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.
NGS - Next-Generation Sequencing - Технологии за секвениране от ново поколение.
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 [2]:

„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.“

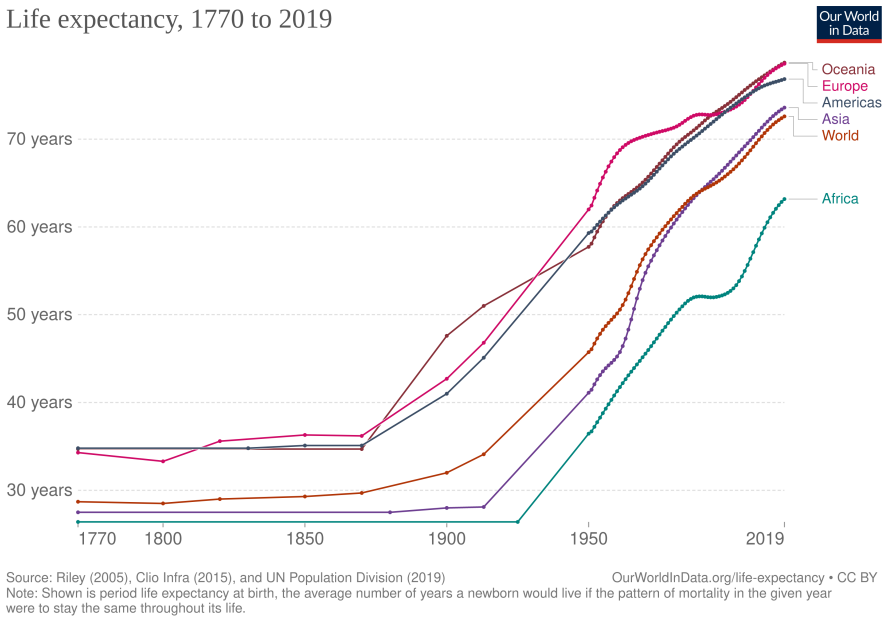
Макар времето да няма каузална връзка с ефектите на стареенето, то корелацията помежду им е причина обикновено да се говори за ефектите на стареенето като за нещо, настъпващо с напредването на възрастта.

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 [20]; загуба на телесна маса [57]; влошен слух[18]; повишен риск от хронични заболявания [31][46]; хронична болка [41]; 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 [63] и др.

2.1.3 Демографски и икономически ефекти

През последния един век очакваната продължителност на живота в целия свят драстично се е повишила [70] (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 [33]. 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% [42]. 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) [43].



Фигура 2.1: Очаквана продължителност на живота за различни региони през периода 1770-2019 [70].

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

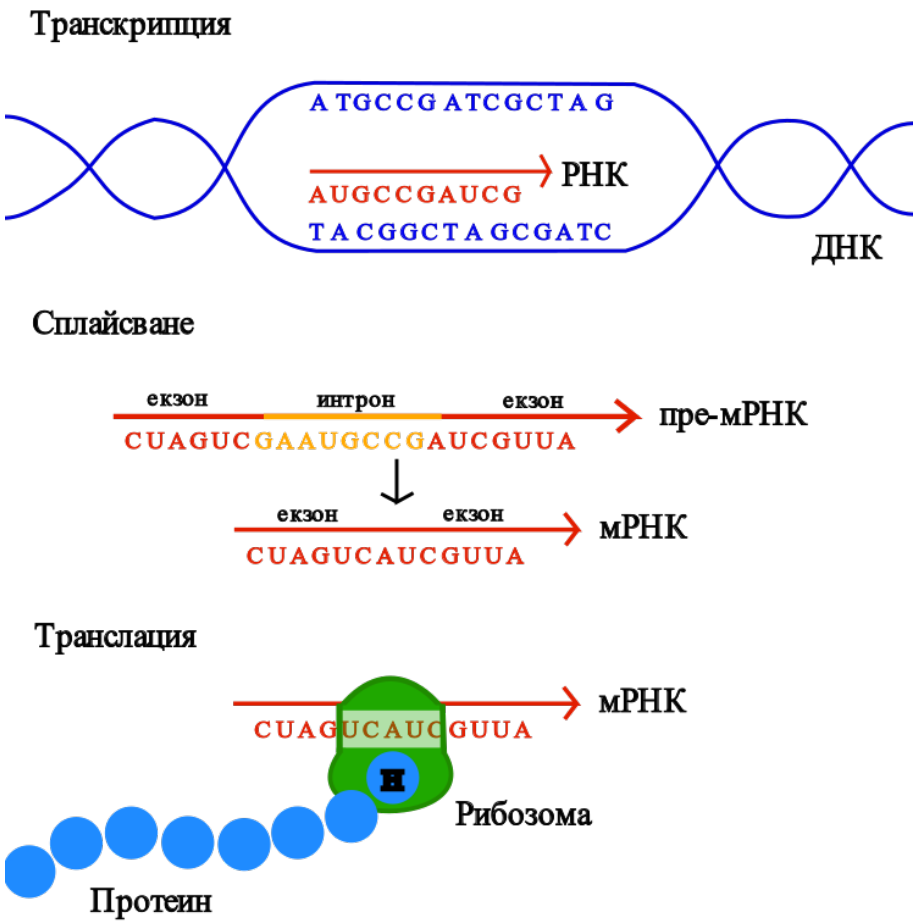
В тази секция ще разгледаме фунадменталните принципи на генетиката. Ще направим кратък обзор на начина на съхранение на генетичната информация и процесите, чрез които тя бива изразена, за да повлияе на фенотипа. С това целим да дадем базов биологически контекст, чрез който да бъдат разбрани по-нататъшните 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 [28, 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 [56]. 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 [28, 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 [36]. 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 [63]. 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 [64]. Външните мутагени могат да бъдат разделени на три

вида - физични, химични и биологични. Пример за физичен мутаген е радиацията [4], 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 [34]. 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 [62].

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 [21]. При всеки цикъл на делене на клетката, теломерите се скъсяват поради непълното синтезиране на изоставащата нишка от ДНК полимеразата [29]. 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 [38]. 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 [67].

Клетъчно стареене

TODO

Епигенетични изменения

TODO

2.3 TODO 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 увреждания в ДНК, активиране на клетъчното остаряване и иницииране на апоптоза (клетъчна смърт) [60]. В изследвания на хора е установено, че полиморфизми в TP53 могат да доведат до удължаване на живота, но да увеличат и смъртността от рак [61]. 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 [15].

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 [59].

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 [11].

```
##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

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 генериране на нови сурови данни и възможностите за извличане на полезна информация и познание от тях [68]. 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 [35]. 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 [7], предсказване на структурните и функционални промени в протеина [35] 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) [7]. 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 [53, стр. 9-10].

VEP

TODO

Annovar

TODO

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 [6]. 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 [53, стр. 9-10].

bcftools

Bcftools е програма, предлагаща множество команди за обработката и анализа на VCF файлове, компресирани VCF файлове и бинарния им еквивалент - BCF файлове. Програмата автоматично засича форматът на входните данни, без да е необходимо потребителят да го индикира. Една от

командите на Bcftools е view. Тя дава възможност за филтриране и конвертиране на VCF и BCF файлове. При филтрирането могат да бъдат зададени голям набор от разнообразни критерии, като например по зададени геномни региони, тип на полиморфизма (точкова мутация, вмъквания-изтривания, и тн.), брой на алелите, и др. Възможно е и да се филтрира по булев израз, включващ кои да е полета от VCF файла [13].

2.4.4 Геномни браузъри

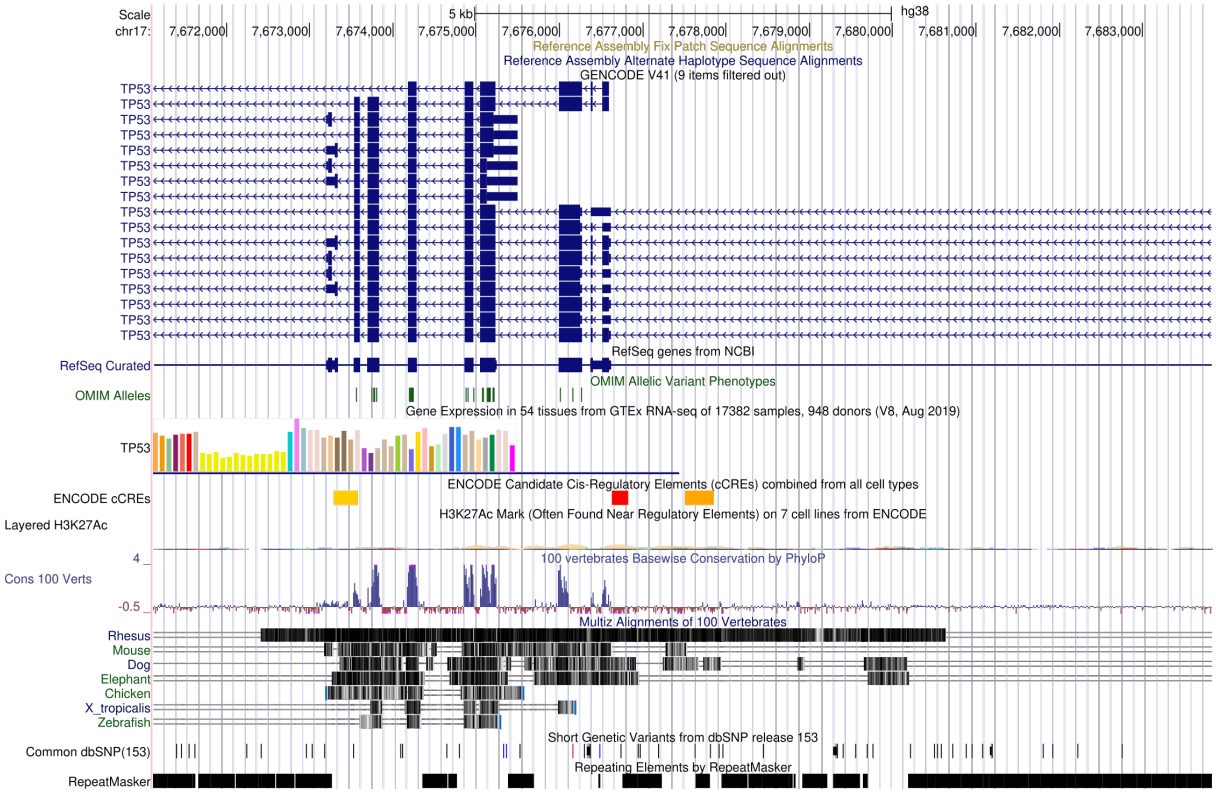
В последните години биологическите науки се превръщат в една от областите генериращи най-голямо количество данни [58]. Развитието на технологии за секвениране от ново поколене (next-generation sequencing или NGS) намалява цената и увеличава скоростта на секвенирането [54]. В резултат, геномите на вече десетки хиляди организми са секвенирани и достъпни в онлайн бази данни, предоставящи информацията публично за свободно ползване [37]. Освен секвенции, тези бази данни съдържат най-различни видове анотации, като например позициите на известни или предсказани гени, мРНК транскрипти, информация за генна експресия, генетични полиморфизми, сравнения с други организми, информация за епигенетични маркери, като нива на метилация, и др.

Есенцията на геномните браузъри е представянето на различните видове данни като отделни информационни ленти една под друга, на които геномните координати са подравнени по вертикалната ос. По този начин потребителят може лесно да види дали има определени съвпадения, или разлики в информацията от различните ленти и да направи изводи за функцията или други характеристики на дадения геномен регион. Интеграцията на секвенции и богат набор от анотации прави геномните браузъри удобна платформа, чрез която молекулярните биолози могат да разглеждат, търсят, извеждат или анализират информация за генома ефективно и удобно [65].

UCSC Genome Browser

На 22 юни 2000 година, UCSC (University of California, Santa Cruz) и другите членове на международния проект за секвениране на човешкия геном завършват първата работна версия на човешкия геном. Няколко седмици по-късно на 7 юли 2000 година, асемблирания геном става публично достъпен на уеб сайта на UCSC. Заедно с него става достъпен и графичен инструмент за разглеждането му - UCSC Genome Browser.

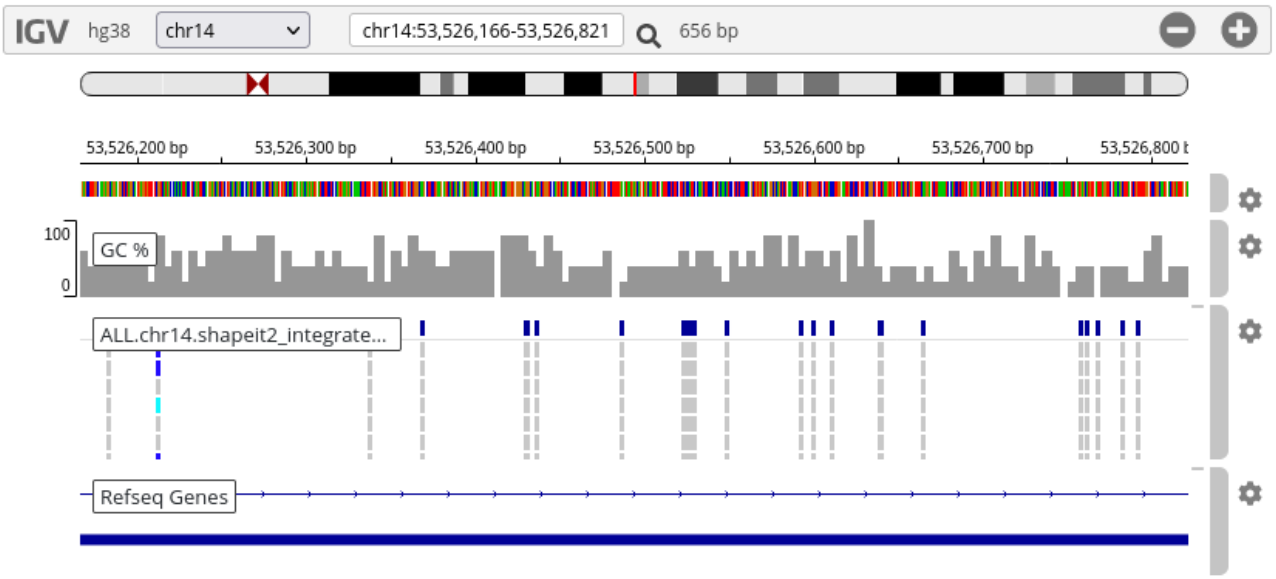
UCSC Genome Browser (фиг. 2.4) предоставя достъп и инструменти за разглеждане, сравняване и анализиране на геномите на голям набор от организми [19]. Анотациите и информацията, представена от UCSC, идва от разнообразни източници, като GENCODE Genes, NCBI RefSeq Genes, Ensembl Genes и др [39]. Също така, потребителят може да зареди анотации от свой собствен файл, като те се добавят като допълнителна информационна лента в браузъра. USCS Genome Browser предоставя прост графичен интерфейс, който дава възможност на потребителя да навигира из генома с функции като увеличаване и намаляване на резолюцията, преместване, търсене по ген, секвенция или геномна локация и др. При натискане върху обект, браузърът показва подробна информация за него. UCSC Genome Browser има възможност да зарежда и обработва огромни по обем файлове с анотации, без да престава да работи гладко [26]. Браузърът поддържа разнообразни видове файлове, включително VCF [39] и BAM [19]. Достъпен е както онлайн, така и като самостоятелна програма.



Фигура 2.4: Примерна визуализация на генът TP53 с UCSC (http://genome.ucsc.edu).

IGV Genome Browser

IGV (Integrative Genomics Viewer) е инструмент за визуализация и интерактивно изследване на разнообразни, големи по обем геномни колекции от данни (фиг 2.5). 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 [48][50]. Данните могат да бъдат зареждани както от публични онлайн бази данни, така и локално от данните, с които разполага потребителя.



Фигура 2.5: Примерна визуализация с IGV

Едно от най-големите предимства на IGV е това, че лесно може да бъде интегриран във външна страница. Достатъчно е да бъде импортирана JavaScript библиотеката igv.js и да се извика функция за вграждане на геномния браузър в конкретен HTML елемент от страницата. IGV браузърът може да работи с различни файлови формати, като например FASTA, BAM и VCF. Може да поддържа големи обеми от данни, като за целта използва индексни файлове, на базата на които прави асинхронни заявки, за да вземе единствено частите от файла, които трябва да бъдат визуализирани. IGV може да се конфигурира и предоставя JavaScript API за взаимодействие между браузъра и съдържащото го приложение [49].

2.4.5 Нагъване на протеини

Протеините са макромолекули, които представляват дълга верига от аминокиселини. Протеините имат роля в множество процеси в организмите, като например катализирането на реакции, репликацията на ДНК, клетъчната комуникация, транспортирането на молекули, образуването на клетъчната структура и др. Както вече бе разгледано в секция 2.2.1, ДНК служи като шаблон за транскрипцията на мРНК молекули, които биват транслирани в протеини. По този начин генетичния код определя първичната структура на протеина, тоест последователността от аминокиселини. За да стана биологически активни, обаче, протеините трябва първо да заемат сгъната (третична) структура в пространството [10].

Триизмерната (третична) структура на протеина се определя от първичната му структура, тоест от последователността от аминокиселини. Движещи сили при нагъването на протеина са хидрофобните интеракции, образуването на междумолекулярни водородни връзки и силите на Ван дер Ваалс. Тези сили въздействат докато протеина не достигне оптимална и стабилна форма [16]. Откритието, че триизмерната структура на протеина зависи почти изцяло от последователността от аминокиселини, поражда търсенето на компютърен алгоритъм, който да може да предскаже триизмерната структура, която определена аминокиселинна поредица би образувала. В контекста на тази дипломна работа, ползата от такъв алгоритъм би била, че ако можем да покажем как определен генетичен вариант на ген, свързан със стареенето, би променил триизмерната структура на кодирания протеин, това би спомогнало за разбирането на това как генетичния вариант се отразява и на функцията му.

Подходи

Съществуват множество различни методи, които се опитват да решат проблема за предсказване на триизмерната протеинова структура. Те могат да бъдат разделени на четири групи: (1) такива, които се основават на базови принципи и не използват информация от външни бази данни; (2) такива, които се основават на базови принципи и използват информация от външни бази данни; (3) threading методи, при които аминокиселиината поредица се моделира спрямо множество от структурни-скелета, взети от готова библиотека с нагъвнания, като за всяко от тях се пресмята оценка за съвместимост; (4) сравнителни (или хомоложни) методи [17]. Сравнителните методи се опитват да предскажат структурата на даден протеин

като идентифицират подобни, вече известни, протеини и ги използват за шаблон, спрямо който да моделират структурата. При тези методи се разчита на наблюдението, че при хомоложните гени (гени, които си приличат поради обща еволюционна история) триизмерната структура на протеина е по-добре запазена от аминикиселинната му последователност [24].

AlphaFold

AlphaFold е създаден от DeepMind, дъщерно дружество на Alphabet. Програмата се базира на метода на хомоложното моделиране, като използва невронни мрежи за дълбоко обучение (deep learning). През 2018 година, на тринадесетото издание на международното състезание за предсказване на протеинова структура CASP AlphaFold се класира първа [1]. Две години по-късно, втората версия на програмата AlphaFold2 отново се класира първа на четиринадесетото издание на CASP, като този път точността на предсказаните структури е далеч по-голяма от постигнатата от всички други участници [25].

2.4.6 Интегрирани софтуерни решения

Galaxy Project

Galaxy е онлайн платформа за научни изчисления, достъпна през уеб браузъра, която позволява на учени да споделят, анализира и визуализират своя собствена информация с минимални технически трудности [9]. Galaxy поддържа голямо разнообразие от стандартни файлове формати, използвани за биологически цели, и дава възможност за интегрирането на различни множества от данни и конструирането на процеси за обработването им [3]. Платформата е достъпна както онлайн, така и като софтуер с отворен код, който може да бъде инсталиран на собствен сървър или компютърен клъстър [40]. Освен, че прави биологическите изследвания по-достъпни, като премахва нуждата от познания по програмиране и компютърни науки за изграждането на биоинформатични процеси за обработка и анализ на данни, Galaxy се стреми да направи всички анализи възпроизводими. За целта, Galaxy записва за всяка стъпка от процеса входните данни, използваните инструменти и техните параметри [52]. Galaxy включва много богат набор от различни инструменти за постигането на различни цели, като например филтриране и подготвяне на входни сурови файлове със секвенции, съпоставяне на секвенции към референтни геноми, категоризиране на генетични варианти, статистически анализи и др [52].

3. Цели и задачи

3.1 Цели на дипломната работа

Дипломната работа има за цел да направи изследването на генетични варианти, свързани със старенето, по-лесно, по-достъпно и по-бързо. За това се предвижда създаването на 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 [51] 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 [47] 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 [23]. 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 [27].

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 [44] 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 [55] 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 [32] 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 [22] 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 [14]. 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 [66] 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

Геномният браузър IGV е разгледан по-подробно в секция 2.4.4. В програмното решение, IGV е интегриран в страницата за изследване на вариантите за определен ген. По този начин, потребителят има възможност да изследва визуално съвпаденията и разликите между геномната секвенция, полиморфизмите в гена и мРНК транскриптите. Всички тези обекти са представени като отделни информационни ленти в генетичния браузър.

Основните причини за избора на IGV пред други геномни браузъри са това, че може лесно да бъде интегриран в уеб приложението, поддържа VCF файлове, чрез които може да визуализира полиморфизми, лесно може да се конфигурира и предоставя програмен интерфейс, чрез който уеб приложението може да взаимодейства с него.

4.10 Външни генетични бази данни

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 [45]. 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 модели организми. Ensembl интегрира експериментални и референтни данни от множество източници на едно място. Към 2020 година, Ensembl съдържа анотираните геноми на 227 различни вида [69]. 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 [30]. 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.

4.10.4 HAGR

Human Ageing Genomic Resources (HAGR) е онлайн колекция от бази данни и инструменти, посветени на изследването на молекулярните и генетичните процеси, свързани със стареенето. HAGR включва бази данни за свързани със стареенето гени при хора и моделни организми, както и различни животински видове. Включва още бази данни за гени, свързани с клетъчното стареене, лекарства и химически съединения, свързани със стареенето, гени, свързани със удължаващи живота ефекти на различни диетични ограничения, а също и за асоциативни изследвания на човешката продължителност на живота. Всички бази данни в HAGR са курирани от експерти, а записите включват богат набор от връзки, както в рамките на HAGR, така и към външни източници на информация [59].

В текущата разработка, HAGR служи за извличване на списък от гени, потенциално асоциирани със стареенето. В страницата за обзор не определен ген е налична връзка към страницата за него в HAGR. Това позволява на потребителят да разбере повече за действието на гена, в контекста на стареенето, и да намери изследванията, които са установили връзка между двете.

5. Резултати

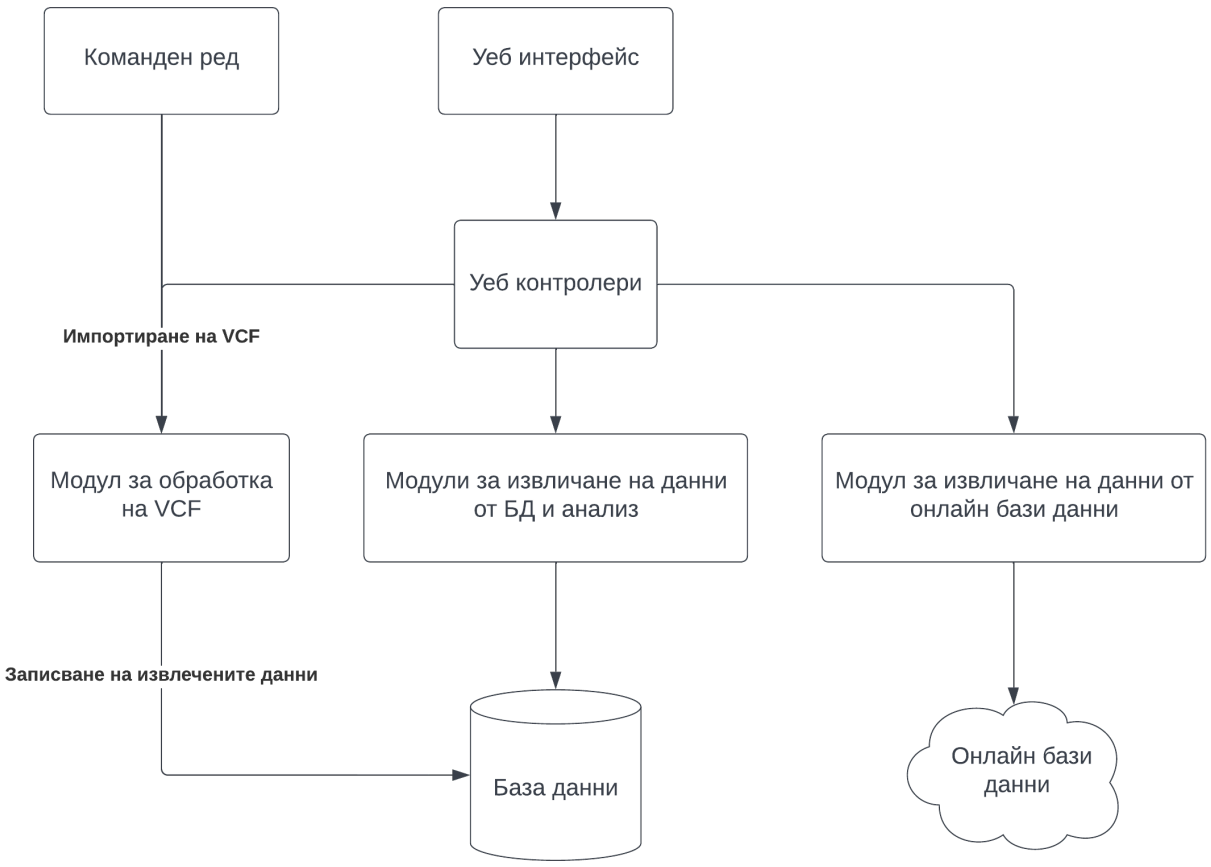
Въпреки огромното значение на процеса на стареене за индивида и обществото като цяло, науката все още не може да го обясни напълно, нито да намери достатъчно ефективни решения за справяне с негативните ефекти, които той предизвиква. Откритите асоциации между определени генетични варианти и стареенето 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

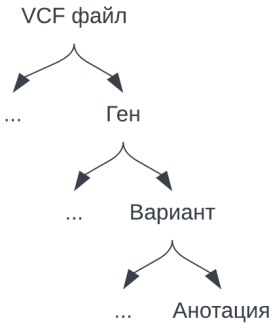
5.2 База данни

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 [12], който не се променя често. Анотациите, предоставяни от SnpEff, също спазват определен стандарт [8]. 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 [5]. 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

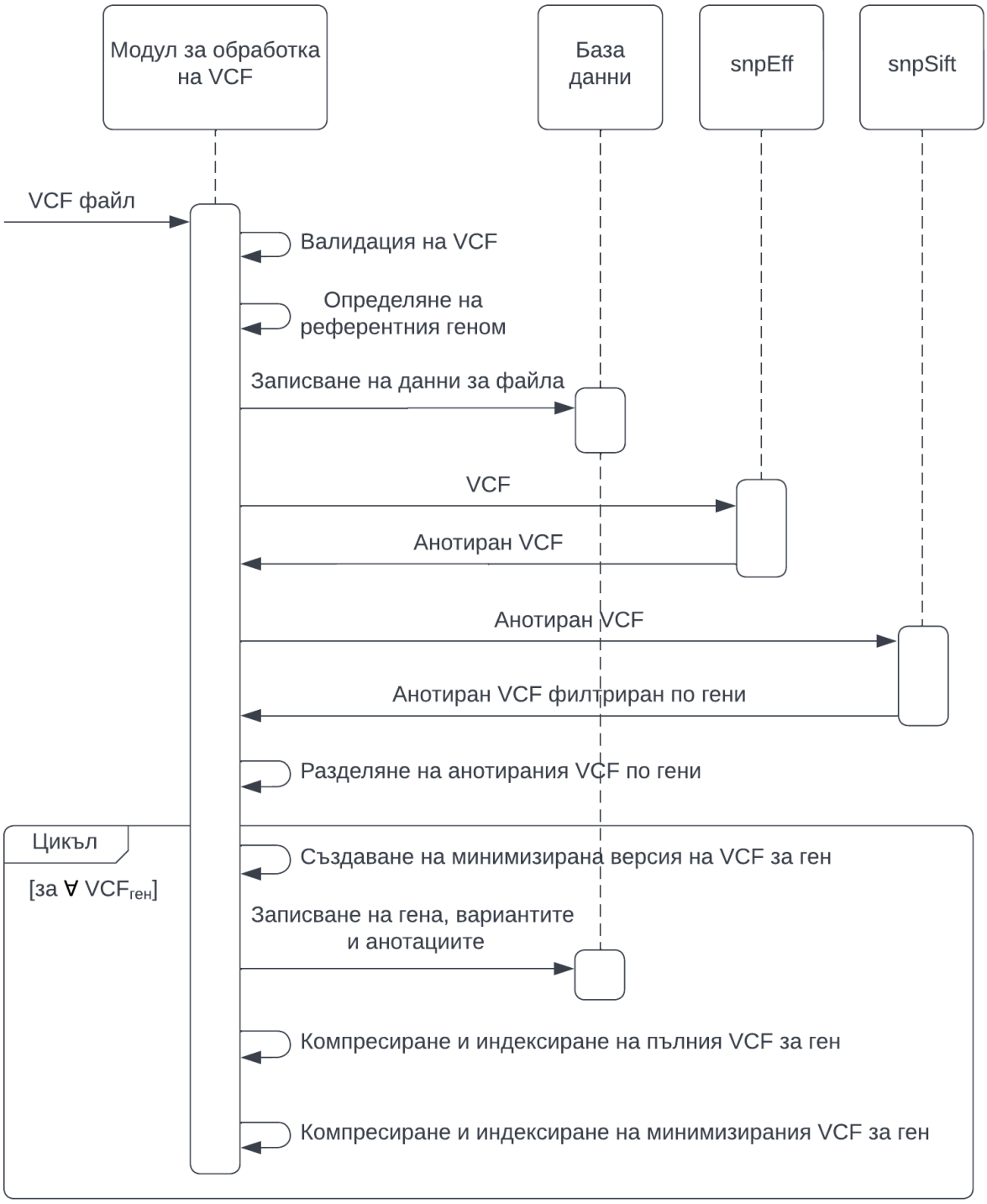
Възможността за управление на генетични множества е проста функционалност, която позволява на потребителя да създава, редактира и изтрива колекции от гени, които представляват цел за някакви изследвания. Всяко генно множество има име, описание и списък от HGNC символи на гени. Дефинирането на генно множество става посредством качване на файл, съдържащ един HGNC символ на ред. В последствие, когато качва VCF файл с генетични множества за обработка, потребителят избира спрямо кое генетично множество да бъде обработен VCF файла. Единствено варианти свързани с гените от това множество ще бъдат запазени в базата данни. Тази, макар и проста функционалност, прави програмното решение по-генерализирано и позволява изследването на различни биологични процеси. Например, ако потребителят качи списък от гени, свързани с ракови заболявания, ще може да изследва полиморфизми за тяхната връзка с рака.

5.4 Обработка на VCF при импортиране

Началната точка при работа с разработения софтуер е качване на VCF файл, съдържащ генетични варианти. Това може да бъде направено както с уеб интерфейса, така и с интерфейсът, използваш командния ред (всъщност, импортирането и обработването на VCF е единствената операция, която командния ред може да извършва към момента). При качването на входен VCF файл за обработка, потребителят подава и генно множество, определящо кои гени да бъдат разгледани при обработката (виж секция 5.3).

Първата стъпка, която се извършва при обработката на входния VCF е проверка, дали същият файл вече е бил изследван. За целта се калкулира неговата SHA256 хеш сума. Тази сума се използва като главен ключ на таблицата с файлове в базата данни. Ако сумата вече съществува в базата данни, значи файлът вече е бил импортиран и обработката приключва. Ако файлът не е срещан досега, обработката продължава. Прави се валидация, която проверява, че файлът не използва твърде стара версия на VCF стандарта (поддържат се версии не по-малки от 4.0), и се извлича информация за това спрямо кой референтен геном е създаден файлът (фиг. 5.4).

Следва същинската обработка на входния VCF файл. Първо, външните програми SnpEff и SnpSift, съответно за анотация и филтриране, биват стартирани като отделни процеси. Двата процеса биват свързани посредством pipe оператор, така че анотираните VCF редове, които SnpEff връща, да се подадат директно като входни данни за SnpSift. По този начин SnpEff добавя анотация към редовете на входния VCF, която включва и HGNC символа на гена, с който дадения вариант е свързан. SnpSift филтрира редовете, които му се подават, като оставя само тези, които се отнасят за гени, включени в генното множество, което потребителят е избрал. Резултатът се чете от основната програма, която разглежда всеки ред за да определи за кой ген се отнася и го записва в нов частичен VCF файл, който съдържа само вариантите за дадения ген. Резултатът от тези стъпки е колекция от VCF файлове, всеки от които е анотиран и съдържа само генетичните варианти за един единствен ген. Следващата стъпка е обхождането на всеки от тези файлове, изчитането на данните в Pandas DataFrame обект, който преминава поредица трансформации, и записването на крайната информация в базата данни. Също така, за всеки от тези VCF файлове се създава минимизирана версия, която изключва вариантите определени като незначителни от SnpEff (по-конкретно, за които полето impact има стойност „MODIFIER“). Тази минимизирана версия по-късно се използва за зареждане на генетичните варианти в геномния браузър, като по този начин се осъществява оптимизация на производителността и намаляване на шума за потребителя. Минимизираният и пълният VCF файлове за всеки ген също така се компресират и индексират посредством tabix функционалността (виж секция 4.5), предоставяна от модула pysam (виж секция 4.6). Компресията служи за оптимизация на изискванията към хардуерните ресурси, а създаването на индексни файлове е необходимо за да могат по-късно файловете да се зареждат и визуализират от геномния браузър.



Фигура 5.4: Диаграма показваща опростено описание на процеса по обработка на VCF файлове с генетични варианти.

5.5 Операции върху импортиран VCF

След като един входен VCF, съдържащ генетични варианти, бъде импортиран и обработката му приключи, потребителят може да изследва резултатите. В тази секция са представени основните функции, които програмното решение предоставя.

5.5.1 Списък на файлове

Началната страница предоставя списък с наличните файлове, които вече са импортирани. Потребителят може да избере дали да изследва данните за някой от наличните файлове, да стартира обработката на нов файл или да изтрие информацията за съществуващ файл (фиг. 8.1).

5.5.2 Обзор на файл

Страницата за обзор на файл дава най-обща информация за вече обработен файл. Това включва референтния геном, спрямо който файлът е създаден, генното множество, спрямо което е изследван, броя гени засегнати от полиморфизми и броят анотации. По-детайлна разбивка е дадена в табличен вид, която показва за всеки ген броя варианти и техните анотации спрямо оценената им тежест (фиг. 8.2).

5.5.3 Обзор на ген

Страницата за обзор на ген дава обща информация за гена и служи като отправна точка за изследването на различните варианти във файла, които го засягат. Страницата предоставя връзки към различни стандартни онлайн бази данни, където потребителят може да открие допълнителна информация за конкретния ген. Тези връзки включват базите данни Ensembl (виж секция 4.10.2), Entrez, HGNC (виж секция 4.10.1) и UniProt. Следва списък от богата анотация за протеина, кодиран от гена, включваща молекулярната функция, процесите и биологичните пътища, в които участва, каталитичната дейност и др (фиг. 8.3). Налична е връзка към страницата, представяща всички варианти за гена.

Страницата включва и по-детайлна разбивка на анотираните варианти за гена, представени в табличен вид. Таблицата показва броят на анотациите за варианти, групирани по оценка на тежестта и генетичен ефект.

Оценката на тежестта може да е четири варианта: модификатор, ниска, средна, висока. Възможните ефекти са разнообразни и включват например: синонимни мутации, придобиване на стоп кодон, мутации, изменящи смисъла или изместващи рамката, мутации в интрони и др (фиг. 8.4). Всяка от стойностите в таблицата представлява връзка, при натискането на която, потребителят бива препратен към страницата за варианти, като се добавя филтър за конкретната стойност, така че потребителят да може да разгледа точно този вид варианти.

5.5.4 Варианти на ген

В страницата за варианти да ген потребителят има достъп до геномен браузър, чрез който може да изследва полиморфизмите засягащи гена. При отварянето на страницата, браузърът автоматично се фокусира върху геномния регион, съдържащ списъка от варианти във VCF файла. Геномният браузър представя няколко различни информационни ленти. Първата лента показва генните варианти. Тъй като VCF файлът може да е за един експеримент, или множество експерименти, тази лента показва също така и комбинациите от изразени варианти за всеки от експериментите във VCF файла. Втората лента показва гените, като са обособени техните екзони и интрони. Третата лента показва транскриптите, които се кодират от даден ген, заедно с техните Ensembl идентификатори (фиг. 8.5).

В страницата също е налична странирана таблица с всички варианти за гена. Таблицата представя началната и крайната точка на полиморфизма, референтната и алтернативната нуклеотидна поредица, типа и подтипа на полиморфизма (точкова мутация, вмъкване, изтриване и тн.). Налични са и контроли, с които могат да бъдат прилагани сложни филтри върху данните от таблицата (фиг. 8.6).

5.5.5 Засегнати транскрипти

За всеки генетичен вариант можем да достъпим страница със засегнатите от него транскрипти. Страницата предоставя таблица с Ensembl идентификатора на транскрипта, неговия тип (например протеинокодиращ, псевдоген и тн), ефекта, който полиморфизма оказва върху този транскрипт, тежеста на промяната и HGVS номенклатура (виж секция 4.7) на промяната на протеина, ако има такава. Потребителят също така може да сравни и референтния и алтернативния протеин за този транскрипт (фиг. 8.7).

6. Дискусия

7. Изводи

Библиография

[1] Mohammed AlQuraishi. AlphaFold at CASP13. *Bioinformatics*, 35(22):4862–4865, 05 2019.

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

[3] D. Blankenberg, N. Coraor, G. Von Kuster, J. Taylor, and A. Nekrutenko. Integrating diverse databases into an unified analysis framework: a Galaxy approach. *Database (Oxford)*, 2011:bar011, 2011.

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

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

[6] 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.

[7] 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.

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

[9] The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):W345–W351, 04 2022.

[10] Thomas E Creighton. Protein folding. *Biochemical journal*, 270(1):1, 1990.

[11] 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.

[12] 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.

[13] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 02 2021. giab008.

[14] 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.

[15] 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.

[16] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annu Rev Biophys*, 37:289–316, 2008.

[17] Márcio Dorn, Mariel Barbachan e Silva, Luciana S. Buriol, and Luis C. Lamb. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry*, 53:251–276, 2014.

[18] 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.

[19] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J.

Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(Database issue):D876–882, Jan 2011.

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

[21] 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.

[22] Marshall J. Heger A. and the open source community, 2009. Available at https://github.com/pysam-developers/pysam.

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

[24] K. Illergård, D. H. Ardell, and A. Elofsson. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins*, 77(3):499–508, Nov 2009.

[25] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Applying and improving alphafold at casp14. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1711–1721, 2021.

[26] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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

[37] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A Chen, Nikos C Kyrpides, and T B K Reddy. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research*, 49(D1):D723–D733, 11 2020.

[38] 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.

[39] J. Navarro Gonzalez, A. S. Zweig, M. L. Speir, D. Schmelter, K. R. Rosenbloom, B. J. Raney, C. C. Powell, L. R. Nassar, N. D. Maulding, C. M. Lee, B. T. Lee, A. S. Hinrichs, A. C. Fyfe, J. D. Fernandes, M. Diekhans, H. Clawson, J. Casper, A. Benet-Pagès, G. P. Barber, D. Haussler, R. M. Kuhn, M. Haeussler, and W. J. Kent. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res*, 49(D1):D1046–D1057, 01 2021.

[40] A. Nekrutenko, D. Baker, N. Coraor, B. Chapman, E. Afgan, and J. Taylor. Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics*, 11 Suppl 12:S4, Dec 2010.

[41] 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.

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

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

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] James T. Robinson, Helga Thorvaldsdóttir, Douglass Turner, and Jill P. Mesirov. igv.js: an embeddable javascript implementation of the integrative genomics viewer (igv). *bioRxiv*, 2020.

[50] 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.

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

[52] M. C. Schatz. The missing graphical user interface for genomics. *Genome Biol*, 11(8):128, 2010.

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

[54] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1):16–18, Jan 2008.

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

[56] 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.

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

[58] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. Big data: Astronomical or genomical? *PLOS Biology*, 13(7):1–11, 07 2015.

[59] 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.

[60] 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.

[61] 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.

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

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

[64] 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.

[65] Jun Wang, Lei Kong, Ge Gao, and Jingchu Luo. A brief introduction to web-based genome browsers. *Briefings in Bioinformatics*, 14(2):131–143, 2013.

[66] 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.

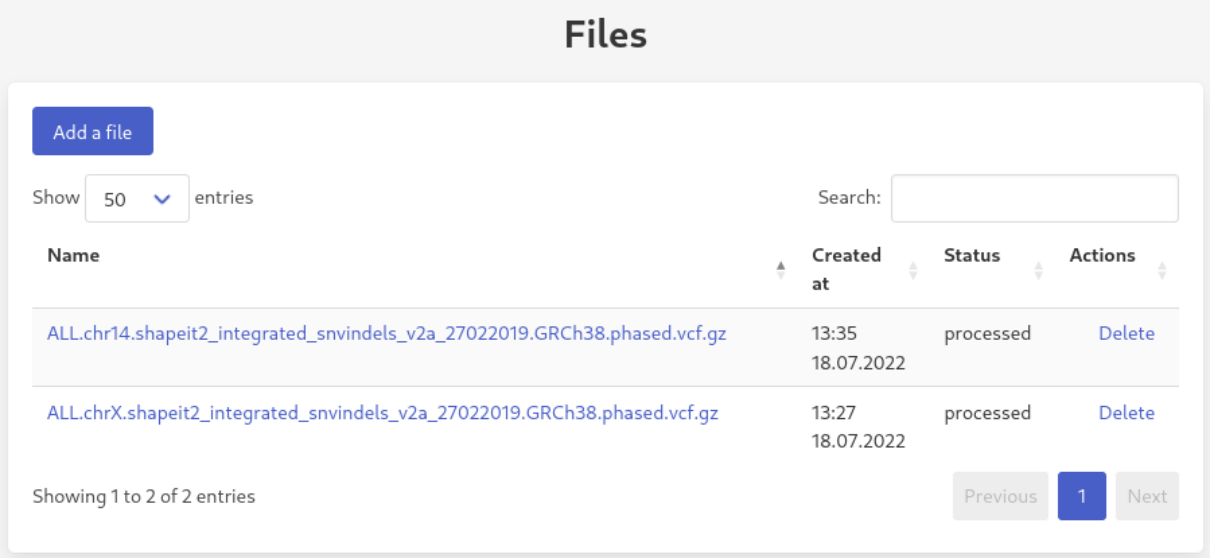
[67] Kurt Whittemore, Elsa Vera, Eva Martínez-Nevado, Carola Sanpera, and Maria A. Blasco. Telomere shortening rate predicts species life span. *Proceedings of the National Academy of Sciences*, 116(30):15122–15127, 2019.

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

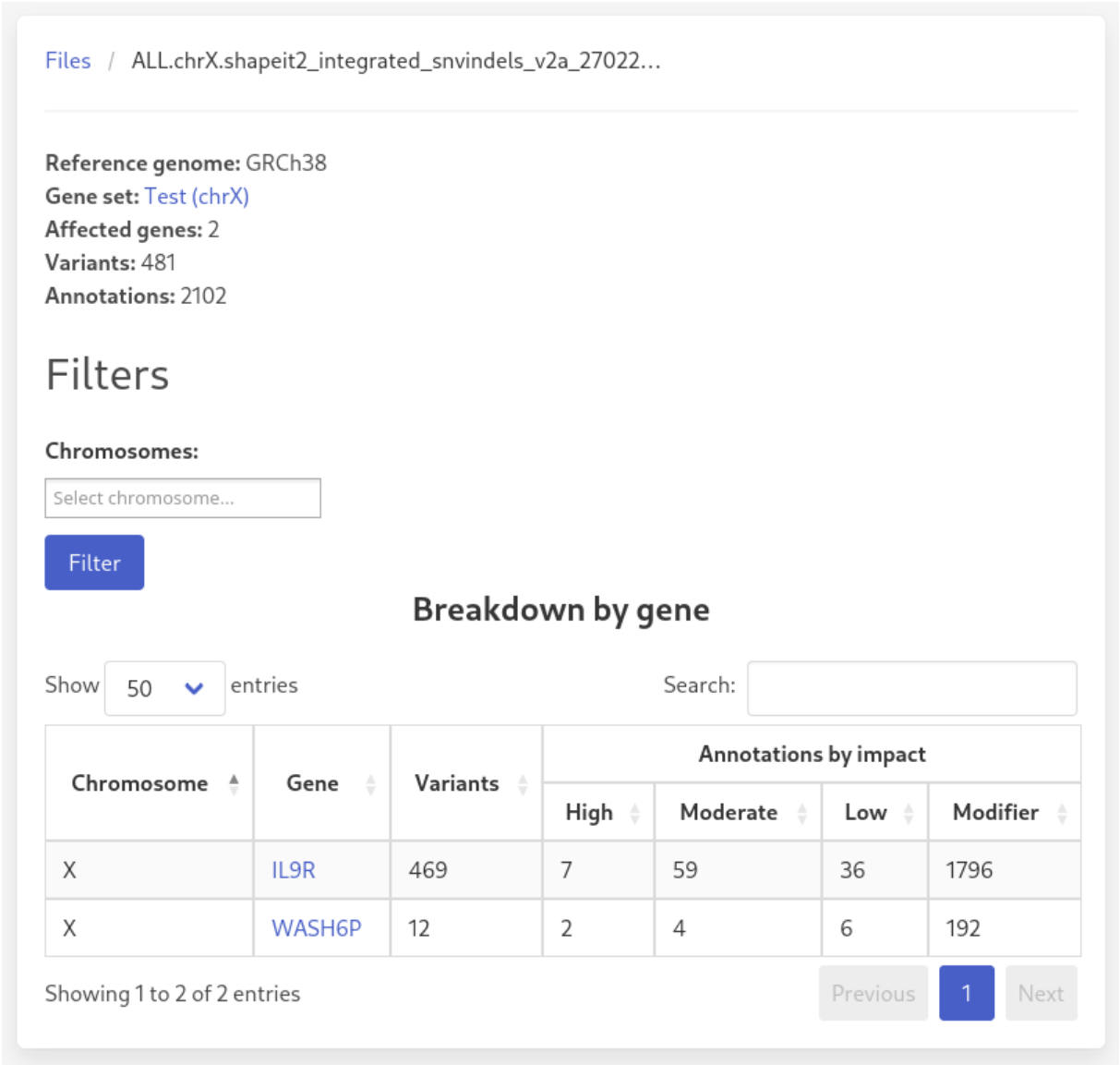
[69] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, and et al. Ensembl 2020. *Nucleic Acids Res*, 48(D1):D682–D688, 01 2020.

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

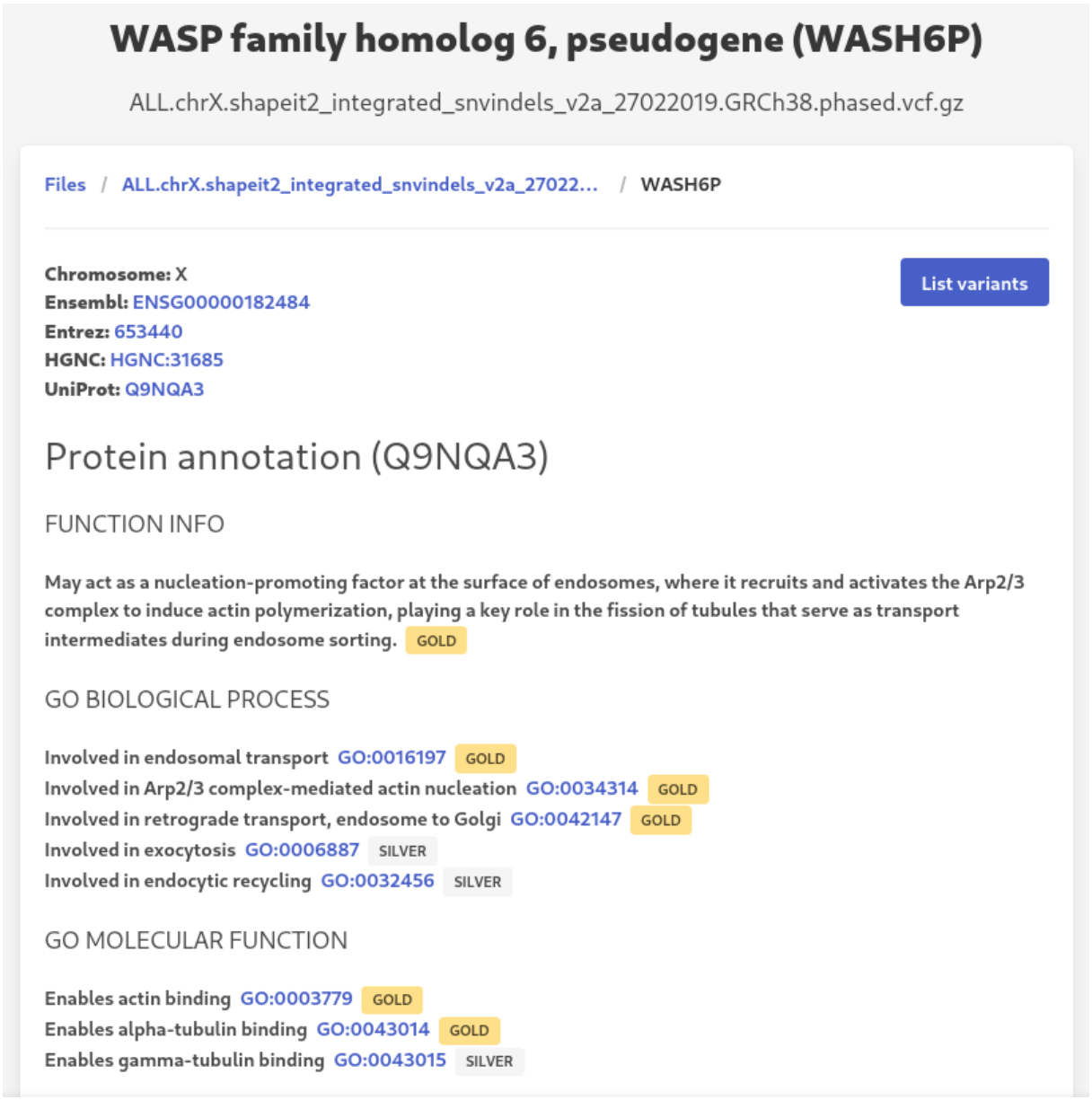
8. Приложение



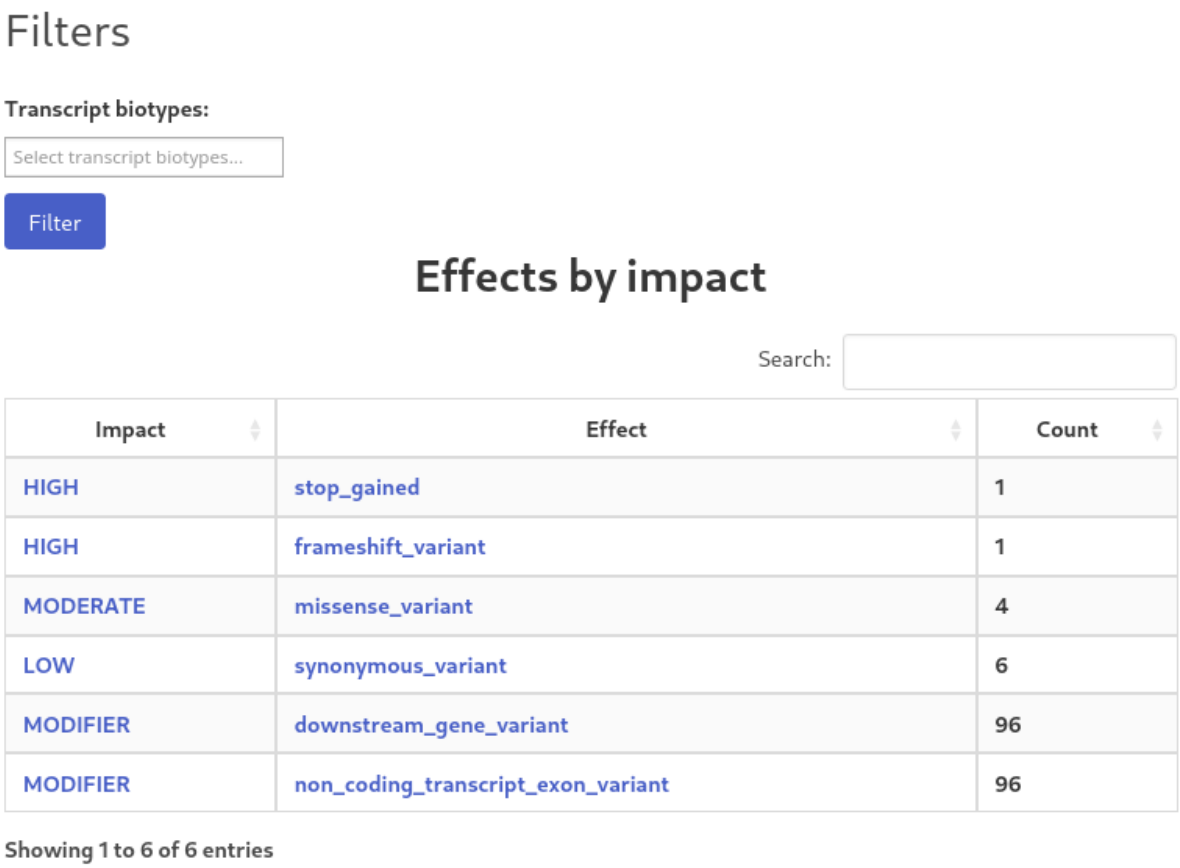
Фигура 8.1: Списък от файлове.



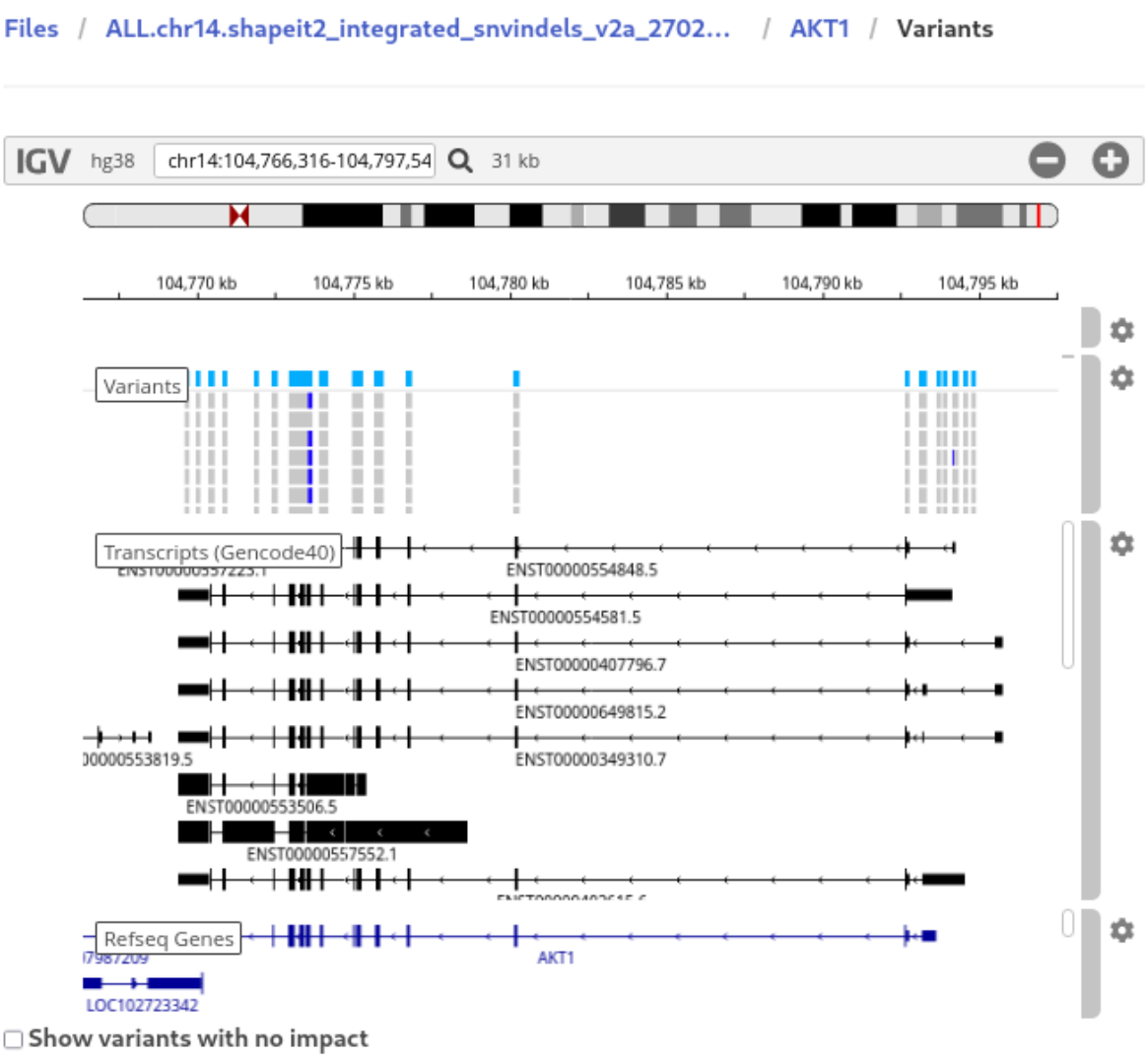
Фигура 8.2: Обзор на файл.



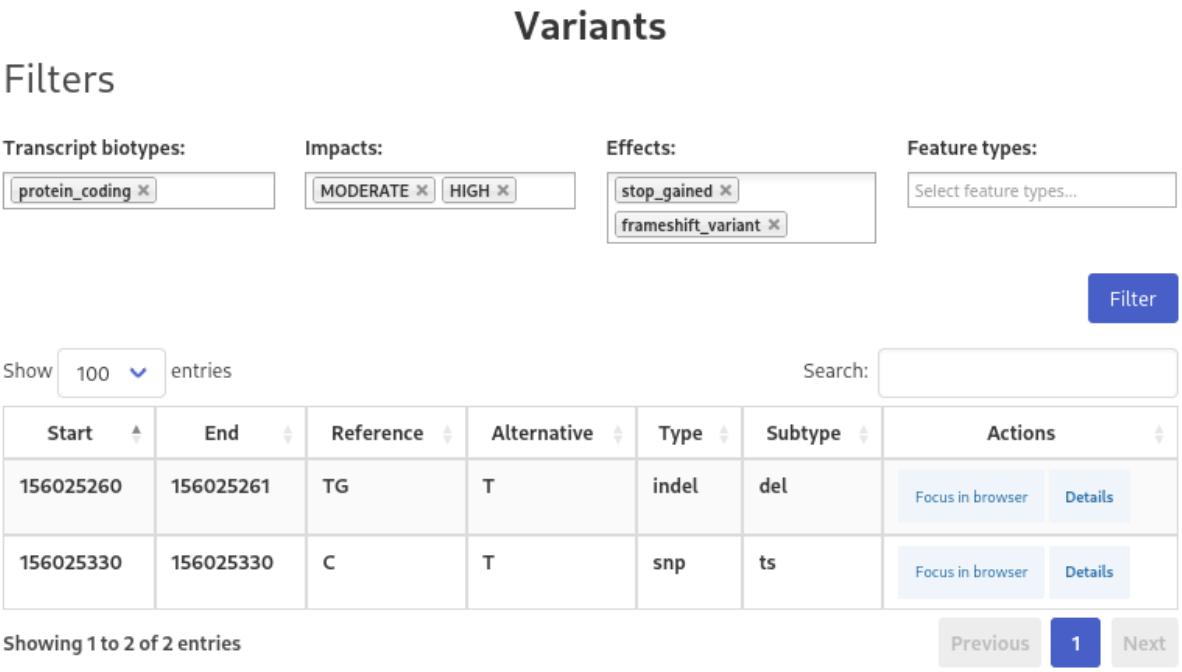
Фигура 8.3: Обща информация за гена.



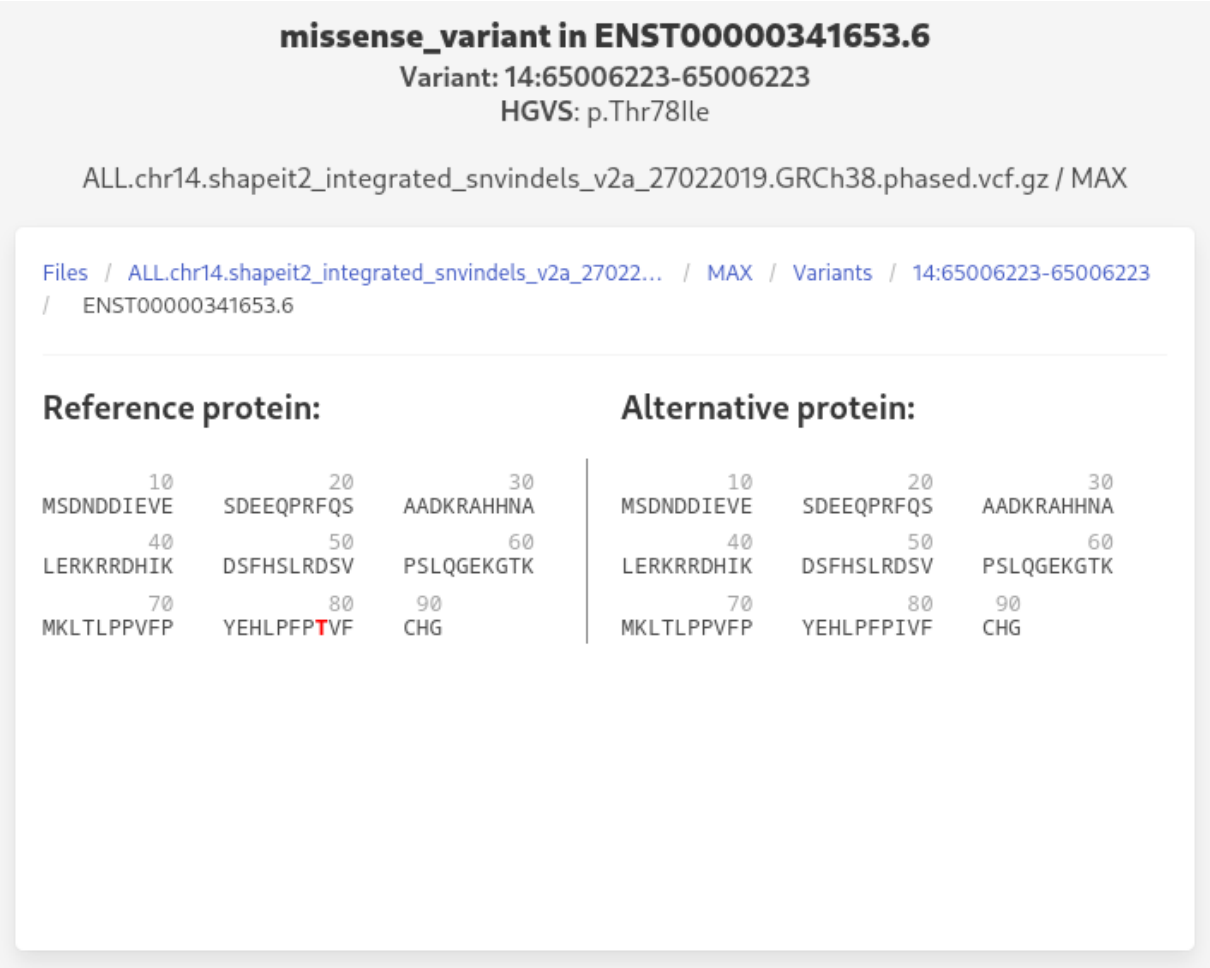
Фигура 8.4: Разбивка на анотираните варианти и техните предсказани ефекти.



Фигура 8.5: Геномен браузър за изследване на генетичните варианти.



Фигура 8.6: Таблица с геномни варианти и филтри.



Фигура 8.7: Сравнение на референтен и алтернативен протеин. Мястото на полиморфизма е оцветено в червено.