A., Gibbs, J. P., Ketmaier, V., Suatoni, E., and Powell, J. R. (1999). Origin and evolutionary relationships of giant Galapagos tortoises. *Proceedings of the National Academy of Sciences of the United States of America*, 96(23):13223–13228.

[Calcinotto et al., 2019] Calcinotto, A., Kohli, J., Zagato, E., Pellegrini, L., Demaria, M., and Alimonti, A. (2019). Cellular senescence: Aging, cancer, and injury. *Physiological Reviews*, 99(2):1047–1078.

[Campbell et al., 2014] Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). *Genome Annotation and Curation Using MAKER and MAKER-P*, volume 2014.

[Campisi and D'Adda Di Fagagna, 2007] Campisi, J. and D'Adda Di Fagagna, F. (2007). Cellular senescence: When bad things happen to good cells. *Nature Reviews Molecular Cell Biology*, 8(9):729–740.

[Cannon, 1934] Cannon, W. B. (1934). Physiological Balance in the Body. *Nature*, 133(3351):82.

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

[Chakkalakal et al., 2012] Chakkalakal, J. V., Jones, K. M., Basson, M. A., and Brack, A. S. (2012). The aged niche disrupts muscle stem cell quiescence. *Nature*, 490(7420):355–360.

[Cheng et al., 2016] Cheng, X., Wang, M., Jiang, M., Bhugul, P. A., Hao, X., and Yang, L. (2016). A protein C and plasminogen compound heterozygous mutation and a compound heterozygote of protein C in two related Chinese families. *Blood Coagulation and Fibrinolysis*, 27(7):838–844.

[Chinen et al., 2015] Chinen, T., Liu, P., Shioda, S., Pagel, J., Cerikan, B., Lin, T. C., Gruss, O., Hayashi, Y., Takeno, H., Shima, T., Okada, Y., Hayakawa, I., Hayashi, Y., Kigoshi, H., Usui, T., and Schiebel, E. (2015). The γ-tubulin-specific inhibitor gatastatin reveals temporal requirements of microtubule nucleation during the cell cycle. *Nature Communications*, 6:1–11.

[Collado et al., 2007] Collado, M., Blasco, M. A., and Serrano, M. (2007). Cellular Senescence in Cancer and Aging. *Cell*, 130(2):223–233.

[Darwin, 1845] Darwin, C. (1845). *Journal of Researches into the Natural History and Geology of the Countries Visited during the Voyage of HMS Beagle round the World, under the Command of Capt. Fitz Roy, R.N.* Cambridge University Press.

[Davit-Béal et al., 2009] Davit-Béal, T., Tucker, A. S., and Sire, J. Y. (2009). Loss of teeth and enamel in tetrapods: Fossil record, genetic data and morphological adaptations. *Journal of Anatomy*, 214(4):477–501.

[De Haan and Lazare, 2018] De Haan, G. and Lazare, S. S. (2018). Aging of hematopoietic stem cells. *Blood*, 131(5):479–487.

[Dear et al., 2000] Dear, T. N., Meier, N. T., Hunn, M., and Boehm, T. (2000). Gene structure, chromosomal localization, and expression pattern of Capn12, a new member of the calpain large subunit gene family. *Genomics*, 68(2):152–160.

[Dimri et al., 1995] Dimri, G. P., Lee, X., Basile, G., Acosta, M., Scott, G., Roskelley, C., Medrano, E. E., Linskens, M., Rubelj, I., Pereira-Smith, O., Peacocke, M., and Campisi, J. (1995). A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 92(20):9363–9367.

[Dobzhansky, 1958] Dobzhansky, T. (1958). Genetics of Homeostasis and Senility. *Annals of the New York Academy of Sciences*, 71(6):1234–1242.

[Eddy, 2011] Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10).

[English et al., 2012] English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. (2012). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE*, 7(11):1–12.

[Eto et al., 2011] Eto, D., Lao, C., DiToro, D., Barnett, B., Escobar, T. C., Kageyama, R., Yusuf, I., and Crotty, S. (2011). IL-21 and IL-6 are critical for different aspects of B cell immunity and redundantly induce optimal follicular helper CD4 T cell (Tfh) differentiation. *PLoS ONE*, 6(3).

[Fais et al., 2016] Fais, A., Johnson, M., Wilson, M., Aguilar Soto, N., and Madsen, P. T. (2016). Sperm whale predator-prey interactions involve chasing and buzzing, but no acoustic stunning. *Scientific Reports*, 6(June):1–13.

[Ferreira et al., 2018] Ferreira, G. S., Bronzati, M., Langer, M. C., and Sterli, J. (2018). Phylogeny, biogeography and diversification patterns of side-necked turtles (Testudines : Pleurodira). *Royal Society Open Science*, 5.

[Fontana et al., 2010] Fontana, L., Partridge, L., and Longo, V. (2010). Extending healthy life span-from yeast to humans. *Science*, 328(5976):321–326.

[Fordyce and Marx, 2018] Fordyce, R. E. and Marx, F. G. (2018). Gigantism Precedes Filter Feeding in Baleen Whale Evolution. *Current Biology*, 28(10):1670–1676.e2.

[Franceschi and Campisi, 2014] Franceschi, C. and Campisi, J. (2014). Chronic inflammation (Inflammaging) and its potential contribution to age-associated diseases. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 69:S4–S9.

[Franceschi et al., 2017] Franceschi, C., Garagnani, P., Vitale, G., Capri, M., and Salvioli, S. (2017). Inflammaging and 'Garb-aging'. *Trends in Endocrinology and Metabolism*, 28(3):199–212.

[Godard et al., 2003] Godard, C., Clark, R., Kerr, I., Madsen, P. T., and Payne, R. (2003). Preliminary report on the sperm whale data collected during the voyage of the Odyssey. (January 2015):1–9.

[Green et al., 2011] Green, D. R., Galluzzi, L., and Kroemer, G. (2011). Mitochondria and the autophagy-inflammation-cell death axis in organismal aging. *Science*, 333(6046):1109–1112.

[Grindel et al., 2018] Grindel, B. J., Martinez, J. R., Tellman, T. V., Harrington, D. A., Zafar, H., Nakhleh, L., Chung, L. W., and Farach-Carson, M. C. (2018). Matrilysin/MMP-7 Cleavage of Perlecan/HSPG2 Complexed with Semaphorin 3A Supports FAK-Mediated Stromal Invasion by Prostate Cancer Cells. *Scientific Reports*, 8(1):1–14.

[Haldane, 1933] Haldane, J. B. S. (1933). *Causes of evolution.pdf*. Longmans Green & Co.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation.

[Harman, 1965] Harman, D. (1965). The free radical theory of ageing: Effect of age on serum copper levels. *Journal of gerontology*, 20:151–153.

[Harrison et al., 2009] Harrison, D. E., Strong, R., Sharp, Z. D., Nelson, J. F., Astle, C. M., Flurkey, K., Nadon, N. L., Wilkinson, J. E., Frenkel, K., Carter, C. S., Pahor, M., Javors, M. A., Fernandez, E., and Miller, R. A. (2009). Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*, 460(7253):392–395.

[Hartl et al., 2011] Hartl, F. U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. *Nature*, 475(7356):324–332.

[Hipp et al., 2019] Hipp, M. S., Kasturi, P., and Hartl, F. U. (2019). The proteostasis network and its decline in ageing.

[Hoeijmakers, 2009] Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *New England Journal of Medicine*, 361(15):1475–1485.

[Hoenicke and Zender, 2012] Hoenicke, L. and Zender, L. (2012). Immune surveillance of senescent cells-biological significance in cancer-and non-cancer pathologies. *Carcinogenesis*, 33(6):1123–1126.

[Hu et al., 2007] Hu, Y., Hosseini, A., Kauwe, J. S., Gross, J., Cairns, N. J., Goate, A. M., Fagan, A. M., Townsend, R. R., and Holtzman, D. M. (2007). Identification and validation of novel CSF biomarkers for early stages of Alzheimer's disease. *Proteomics - Clinical Applications*, 1(11):1373–1384.

[Irmscher et al., 2018] Irmscher, S., Döring, N., Halder, L. D., Jo, E. A., Kopka, I., Dunker, C., Jacobsen, I. D., Luo, S., Slevogt, H., Lorkowski, S., Beyersdorf, N., Zipfel, P. F., and Skerka, C. (2018). Kallikrein Cleaves C3 and Activates Complement. *Journal of Innate Immunity*, 10(2):94–105.

[Ishii et al., 2014] Ishii, T., Abboud, R. T., Wallace, A. M., English, J. C., Coxson, H. O., Finley, R. J., Shumansky, K., Paré, P. D., and Sandford, A. J. (2014). Alveolar macrophage proteinase/antiproteinase expression in lung function and emphysema. *European Respiratory Journal*, 43(1):82–91.

[Janzen et al., 2006] Janzen, V., Forkert, R., Fleming, H. E., Saito, Y., Waring, M. T., Dombkowski, D. M., Cheng, T., DePinho, R. A., Sharpless, N. E., and Scadden, D. T. (2006). Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16 INK4a. *Nature*, 443(7110):421–426.

[Jiang et al., 2015] Jiang, H. Y., Li, G. D., Dai, S. X., Bi, R., Zhang, D. F., Li, Z. F., Xu, X. F., Zhou, T. C., Yu, L., and Yao, Y. G. (2015). Identification of PSEN1 mutations p.M233L and p.R352C in Han Chinese families with early-onset familial Alzheimer's disease. *Neurobiology of Aging*, 36(3):1602.e3–1602.e6.

[Kang et al., 2011] Kang, T. W., Yevsa, T., Woller, N., Hoenicke, L., Wuestefeld, T., Dauch, D., Hohmeyer, A., Gereke, M., Rudalska, R., Potapova, A., Iken, M., Vucur, M., Weiss, S., Heikenwalder, M., Khan, S., Gil, J., Bruder, D., Manns, M., Schirmacher, P., Tacke, F., Ott, M., Luedde, T., Longerich, T., Kubicka, S., and Zender, L. (2011). Senescence surveillance of pre-malignant hepatocytes limits liver cancer development. *Nature*, 479(7374):547–551.

[Keane et al., 2015] Keane, M., Semeiks, J., Webb, A. E. E., Li, Y. I. I., Quesada, V., Craig, T., Madsen, L. B. B., van Dam, S., Brawand, D., Marques, P. I. I., Michalak,

P., Kang, L., Bhak, J., Yim, H.-S. S., Grishin, N. V. V., Nielsen, N. H. H., Heide-Jørgensen, M. P. P., Oziolor, E. M. M., Matson, C. W. W., Church, G. M. M., Stuart, G. W. W., Patton, J. C. C., George, J. C. C., Suydam, R., Larsen, K., López-Otín, C., O'Connell, M. J., Bickham, J. W. W., Thomsen, B., DeMagalhães, J. P., van Dam, S., Brawand, D., Marques, P. I. I., Michalak, P., Kang, L., Bhak, J., Yim, H.-S. S., Grishin, N. V. V., Nielsen, N. H. H., Heide-Jørgensen, M. P. P., Oziolor, E. M. M., Matson, C. W. W., Church, G. M. M., Stuart, G. W. W., Patton, J. C. C., George, J. C. C., Suydam, R., Larsen, K., López-Otín, C., O'Connell, M. J., Bickham, J. W. W., Thomsen, B., and de Magalhães, J. P. (2015). Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports*, 10(1):112–122.

[Khan et al., 2007] Khan, S. Q., Dhillon, O., Struck, J., Quinn, P., Morgenthaler, N. G., Squire, I. B., Davies, J. E., Bergmann, A., and Ng, L. L. (2007). C-terminal pro-endothelin-1 offers additional prognostic information in patients after acute myocardial infarction. Leicester Acute Myocardial Infarction Peptide (LAMP) Study. *American Heart Journal*, 154(4):736–742.

[Kir et al., 2011] Kir, S., Beddow, S. A., Samuel, V. T., Miller, P., Previs, S. F., Suino-Powell, K., Xu, H. E., Shulman, G. I., Kliewer, S. A., and Mangelsdorf, D. J. (2011). FGF19 as a postprandial, insulin-independent activator of hepatic protein and glycogen synthesis. *Science*, 331(6024):1621–1624.

[Kirkwood, 2005] Kirkwood, T. B. (2005). Understanding the odd science of aging. *Cell*, 120(4):437–447.

[Kishibe et al., 2007] Kishibe, M., Bando, Y., Terayama, R., Namikawa, K., Takahashi, H., Hashimoto, Y., Ishida-Yamamoto, A., Jiang, Y. P., Mitrovic, B., Perez, D., Iizuka, H., and Yoshida, S. (2007). Kallikrein 8 is involved in skin desquamation in cooperation with other kallikreins. *Journal of Biological Chemistry*, 282(8):5834–5841.

[Klaips et al., 2018] Klaips, C. L., Jayaraj, G. G., and Hartl, F. U. (2018). Pathways of cellular proteostasis in aging and disease.

[Koga et al., 2011] Koga, H., Kaushik, S., and Cuervo, A. M. (2011). Protein homeostasis and aging: The importance of exquisite quality control. *Ageing Research Reviews*, 10(2):205–215.

[Kondrashov et al., 2002] Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23):14878–14883.

[Koskensalo et al., 2011] Koskensalo, S., Louhimo, J., Nordling, S., Hagström, J., and Haglund, C. (2011). MMP-7 as a prognostic marker in colorectal cancer. *Tumor Biology*, 32(2):259–264.

[Kuilman et al., 2010] Kuilman, T., Michaloglou, C., Mooi, W. J., and Peeper, D. S. (2010). The essence of senescence. *Genes and Development*, 24(22):2463–2479.

[Kumar et al., 2017] Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular biology and evolution*, 34(7):1812–1819.

[Kuwae et al., 2002] Kuwae, K., Matsumoto-Miyai, K., Yoshida, S., Sadayama, T., Yoshikawa, K., Hosokawa, K., and Shiosaka, S. (2002). Epidermal expression of serine protease, neuropsin (KLK8) in normal and pathological skin samples. *Journal of Clinical Pathology - Molecular Pathology*, 55(4):235–241.

[Laplante and Sabatini, 2012] Laplante, M. and Sabatini, D. M. (2012). MTOR signaling in growth control and disease. *Cell*, 149(2):274–293.

[Le et al., 2006] Le, M., Raxworthy, C. J., McCord, W. P., and Mertz, L. (2006). A molecular phylogeny of tortoises (Testudines: Testudinidae) based on mitochondrial and nuclear genes. *Molecular Phylogenetics and Evolution*, 40(2):517–531.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li et al., 2014] Li, Z., Zhang, D., Zhang, H., Miao, Z., Tang, Y., Sun, G., and Dai, D. (2014). Prediction of peritoneal recurrence by the mRNA level of CEA and MMP-7 in peritoneal lavage of gastric cancer patients. *Tumor Biology*, 35(4):3463–3470.

[López-Otín et al., 2013] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194.

[López-Otín and Bond, 2008] López-Otín, C. and Bond, J. S. (2008). Proteases: Multifunctional enzymes in life and disease.

[López-Otín and Overall, 2002] López-Otín, C. and Overall, C. M. (2002). Protease degradomics: A new challenge for proteomics. *Nature Reviews Molecular Cell Biology*, 3(7):509–519.

[Lord and Ashworth, 2012] Lord, C. J. and Ashworth, A. (2012). The DNA damage response and cancer therapy.

[Lv et al., 2015] Lv, F. Z., Wang, J. L., Wu, Y., Chen, H. F., and Shen, X. Y. (2015). Knockdown of MMP12 inhibits the growth and invasion of lung adenocarcinoma cells. *International Journal of Immunopathology and Pharmacology*, 28(1):77–84.

[Mancia, 2018] Mancia, A. (2018). On the revolution of cetacean evolution. *Marine Genomics*, 41(August):1–5.

[Mannen and Li, 1999] Mannen, H. and Li, S. S. (1999). Molecular Evidence for a Clade of Turtles. *Molecular Phylogenetics and Evolution*, 13(1):144–148.

[Mattison et al., 2017] Mattison, J. A., Colman, R. J., Beasley, T. M., Allison, D. B., Kemnitz, J. W., Roth, G. S., Ingram, D. K., Weindruch, R., De Cabo, R., and Anderson, R. M. (2017). Caloric restriction improves health and survival of rhesus monkeys. *Nature Communications*, 8(1):1–12.

[McIlwain et al., 2015] McIlwain, D. R., Berger, T., and Mak, T. W. (2015). Caspase functions in cell death and disease. *Cold Spring Harbor Perspectives in Biology*, 7(4).

[Meredith et al., 2014] Meredith, R. W., Zhang, G., Gilbert, M. T. P., Jarvis, E. D., and Springer, M. S. (2014). Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science*, 346(6215).

[Mitsui et al., 2013] Mitsui, S., Hidaka, C., Furihata, M., Osako, Y., and Yuri, K. (2013). A mental retardation gene, motopsin/prss12, modulates cell morphology by interaction with seizure-related gene 6. *Biochemical and Biophysical Research Communications*, 436(4):638–644.

[Mizushima et al., 2008] Mizushima, N., Levine, B., Cuervo, A. M., and Klionsky, D. J. (2008). Autophagy fights disease through cellular self-digestion. *Nature*, 451(7182):1069–1075.

[Modesto and Anderson, 2004] Modesto, S. P. and Anderson, J. S. (2004). The phylogenetic definition of reptilia. *Systematic Biology*, 53(5):815–821.

[Møhl, 2001] Møhl, B. (2001). Sound transmission in the nose of the sperm whale Physeter catodon. A post mortem study. *Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology*, 187(5):335–340.

[Morgulis et al., 2006] Morgulis, A., Gertz, E. M., Schäffer, A. A., and Agarwala, R. (2006). WindowMasker: Window-based masker for sequenced genomes. *Bioinformatics*, 22(2):134–141.

[Morito and Nagata, 2015] Morito, D. and Nagata, K. (2015). Pathogenic Hijacking of ER-Associated Degradation: Is ERAD Flexible? *Molecular Cell*, 59(3):335–344.

[Moskalev et al., 2013] Moskalev, A. A., Shaposhnikov, M. V., Plyusnina, E. N., Zhavoronkov, A., Budovsky, A., Yanai, H., and Fraifeld, V. E. (2013). The role of DNA damage and repair in aging through the prism of Koch-like criteria.

[Osorio et al., 2010] Osorio, F. G., Varela, I., Lara, E., Puente, X. S., Espada, J., Santoro, R., Freije, J. M., Fraga, M. F., and López-Otín, C. (2010). Nuclear envelope alterations generate an aging-like epigenetic pattern in mice deficient in Zmpste24 metalloprotease. *Aging Cell*, 9(6):947–957.

[Pal et al., 2012] Pal, D., Dasgupta, S., Kundu, R., Maitra, S., Das, G., Mukhopadhyay, S., Ray, S., Majumdar, S. S., and Bhattacharya, S. (2012). Fetuin-A acts as an endogenous ligand of TLR4 to promote lipid-induced insulin resistance. *Nature Medicine*, 18(8):1279–1285.

[Pal and Tyler, 2016] Pal, S. and Tyler, J. K. (2016). Epigenetics and aging.

[Pérez-Silva et al., 2016] Pérez-Silva, J. G., Español, Y., Velasco, G., and Quesada, V. (2016). The Degradome database: Expanding roles of mammalian proteases in life and disease. *Nucleic Acids Research*, 44(D1):D351–D355.

[Powers et al., 2009] Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W., and Balch, W. E. (2009). Biological and Chemical Approaches to Diseases of Proteostasis Deficiency. *Annual Review of Biochemistry*, 78(1):959–991.

[Prattichizzo et al., 2017] Prattichizzo, F., Micolucci, L., Cricca, M., De Carolis, S., Mensà, E., Ceriello, A., Procopio, A. D., Bonafè, M., and Olivieri, F. (2017). Exosome-based immunomodulation during aging: A nano-perspective on inflammaging. *Mechanisms of Ageing and Development*, 168:44–53.

[Proietta et al., 2014] Proietta, M., Tritapepe, L., Cifani, N., Ferri, L., Taurino, M., and Del Porto, F. (2014). MMP-12 as a new marker of Stanford-A acute aortic dissection. *Annals of Medicine*, 46(1):44–48.

[Puente et al., 2006] Puente, X. S., Velasco, G., Gutiérrez-Fernández, A., Bertranpetit, J., King, M. C., and López-Otín, C. (2006). Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics*, 7:1–9.

[Quélin et al., 2006] Quélin, F., Mathonnet, F., Potentini-Esnault, C., Trigui, N., Peynet, J., Bastenaire, B., Guillon, L., Bigel, M. L., Sauger, A., Mazurier, C., and De Mazancourt, P. (2006). Identification of five novel mutations in the factor XI gene (F11) of patients with factor XI deficiency. *Blood Coagulation and Fibrinolysis*, 17(1):69–73.

[Quesada et al., 2009] Quesada, V., Ordóñez, G. R., Sánchez, L. M., Puente, X. S., and López-Otín, C. (2009). The degradome database: Mammalian proteases and diseases of proteolysis. *Nucleic Acids Research*, 37(SUPPL. 1):D239–43.

[Quirós et al., 2015] Quirós, P. M., Langer, T., and López-Otín, C. (2015). New roles for mitochondrial proteases in health, ageing and disease. *Nature Reviews Molecular Cell Biology*, 16(6):345–359.

[R. Clarke, 1970] R. Clarke, M. (1970). Function of the spermaceti organ of the sperm whale. *Nature Publishing Group*, 228:726–734.

[Rando and Chang, 2012] Rando, T. A. and Chang, H. Y. (2012). Aging, rejuvenation, and epigenetic reprogramming: Resetting the aging clock. *Cell*, 148(1-2):46–57.

[Reinhard et al., 2015] Reinhard, S. M., Razak, K., and Ethell, I. M. (2015). A delicate balance: Role of MMP-9 in brain development and pathophysiology of neurodevelopmental disorders.

[Rera et al., 2011] Rera, M., Bahadorani, S., Cho, J., Koehler, C. L., Ulgherait, M., Hur, J. H., Ansari, W. S., Lo, T., Jones, D. L., and Walker, D. W. (2011). Modulation of longevity and tissue homeostasis by the drosophila PGC-1 homolog. *Cell Metabolism*, 14(5):623–634.

[Rhodin et al., 2017] Rhodin, A. G., Iverson, J. B., Bour, R., Fritz, U., Georges, A., Shaffer, H. B., and van Dijk, P. P. (2017). *Turtles of the World: Annotated Checklist*

and Atlas of Taxonomy, Synonymy, Distribution, and Conservation Status. 8th editio edition.

[Richard et al., 1994] Richard, K. R., McCarrey, S. W., and Wright, J. M. (1994). DNA sequence from the SRY gene of the sperm whale (Physeter macrocephalus) for use in molecular sexing . *Canadian Journal of Zoology*, 72(5):873–877.

[Rossi et al., 2007] Rossi, D. J., Bryder, D., Seita, J., Nussenzweig, A., Hoeijmakers, J., and Weissman, I. L. (2007). Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature*, 447(7145):725–729.

[Ryazantseva et al., 2016] Ryazantseva, M., Skobeleva, K., Glushankova, L., and Kaznacheyeva, E. (2016). Attenuated presenilin-1 endoproteolysis enhances store-operated calcium currents in neuronal cells. *Journal of Neurochemistry*, 136(5):1085–1095.

[Saleh et al., 2004] Saleh, M., Vaillancourt, J. P., Graham, R. K., Huyck, M., Srinivasula, S. M., Alnemri, E. S., Steinberg, M. H., Holan, V., Baldwin, C. T., Hotchkiss, R. S., Buchman, T. G., Zehnbauer, B. A., Hayden, M. R., Farrer, L. A., Roy, S., and Nicholson, D. W. (2004). Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature*, 429(6987):75–79.

[Salminen et al., 2012] Salminen, A., Kaarniranta, K., and Kauppinen, A. (2012). Inflammaging: Disturbed interplay between autophagy and inflammasomes. *Aging*, 4(3):166–175.

[Sanchis-Gomar et al., 2015] Sanchis-Gomar, F., Pareja-Galeano, H., Santos-Lozano, A., Garatachea, N., Fiuza-Luces, C., Venturini, L., Ricevuti, G., Lucia, A., and Emanuele, E. (2015). A preliminary candidate approach identifies the combination of chemerin, fetuin-A, and fibroblast growth factors 19 and 21 as a potential biomarker panel of successful aging. *Age*, 37(3).

[Scheper and Hoozemans, 2015] Scheper, W. and Hoozemans, J. J. (2015). The unfolded protein response in neurodegenerative diseases: a neuropathological perspective. *Acta Neuropathologica*, 130(3):315–331.

[Schorr et al., 2014] Schorr, G. S., Falcone, E. A., Moretti, D. J., and Andrews, R. D. (2014). First long-term behavioral records from Cuvier's beaked whales (Ziphius cavirostris) reveal record-breaking dives. *PLoS ONE*, 9(3).

[Schulz et al., 2011] Schulz, T. M., Whitehead, H., Gero, S., and Rendell, L. (2011). Individual vocal production in a sperm whale (Physeter macrocephalus) social unit. *Marine Mammal Science*, 27(1):149–166.

[Seppey et al., 2019] Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In *Methods in Molecular Biology*, volume 1962, pages 227–245. Humana Press Inc.

[Shaw et al., 2010] Shaw, A. C., Joshi, S., Greenwood, H., Panda, A., and Lord, J. M. (2010). Aging of the innate immune system. *Current Opinion in Immunology*, 22(4):507–513.

[Shay, 2016] Shay, J. W. (2016). Role of telomeres and telomerase in aging and cancer.

[Singh et al., 2019] Singh, P. P., Demmitt, B. A., Nath, R. D., and Brunet, A. (2019). The Genetics of Aging: A Vertebrate Perspective. *Cell*, 177(1):200–220.

[Smit et al.,] Smit, A., Hubley, R., and Green, P. RepeatMasker Open-4.0.

[Sokolov, 1982] Sokolov, V. E. (1982). *Mammal skin*. University of California Press.

[Speakman, 2005] Speakman, J. R. (2005). Body size, energy metabolism and lifespan. *Journal of Experimental Biology*, 208(9):1717–1730.

[Stengaard-Pedersen et al., 2003] Stengaard-Pedersen, K., Thiel, S., Gadjeva, M., Møller-Kristensen, M., Sørensen, R., Jensen, L. T., Sjöholm, A. G., Fugger, L., and Jensenius, J. C. (2003). Inherited deficiency of mannan-binding lectin-associated serine protease 2. *New England Journal of Medicine*, 349(6):554–560.

[Stenvinkel et al., 2017] Stenvinkel, P., Luttropp, K., Mcguinness, D., Witasp, A., Qureshi, A. R., Wernerson, A., Nordfors, L., Schalling, M., Ripsweden, J., Wennberg, L., Söderberg, M., Bárány, P., Olauson, H., and Shiels, P. G. (2017). CDKN2A p16INK4 expression is associated with vascular progeria in CKD. *Aging*, 9(2):494–507.

[Stewart, 2009] Stewart, B. S. (2009). Diving behavior. *Encyclopedia of Marine Mammals*, pages 321–327.

[Strauss, 1969] Strauss, M. B. (1969). Mammalian adaptations to diving.

[Suh et al., 2014] Suh, A., Churakov, G., Ramakodi, M. P., Platt, R. N., Jurka, J., Kojima, K. K., Caballero, J., Smit, A. F., Vliet, K. A., Hoffmann, F. G., Brosius, J., Green, R. E., Braun, E. L., Ray, D. A., and Schmitz, J. (2014). Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biology and Evolution*, 7(1):205–217.

[Sun et al., 2009] Sun, Q., Jin, H. J., and Bond, J. S. (2009). Disruption of the meprin α and β genes in mice alters homeostasis of monocytes and natural killer cells. *Experimental Hematology*, 37(3):346–356.

[Swee et al., 2008] Swee, M., Wilson, C. L., Wang, Y., McGuire, J. K., and Parks, W. C. (2008). Matrix metalloproteinase-7 (matrilysin) controls neutrophil egress by generating chemokine gradients. *Journal of Leukocyte Biology*, 83(6):1404–1412.

[Tabuce et al., 2008] Tabuce, R., Asher, R. J., and Lehmann, T. (2008). Afrotherian mammals: A review of current data. *Mammalia*, 72(1):2–14.

[Tollis et al., 2017a] Tollis, M., Boddy, A. M., and Maley, C. C. (2017a). Peto's Paradox: How has evolution solved the problem of cancer prevention? *BMC Biology*, 15(1):1–5.

[Tollis et al., 2017b] Tollis, M., Denardo, D. F., Cornelius, J. A., Dolby, G. A., Edwards, T., Henen, B. T., Karl, A. E., Murphy, R. W., and Kusumi, K. (2017b). The Agassiz ' s desert tortoise genome provides a resource for the conservation of a threatened species. pages 1–25.

[Trapnell et al., 2009] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

[Tsai et al., 2010] Tsai, I. J., Otto, T. D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, 11(R41).

[Tümpel and Rudolph, 2019] Tümpel, S. and Rudolph, K. L. (2019). Quiescence: Good and Bad of Stem Cell Aging. *Trends in Cell Biology*, 29(8):672–685.

[Turk et al., 2012] Turk, B., Turk, D., and Turk, V. (2012). Protease signalling: The cutting edge.

[Turner et al., 2019] Turner, K., Vasu, V., and Griffin, D. (2019). Telomere Biology and Human Phenotype. *Cells*, 8(1):73.

[Uria and Lopez-Otin, 2000] Uria, J. A. and Lopez-Otin, C. (2000). Matrilysin-2, a new matrix metalloproteinase expressed in human tumors and showing the minimal domain organization required for secretion, latency, and activity. *Cancer Research*, 60(17):4745–4751.

[Verweij et al., 2013] Verweij, N., Mahmud, H., Leach, I. M., De Boer, R. A., Brouwers, F. P., Yu, H., Asselbergs, F. W., Struck, J., Bakker, S. J., Gansevoort, R. T., Munroe, P. B., Hillege, H. L., Van Veldhuisen, D. J., Van Gilst, W. H., Silljé, H. H., and Van Der Harst, P. (2013). Genome-wide association study on plasma levels of midregional-proadrenomedullin and C-terminal-pro-endothelin-1. *Hypertension*, 61(3):602–608.

[Voskoboinik et al., 2015] Voskoboinik, I., Whisstock, J. C., and Trapani, J. A. (2015). Perforin and granzymes: Function, dysfunction and human pathology. *Nature Reviews Immunology*, 15(6):388–400.

[Wada et al., 2003] Wada, S., Oishi, M., and Yamada, T. K. (2003). A newly discovered species of living baleen whale. *Nature*, 426(6964):278–281.

[Wang et al., 2014] Wang, L., Karpac, J., and Jasper, H. (2014). Promoting longevity by maintaining metabolic and proliferative homeostasis. *Journal of Experimental Biology*, 217(1):109–118.

[Wang et al., 2013] Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., Li, C., White, S., Xiong, Z., Fang, D., Wang, B., Ming, Y., Chen, Y., Zheng, Y., Kuraku, S., Pignatelli, M., Herrero, J., Beal, K., Nozawa, M., Li, Q., Wang, J., Zhang, H., Yu, L., Shigenobu, S., Wang, J., Liu, J., Flicek, P., Searle, S., Weng, J., Kuratani, S., Yin, Y., Aken, B., Zhang, G., and Irie, N. (2013). The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nature Genetics*, 45(6):701–706.

[Warren et al., 2017] Warren, W. C., Kuderna, L., Alexander, A., Catchen, J., Pérez-Silva, J. G., López-Otín, C., Quesada, V., Minx, P., Tomlinson, C., Montague, M. J., Farias, F. H., Walter, R. B., Marques-Bonet, T., Glenn, T., Kieran, T. J., Wise, S. S., Wise, J. P., and Waterhouse, R. M. (2017). The Novel Evolution of the Sperm Whale Genome. *Genome Biology and Evolution*, 9(12):3260–3264.

[Whitehead, 2003] Whitehead, H. (2003). *Sperm whales : social evolution in the ocean*. University of Chicago Press.

[Wilcock et al., 2019] Wilcock, A., Bahri, R., Bulfone-Paus, S., and Arkwright, P. D. (2019). Mast cell disorders: From infancy to maturity. *Allergy: European Journal of Allergy and Clinical Immunology*, 74(1):53–63.

[Wong and Takei, 2013] Wong, M. K. S. and Takei, Y. (2013). Lack of plasma kallikrein-kinin system cascade in teleosts. *PLoS ONE*, 8(11):1–13.

[Worley et al., 2014] Worley, K. C., Warren, W. C., Rogers, J., Locke, D., Muzny, D. M., Mardis, E. R., Weinstock, G. M., Tardif, S. D., Aagaard, K. M., Archidiacono, N., Arul Rayan, N., Batzer, M. A., Beal, K., Brejova, B., Capozzi, O., Capuano, S. B., Casola, C., Chandrabose, M. M., Cree, A., Diep Dao, M., De Jong, P. J., Cruz-Herrera del Rosario, R., Delehaunty, K. D., Dinh, H. H., Eichler, E. E., Fitzgerald, S., Flicek, P., Fontenot, C. C., Fowler, R. G., Fronick, C., Fulton, L. A., Fulton, R. S., Gabisi, R. A., Gerlach, D., Graves, T. A., Gunaratne, P. H., Hahn, M. W., Haig, D., Han, Y., Harris, R. A., Herrero, J., Hillier, L. D. W., Hubley, R., Hughes, J. F., Hume, J., Jhangiani, S. N., Jorde, L. B., Joshi, V., Karakor, E., Konkel, M. K., Kosiol, C., LKovar, C., Kriventseva, E. V., Lee, S. L., Lewis, L. R., Liu, Y. S., Lopez, J., Lopez-Otin, C., Lorente-Galdos, B., Mansfield, K. G., Marques-Bonet, T., Minx, P., Misceo, D., Moncrieff, J. S., Morgan, M. B., Nazareth, L. V., Newsham, I., Nguyen, N. B., Okwuonu, G. O., Prabhakar, S., Perales, L., Pu, L. L., Puente, X. S., Quesada, V., Ranck, M. C., Raney, B. J., Raveendran, M., Deiros, D. R., Rocchi, M., Rodriguez, D., Ross, C., Ruffier, M., Ruiz, S. J., Sajjadian, S., Santibanez, J., Schrider, D. R., Searle, S., Skaletsky, H., Soibam, B., Smit, A. F., Tennakoon, J. B., Tomaska, L., Ullmer, B., Vejnar, C. E., Ventura, M., Vilella, A. J., Vinar, T., Vogel, J. H., Walker, J. A., Wang, Q., Warner, C. M., Wildman, D. E., Witherspoon, D. J., Wright, R. A., Wu, Y., Xiao, W., Xing, J., Zdobnov, E. M., Zhu, B., Gibbs, R. A., and Wilson, R. K. (2014). The common marmoset genome provides insight into primate biology and evolution. *Nature Genetics*, 46(8):850–857.

[Wu et al., 2019] Wu, G., Quek, A. J., Caradoc-Davies, T. T., Ekkel, S. M., Mazzitelli, B., Whisstock, J. C., and Law, R. H. (2019). Structural studies of plasmin inhibition. *Biochemical Society Transactions*, 47(2):541–557.

[Xue et al., 2007] Xue, W., Zender, L., Miething, C., Dickins, R. A., Hernando, E., Krizhanovsky, V., Cordon-Cardo, C., and Lowe, S. W. (2007). Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature*, 445(7128):656–660.

[Yang, 2007] Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

[Yim et al., 2014] Yim, H. S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J. Y., Cha, S. S., Oh, H. M., Lee, J. H. J. H. J. H., Yang, E. C., Kwon, K. K., Kim, Y. J., Kim, T. W. H. W., Kim, W., Jeon, J. H., Kim, S. J., Choi, D. H., Jho, S., Kim, H. S. H. H. M. H. S., Ko, J., Kim, H. S. H. H. M. H. S., Shin, Y. A., Jung, H. J., Zheng, Y., Wang, Z., Chen, Y., Chen, M., Jiang, A., Li, E., Zhang, S., Hou, H., Kim, T. W. H. W., Yu, L., Liu, S., Ahn, K., Cooper, J., Park, S. G., Hong, C. P., Jin, W., Kim, H. S. H. H. M. H. S., Park, C., Lee, K., Chun, S., Morin, P. A., O'Brien, S. J., Lee, H. S. H. H. S., Kimura, J., Moon, D. Y., Manica, A., Edwards, J., Kim, B. C., Kim, S. J., Wang, J., Bhak, J., Lee, H. S. H. H. S., and Lee, J. H. J. H. J. H. (2014). Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*, 46(1):88–92.

[Yoon et al., 2012] Yoon, S. H., Bae, Y. S., Mun, M. S., Park, K. Y., Ye, S. K., Kim, E., and Kim, M. H. (2012). Developmental retardation, microcephaly, and peptiduria in mice without aminopeptidase P1. *Biochemical and Biophysical Research Communications*, 429(3-4):204–209.

[Zhang and Cuervo, 2008] Zhang, C. and Cuervo, A. M. (2008). Restoration of chaperone-mediated autophagy in aging liver improves cellular maintenance and hepatic function. *Nature Medicine*, 14(9):959–965.

[Zhang et al., 2013] Zhang, G., Cowled, C., Shi, Z., Huang, Z., Bishop-Lilly, K. a., Fang, X., Wynne, J. W., Xiong, Z., Baker, M. L., Zhao, W., Tachedjian, M., Zhu, Y., Zhou, P., Jiang, X., Ng, J., Yang, L., Wu, L., Xiao, J., Feng, Y., Chen, Y., Sun, X., Zhang, Y., Marsh, G. a., Crameri, G., Broder, C. C., Frey, K. G., Wang, L.-F., and Wang, J. (2013). Comparative Analysis of Bat Genomes. *Science*, 339(January):456 – 460.

[Zhang et al., 2000] Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7(1-2):203–214.

[Zimmerman et al., 2010] Zimmerman, L. M., Vogel, L. A., and Bowden, R. M. (2010). Commentary: Understanding the vertebrate immune system: Insights from the reptilian perspective. *Journal of Experimental Biology*, 213(5):661–671.

Supplementary info: Figures

Fig. A.1: Cetacean truncations and alterations in proteases. Each mark represents an adaptative event, either truncations or duplications (represented as a doble mark), in the branch in which data suggests it occurred. Events of convergent evolution are represented by a doted line joining the pertinent marks. Those marks filled with **green** are those that, although mentioned in the main text, most likely played a late role once a main event of adaptation had already occured (said event will be represented by **orange** as the rest). The grey mark represents an absence. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*. The tree itself is based in the relations shown by "Time tree" [Kumar et al., 2017].

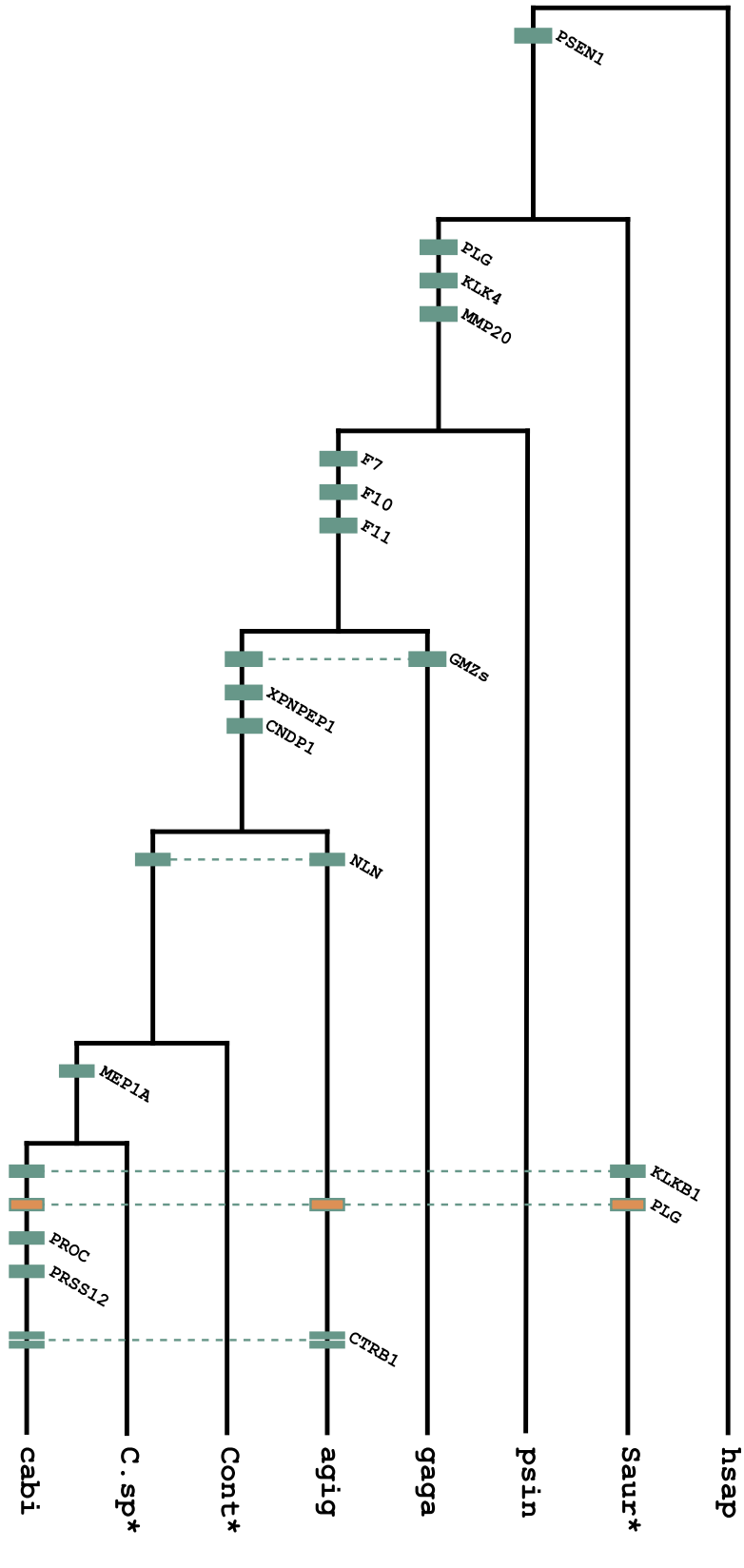

Fig. A.2: Testudines truncations and alterations in proteases. Each mark represents an adaptative event, either truncations or duplications (represented as a doble mark), in the branch in which data suggests it occurred. Events of convergent evolution are represented by a doted line joining the pertinent marks. Those marks filled with orange are those that, although mentioned in the main text, most likely played a late role once a main event of adaptation had already occured (said event will be represented by green as the rest). `hsap`, *H. sapiens*; `Saur*` represents other Sauropsids (such as *G. gallus*, or *A. carolinensis*); `psin`, *P. sinensis*; `gaga`, *G. agassizii*; `agig`, *A. gigantea*; `Cont*` reffers to the continental outgroups of the Galápagos giant tortoises, *C. denticulata* y *C. carbonaria*; `C.sp*` (*C. sp*) refers to all Galápagos giant tortoises but *C. abingdoni*, which is referenced as `cabi`. The tree itself is based in the relations shown by "Time tree" [Kumar et al., 2017].

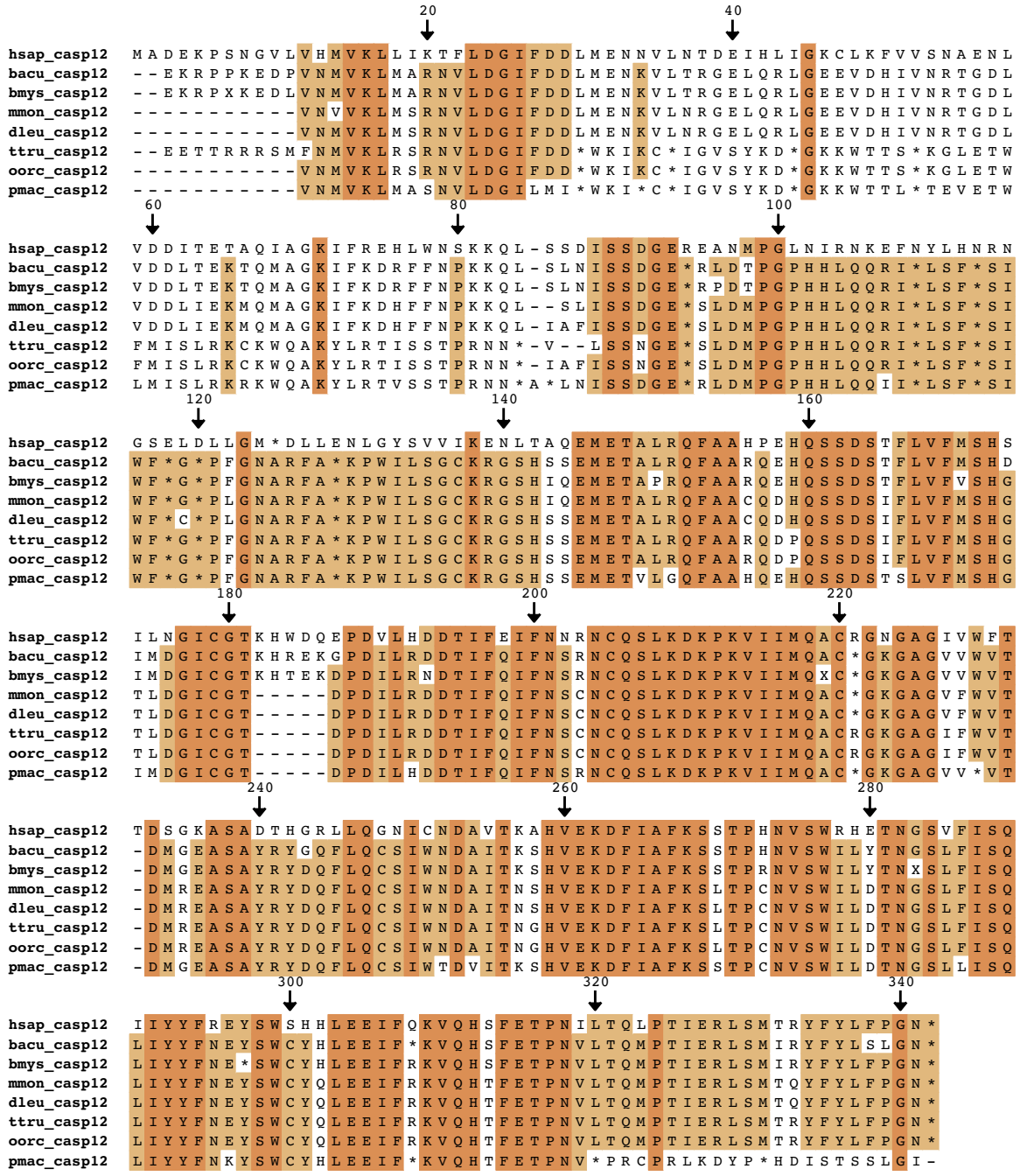

Fig. A.3: Complete alignment of CASP12 in cetaceans, showing the complex scheme of evolutive history that lead to its truncation in all cetaceans. Using as reference the truncated human protein referred to in the main text. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

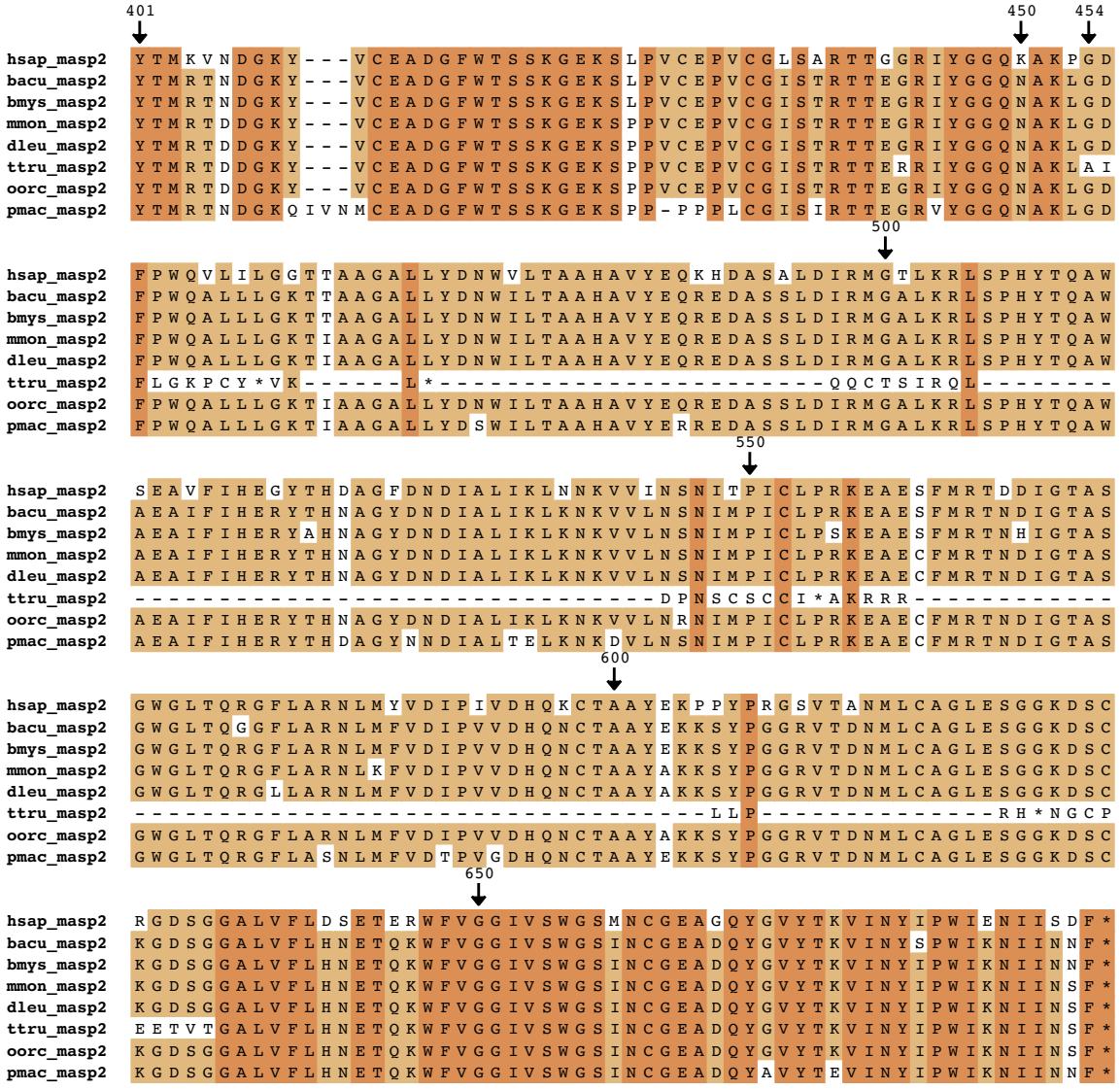

Fig. A.4: Alignment of MASP2 n cetaceans, showing the framshift mentioned in the main text, occurred exclusively on the bottlenose dolphin, starting in p.G454 and causing several premature stop codons to appear. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

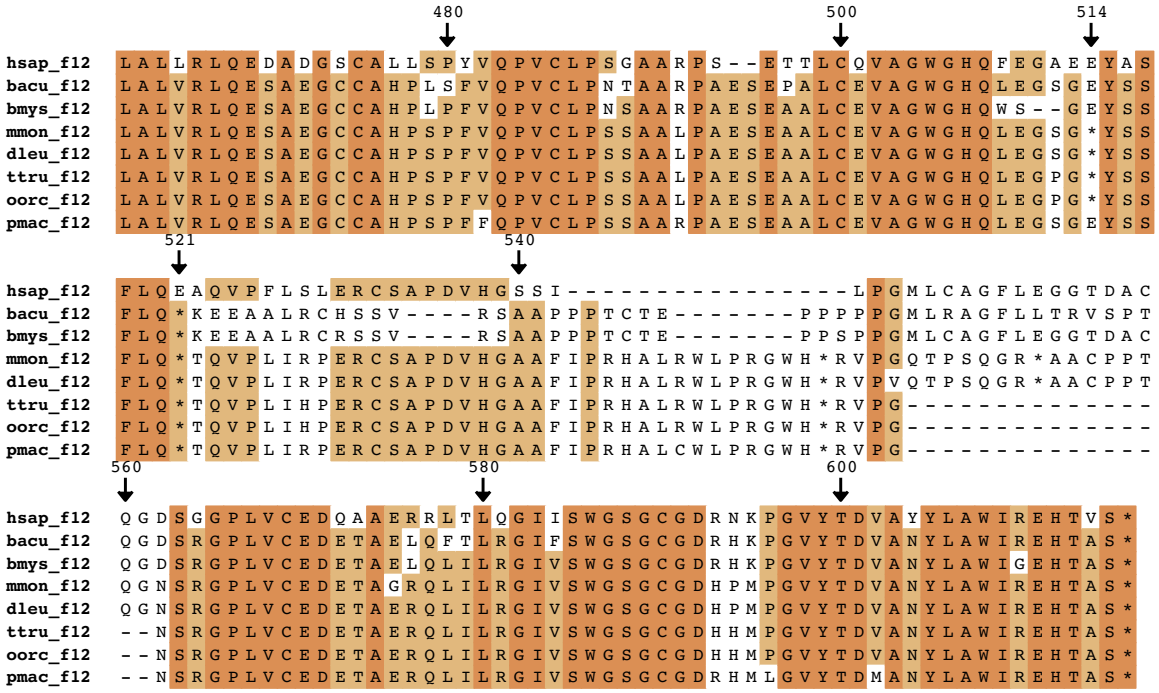

Fig. A.5: Alignment of F12 in cetaceans, showing the different premature stop codons shared by the different groups, caused by different patterns of frameshifts. A different common patter of frameshift can be seen also in the two Misticeti, despite the fact that it does not provoke a premature stop codon as a result. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

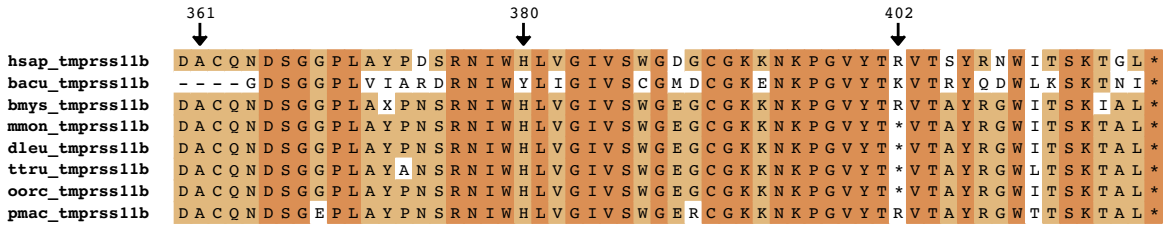

Fig. A.6: Alignment of TMPRSS11B in cetaceans, showing the shared premature stop codon common to all Delphinoidea mentiopned in the main text. `hsap`, *H. sapiens*; `bacu`, *B. acutorostrata*; `bmys`, *B. mysticetus*; `mmon`, *M. monoceros*; `dleu`, *D. leucas*; `ttru`, *T. truncatus*; `oorc`, *O. orca*; and `pmac`, *P. macrocephalus*.

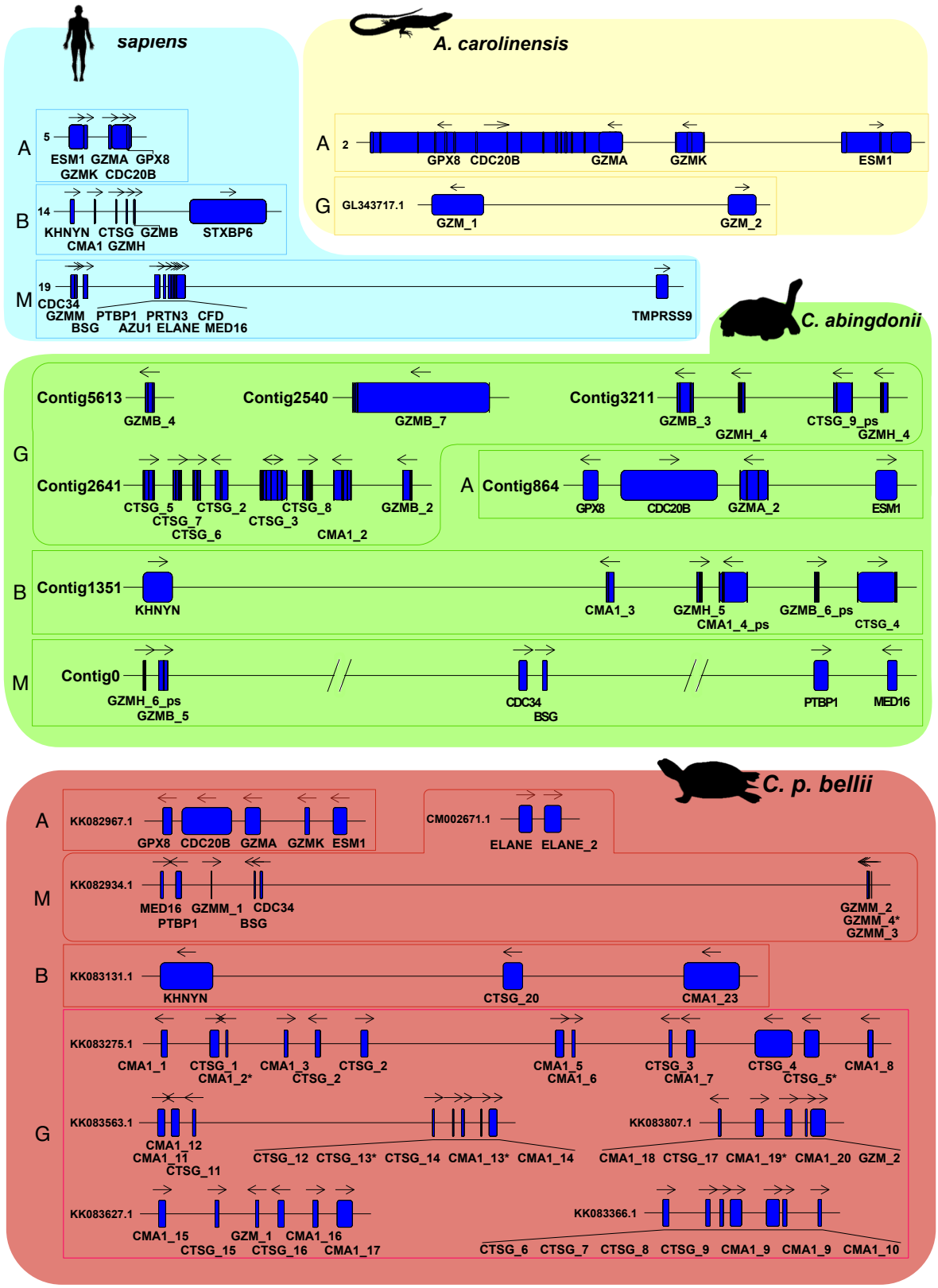

Fig. A.7: Expansion of the granzyme clusters in *C. abingdonii* and other organisms (*P. sinensis*, *A. carolinensis*, and *H. sapiens*).

Supplementary info: Tables

Table B.1: Possitive selection analysis results in *C. abingdonii*. Only shown those with an $\omega 2$ greater than 1. Also removed those with incomplete data, and those with "too high" $\omega 2$ values. The complete table can be checked here.

Gene	$\omega 2$	$\omega 1$	$\omega 0$	$\omega 2$-$\omega 1$
TUBE1	153,06935	0,07402	0,04658	152,99533
FRRS1L	30,98228	0,01332	0,01626	30,96896
TUBG1	7,98622	0,00904	0,00918	7,97718
CLRN2	3,88988	0,15350	0,15807	3,73638
MVK	3,77348	0,17305	0,17735	3,60043
CMAS	3,10102	0,13879	0,15952	2,96223
CREBL2	3,09570	0,08965	0,04218	3,00605
IRAK1BP1	2,35841	0,17743	0,13865	2,18098
ALG1	2,12288	0,16317	0,17949	1,95971
COX18	1,71757	0,19049	0,21782	1,52708
NTF3	1,61603	0,08596	0,08035	1,53007
IL1R2	1,60957	0,30844	0,30517	1,30113
AHSG	1,54836	0,43872	0,36735	1,10964
PTCD2	1,42073	0,37172	0,37426	1,04901
FGF19	1,40904	0,14895	0,16238	1,26009
UBE2J1	1,39297	0,13930	0,01943	1,25367
F2R	1,33392	0,21719	0,167	1,11673
EMC2	1,27666	0,03209	0,05771	1,24457
MED19	1,26411	0,01153	0,00912	1,25258
VPS35	1,26021	0,00873	0,00869	1,25148
ATP4B	1,24188	0,17286	0,18168	1,06902
CRLS1	1,23942	0,06322	0,52669	1,17620
ITM2B	1,23539	0,07216	0,07512	1,16323
HRH1	1,20455	0,23600	0,16651	0,96855
PCP4	1,18271	0,01507	0,01507	1,16764
PIM3	1,17259	0,02294	0,02177	1,14965
GTPBP10	1,17153	0,18357	0,23034	0,98796
BAG2	1,16059	0,07724	0,07518	1,08335
GDPD1	1,13870	0,07981	0,08303	1,05889
WDR27	1,07450	0,26707	0,17619	0,80743
COTL1	1,06762	0,02572	0,02609	1,04190
UQCRB	1,06365	0,13591	0,12339	0,92774
TDO2	1,06041	0,08567	0,04462	0,97474

Table B.1 – *Continues on next page*

Table B.1 – *Continued from previous page*

Gene	$\omega 2$	$\omega 1$	$\omega 0$	$\omega 2$-$\omega 1$
MALSU1	1,05997	0,16755	0,18792	0,89242
SGPP2	1,05829	0,17148	0,16672	0,88681
CHIC2	1,05334	0,05987	0,00887	0,99347
DYNLRB1	1,04162	0,03182	0,03816	1,00980
TMEM38A	1,02214	0,08349	0,05464	0,93865
KCNK18	1,01755	0,24701	0,21284	0,77054
KDSR	1,00704	0,08597	0,05731	0,92107

Table B.2: Results of the positive selected sites in the positive selected genes, including their *equivalent* in human, and their status in Aldabra giant tortoises (where "Conserved" reffers to its status in Galápagos giant tortoises.

Gen	Residue	Contig	Position	Human *eq.*	Status in *A. gigantea*
BCL2A1	(-113)	Contig989	18520		Conserved
BCL2A1	(K102)	Contig989	18553	K102K	Conserved
BCL2A1	(K146)	Contig989	16839	K146T	Conserved
BCL2A1	(L99)	Contig989	18562	L99I	Conserved
BCL2A1	(N129)	Contig989	18472		Conserved
BCL2A1	(Q116)	Contig989	18511		Conserved
BAG2	(A94)	Contig1585	74920	A131Y	Conserved
BAG2	(G52)	Contig1585	75046	G89D	Mutated to G
BAG2	(H156)	Contig1585	74734	H193R	Conserved
BAG2	(I144)	Contig1585	74770	I181L	Conserved
BAG2	(N93)	Contig1585	74923	N130G	Conserved
BAG2	(Q116)	Contig1585	74854	Q153H	Mutated to Q
BAG2	(T56)	Contig1585	75034	T93S	Conserved
CCKAR	(A304)	Contig623	335179	A303V	Conserved
CCKAR	(E403)	Contig623	335476		Conserved
CCKAR	(S210)	Contig623	334255	S208E	Conserved
CCKAR	(W211)	Contig623	334258	W209R	Conserved
CLRN2	(F189)	Contig471	332683	F189V	Mutated to *,Q
CLRN2	(G146)	Contig471	332812	G146S	Mutated to P,S
CLRN2	(I208)	Contig471	332626	I208V	Conserved
CLRN2	(I22)	Contig471	337663		Conserved
CLRN2	(K217)	Contig471	332599	K217R	Conserved
CLRN2	(L107)	Contig471	333408	L107G	Conserved
CLRN2	(P212)	Contig471	332614	P212R	Conserved
CLRN2	(S14)	Contig471	337687	S14A	Conserved
CLRN2	(V90)	Contig471	333459	V90M	Mutated to H
CLRN2	(W9)	Contig471	337699	W9F	Mutated to N
COX18	(E57)	Contig2388	18753	E241Q	Conserved
COX18	(F147)	Contig2388	20469		Conserved
COX18	(F148)	Contig2388	20472		Conserved
COX18	(I82)	Contig2388	18828		Conserved
COX18	(I85)	Contig2388	18837		Conserved
COX18	(R123)	Contig2388	20397		Conserved

Table B.2 – *Continues on next page*

Table B.2 – *Continued from previous page*

Gen	Residue	Contig	Position	Human *eq.*	Status in *A. gigantea*
COX18	(V56)	Contig2388	18750	V240L	Conserved
CREBL2	(-100)	Contig520	637897	Q101H	Mutated to V
CREBL2	(-106)	Contig520	637879		Conserved
CREBL2	(-99)	Contig520	637900		Conserved
CRLS1	(A109)	Contig1626	302674		Conserved
CRLS1	(I126)	Contig1626	302725	I241V	Conserved
CRLS1	(G110)	Contig1626	302677	A224G	Conserved
CRLS1	(L108)	Contig1626	302671	L223V	Conserved
CRLS1	(L34)	Contig1626	297486		Conserved
CRLS1	(V77)	Contig1626	302042		Conserved
EMC2	(I53)	Contig465	1306819	I54V	Mutated to I
EMC2	(R95)	Contig465	1312717	R96K	Conserved
EMC2	(T183)	Contig465	1338091	T184A	Conserved
EMC2	(Y255)	Contig465	1341115	Y256F	Conserved
EMC2	(Y269)	Contig465	1341157		Conserved
ERP27	(D127)	Contig2026	58280	D250S	Conserved
ERP27	(E108)	Contig2026	58223	E231Q	Conserved
ERP27	(H121)	Contig2026	58262		Conserved
ERP27	(K136)	Contig2026	58307	K259I	Conserved
ERP27	(S131)	Contig2026	58292	S254K	Conserved
ERP27	(T7)	Contig2026	47324	T130A	Conserved
ERP27	(V27)	Contig2026	47381	V150L	Conserved
FAM69C	(E282)	Contig976	489334	E413Q	Conserved
FAM69C	(G36)	Contig976	498104	G169T	Heterozygous
FAM69C	(G44)	Contig976	498080	G177S	Conserved
FAM69C	(P78)	Contig976	497978	P211Q	Conserved
FAM69C	(S33)	Contig976	498113	S166T	Conserved
AHSG	(-317)	Contig54	2421121		Conserved
AHSG	(D203)	Contig54	2418218	D208T	Conserved
AHSG	(E221)	Contig54	2418272		Mutated to K
AHSG	(G284)	Contig54	2421022		Conserved
AHSG	(H335)	Contig54	2421175	H365Y	Conserved
AHSG	(I42)	Contig54	2411438	I45V	Conserved
AHSG	(K121)	Contig54	2415816	K124R	Conserved
AHSG	(K128)	Contig54	2415837	K131Q	Conserved
AHSG	(L8)	Contig54	2411324	L8I	Conserved

Table B.2 – *Continues on next page*

Table B.2 – *Continued from previous page*

Gen	Residue	Contig	Position	Human *eq.*	Status in *A. gigantea*
AHSG	(N175)	Contig54	2417838	N180F	Conserved
AHSG	(S306)	Contig54	2421100		Conserved
AHSG	(T92)	Contig54	2413074	T95V	Mutated to L
AHSG	(V107)	Contig54	2415774	V110I	Mutated to V
AHSG	(V139)	Contig54	2417730	V146I	Conserved
AHSG	(V199)	Contig54	2418206	V204I	Conserved
AHSG	(V285)	Contig54	2421037		Conserved
AHSG	(V5)	Contig54	2411315	V5I	Conserved
FGF19	(E15)	Contig4856	79636	E117A	Conserved
FGF19	(E72)	Contig4856	79807	E174I	Mutated to V
FGF19	(M91)	Contig4856	79864	M187V	Conserved
FGF19	(S14)	Contig4856	79633	S116A	Conserved
FRRS1L	(S85)	Contig1385	415653	S289K	Conserved
GDPD1	(E25)	Contig2053	522028	E81Q	Conserved
GDPD1	(N101)	Contig2053	519402	N157S	Conserved
GDPD1	(N126)	Contig2053	517280	N182S	Mutated to L
GDPD1	(S191)	Contig2053	513811	S247R	Conserved
GDPD1	(V43)	Contig2053	521974	V99I	Mutated to Y
GPR39	(S213)	Contig1955	248729	S214A	Conserved
GTPBP10	(A44)	Contig401	99027		Conserved
GTPBP10	(E238)	Contig401	116152	E267D	Conserved
GTPBP10	(F151)	Contig401	108563		Conserved
GTPBP10	(G41)	Contig401	99018		Conserved
GTPBP10	(I225)	Contig401	113162	I254V	Conserved
HRH1	(-202)	Contig1259	121479		Conserved
HRH1	(-209)	Contig1259	121500		Conserved
HRH1	(H265)	Contig1259	121704	H309K	Conserved
HRH1	(I78)	Contig1259	121104	I123L	Conserved
HRH1	(K221)	Contig1259	121569		Mutated to N
HRH1	(N121)	Contig1259	121233	N166R	Conserved
HRH1	(V18)	Contig1259	120918	V61D	Heterozygous
HSPA14	(N30)	Contig455	1023422	N153H	Conserved
HSPA14	(V13)	Contig455	1023371	A136A	Conserved
IL20RA	(-307)	Contig11	1428874		Conserved
IL20RA	(D370)	Contig11	1429087		Conserved
IL20RA	(D473)	Contig11	1429444	D528N	Conserved

Table B.2 – *Continues on next page*

Table B.2 – *Continued from previous page*

Gen	Residue	Contig	Position	Human *eq.*	Status in *A. gigantea*
IL20RA	(E421)	Contig11	1429273	E473G	Mutated to E
IL20RA	(E496)	Contig11	1429516	E552G	Conserved
IL20RA	(G374)	Contig11	1429099	G413V	Conserved
IL20RA	(G490)	Contig11	1429498	G546E	Conserved
IL20RA	(V493)	Contig11	1429507	V549I	Conserved
IRAK1BP1	(A58)	Contig331	1108671	V167I	Conserved
IRAK1BP1	(D115)	Contig331	1107447	S225N	Conserved
IRAK1BP1	(D143)	Contig331	1107363	K253Q	Conserved
IRAK1BP1	(G142)	Contig331	1107366	G252R	Conserved
IRAK1BP1	(I45)	Contig331	1108710	V154T	Conserved
IRAK1BP1	(V37)	Contig331	1108734	V146I	Conserved
IL1R2	(R227)	Contig315	1381393	R295Y	Mutated to H
IL1R2	(W205)	Contig315	1381327	W273T	Mutated to M
ITM2B	(A49)	Contig2350	230502	A105G	Conserved
ITM2B	(I181)	Contig2350	236814	I237L	Conserved
ITM2B	(K96)	Contig2350	233143		Conserved
ITM2B	(V148)	Contig2350	236715	V204I	Conserved
KCNK18	(-208)	Contig172	2027910		Mutated to G
KCNK18	(F309)	Contig172	2027613	F303V	Mutated to F
KCNK18	(F97)	Contig172	2034818	F95L	Conserved
KCNK18	(K373)	Contig172	2027421	K367R	Conserved
KDSR	(I136)	Contig295	233996	A276V	Conserved
KDSR	(Q120)	Contig295	234044		Conserved
KDSR	(S146)	Contig295	233966	S286T	Conserved
MVK	(A136)	Contig593	589742	A141T	Mutated to A
MVK	(D165)	Contig593	589655	D170T	Conserved
MVK	(D95)	Contig593	590372	D100E	Conserved
MVK	(G244)	Contig593	576336	G249S	Conserved
MVK	(H22)	Contig593	591616	H24Y	Conserved
MVK	(I216)	Contig593	578751	I221F	Conserved
MVK	(L163)	Contig593	589661	L168V	Conserved
MVK	(L177)	Contig593	580412	L182M	Conserved
MVK	(T173)	Contig593	580424	T178S	Conserved
MVK	(T99)	Contig593	590360	T104A	Conserved
MALSU1	(A86)	Contig12	3995739	A204V	Conserved
MALSU1	(L17)	Contig12	4003223	L135I	Conserved

Table B.2 – *Continues on next page*

Table B.2 – *Continued from previous page*

Gen	Residue	Contig	Position	Human *eq.*	Status in *A. gigantea*
CMAS	(A190)	Contig409	364190	A352Q	Conserved
CMAS	(C260)	Contig409	364934		Conserved
CMAS	(G26)	Contig409	358568	G186T	Conserved
CMAS	(G68)	Contig409	358694	G228N	Conserved
CMAS	(I259)	Contig409	364931	I421V	Conserved
CMAS	(K220)	Contig409	364814	K382R	Conserved
CMAS	(L175)	Contig409	364145	L334I	Conserved
CMAS	(R249)	Contig409	364901	R411H	Conserved
CMAS	(T171)	Contig409	364133	T333A	Mutated to I
CMAS	(V192)	Contig409	364196	V354I	Conserved
NTF3	(D57)	Contig636	963965		Conserved
NTF3	(E67)	Contig636	963995	E80G	Mutated to E
NTF3	(Q96)	Contig636	964082	Q109R	Conserved
NTF3	(T48)	Contig636	963938	L61F	Mutated to L
F2R	(L157)	Contig1017	589497	L216S	Conserved
PTCD2	(D102)	Contig888	45273	D144G	Conserved
PTCD2	(E159)	Contig888	37994	D201E	Conserved
PTCD2	(F156)	Contig888	38003	F198V	Mutated to I
PTCD2	(I154)	Contig888	38009		Conserved
PTCD2	(K190)	Contig888	34556	K232T	Conserved
PTCD2	(L42)	Contig888	46085	L84M	Conserved
PTCD2	(L68)	Contig888	46007		Conserved
PTCD2	(Q188)	Contig888	34562	L230H	Mutated to Q
PTCD2	(Q61)	Contig888	46028	E103K	Conserved
SGPP2	(-175)	Contig53	792630		Conserved
SGPP2	(-178)	Contig53	792639		Conserved
SGPP2	(A234)	Contig53	792816		Conserved
SGPP2	(C118)	Contig53	792459	C241R	Conserved
SGPP2	(E233)	Contig53	792813		Conserved
SGPP2	(E36)	Contig53	777142	E159D	Conserved
SGPP2	(I116)	Contig53	792453	I239V	Mutated to I
SGPP2	(K31)	Contig53	777127	K154V	Mutated to M
SGPP2	(L216)	Contig53	792762		Conserved
SGPP2	(P185)	Contig53	792660		Conserved
SGPP2	(T229)	Contig53	792801		Conserved
SGPP2	(V107)	Contig53	792426	V230I	Conserved

Table B.2 – *Continues on next page*

Table B.2 – *Continued from previous page*

Gen	Residue	Contig	Position	Human *eq.*	Status in *A. gigantea*
SGPP2	(V3)	Contig53	777046		Conserved
SRA1	(E11)	Contig498	530830		Mutated to T
SRA1	(H146)	Contig498	533764		Conserved
SRA1	(L103)	Contig498	533245		Conserved
TUBE1	(E94)	Contig2794	74108	E169D	Conserved
TUBE1	(I111)	Contig2794	74159	I186V	Conserved
TMEM38A	(A114)	Contig2922	57828	V159I	Conserved
TMEM38A	(D254)	Contig2922	59654	D299Q	Conserved
TMEM38A	(I88)	Contig2922	56328	I134I	Conserved
TMEM38A	(L135)	Contig2922	57891	L181I	Conserved
TMEM38A	(V108)	Contig2922	56388	V154I	Conserved
VPS35	(E763)	Contig1555	330070	E792D	Conserved
VPS35	(G664)	Contig1555	332054	G693T	Conserved
VPS35	(I767)	Contig1555	330058		Conserved
WDR27	(G465)	Contig131	944368		Conserved

Publications

Published online 8 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D351–D355***
doi: 10.1093/nar/gkv1201

The Degradome database: expanding roles of mammalian proteases in life and disease

José G. Pérez-Silva, Yaiza Español, Gloria Velasco and Víctor Quesada*

From the Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, 33006 Oviedo, Spain

Received September 14, 2015; Revised October 23, 2015; Accepted October 26, 2015

ABSTRACT

Since the definition of the degradome as the complete repertoire of proteases in a given organism, the combined effort of numerous laboratories has greatly expanded our knowledge of its roles in biology and pathology. Once the genomic sequences of several important model organisms were made available, we presented the Degradome database containing the curated sets of known protease genes in human, chimpanzee, mouse and rat. Here, we describe the updated Degradome database, featuring 81 new protease genes and 7 new protease families. Notably, in this short time span, the number of known hereditary diseases caused by mutations in protease genes has increased from 77 to 119. This increase reflects the growing interest on the roles of the degradome in multiple diseases, including cancer and ageing. Finally, we have leveraged the widespread adoption of new webtools to provide interactive graphic views that show information about proteases in the global context of the degradome. The Degradome database can be accessed through its web interface at http://degradome.uniovi.es.

INTRODUCTION

Proteases catalyze the hydrolysis of peptide bonds in a fundamentally irreversible reaction. This means that these enzymes must be tightly regulated in terms of activation and specificity to avoid massive homeostatic disorders. In turn, this need for specificity has led to the evolutionary expansion of protease genes to regulate the correct proteolysis of a large set of substrates. The parallel expansion of protease inhibitor genes added an additional level of complexity to this biochemical process. In recognition of this complex and interwoven system, the degradome of an organism was defined as the complete set of proteases in that organism (1). The degradome has been shown to affect most of the characterized biochemical pathways. Thus, different proteases are known to play key roles in such biological processes as cell cycle progression, tissue remodelling, neuronal outgrowth, haemostasis, wound healing, immunity, angiogenesis and apoptosis (2–6). Conversely, failures in the regulation of the degradome underlie diverse pathological conditions, including cancer, arthritis, progeria and neurological diseases (7–10).

Since the definition of the degradome relies on a global appraisal of the proteolytic processes, it follows that degradomics, the set of techniques specifically aimed at characterizing the degradome, must manage and integrate high-throughput data. In this regard, the completion of multiple genome projects allowed researchers to extend the degradomes of several species *in silico* from known protease sequences (11,12). Our experience in degradomics led us to tackle this problem with methods that relied heavily on manual curation after automatic predictions (13). As an additional advantage of this approach, a part of these projects consisted in the mining of the literature looking for known relationships between protease alterations and hereditary diseases, termed degradomopathies.

With this information, we described the Degradome database, containing the results of the manual annotation of every protease gene in the genomes of human, chimpanzee, mouse, and rat, along with relationships between protease alterations and hereditary diseases (14). This database complemented existing databases devoted to proteases, by providing a different focus. For instance, CutDB documents actual and predicted proteolytic events, but does not provide a global view of the proteases themselves (15). Also, MEROPS is a comprehensive and excellent database which relies on large-scale experiments and automatic predictions (16). By contrast, the Degradome database relies on manual annotation and exhaustive curation of genes, in multiple cases supported by direct cloning and sequencing experiments. For this reason, multiple instances of non-functional gene expansions, which our approach has filtered out, are annotated as putative proteases in the MEROPS database. In fact, the number of putative human proteases according to MEROPS is 990, whereas the Degradome database describes 588. In exchange, the MEROPS database features a very large number of species. Finally, our emphasis in diseases adds important information on the pathological rel-

*To whom correspondence should be addressed. Tel: +34-985 105025; Fax: +34 985 103564; Email: quesadavictor@uniovi.es

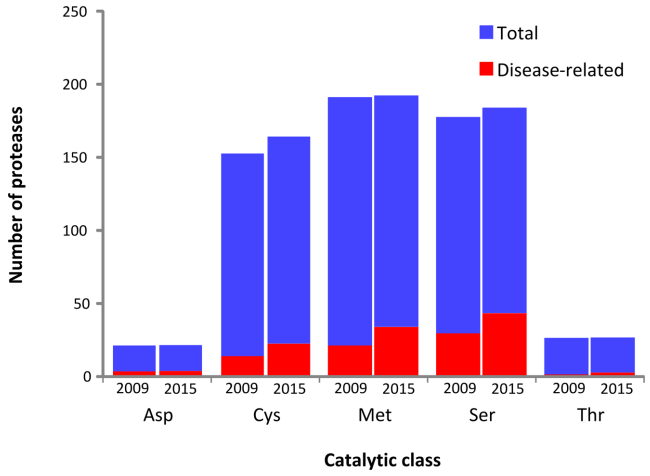

Figure 1. New annotations of proteases and degradomopathies. The number of proteases and degradomopathies annotated in each catalytic class is represented at the time of creation of the database (2009) and at the current version (2015).

evance of some proteases, which is not directly available in other databases. In our view, these databases taken together provide an accurate depiction of the current knowledge about degradomics.

In these last years, the field of degradomics has experienced a remarkable expansion, not so much in the number of known proteases as in the biological and pathological roles played by the degradome. Thus, almost all of the growth in the number of known proteases has developed through the inclusion of new protein families which were not previously known to display proteolytic activity. This suggests that the initial annotation of proteases in these model genomes was highly successful. However, the number of known degradomopathies has undergone a sharp increase (>50%) in the last six years. In this manuscript, we present the updates to the Degradome database, including new interactive representations which, in our opinion, provide a useful global depiction of the degradomes.

DATABASE ACCESS

Annotation of individual proteases

The overall organization of the database remains unchanged, with information about five catalytic classes (aspartyl-, cysteine-, metallo-, serine- and threonine-proteases) encompassing 82 protease families in four species. Compared to the previous version of this database, we have annotated seven additional families with 18 new human and chimpanzee proteases, 21 new mouse proteases and 17 new rat proteases (Figure 1). This growth reflects a body of biochemical research by multiple laboratories which has uncovered previously unknown proteolytic activities in known proteins (17–19).

The first mode of access is contained in five tables, one for each catalytic class. Each table displays the names of the protease families using the MEROPS classification system, the name of each protease, and the gene symbol for the protease in each species (Figure 2A). The table directly represents the orthology between proteases in different species, as well as the pseudogenes for which a functional ortholog exists in at least one of the selected mammalian species. Clicking each table cell gives the user access to additional information. Thus, the family name leads to a summary containing selected publications about that family of proteases, and, when available, a description of its structural features. In addition, protease-specific links open *popup* tables with information about the status and activity of each protease and, if available, hyperlinks to other databases, including MEROPS (16), NCBI Gene (20) and Ensembl (21). These tables also show if the protease is involved in a degradomopathy, along with a link to the OMIM database entry describing the disease.

Hereditary diseases of proteolysis

The information about mutated proteases in hereditary diseases is kept into a separate table (http://degradome.uniovi.es/diseases.html), to bypass the need to browse or search the individual annotations. This table of degradomopathies contains information about causal gene locus, mode of inheritance, pathologic protease alteration (gain/loss of proteolytic activity), and availability of described animal models containing the same protease anomaly. A link to related OMIM entries is also provided.

To our knowledge, this remains the only summary of the relationships between degradomics and pathology. In six years, the number of entries has grown from 77 to 119, reflecting the intense research into the pathological implications of the degradome. In fact, several proteases, such as ADAM10 and AFG3L2, have been related to more than one hereditary disease through different alterations. This table does not reflect the whole contribution of the degradome to human disease, as it does not encompass the numerous examples of non-hereditary diseases in which proteases are known to play an important role through alterations in their spatio-temporal patterns of expression. In this regard, the Degradome database has also demonstrated its usefulness in the analysis of proteases associated with cancer (22,23) and ageing (4,24).

Graphic interface

Ever since its definition, a recurrent global representation of the degradome has featured a characteristic circular drawing (Figure 2B). In this new version of the Degradome database, we have developed a tool that depicts the whole database as an svg figure. In addition to being customizable and constantly updated, this format allows a fair level of interaction with the user. Thus, each individual protease object contains a hyperlink to the corresponding annotation table. This figure can also be modified to represent the results of a search.

More importantly, this figure provides a global view of the degradomes of the included species. This not only provides a sense of the sizes of protease families, but also a fast and intuitive comparison between the degradomes of different species, including family expansions (i.e. mouse and rat expansions of the S01 family) and pseudogenization events. This interactive figure is used at this time in the home page

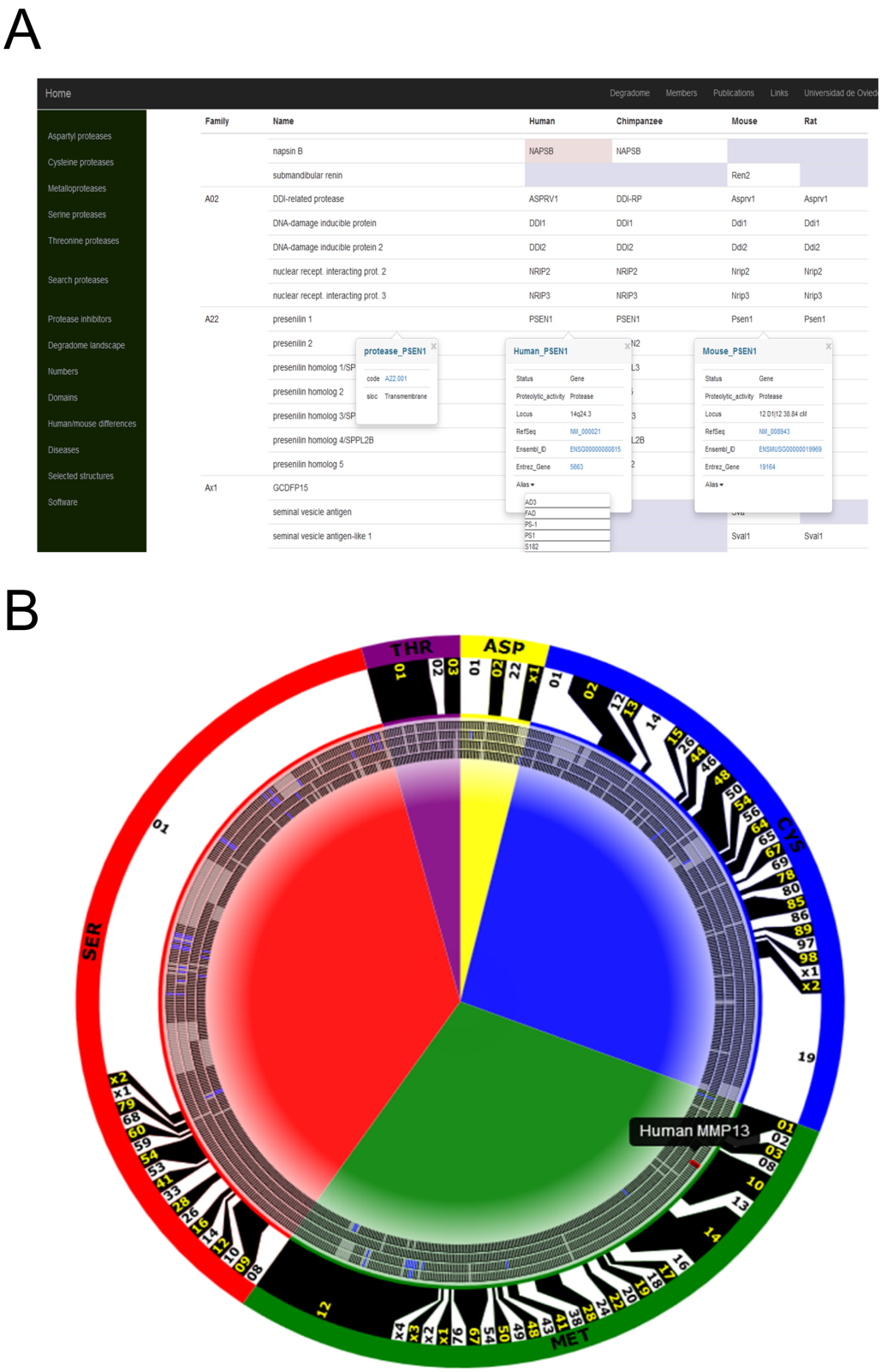

Figure 2. New features of the Degradome database. (**A**) Individual annotations of aspartyl proteases. The first column contains the name of the family, with a hyperlink to a web page where the user can find related selected publications and structures. The second column contains the name of each protease, with a hyperlink which opens a *popup* table with further general information—in this example, *presenilin 1*. Pseudogenes are shown over a *pink background*. Proteases absent in a species are shown as *empty grey cells*. (**B**) Interactive representation of the degradomes of (from outer to inner ticks) human, chimpanzee, mouse and rat. Protease families are limited with *black* and *white boxes*. Catalytic classes are shown as background *colored arcs*. Proteases which have been pseudogenized are depicted as *blue ticks*, and proteases absent in an organism are shown as *grey ticks*. Human collagenase-3 is highlighted to show the interactivity of the ticks.

(http://degradome.uniovi.es/dindex.html), as an alternative entry method for the database, and as a snapshot of the differences between the degradomes of human and mouse (http://degradome.uniovi.es/hmd.html).

Additional contents

In addition to the Degradome database, the web site also keeps offering several summaries of the characteristics of mammalian degradomes. Thus, a static table listing human, mouse and rat protease inhibitors can be found at http://degradome.uniovi.es/inhibitors.html. A count of proteases in these species, itemized by catalytic class, is shown at http://degradome.uniovi.es/numbers.html. These numbers are kept updated as novel catalytic classes are discovered and added to the Degradome database. Additionally, we also keep a figure showing the different ancillary domains present in proteases (http://degradome.uniovi.es/domains.html). Interactive structures for different protease families are also kept in pdf format for teaching purposes (http://degradome.uniovi.es/structures.html).

IMPLEMENTATION

The database with the annotations of individual proteases has been migrated to a single JSON file, which is freely available upon request. The information is queried from the web interface using the AJAX technology through JQuery. Therefore, if the browser lacks Javascript or Javascript is blocked, the user is offered a link to a static table displaying all of the information at once. The style of the web pages is implemented with the Bootstrap library to increase accessibility from multiple devices. The new graphical interface is written in svg, which is directly generated from the degradome JSON file using a custom Perl script.

Finally, *selected structures* are displayed in pdf format. Thus, the user needs Adobe Reader v7.0 or higher. Several reasons may hamper the viewing for these files from the common browsers. If this happens, the user can download the pdf file and access its contents locally.

CONCLUSION AND FUTURE DIRECTION

The Degradome database has grown in the last years reflecting the advances in our understanding of proteases in biological and pathological processes. As a part of our involvement in degradomics, we will continue updating this database as more results become available. Based on our experience, we expect this increase to continue, driven mainly by two sources: biochemical studies that uncover novel proteolytic activities and high-throughput association studies which relate additional protease mutations with hereditary diseases.

FUNDING

European Union [CancerDegradome-FP6 and MicroEnviMet-FP7]; Ministerio de Economía y Competitividad-Spain, Principado de Asturias and Fundación Botín (supported by Banco Santander through its Santander Universities Global Division). Funding for open access charge: Ministerio de Economía y Competitividad-Spain.

Conflict of interest statement. None declared.

REFERENCES

1. Lopez-Otin,C. and Overall,C.M. (2002) Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell. Biol.*, **3**, 509–519.
2. Lopez-Otin,C. and Bond,J.S. (2008) Proteases: multifunctional enzymes in life and disease. *J. Biol. Chem.*, **283**, 30433–30437.
3. Reinhard,S.M., Razak,K. and Ethell,I.M. (2015) A delicate balance: role of MMP-9 in brain development and pathophysiology of neurodevelopmental disorders. *Front. Cell Neurosci.*, **9**, 280.
4. Quiros,P.M., Langer,T. and Lopez-Otin,C. (2015) New roles for mitochondrial proteases in health, ageing and disease. *Nat. Rev. Mol. Cell. Biol.*, **16**, 345–359.
5. Voskoboinik,I., Whisstock,J.C. and Trapani,J.A. (2015) Perforin and granzymes: function, dysfunction and human pathology. *Nat. Rev. Immunol.*, **15**, 388–400.
6. Turk,B., Turk,D. and Turk,V. (2012) Protease signalling: the cutting edge. *EMBO J.*, **31**, 1630–1643.
7. Nalivaeva,N.N., Belyaev,N.D., Kerridge,C. and Turner,A.J. (2014) Amyloid-clearing proteins and their epigenetic regulation as a therapeutic target in Alzheimer's disease. *Front. Aging Neurosci.*, **6**, 235.
8. Gordon,L.B., Rothman,F.G., Lopez-Otin,C. and Misteli,T. (2014) Progeria: a paradigm for translational medicine. *Cell*, **156**, 400–407.
9. Gutierrez-Fernandez,A., Soria-Valles,C., Osorio,F.G., Gutierrez-Abril,J., Garabaya,C., Aguirre,A., Fueyo,A., Fernandez-Garcia,M.S., Puente,X.S. and Lopez-Otin,C. (2015) Loss of MT1-MMP causes cell senescence and nuclear defects which can be reversed by retinoic acid. *EMBO J.*, **34**, 1875–1888.
10. Lopez-Otin,C. and Hunter,T. (2010) The regulatory crosstalk between kinases and proteases in cancer. *Nat. Rev. Cancer*, **10**, 278–292.
11. Puente,X.S., Sanchez,L.M., Overall,C.M. and Lopez-Otin,C. (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.*, **4**, 544–558.
12. Puente,X.S., Sanchez,L.M., Gutierrez-Fernandez,A., Velasco,G. and Lopez-Otin,C. (2005) A genomic view of the complexity of mammalian proteolytic systems. *Biochem. Soc. Trans.*, **33**, 331–334.
13. Ordonez,G.R., Puente,X.S., Quesada,V. and Lopez-Otin,C. (2009) Proteolytic systems: constructing degradomes. *Methods Mol. Biol.*, **539**, 33–47.
14. Quesada,V., Ordonez,G.R., Sanchez,L.M., Puente,X.S. and Lopez-Otin,C. (2009) The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res.*, **37**, D239–243.
15. Igarashi,Y., Heureux,E., Doctor,K.S., Talwar,P., Gramatikova,S., Gramatikoff,K., Zhang,Y., Blinov,M., Ibragimova,S.S., Boyd,S. *et al.* (2009) PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res.*, **37**, D611–D618.
16. Rawlings,N.D., Waller,M., Barrett,A.J. and Bateman,A. (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, **42**, D503–D509.
17. Bujalka,H., Koenning,M., Jackson,S., Perreau,V.M., Pope,B., Hay,C.M., Mitew,S., Hill,A.F., Lu,Q.R., Wegner,M. *et al.* (2013) MYRF is a membrane-associated transcription factor that autoproteolytically cleaves to directly activate myelin genes. *PLoS Biol.*, **11**, e1001625.
18. Liang,J., Saad,Y., Lei,T., Wang,J., Qi,D., Yang,Q., Kolattukudy,P.E. and Fu,M. (2010) MCP-induced protein 1 deubiquitinates TRAF proteins and negatively regulates JNK and NF-kappaB signaling. *J Exp. Med.*, **207**, 2959–2973.
19. Stingele,J., Habermann,B. and Jentsch,S. (2015) DNA-protein crosslink repair: proteases as DNA repair enzymes. *Trends Biochem. Sci.*, **40**, 67–71.
20. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–42.
21. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.

22. Fraile,J.M., Quesada,V., Rodriguez,D., Freije,J.M. and Lopez-Otin,C. (2012) Deubiquitinases in cancer: new functions and therapeutic options. *Oncogene*, **31**, 2373–2388.
23. Sevenich,L. and Joyce,J.A. (2014) Pericellular proteolysis in cancer. *Genes Dev.*, **28**, 2331–2347.
24. Fernandez,A.F. and Lopez-Otin,C. (2015) The functional and pathologic relevance of autophagy proteases. *J. Clin. Invest.*, **125**, 33–41.

The Novel Evolution of the Sperm Whale Genome

Wesley C. Warren[1,*], Lukas Kuderna[2], Alana Alexander[3], Julian Catchen[4], José G. Pérez-Silva[5], Carlos López-Otín[5], Víctor Quesada[5], Patrick Minx[1], Chad Tomlinson[1], Michael J. Montague[6], Fabiana H.G. Farias[1], Ronald B. Walter[7], Tomas Marques-Bonet[2], Travis Glenn[8], Troy J. Kieran[8], Sandra S. Wise[9], John Pierce Wise Jr [9], Robert M. Waterhouse[10], and John Pierce Wise Sr [9]

[1]McDonnell Genome Institute, Washington University, St Louis

[2]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain

[3]Biodiversity Institute, University of Kansas

[4]Department of Animal Biology, University of Illinois, Urbana

[5]Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, Spain

[6]Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania

[7]Department of Chemistry and Biochemistry, Texas State University

[8]Department of Environmental Health Science, University of Georgia, Environmental Health Science Bldg, Athens, Georgia

[9]Wise Laboratory of Environmental and Genetic Toxicology, Department of Pharmacology and Toxicology, School of Medicine, University of Louisville

[10]Department of Ecology and Evolution, University of Lausanne, Switzerland

*Corresponding author: E-mail: wwarren@wustl.edu.

Accepted: September 12, 2017

Data deposition: This project has been deposited at NCBI under the Bioproject accession numbers PRJNA89089 and PRJNA237226.

Abstract

The sperm whale, made famous by *Moby Dick*, is one of the most fascinating of all ocean-dwelling species given their unique life history, novel physiological adaptations to hunting squid at extreme ocean depths, and their position as one of the earliest branching toothed whales (Odontoceti). We assembled the sperm whale (*Physeter macrocephalus*) genome and resequenced individuals from multiple ocean basins to identify new candidate genes for adaptation to an aquatic environment and infer demographic history. Genes crucial for skin integrity appeared to be particularly important in both the sperm whale and other cetaceans. We also find sperm whales experienced a steep population decline during the early Pleistocene epoch. These genomic data add new comparative insight into the evolution of whales.

Key words: sperm whale, cetaceans, genome.

Introduction

The sperm whale, made famous by *Moby Dick*, makes some of the deepest and longest dives of any marine mammal: >73 min long and up to 2,035 m deep (Watkins et al. 1993; Watwood et al. 2006) to feed on squid, including the infamous giant and colossal squids (Best 1979; Whitehead 2003). Previous comparative genomic analyses of cetaceans indicated genic adaptation to a marine existence (Foote et al. 2015; Yim et al. 2014), including convergent pathways of metabolism regulation for deep diving (Foote et al. 2015). However, to date, the sperm whale—one of the deepest diving and earliest branching toothed whales (Odontoceti; Whitehead 2003)—has been excluded from these comparisons. We sequenced and assembled multiple sperm whale genomes to explore genic adaptation. Given the important and broad physiological roles played by proteases, our explorations mostly focused on examining protease loss-of-function (LoF) events important in sperm whale, and cetacean, evolution. We also sought to discover which genes showed signs of positive selection shared with other cetaceans or unique to sperm whale. Finally, as previous analyses suggested that sperm whales

Table 1
Genes under Positive Selection Enriched by Pathway, Phenotype, or Protein Interactions

Pathway	Source	Genes	Ratio of Enrichment	Adjusted *P*-value
Focal adhesion	Wiki	*CHAD, COL1A2, THBS2, TNC, FLT1*	6.9	0.015
Focal adhesion	KEGG	*CHAD, COL1A2, PARVG, THBS2, TNC, FLT1*	7.5	0.0076
Pemphigus	Disease	*PPL, EVPL, DSP, DSG3*	38.8	0.0006
Calcium signaling	KEGG	*PTK2B, ADCY3, RYR1, NOS2, P2RX3, PHKB*	7.6	0.0076
Blood circulation	GO	*ALOX5, CHD7, WNK1, EPAS1, CX3CL1, PPP1R13L, COL1A2, AZU1, DSP, GUCY1A3, MYBPC3, TBC1D8*	3.4	0.036
Cornified envelope	Reactome	*DSP, TGM1, KRT4, PKP1, DSG3, PPL, EVPL*	NA	[a]FDR 3.65×10^{-11}

[a]The false discovery rate (FDR) calculated within the Reactome software (Croft et al. 2014) is the probability corrected for multiple comparisons. Adjusted *P* values are not provided.

experienced a global expansion <80,000 years ago (Alexander et al. 2016), we examine the estimated historical effective population size using samples from throughout the sperm whale's range.

Materials and Methods

We sequenced a Gulf of Mexico female sperm whale (GMX) to high coverage (72×) using short-insert and mate-pair libraries of 100 bp length (detailed in the supplementary material S1, Supplementary Material online) on an Illumina HiSeq2000. We assembled the draft genome of all sequences with ALLPATHS (Gnerre et al. 2011) using default parameter settings, subjecting assembly input reads to quality control as detailed in the ALLPATHS documentation (Gnerre et al. 2011). We obtained RNAseq data from skin biopsies of a different GMX sperm whale to aid gene annotation as described in the supplementary material S1, Supplementary Material online. Gene annotation was performed according to the NCBI gene annotation pipeline as described here: http://www.ncbi.nlm.nih.gov/books/NBK169439/. After aligning genes from the sperm whale with other taxa (detailed in supplementary material S1, Supplementary Material online) to establish 1:1 gene orthology, positive selection was detected using PAML4.0 (Yang 2007) and impact on protein structure tested with Provean (Choi and Chan 2015). Canonical pathway enrichment of gene clusters under positive selection was established as detailed in the supplementary material S1, Supplementary Material online. Protease genes were manually annotated and validated for loss/duplication events using BATI (http://degradome.uniovi.es/downloads.html). Four additional sperm whale individuals (supplementary table S1, Supplementary Material online) were resequenced to moderate depth (21–28×) and reads were mapped to the draft genome as described in the supplementary material S1, Supplementary Material online. We calculated heterozygosity on a per-individual basis using VCFtools (Danecek et al. 2011). Effective population size was reconstructed with PSMC (Li and Durbin 2011) using the parameters specified in the supplementary material S1, Supplementary Material online.

Results and Discussion

Our sperm whale total assembled sequence was similar in size to other assembled cetacean genomes (supplementary table S2, Supplementary Material online). Using our GMX individual reference assembly (Genbank assembly accession GCA_000472045.1) we inferred 18,686 protein-coding genes—second only to the baiji (*Lipotes vexillifer*) among sequenced cetaceans at 18,906 genes. Using a core eukaryotic mapping method (Simao et al. 2015) we also demonstrated >94.7% of conserved genes were complete in our assembly (supplementary table S3, Supplementary Material online). Of the 18,686 protein-coding genes, 12,717 had single-copy orthologs in both human and other cetartiodactyls (supplementary table S4, Supplementary Material online; additional methods/results can be found in the supplementary material S1, Supplementary Material online). A total of 45 genes found across eight taxa were identified as being under positive selection in the sperm whale lineage; these genes also passed our stringent functional impact tests (default cutoff < -2.5) using Provean (Choi and Chan 2015; supplementary file S1, Supplementary Material online). Several significant pathways emerged from enrichment analyses, which included genes associated with blood-circulation and skin stress responses (table 1). Cetaceans, including the sperm whale, exhibit molting or skin sloughing (Amos et al. 1992), potentially as an adaptive response to fouling by barnacles and other organisms. However, sperm whales face the additional challenge of maintaining skin integrity and blood homeostasis at high water pressures during deep foraging dives.

To complement the analysis of genes under positive selection, we manually annotated the complete set of proteases (i.e., degradome) of the sperm whale. This independent analysis identified several proteases involved in skin function and blood homeostasis that showed LoF events along the lineage leading to sperm whales (fig. 1; additional methods/results in supplementary material S1, Supplementary Material online). We also detected LoF in proteases involved in inflammation, immunity and metabolism within cetaceans, and specifically

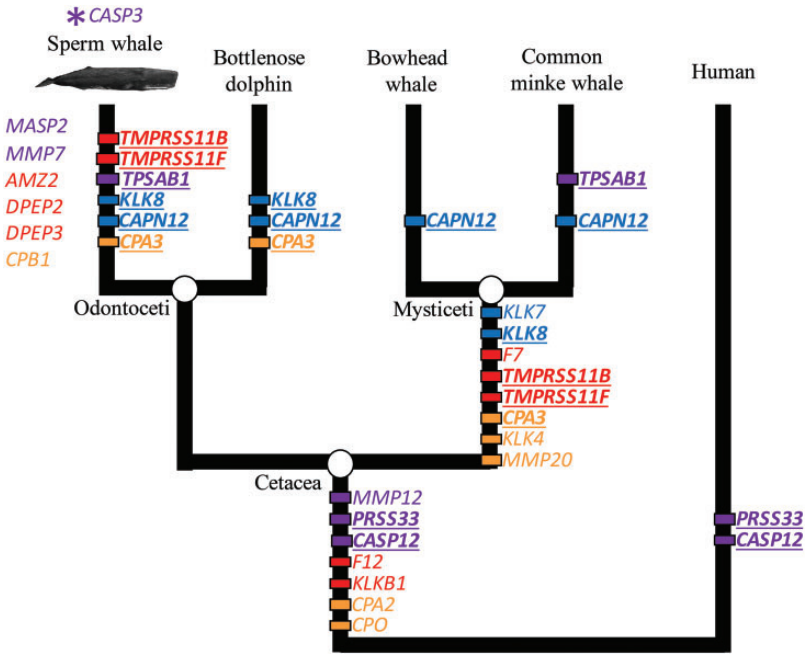

Fig. 1.—Cetacean-specific losses of protease genes. Proteases that have undergone loss-of-function in sperm whales, specifically, are shown to the left of the phylogeny whereas those that are inferred to be convergent, or inferred to have occurred in ancestral lineages, are mapped on to the phylogeny. Each event is depicted along the branch where loss events have been inferred to occur. Genes expected to impact skin function are colored blue; immune system: purple; blood homeostasis: red; digestion: orange, and those showing convergent loss-of-function as underlined bold. The unique duplication of sperm whale *CASP3* is shown above the phylogeny and marked by an asterisk.

within the sperm whale. A loss of several proteases in cetaceans suggests a trend towards a milder inflammatory response relevant to Peto's paradox: A theory postulated to explain the lower relative incidence of cancer in large mammals (Caulin and Maley 2011). In addition, *MMP7*—which promotes metastasis when expressed at high levels (Li et al. 2014; Koskensalo et al. 2011)—contains a premature stop codon in sperm whales, a putative sperm whale-specific mechanism to reduce cancer incidence. We also found that *CASP12* and *PRSS33* were independently lost in cetaceans and some hominoids, suggesting a case of convergent evolution of the immune system in very different environments. Several proteases involved in digestion (*CPA2*, *CPA3*, *CPO*) were also lost in cetaceans (fig. 1). In some cases these losses were independent, suggesting convergent evolution driven by trophic level. As expected, odontocetes retain functional orthologs of proteases involved in dentition (*KLK4*, *MMP20*), which were lost in mysticetes, who use baleen—not teeth—to filter food (Keane et al. 2015).

To better understand patterns of genetic diversity among sperm whales from different ocean basins, we carried out medium-coverage resequencing of individuals from the Pacific Ocean and Indian Ocean. Average genome-wide heterozygosity per base, corrected for callable sequence space, was 0.0011. This value is low in comparison with the fin whale (0.0015) and bottlenose dolphin (0.0014; Yim et al. 2014), suggesting the sperm whale has a smaller effective population size (N_e). A pairwise sequentially Markovian coalescent (PSMC) analysis (Li and Durbin 2011) indicated a rapid decline in N_e during the transition from the Pliocene to Pleistocene epoch, inferred consistently regardless of the ocean origin of samples (fig. 2*A*). The increase in upwelling associated with the Pliocene and/or cycles of glaciation within the Pleistocene have been implicated in the evolution of gigantism in mysticetes (Slater et al. 2017), as well as the diversification of marine dolphins (do Amaral et al. 2016). This suggests that changes in ocean dynamics during this time period have had a strong impact on cetaceans in general, and we suggest are also the likely cause of the inferred sperm whale population decline. The GMX sample had significantly lower heterozygosity than Pacific and Indian Ocean samples (fig. 2*B*, supplementary table S1, Supplementary Material online). Future sequencing will clarify whether lower diversity is restricted to GMX, or characteristic of the entire Atlantic. However, the isolation of GMX due to high levels of female philopatry (inferred from differentiation of the maternally-inherited mitochondrial DNA, Engelhaupt et al. 2009; Alexander et al. 2016), and the limited census size (763 sperm whales in 2009, Waring et al. 2013), suggest that GMX could be subjected to greater levels of genetic drift associated with a small and maternally-isolated population. The ability of the sperm whale to respond to future selective pressures, including climate change, in the face of such reduced genetic diversity should be a focus of ongoing study.

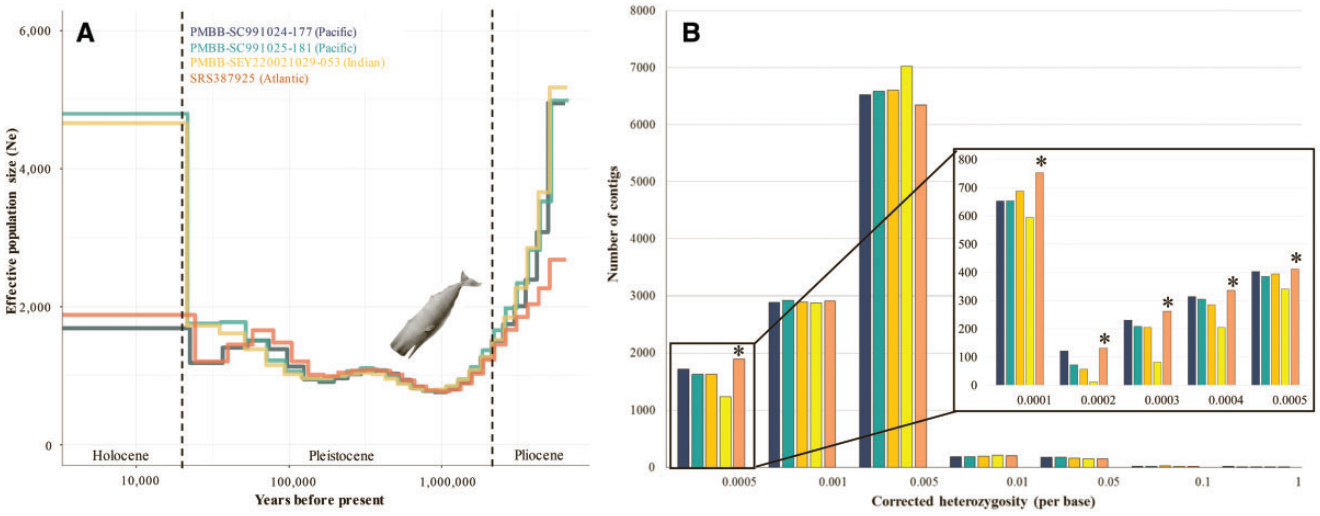

Fig. 2.—Estimated effective population size history and heterozygosity of sperm whales from different ocean basins. Samples are color coded by the key, with blue/green = Pacific, orange/yellow = Indian Ocean, and dark orange = Gulf of Mexico, Atlantic. (*A*) PSMC reconstruction of effective population size through time by sample (excluding SEY420021031-063, see supplementary material S1, Supplementary Material online), dashed lines represent the estimated start dates for each epoch; (*B*) Genome wide distribution of heterozygosity for each sample, by contig/scaffold. The Gulf of Mexico sample—characterized by low heterozygosity—is marked by an asterisk where it has the largest number of contigs in a category. The insert emphasizes that this sample has the largest number of contigs with low heterozygosity (<0.0005). Bright yellow in panel (b) is additional Indian ocean sample

Overall, our results suggest positive selection has differentially affected localized portions of the sperm whale genome. In particular, the complex pattern of convergent gene evolution involving skin-related genes suggests they have played an important role in aquatic adaptation, possibly influenced by the somewhat contradictory requirements of heat insulation, buoyancy and deep diving. In comparison to the localized effects of selection on the genome, we infer that the sperm whale experienced a rapid population decline, potentially in response to glaciation, which had a broad effect on genome-wide diversity. Given the apparent influence of past climate change, monitoring the on-going response of sperm whales to anthropogenically mediated climate change will be paramount.

Authors Contributions

S.S.W., J.W.J., J.W.S. isolated genomic DNA and sexed all samples. L.K., T.M.B. performed all population history analyses. J.G.P., C.L.O., V.Q. performed all protease analyses. C.T., P.M. completed genome assembly and curation. J.C., T.J.K., T.G. analyzed population sequences. M.J.M., F.H.F. analyzed gene selection. R.M.W. completed all gene orthology analyses. W.C.W., A.A. wrote the paper and reviewed all analyses. All authors have read and approved the manuscript.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the National Institute of Health grant 2R24OD011198-04A1 (W.C.W., PI); National Institute of Environmental Health Sciences [ES016893 (J.W. Sr, PI)]; Army Research Office [W911NF-09-1-0296 (J.W. Sr, PI)]; Swiss National Science Foundation grant PP00P3_170664 (R.M.W.); Ministerio of Economia and Competitividad (Spain); and European Union (ERC-Advanced Grant DeAge) (C.L.-O., PI); the Campbell Foundation; the Ocean Foundation; Ocean Alliance; and the many individual and anonymous Wise Laboratory donors. The authors declare no competing financial interests. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences, the National Institutes of Health, the Army Research Office or the Department of Defense. Work was conducted under National Marine Fisheries Service permit #1008-1637-03 (J.W. Sr, PI) and permit #751-1614 (I.K., PI). We thank Kyung Kim for computational support. We would like to thank Catherine Wise, James Wise, Chris Gianios, and all the Wise Laboratory/*Odyssey* science team volunteers for their help with technical support, whale spotting and sample collection. We thank Iain Kerr, Roger Payne, Bob Wallace, Derek Walker, and all of the *Odyssey* boat crew for their help with sample collection and logistics. We thank C. Scott Baker for comments on this manuscript. Finally, we thank all the volunteers who supported us in this project.

Literature Cited

Alexander A, et al. 2016. What influences the worldwide genetic structure of sperm whales (*Physeter macrocephalus*)? Mol Ecol. 25(12):2754–2772.

Amos W, et al. 1992. Restrictable DNA from sloughed cetacean skin; its potential for use in population analysis. Mar Mamm Sci. 8(3):275–283.

Best PB. 1979. Social organization in sperm whales, Physeter macrocephalus. In Behavior of marine animals. pp. 227–289. Springer US.

Caulin AF, Maley CC. 2011. Peto's Paradox: evolution's prescription for cancer prevention. Trends Ecol Evol. 26(4):175–182.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31(16):2745–2747.

Croft D, et al. 2014. The Reactome pathway knowledgebase. Nucleic Acids Res. 42(Database issue):D472–D477.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

do Amaral KB, Amaral AR, Fordyce RE, Moreno IB. 2016. Historical biogeography of delphininae dolphins and related taxa (Artiodactyla: Delphinidae). J Mamm Evol. 1–19. https://link.springer.com/article/10.1007/s10914-016-9376-3

Engelhaupt D, et al. 2009. Female philopatry in coastal basins and male dispersion across the North Atlantic in a highly mobile marine species, the sperm whale (*Physeter macrocephalus*). Mol Ecol. 18(20):4193–4205.

Foote AD, et al. 2015. Convergent evolution of the genomes of marine mammals. Nat Genet. 47(3):272–275.

Gnerre S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 108(4):1513–1518.

Keane M, et al. 2015. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. 10(1):112–122.

Koskensalo S, Louhimo J, Nordling S, Hagström J, Haglund C. 2011. MMP-7 as a prognostic marker in colorectal cancer. Tumour Biol. 32(2):259–264.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475(7357):493–496.

Li Z, et al. 2014. Prediction of peritoneal recurrence by the mRNA level of CEA and MMP-7 in peritoneal lavage of gastric cancer patients. Tumour Biol. 35(4):3463–3470.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.

Slater GJ, Goldbogen JA, Pyenson ND. 2017. Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. Proc Biol Sci. 284(1855):20170546.

Waring GTJE, Maze-Foley K, Rosel PE. 2013. NOAA Tech Memo NMFS-NE-223. (ed. NOAA).

Watkins WA, Daher MA, Fristrup KM, Howald TJ, di Sciara GN. 1993. Sperm whales tagged with transponders and tracked underwater by sonar. Mar Mamm Sci. 9(1):55–67.

Watwood SL, Miller PJ, Johnson M, Madsen PT, Tyack PL. 2006. Deep-diving foraging behaviour of sperm whales (*Physeter macrocephalus*). J Anim Ecol. 75(3):814–825.

Whitehead H. 2003. Sperm whales: social evolution in the ocean. Chicago: The University of Chicago Press.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Yim HS, et al. 2014. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 46(1):88–92.

Associate editor: Tal Dagan

Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty109
Advance Access Publication Date: 23 February 2018
Applications Note

OXFORD

Data and text mining

nVenn: generalized, quasi-proportional Venn and Euler diagrams

José G. Pérez-Silva, Miguel Araujo-Voces and Víctor Quesada*

Departamento de Bioquímica y Biología Molecular, Universidad de Oviedo, Oviedo 33006, Spain

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 7, 2017; revised on February 20, 2018; editorial decision on February 21, 2018; accepted on February 21, 2018

Abstract

Motivation: Venn and Euler diagrams are extensively used for the visualization of relationships between experiments and datasets. However, representing more than three datasets while keeping the proportions of each region is still not feasible with existing tools.

Results: We present an algorithm to render all the regions of a generalized n-dimensional Venn diagram, while keeping the area of each region approximately proportional to the number of elements included. In addition, missing regions in Euler diagrams lead to simplified representations. The algorithm generates an n-dimensional Venn diagram and inserts circles of given areas in each region. Then, the diagram is rearranged with a dynamic, self-correcting simulation in which each set border is contracted until it contacts the circles inside. This algorithm is implemented in a C++ tool (nVenn) with or without a web interface. The web interface also provides the ability to analyze the regions of the diagram.

Availability and implementation: The source code and pre-compiled binaries of nVenn are available at https://github.com/vqf/nVenn. A web interface for up to six sets can be accessed at http://degradome.uniovi.es/cgi-bin/nVenn/nvenn.cgi.

Contact: quesadavictor@uniovi.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A recurrent task in data mining is set visualization (Alsallakh *et al.*, 2016). Ideally, the aim of this analysis is to find the most important relationships between datasets at a glance (Lex *et al.*, 2014). Venn and Euler diagrams are a popular tool for this purpose, as they represent in a single figure all the relevant overlaps between sets. Venn diagrams are similar to Euler representations, but they show all the possible intersections between sets, even if they do not exist in the input. In the field of bioinformatics, sets can for instance contain genes that are differentially expressed in multiple conditions. Similarities in the response to those conditions will be immediate apparent as intersections containing a larger-than-expected number of elements. For this reason, making the area of each region proportional to the number of elements it contains is particularly useful.

Multiple tools exist for the automatic creation of Euler diagrams, reflecting the extensive use of this representation in research. Most of these tools represent up to three sets, and keep the regions approximately proportional to the number of elements (e. g., Hulsen *et al.*, 2008, Micallef and Rodgers, 2014a). Representing more than three sets while keeping proportionality is not trivial, as symmetric set shapes are not flexible enough. Several tools accomplish approximate proportionality by using penalty functions or other transformations (e. g., Kestler *et al.*, 2008). However, most tools simply present a pre-drawn n-set Venn diagram with numbers inserted, which, while useful, is hard to interpret (e. g., Bardou *et al.*, 2014, Heberle *et al.*, 2015).

An intriguing development in the creation of proportional Euler diagrams has been used in eulerForce (Micallef and Rodgers, 2014b). The algorithm used inside this tool performs a physical simulation on a system which is attuned to generate the desired layout for an Euler diagram. Each curve enclosing a set is represented conceptually as a number of charges joined by springs. By manipulating the forces between those virtual charges, eulerForce creates Euler diagrams that are regular, smooth and aesthetically pleasing.

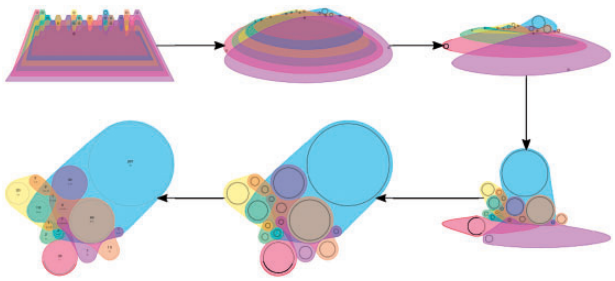

Fig. 1. nVenn algorithm. The figure is created through simulation from a generalized Venn diagram (top left) up to a quasi-static solution (bottom center). Then, the lines are softened to obtain the final figure (bottom left)

However, the areas of the resulting regions are not easy to manipulate, and using more than five curves is too computationally expensive. These limitations illustrate the complexity of representing multiple sets with symmetric or simple convex shapes and respect the proportionality of each region.

Here, we present an algorithm to create Venn and Euler diagrams with an arbitrary number of sets. Each region in the diagram contains a circle whose area is proportional to the number of elements in that intersection. We have also prepared a web interface to create proportional diagrams with up to six sets.

2 Materials and methods

The nVenn program, coded in C++, accepts a text input describing the sizes of each region in a Venn diagram and outputs a figure in SVG format. The steps of this algorithm are summarized in Figure 1. Briefly,

- A generalized Venn diagram for the desired number of sets (n) is generated based on a symmetric chain decomposition (Greene and Kleitman, 1976; Griggs *et al.*, 2004). First, the boolean lattice for n groups is generated based on the depiction in Ruskey *et al.* (2006). This lattice expresses each region as a boolean vector where each element represents a set. If that position is filled with a 1, the region belongs to the set, whereas if it is filled with a 0 it does not belong to the set. The symmetric chain decomposition ensures that all the regions belonging to a set can be enclosed with a simple curve while excluding all the regions not belonging to the set (Supplementary Material, Section 1.1).
- Then, each region is shrunk through simulation (Fig. 1, top center, top right, bottom right and bottom center). In this process, each line of the diagram is replaced by a large number of points joined by springs (Supplementary Material, Section 1.2). Internal circles move when contacted by line points and line springs. This system is simulated with a naïve engine based on a small delta time. It includes friction and damping forces to speed up the contraction of lines.
- Finally, the contracted diagram is embellished through a simulation where inner circles are fixed and an attractive spring force (Supplementary Material, Section 1.1) between line points and circles is added, so that lines represent each region more closely (Supplementary Movie S1, starting at 20 seconds).

The end of each step is controlled by the user in different ways, depending on the interface.

2.1 Interfaces

The current version of nVenn can be used with three different interfaces: command line, OpenGL graphical output and web interface. Although the core methods are the same, each flavor has distinct requirements and modes of use.

2.1.1 Command line

This version accepts a text input file describing each region in the diagram (Supplementary Fig. S1) and performs all three steps automatically. The number of cycles per step is fixed in the code, and fits most purposes for up to six sets. In the current version, the main step consists of 7000 cycles. The program automatically saves the intermediate result, so that more cycles can be added by simply repeating this procedure on the same input file. The final embellishment runs for 200 cycles. The syntax to run this version is:

```
./nVenn input_file [output_file_name=result]
```

By repeated execution, an unlimited number of sets can be processed. However, it must be noted that the time it takes for the simulation to complete grows quickly with the number of sets. This version uses standard libraries, and therefore it is easy to compile in most operative systems. x64 Linux Debian and Microsoft Windows pre-compiled versions are available for download.

2.1.2 Graphical output

A graphical version using OpenGL is also provided. The input for this tool has the same format as that of the command-line version, although the names of input and output files are fixed. By contrast, the user can decide in real time when to jump from one phase to the next. This interface also gives a simple overview of the simulation process, as shown in Supplementary Movie S1.

The OpenGL interface uses Microsoft Windows-specific libraries. A pre-compiled version for this platform is available for download.

2.1.3 Web interface

This version uses a more simple input and allows further analysis of the output diagram. Thus, users can directly enter the members of each set in text boxes. The interface then calculates the number of elements in each region of the diagram, runs nVenn and renders the output. In addition, users can query their data for any intersection between sets by checking boxes or by directly clicking the corresponding region in the output figure (Fig. 2). The length of the simulation is fixed, but more cycles can be added by repeated execution. At this time, the interface allows up to six sets, but the system is easily scalable to a higher number of sets.

The interface also allows the customization of the final figure. Users can add or remove labels to describe each region and to show how many elements are included in it. The color and opacity of each set, as well as the width of their borders, can also be tweaked. The output figure can be saved in a vectorial format (scalable vector graphics, SVG) and in a bitmap format (portable network graphics, PNG).

The web page interfacing nVenn is coded in standard HTML, CSS (using min. Bootstrap v.3.3.1) and javascript (using min. jQuery v.1.11.1) We have also added a step-by-step tutorial at the top navigation bar.

2.2 Test

As a proof of concept, we performed an Euler diagram with the web interface of nVenn using genes included in the *innate immune system* GO category (GO: 0045087). Six subsets with different GO evidence codes were generated: IBA (Inferred from Biological aspect of Ancestor), IC (Inferred by Curator), IDA (Inferred from Direct

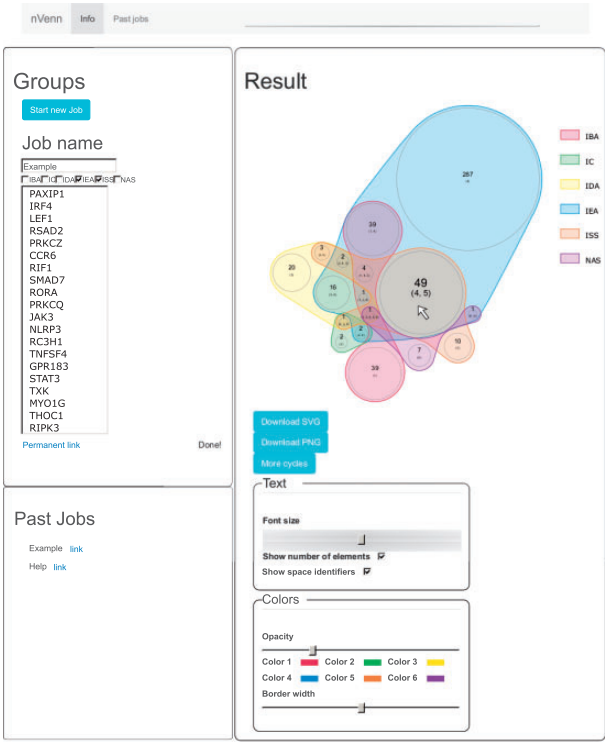

Fig. 2. Test nVenn diagram in web interface. The figure is shown at the right panel, with the tools provided for manipulation below. The text area at the top left panel shows the elements present only in sets 4 and 5. This region can be selected from the checkboxes (above the text area) and by clicking the corresponding area in the figure (arrow)

Assay), IEA (Inferred from Electronic Annotation), ISS (Inferred from Sequence or Structural Similarity) and NAS (Non-traceable Author Statement). The execution took three submissions for a satisfying result (about 12 min).

The result is shown in Figure 2, and gives a quick overview of the different sizes of each subset, as well as conspicuous correlations between them. Thus, the intersection of the IEA and ISS groups is so large as to make ISS almost a subset of IEA. Also, the number of elements shared only by those subsets is disproportionately large. Figure 2 also shows how users can analyze each region by just clicking on it.

3 Discussion

The aim of nVenn is to produce easy-to-interpret Euler diagrams that convey information about an unlimited number of sets. Although this capability exists, in practice it is very hard to interpret Venn diagrams for more than six sets.

The simulation-based algorithm for set drawing is conceptually similar to that of eulerForce, although the latter only simulates lines. In fact, the aim of eulerForce is building well-formed Euler diagrams. Achieving proportionality with this program would be very hard, requiring users to provide the coordinates of a starting Euler diagram and to perfectly balance the internal forces. By contrast, the use of inner circles in nVenn allows users to directly control the size of each region. This means that the resulting diagrams are not necessarily convex or strictly well-formed as with eulerForce. However, we have added specific steps to ease the interpretation of the results.

Thus, in the development of nVenn, some aesthetic qualities of the final diagram have taken precedence over strict proportionality. First, the area of each region is larger than the inner proportional circle, which produces smoother curves for each set. Therefore, users must take into account the areas of circles, and not empty spaces. In this regard, large numbers of sets with few intersections will frequently lead to diagrams with large empty spaces (Supplementary Fig. S2). Since this algorithm tackles a hard circle-packing problem with added constraints, this drawback is expected and accepted. Future versions of nVenn may incorporate random deviations in the initial conditions so that, after multiple runs, several solutions can be explored.

Further deviations from proportionality occur when some of the regions are too small in the final figure. To avoid invisible regions, the minimal circle radius is set at 1% of the width or height of the diagram. Circles whose radius should be lower than that will appear larger than expected. Since those regions would be even harder to interpret in a strict diagram, this caveat is also accepted. Finally, the set lines are separated by different distances from the inner circles. This feature minimizes the overlaps between lines so that each set line can be easily followed.

The web interface to nVenn offers additional tools for the analysis of diagrams, so that users can quickly find out which elements correspond to each region. We have designed this interface to be easy and intuitive, in the hope that it may become a valuable tool for researchers trying to visualize complex relationships between sets.

Acknowledgements

We thank Drs. Carlos López-Otín, Gloria Velasco and Magda R. Hamczyk for helpful discussions during the development of this manuscript.

Funding

This work has been supported by the Ministerio de Economía y Competitividad-Spain (SAF2014-59986-R, including FEDER funding, and Ramón y Cajal program), Instituto de Salud Carlos III and Principado de Asturias, including FEDER funding.

Conflict of Interest: none declared.

References

Alsallakh,B. *et al.* (2016) The state-of-the-art of set visualization. *Comput. Graph. Forum*, **35**, 234–260.

Bardou,P. *et al.* (2014) jvenn: an interactive venn diagram viewer. *BMC Bioinformatics*, **15**, 293.

Greene,C. and Kleitman,D.J. (1976) Strong versions of sperner's theorem. *J. Combin. Theory Ser. A*, **20**, 80–88.

Griggs,J. *et al.* (2004) Venn diagrams and symmetric chain decompositions in the boolean lattice. *Electronic J. Combin.*, **11**. Research Paper #R2.

Heberle,H. *et al.* (2015) Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC Bioinformatics*, **16**, 169.

Hulsen,T. *et al.* (2008) Biovenn – a web application for the comparison and visualization of biological lists using area-proportional venn diagrams. *BMC Genomics*, **9**, 488.

Kestler,H.A. *et al.* (2008) Vennmaster: area-proportional euler diagrams for functional go analysis of microarrays. *BMC Bioinformatics*, **9**, 67.

Lex,A. *et al.* (2014) Upset: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, **20**, 1983–1992.

Micallef,L. and Rodgers,P. (2014a) eulerape: drawing area-proportional 3-venn diagrams using ellipses. *PLoS One*, **9**, e101717.

Micallef,L. and Rodgers,P. (2014b) eulerForce: force-directed layout for euler diagrams. *J. Vis. Lang. Comput.*, **25**, 924–934.

Ruskey,F. *et al.* (2006) The search for simple symmetric venn diagrams. *Notices Am. Math. Soc.*, **53**, 1304–1312.

nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-018-0733-x

OPEN

Giant tortoise genomes provide insights into longevity and age-related disease

Víctor Quesada[1,19], Sandra Freitas-Rodríguez[1,19], Joshua Miller[2,19], José G. Pérez-Silva[1,19], Zi-Feng Jiang[3], Washington Tapia[4,5], Olaya Santiago-Fernández[1], Diana Campos-Iglesias[1], Lukas F. K. Kuderna[6,7], Maud Quinzin[2], Miguel G. Álvarez[1], Dido Carrero[1], Luciano B. Beheregaray[8], James P. Gibbs[9], Ylenia Chiari[10], Scott Glaberman[10], Claudio Ciofi[11], Miguel Araujo-Voces[1], Pablo Mayoral[1], Javier R. Arango[1], Isaac Tamargo-Gómez[1], David Roiz-Valle[1], María Pascual-Torner[1], Benjamin R. Evans[2], Danielle L. Edwards[12], Ryan C. Garrick[13], Michael A. Russello[14], Nikos Poulakakis[15,16], Stephen J. Gaughran[2], Danny O. Rueda[4], Gabriel Bretones[1], Tomàs Marquès-Bonet[6,7,17,18], Kevin P. White[3], Adalgisa Caccone[2*] and Carlos López-Otín[1*]

Giant tortoises are among the longest-lived vertebrate animals and, as such, provide an excellent model to study traits like longevity and age-related diseases. However, genomic and molecular evolutionary information on giant tortoises is scarce. Here, we describe a global analysis of the genomes of Lonesome George—the iconic last member of *Chelonoidis abingdonii*—and the Aldabra giant tortoise (*Aldabrachelys gigantea*). Comparison of these genomes with those of related species, using both unsupervised and supervised analyses, led us to detect lineage-specific variants affecting DNA repair genes, inflammatory mediators and genes related to cancer development. Our study also hints at specific evolutionary strategies linked to increased lifespan, and expands our understanding of the genomic determinants of ageing. These new genome sequences also provide important resources to help the efforts for restoration of giant tortoise populations.

Comparative genomic analyses leverage the mechanisms of natural selection to find genes and biochemical pathways related to complex traits and processes. Multiple works have used these techniques with the genomes of long-lived mammals to shed light on the signalling and metabolic networks that might play a role in regulating age-related conditions[1,2]. Similar studies on unrelated longevous organisms might unveil novel evolutionary strategies and genetic determinants of ageing in different environments. In this regard, giant tortoises constitute one of the few groups of vertebrates with an exceptional longevity: in excess of 100 years according to some estimates.

In this manuscript, we report the genomic sequencing and comparative genomic analysis of two long-lived giant tortoises: Lonesome George—the last representative of *Chelonoidis abingdonii*[3], endemic to the island of Pinta (Galapagos Islands, Ecuador)—and an individual of *Aldabrachelys gigantea*, endemic to the Aldabra Atoll and the only extant species of giant tortoises in the Indian Ocean[4] (Fig. 1a). Unsupervised and supervised comparative analyses of these genomic sequences add new genetic information on the evolution of turtles, and provide novel candidate genes that might underlie the extraordinary characteristics of giant tortoises, including their gigantism and longevity.

Results and discussion

The genome of Lonesome George was sequenced using a combination of Illumina and PacBio platforms (Supplementary Section 1.1). The assembled genome (CheloAbing 1.0) has a genomic size of 2.3 gigabases and contains 10,623 scaffolds with an N50 of 1.27 megabases (Supplementary Section 1.1 and Supplementary Tables 1–3). We also sequenced, with the Illumina platform, the closely related tortoise *A. gigantea* at an average read depth of 28×. These genomic sequences were aligned to CheloAbing 1.0.

TimeTree database estimations (http://www.timetree.org) indicate that Galapagos and Aldabra giant tortoises shared a last common ancestor about 40 million years ago, while both diverged from the human lineage more than 300 million years ago (Supplementary Section 1.4). A preliminary analysis of demographic history using the pairwise sequentially Markovian coalescent (PSMC)[5] model

[1]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología del Principado de Asturias, CIBERONC, Universidad de Oviedo, Oviedo, Spain. [2]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [3]Institute for Genomics and Systems Biology, The University of Chicago, Chicago, IL, USA. [4]Galapagos National Park Directorate, Galapagos Islands, Ecuador. [5]Galapagos Conservancy, Fairfax, VA, USA. [6]Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain. [7]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology , Barcelona, Spain. [8]College of Science and Engineering, Flinders University, Adelaide, South Australia, Australia. [9]College of Environmental Science and Forestry, State University of New York, Syracuse, NY, USA. [10]Department of Biology, University of South Alabama, Mobile, AL, USA. [11]Department of Biology, University of Florence, Florence, Italy. [12]School of Natural Sciences, University of California, Merced, CA, USA. [13]Department of Biology, University of Mississippi, Oxford, MS, USA. [14]Department of Biology, The University of British Columbia, Kelowna, British Columbia, Canada. [15]Department of Biology, School of Sciences and Engineering, University of Crete, Heraklion, Greece. [16]Natural History Museum of Crete, Heraklion, Greece. [17]Catalan Institution of Research and Advanced Studies, Barcelona, Spain. [18]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. [19]These authors contributed equally: Víctor Quesada, Sandra Freitas-Rodríguez, Joshua Miller, José G. Pérez-Silva. *e-mail: adalgisa.caccone@yale.edu; clo@uniovi.es

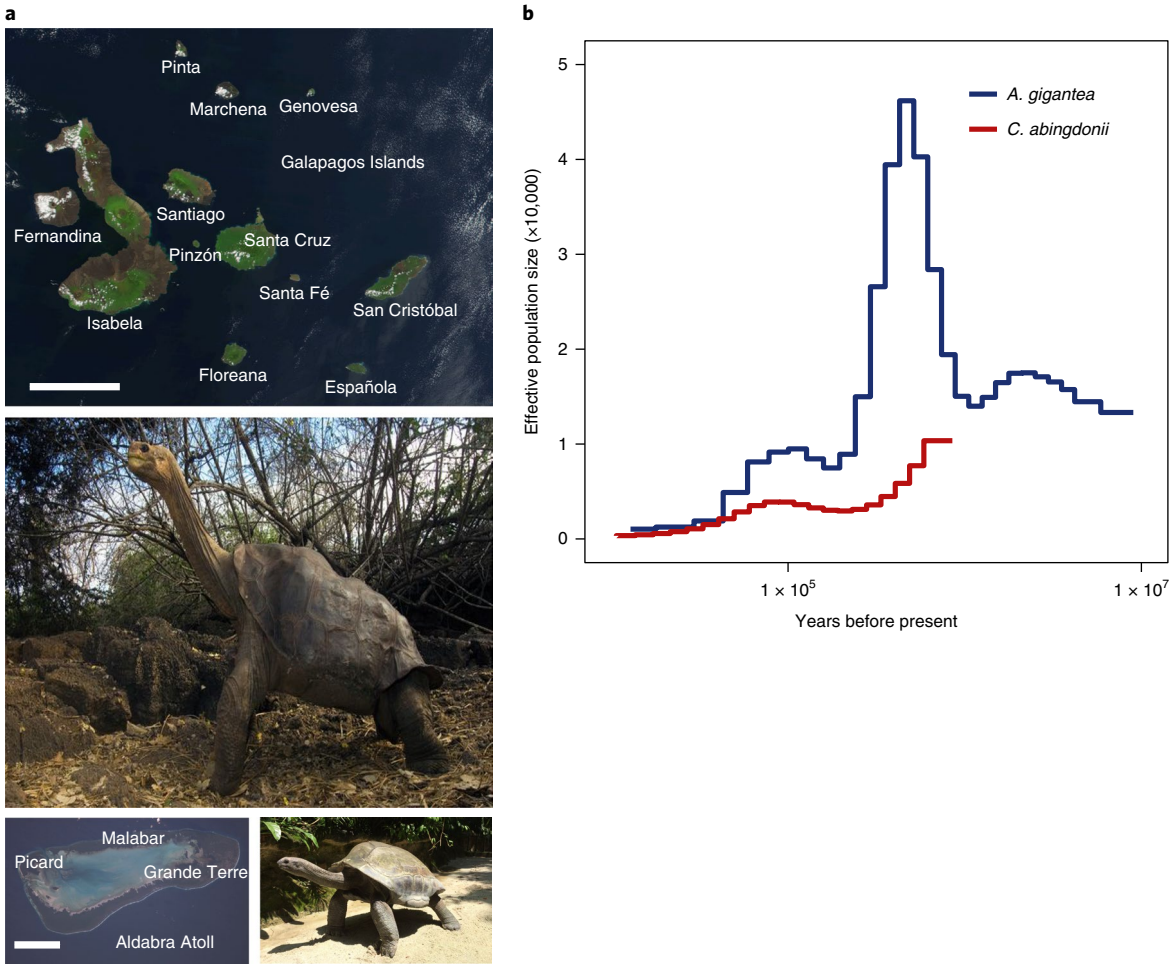

Fig. 1 | Geographical and temporal distribution of giant tortoises. a, Satellite view of the Galapagos Islands (top; scale bar: 50 km) and Aldabra Atoll (bottom left; scale bar: 10 km), and pictures of *C. abingdonii* (middle) and *A. gigantea* (bottom right). Both pictures are from http://eol.jsc.nasa.gov. **b**, Demographic history of giant tortoises, inferred using a hidden Markov model approach as implemented in the PSMC model. The default mutation rate (μ) for humans of 2.5×10^{-8} and an average generation time (g) of 25 years were used in the calculations.

showed that while the effective population size of *C. abingdonii* has been steadily declining for the past million years, with a slight uptick about 90,000 years ago, the population of Aldabra giant tortoises experienced substantial fluctuations over this period (Fig. 1b). Effective population size reconstructions for *C. abingdonii* lose statistical power at the million-year time frame, probably due to complete coalescence. In turn, this suggests that overall diversity in these giant tortoises must have been low throughout many generations. Together, these results prompt us to propose that the populations of these insular giant tortoises were vulnerable at the time of human discovery of the Galapagos Islands, probably elevating their extinction risk.

Using homology searches with known gene sets from humans and *Pelodiscus sinensis* (the Chinese soft-shell turtle), along with RNA sequencing (RNA-Seq) data from *C. abingdonii* blood and an *A. gigantea* granuloma, we automatically predicted a primary set of 27,208 genes from the genome assembly using the MAKER2 algorithm[6]. We then performed pairwise alignments between each of the primary predicted protein sequences and the UniProt databases for humans and *P. sinensis*, whose annotated sequences show relatively high quality when compared with data available for other turtles[7]. Using alignments spanning at least 80% of the longest protein and showing more than 60% identity, we constructed sets of protein families shared among these species. This preliminary analysis singled out several protein families that seem to have undergone moderate expansion in a common ancestor of *C. abingdonii* and *A. gigantea*. Almost all of these expansions were also confirmed in the genome of the related, long-lived tortoise *Gopherus agassizii* (Supplementary Section 1.2 and Supplementary Table 4). Most of these genes have been linked to exosome formation, suggesting that this process may have been important in tortoise evolution.

We also interrogated the predicted gene set for evidence of positive selection in giant tortoises. This analysis singled out 43 genes with evidence of giant-tortoise-specific positive selection (Supplementary Section 1.2, Supplementary Table 5 and Supplementary Fig. 1). This list includes genes with known roles in the dynamics of the tubulin cytoskeleton (*TUBE1* and *TUBG1*) and intracellular vesicle trafficking (*VPS35*). Importantly, the analysis of genes showing evidence of positive selection also includes *AHSG* and *FGF19*, whose expression levels have been linked to successful ageing in humans[8]. The role of both factors in metabolism regulation[9,10]—another hallmark of ageing[11,12]—suggests that the specific changes observed in these proteins may have arisen to accommodate the challenges that longevity poses on this system. The list of genes with signatures of positive selection also features *TDO2*, whose inhibition has been proposed to protect against age-related diseases through regulation of tryptophan-mediated proteostasis[13]. In addition, we found evidence for positive selection affecting several genes involved in immune system modulation, such as *MVK*, *IRAK1BP1* and *IL1R2*. Taken together, these results identify

proteostasis, metabolism regulation and immune response as key processes during the evolution of giant tortoises via effects on longevity and resistance to infection.

Parallel to this automatic analysis, we used manually supervised annotation on more than 3,000 genes selected a priori for a series of hypothesis-driven studies on development, physiology, immunity, metabolism, stress response, cancer susceptibility and longevity (Supplementary Section 1.3 and Supplementary Fig. 2). We searched for truncating variants, variants affecting known motifs and variants whose human counterparts are related to known genetic diseases (Supplementary Section 1.3 and Supplementary Table 6). These variants were first confirmed with the RNA-Seq data. Then, more than 100 of the most interesting variants in terms of putative functional relevance were also validated by PCR amplification followed by Sanger sequencing. To this end, we used a panel of genomic DNA samples of 11 different species of giant tortoises endemic to different islands from the Galapagos Archipelago (Supplementary Section 1, Supplementary Table 7 and Supplementary Fig. 3).

The manually supervised annotation of development-related genes showed the complete conservation of the Hox gene set among giant tortoises, with the exception of *HOXC3*, which seems to have been lost in the radiation of Archelosauria[14,15] (Supplementary Section 2, Supplementary Table 8 and Supplementary Fig. 4). *BMP* and *GDF* gene families were also found to be conserved, although the duplication event that gave rise to *GDF1* and *GDF3* in mammals did not occur in turtles, birds and crocodiles. In contrast, we found a duplication of the ParaHox gene *CDX4* in giant tortoises, also present in other reptiles as well as avian reptiles (birds). This annotation also showed the duplication of *WNT11* in turtles and chickens (but not in the lizard *Anolis carolinensis*), and the specific duplication of *WNT4* in turtles. Given the roles of these duplicated genes and their conservation in most vertebrate species, they could prove to be useful candidates to study the morphological development of turtles, particularly in relation to shell formation. Of note, *KDSR*—one of the genes possibly under positive selection in giant tortoises—has been linked to hyperkeratinization disorders[16]. Also, in this regard, we annotated 30 β-keratins in *C. abingdonii*, 26 of which seem to be functional. These numbers are lower than those previously reported for β-keratins in other turtles[17]. Finally, we did not find in *C. abingdonii* or *A. gigantea* any functional orthologues of genes specifically involved in tooth development (such as *ENAM*, *AMEL*, *AMBN*, *DSPP*, *KLK4* and *MMP20*). This finding confirms a pattern in the evolutionary molecular mechanisms for tooth loss, which seems to have been followed consistently and independently across vertebrates. Taken together, these results offer multiple candidates to study developmental traits in tortoises (Supplementary Section 2 and Supplementary Figs. 5–8).

In most species, the immune function is an evolutionary driver that is under strong selective pressure and has important implications in ageing and disease[18]. The specific components and functionality of immune system components in Reptilia, however, have not been extensively characterized beyond the major histocompatibility complex (MHC)[19,20]. Our detailed analysis of 891 genes involved in immune function consistently found duplications affecting immunity genes in giant tortoises compared with mammals (Supplementary Section 3, Supplementary Table 9 and Supplementary Figs. 9–13). We found a genomic expansion of *PRF1* (encoding perforin) in giant tortoises and other turtles, compared with chickens (one copy), *A. carolinensis* (two copies) and most mammals (one copy). Both *C. abingdonii* and *A. gigantea* possess 12 copies of this gene (validated by Sanger sequencing), although three of them have been pseudogenized in *C. abingdonii*. In addition, we detected and validated, by Sanger sequencing, an expansion of the chymase locus, containing granzymes, in giant tortoises (Supplementary Section 3.1 and Supplementary Fig. 10). Both expansions are expected to affect cytotoxic T lymphocyte and natural killer functions, which play important roles in defence against both pathogens and cancer[21,22]. Other concurrent expansions involve *APOBEC1*, *CAMP*, *CHIA* and *NLRP* genes, which participate in viral, microbial, fungal and parasite defence, respectively. These results suggest that the innate immune system in turtles, and especially in giant tortoises, may play a more relevant role than in mammals, consistent with the less important role that adaptive immunity seems to play[19]. We found that class I and II MHC genes probably underwent a duplication event in a common ancestor between giant tortoises and painted turtles (*Chrysemys picta bellii*). We also annotated 40 class III MHC genes, thus confirming the conservation of this cluster in giant tortoises. The large number of MHC genes in giant tortoises is consistent with the suggestion that ancestors of archosaurs and chelonians did not possess a minimal essential MHC as found in the chicken genome[20] (Supplementary Section 3.3, Supplementary Table 10 and Supplementary Figs. 14–16).

Giant tortoises are at the upper end of the size scale for extant Chelonii, and have often been used as an example of gigantism[23]. We analysed a series of genes involved in size regulation in vertebrates, most notably dogs (Supplementary Section 2, Supplementary Table 8 and Supplementary Fig. 6). Our results on genes related to growth hormone, the insulin-like growth factor (IGF) system and stanniocalcins suggest that these genes are well conserved; therefore, additional size determinants may exist in giant tortoises. As a complex phenotype, gigantism in tortoises is expected to be caused by interactions between different genetic and environmental factors. An interesting finding in this regard is the presence of several gene variants in tortoises (including *G. agassizii*) probably affecting the activities of glucose metabolism genes, such as *MIF* (p.N111C; expected to yield a locked trimer) and *GSK3A* (p.R272Q in the activation loop). Given the roles of these positions in the mammalian orthologues of these genes, tortoise-specific changes could point to differences in the regulation of glucose intake and tolerance (Supplementary Section 4, Supplementary Table 11, and Supplementary Figs. 17 and 18). We also found expansions and inactivations in other genes involved in energy metabolism. Thus, glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*)—a glycolytic enzyme with a key role in energy production, as well as in DNA repair and apoptosis[24]—is expanded in giant tortoises. Conversely, the *NLN* gene encoding neurolysin is pseudogenized in tortoises. The loss of this gene in mice has been related to improved glucose uptake and insulin sensitivity[25]. Taken together, these results led us to hypothesize that genomic variants affecting glucose metabolism may have been a factor in the development of tortoises.

The analysis of genes related to the stress response has also highlighted several putative variants in giant tortoises affecting globins and DNA repair factors (Supplementary Section 5, Supplementary Tables 12 and 13, and Supplementary Figs. 19–22, 32 and 33). We found that, despite living terrestrially, giant tortoises conserve the hypoxia-related globin *GbX*[26]. Together with coelacanths, turtles, including giant tortoises, are the only organisms known to possess all eight different types of globins[27]. Consistent with this, we found in both giant tortoise genomes a variant in the transcription factor *TP53* (p.S106E) that has been linked to hypoxia resistance in some mammals and fishes[28]. The presence of the same residue in Testudines strongly suggests a process of convergent evolution in the adaptation to hypoxia, probably driven by an ancestral aquatic environment, which left this footprint in the genomes of terrestrial giant tortoises.

An important trait of large, long-lived vertebrates is their need for tighter cancer protection mechanisms, as illustrated by Peto's paradox[29,30]. In turn, this need for additional protection illustrates the deep relationship and interdependence between cancer and longevity (Fig. 2). Notably, tumours are believed to be very rare in turtles[31]. Therefore, we analysed more than 400 genes classified in

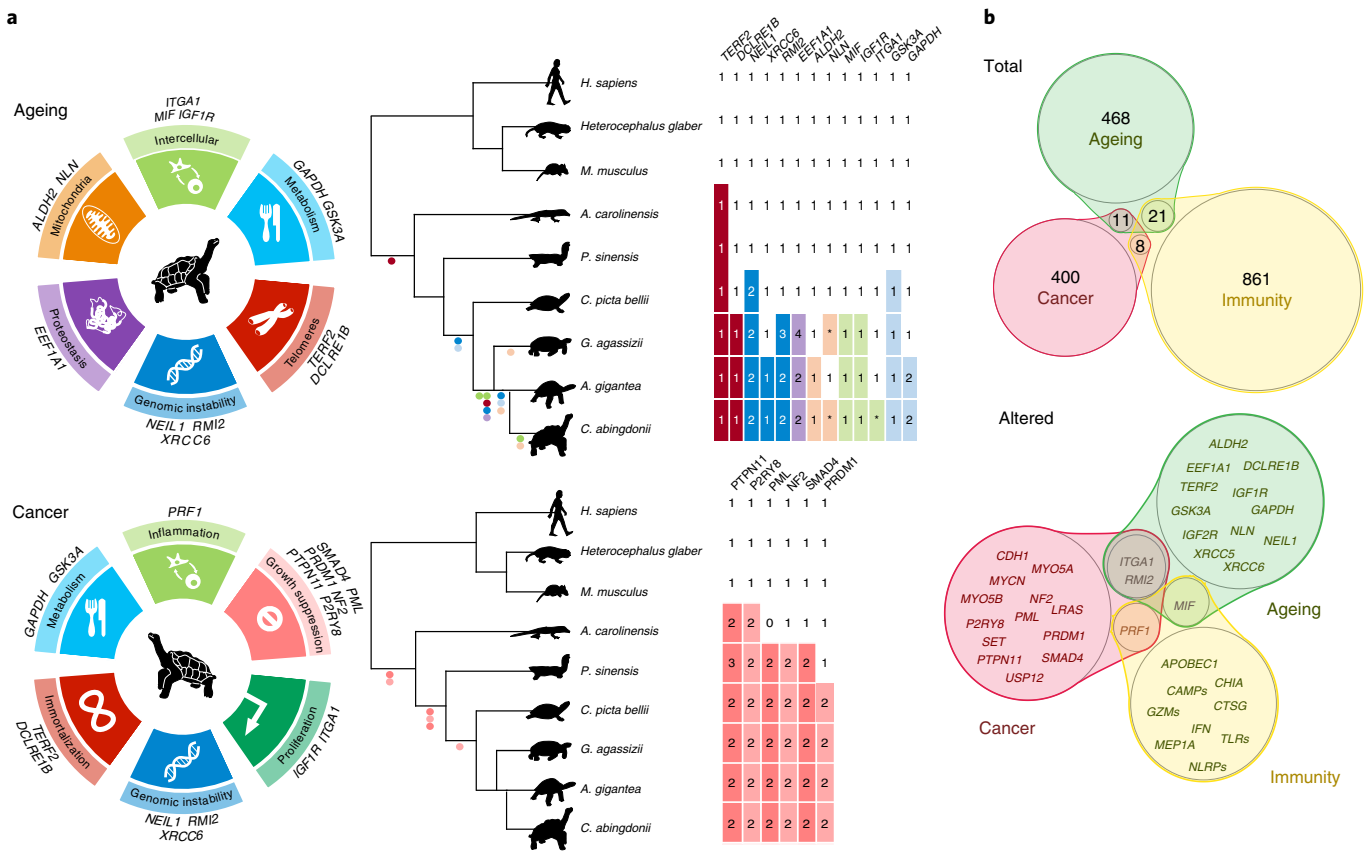

Fig. 2 | Genomic basis of longevity and cancer in giant tortoises. a, Genes potentially implicated in *C. abingdonii* and *A. gigantea* longevity extension and cancer resistance, classified according to their putative role in the different hallmarks. Tables indicate copy-number variations and relevant variants of age-related genes and tumour suppressors found in *C. abingdonii*, *A. gigantea* and other species. Within these tables, numbers indicate gene copy numbers, and asterisks represent pseudogenization events. Dots in colours relating to each hallmark represent presence of the variant. **b**, Venn diagrams showing the relationships between cancer-, ageing- and immunity-related genes, as classified before annotation. Top, all of the genes related to each category that have been manually annotated, including the number of genes in each group. Bottom, those genes showing potentially interesting variations after annotation.

a well-established census of cancer genes as oncogenes and tumour suppressors[32]. Although most presented a highly conserved amino acid sequence when compared with the sequences of other organisms, we uncovered alterations in several tumourigenesis-related genes (Fig. 2a, Supplementary Section 6, Supplementary Table 14 and Supplementary Figs. 23–29). First, we found that several putative tumour suppressors are expanded in turtles compared with other vertebrates, including duplications in *SMAD4*, *NF2*, *PML*, *PTPN11* and *P2RY8*. In addition, the aforementioned expansion of *PRF1*, together with the tortoise-specific duplication of *PRDM1*, suggests that immunosurveillance may be enhanced in turtles. Likewise, we found giant-tortoise-specific duplications affecting two putative proto-oncogenes—*MYCN* and *SET*. Notably, the SET complex mediates oxidative stress responses induced by mitochondrial damage through the action of PRF1 and GZMA in cytotoxic T lymphocyte- and natural killer-mediated cytotoxicity[33]. Taken together, these results suggest that multiple gene copy-number alterations may have influenced the mechanisms of spontaneous tumour growth. Nevertheless, further studies are needed to evaluate the genomic determinants of putative giant-tortoise-specific cancer mechanisms.

Finally, we selected, for manually supervised annotation, a set of 500 genes that may be involved in ageing modulation (Supplementary Section 7 and Supplementary Table 15). The extreme longevity of giant tortoises is expected to involve multiple genes affecting different hallmarks of ageing[11]. We found several alterations in the genomes of giant tortoises that may play a direct role in six of them, and impinge on other ageing hallmarks and processes, such as cancer progression[34] (Fig. 2b). First, we identified changes in three candidate factors (NEIL1, RMI2 and XRCC6) related to the maintenance of genome integrity, a primary hallmark of ageing[11] (Fig. 3a). Thus, we found and validated a duplication affecting NEIL1, a key protein involved in the base-excision repair process whose expression has been linked to extended lifespans in several species[35]. Likewise, *RMI2* is duplicated in tortoises, suggesting an enhanced ability to resolve homologous recombination intermediates to limit DNA crossover formation in cells[36]. In a preliminary exploration of this hypothesis, we overexpressed *NEIL1* and *RMI2* in HEK-293T cells and exposed the infected cells to a sublethal dosage of H_2O_2 or ultraviolet light, monitoring DNA damage by western blot analysis at 24 and 48 h after treatment. As shown in Supplementary Figs. 22, 32 and 33, the expression of both genes results in reduced levels of phosphorylated histone H2AX and cleaved poly (ADP-ribose) polymerase (PARP), suggesting reduced levels of DNA damage[37]. In turn, this result is consistent with the hypothesis that NEIL1 and RMI2 levels may regulate the strength of DNA repair mechanisms. Also in relation to DNA repair mechanisms, we identified and validated a variant affecting *XRCC6*—encoding a helicase involved in non-homologous end joining of double-strand DNA breaks—which may affect a known sumoylation site (p.K556R). This lysine is conserved in diverse vertebrates but, notably, is changed in giant tortoises, and also in the naked mole rat (p.K556N), the longest-lived rodent, which suggests a putative process of convergent evolution (Fig. 3b). Since

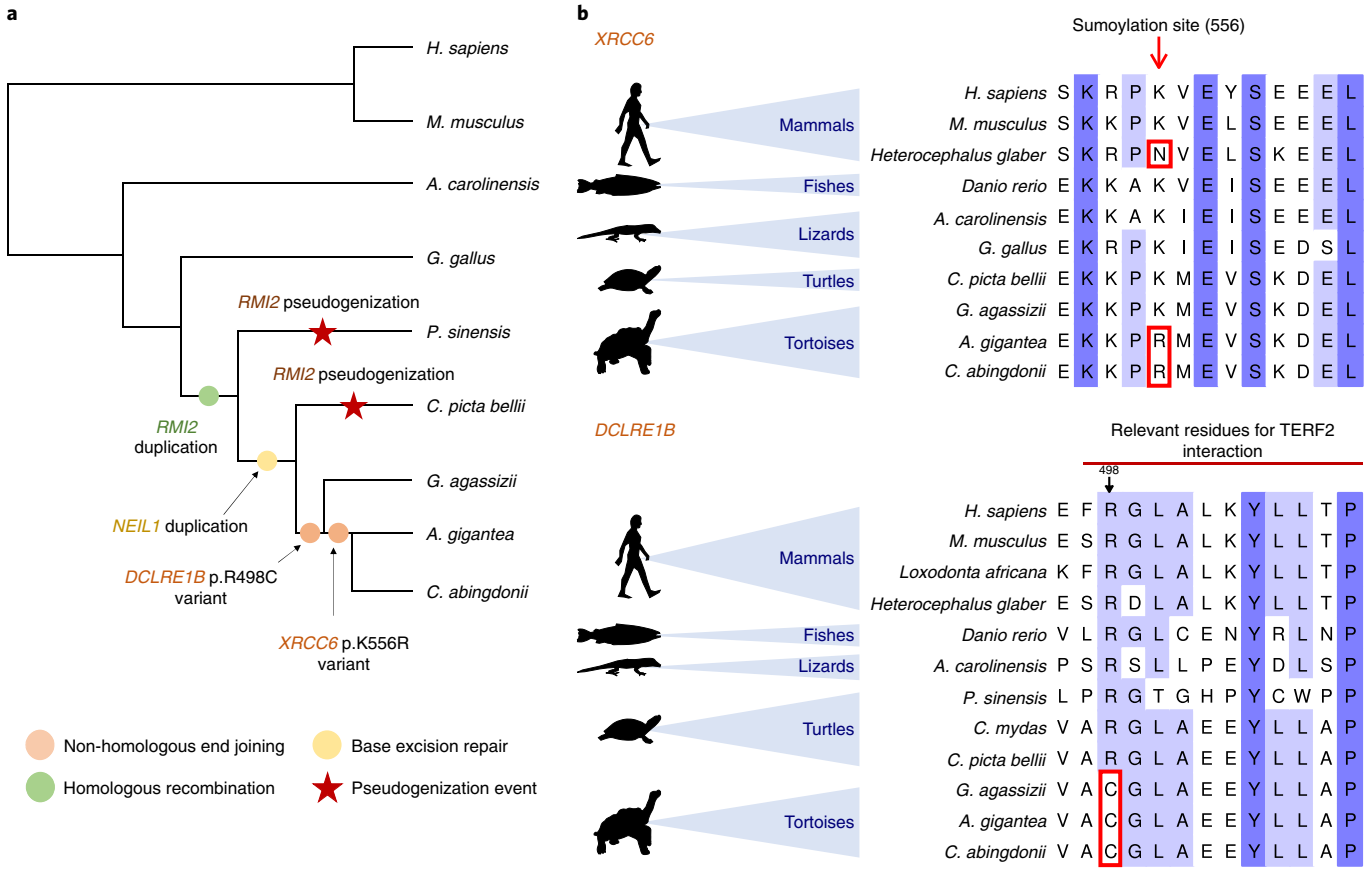

Fig. 3 | DNA repair response in giant tortoises. a, Copy-number variations and putative function-altering point variants found in *C. abingdonii*, *A. gigantea* and closely related species. **b**, Alignments showing the variants highlighted in *XRCC6* and *DCLRE1B*.

sumoylation is induced following DNA damage and plays a key role in DNA repair response and multiple regulatory processes[38], this variant may reflect selective pressures acting on the regulation of the repair of double-strand DNA breaks in long-lived organisms (Supplementary Section 5.5).

Regarding telomere attrition—another primary hallmark of ageing[11]—we uncovered in giant tortoises one variant in DCLRE1B (p.R498C) potentially affecting its binding interface with telomeric repeat binding factor 2 (TERF2) (Fig. 3b and Supplementary Section 7.2). This change, together with the aforementioned variants affecting DNA repair genes that may also impinge on telomere dynamics[39–41], highlights the relevance of telomere maintenance as a regulatory mechanism of longevity in tortoises. Moreover, we found changes potentially affecting proteostasis (Fig. 2a). We independently found specific expansions of the elongation factor gene *EEF1A1* in *C. abingdonii*, *A. gigantea* and *G. agassizii*, as described with the automatic annotation. Importantly, overexpression of *EEF1A1* homologues in *Drosophila melanogaster* has been linked to an increased lifespan in this species[42].

Over time, nutrient sensing deregulation—another hallmark of ageing—can result from alterations in metabolic control mechanisms and signalling pathways[12]. The aforementioned variant affecting the activation loop of GSK3A (Supplementary Section 4.1), which is present in *C. abingdonii* and all tested tortoises from the Galapagos Islands and Aldabra Atoll, as well as their continental outgroups, *G. agassizii* and *C. picta bellii*, may be involved in the maintenance of glucose homoeostasis. Interestingly, the inhibition of GSK3 can extend lifespan in *D. melanogaster*[43]. Likewise, the identified alterations in other giant tortoise genes implicated in glucose metabolism, such as the aforementioned inactivation of *NLN*, may provide interesting candidates to study nutrient sensing in these long-lived species (Supplementary Section 7.4).

Regarding the mitochondrial function, we found two variants (p.Q366M and p.M487T) potentially affecting the function of ALDH2, a mitochondrial aldehyde dehydrogenase involved in alcohol metabolism and lipid peroxidation, among other detoxification processes[44]. Notably, the p.Q366M variant, which may alter the NAD-binding site of ALDH2, is exclusively found in Galapagos giant tortoises, but not in their continental close relative *Chelonoidis chilensis*, nor in the more distantly related Aldabra or Agassiz's tortoises. Thus, these changes could also alter the detoxification process and contribute to pro-longevity mechanisms. Together with the above described specific alterations in other genes of giant tortoises, such as *NLN* and *GAPDH*, which encode enzymes associated with mitochondrial functions[45,46], these variants may also impinge on mitochondrial dysfunction, an antagonistic hallmark of ageing[11] (Supplementary Section 7.5).

We have also found evidence in tortoises of some variants related to altered intercellular communication (Supplementary Section 7.6 and Supplementary Fig. 30), an integrative hallmark of ageing[11]. Thus, we have detected exclusively in *C. abingdonii* a premature stop codon affecting *ITGA1* (p.R990*), an essential integrin involved in cell–matrix and cell–cell interactions. In addition, the aforementioned variant affecting *MIF* is also expected to cause the formation of inactivating interchain disulfide bonds, inhibiting intracellular signalling cascades[47]. Moreover, *MIF* deficiency reduces chronic inflammation in white adipose tissue and expands lifespan, especially in response to caloric restriction[48,49]. Finally, we have annotated a specific variant in *IGF1R* that is expected to affect the interaction between this receptor and the IGF1/2 growth

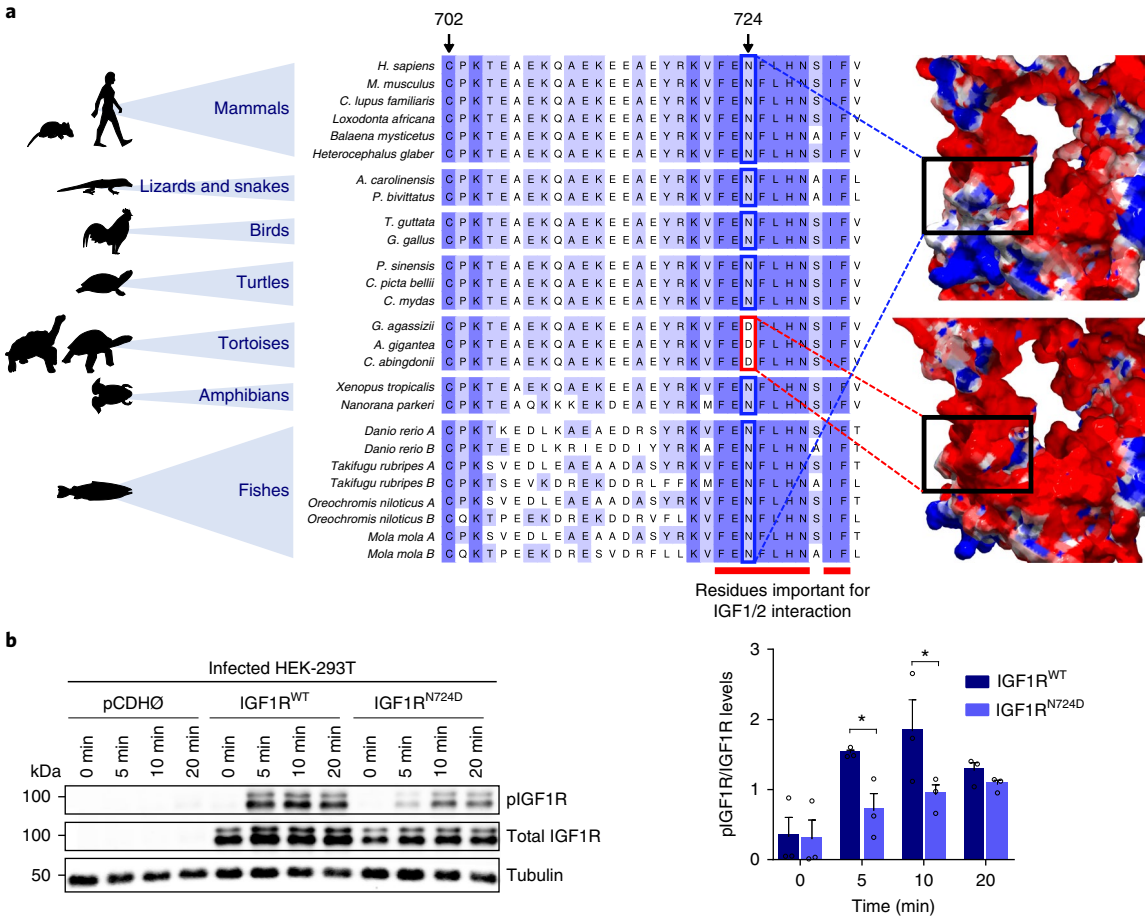

Fig. 4 | Functional relevance of IGF1R^{N724D} in the IGF1 signalling pathway. a, Alignment of IGF1R around residue p.N724 in *C. abingdonii*, *A. gigantea* and other representative species. The predicted electrostatic surfaces of human (top right) and modelled *C. abingdonii* (bottom right) IGF1R around the same residue are shown for comparison. Negatively charged areas are depicted in red, while positively charged areas are depicted in blue. **b**, Western blot analysis and densitometry quantification of the phospho-IGF1R (pIGF1R)/total IGF1R ratio at 5, 10 and 20 min intervals after IGF1 addition in HEK-293T cells infected with pCDH, pCDH-IGF1R^{WT} and pCDH-IGF1R^{N724D} plasmids. Bars indicate means ± s.e.m. *$P < 0.05$, Fisher's least significant difference test ($n = 3$ independent experiments).

factors[50]. Notably, a homology model of this region in IGF1R in *C. abingdonii* suggests that position 724 is located at the surface of the protein, and the presence of an aspartic acid residue changes the local electrostatic field (Fig. 4a). The extended lifespan in different species correlates with IGF signalling decrease[51,52], which suggests that this unique change in *IGF1R* may provide an attractive target to study the cellular mechanisms underlying the exceptional lifespan of these animals. To explore the functional consequences of differential IGF1 signalling caused by the p.N724D variant found in the IGF1 receptor (IGF1R), we infected HEK-293T cells with pCDH, pCDH-IGF1R^{WT} and pCDH-IGF1R^{N724D} plasmids. Cells expressing the mutant receptor showed an attenuation of IGF1 signalling, compared with those expressing the wild-type protein, measured as a significant reduction in the phosphorylation levels of IGF1R at 5 min (95% confidence interval of difference: 0.1119–1.5330, $t = 2.454$, $P = 0.026$) and 10 min (95% confidence interval of difference: 0.1991–1.6200, $t = 2.714$, $P = 0.0153$) after IGF1 treatment (Fig. 4b, Supplementary Section 7.6.2 and Supplementary Fig. 31). According to a two-way analysis of variance, the exogenous IGF1R form accounted for 16.07% of total variation ($F_{1,4} = 20.91$, $P = 0.0102$), while time accounted for 44.23% of total variation ($F_{3,12} = 6.57$, $P = 0.0071$). Interestingly, we also found in tortoises a short deletion in the coding region of *IGF2R* that results in the loss of two amino acids. The fact that *IGF2R* variants have been associated with human longevity[53] opens the possibility that the variant found in tortoises could also contribute to increasing the lifespan of these long-lived animals.

In summary, in this work, we report the preliminary characterization of giant tortoise genomes. We complemented the automatic annotation of genomes from two giant tortoise species with a hypothesis-driven strategy using manually supervised annotation of a large set of genes. The analysis of the resulting sequences offers candidate genes and pathways that may underlie the extraordinary characteristics of these iconic species, including their development, gigantism and longevity. A better understanding of the processes that we have studied may help to further elucidate the biology of these species and therefore aid the ongoing efforts to conserve these dwindling lineages. Lonesome George—the last representative of *C. abingdonii*, and a renowned emblem of the plight of endangered species—left a legacy including a story written in his genome whose unveiling has just started.

Methods

Genome sequencing and assembly. We obtained DNA from a blood sample from Lonesome George—the last member of *C. abingdonii*. This DNA was sequenced, using the Illumina HiSeq 2000 platform, from a 180-base pair-insert paired-end library, a 5-kilobase (kb)-insert mate-pair library and a 20-kb-insert mate-pair library. These libraries were assembled with the AllPaths algorithm[54] for a draft genome containing 64,657 contigs with an N50 of 74 kb. Then, we scaffolded the

contigs with SSPACE version 3.0 (ref. [55]) using the long-insert mate-pair libraries. Finally, we filled the gaps with PBJelly version 15.8.24 (ref. [56]) using the reads obtained from 18 BioPac cells. This step yielded 10,623 scaffolds with an N50 of 1.27 megabases, for a final assembly 2.3 gigabases long. Then, we soft-masked repeated regions using RepeatMasker (http://www.repeatmasker.org) with a database containing chordate repeated elements (included in the software) as a reference. Additionally, we assessed the completeness of assembly by their estimated gene content, using Benchmarking Universal Single-Copy Orthologs (BUSCO version 3.0.0)[57], which tested the status of a set of 2,586 vertebrata genes from the comprehensive catalogue of orthologues[58]. We also performed RNA-Seq from *C. abingdonii* blood and *A. gigantea* granuloma, and aligned the resulting reads to the assembled genome using TopHat[59] (version 2.0.14). Finally, we obtained whole-genome data from *A. gigantea* with one Illumina lane of a 180-base pair paired-end library. The resulting reads were aligned to the *C. abingdonii* genome with BWA[60] (version 0.7.5a). Raw reads from *C. abingdonii* were also aligned to the genome for manual curation of the results. All work on field samples was conducted at Yale University under Institutional Animal Care and Use Committee permit number 2016-10825, Galapagos Park Permit PC-75-16 and Convention on International Trade in Endangered Species number 15US209142/9.

Genome annotation. Using the genome assembly of *C. abingdonii* and the RNA-Seq reads from *C. abingdonii* and *A. gigantea*, we performed de novo annotation with MAKER2. The algorithm was also fed both human and *P. sinensis* reference sequences, and performed two runs in a Microsoft Azure virtual machine (Supplementary Table 16). In parallel, we used selected genes from the human protein database in Ensembl as a reference to manually predict the corresponding homologues in the genome of *C. abingdonii* using the BATI algorithm (Blast, Annotate, Tune, Iterate)[61]. Briefly, this algorithm allows a user to annotate the position and intron/exon boundaries of genes in novel genomes from tblastn results. In addition, tblastn results are integrated to search for novel homologues in the explored genome. Sequencing data have been deposited at the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra), with comments showing which regions were filled with the BioPac reads and therefore may contain frequent errors.

Effective population size changes and diversity. We reconstructed changes in the effective population over time using the PSMC model[5] in the following way: the reads of both individuals were aligned to the reference assembly using bwa mem (version 0.7.15-r1140). We then constructed pseudodiploid sequences using variant calls generated with SAMtools and BCFtools[62], requiring minimal base and mapping qualities of 30. We additionally masked out any region with coverage below 36 or above 216 for the *C. abingdonii* sample, and below 8 or above 52 for the *A. gigantea* sample, as a function of their respective genome-wide average coverage. The resulting sequences were used to run 100 PSMC bootstrap replicates per individual, using the following parameters: -N25 -t15 -r5 -p '4+25*2+4+6'. The result was averaged and scaled to real time assuming a mutation rate (μ) of 2.5×10^{-8} and a generation time (g) of 25 years.

Expansion of gene families. To detect expansion of gene families, we aligned pairwise all the predicted proteins from the automatic annotation to the UniProt[63] database of human proteins and the UniProt database of *P. sinensis* proteins using BLAST[64] (version 2.6.011). Then, we used in-house Perl scripts to group these proteins in one-to-one, one-to-many and many-to-many orthologous relationships. Only alignments spanning at least 80% of the longer protein, and with more than 60% identities, were considered. Finally, we interrogated the resulting database to find families with *C. abingdonii*-specific expansions and curated the results manually. This way, we constructed extended orthology sets that may contain more than one sequence per species. These sets recapitulate most of the known families, although some of these families appear split according to sequence similarity.

Phylogenetic, evolutionary and structural analyses. Next, we assessed evidence for signatures of positive selection affecting the predicted set of genes. For this purpose, we used databases from the human (*Homo sapiens*), mouse (*Mus musculus*), dog (*Canis lupus familiaris*), gecko (*Gekko japonicus*), green anole lizard (*A. carolinensis*), python snake (*Python bivittatus*), common garter snake (*Thamnophis sirtalis*), Habu viper (*Trimeresurus mucrosquamatus*), budgerigar (*Melopsittacus undulatus*), zebra finch (*Taeniopygia guttata*), flycatcher (*Ficedula albicollis*), duck (*Anas platyrhynchos*), turkey (*Meleagris gallopavo*), chicken (*Gallus gallus*), Chinese soft-shell turtle (*P. sinensis*), green sea turtle (*Chelonia mydas*) and painted turtle (*C. picta bellii*) to generate pairwise alignments of all available genes one by one. To this end, we used BLAST and simple in-house Perl scripts (https://github.com/vqf/LG), which allowed us to group the genes by identity (focusing only on those presenting one-to-one orthology). We then discarded those groups in which there were more than three species missing (always excluding those in which *C. abingdonii* was missing). This way, we obtained 1,592 groups of sequences (similar to other studies). We then aligned them with PRANK version 150803 using the codon model and analysed the alignments with codeml from the PAML package[65]. To search for genes with signatures of positive selection affecting genes specific to *C. abingdonii*, we executed two different branch models—M0, with a single $\omega0$ value (where ω represents the ratio of non-synonymous to synonymous substitutions) for all the branches (nested), and M2a, with a foreground $\omega2$ value exclusive for *C. abingdonii* and a background $\omega1$ value for all the other branches. As a control, the second model was repeated using *P. sinensis* as the foreground branch. Genes with a high $\omega2$ value (>1) and a low $\omega1$ value ($\omega1<0.2$ and $\omega1\sim\omega0$) in *C. abingdonii*, but not in *P. sinensis* (Supplementary Section 1.2 and Supplementary Tables 5 and 17), were then considered to be under positive selection. After this, we used the M8 model to assess the individual importance of every site in these positively selected genes, obtaining a list of sites of special interest in this evolutionary effect. These results were compared with those of the Aldabra tortoise through alignments, to evaluate which of these important residues were altered (Supplementary Table 18). Homology models were performed with SWISS-MODEL[66] from the closest template available. The results were inspected and rendered with DeepView version 4.0.1. Electric potentials were calculated with DeepView using the Poisson–Boltzmann computation method. Figures were generated with PovRay (http://povray.org).

Functional analyses. HEK-293T cells were infected with pCDH, pCDH-NEIL1, pCDH-RMI2 or pCDH-NEIL1 + pCDH-RMI2 in the case of repair studies, and pCDH, pCDH-IGF1R^{WT} or pCDH-IGF1R^{N724D} in the case of IGF1R analyses. For the repair studies, we isolated clones of infected HEK-293T cells with proper expression levels of *NEIL1* and *RMI2*. Cells were exposed to ultraviolet light (20 J m^{-2}) or H_2O_2 (500 µM) 24 and 48 h before being lysed in NP-40 lysis buffer containing 50 mM Tris-HCl pH 7.4, 150 mM NaCl, 10 mM EDTA pH 8 and 1% NP-40, and supplemented with protease inhibitor cocktail (cOmplete, EDTA-free; Roche), as well as phosphatase inhibitors (PhosSTOP; Roche/NaF; Merck). For the *IGF1R* variant analyses, cells were serum starved for 14 h, then treated with 100 nM IGF1 for 5, 10 and 20 min before lysis in the same buffer. Equal amounts of protein were resolved by 8 to 13% sodium dodecyl sulfate polyacrylamide gel electrophoresis and transferred to PVDF membranes (GE Healthcare Life Sciences). Membranes were blocked for 1 h at room temperature with TBS-T (0.1% Tween 20) containing 5% bovine serum albumin. Immunoblotting was performed with primary antibodies diluted 1:500 to 1:1000 in TBS-T and 1% bovine serum albumin and incubated overnight at 4 °C. The primary antibodies used were: anti-phospho-Histone H2AX (Ser139) (EMD Millipore; 05-636, clone JBW301, lot 2854120), anti-PARP (Cell Signaling Technology; 9542S, rabbit polyclonal, lot 15), anti-FLAG (Cell Signaling Technology; 2368S, rabbit polyclonal, lot 12), anti-IGF1R (Abcam; ab182408, clone EPR19322, lot GR312678-8), anti-IGF1R (p Tyr1161) (Novus Biologicals; NB100-92555, rabbit polyclonal, lot CJ36131), anti-β-actin (Sigma–Aldrich, A5441, clone AC-15, lot 014M4759) and anti-α-tubulin (Sigma–Aldrich, T6074, clone B-5-1-2, lot 075M4823V). After washing with TBS-T, membranes were incubated with secondary antibodies conjugated with IRDye 680RD (LI-COR Biosciences; 926-68071, polyclonal goat-anti-rabbit, lot C41217-03; and 926-32220, polyclonal goat-anti-mouse, lot C00727-03) or IRDye 800CW (LI-COR Biosciences; 926-32211, polyclonal goat-anti-rabbit, lot C60113-05; and 926-32210, polyclonal goat-anti-mouse, lot C50316-03) for 1 h at room temperature. Protein bands were scanned on an Odyssey infrared scanner (LI-COR Biosciences). Band intensities were quantified by ImageJ and used to calculate the phospho-IGF1R/IGF1R ratio in the case of the IGF1R assay. In each replicate, cells were infected independently. For the samples from ultraviolet treatment, Flag (RMI2) was detected on the same samples used for the remaining western blots shown in this panel, run in parallel on an identical blot. Similarly, for the samples from H_2O_2 treatment, the western blots shown were carried out with the same samples run in parallel in three identical blots (one for PARP and actin, a second for Flag (NEIL1 and RMI2) and a third for pH2AX). Each sample contained one replicate. Statistical comparisons consisted of two-way analysis of variance performed using GraphPad Prism 7.0 software. Differences were considered statistically significant when $P<0.05$. Effect sizes are expressed as group sum-of-squares divided by the total sum-of-squares (R^2). At each time point, both groups were also compared with Fisher's least significant difference test (uncorrected; $\alpha=0.05$).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The scripts for manual annotation (BATI) can be accessed at http://degradome.uniovi.es/downloads.html. Custom scripts used to produce multiple alignments for positive selection and copy-number studies are freely available at https://github.com/vqf/LG.

Data availability

Data supporting the findings of this study are available within the paper and its Supplementary Information. Sequencing data have been deposited at the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) with BioProject accession number PRJNA416050. The accession number of the assembled genomic sequence is PKMU00000000. MAKER2-predicted protein sequences can be downloaded from https://github.com/vqf/LG.

Received: 24 January 2018; Accepted: 25 October 2018;
Published online: 03 December 2018

References

1. Kim, E. B. et al. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223–227 (2011).
2. Keane, M. et al. Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* **10**, 112–122 (2015).
3. Nicholls, H. The legacy of Lonesome George. *Nature* **487**, 279–280 (2012).
4. Kehlmaier, C. et al. Tropical ancient DNA reveals relationships of the extinct Bahamian giant tortoise *Chelonoidis alburyorum*. *Proc. R. Soc. B* **284**, 20162235 (2017).
5. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
6. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics* **48**, 1–39 (2014).
7. Wang, Z. et al. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat. Genet.* **45**, 701–706 (2013).
8. Sanchis-Gomar, F. et al. A preliminary candidate approach identifies the combination of chemerin, fetuin-A, and fibroblast growth factors 19 and 21 as a potential biomarker panel of successful aging. *Age* **37**, 9776 (2015).
9. Pal, D. et al. Fetuin-A acts as an endogenous ligand of TLR4 to promote lipid-induced insulin resistance. *Nat. Med.* **18**, 1279–1285 (2012).
10. Kir, S. et al. FGF19 as a postprandial, insulin-independent activator of hepatic protein and glycogen synthesis. *Science* **331**, 1621–1624 (2011).
11. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
12. López-Otín, C., Galluzzi, L., Freije, J. M., Madeo, F. & Kroemer, G. Metabolic control of longevity. *Cell* **166**, 802–821 (2016).
13. Van der Goot, A. T. et al. Delaying aging and the aging-associated decline in protein homeostasis by inhibition of tryptophan degradation. *Proc. Natl Acad. Sci. USA* **109**, 14912–14917 (2012).
14. Crawford, N. G. et al. A phylogenomic analysis of turtles. *Mol. Phylogenet. Evol.* **83**, 250–257 (2015).
15. Chiari, Y., Cahais, V., Galtier, N. & Delsuc, F. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* **10**, 65 (2012).
16. Boyden, L. M. et al. Mutations in *KDSR* cause recessive progressive symmetric erythrokeratoderma. *Am. J. Hum. Genet.* **100**, 978–984 (2017).
17. Li, Y. I., Kong, L., Ponting, C. P. & Haerty, W. Rapid evolution of beta-keratin genes contribute to phenotypic differences that distinguish turtles and birds from other reptiles. *Genome Biol. Evol.* **5**, 923–933 (2013).
18. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).
19. Zimmerman, L. M., Vogel, L. A. & Bowden, R. M. Understanding the vertebrate immune system: insights from the reptilian perspective. *J. Exp. Biol.* **213**, 661–671 (2010).
20. Balakrishnan, C. N. et al. Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biol.* **8**, 29 (2010).
21. Dotiwala, F. et al. Killer lymphocytes use granulysin, perforin and granzymes to kill intracellular parasites. *Nat. Med.* **22**, 210–216 (2016).
22. Voskoboinik, I., Whisstock, J. C. & Trapani, J. A. Perforin and granzymes: function, dysfunction and human pathology. *Nat. Rev. Immunol.* **15**, 388–400 (2015).
23. Jaffe, A. L., Slater, G. J. & Alfaro, M. E. The evolution of island gigantism and body size variation in tortoises and turtles. *Biol. Lett.* **7**, 558–561 (2011).
24. Chuang, D. M., Hough, C. & Senatorov, V. V. Glyceraldehyde-3-phosphate dehydrogenase, apoptosis, and neurodegenerative diseases. *Annu. Rev. Pharmacol. Toxicol.* **45**, 269–290 (2005).
25. Cavalcanti, D. M. et al. Neurolysin knockout mice generation and initial phenotype characterization. *J. Biol. Chem.* **289**, 15426–15440 (2014).
26. Corti, P. et al. Globin X is a six-coordinate globin that reduces nitrite to nitric oxide in fish red blood cells. *Proc. Natl Acad. Sci. USA* **113**, 8538–8543 (2016).
27. Schwarze, K., Singh, A. & Burmester, T. The full globin repertoire of turtles provides insights into vertebrate globin evolution and functions. *Genome Biol. Evol.* **7**, 1896–1913 (2015).
28. Zhao, Y. et al. Codon 104 variation of *p53* gene provides adaptive apoptotic responses to extreme environments in mammals of the Tibet plateau. *Proc. Natl Acad. Sci. USA* **110**, 20639–20644 (2013).
29. Caulin, A. F. & Maley, C. C. Peto's paradox: evolution's prescription for cancer prevention. *Trends Ecol. Evol.* **26**, 175–182 (2011).
30. Chiari, Y., Glaberman, S. & Lynch, V. J. Insights on cancer resistance in vertebrates: reptiles as a parallel system to mammals. *Nat. Rev. Cancer* **18**, 525 (2018).
31. Garner, M. M., Hernandez-Divers, S. M. & Raymond, J. T. Reptile neoplasia: a retrospective study of case submissions to a specialty diagnostic service. *Vet. Clin. North Am. Exot. Anim. Pract.* **7**, 653–671 (2004).
32. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
33. Martinvalet, D., Zhu, P. & Lieberman, J. Granzyme A induces caspase-independent mitochondrial damage, a required first step for apoptosis. *Immunity* **22**, 355–370 (2005).
34. Gorbunova, V., Seluanov, A., Zhang, Z., Gladyshev, V. N. & Vijg, J. Comparative genetics of longevity and cancer: insights from long-lived rodents. *Nat. Rev. Genet.* **15**, 531–540 (2014).
35. MacRae, S. L. et al. DNA repair in species with extreme lifespan differences. *Aging* **7**, 1171–1184 (2015).
36. Daley, J. M., Chiba, T., Xue, X., Niu, H. & Sung, P. Multifaceted role of the Topo IIIα–RMI1-RMI2 complex and DNA2 in the BLM-dependent pathway of DNA break end resection. *Nucleic Acids Res.* **42**, 11083–11091 (2014).
37. Ivashkevich, A., Redon, C. E., Nakamura, A. J., Martin, R. F. & Martin, O. A. Use of the gamma-H2AX assay to monitor DNA damage and repair in translational cancer research. *Cancer Lett.* **327**, 123–133 (2012).
38. Cremona, C. A. et al. Extensive DNA damage-induced sumoylation contributes to replication and repair and acts in addition to the mec1 checkpoint. *Mol. Cell* **45**, 422–432 (2012).
39. Wang, Y., Ghosh, G. & Hendrickson, E. A. Ku86 represses lethal telomere deletion events in human somatic cells. *Proc. Natl Acad. Sci. USA* **106**, 12430–12435 (2009).
40. Tong, A. S. et al. ATM and ATR signaling regulate the recruitment of human telomerase to telomeres. *Cell Rep.* **13**, 1633–1646 (2015).
41. Ribes-Zamora, A., Indiviglio, S. M., Mihalek, I., Williams, C. L. & Bertuch, A. A. TRF2 interaction with Ku heterotetramerization interface gives insight into c-NHEJ prevention at human telomeres. *Cell Rep.* **5**, 194–206 (2013).
42. Shikama, N., Ackermann, R. & Brack, C. Protein synthesis elongation factor EF-1 alpha expression and longevity in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **91**, 4199–4203 (1994).
43. Castillo-Quan, J. I. et al. Lithium promotes longevity through GSK3/NRF2-dependent hormesis. *Cell Rep.* **15**, 638–650 (2016).
44. Ohta, S., Ohsawa, I., Kamino, K., Ando, F. & Shimokata, H. Mitochondrial ALDH2 deficiency as an oxidative stress. *Ann. NY Acad. Sci.* **1011**, 36–44 (2004).
45. Serizawa, A., Dando, P. M. & Barrett, A. J. Characterization of a mitochondrial metallopeptidase reveals neurolysin as a homologue of thimet oligopeptidase. *J. Biol. Chem.* **270**, 2092–2098 (1995).
46. Tristan, C., Shahani, N., Sedlak, T. W. & Sawa, A. The diverse functions of GAPDH: views from different subcellular compartments. *Cell. Signal.* **23**, 317–323 (2011).
47. Fan, C. et al. MIF intersubunit disulfide mutant antagonist supports activation of CD74 by endogenous MIF trimer at physiologic concentrations. *Proc. Natl Acad. Sci. USA* **110**, 10994–10999 (2013).
48. Verschuren, L. et al. MIF deficiency reduces chronic inflammation in white adipose tissue and impairs the development of insulin resistance, glucose intolerance, and associated atherosclerotic disease. *Circ. Res.* **105**, 99–107 (2009).
49. Harper, J. M., Wilkinson, J. E. & Miller, R. A. Macrophage migration inhibitory factor-knockout mice are long lived and respond to caloric restriction. *FASEB J.* **24**, 2436–2442 (2010).
50. Whittaker, J. et al. Alanine scanning mutagenesis of a type 1 insulin-like growth factor receptor ligand binding site. *J. Biol. Chem.* **276**, 43980–43986 (2001).
51. Kenyon, C. J. The genetics of ageing. *Nature* **464**, 504–512 (2010).
52. Brohus, M., Gorbunova, V., Faulkes, C. G., Overgaard, M. T. & Conover, C. A. The insulin-like growth factor system in the long-lived naked mole-rat. *PLoS ONE* **10**, e0145587 (2015).
53. Soerensen, M. et al. Human longevity and variation in GH/IGF-1/insulin signaling, DNA damage signaling and repair and pro/antioxidant pathway genes: cross sectional and longitudinal studies. *Exp. Gerontol.* **47**, 379–387 (2012).
54. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA* **108**, 1513–1518 (2011).
55. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
56. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
57. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
58. Zdobnov, E. M. et al. OrthoDBv9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45**, D744–D749 (2017).

59. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
60. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
61. Quesada, V., Velasco, G., Puente, X. S., Warren, W. C. & López-Otín, C. Comparative genomic analysis of the zebra finch degradome provides new insights into evolution of proteases in birds and mammals. *BMC Genomics* **11**, 220 (2010).
62. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. The UniProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
64. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
65. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
66. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–W258 (2014).

Acknowledgements

We thank J. R. Obeso for support, J. M. Freije, X. S. Puente, R. Valdés-Mas, F. G. Osorio, D. López-Velasco, A. Corrales, P. Salinas, D. Rodríguez, A. López-Soto, A. R. Folgueras and M. Mittelbrunn for helpful comments and advice, M. Garaña, O. Sanz, J. Isla and A. Marcos (Microsoft) for computing facilities, and F. Rodríguez, D. A. Puente and S. A. Miranda for excellent technical assistance. We also acknowledge generous support from J. I. Cabrera. We thank Banco Santander for funding a short stay of S.F.-R. and D.C.-I. at Yale University. V.Q. is supported by grants from the Principado de Asturias and Ministerio de Economía y Competitividad, including FEDER funding. L.F.K.K. is supported by an FPI fellowship associated with BFU2014-55090-P (FEDER). T.M.-B. is supported by MINECO BFU2017-86471-P (MINECO/FEDER, UE), an NIH U01 MH106874 grant, the Howard Hughes International Early Career programme, Obra Social 'La Caixa' and Secretaria d'Universitats i Recerca, and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya. C.L.-O. is supported by grants from the European Research Council (DeAge; ERC Advanced Grant), Ministerio de Economía y Competitividad, Instituto de Salud Carlos III (RTICC) and Progeria Research Foundation. The Instituto Universitario de Oncología is supported by Fundación Bancaria Caja de Ahorros de Asturias. We also thank staff at the Galapagos National Park and Galapagos Conservancy for logistic and financial support.

Author contributions

V.Q. and J.G.P.-S. performed the automatic analysis of genomes. S.F.-R. coordinated the manual genomic annotation, which was performed by J.G.P.-S., O.S.-F., D.C.-I., M.G.A., M.A.-V., D.C., P.M., J.R.A., I.T.-G., D.R.-V. and M.P.-T. S.F.-R. and D.C.-I. performed the validation of the identified genomic variants. G.B. coordinated the functional analyses of the identified genomic variants, which were carried out by O.S.-F., D.C.-I., M.G.A., M.A.-V., D.C., P.M., J.R.A. and I.T.-G. J.M. helped to screen the wild samples for SNP validation, and contributed to results interpretation. M.Q., L.B.B., J.P.G., Y.C., S.G., C.C., B.R.E., S.J.G., D.L.E., R.C.G., M.A.R. and N.P. contributed to early data collection and analyses. W.T., D.O.R. and J.P.G. helped to obtain material-securing permits and biological samples. K.P.W. partly supported data collection and supervised the initial analysis. Z.-F.J. prepared DNA and RNA samples for genomic analyses and conducted raw data quality checks. L.F.K.K. and T.M.-B. performed population history and diversity studies. V.Q., A.C. and C.L.-O. directed the research, analysed the data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0733-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.C. or C.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 22

Genome Sequencing and Analysis Methods in Chronic Lymphocytic Leukemia

Víctor Quesada, Miguel Araujo-Voces, José G. Pérez-Silva, Gloria Velasco, and Carlos López-Otín

Abstract

The genomic sequencing of chronic lymphocytic leukemia (CLL) samples has provided exciting new venues for the understanding and treatment of this prevalent disease. This feat is possible thanks to high-throughput sequencing methods, such as Illumina sequencing. The interpretation of these data sources requires not only appropriate software and hardware, but also understanding the biology and technology behind the sequencing process. Here, we provide a primer to understand each step in the analysis of point mutations from whole-genome or whole-exome sequencing experiments of tumor and normal samples.

Key words Bioinformatics, Genomics, Cancer, Next-generation sequencing, Leukemia

1 Introduction

Genomic studies on hematological neoplasias have provided important insights into the molecular mechanisms driving initiation and evolution of these diseases [1]. This is particularly the case of chronic lymphocytic leukemia (CLL), which has benefited enormously from Cancer Genomic initiatives aimed at elucidating the mutational landscape of this prevalent disease [2–4]. Multiple programs exist for the interpretation of high-throughput sequencing (HTS) data, including graphical [5] and commercial tools. The search for somatic mutations in paired tumor/normal samples can be roughly divided into three phases with dedicated tools: alignment of reads (frequently using BWA [6]), mutation discovery (using for instance GATK [7] or SomaticSniper [8]), and mutation characterization (using for instance VEP [9]). Although not exclusively, this type of analysis is mainly used with whole-genome (WGS) or whole-exome (WES) sequencing.

In the near future, HTS is very likely to become a fixture in research laboratories and clinical institutions. A foreseeable

Sami N. Malek (ed.), *Chronic Lymphocytic Leukemia: Methods and Protocols*, Methods in Molecular Biology, vol. 1881, https://doi.org/10.1007/978-1-4939-8876-1_22,

consequence of this trend will be the tight integration and standardization of every step of HTS analysis. While this will allow non-specialists to benefit from these powerful techniques, it also means that users will be separated from the analytical process. However, in our experience, the understanding of the challenges posed by HTS improves the interpretation of results, independently of which tools are used. For this reason, we provide here a typical analysis with Sidrón, our mutation discovery pipeline [10, 11]. To simplify this primer, only one sample will be considered, and the existing variants will be obtained.

2 Materials

All the necessary files to follow this pipeline are provided at http://github.com/vqf/sidron. These files are designed for Unix-based systems. Most executables are written in Perl, and therefore can be run in Windows-based systems. However, adapting the pipeline to Windows systems requires some programming experience. This tutorial includes small input (fastq and bam) files, so it does not require special hardware. Actual work with WES and particularly WGS files requires at least large memory storage capacities, and in practice also multiple CPUs and access to large RAM. As a reference, each full WGS file will require permanent memory in the order of hundreds of Gb.

In addition, other external programs are necessary to follow the procedure. The first part of the tutorial includes the alignment of the sequences, which is performed with BWA. Once the reads are aligned, Sidrón uses Samtools to extract information from the BAM files. Finally, one of the filtering procedures uses a second aligner, named BLAT. All these programs are public and free:

1. BWA installation files and procedures can be found at http://bio-bwa.sourceforge.net/.
2. To install Samtools, follow the instructions at http://www.htslib.org/. You will also need the corresponding Perl library (install distribution *LDS/Bio-SamTools-1.43.tar.gz* from CPAN).
3. To install BLAT, download the file https://users.soe.ucsc.edu/~kent/src/blatSrc.zip and follow the instructions within. Also download http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/faToTwoBit.
4. The example uses the human genome as a template. The corresponding FASTA sequence can be downloaded from ftp://ftp.ensembl.org/pub/release-91/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.toplevel.fa.gz. Download the file and decompress it with *gzip –d Homo_sapiens.GRCh38.dna.toplevel.fa.gz* at the download folder.

5. Index the reference genome with *faToTwoBit*. *Homo_sapiens.GRCh38.dna.toplevel.fa Homo_sapiens.GRCh38.dna.toplevel.fa.2bit*. Start a BLAT server with the script provided in the example (*./startHumanServer Homo_sapiens.GRCh38.dna.toplevel.fa.2bit [port_number]*). If *port_number* is not specified, the script will use 9006.
   The tutorial assumes that the executables *bwa*, *samtools*, *gfClient*, *gfServer*, and *faToTwoBit* are available from any folder. You can create softlinks to those executables in a folder inside the PATH environment or change the corresponding commands to include the path to the executable.

3 Methods

The files provided at http://github.com/vqf/sidron include two small input fastq files (*ex_1.fastq.gz* and *ex_2.fastq.gz*). You can see the format of this file by typing *zcat ex_1.fastq.gz | head* at the download folder. Although this file is small, its format is identical to that of the typical output from HTS machines.

3.1 Alignment

1. Create a custom alignment file. In the folder where Sidrón was downloaded, run *perl align.pl*. The script will ask for several pieces of data:
   (a) Basename: name of the output file (i.e., align_mreads).
   (b) Ref_genome: Path to the fasta file containing the reference genome.
   (c) Read_folder: Path to the folder containing the fastq files to align (single-end or paired-end). You can accept the default value pointing at the current folder.
   (d) File_pattern: Common part of the fastq file names (e.g., *ex*). If reads are paired, the script searches for files whose names match this pattern and afterward contain "_1" and "_2." This will identify the input files *ex_1.fastq.gz* and *ex_2.fastq.gz*.

   If all data are correct, the script will create an executable file called *align_mreads.sh*. Its contents automate the alignment process. At the beginning of the file, two lines provide the names of the executables for *bwa* and *samtools*. These can be changed manually.

2. Execute the custom alignment file with. */align_mreads.sh*. This will create a folder called *align_mreads* with two files named *align_mreads.sorted.bam* and *align_mreads.nodups.bam*. The second file contains the alignments with duplicates removed (**Note 1**). In this example, we will use *align_mreads.sorted.bam*.

3.2 Extraction and Calculation of Putative Variants

1. Enter the folder containing the bam file (*cd align_mreads*).
2. Use samtools to feed the pileup of the bam file into the perl script (*samtools mpileup -f* path_to_genome_fasta *align_mreads.sorted.bam | perl ../extract_mq.pl>mreads.mq*) (**Note 2**). The file *mreads.mq* contains all the positions in the alignment that show any change that may suggest that there is a variant. Most of these positions will in fact not contain variants, but sequencing or alignment errors. It is important to notice that for the rest of the positions there is no indication of variant. Even if other variants exist, the current sequencing data cannot find them.
3. Use Sidrón to assign a score to each putative variant (*perl ../sidron.pl mreads.mp ../table.hsh>mreads.sidron*) (**Note 3**). The output file contains three additional columns: genotypes considered, Sidrón score, and reserved (**Note 4**).

3.3 Filtering of Variants

1. Get positions with high S values (*../downstream_onesample.sh mreads.sidron*) (**Note 5**). This will create a file called *mreads.sidron.variants*.
2. Filter by nonlocal criteria (bad alignment, repetitive regions, etc.) with polyfilter (*perl ../polyfilter.pl mreads.sidron.variants align_mreads.sorted.bam > mreads.polyfilter*) (**Note 6**).
3. Repeat steps in Subheading 2, **item 3** and Subheading 3.1 with the filtered positions:
   perl ../sidron.pl mreads.polyfilter ../table.hsh > mreads.polyfilter.sidron
   ../downstream_onesample.sh mreads.polyfilter.sidron.
   At the end, we obtain a file called *mreads.polyfilter.sidron.variants* with the filtered variants.

3.4 Exploration of Variants

1. Create files with the genomic coordinates to explore. For instance, we can run *head mreads.polyfilter.sidron.variants>ex*. This will create a file named *ex* with the first ten variants. The only columns needed are the first and the second (chromosome and position), the script will not read the rest of the columns.
2. Create snapshots of the interesting positions with *perl ../snapshot.pl align_mreads.sorted.bam ex*. Each position will yield an html file that can be examined with any web browser (**Note 7**). The reference genome appears at the top in green, and each read appears aligned below. When the read base is the same as the corresponding base in the reference genome, we have points (read in the +strand) or commas (read in the −strand). High-confidence bases are in blue, and low-confidence bases are in red. Each read is clickable for more information (**Notes 8–9**).

4 Notes

1. Depending on the type of sequencing, we may want to remove duplicates (whole-genome, whole-exon) or not (pooled sequencing). In general, we want to remove duplicates when the read depth is relatively low and the probability of independently getting exactly the same DNA fragment more than once is low. If we sequence a part of the genome at very high read depth, this probability is much higher, and most duplicates will not be artifacts.
2. By default, *samtools mpileup* cuts the read depth at 250. If necessary, this limitation can be circumvented by adding the –d option (e.g., *samtools mpileup -d 1e8 -f* path_to_genome_fasta *align_mreads.sorted.bam...*).
3. The *table.hsh* file contains the expected rates of error for each base the sequencer reads (i.e., the probability that the machine reads A when in fact it should read C). Ideally, one should determine those error rates with orthogonal methods, such as genotyping microarrays. However, we have also developed specific methods to estimate those error rates directly from the reads.
4. The genotypes in the Sidrón file are given as a pair of bases N_1N_2. The first base is the most represented in the pileup, and the second base is the second most frequent base in the same position. Sidrón considers and compares two genotypes: homozygous (N_1 N_1) and heterozygous (N_1 N_2). The S score is defined as

   S=log10pcHetpcHz

   Here, $\log_{10}$ is the logarithm in base 10, *c* is the configuration (which bases were read and with which qualities), *Het* is the heterozygous genotype, and *Hz* is the homozygous genotype. Each probability is computed from the configuration and the error table (*table.hsh*). For instance, the probability that a configuration contains 3 As and one G given a *Hz* genotype is the probability that 3 bases were correctly read and in one the machine gave a G when it should have given an A.
5. The criteria to filter variants with *downstream_onesample.sh* are complex. An explanation of the cutoff points can be found inside the script. The cutoff values depend on the read depth of each position, as high-depth positions contain more information and allow finer distinctions. Each parameter in this file can be overridden by creating a file called *config.txt* in the run folder with the definitions. The operations and values are stored in a file called *log.txt*.

6. As the name implies, polyfilter contains several filters. The most important ones consider the probability that the variant occurs only at specific positions in each read and the possibility that the reads with the variants can be aligned to different positions in the genome. The first of those filters finds the position of each variant base inside its read. Then, it calculates the maximum distance between those positions (d). The probability that n bases chosen at random in a read of length l yield a maximum distance of d or less is

$$p(n,d)=(l-d)(d+1)^n-(l-d-1)d^n,\forall d\in(0,l-1)$$

If the computed probability for a position is lower than 0.2%, the configuration is considered spurious and filtered out. The second filter performs a BLAT alignment of each read containing a variant and filters the read out if it can be aligned with the same or higher quality at some other place in the genome. Since BLAT has a slightly different algorithm than BWA, this filter can improve the sensitivity to misaligned reads. On the other hand, this filter does not remove positions, only reads. This is the reason why Sidrón must be executed again after this step. Filtered positions are written to a file called *filtered_out*.

7. We developed *snapshot.pl* when few alternatives existed to examine a genomic position, and we still use it as a lightweight tool. Currently, other more sophisticated tools exist, such as IGV (http://software.broadinstitute.org/software/igv/).
8. This primer only explores how to get variants from a single sample. To compare tumor and normal samples, the procedure can be followed with these techniques. First, we obtain the *mq* file from the tumor sample. Then, we extract the corresponding positions from the normal sample with a different script (not provided). The Sidrón script then computes the S values for each position in both the tumor and normal sample. The rest of the procedure is similar to the one described above, with the only distinction that we will look for high S values in the tumoral sample (Het) and low S values in the normal sample (Hz).
9. We have only considered point mutations in this procedure, where Sidrón adds resolving power. For small insertions and deletions, other considerations make this technique insufficient. For a primer on how to find insertions and deletions, *see* http://samtools.sourceforge.net/cns0.shtml.

Acknowledgment

We thank J.M.P. Freije and X.S. Puente for helpful comments and advice. The Instituto Universitario de Oncología is supported by Fundación Bancaria Caja de Ahorros de Asturias. V.Q. is supported by Ministerio de Economía y Competitividad and Gobierno del Principado de Asturias, including FEDER funding. C.L.-O. is supported by grants from European Research Council (DeAge, ERC Advanced Grant), Ministerio de Economía y Competitividad, Instituto de Salud Carlos III (RTICC) and Progeria Research Foundation.

References

1. Ferrando AA, López-Otín C (2017) Clonal evolution in leukemia. Nat Med 23 (10):1135–1145
2. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI et al (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 526 (7574):519–524
3. Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J et al (2015) Mutations driving CLL and their evolution in progression and relapse. Nature 526 (7574):525–530
4. Valdés-Mas R, Gutiérrez-Abril J, Puente XS, López-Otín C (2016) Chronic lymphocytic leukemia: looking into the dark side of the genome. Cell Death Differ 23(1):7–9
5. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M et al (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44(W1): W3–W10
6. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595
7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al (2010 Sep) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20 (9):1297–1303
8. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ et al (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28(3):311–317
9. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A et al (2016) The Ensembl variant effect predictor. Genome Biol 17 (1):122
10. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N et al (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475(7354):101–105
11. Puente XS, Quesada V, Osorio FG, Cabanillas R, Cadiñanos J, Fraile JM et al (2011) Exome sequencing and functional analysis identifies BANF1 mutation as the cause of a hereditary progeroid syndrome. Am J Hum Genet 88(5):650–656

SCIENTIFIC REPORTS
nature research

OPEN

Differential mechanisms of tolerance to extreme environmental conditions in tardigrades

Dido Carrero*, José G. Pérez-Silva, Víctor Quesada & Carlos López-Otín*

Tardigrades, also known as water bears, are small aquatic animals that inhabit marine, fresh water or limno-terrestrial environments. While all tardigrades require surrounding water to grow and reproduce, species living in limno-terrestrial environments (e.g. *Ramazzottius varieornatus*) are able to undergo almost complete dehydration by entering an arrested state known as anhydrobiosis, which allows them to tolerate ionic radiation, extreme temperatures and intense pressure. Previous studies based on comparison of the genomes of *R. varieornatus* and *Hypsibius dujardini* - a less tolerant tardigrade - have pointed to potential mechanisms that may partially contribute to their remarkable ability to resist extreme physical conditions. In this work, we have further annotated the genomes of both tardigrades using a guided approach in search for novel mechanisms underlying the extremotolerance of *R. varieornatus*. We have found specific amplifications of several genes, including *MRE11* and *XPC*, and numerous missense variants exclusive of *R. varieornatus* in *CHEK1*, *POLK*, *UNG* and *TERT*, all of them involved in important pathways for DNA repair and telomere maintenance. Taken collectively, these results point to genomic features that may contribute to the enhanced ability to resist extreme environmental conditions shown by *R. varieornatus*.

Tardigrades are small animals classically included in the clade Panarthropoda, together with Arthropoda and Onychophora. More than 1,200 species of tardigrades have been reported to inhabit all kinds of water environments. Even though they require surrounding water to grow and reproduce, limno-terrestrial tardigrades are well known for their remarkable capacity to endure extreme circumstances (such as dehydration, radiation, high and low temperature, high pressure, heavy metals and even outer-space conditions) when entering the anhydrobiotic state[1–6]. Nevertheless, some marine tardigrade species, such as *Echiniscoides sigismundi*, also present the ability to resist extreme dessication and intense gamma radiation[7,8]. Studies focused on survival and reproduction indicate that *R. varieornatus* presents a longer lifespan than *H. dujardini*[5].

The study of the genomic sequence of one of the most stress-tolerant limno-terrestrial tardigrade species, *R. varieornatus*, has reported genomic alterations such as the expansion of several stress-related genes and the selective loss of peroxisomal oxidative and autophagy-related pathways, which can contribute to their tolerance to extreme environmental conditions[9]. Parallel studies have addressed the genome characterization of freshwater tardigrades, such as *H. dujardini*, which are among the least desiccation-resistant members of the phylum Tardigrada[10], since they require previous conditioning to desiccation before entering anhydrobiosis. Such studies have also revealed various modifications in genes involved in macromolecule protection and stress signaling pathways that could contribute to the biological features exhibited by this tardigrade species, which lacks the extreme tolerance of *R. varieornatus*[11]. Other genomic comparative analyses have previously contributed to elucidate the mechanisms underlying aspects such as cancer resistance or longevity in different species[11–16].

These genomic data have also revealed in *R. varieornatus* the presence of a novel tardigrade-unique protein called Dsup (damage suppressor) that suppresses X-ray-induced DNA damage and improves radiotolerance[9]. Nonetheless, recent studies found a Dsup homologue in *H. dujardini* that, despite its weak similarity with *R. varieornatus* Dsup, also presents nuclear localization and similar profiles in hydrophobicity and charge distribution

Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología del Principado de Asturias (IUOPA), Universidad de Oviedo, 33006, Oviedo, Spain. *email: didocarrero94@gmail.com; clo@uniovi.es

along the protein[17]. This finding suggests that additional factors are involved in *R. varieornatus* extraordinary resistance to extreme conditions in comparison to *H. dujardini*, therefore encouraging the search for new hypotheses that explain the extremotolerance differences shown by these tardigrade species.

In this work, we have further explored the molecular mechanisms conferring extreme tolerance to limno-terrestrial tardigrades by comparing the genomes of *R. varieornatus* and *H. dujardini*, as well as that of a distant arthropod (*Drosophila melanogaster*). To this purpose, we have performed exhaustive manual annotation in these genomes of more than 250 genes involved in different DNA repair mechanisms. This comparative genomic analysis, together with the experimental validation of the identified alterations, has allowed us to detect specific gene amplifications and residue alterations in proteins involved in DNA repair pathways that may contribute to the enhanced tolerance to extreme environments exhibited by *R. varieornatus*.

Methods

Gene selection. Prior to genome annotation, we curated a list of more than 250 genes involved in oxygen homeostasis, stress response, telomere maintenance and DNA repair. Each gene was selected based on the experience of our laboratory in these fields[18–23], and following a detailed revision of the available publications on each subject.

Genomic analysis. We performed manual annotation of genomes *H. dujardini* (assembly 3.1, GCA_002082055.1) and *R. varieornatus* (assembly 4.0, GCA_001949185.1) using the BATI algorithm (Blast, Annotate, Tune, Iterate)[24], that allows researchers to annotate the coordinates and intron/exon boundaries of genes in novel genomes from Tblastn results. This procedure also enables the user to identify novel homologues. In addition to each genome, the algorithm was fed reference sequences from *D. melanogaster* and automatically-annotated *H. dujardini* (obtained from Ensembl and NCBI databases). This supporting information contributes to generate homology-based alignments that are later interpreted and revised manually, thus allowing the researcher to apply the experience in defying genes and obtaining better and more precise genomic structures (especially in the case of the aforementioned exon/intron boundaries). Once the selected genes were properly annotated, we compared the resulting sequences of *R. varieornatus* and *H. dujardini* to those of human, chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*), naked mole rat (*Heterocephalus glaber*), dog (*Canis lupus familiaris*), chicken (*Gallus gallus*), zebrafish (*Danio rerio*), Japanese rice fish (*Oryzias latipes*), coelacanth (*Latimeria chalumnae*), fruit fly *(D. melanogaster*) and roundworm (*Caenorhabditis elegans*) when available. This allowed the identification of gene expansions and losses, as well as residue changes specific of *R. varieornatus* and *H. dujardini*. In the alignment of TERT, we also included the HIV-1 RT sequence. In the alignment of POLK, we also included the prokaryotic species *Bdellovibrio bacteriovorus*, *Clostridium tetani*, *Escherichia coli*, *Mesorhizobium japonicum* and *Mycobacterium tuberculosis*. We evaluated the putative effects of these residue changes using data from NCBI Conserved Domains, UniProt and ClinVar databases.

PCR analysis. To validate copy-number variations of genes of interest that we obtained through manual annotation, we performed PCR reactions with primer pairs that amplified a target region of the genomes of *R. varieornatus* and *H. dujardini* with different nucleotide sequences in each copy (Supplementary Table 1), and then examined the resulting electropherogram for evidence of both copies. *R. varieornatus* tardigrades were kindly provided by Dr. Takekazu Kunieda, University of Tokyo, Japan, while *H. dujardini* tardigrades were obtained from Sciento. Samples consisted of 50 tardigrades per species, which were snap frozen with liquid nitrogen. DNA was extracted using the QIAamp DNA Micro Kit (Qiagen). We tested the success of the PCR reactions by electrophoresis of the resulting products in a 1.5% agarose gel. Finally, the products were sequenced using the Sanger method and an ABI PRISM 3130xl Genetic Analyzer (Thermofisher). The results of the manual annotation and PCR analysis were also confirmed through RNA-Seq data from *H. dujardini* and *R. varieornatus* present into the NCBI Sequence Read Archive (SRA).

Homology models. Homology models of selected proteins were performed with SWISS-MODEL[25] and used to evaluate the potential function of the residues analysed in this manuscript. The sequences of CHK1 and POLK from *R. varieornatus* were modelled using structure 1jx4 and 3jvr as a template, respectively. Similarly, the sequence of UNG from *R. varieornatus* was modelled using structure 1q3f as a template. The resulting structure was aligned to structure 1ssp to study its putative mode of interaction with a DNA substrate. The results were inspected and rendered with DeepView v4.1.0. Electric potentials were calculated with DeepView using the Poisson-Boltzmann computation method. Figures were generated with PovRay (http://povray.org) and UCSF Chimera[26].

Results and Discussion

Manual annotation of genes involved in DNA repair, stress response, telomere maintenance and oxygen homeostasis in tardigrades. To study the molecular mechanisms linked to the increased resistance to extreme environmental conditions shown by the tardigrade species *R. varieornatus* in comparison to *H. dujardini*, we analyzed a set of more than 250 genes involved in stress response, oxygen homeostasis, telomere maintenance and DNA repair (Table 1). Manual annotation of this gene set allowed us to find copy-number variations in genes related to DNA repair pathways, as well as to verify the previously described variations for both species of tardigrades. Interestingly, our analysis only revealed copy number variations between the two species of tardigrades in genes related to DNA repair mechanisms, particularly in genome maintenance during replication, double-strand break (DSB) repair, and nucleotide excision repair (NER) pathways (Table 2; Supplementary Table 2). However, no relevant copy number alterations were found in genes related to telomere maintenance, stress response or oxygen homeostasis when comparing the genomes of *R. varieornatus* and *H. dujardini*. Moreover, our analysis of DNA repair pathways in tardigrades and their comparison with reported data on human sequences led us to identify a series of residue changes that are exclusive of *R. varieornatus* and/or *H. dujardini* (Supplementary Table 3).

ADGB	CCNH	EIF2AK4	FANCD2	GTF2H1	JUNB	NGB	POLM	REV3	TP53
ALKBH2	CDK7	EIF2S1	FANCE	GTF2H2	JUND	NHEJ1	POLN	RIF1	TPP1
ALKBH3	CETN2	EIF2S2	FANCF	GTF2H3	LIG1	NHP2	POLQ	RMI2	TREX1
APEX1	CHAF1A	EIF2S3	FANCG	GTF2H4	LIG3	NOP10	POT1	RNF168	TREX2
APEX2	CHEK1	EME1	FANCI	GTF2H5	LIG4	NTHL1	PRKDC	RNF4	TSC1
APOLD1	CHEK2	EME2	FANCL	H2AFX	MAD2L2	NUDT1	PROC	RNF8	TSC2
APTX	CLK2	ENDOV	FANCM	HBA1	MB	ODF1	PRPF19	RPA1	UBE2A
ARNTL	CLOCK	ENOX1	FEN1	HBB	MBD4	OGG1	PTGS1	RPA2	UBE2B
ATM	CRY1	ENOX2	FOS	HBZ	MDC1	PALB2	PTGS2	RPA3	UBE2N
ATR	CRY2	EPAS1	FOSB	HELQ	MGMT	PARP1	RAD1	RPA4	UBE2V2
ATRIP	CRYAA	ERCC1	FOSL1	HIF1A	MLH1	PARP2	RAD17	RRP1	UNG
BAD	CRYAB	ERCC2	FOSL2	HIF1AN	MLH3	PARP3	RAD18	SEM1	UVSSA
BAK1	CTC1	ERCC3	FOXO1	HIF3A	MMS19	PCNA	RAD23A	SETMAR	VHL
BCL2A1	CYGB	ERCC4	FOXO3	HLTF	MNAT1	PER1	RAD23B	SHPRH	VHLL
BCL2L1	DCLRE1A	ERCC5	FOXO4	HP	MPG	PER2	RAD50	SLX1A	WRN
BCL2L10	DCLRE1B	ERCC6	FOXO6	HSBP1	MPLKIP	PLAT	RAD51	SLX4	XAB2
BCL2L11	DCLRE1C	ERCC8	GADD45A	HSF1	MRE11	PLAU	RAD51B	SMUG1	XPA
BCL2L12	DDB1	ERN1	GADD45B	HSF2	MSH2	PLG	RAD51C	SPO11	XPC
BCL2L13	DDB2	ERN2	GADD45G	HSF3	MSH3	PMS1	RAD51D	SPRTN	XRCC1
BCL2L14	DKC1	EXO1	GAR1	HSF4	MSH4	PMS2	RAD52	STN1	XRCC2
BCL2L15	DMC1	F10	GEN1	HSF5	MSH5	PNKP	RAD54B	TDG	XRCC3
BCL2L2	Dsup	F11	GPX1	HSPA	MSH6	POLB	RAD54L	TDP1	XRCC4
BLM	DUT	F7	GPX2	HSPA12A	MUS81	POLD1	RAD9A	TDP2	XRCC5
BNIP2	EGLN1	FAAP20	GPX3	HSPA12B	MUTYH	POLE	RBBP8	TEN1	XRCC6
BOK	EGLN2	FAAP24	GPX4	HSPB	NABP2	POLG	RDM1	TERF1	ZFAND2A
BRCA1	EGLN3	FAN1	GPX5	HSPH1	NBN	POLH	RECQL	TERF2	ZFAND2B
BRCA2	EIF2AK1	FANCA	GPX6	HUS1	NEIL1	POLI	RECQL4	TERT	
BRIP1	EIF2AK2	FANCB	GPX7	HYOU1	NEIL2	POLK	RECQL5	TINF2	
CAT	EIF2AK3	FANCC	GPX8	JUN	NEIL3	POLL	REV1	TOPBP1	

Table 1. List of genes analysed in this study.

Gene	Status in *R. varieornatus*	Status in *H. dujardini*	DNA repair mechanism
CHEK1	Residue change (p.F93Y)	No changes	DNA repair during replication, homologous recombination
LIG4	Amplification (two copies)	No changes	DNA repair during replication, non-homologous end joining
XPC	Amplification (two copies)	No changes	Nucleotide excision repair
MRE11	Amplification (four copies)	No changes	Non-homologous end joining, homologous recombination
UNG	Residue change (p.P177R)	No changes	Base excision repair
RAD51	Amplification (three copies)	No changes	Homologous recombination
ERCC4	Amplification (two copies)	No changes	Homologous recombination
POLK	Residue change (p.S132G)	No changes	Translesion synthesis
REV1	Residue change (p.A509S)	No changes	Translesion synthesis

Table 2. Genes showing copy-number variations or residue changes in *R. varieornatus* in comparison to *H. dujardini*, classified into the main repair mechanisms that they are involved in.

In this study, we focused on the description of copy number variations and residue changes exclusive of the extreme tolerant *R. varieornatus* that lay in active sites or DNA binding sites, and involve genes important for homologous recombination, base excision repair, nucleotide excision repair, non-homologous end-joining, translesion synthesis, DNA repair during replication (Table 2), and for telomere dynamics.

Telomere dynamics in *R. varieornatus* and *H. dujardini*. Telomeres have been widely studied in all Arthropoda, being their ancestral sequence $(TTAGG)_n$ common to hexapods, crustaceans, myriapods, pycnogonids and most chelicerates, but not to spiders[27]. Nonetheless, such repeat sequence is absent in Tardigrada and Onychophora, which are closely related to Arthropoda. Thus, Onychophora present the vertebrate motif $(TTAGGG)_n$, while tardigrades do not exhibit this telomere sequence either[27]. Further analysis of repeat sequences in the genome of *H. dujardini* revealed the presence of $(GATGGGTTTT)_n$ repeats, which were exclusively found at 9 scaffold ends and are thought to correspond to telomeric sequences[11] located in its 5 pairs of chromosomes[28]. Moreover, tardigrades and most arthropods lack the TERT motif CP, with the exceptions of hymenopterans and

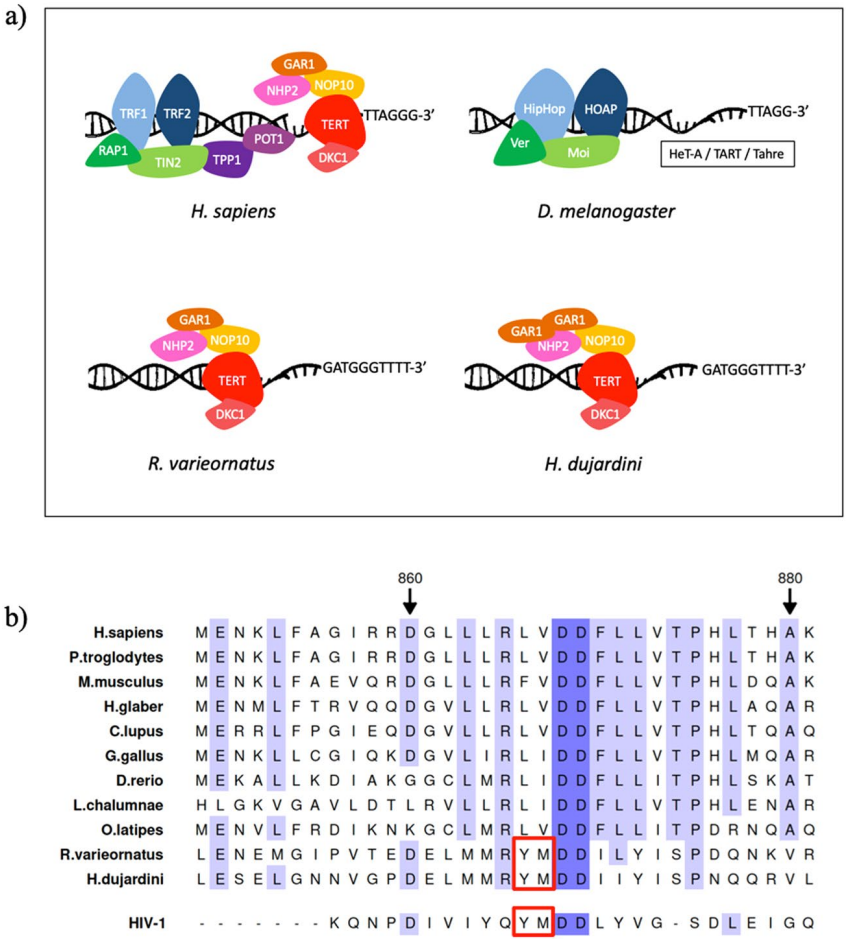

Figure 1. Telomere architecture in tardigrades compared to human and fruitfly. (**a**) Telomerase and telomere-capping complexes of human, fruitfly and tardigrades. Humans possess the shelterin complex (TRF1, TRF2, RAP1, TIN2, TPP1 and POT1), while *Drosophila* has the terminin complex (HipHop, HOAP, Ver and Moi), and tardigrades (*R. varieornatus* and *H. dujardini*) lack a telomere-capping complex. The telomerase complexes of humans and tardigrades are very similar, while in *Drosophila* telomeres replicate using a retrotransposon machinery composed of the elements HeT-A, TART and Tahre. (**b**) Partial amino acid sequence alignment of the TERT sequence in *R. varieornatus*, *H. dujardini* and other species of interest. Variants p.L866Y and p.V867M present in *R. varieornatus*, *H. dujardini* and HIV-1 reverse transcriptase are indicated with a red rectangle.

some centipedes[29]. This motif, together with the T motif, forms an extended pocket (T-CP pocket) on the surface of the protein implicated in RNA recognition and binding[30]. Remarkably, telomere elongation in *D. melanogaster* is carried out by three specialized retrotransposable elements (HeT-A, TART and Tahre)[31], while no ortholog for the human gene *TERT* has been reported. In addition, fruit fly telomeres are capped by the complex terminin, functionally but not structurally analogous to shelterin, which includes the proteins HOAP, HipHop, Moi and Ver[32,33] (Fig. 1a). These data indicate that telomere elongation and maintenance are carried out through different mechanisms in this species in contrast to other members of the Metazoa group.

In this work, we manually annotated several genes that encode proteins belonging to the telomerase, shelterin and CST complexes in tardigrades (Fig. 1a). Except for *TPP1*, none of the other components from the shelterin (*TERF1*, *TERF2*, *RAP1*, *POT1*, and *TINF2*) and CST (*CTC1*, *STN1* and *TEN1*) complexes were identified (Fig. 1a, CST complex not shown). Interestingly, we found in tardigrades a *bona fide TERT* ortholog, together with copies encoding all the elements of the telomerase complex, namely *NHP2*, *NOP10*, *DKC1* and *GAR1* (the latter being duplicated in *H. dujardini*) (Fig. 1a). Remarkably, two residue changes in TERT protein - p.L866Y and p.V867M - were found to be exclusively present in *H. dujardini* and *R. varieornatus* (Fig. 1b). Both residues are part of a tetrapeptide that includes a catalytically essential aspartate dyad (residues D868 and D869)[34]. These residues have been studied based on the previous discovery of the function of Y183 and M184, cognate amino acids to human TERT L866 and V867 in HIV-1 reverse transcriptase (Fig. 1b), which play important roles in processing, fidelity, enzymatic activity, dNTP utilization and nucleoside analogue inhibitor resistance[35]. These functional studies in human TERT have shown that the first variant alone (p.L866Y) results in a moderate reduction in telomerase activity, but produces no changes in repeat extension rate or in nucleotide incorporation fidelity[34]. The second variant (p.V867M) causes a 75% reduction in telomerase activity, 50% reduction in repeat extension rate, and

5.2-fold increase in nucleotide incorporation fidelity[34]. However, when both variants are present, they result in a slight reduction in telomerase activity and 13.5-fold increase in nucleotide incorporation fidelity[34]. This finding suggests that telomere dynamics in tardigrades may display reduced telomerase activity but also enhanced replication fidelity to prevent genomic instability caused by defects in telomere maintenance[20].

Alterations in genes involved in DNA repair and genome maintenance during replication in tardigrades. DNA ligation is essential for replication and repair, and genetic deficiencies in human DNA ligases have been associated with clinical syndromes characterized by radiation sensitivity and defects in DNA repair during replication through nonhomologous end joining (NHEJ)[36]. In mammals, this functional role is carried out by a protein family encoded by three genes (*LIG1*, *LIG3* and *LIG4*), all of them also present in *D. melanogaster*. While both tardigrade species seem to have one copy of *LIG1* and none of *LIG3*, we found two copies of *LIG4* in the genome of *R. varieornatus* (called *LIG4_1* and *LIG4_2*), while only one full copy and what could be one exon of another copy were detected in the genome of *H. dujardini*. The presence of this second *LIG4* copy in *H. dujardini* could not be verified by RNA-Seq nor Sanger sequencing due to the shortness of its contig (Supplementary Table 4), even though a putative expansion of *LIG4* in *H. dujardini* has been previously suggested[11]. Nevertheless, supporting data in this regard are not available in public repositories of genomic data[11]. Importantly, patients with null mutations in *LIG4* show increased sensitivity to ionizing radiation, as well as immunodeficiency, growth failure, and microcephaly[37]. In mice, Lig4 deficiency causes embryonic lethality due to a defective p53-dependent response to unrepaired DNA damage, as well as neuronal apoptosis and arrested lymphogenesis[38]. Moreover, mice with a hypomorphic mutation in *Lig4* show high levels of DNA DSBs during embryonic development and a deficient DSB repair response[39]. Accordingly, LIG4 mediates Wnt/β-catenin signaling activation during radiation-induced intestinal regeneration and blocking LIG4 sensitizes colorectal cancer cells to radiation[40]. Since the second copy of *H. dujardini* is not experimentally supported, it is plausible that the exclusive presence of two copies of *LIG4* in *R. varieornatus* might contribute to its enhanced resistance to DNA damage.

Moreover, we found several remarkable residue changes in *CHEK1* (Supplementary Table 3), which codes for the protein kinase CHK1 involved in DNA damage response (DDR), cell cycle arrest, and homologous recombination (HR)[41]. Among these *CHEK1* variants, we focused our attention on p.F93Y, exclusive of *R. varieornatus* (Fig. 2a), which affects an active site that functions as an allosteric inhibitor binding site and as a polypeptide substrate binding site[42]. To explore the putative functional relevance of this change, we generated a homology model of this protein in *R. varieornatus* (Fig. 2b). This model revealed that position 93 is located at the surface of a pocket in which allosteric inhibitors can be fitted, and showed the potential of the residue Y93 to form an H-bond with a synthetic allosteric inhibitor (Fig. 2b)[42]. This amino acidic change might influence the allosteric regulation of CHEK1 in *R. varieornatus* in comparison to *H. dujardini*. This regulatory mechanism may be important for its function, since CHK1 is involved in DNA damage response (DDR), cell cycle arrest, and homologous recombination (HR)[41].

We also found an alteration (p.S132G) in the polymerase POLK exclusive of *R. varieornatus* (Fig. 3a), together with other residue changes shared with *H. dujardini* (Supplementary Table 3). The p.S132G variant affects a residue involved in DNA binding[43]. POLK is an error-prone DNA polymerase specifically involved in translesion synthesis during DNA replication, which preferentially incorporates adenine residues opposite to 8-oxoguanine lesions. These lesions frequently appear as a result of ionizing radiation, therefore producing missense mutations and frameshifts[43,44]. POLK appears to be absent in all arthropods. Its prokaryotic ortholog, DNA polymerase IV[45], is also involved in repair of 8-oxoguanine lesions, but incorporates cytosine instead of adenine opposite to 8-oxoguanine with high efficiency, thus avoiding potential mutations[46]. Notably, prokaryotic DNA polymerase IV also presents glycine instead of serine in residue 132 (Fig. 3a), which suggests that the presence of glycine may contribute to incorporating the right nucleotide during repair of 8-oxoguanine lesions, resulting in higher fidelity and decreasing the occurrence of point mutations. The homology model of this protein in *R. varieornatus* suggests that, although the position 132 is not strictly close to the 8-oxoguanine lesion, it contributes to creating a more acute beta turn (Fig. 3b). Finally, REV1 - another protein involved in translesion synthesis[47] - presents a variant exclusive of *R. varieornatus* affecting a DNA binding site (p.A509S)[48,49]. Additionally, *R. varieornatus* REV1 presents other changes in DNA binding sites that are also found in *H. dujardini* (Supplementary Table 3).

Finally, the gene *MGMT*, which encodes a methyltransferase involved in repairing the naturally occurring mutations O^6-methylguanine and O^4-methylthymine during replication[50], is present in *H. dujardini* but the corresponding ortholog in *R. varieornatus* had not been previously identified in manual and automatic annotations. However, we could confirm the presence of *MGMT* when performing PCR on the genome of *R. varieornatus* using oligonucleotides based on the corresponding *MGMT* sequence of *H. dujardini* (Supplementary Table 4). Accordingly, its apparent absence in *R. varieornatus* genome is likely due to errors in the currently available genome assembly for this tardigrade.

Expansion of genes involved in double-strand break repair in tardigrades. DSBs are particularly damaging alterations, since they can lead to chromosome rearrangements and losses. These genomic lesions can be repaired through three mechanisms: NHEJ, HR and microhomology-mediated end joining (MMEJ)[51]. We confirmed that the human *MRE11* ortholog, involved in NHEJ and HR[52], is at least quadrupled in *R. varieornatus*, while *H. dujardini* displays one copy (Supplementary Table 4), as it has previously been reported[9,11]. The remarkable expansion of this gene may be responsible for an enhanced ability to repair DNA damage[53]. Moreover, knockdown of *MRE11* impaired DSB repair in HeLa and CNE2 cells[54], and upregulation of this protein in cancer cells following ionizing radiation promoted DNA repair[54]. Altogether, these data suggest an important role of *MRE11* ortholog in *R. varieornatus* in promoting DNA repair after exposure to ionizing radiation.

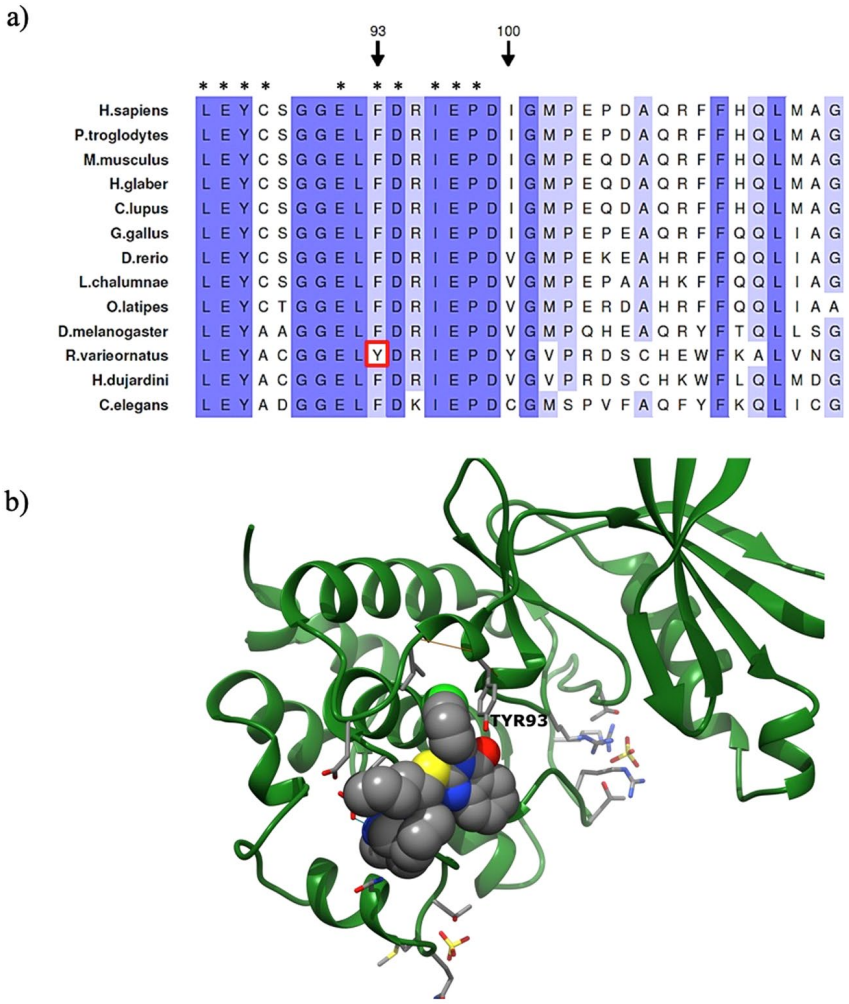

Figure 2. Comparative sequence analysis and homology modeling of CHK1 from *R. varieornatus*. (**a**) Partial amino acid sequence alignment of the CHK1 sequence in *R. varieornatus*, *H. dujardini* and other species of interest. Variant p.F93Y present in *R. varieornatus* is highlighted with a red rectangle. Important residues for its function are marked with *. (**b**) Representative image of the residue Y93 in the homology model of CHK1 from *R. varieornatus*. The homology model shows that the residue Y93, exclusive of *R. varieornatus*, that is defined in its wild-type form (F93) as an allosteric inhibitor binding site, is able to form an H-bond with the allosteric inhibitor that cannot be formed in its wild-type form (F93).

We also confirmed the previous finding that the RAD51 protein family, involved in DSB repair through HR[55], is expanded in *R. varieornatus*[9]. However, according to our data, we propose that one of the four copies annotated in this tardigrade's genome by Hashimoto *et al.* actually corresponds to the *XRCC2* ortholog, as assessed by performing blast of these sequences (deposited in the NCBI database) against the human genome. Therefore, according to our annotation, the genome of *R. varieornatus* contains three copies of RAD51. We independently found the presence of the other three copies in both tardigrades. Expansion of the DNA repair endonuclease XPF (encoded by the gene *ERCC4*), also involved in HR[56], was reported in *H. dujardini*, since five copies of this gene were found in its genome[11]. However, only three sequences out of these five could be found in the NCBI database, two of which belong to very small polypeptides (<100 aa); and only one is available at Ensembl Tardigrades[11]. In turn, manual annotation of this gene revealed two copies of *ERCC4* in this species (named *ERCC4_1* and *ERCC4_2*), while only one copy was found in *R. varieornatus*. This duplication could be verified by RNA-Seq, but not using Sanger sequencing due to the high similarity between both copies, and the presence of repetitive sequences (Supplementary Table 4). Finally, and similarly to the case of *MGMT*, one copy of the gene *XRCC3*, also involved in HR, could be found in the genome of *H. dujardini*. Although this gene seemed to be absent in the genome of *R. varieornatus*, we detected it by PCR using oligonucleotides designed for *H. dujardini* (Supplementary Table 4).

Changes in genes related to base excision repair in *R. varieornatus*. Among all genes involved in base excision repair (BER) analysed in *R. varieornatus* and *H. dujardini*, we found a variant in an active site and UGI (uracil-DNA glycosylase inhibitor protein) interface site (p.P177R) in the protein encoded by *UNG* that is

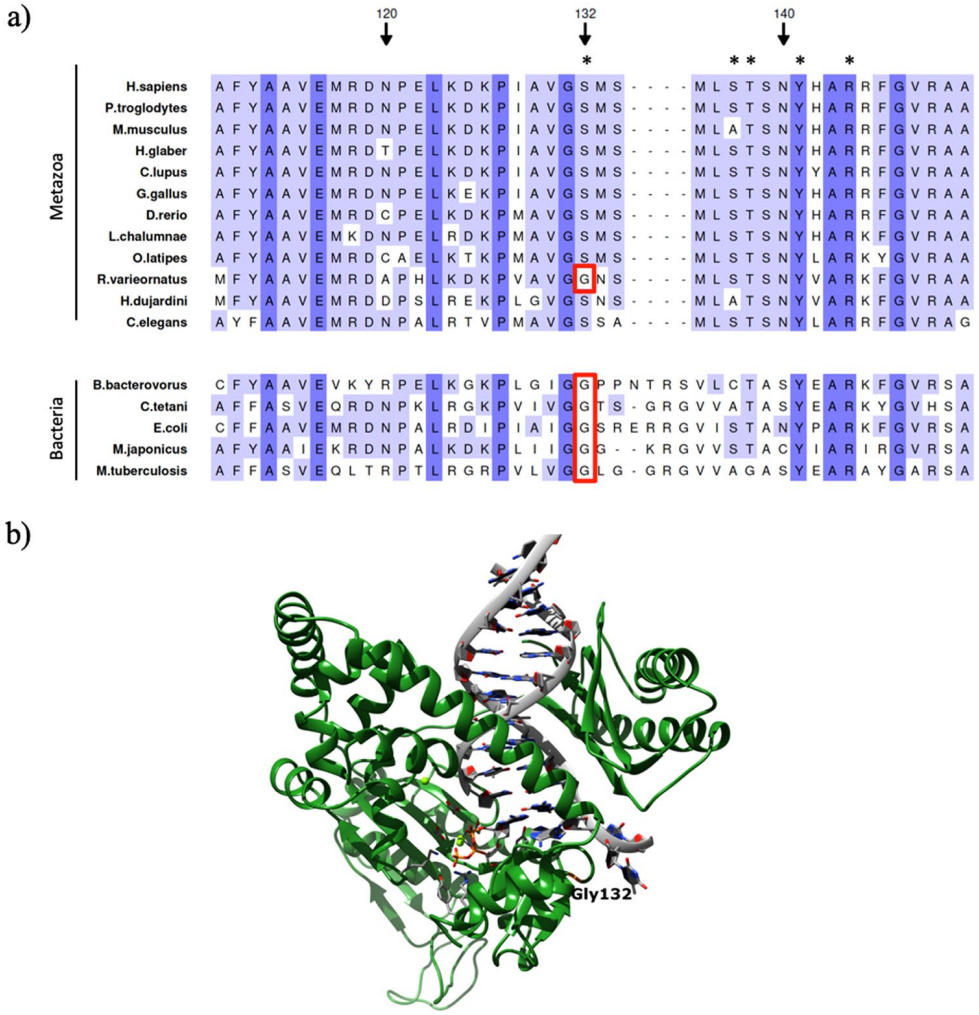

Figure 3. Comparative sequence analysis and homology modeling of POLK from *R. varieornatus*. (**a**) Partial amino acid sequence alignment of the POLK sequence in *R. varieornatus*, *H. dujardini* and other species of interest. Variant p.S132G present in *R. varieornatus* is indicated with a red rectangle. Important residues for its function are marked with *. (**b**) Representative image of the residue G132 in the homology model of POLK from *R. varieornatus*. The homology model shows that the residue G132, exclusive of *R. varieornatus*, that is defined in its wild-type form (S132) as DNA binding site, creates a more acute beta turn in the protein.

exclusive of *R. varieornatus*[57] (Fig. 4a). This protein is a DNA glycosylase that excises uracil residues from DNA when misincorporation of uracil occurs during DNA replication or due to deamination of cytosine[58]. The model predicts that Arg 177 fits the minor groove of the DNA molecule, very close to the everted base (Fig. 4b). This mode of interaction has been described previously in the context of the nucleosome, and it was found to be independent of the DNA sequence[59], which suggests that this variant might contribute to the association of UNG to substrate DNA. In this regard, another tardigrade-specific arginine at position 256 (Fig. 4a) interacts with a phosphate group at the other side of the everted base. However, given the proximity of Arg177 to the substrate base, this model cannot rule out the possibility that this residue might also play a role in base eversion, as proposed in similar contexts for other enzymes[60].

Nucleotide excision repair in *R. varieornatus*. Oxidative DNA damage is considered as a leading cause of both neurodegeneration and cancer development as illustrated by syndromes that result from NER defects, such as Xeroderma pigmentosum (XP) and Cockayne syndrome (CS)[61,62]. Among all the genes involved in NER, *XPC* appears to be duplicated in *R. varieornatus* (with copies we have named *XPC_1* and *XPC_2*) but not in *H. dujardini* (Supplementary Table 4). This protein is involved in repair of damage caused by UV light, since mutations in the gene encoding this protein in humans lead to XP[61], and *Xpc* knockout mice show an increased susceptibility to UVB induced squamous cell carcinomas[63]. Therefore, this duplication in the *XPC* ortholog in *R. varieornatus* may also contribute to the enhanced tolerance to radiation in this species by improving its NER response pathway.

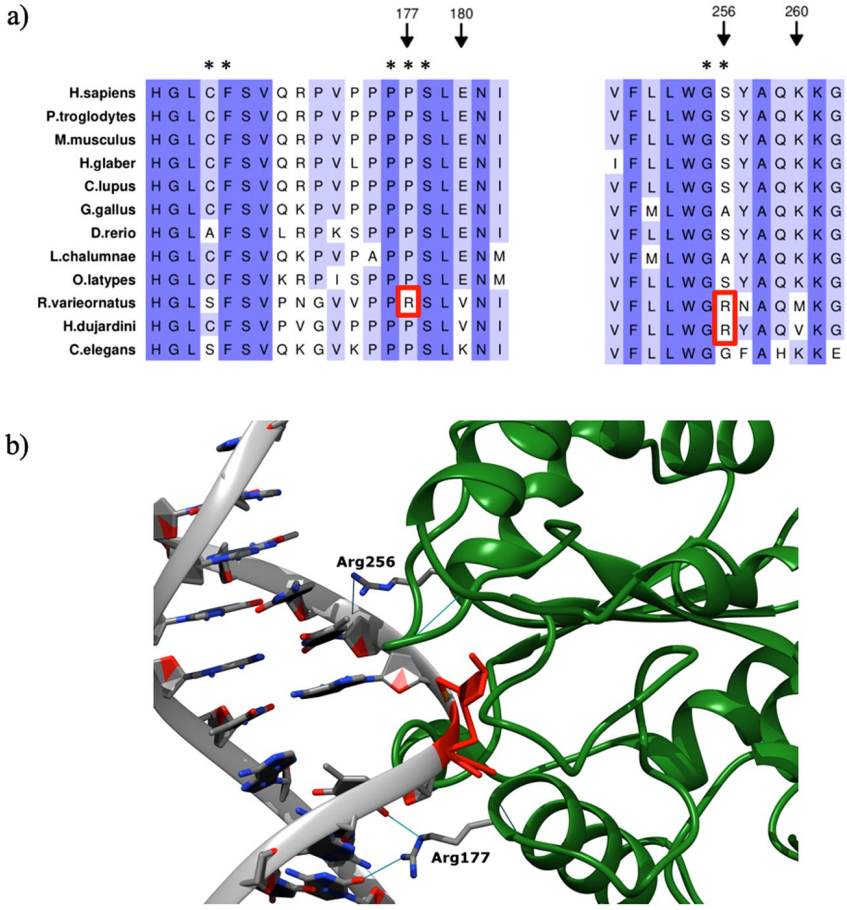

Figure 4. Comparative sequence analysis and homology modeling of *R. varieornatus* UNG bound to DNA. (**a**) Partial amino acid sequence alignment of the UNG sequence in *R. varieornatus*, *H. dujardini* and other species of interest. Variants p.P177R and p.S256R present in *R. varieornatus* are highlighted with a red rectangle. Important residues for its function are marked with *. (**b**) The enzyme is shown as a green ribbon. A DNA intermediate from structure 1ssp is shown in grey. The sugar from the substrate base is shown in red. UNG arginines 177 (specific of *R. varieornatus*) and 256 (specific of tardigrades) are labelled. Putative interactions involving R177 or R256 are shown as blue lines.

Summary. In this manuscript, we describe several gene expansions of pivotal elements in DNA repair pathways observed in the genomes of *R. varieornatus* and *H. dujardini* through manual annotation, including previously described expansions, such as *XPC*, *LIG4*, *ERCC4* and *MRE11*[9,11]. Manual genomic comparative analyses also revealed residue changes in key elements in DNA repair pathways that in the corresponding human orthologs are known to cause an effect in the function of the protein (Supplementary Table 3), among which we highlight the ones exclusively found in *R. varieornatus* in the genes *TERT*, *CHEK1*, *POLK* and *UNG*. However, considering the phylogenetic distance between tardigrades and humans, in most cases it is difficult to define the consequences of such variants in tardigrade proteins, and further experimental work is required to raise definitive conclusions in this regard. Nonetheless, these findings show that combining both manual and automatic annotation approaches is an advantageous strategy to better determinate the precise number of gene copies and to find residue changes when analyzing a genome *de novo*.

In short, all the changes we observed in *R. varieornatus* suggest an enhanced ability to maintain genomic stability, which may explain its resistance to extreme conditions, as well as its longer lifespan in comparison to *H. dujardini*. Additionally, the recent finding of a Dsup homologue in *H. dujardini*[17] reinforces our proposal that specific features in DNA repair genes are important elements in the extraordinary resistance shown by this limno-terrestrial tardigrade species.

Data availability

The manually annotated dataset of genes and proteins generated and analyzed during the current study supporting the conclusions of this article are available in a public repository in GitHub(https://github.com/EreboPSilva/rvar.hduj.prots).

Received: 21 March 2019; Accepted: 29 September 2019;

Published online: 17 October 2019

References

1. Jönsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M. & Rettberg, P. Tardigrades survive exposure to space in low Earth orbit. *Curr Biol* **18**, R729–R731, https://doi.org/10.1016/j.cub.2008.06.048 (2008).
2. Hengherr, S., Worland, M. R., Reuner, A., Brummer, F. & Schill, R. O. Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. *J Exp Biol* **212**, 802–807, https://doi.org/10.1242/jeb.025973 (2009).
3. Hengherr, S., Worland, M. R., Reuner, A., Brummer, F. & Schill, R. O. High-temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. *Physiol Biochem Zool* **82**, 749–755, https://doi.org/10.1086/605954 (2009).
4. Mobjerg, N. *et al.* Survival in extreme environments - on the current knowledge of adaptations in tardigrades. *Acta Physiol (Oxf)* **202**, 409–420, https://doi.org/10.1111/j.1748-1716.2011.02252.x (2011).
5. Horikawa, D. D. *et al.* Analysis of DNA repair and protection in the Tardigrade Ramazzottius varieornatus and Hypsibius dujardini after exposure to UVC radiation. *PLoS One* **8**, e64793, https://doi.org/10.1371/journal.pone.0064793 (2013).
6. Hygum, T. L. *et al.* Comparative investigation of copper tolerance and identification of putative tolerance related genes in tardigrades. *Front Physiol* **8**, 95, https://doi.org/10.3389/fphys.2017.00095 (2017).
7. Jönsson, K. I., Hygum, T. L., Andersen, K. N., Clausen, L. K. & Mobjerg, N. Tolerance to gamma radiation in the marine heterotardigrade, Echiniscoides sigismundi. *PLoS One* **11**, e0168884, https://doi.org/10.1371/journal.pone.0168884 (2016).
8. Sorensen-Hygum, T. L., Stuart, R. M., Jorgensen, A. & Mobjerg, N. Modelling extreme desiccation tolerance in a marine tardigrade. *Sci Rep* **8**, 11495, https://doi.org/10.1038/s41598-018-29824-6 (2018).
9. Hashimoto, T. *et al.* Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat Commun* **7**, 12808, https://doi.org/10.1038/ncomms12808 (2016).
10. Boothby, T. C. *et al.* Tardigrades use intrinsically disordered proteins to survive desiccation. *Mol Cell* **65**, 975–984 e975, https://doi.org/10.1016/j.molcel.2017.02.018 (2017).
11. Yoshida, Y. *et al.* Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus. *PLoS Biol* **15**, e2002266, https://doi.org/10.1371/journal.pbio.2002266 (2017).
12. Seluanov, A., Gladyshev, V. N., Vijg, J. & Gorbunova, V. Mechanisms of cancer resistance in long-lived mammals. *Nat Rev Cancer* **18**, 433–441, https://doi.org/10.1038/s41568-018-0004-9 (2018).
13. Gorbunova, V., Seluanov, A., Zhang, Z., Gladyshev, V. N. & Vijg, J. Comparative genetics of longevity and cancer: insights from long-lived rodents. *Nat Rev Genet* **15**, 531–540, https://doi.org/10.1038/nrg3728 (2014).
14. Ma, S. *et al.* Comparative transcriptomics across 14 Drosophila species reveals signatures of longevity. *Aging Cell*, e12740, https://doi.org/10.1111/acel.12740 (2018).
15. Tollis, M., Schiffman, J. D. & Boddy, A. M. Evolution of cancer suppression as revealed by mammalian comparative genomics. *Curr Opin Genet Dev* **42**, 40–47, https://doi.org/10.1016/j.gde.2016.12.004 (2017).
16. Doherty, A. & de Magalhaes, J. P. Has gene duplication impacted the evolution of Eutherian longevity? *Aging Cell* **15**, 978–980, https://doi.org/10.1111/acel.12503 (2016).
17. Hashimoto, T. & Kunieda, T. DNA Protection Protein, a Novel Mechanism of Radiation Tolerance: Lessons from Tardigrades. *Life (Basel)* **7**, https://doi.org/10.3390/life7020026 (2017).
18. Warren, W. C. *et al.* The novel evolution of the sperm whale genome. *Genome Biol Evol* **9**, 3260–3264, https://doi.org/10.1093/gbe/evx187 (2017).
19. Keane, M. *et al.* Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep* **10**, 112–122, https://doi.org/10.1016/j.celrep.2014.12.008 (2015).
20. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217, https://doi.org/10.1016/j.cell.2013.05.039 (2013).
21. Puente, X. S. *et al.* Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics* **7**, 15, https://doi.org/10.1186/1471-2164-7-15 (2006).
22. Ferrando, A. A. & Lopez-Otin, C. Clonal evolution in leukemia. *Nat Med* **23**, 1135–1145, https://doi.org/10.1038/nm.4410 (2017).
23. Quesada, V. *et al.* Giant tortoise genomes provide insights into longevity and age-related disease. *Nat Ecol Evol*, https://doi.org/10.1038/s41559-018-0733-x (2018).
24. Quesada, V., Velasco, G., Puente, X. S., Warren, W. C. & Lopez-Otin, C. Comparative genomic analysis of the zebra finch degradome provides new insights into evolution of proteases in birds and mammals. *BMC Genomics* **11**, 220, https://doi.org/10.1186/1471-2164-11-220 (2010).
25. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* **42**, W252–258, https://doi.org/10.1093/nar/gku340 (2014).
26. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612, https://doi.org/10.1002/jcc.20084 (2004).
27. Vitkova, M., Kral, J., Traut, W., Zrzavy, J. & Marec, F. The evolutionary origin of insect telomeric repeats, (TTAGG)n. *Chromosome Res* **13**, 145–156, https://doi.org/10.1007/s10577-005-7721-0 (2005).
28. Gabriel, W. N. *et al.* The tardigrade Hypsibius dujardini, a new model for studying the evolution of development. *Dev Biol* **312**, 545–559, https://doi.org/10.1016/j.ydbio.2007.09.055 (2007).
29. Lai, A. G. *et al.* The protein subunit of telomerase displays patterns of dynamic evolution and conservation across different metazoan taxa. *BMC Evol Biol* **17**, 107, https://doi.org/10.1186/s12862-017-0949-4 (2017).
30. Gillis, A. J., Schuller, A. P. & Skordalakes, E. Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature* **455**, 633–637, https://doi.org/10.1038/nature07283 (2008).
31. Mason, J. M., Frydrychova, R. C. & Biessmann, H. Drosophila telomeres: an exception providing new insights. *Bioessays* **30**, 25–37, https://doi.org/10.1002/bies.20688 (2008).
32. Cicconi, A. *et al.* The Drosophila telomere-capping protein Verrocchio binds single-stranded DNA and protects telomeres from DNA damage response. *Nucleic Acids Res* **45**, 3068–3085, https://doi.org/10.1093/nar/gkw1244 (2017).
33. Raffa, G. D., Ciapponi, L., Cenci, G. & Gatti, M. Terminin: a protein complex that mediates epigenetic maintenance of Drosophila telomeres. *Nucleus* **2**, 383–391, https://doi.org/10.4161/nucl.2.5.17873 (2011).
34. Drosopoulos, W. C. & Prasad, V. R. The active site residue Valine 867 in human telomerase reverse transcriptase influences nucleotide incorporation and fidelity. *Nucleic Acids Res* **35**, 1155–1168, https://doi.org/10.1093/nar/gkm002 (2007).
35. Harris, D., Yadav, P. N. & Pandey, V. N. Loss of polymerase activity due to Tyr to Phe substitution in the YMDD motif of human immunodeficiency virus type-1 reverse transcriptase is compensated by Met to Val substitution within the same motif. *Biochemistry* **37**, 9630–9640, https://doi.org/10.1021/bi980549z (1998).
36. Ben-Omran, T. I., Cerosaletti, K., Concannon, P., Weitzman, S. & Nezarati, M. M. A patient with mutations in DNA Ligase IV: clinical features and overlap with Nijmegen breakage syndrome. *Am J Med Genet A* **137A**, 283–287, https://doi.org/10.1002/ajmg.a.30869 (2005).
37. IJspeert, H. *et al.* Clinical spectrum of LIG4 deficiency is broadened with severe dysmaturity, primordial dwarfism, and neurological abnormalities. *Hum Mutat* **34**, 1611–1614, https://doi.org/10.1002/humu.22436 (2013).
38. Frank, K. M. *et al.* DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. *Mol Cell* **5**, 993-1002 (2000).

39. Barazzuol, L. & Jeggo, P. A. *In vivo* sensitivity of the embryonic and adult neural stem cell compartments to low-dose radiation. *J Radiat Res* **57**(Suppl 1), i2–i10, https://doi.org/10.1093/jrr/rrw013 (2016).
40. Jun, S. *et al*. LIG4 mediates Wnt signalling-induced radioresistance. *Nat Commun* **7**, 10994, https://doi.org/10.1038/ncomms10994 (2016).
41. McNeely, S., Beckmann, R. & Bence Lin, A. K. CHEK again: revisiting the development of CHK1 inhibitors for cancer therapy. *Pharmacol Ther* **142**, 1–10, https://doi.org/10.1016/j.pharmthera.2013.10.005 (2014).
42. Vanderpool, D. *et al*. Characterization of the CHK1 allosteric inhibitor binding site. *Biochemistry* **48**, 9823–9830, https://doi.org/10.1021/bi900258v (2009).
43. Vasquez-Del Carpio, R. *et al*. Structure of human DNA polymerase kappa inserting dATP opposite an 8-OxoG DNA lesion. *PLoS One* **4**, e5766, https://doi.org/10.1371/journal.pone.0005766 (2009).
44. Pillaire, M. J., Betous, R. & Hoffmann, J. S. Role of DNA polymerase kappa in the maintenance of genomic stability. *Mol Cell Oncol* **1**, e29902, https://doi.org/10.4161/mco.29902 (2014).
45. Gerlach, V. L., Feaver, W. J., Fischhaber, P. L. & Friedberg, E. C. Purification and characterization of pol kappa, a DNA polymerase encoded by the human DINB1 gene. *J Biol Chem* **276**, 92–98, https://doi.org/10.1074/jbc.M004413200 (2001).
46. Raper, A. T., Gadkari, V. V., Maxwell, B. A. & Suo, Z. Single-molecule investigation of response to oxidative DNA damage by a Y-family DNA polymerase. *Biochemistry* **55**, 2187–2196, https://doi.org/10.1021/acs.biochem.6b00166 (2016).
47. Tellier-Lebegue, C. *et al*. The translesion DNA polymerases Pol zeta and Rev1 are activated independently of PCNA ubiquitination upon UV radiation in mutants of DNA polymerase delta. *PLoS Genet* **13**, e1007119, https://doi.org/10.1371/journal.pgen.1007119 (2017).
48. Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S. & Aggarwal, A. K. Rev1 employs a novel mechanism of DNA synthesis using a protein template. *Science* **309**, 2219–2222, https://doi.org/10.1126/science.1116336 (2005).
49. Swan, M. K., Johnson, R. E., Prakash, L., Prakash, S. & Aggarwal, A. K. Structure of the human Rev1-DNA-dNTP ternary complex. *J Mol Biol* **390**, 699–709, https://doi.org/10.1016/j.jmb.2009.05.026 (2009).
50. Iyama, T. & Wilson, D. M. 3rd DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst)* **12**, 620–636, https://doi.org/10.1016/j.dnarep.2013.04.015 (2013).
51. Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol* **26**, 52–64, https://doi.org/10.1016/j.tcb.2015.07.009 (2016).
52. Stracker, T. H. & Petrini, J. H. The MRE11 complex: starting from the ends. *Nat Rev Mol Cell Biol* **12**, 90–103, https://doi.org/10.1038/nrm3047 (2011).
53. Takeda, S., Hoa, N. N. & Sasanuma, H. The role of the Mre11-Rad50-Nbs1 complex in double-strand break repair-facts and myths. *J Radiat Res* **57**(Suppl 1), i25–i32, https://doi.org/10.1093/jrr/rrw034 (2016).
54. Deng, R. *et al*. PKB/Akt promotes DSB repair in cancer cells through upregulating Mre11 expression following ionizing radiation. *Oncogene* **30**, 944–955, https://doi.org/10.1038/onc.2010.467 (2011).
55. Inano, S. *et al*. RFWD3-mediated ubiquitination promotes timely removal of both RPA and RAD51 from DNA damage sites to facilitate homologous recombination. *Mol Cell* **66**, 622–634 e628, https://doi.org/10.1016/j.molcel.2017.04.022 (2017).
56. Svendsen, J. M. *et al*. Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. *Cell* **138**, 63–77, https://doi.org/10.1016/j.cell.2009.06.030 (2009).
57. Roberts, V. A., Pique, M. E., Hsu, S. & Li, S. Combining H/D exchange mass spectrometry and computational docking to derive the structure of protein-protein complexes. *Biochemistry* **56**, 6329–6342, https://doi.org/10.1021/acs.biochem.7b00643 (2017).
58. Zharkov, D. O., Mechetin, G. V. & Nevinsky, G. A. Uracil-DNA glycosylase: Structural, thermodynamic and kinetic aspects of lesion search and recognition. *Mutat Res* **685**, 11–20, https://doi.org/10.1016/j.mrfmmm.2009.10.017 (2010).
59. West, S. M., Rohs, R., Mann, R. S. & Honig, B. Electrostatic interactions between arginines and the minor groove in the nucleosome. *J Biomol Struct Dyn* **27**, 861–866, https://doi.org/10.1080/07391102.2010.10508587 (2010).
60. Shieh, F. K., Youngblood, B. & Reich, N. O. The role of Arg165 towards base flipping, base stabilization and catalysis in M.HhaI. *J Mol Biol* **362**, 516–527, https://doi.org/10.1016/j.jmb.2006.07.030 (2006).
61. Carrero, D., Soria-Valles, C. & Lopez-Otin, C. Hallmarks of progeroid syndromes: lessons from mice and reprogrammed cells. *Dis Model Mech* **9**, 719–735, https://doi.org/10.1242/dmm.024711 (2016).
62. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol* **15**, 465–481, https://doi.org/10.1038/nrm3822 (2014).
63. Sands, A. T., Abuin, A., Sanchez, A., Conti, C. J. & Bradley, A. High susceptibility to ultraviolet-induced carcinogenesis in mice lacking XPC. *Nature* **377**, 162–165, https://doi.org/10.1038/377162a0 (1995).

Acknowledgements

We thank Alicia R. Folgueras and J.M. Freije for helpful comments and advice, and T. Kunieda for providing valuable biological materials. This work was supported by grants from European Research Council (DeAge, ERC Advanced Grant), Ministerio de Economía y Competitividad, Instituto de Salud Carlos III (Ciberonc) and Progeria Research Foundation. The Instituto Universitario de Oncología is supported by Fundación Bancaria Caja de Ahorros de Asturias.

Author contributions

D.C. performed manual annotation of genomes, data interpretation and preparation of the manuscript. J.G.P.S. prepared the analyzed genomes for their manual annotation and performed bioinformatic analyses. V.Q. performed protein modelling and supervised data interpretation. C.L.O. supervised research and project planning, data interpretation and preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-51471-8.

Correspondence and requests for materials should be addressed to D.C. or C.L.-O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.