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

HGVS - Human Genome Variation Society. Организация, занимаваща се с генетични варианти при хората, дала името и на стандартна номенклатура за описване на варианти на ДНК, РНК, протеини и други свързани с генетиката макромолекули.
Indel - Isertion/Deletion. Генни варианти, при които определена нуклеотидна последователност е изтрита или вмъкната.
MNP - Multiple Nucleotide Polymorphism. Множествен нуклеотиден полиморфизъм се нарича когато варианта и референтната поредица имат еднаква дължина, различна от 1.
OLAP - Online Analytical Processing. Анализът в реално време е подход за бързо обработване на многомерни аналитични заявки.
SNP - Single Nucleotide Polymorphism. Единичен нуклеотиден полиморфизъм е тип мутация, наричана още точкова мутация, при която една единствена нуклеотидна база е променена.
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.“

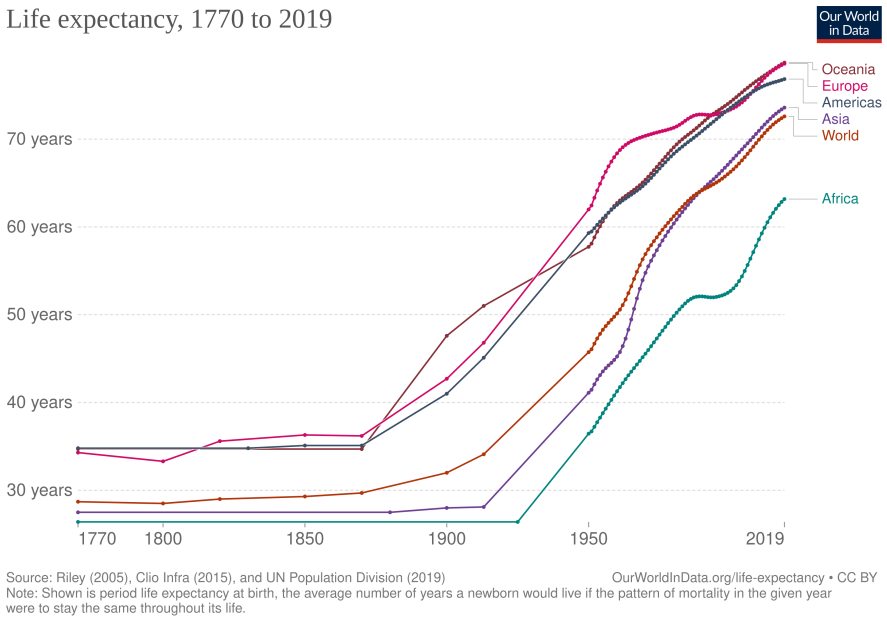
Макар времето да няма каузална връзка с ефектите на стареенето, то корелацията помежду им е причина обикновено да се говори за ефектите на стареенето като за нещо, настъпващо с напредването на възрастта.

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 [12]; загуба на телесна маса [40]; влошен слух[11]; повишен риск от хронични заболявания [20][32]; хронична болка [27]; 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 [45] и др.

2.1.3 Демографски и икономически ефекти

През последния един век очакваната продължителност на живота в целия свят драстично се е повишила [51] (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 [22]. 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% [28]. 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) [29].



Фигура 2.1: Очаквана продължителност на живота за различни региони през периода 1770-2019 [51].

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

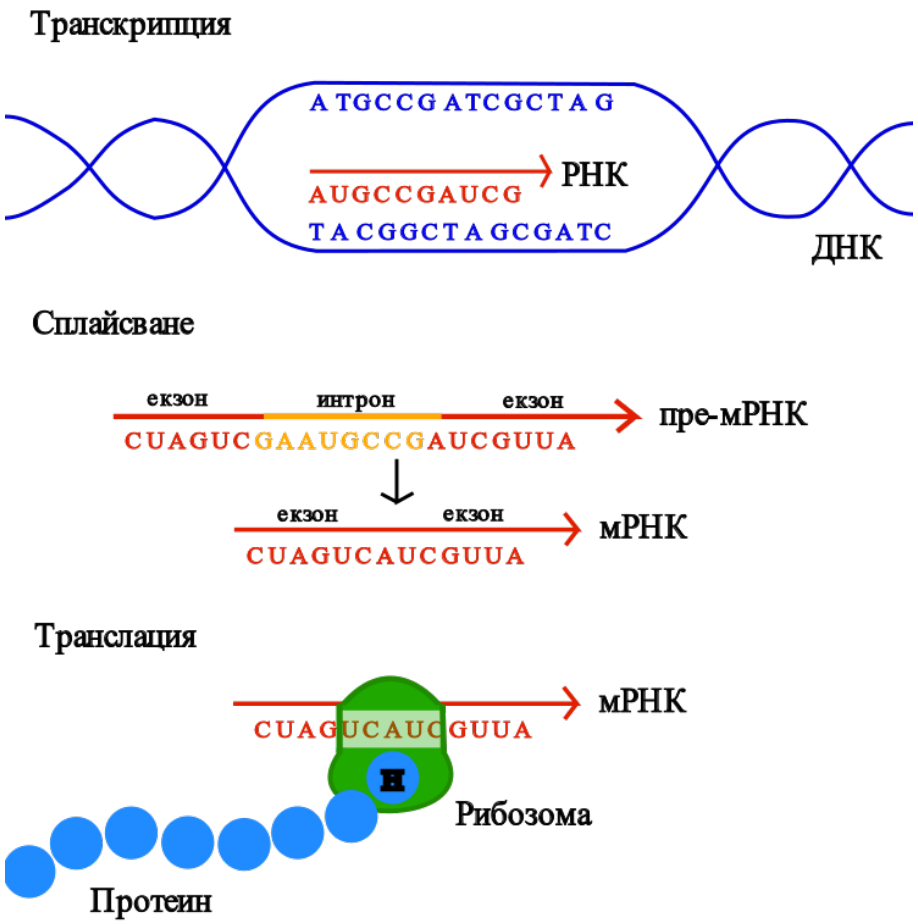
В тази секция ще разгледаме фунадменталните принципи на генетиката. Ще направим кратък обзор на начина на съхранение на генетичната информация и процесите, чрез които тя бива изразена, за да повлияе на фенотипа. С това целим да дадем базов биологически контекст, чрез който да бъдат разбрани по-нататъшните 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 използват 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 [17, 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 [39]. 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 [17, 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 [25]. 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 [45]. 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 [46]. 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 [23]. 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 [44].

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 [13]. При всеки цикъл на делене на клетката, теломерите се скъсяват поради непълното синтезиране на изоставащата нишка от ДНК полимеразата [18]. 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 [26]. 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 [48].

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) [42]. В изследвания на хора е установено, че полиморфизми в TP53 могат да доведат до удължаване на живота, но да увеличат и смъртността от рак [43]. 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 [10].

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 [41].

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 [7].

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

```
##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

за генериране на нови сурови данни и възможностите за извличане на полезна информация и познание от тях [49]. 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 [24]. 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 [5], предсказване на структурните и функционални промени в протеина [24] и др.

snpEff

SnpEff е софтуер с отворен код, който може бързо да анотира и категоризира генни варианти на база на ефекта, който те биха имали върху анотираните гени. SnpEff поддържа анотацияа на различни видове полиморфизми, като например единични нуклеотидни полиморфизми (SNPs), множествени нуклеотидни полиморфизми (MNPs) и вмъквания-изтривания (Indels) [5]. SnpEff разполага с много богата база данни от различни референтни геноми, с които може да работи, а дава възможност на потребителя да използва и свой собствен референтен геном. Основният формат, с който SnpEff работи е VCF. След обработката на входния VCF файл, съдържащ по един полиморфизъм на ред, SnpEff добавя една или повече анотации за всеки полиморфизъм в полето INFO, като всяка анотация има ключ „ANN“. Някои от по-важните анотации, които SnpEff предоставя са: идентификация на генът, с който е свързан полиморфизма; идентификатори на транскриптите, които полиморфмизмът засяга; оценка на ефекта на полиморфизма и

на промените, които би причинил в аминокиселинния състав на протеина, който кодира. SnpEff е имплементиран на програмния език Java, което го прави лесно преносим и му дава възможност да работи на изключително голям набор от операционни системи и устройства [37, стр. 9-10].

VEP

Annovar

2.4.3 Филтриране и анализ на генетични варианти

snpSift

SnpSift е софтуер за филтриране и промяна на VCF файлове, съдържащи анотирани генетични варианти. Чрез SnpSift могат да се извършат различни операции, като например филтриране по геномен регион, разделяне на файла по хромозома, извличане на определени полета, допълнително анотиране спрямо външни бази данни, както и филтриране с потребителски дефиниран логически израз. Филтрирането с потребителски израз работи посредством рекурсивна граматика, която може да обработва изрази с произволна сложност [4]. Това прави snpSift доста мощен инструмент за лесна обработка и анализ на генетични варианти и извличане на информация от тях. SnpSift, също като SnpEff, е имплементиран на програмния език Java, което му дава голяма преносимост върху различни операционни системи и платформи [37, стр. 9-10].

??? TODO

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

3.1 Цели на дипломната работа

Дипломната работа има за цел създаването интегрирана софтуерна система за биоинформатичен анализ на геномни варианти, която има следните характеристики:

- Да приема входни данни за генетични варианти посредством стандартен VCF файлов формат.
- Да може да анализира генетични варианти и да предоставя подробен доклад, съдържащ:
  - Идентификация на гените, засегнати от полиморфизмите.
  - Асоцииране на тези гени с процеса на стареенето.
  - Оценка на тежестта на откритите варианти.
  - Предсказване на откритите варианти върху процеса на транслация и протеиновата структура.
  - Предсказване и визуализация на на триизмерните протеинови (третични) структури.
- Да разполага с уеб-базиран потребителски интерфейс, за улеснено ползване от потребители, които не са компютърни специалисти.
- Да разполага с потребителски интерфейс, работещ в командния ред на операционната система, позволяващ интеграцията на софтуера в други биоинформатични системи.

3.2 Задачи

За постигане на целите, описани в предишната секция се предвижда следния списък от задачи:

1. Интегриране на софтуер за анотация на генетични варианти към програмното решение.

2. Филтрация на анотираният VCF с генетични варианти, така че да съдържа единствено полиморфизми, засягащи гени, които потребителят е решил да изследва.

3. Анализ на наличните данни и моделиране на релационна база данни, в която данни да бъдат съхранявани с цел последващо изпълняване на разнообразни аналитични заявки.

4. Разработване на уеб-базиран потребителски интерфейс, който да включва:

   (а) Възможност за създаване и управление на генетични множества, спрямо които да бъдат изследвани входните VCF файлове с генетични варианти.

   (б) Възможност за качване на входен VCF файл, съдържащ генетични варианти.

   (в) Набор от страници за изследване на резултатите за обработен VCF файл.

5. Намиране на модифицираната полипептидната поредица на модифицираните от генетичен вариант протеини.

6. Интегриране на софтуер за предсказване на третичната (триизмерна) структура на протеини към програмното решение.

7. Интегриране на решение за визуализация на биолгоични макромолекули, с цел представяне на триизмерната третична структура на референтния и модифицирания протеин с цел сравнението им.

4. Използвани софтуерни решения

4.1 Flask

Flask [36] е минималистичен framework (преизползваема платформа, която подпомага разработването на софтуер) за създаване на уеб приложения с програмния език Python. Класифицира се като минималистичен, тъй като не налага използването на определени библиотеки и инструменти, а оставя избора да бъде направен от програмиста. За разлика от много други уеб framework решения, Flask не включва определени компоненти и библиотеки за стандартни нужди като връзка с бази данни, валидация на потребителски входни данни, автентикация и др. Вместо това, предоставя възможности за разширяване, към които могат да се вградят произволни външни библиотеки и компоненти.

Flask (v2.1.2) е основен компонент в програмното решение, като го използваме за изграждането на уеб-базирания потребителски интерфейс. Flask се грижи за обработката на HTTP заявките, насочването им към правилния контролер-метод, изграждането на статичните страници (чрез шаблонни страници, рендерирани от библиотеката Jinja), поддържането на потребителската сесия и др.

4.2 DuckDB

DuckDB [33] е система за управление на бази данни, която позволява изпълняването на SQL заявки, докато е вградена в друг процес. Това означава, че базата данни не се нуждае от отделен сървър, който да управлява базата данни и да изпълнява заявките. Вместо това, системата може да бъде вградена в програма под формата на библиотека с функции, позволява-

щи работа с базата данни, която представлява отделен файл на файловата система. В този аспект, DuckDB прилича на популярната база данни SQLite [15]. DuckDB е предвидена за изпълняването на аналитични заявки, известни още като OLAP (Online Analytical Processing). При този тип заявки, най-често се използва сравнително малко подмножество от наличните колони, но за сметка на това се прави обработка на всички налични редове. За да се осгури добра производителност за този тип употреба, DuckDB използва техниката за векторизация на заявките, при която множество стойности биват изчитани и обработвани накуп, вместо една по една, като по този начин се амортизира сложността на итерацията [16].

За нашето решение, избрахме да използваме DuckDB (v0.4.0), защото основните заявки, които се изпълняват са аналитични по своята природа и обработват голямо количество от редове. Заявките за писане, и особено транзакционните такива, са редки. В допълнение, липсата на отделен сървър за базата данни улеснява инсталацията на софтуера за потенциалните му потребители.

4.3 Pandas

Pandas [30] е популярна библиотека за програмния език Python за анализ и манипулация на данни, създадена през 2008 година. Pandas предоставя DataFrame стуктура от данни, представляваща колекция от данни, представени в табличен вид. Pandas включва богат набор от функции за операции като анализ, трансформация, групиране, агрегация, преоразмероване, сливане и разделяне, четене/писане от и към различни видове файлови формати и др. Основните критични пътища са имплементирани на C и CPython, значително оптимизирайки производителността.

Повечето от операциите в програмното решение засягат обработката на голям обем от данни от структурирани данни, което прави използването на Pandas удачно. Също така, DuckDB предоставя интерфейс за извличане на резултати от заявки и за подаване на данни, при който се използват Pandas DataFrame обекти. Разработчиците на DuckDB препоръчват използването на тези интерфейси, когато обемът на данните е по-голям, поради по-голямата им ефективност. Използвана е версия 1.4.2.

4.4 PyVCF3

PyVCF3 [38] е наследник на библиотеката PyVCF, адаптирано за съвместимост с версия 3 на програмния език Python. PyVCF е библиотека за синтактичен анализ на VCF файлове, която позволява прочичтането на VCF файлове и зареждането им в Python класове и структури от данни. PyVCF се опитва да обработи „##INFO“ и „##FORMAT“ редовете, за да отгатне структурата на данните във входния файл. Ако това не е успешно, се обръща към дефинициите в VCF стандарта.

Програмното решение използва PyVCF3 (v1.0.3) за да прочете входните или междинните VCF файлове. Веднага след прочитането, данните се зареждат в Pandas DataFrame, чрез който биват обработвани.

4.5 Samtools и Tabix

Форматът SAM (Sequence Alignment/Map) е стандартен формат за съхраняване на голям брой подравнявания на нуклеотидни последователности. BAM е бинарен файл, представляващ компресирана версия на съответния му SAM файл. Samtools [21] софтуер за обработка на SAM и BAM файлове, включващ редица операции като сортиране, инспектиране, обединяване на файлове и др.

Tabix е програма за индексиране на делимитирани с табулация файлове, съдържащи геномни позиции. След индексирането на файл, Tabix позволява бързото намиране на редове, съдържащи информация за конкретни геномни региони. Важно необходимо условие за използването на Tabix за индексиране на файл е файлът да бъде сортиран предварително спрямо геномната позиция.

4.6 Pysam

Pysam [14] е библиотека за Python, която позволява четенето и манипулирането на файлове в SAM и BAM формат. Pysam представлява тънка обвивка около samtools и tabix и позволява лесното взаимодействие с тези инструменти от Python програми.

В програмното решение, Pysam (v0.19.1) се използва за компресиране и индексиране на междинните VCF файлове, които програмата извлича след анотирането и филтрирането на входния VCF.

4.7 HGVS

Human Genome Variation Society (HGVS) е организация, занимаваща се с генетични варианти при хората, дала името и на стандартна номенклатура за описване на варианти на ДНК, РНК, протеини и други свързани с генетиката макромолекули. HGVS стандартът е широкоприет [9]. Общият формат за описване на вариант е „референция:описание“. Например, „NM_004006.2:c.4375C>T“ описва вариант на транскрипта NM_004006.2, при който в кодиращото ДНК цитозина от позиция 4375 е подменен от тимин.

Hgvs [47] е също така и името на библиотека за Python за манипулирането на варианти на поредици, използвайки HGVS номенклатурата. Hgvs позволява синтактичен анализ, форматиране, валидация и нормализиране на варианти на поредици от ниво геном, транскриптом или протеом.

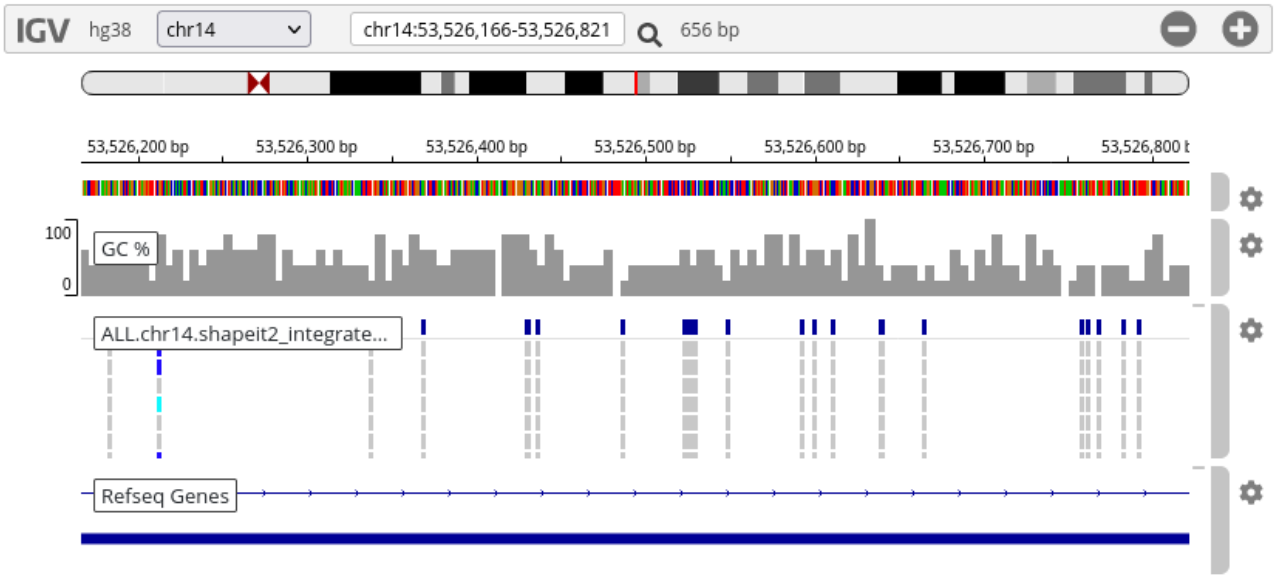
В програмното решение се използва версия 1.5.2 на пакета за да се направи синтактичен анализ на HGVS номенклатурата, получена при анотацията посредством SnpEff, и да се изведе модифицирания протеин.

4.8 Bulma

Bulma е CSS framework, който предоставя готови компоненти и функционалности за стилизиране на уеб страници и разработване на потребителски интерфейси. Bulma се съдържа изцяло в един единствен CSS файл, който предоставя богата колекция от класове за стилизране на HTML. Bulma предоставя и множество стандартни компоненти за изграждане на потребителски интерфейси като: менюта, контейнери за съобщения, навиганционни ленти, модални прозорци, страниране, етикети и др. В програмното решение е използвана версия 0.9.4 на Bulma.

4.9 IGV

IGV (Integrative Genomics Viewer) е инструмент за визуализация и интерактивно изследване на разнообразни, големи по обем геномни колекции от данни . Посредством IGV могат гъвкаво да се интегрират широк спектър от различни типове геномни данни, като например подравнени поредици, мутации, експресия на гени, метилация, геномни анотации и др. IGV позволява на потребителя да изследва данните си на различни нива на резолюция, като дава възможност за приближаване и отдалечаване в реално време [34][35]. Данните могат да бъдат зареждани както от публични онлайн бази данни, така и локално от данните, с които разполага потребителя.



Фигура 4.1: Примерна визуализация с IGV

В програмното решение, IGV е интегриран в страницата за изследване на вариантите за определен ген.

4.10 Външни генетични библиотеки

4.10.1 HGNC

HGNC (HUGO Gene Name Nomenclature Committee) е организация отговорна за одобрението на уникални символи и имена за човешки гени, протеини, некодиращи РНК гени и псевдогени, с цел улесняване на комуникацията между учените [31]. HGNC също предоставят публична база

данни с пълната, одобрена от организацията, номенклатура. Тя е достъпна на адрес „genenames.org“ и също така предоставя REST API, чрез който информацията може да бъде достъпвана програматично. Освен номенклатурата, базата данни предоставя и идентификатори за артефакта на различни често използвани външни бази данни, като например Ensembl и UniProt.

Програмното решение използва REST API услугата на HGNC за да извлича името и обща информация за гени, включително идентификатори от други бази данни. Тези идентификатори се използват за генериране на линкове към, с които потребителя да може да изследва въпросният ген в по-голяма дълбочина.

4.10.2 Ensembl

Ensembl е система за генериране и дистрибуция на геномна анотация, като например гени, генни варианти, генетична регулация и сравнителна геномика за множество видове сред гръбначните организми и други моделни организми. Ensembl интегрира експериментални и референтни данни от множество източници на едно място. Към 2020 година, Ensembl съдържа анотираните геноми на 227 различни вида [50]. Информацията, предоставяна от Ensembl, също така може да бъде извличана програматично чрез REST API.

Програмното решение използва REST API услугите, предоставяни от Ensembl, за да извлича референтната аминокиселинна поредица на протеините, които потребителя изследва.

4.10.3 NextProt

NextProt е платформа, която се стреми да дава цялостна и богата информация за човешките протеини. Разработва се от Швейцарския Институт по Биоинформатика. Използва както информация от експертно-модерираната база данни UniProtKB/Swiss-Prot, така и внимателно селектирана информация от експерименти с високопроизводително секвениране [19]. NextProt съдържа богата информация за функцията на протеините, биологичните пътища, в които те участват, както и молекулярните процеси, с които са свързани.

Програмното решение използва REST API услугата на NextProt за да извлича функциите, молекулярните процеси и биологичните пътища, с които се свързва даден протеин. Информацията се предоставя на потребителя, за да даде контекст за значението на протеина, който бива изследван.

5. Резултати

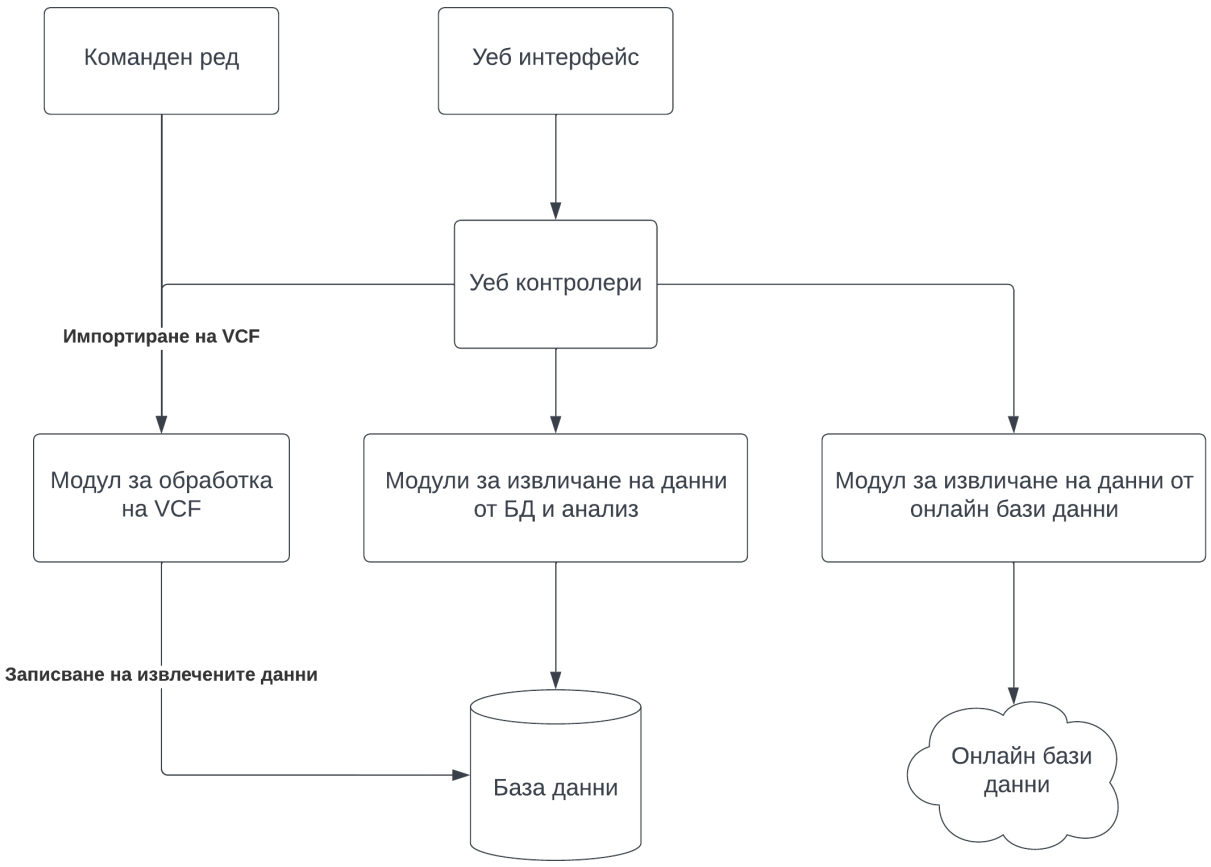
Въпреки огромното значение на процеса на стареене за индивида и обществото като цяло, науката все още не може да го обясни напълно, нито да намери достатъчно ефективни решения за справяне с негативните ефекти, които той предизвиква. Откритите асоциации между определени генетични варианти и стареенето означава, че бъдещите изследвания трябва да включват и изследване на генетични полиморфизми. Този тип изследвания генерират големи обеми от данни, но съществуващият биоинформатичен софтуер за целта изисква високо ниво на подготовка в областта на информационните технологии. Биолозите и генетиците, които искат да разберат данните, получени от геномните секвенирания, често трябва да изграждат сложни системи от различни програми. Свързването им често включва писането на програмни скриптове.

С цел да облекчим работата при бъдещи изследвания на генетични полиморфизми, свързани със стареенето, и да направим такива изследвания по-достъпни, разработихме интегрирана биоинформатична система, която позволява анализирането и интерпретирането на данни за генетични варианти посредством удобен, уеб-базиран, графичен интерфейс. Софтуерът ни е имплементиран на програмния език Python и е достъпен за свободно ползване без ограничения, като кодът му е отворен и достъпен на адрес `github.com/mzdravkov/gene_variants`. Софтуерът има за цел да бъде лесен за инсталиране и използване.

В процеса на разработка установихме, че същите принципи, които прилагаме за изследване на полиморфизми, свързани със стареенето, могат да бъдат приложени и върху други видове изследвания, като това изискваше съвсем малка генерализация на предвиденото софтуерно решение. В резултат, разработихме възможност за управление на различни множества от гени, като при подаване на VCF файл с генетични варианти, потребителят може да избере кое генно множество да бъде анализирано при обработката на полиморфизмите.

5.1 Софтуерна архитектура

Софтуерната архитектура на системата може да се раздели на четири слоя. Интерфейсен, включващ командния ред и уеб интерфейса с неговите страници, стилове и JavaScript код, който се изпълнява от браузъра на потребителя. Вторият слой включва контролерите, които обработват HTTP заявките, изпращани от уеб интерфейса. Третият слой се занимава с обработката или извличането на данни. В него се включват модулът за обработка на входни VCF файлове, модулите за извличане и анализ на данни от базата данни и модулът за извличане на данни от външни, онлайн бази данни (фиг. 5.1).



Фигура 5.1: Диаграма на софтуерната архитектура на разработеното решение

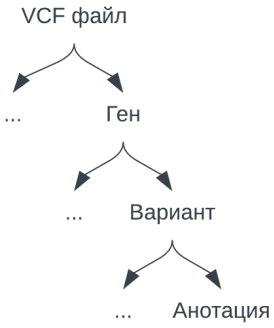
5.2 База данни

При изборът на тип на базата данни се спряхме на релационна база данни. Причините за това са няколко. Първата съществена причина е това, че структурата на данните е предварително ясна и промени по нея са малко вероятни. Това е така, тъй като структурата е базирана на информацията налична във VCF файловете и тази, предоставяна от софтуера за анотация (SnpEff). VCF форматът е дефиниран в общоприет стандарт [8], който не се променя често. Анотациите, предоставяни от SnpEff, също спазват определен стандарт [6]. Втората причина е, че релационните бази данни позволяват моделиране на базата спрямо концептуалния модел на данните, докато заявките, които ще се изпълняват, могат да имат по-малко значение [3]. Това свойство на релационните бази данни е подходящо за нашия случай, при който от самото начало е ясно с какви данни разполагаме, но не и как точно можем да ги анализираме. Релационната база данни ни позволява лесно да добавим нови заявки, без да са необходими промени по структурата на базата данни. Третата причина е, че не предвиждаме работа с толкова големи обеми от данни, че да изискват дистрибутиране на базата данни върху различни сървъри.

Измежду множеството релационни бази данни се избрахме DuckDB. Причините за това са свойството ѝ да се вгражда в друга програма, без да е необходим отделен сървър за управление на базата данни, което улеснява инсталацията, и поради приложимостта ѝ за изпълняване на аналитични заявки (виж секция 4.2).

Таблиците, използвани в програмното решение, могат да се разделят на две групи: такива, които са свързани с дефинирането на генни множества, и такива, които съхраняват информацията от анотирани VCF файлове с генетични варианти. За дефинирането на генни множества се използва таблица за множеството и втора таблица за членовете на множествата, като съществува 1-към-N релация между двете (фиг. 5.2). Информацията за всеки анотиран VCF файл с генетични варианти може да се представи като йерархична структура: файлът включва множество гени, всеки от които е засегнат от множество варианти, а всеки варинт може да притежава много анотации (фиг. 5.3). Съответно, тази йерархия е моделирана поседством четири таблици: файлове, гени, варианти и анотации. Всяка от тях съдържа външен ключ към предишната и по този начин образува N-към-1 релация с нея.

Фигура 5.2: Диаграма на структурата на базата данни



Фигура 5.3: Йерархично представяне на информацията от анотиран VCF файл

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] Artem Chebotko, Andrey Kashlev, and Shiyong Lu. A big data modeling methodology for apache cassandra. pages 238–245, 06 2015.

[4] P. Cingolani, V. M. Patel, M. Coon, T. Nguyen, S. J. Land, D. M. Ruden, and X. Lu. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet*, 3:35, 2012.

[5] 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.

[6] Pablo Cingolani, Fiona Cunningham, Will Mclaren, and Kai Wang. Variant annotations in vcf format. *January (January)*, 2018.

[7] 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.

[8] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[9] Johan T. den Dunnen, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux,

Timothy Smith, Stylianos E. Antonarakis, and Peter E.M. Taschner. Hgvs recommendations for the description of sequence variants: 2016 update. *Human Mutation*, 37(6):564–569, 2016.

[10] 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.

[11] 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.

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

[13] 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.

[14] Marshall J. Heger A. and the open source community.

[15] Richard D Hipp. SQLite, 2020. Available at https://www.sqlite.org.

[16] Timo Kersten, Viktor Leis, Alfons Kemper, Thomas Neumann, Andrew Pavlo, and Peter Boncz. Everything you always wanted to know about compiled and vectorized queries but were afraid to ask. *Proc. VLDB Endow.*, 11(13):2209–2222, sep 2018.

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

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

[19] Lydie Lane, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D. Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, Catherine Zwahlen, and Amos Bairoch. neXtProt: a knowledge platform for human proteins. *Nucleic Acids Research*, 40(D1):D76–D83, 12 2011.

[20] 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.

[21] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

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

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

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

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

[30] The pandas development team. pandas-dev/pandas: Pandas, February 2020. Available at https://doi.org/10.5281/zenodo.3509134.

[31] Sue Povey, Ruth Lovering, Elspeth Bruford, Mathew Wright, Michael Lush, and Hester Wain. The hugo gene nomenclature committee (hgnc). *Human genetics*, 109(6):678–680, 2001.

[32] 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.

[33] Mark Raasveldt and Hannes Mühleisen. Duckdb: An embeddable analytical database. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1981–1984, New York, NY, USA, 2019. Association for Computing Machinery.

[34] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat Biotechnol*, 29(1):24–26, Jan 2011.

[35] James T. Robinson, Helga Thorvaldsdóttir, Aaron M. Wenger, Ahmet Zehir, and Jill P. Mesirov. Variant Review with the Integrative Genomics Viewer. *Cancer Research*, 77(21):e31–e34, 10 2017.

[36] A. Ronacher and open source community. Flask, 2010. Available at https://flask.palletsprojects.com/en/2.1.x/.

[37] Herbert Schildt. The complete reference java, 2020.

[38] Dougherty J. Schutz S., Casbon J., 2012. Available at https://github.com/dridk/PyVCF3.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

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

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

[46] 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.

[47] M. Wang, K. M. Callenberg, R. Dalgleish, A. Fedtsov, N. K. Fox, P. J. Freeman, K. B. Jacobs, P. Kaleta, A. J. McMurry, A. Prlić, V. Rajaraman, and R. K. Hart. hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update. *Hum Mutat*, 39(12):1803–1813, 12 2018.

[48] 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.

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

[50] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, and et al. Ensembl 2020. *Nucleic Acids Res*, 48(D1):D682–D688, 01 2020.

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