НОВ БЪЛГАРСКИ УНИВЕРСИТЕТ

Департамент Информатика

Бакалавърска програма Информатика

Автоматизиран биоинформатичен анализ на генетични варианти, потенциално свързани със стареенето

Дипломна работа на

Михаил М. Здравков

Научни ръководители:
доц. д-р Милена Георгиева
Момчил Топалов

Дипломен консултант:
гл. ас. д-р Методи Трайков

София 2022

Съдържание

Използвани съкращения

VCF - Variant Call Format. Стандартен файлов формат за описване на генни варианти спрямо определен референтен геном.

1. Увод

Стареенето е естествен процес, който има огромно значение както за отделния индивид, така и за обществото като цяло. С напредването на възрастта, рискът от разнообразни заболявания като рак, болест на Алцхаймер, диабет, сърдечно-съдови заболявания и др. нараства значително. Смята се, че около две-трети от смъртните случаи при хора се дължат на заболявания, свързани с възрастта. Същевременно, с глобалното нарастване на средната продължителност на живота, проблемите на стареенето засягат все повече хора и имат все по-голямо обществено значение. От социална гледна точка, стареенето оказва значителен икономически и демографски ефект.

Установено е, че процесът на стареене се влияе както от генетични, така и от епигенетични фактори. Въпреки това, този процес все още не е достатъчно добре разбран от науката, поради което е трудно да се създадат ефективни методи за терапия и справяне с негативните му ефекти.

Настоящата дипломна работа се фокусира върху генетичната основа на стареенето. Основен подход при нейното изследване е анализът на генетични варианти. При такива изследвания е необходима обработката на големи обеми от данни, което налага нуждата от използване на специализиран биоинформатичен софтуер. Налични са множество различни инструменти, покриващи различни аспекти от обработката на файлове с генетични варианти - анотация, филтриране, анализ и тн. Повечето от тях, обаче, изискват значителни технически познания, което ги прави трудни за използване от специалисти в други области, като биология и генетика.

Целта на настоящата дипломна работа е създаването на интегрирана софтуерна система за биоинформатични изследвания на генетични варианти и предсказване на тяхната потенциална асоциация с процеса на стареене. Надяваме се, чрез създаване на по-достъпен инструмент, да допринесем за

бъдещи изследвания на процеса на стареене и за търсенето на ефективни терапии против негативните му ефекти.

2. Литературен обзор

2.1 Значение на стареенето

2.1.1 Дефиниция

Въпреки, че концепцията за стареене е универсално разбираема, формалната ѝ дефиниция не е тривиална и множество автори дават твърде различни определения за този термин. Аркинг (2006, стр. 11) прави преглед на наличната литература и, в резултат, предлага следната дефиниция [1]:

„Стареенето е независима от времето поредица от кумулативни, прогресивни, свойствени и вредящи структурни и функционални промени, които обикновено започват да се изразяват при репродуктивната зрялост и приключват със смъртта.“

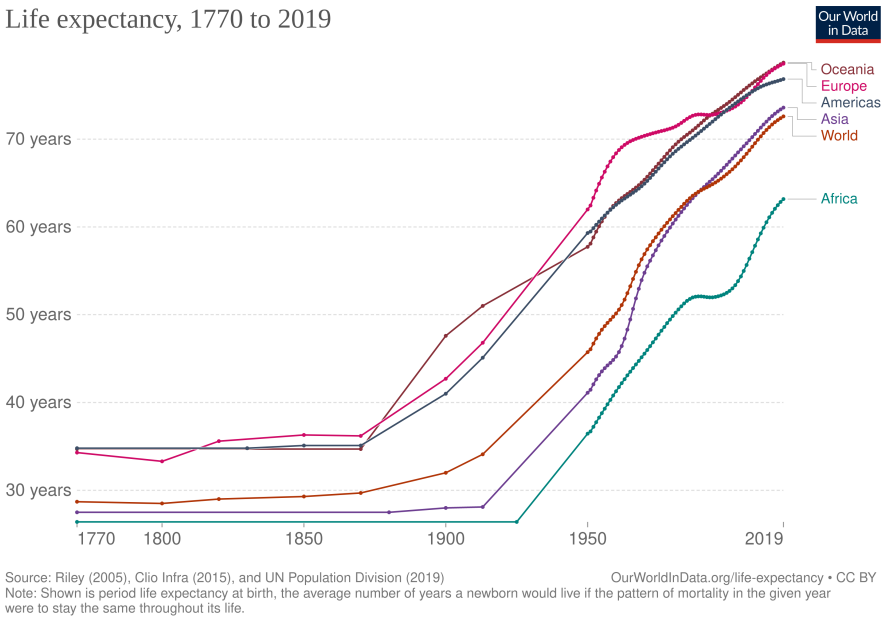
Макар времето да няма каузална връзка с ефектите на стареенето, то корелацията помежду им е причина обикновено да се говори за ефектите на стареенето като за нещо, настъпващо с напредването на възрастта.

2.1.2 Физиологични ефекти

Стареенето оказва изключително голям ефект върху човешкото тяло. То обикновено включва широк спектър от различни физиологични промени, които влошават жизнеността и качеството на живот на индивида. Примери за това са понижена фертилност при жените [7]; загуба на телесна маса [22]; влошен слух[6]; повишен риск от хронични заболявания [11][20]; хронична болка [17]; загуба на сила и еластичност в мускулно-скелетната система; понижената способност за устояване на инфекции, екстремни температури и др. видове стрес; влошаване на зрението; загуба на неврологични функции [27] и др.

2.1.3 Демографски и икономически ефекти

През последния един век очакваната продължителност на живота в целия свят драстично се е повишила [31] (виж фиг. 2.1). Освен безспорните ползи, това води и до редица проблеми. Удължаването на продължителността на живота, в комбинация с наблюдавания спад на раждаемостта, се очаква да доведе до застаряване на населението [12]. Световната Здравна Организация (СЗО) предупреждава, че се очаква между 2015 и 2050 броят на хората над 60-годишна възраст да се повиши от 12% от населението до 22% [18]. Същевременно, по данни на СЗО, увеличаването на продължителността на живота (с 6 години за периода между 2000 и 2019) изпреварва увеличаването в продължителността на здравословния живот (с 5.4 години за същия период) [19].



Фигура 2.1: Очаквана продължителност на живота за различни региони през периода 1770-2019 [31].

Застаряването на населението би оказало неблагоприятен ефект и върху икономиката на държавите. Първо, заради увеличаването на дяла на хора, които не участват в работната сила. Второ, поради това, че здравните системи ще бъдат допълнително натоварени с по-голям брой хора в напреднала възраст, за които рисковете от хронични заболявания са значително по-големи.

2.2 Молекулярно-биологични теории за стареенето

2.2.1 Общи молекулярно-биологични процеси

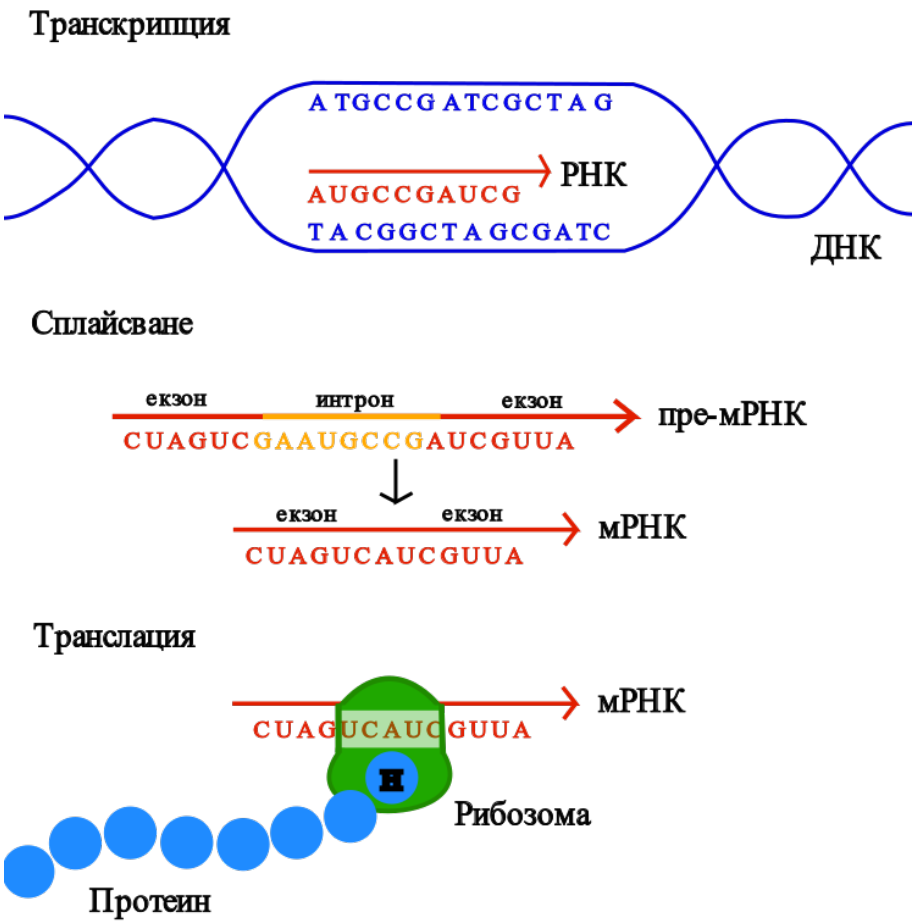
В тази секция ще разгледаме фунадменталните принципи на генетиката. Ще направим кратък обзор на начина на съхранение на генетичната информация и процесите, чрез които тя бива изразена, за да повлияе на фенотипа. С това целим да дадем базов биологически контекст, чрез който да бъдат разбрани по-нататъшните разработки и биоинформатични анализи.

Дезоксирибонуклеиновата киселина (ДНК) представлява две вериги от спираловидно преплетени полимери, които съдържат генетичната информация при всички клетъчни форми на живот. Полимерите са създадени от последователности от мономери - нуклеотидни бази. В ДНК се изпозлват четири вида бази - аденин (A), цитозин (C), гуанин (G) и тимин (T). Базите A и T образуват двойки помежду си, както и базите C и G. Казваме, че двете нишки на ДНК са комплиментарни. Всеки ген може да бъде разположен на коя да е от двете нишки на ДНК и е описан от дълга последователност от нуклеотидни бази [9, стр. 301-310].

Най-често крайната цел на един ген е кодирането на протеин. Първата стъпка към това е транскрипцията, при която ензимът ДНК-полимераза копира информацията от ДНК в комплиментарна РНК молекула [21]. При РНК, базата T е заменена с урацил (U). Първичната РНК молекула (pre-mRNA) преминава прецес на сплайсване, при който части от нея (интрони) биват изрязвани и остранени. Останалите части (екзони) се свързват отново. Така се образува зрялата mRNA, която бива транслирана в рибозомите, като на всеки кодон (група от три нуклеотидни бази) се съпоставя определена амино-киселина (виж фиг. 2.2). Верига от амино-киселини образува протеин [9, стр. 412-420].

2.2.2 Теории за стареенето

Стареенето е въпрос, който вълнува учените от дълго време. През 1990-та, Медведев твърди, че вече съществуват над 300 теории за стареенето [15]. Въпреки постигнатият значителен прогрес през последните години в



Фигура 2.2: Графична репрезентация на процесите на транскрипция, сплайсване и транслация

областта на геронтологията, причините за стареенето все още остават ненапълно обяснени. Това се дължи на факта, че стареенето е сложен процес, в който са намесени множество фактори. Все още липсва голяма обединяваща теория на стареенето, която да обясни изцяло процеса, но съществуват множество теории, които дават добра представа за различни негови аспекти [27]. Следва кратък преглед на основните теории:

Натрупване на геномни изменения

Изменения в ДНК молекулите могат да настъпят както в следствие на вътрешноклетъчни фактори, така и поради въздействието на външни мутагени. Примери за вътрешноклетъчни фактори са случайни грешки при репликация и оксидативния стрес, предизвикан от натрупването на свободни радикали [28]. Външните мутагени могат да бъдат разделени на три

вида - физични, химични и биологични. Пример за физичен мутаген е радиацията [2], а за биологичен вирусните инфекции, които също могат да предизвикат генетични мутации. Измененията в ДНК молекулите включват различни видове мутации като точкови мутации, делеции и инсерции, транслокации, инверсии и др.

Съществуват механизми, чрез които клетките засичат мутациите и ги поправят. Основни такива механизми са гените ATM и TP53. Все пак, тези механизми не са ефективни на 100% и ефективността им допълнително спада с възрастта [13]. В резултат, в течение на времето, ДНК молекулите акумулират все повече мутации. Смята се, че тази геномна нестабилност е един от основните фактори, допринасящи за процеса на стареенето [26].

Скъсяване на теломерите

Теломерите са регион, намиращ се в края на хромозомите, в който се съдържат повтарящи се поредици от нуклеотидни бази. Те служат за предпазване на хромозомата от рекомбинация и постепенна деградация и дават възможност на клетката да различава края на хромозомата от случайни прекъсвания, при които биха били активирани механизмите за поправка на ДНК [8]. При всеки цикъл на делене на клетката, теломерите се скъсяват поради непълното синтезиране на изоставащата нишка от ДНК полимеразата [10]. Този проблем се компенсира донякъде от ензима теломераза, който пренася своя собствена РНК молекула и я използва като шаблон, спрямо който да удължи скъсения теломер. Въпреки това, недостатъчната експресия на теломеразата води до постепенното скъсяване на теломерите. Това може да доведе до загуба на репликативна способност на клетката и блокирането на клетъчния ѝ цикъл, процес известен като клетъчно стареене [16]. Установено е, че първоначалната дължина на теломерите няма връзка със стареенето при различни видове, но скоростта на тяхното скъсяване има значителна корелация със продължителността на живота им [29].

Клетъчно стареене

TODO

Епигенетични изменения

TODO

2.3 Генетични фактори, влияещи на процеса на стареене

В секция 2.2.2 беше представен кратък обзор на различните биологични процеси, които способстват процеса на стареене. Уместен е въпросът дали има определени генетични фактори, които оказват въздействие на тези процеси. Ако това е така, бихме могли да очакваме, че съществуват генни алели, които забързват или забавят стареенето. В текущата глава ще разгледаме въпроса за съществуването на такива генни алели, както и за начините им на действие и методите за изследването им.

2.3.1 Видове генетични мутации, влияещи на стареенето

Два от биологичните процеси, разгледани в секция 2.2, за които се смята, че причиняват стареенето, са натрупването на геномни мутации и клетъчното стареене. Един протеин, който играе важна роля и в двата процеса е p53. Той се кодира от хомолози на един и същи ген в различни организми. При хората това е генът TP53. Протеинът p53 има роля за предотвратяването на натрупване на геномни мутации и спирането на туморогенезиса. Той бива активиран в отговор на увреждания на ДНК, експресия на онкогени и дисфункция на рибозомите. Функциите на p53 включват активиране на гени, свързани с поправката на ДНК, спиране на клетъчния цикъл, за да се предотврати размножаване на клетката, докато има уреждания в ДНК, активиране на клетъчното остаряване и иницииране на апоптоза (клетъчна смърт) [24]. В изследвания на хора е установено, че полиморфизми в TP53 могат да доведат до удължаване на живота, но да увеличат и смъртността от рак [25]. Това демонстрира и крехкия баланс между ползи и вреди, които дадени алели могат да носят.

Протеинът Telomeric repeat-binding factor 1, кодиран от генът TERF1 при хората, е основен компонент от shelterin комплекса, който има важна роля в защитата и репликацията на теломерите. Изследвания показват, че увеличаването на експресията на TRF1 в зрели мишки (на 1 година) и възрастни мишки (на 2 години), посредством генна терапия, може да забави настъпването на патологии, свързани със стареенето [5].

TODO: пример свързан с епигенетични процеси

Тези примери не са изолирани изключения. В научната литература могат да бъдат намерени много гени, за които изследвания са открили асоциация със стареенето. Публичната база данни Human Ageing Genomic Resources (HAGR), представлява колекция от ресурси за изследването на стареенето при хората. Някои записи в HAGR са включени на база установена директна връзка между даден ген и стареенето, докато други са включени на база ролята им в различни човешки патологии. Много от записите са подбрани, тъй като за техни хомолози в други организми е била открита връзка със стареенето. HAGR предоставя и набор от софтуерни инструменти (предимно Perl и SPSS скриптове) за различни видове биоинформатичен анализ. Към момента в HAGR са налични над 300 човешки гена, за които се предполага, че имат потенциална връзка със стареенето [23].

2.4 Обзор на съществуващи биоинформатични решения

2.4.1 VCF файлове

Variant Call Format (VCF) е стандартен файлов формат, който се използва за описване на генетичните полиморфизми за дадена секвенция (за примерен файл виж фиг. 2.3). VCF е текстов файлов формат с разделители-табулации (tab-delimited), който често бива съхраняван в компресиран вид, с цел оптимизиране на хардурерните ресурси, като дори компресиран може да бъде индексиран за бързо търсене. Във VCF могат да бъдат описани различни видове полиморфизми, от прости като точкови мутации, инсерции и делеции до по-сложни като например инверсии. VCF може да съдържа коментари, заглавен ред и редове за данни. В редовете за данни, всеки ред показва един полиморфизъм, като обикновено се използват стандартни референтни геноми, спрямо които се определят полиморфизмите. Файловият формат позволява и добавянето на богата анотация и потребителски-дефинирани полета. VCF стандартът е разработен за 1000 Genomes Project, а впоследствие е добил широка приемственост в биоинформатичната общност [4].

2.4.2 Анотация на генетични варианти

С напредъка на технологиите за секвениране способността за бързо генериране на големи обеми от данни за генетични варианти бързо расте. Същевременно се образува все по-голяма пропаст между възможностите за

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Фигура 2.3: Примерен VCF файл

генериране на нови сурови данни и възможностите за извличане на полезна информация и познание от тях [30]. Основна стъпка за разбирането на суровите данни с генетични варианти е анотирането им. Анотацията представлява процес, при който към генетичните варианти се добавя допълнителна функционална информация [14]. Това може да бъде информация към кои кодиращи секвенции и гени се отнася варианта, оценка на степента му на въздействие, индикация дали се променят аминокиселините на кодирания протеин [3], предсказване на структурните и функционални промени в протеина [14] и др.

snpEff

VEP

Annovar

2.4.3 Филтриране и анализ на генетични варианти

snpSift

???

2.4.4 Геномни браузъри

UCSC Genome Browser

IGV Genome Browser

2.4.5 Нагъване на протеини

Подходи

AlphaFold

2.4.6 Интегрирани софтуерни решения

Galaxy Project

3. Цели и задачи

4. Използвани софтуерни решения

5. Резултати

5.1 Софтуерна архитектура

5.2 Структура на базата данни

5.3 Управление на генни множества

5.4 Обработка на VCF при импортиране

5.5 Операции върху импортиран VCF

6. Дискусия

7. Изводи

Библиография

[1] R. Arking. *Biology of Aging: Observations and Principles.* Oxford University Press, 2006.

[2] L. H. Breimer. Ionizing radiation-induced mutagenesis. *Br J Cancer*, 57(1):6–18, Jan 1988.

[3] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[4] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, and et al Handsaker. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug 2011.

[5] A. Derevyanko, K. Whittemore, R. P. Schneider, V. Jiménez, F. Bosch, and M. A. Blasco. Gene therapy with the TRF1 telomere gene rescues decreased TRF1 levels with aging and prolongs mouse health span. *Aging Cell*, 16(6):1353–1368, 12 2017.

[6] K. Feder, D. Michaud, P. Ramage-Morin, J. McNamee, and Y. Beauregard. Prevalence of hearing loss among Canadians aged 20 to 79: Audiometric results from the 2012/2013 Canadian Health Measures Survey. *Health Rep*, 26(7):18–25, Jul 2015.

[7] K. George and M. S. Kamath. Fertility and age. *J Hum Reprod Sci*, 3(3):121–123, Sep 2010.

[8] Jack D Griffith, Laurey Comeau, Soraya Rosenfield, Rachel M Stansel, Alessandro Bianchi, Heidi Moss, and Titia de Lange. Mammalian telomeres end in a large duplex loop. *Cell*, 97(4):503–514, 1999.

[9] William S. Klug, Michael R. Cummings, Spencer Charlotte A., and Michael A. Palladino. *Concepts of Genetics: Pearson New International Edition*. 2014.

[10] Alexander K. Koliada, Dmitry S. Krasnenkov, and Alexander M. Vaiserman. Telomeric aging: mitotic clock or stress indicator? *Frontiers in Genetics*, 6, 2015.

[11] E. B. Larson, K. Yaffe, and K. M. Langa. New insights into the dementia epidemic. *N Engl J Med*, 369(24):2275–2277, Dec 2013.

[12] W. Lutz, W. Sanderson, and S. Scherbov. The coming acceleration of global population ageing. *Nature*, 451(7179):716–719, Feb 2008.

[13] Mark T. Mc Auley, Alvaro Martinez Guimera, David Hodgson, Neil Mcdonald, Kathleen M. Mooney, Amy E. Morgan, and Carole J. Proctor. Modelling the molecular mechanisms of aging. *Bioscience Reports*, 37(1), 02 2017. BSR20160177.

[14] D. J. McCarthy, P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J. B. Cazier, and P. Donnelly. Choice of transcripts and software has a large effect on variant annotation. *Genome Med*, 6(3):26, 2014.

[15] Zhores A Medvedev. An attempt at a rational classification of theories of ageing. *Biological Reviews*, 65(3):375–398, 1990.

[16] Keiko Muraki, Kristine Nyhan, Limei Han, and John P Murnane. Mechanisms of telomere loss and their consequences for chromosome instability. *Frontiers in oncology*, 2:135, 2012.

[17] AGS Panel on Persistent Pain in Older Persons. The management of persistent pain in older persons. *Journal of the American Geriatrics Society*, 50(6 Suppl):S205–224, Jun 2002.

[18] World Health Organization. *World report on ageing and health*. World Health Organization, 2015.

[19] World Health Organization. Life expectancy and healthy life expecancy - data by country. *Published online*, 2020.

[20] Sahdeo Prasad, Bokyung Sung, and Bharat B. Aggarwal. Age-associated chronic diseases require age-old medicine: Role of chronic inflammation. *Preventive Medicine*, 54:S29–S37, 2012. Dietary Nutraceuticals and Age Management Medicine.

[21] R. J. Sims, S. S. Mandal, and D. Reinberg. Recent highlights of RNA-polymerase-II-mediated transcription. *Curr Opin Cell Biol*, 16(3):263–271, Jun 2004.

[22] R. P. Spencer. Organ/body weight loss with aging: evidence for co-ordinated involution. *Med Hypotheses*, 46(2):59–62, Feb 1996.

[23] R. Tacutu, D. Thornton, E. Johnson, A. Budovsky, D. Barardo, T. Craig, E. Diana, G. Lehmann, D. Toren, J. Wang, V. E. Fraifeld, and J. P. de Magalhães. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res*, 46(D1):D1083–D1090, 01 2018.

[24] E. Toufektchan and F. Toledo. The Guardian of the Genome Revisited: p53 Downregulates Genes Required for Telomere Maintenance, DNA Repair, and Centromere Structure. *Cancers (Basel)*, 10(5), May 2018.

[25] D. van Heemst, S. P. Mooijaart, M. Beekman, J. Schreuder, A. J. de Craen, B. W. Brandt, P. E. Slagboom, and R. G. Westendorp. Variation in the human TP53 gene affects old age survival and cancer mortality. *Exp Gerontol*, 40(1-2):11–15, 2005.

[26] J. Vijg and Y. Suh. Genome instability and aging. *Annu Rev Physiol*, 75:645–668, 2013.

[27] Jose Viña, Consuelo Borrás, and Jaime Miquel. Theories of ageing. *IUBMB life*, 59(4-5):249–254, 2007.

[28] David Wang, Deborah A. Kreutzer, and John M. Essigmann. Mutagenicity and repair of oxidative dna damage: insights from studies using defined lesions. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 400(1):99–115, 1998.

[29] Kurt Whittemore, Elsa Vera, Eva Martínez-Nevado, Carola Sanpera, and Maria A. Blasco. Telomere shortening rate predicts species life span. *Proceedings of the National Academy of Sciences*, 116(30):15122–15127, 2019.

[30] H. Yang and K. Wang. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc*, 10(10):1556–1566, Oct 2015.

[31] Richard Zijdeman and Filipa Ribeira da Silva. Life Expectancy at Birth (Total). 2015.