UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Anđela R. Mijailović

IMPLEMENTACIJA DODATKA ZA SOFTVER ANNOVAR ZA PRIKAZ FUNKCIJE I FENOTIPA GENA

master rad

Beograd, 2021.

Mentor:

dr Jovana Kovačević, docent
Matematički fakultet, Univerzitet u Beogradu

Članovi komisije:

prof. dr Gordana Pavlović-Lažetić, redovni profesor
Matematički fakultet, Univerzitet u Beogradu

dr Nevena Veljković, naučni savetnik
INN Vinča, Univerzitet u Beogradu

Datum odbrane: _____

Najmilijima

Naslov master rada: Implementacija dodatka za softver Annovar za prikaz funkcije i fenotipa gena

Rezime: U ovom radu implementiran je dodatak za softver *Annovar* pod nazivom *AnnoPI* (skraćeno od *Annovar Plug In*). Ovaj softver omogućava anotaciju datoteka sa genetskim varijacijama što podrazumeva generisanje nove datoteke sa različitim podacima o genetskim varijacijama koje se nalaze rasute po različitim bazama podataka. Pored podataka koje *Annovar* dodeljuje, od značaja mogu biti i drugi podaci o genima kao što su funkcija i fenotip gena. *AnnoPI* omogućava automatsko pronalaženje ovih podataka u javno dostupnim bazama podataka i zajedno sa podacima iz datoteke generisane od strane *Annovar*-a kreira datoteku sa objedinjenim podacima o genetskim varijacijama. Aplikacija je testirana nad podacima vezanim za Aškenazi trio koji su deo značajnog projekta Genom u boci. Pisana je u programskom jeziku *Python*, a rezultat predstavlja *.html* stranice koje pregledno prikazuju objedinjene informacije.

Ključne reči: gen, genom, nukleotid, varijacija, SNV, *Annovar*, ontologije, *GO*, *HPO*, funkcija, fenotip

Sadržaj

Glava 1

Uvod

Humani genom predstavlja ogroman izazov, kako za istraživanje, tako i za reprezentaciju. Da bi se pročitao, potrebno je prvo podeliti ga na segmente. Njihove dužine moraju biti dužine najviše 100-200 baznih parova jer sekvenceri ne mogu očitati duže segmente. Iscepkan humani genom potom treba sastaviti kao slagalicu, s tim što nam često nije poznata slika koju treba dobiti. Rekonstrukcija genoma je olakšana ukoliko su delovi za slaganje veći. Posebni alati koji služe za sklapanje genomske slagalice na osnovu očitavanja nazivaju se asembleri.

Brzi razvoj tehnologija za sekvenciranje genoma doveo je do mogućnosti da se sekvenciraju genomi različitih organizama. Pored toga, za neke vrste postoje takozvani referentni genomi koji predstavljaju genome nastale kombinovanjem genoma određenog broja uzoraka te vrste. U genomskim sekvencama kod različitih jedinki iste vrste, na nivou populacije, može doći do razlika u nukleotidima na istim pozicijama. Ove razlike nazivamo genetskim varijacijama. Naučnicima koji se bave ovom oblašću od značaja su različite informacije o genetskim varijacijama kao što su pozicija na hromozomu na kojoj se nalazi varijacija, kom genu pripada, kakvi su njeni predviđeni funkcionalni efekti, itd. Ove informacije se nalaze rasute po raznim bazama podataka, a za neke od njih je potrebno pokrenuti i različite alate. Jedan od takvih alata je softver *Annovar*. Cilj ovog rada jeste implementacija dodatka *AnnoPI* za softver *Annovar* koji bi omogućio prikaz informacija o genetskim varijacijama koje prikuplja Annovar zajedno sa dodatnim informacijama koje *Annovar* ne uključuje, konkretno informacijama o funkciji i fenotipu gena na kom se varijacije nalaze. Implementirani softver je javno dostupan na *GitHub* repozitorijumu projekta *AnnoPI* [2].

Na početku rada, u poglavlju 2, dat je prikaz bioloških pojmova koji se koriste

u ovom radu, pre svega genoma i genetskih varijacija. Poglavlje 3 posvećeno je bioinformatičkim podacima gde centralno mesto zauzimaju ontologije u kojima su zapisani podaci o funkciji i fenotipu gena (*GO* i *HPO*). Prikazan je i projekat „Genom u boci" u okviru koga se i nalaze podaci testirani u ovom radu. Tehnički okvir i upotreba softvera *Annovar* objašnjeni su detaljnije u poglavlju 4, dok je peto poglavlje posvećeno rezultatima i detaljnom opisu praktičnog dela rada.

Glava 2

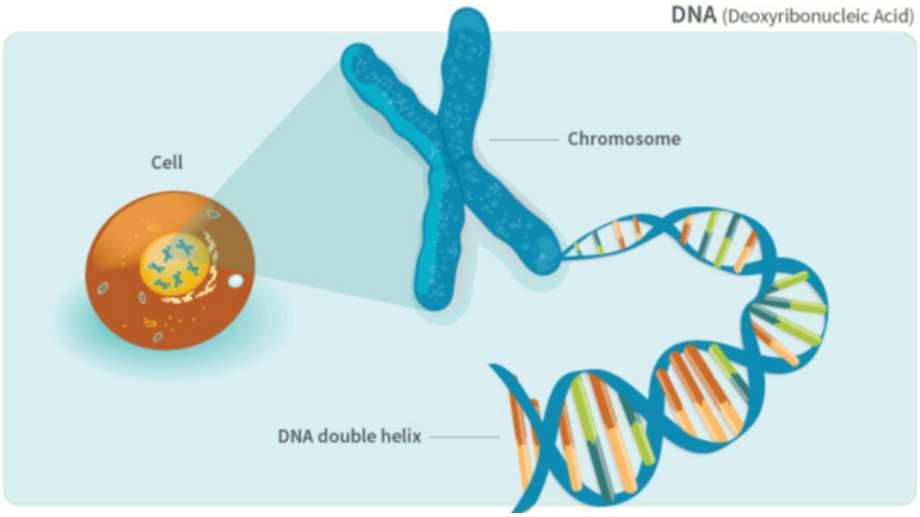
Genom i SNV

Ćelija je osnovna jedinica funkcije i građe svakog živog bića. U zavisnosti od toga kako je u ćeliji organizovan genetski materijal, postoje prokariotske i eukariotske ćelije. Kod prokariotskih ćelija genetski materijal je rasut po ćeliji, dok je kod eukariotskih ćelija genetski materijal grupisan u jezgru koje je od ostatka ćelije odvojeno membranom. Eukaritoske ćelije sadrže organele u kojima se odvijaju ćelijski procesi. Jezgro eukariotske ćelije naziva se jedro i predstavlja najveću organelu zaduženu za regulaciju svih ćelijskih procesa. U tim procesima učestvuju jedinjenja kao što su nukleinske kiseline, sastavljene od 4 nukleotida, i proteini, sastavljeni od 20 esencijalnih aminokiselina.

U ćeliji postoje dve vrste nukleinskih kiselina - dezoksiribonukleinska kiselina (DNK) i ribonukleinska kiselina (RNK). Glavni nosioci genetičke informacije su molekuli DNK. Osnovu svakog nukleotida čini jedna od četiri azotne baze: adenin, citozin, guanin i timin. Nukleotidi RNK sadrže uracil umesto timina. Azotne baze se skraćeno obeležavaju sa $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{G}$ i $\boldsymbol{T}$, odnosno $\boldsymbol{U}$. Postoji pravilo po kom se ove baze spajaju i to je adenin sa timinom, citozin sa guaninom, a sve sa ciljem formiranja jedinica koje nazivamo bazni parovi (skraćeno ***bp***). Sa stanovišta računarstva, DNK možemo posmatrati kao nisku nad azbukom $\{A, T, C, G\}$, a proteine kao nisku nad azbukom od 20 aminokiselina [6].

Za vreme deobe ćelije u jedru se mogu uočiti hromozomi koji imaju ključnu ulogu u nasleđivanju. Ćelija može imati jednostruki skup hromozoma (haploidna ćelija) i dvostruki skup hromozoma (diploidna ćelija), pri čemu je svaki skup dobijen od jednog roditelja. Molekul DNK se nalazi u hromozomima. Celovit deo DNK potreban za sintezu jednog proteina ili jednog molekula RNK naziva se ***gen***. On predstavlja fizičku i funkcionalnu jedinicu nasleđivanja. Ilustracija ćelije, hromozoma i DNK

data je na slici 2.1.



Slika 2.1: Ćelija, hromozom i DNK [1]

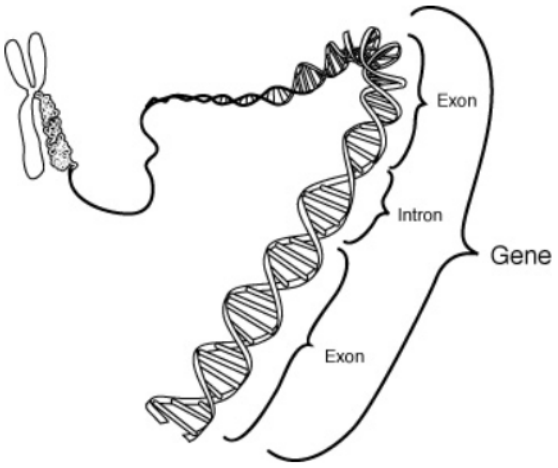
Svaki gen zauzima određeno mesto na hromozomu, ali kako u diploidnoj ćeliji postoje dva hromozoma sa tim genom na istom mestu, može se reći da se jedan gen javlja u dva oblika. Ovi oblici nazivaju se ***aleli***. Nameće se pitanje kako se utvrđuje položaj gena na hromozomu. Ovaj postupak se naziva ***mapiranje gena*** [4]. Jedinstveni skup svih gena jedinke naziva se ***genotip***. Na složeni način, uz uslove spoljašnje sredine, genotip upravlja skupom svih osobina jednog organizma, funkcije i ponašanja organizma koji nazivamo ***fenotip*** [5].

Geni nose informaciju o broju, vrsti i redosledu aminokiselina u proteinskom lancu. Informaciju o redosledu aminokiselina zapravo daje redosled nukleotida u DNK i predstavlja recept za sintezu proteina. Međutim, ne nose svi delovi gena informaciju za sintezu proteina. Oni segmenti gena koji nose informaciju se nazivaju ***egzoni***, a oni koji ne nose ***introni***.

Možemo uočiti da je građa gena eukariota mozaična: deo gena koji nosi šifru ispresecan je delovima koji ne nose šifru. Stoga, za gene je karakterističan diskontinuitet genetičke informacije. Ilustracija je data na slici 2.2.

Kod prokariota introni ne postoje već su njihovi geni neprekinuti nizovi kodirajućih nukleotida. Biološki značaj introna i njihova funkcija nisu još uvek razjašnjeni. Introni su našli praktičnu primenu u kriminologiji i sudskoj medicini poznatoj kao genetički otisci prstiju.

[1] https://www.ancestry.com/lp/where-is-dna-found



Slika 2.2: Ilustracija mozaične strukture gena eukariota [9]

2.1 Genom

Kompletan nasledni materijal sadržan u jednoj haploidnoj ćeliji predstavlja ***genom***. Humani genom sadrži oko tri milijarde i 200 hiljada baznih parova. Svaka jedinka ima jedinstveni genom pa tako među genomima različitih jedinki ima razlika. Različitosti na nivou genoma nazivamo ***genetskim varijacijama***. One mogu biti na pojedinačnom nukleotidu i skraćeno ih označavamo sa ***SNV*** (eng. *Single Nucleotide Variation*). Drugi tip varijacija je ***mala varijacija*** (eng. *small indel*). Postoje i ***uzastopne varijacije*** (eng. *Copy Nucleotide Variation*, skraćeno *CNV*) i ***strukturne varijacije*** (eng. *Structural Variations*, skraćeno *SV*).

Sa jedne strane, varijacije utiču na fenotip i samim tim na to kako izgledamo. Sa druge strane, uzročnici su bolesti, stoga je važno otkriti ih. Koristeći prednosti sekvencioniranja nove generacije, genetske varijacije se mogu posmatrati na nivou čitavog genoma. Ono što želimo je da razumemo tehnike koje se odnose na pronalaženje varijacija na pojedinačnim nukleotidima.

Postupak određivanja genomske sekvence naziva se ***sekvenciranje genoma***. Suštinski, predstavlja određivanje redosleda nukleotida na nivou molekula DNK. Sekvenciranje se vrši tako što se iz uzorka očitaju podsekvence DNK koje se nazivaju ***očitavanja*** (eng. *reads*) a koje je nakon toga neophodno sastaviti u polaznu DNK sekvencu pomoću asemblera [7].

Humani referentni genom predstavlja „prosečan" humani genom koji je izračunat na određenom broju uzoraka. Kako je ovakav genom sastavljen sekvencioniranjem

DNK većeg broja davalaca, referentni genom ne predstavlja skup gena nijedne pojedinačne osobe. Humani referentni genom u oznaci ***GRCh37*** (*The Genome Reference Consortium human genome (build 37)*) je trenutni referentni genom i baziran je na genomima 13 dobrovoljaca iz SAD. Na njegovom usavršavanju se i dalje radi.

2.2 SNV

Pojedinačna varijacija nukleotida (*SNV*) je tačka mutacije koja se javlja na određenom mestu u našem genomu. To mesto se naziva ***lokus*** [7]. SNV je najčešća varijacija genoma. Svaka jedinka ima oko nekih 3 miliona ovih pojedinačnih nukleotidnih varijacija, dakle, jedna na svakih hiljadu nukleotida. Pojedinačne varijacije nukleotida se mogu javiti u kodirajućim i nekodirajućim regionima genoma. Kodirajući regioni genoma se sastoje od ***kodona***, nukleotidnih tripleta od kojih svaki određuje pojedinačnu aminokiselinu. Ako *SNV* ne menja tip aminokiseline, onda je u pitanju ***sinonimni SNV***, u suprotnom je ***nesinonimni SNV***. Nesinonimni *SNV* mogu menjati kodon jedne aminokiseline u kodon druge aminokiseline (eng. *missense SNV*) ili kodon jedne aminokiseline u stop kodon (eng. *nonsense SNV*). Pojedinačne varijacije nukleotida mogu značajno izmeniti 3D strukturu i funkciju proteina.

Kada se *SNV* javlja u nekodirajućim regionima, uglavnom je neutralan. Međutim, mogu uticati na ekspresiju gena ako se javljaju na mestima vezivanja transkripcionog faktora (proteina koji se vezuje za specifičnu DNK sekvencu i koji ima uloge kontrole prenosa genetičke informacije sa DNK na iRNK).

2.2.1 Metode za pronalaženje SNV

Polazeći od datog skupa očitavanja dobijenog nakon sekvenciranja genoma, jedna od procedura za pronalaženje genetskih varijacija je sledeća [7]:

1. Poravnavaju se sva očitavanja u odnosu na referentni genom i kreiraju se takozvane *.bam* datoteke (ove datoteke predstavljaju kompresovane *.sam* datoteke koje sadrže informacije o poravnatim očitavanjima)

2. U slučaju pojave očitavanja koja nisu ispravna vrši se njihovo filtriranje

3. U ovom koraku sakupljaju se aleli na svim lokusima. Postoji mogućnost da su neki aleli koji se pojavljuju na određenom lokusu različiti u odnosu na referentne alele. Na jednom takvom lokusu može postojati SNV.

4. Za lokuse sa nereferentnim alelima primenjuju se različite statistike radi određivanja postojanja SNV na takvom lokusu.

5. Nakon pronalaženja svih SNV, dobijene informacije se smeštaju u *.vcf* (eng. *Variant Calling Format*) datoteku.

Glava 3

Podaci

Bioinformatika kao naučna oblast podrazumeva analizu i interpretaciju različitih tipova bioloških podataka sa ciljem istraživanja i boljeg razumevanja bioloških procesa, što se postiže razvijanjem računarskih metoda i alata. Najvažnije institucije, u čijim okvirima se i nalaze bioinformatičke baze podataka su:

- ***NCBI*** - Nacionalni centar za biološke informacije, SAD [17]
- ***EBI*** - Evropski bioinformatički insitut, Velika Britanija [11]
- ***SIB*** - Švajcarski bioinformatiči institut, Švajcarska [10]
- ***KEGG*** - Kjoto enciklopedija gena i genoma, Japan [15]

Jedna od važnih procedura u bioinformatičkim istraživanjima jeste anotiranje genetskih varijacija. Informacije o genetskim varijacijama se nalaze u datotekama. Anotacija predstavlja pridruživanje relevantnih informacija svakoj genetskoj varijaciji iz različitih baza podataka, često uz pomoć različitih alata. Primeri relevantnih informacija mogu biti: na kojoj se poziciji na hromozomu nalazi varijacija u odnosu na referentni genom, u kom je genu varijacija, kakvi su njeni predviđeni funkcionalni efekti, itd.

3.1 Ontologije

Pojam ontologije u računarstvu se odnosi na predstavljanje znanja u obliku formalno definisanog sistema pojmova (klasa, izraza, koncepata) i relacija između njih. U biološkim okvirima ovakav koncept služi za organizaciju i objavljivanje bioloških podataka.

3.1.1 *GO*

Ontologija gena, u oznaci *GO* (engl. *Gene Ontology*) predstavlja najveći izvor informacija vezanih za funkcije gena odnosno genskih produkata[1]. Podaci u *GO* su prilagođeni za čitanje i od strane čoveka ali i za računarsku obradu[12]. Funkcija genskog produkta se iz ugla ontologije gena sagledava iz tri različita aspekta: molekulskih funkcija (eng. *Molecular Function, MF*), ćelijskih komponenti (eng. *Cellular Component, CC*) i bioloških procesa (eng. *Biological Process, BP*). Molekulska funkcija podrazumeva biohemijsku aktivnost genskog produkta. Ćelijska komponenta predstavlja mesto u ćeliji gde je genski produkt aktivan. Biološki procesi predstavljaju metaboličke ili regulacione procese kojima genski produkt doprinosi. Ako uzmemo za primer *GO* anotaciju genskog produkta *cytochrome c*, sa aspekta molekulske funkcije govori se o aktivaciji enzima oksidoreduktaze. Kada je reč o biološkom procesu, onda je to oksidativna fosforilacija. Ćelijska komponenta ovog produkta je mitohondrijski matriks.

Kako *GO* projekat nastoji da predstavlja aktuelne informacije, konstatno se ažurira i podleže revizijama. Izmene su često na nedeljnom nivou. Na zvaničnoj internet stranici *GO* projekta [13] moguće je preuzeti aktuelne verzije datoteka sa ontologijama. Ove datoteke imaju ekstenziju *.gaf*. Struktura datoteke je takva da su kolone razdvojene tabovima. Svaki red u *.gaf* datoteci predstavlja jednu vezu između genskog produkta i *GO* funkcije.

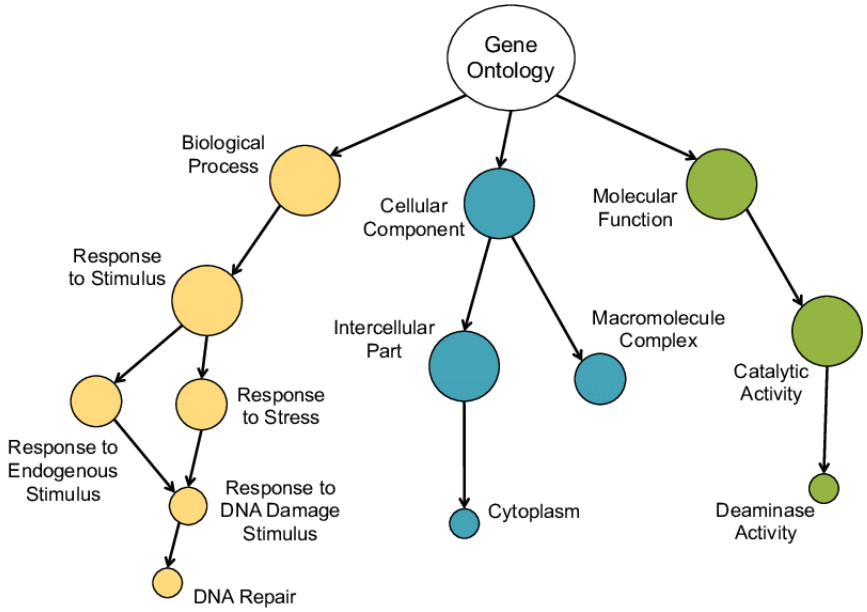
GO termi

Elementi ontologije gena su *GO* klase ili termi. Njihovi suštinski elementi su jedinstveno određeni kod i naziv terma. Jedinstveno određeni kod (u oznaci *GO ID*) predstavlja identifikator koji se sastoji od sedam cifara sa *GO:* prefiksom. Ime terma je takvo da je razumljivo čoveku, npr. vezivanje aminokiselina (engl. *amino acid binding*). Term, takođe, ima aspekt i može biti jedan od tri pomenuta: molekulska funkcija, ćelijska komponenta i biološki proces. Term ima i definiciju. Ona predstavlja opis onoga što term predstavlja. Između termova postoje relacije. One mogu biti *je, je deo* i *reguliše*.

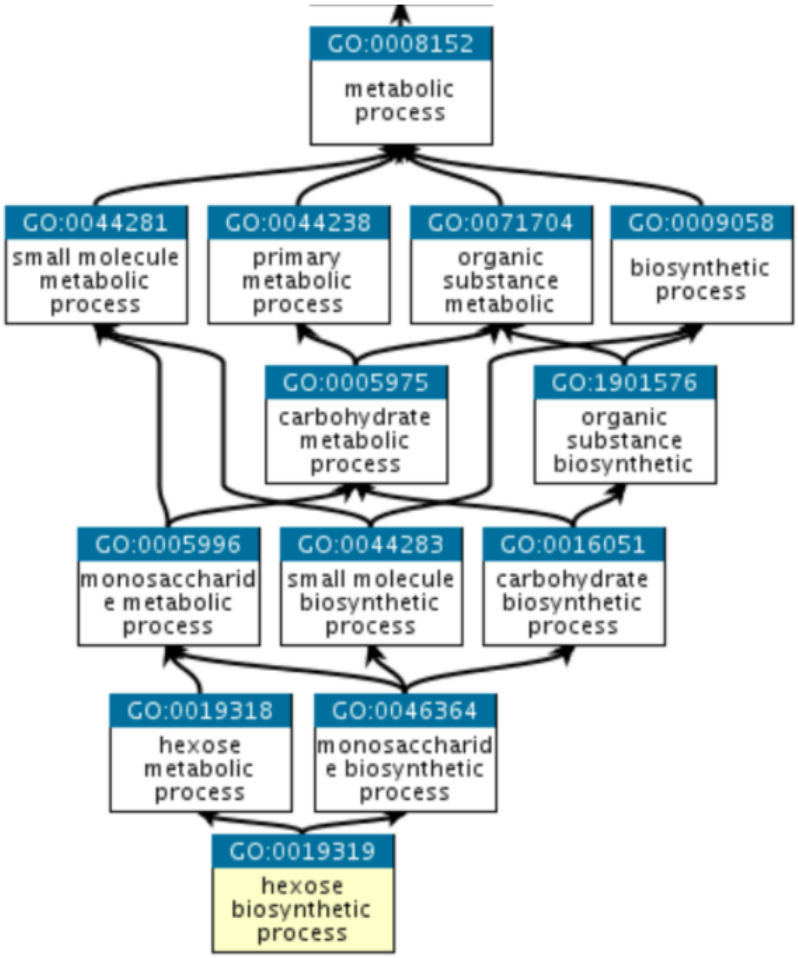
Sa aspekta računarstva, *GO* možemo posmatrati kao usmereni aciklički graf. U takvom grafu, *GO* termi odnosno funkcije predstavljaju čvorove, a relacije između njih predstavljaju grane. *GO* struktura je hijerarhijska. Relacija "dete-

[1]Pod genskim produktom se podrazumeva protein, nekodirajuća RNK ili makromolekulski kompleks

roditelj" je relacija "je" (engl. "is-a"). To, zapravo, predstavlja odnos specijalizacija-generalizacija. Svaki čvor opisuje specifičniju funkciju u odnosu na svog pretka. U ovoj hijerarhiji moguće je da dete ima više roditelja. Koren ovakvog grafa predstavlja najopštiju funkciju i nosi naziv ontologije (***MFO, BPO, CCO***) dok listovi predstavljaju najspecifičnije funkcije. Ilustracije *GO* ontologije date su na slikama 3.1 i 3.2.



Slika 3.1: Deo ontologije *GO* [1]



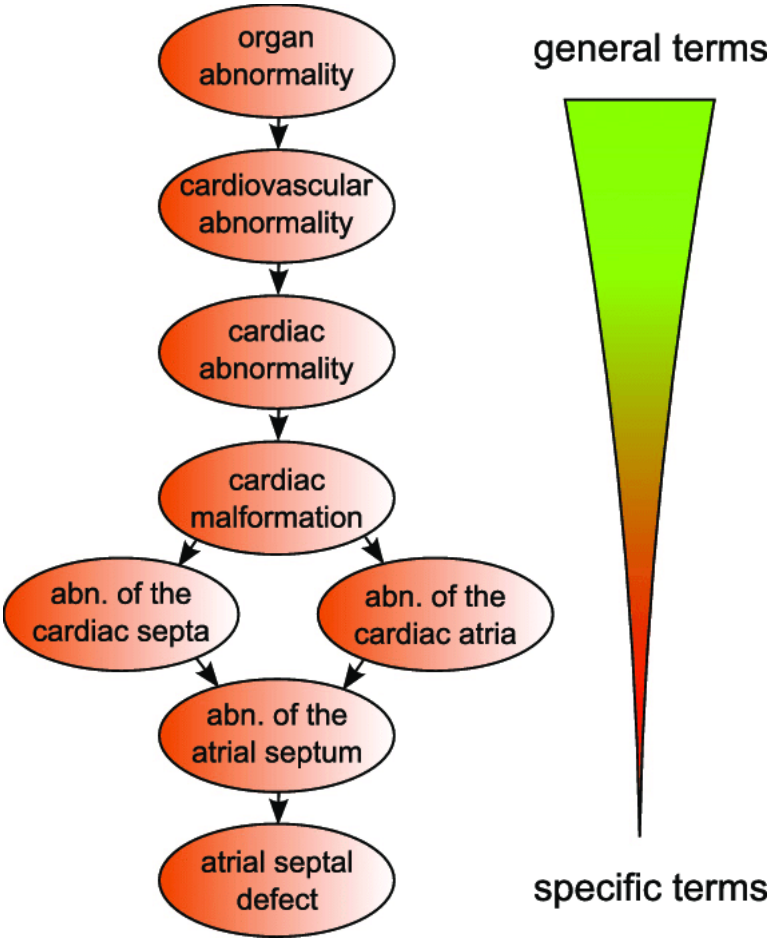
Slika 3.2: Primer podgrafa ontologije *GO* [12]

3.1.2 HPO

The Human Phenotype Ontology u oznaci ***HPO*** obezbeđuje standardizovan rečnik fenotipskih abnormalnosti koje se sreću kod ljudskih bolesti. Svaki izraz u *HPO* opisuje jednu fenotipsku abnormalnost. To mogu biti opštiji izrazi kao što je npr. abnormalnost građe uha, ali, takođe, mogu biti i prilično specifični kao što je korioretinalna atrofija. *HPO* pruža ontologiju medicinski relevantnih fenotipova, anotacije vezane za fenotip koji se odnosi na bolesti i algoritama koji se primenjuju u ovom kontekstu. Institucije poput međunarodne organizacije za retke bolesti, registri kliničkih laboratorija, biomedicinskih resursa i kliničkih softverskih alata sve više usvajaju *HPO* kao standard za fenotipske abnormalnosti čime u velikoj meri doprinose novonastalim naporima u globalnoj razmeni podataka za identifikovanje uzroka bolesti[14].

Elementi ontologije fenotipa imaju svoj identifikator. To je sedmocifreni kod sa prefiksom ***HP:***.

Kao i prethodno opisanu ontologiju, ontologiju fenotipa možemo posmatrati kao usmereni aciklički graf. Čvorovi predstavljaju fenotipske pojmove a grane su relacije između njih. Razlika u odnosu na stabla se ogleda u tome što čvor koji predstavlja specijalizovaniji term (potomak) može imati više predaka - manje specijalizovanih terma. Upravo ovakva relacija ilustruje relaciju „is-a", odnosno „je". Na slici 3.3 dat je prikaz dela *HPO* ontologije.



Slika 3.3: Deo ontologije HPO [3]

3.2 Genom u boci

Konzorcijum „Genom u boci"[16] (eng. *Genome In A Bottle*, skraćeno *GIAB*) je javno-privatno-akademski konzorcijum pod pokroviteljstvom američkog Nacionalnog instituta za standarde i tehnologiju (eng. *National Institute of Standards and Technology, NIST*). Cilj ovog konzorcijuma je razvoj kompletne tehničke infrastrukture (referentnih standarda, referentnih metoda, referentnih podataka) koja bi omogućila korišćenje rezultata sekvencioniranja humanog genoma u kliničkoj praksi kao i u tehnološkim inovacijama. Glavni zadatak *GIAB*-a je karakterizacija humanih genoma od strane stručnjaka za dalje korišćenje, međusobno upoređivanje i računarsku obradu. Pored ostalih podataka, *GIAB* obezbeđuje podatke o genetskim varijacijama za tri humana genoma, takozvani ***Aškenazi trio***, nad kojima je testiran praktični deo ovog rada. U poglavlju 5 je naveden detaljan opis korišćenih datoteka.

Glava 4

Annovar

Postoji sve veći jaz između stvaranja sirovih podataka o sekvenciranju i izdvajanja značajnih bioloških informacija. Jedan od načina za efikasno dobijanje ovih informacija je upotreba softvera *Annovar* (ime predstavlja akronim engleskih reči *ANNOtate VARiation*). Ovaj alat omogućava korišćenje najsvežijih informacija vezanih za anotiranje genetskih varijacija koje se odnose na genome čoveka, kao i miša, muve, crva... *Annovar* može dodeliti anotacije na nivou gena, regiona, filtera, a ima i druge funkcionalnosti.

Na nivou gena možemo doći do informacije da li *SNP* ili *CNV* uzrokuju izmene u proteinu, kao i koje su aminokiseline zapravo izmenjene. Za nazive gena mogu biti korišćeni identifikatori iz različitih nomenklatura: *RefSeq, UCSC, ENSEMBL, GENCODE, AceView*. Na nivou regiona govorimo o nalaženju varijacija u posebnim regionima genoma, kao što su npr. zaštićeni regioni u okviru 44 vrste, predviđena mesta vezivanja faktora transkripcije, duplirani regioni u određenim segmentima, itd. U okviru filtera govorimo o identifikovanju varijacija koje se mogu naći u specijalizovanim bazama podataka, na primer je li varijacija prijavljena u *dbSNP*[1], koja je učestalost alela u „Genom projektu 1000" i drugo[18].

4.1 Tehnički okvir i upotreba

Annovar je napisan u programskom jeziku *Perl*. Najjednostavniji način za upotrebu *Annovar*-a je korišćenjem programa `table_annovar.pl` koji se poziva iz komandne linije. Obavezni argumenti komandne linije su:

[1] baza podataka koja sadrži, između ostalog, pojedinačne nukleotidne varijacije (https://www.ncbi.nlm.nih.gov/snp/)

- putanja do ulazne datoteke
- putanja do direktorijuma u kom se nalaze baze podataka u vidu tekstualnih datoteka

Neki od opcionih argumenata su:

- `-h` (prikazivanje poruke sa uputstvom za korišćenje)
- `-buildver <string>` (verzija genoma)
- `-out <string>` (prefiks imena izlazne datoteke)
- `-remove` (brisanje pomoćnih datoteka)
- `-protocol <string>` (upućuje na imena datoteka koje predstavljaju baze podataka, a koje se nalaze u direktorijumu zadatom kao drugi obavezni argument komandne linije)
- `-operation <string>` predstavlja operacije nad navedenim datotekama argumenta `-protocol` koje upućuju na nivo na kom se vrši anotiranje:
  - *g* (nivo gena)
  - *gx* (nivo gena uz dodatne anotacije)
  - *r* (nivo regiona)
  - *f* (nivo filtera)
- `-nastring <string>` (znak za prikaz praznog polja)
- `-vcfinput` (naznačavanje da je ulazna datoteka u *.vcf* formatu i da će se pored izlazne *.txt* datoteke generisati i izlazna *.vcf* datoteka)
- `-polish` (proteinska notacija)
- `-otherinfo` (prikaz dodatnih informacija)
- `-csvout` (generisanje izlazne datoteke u *.csv* formatu)

Kompletna lista argumenata sa detaljnim objašnjenjima može se naći u uputstvu za korišćenje softvera *Annovar* [8].

Primer pozivanja programa je:

```
perl table_annovar.pl mom/mom.vcf humandb/ -buildver hg19 -out mom
-remove -protocol refGene -operation g -nastring .  -vcfinput -polish
```

Ulazna datoteka za ovaj program je datoteka koja sadrži informacije o genetskim varijacijama, najčešće tipa *.vcf*. Svaki red ovakve datoteke sadrži poziciju na kojoj postoji varijacija u odnosu na referentni genom u odnosu na koji je vršeno mapiranje, identifikator varijacije, tip varijacije, itd. Primer ulazne datoteke u *.vcf* formatu je dat na slici 4.1.

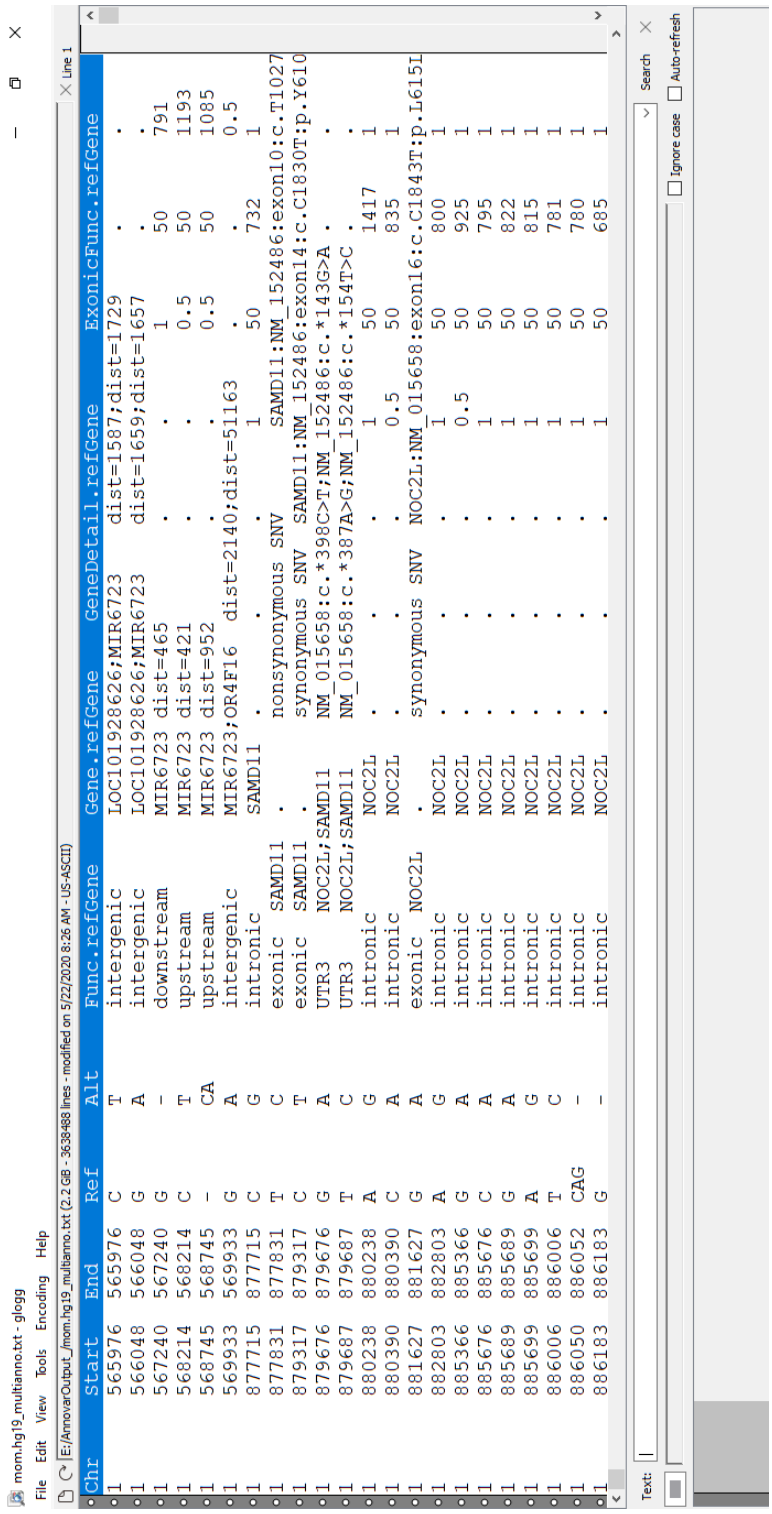
Izlazna datoteka programa je tabelarno organizovana datoteka u kojoj svaki red predstavlja skup anotacija za jednu genetsku varijaciju. Podrazumevani format je *.txt*, a navođenjem odgovarajućih argumenata u komandnoj liniji moguće je generisanje datoteke i u *.vcf* ili *.csv* formatu. Prvih nekoliko kolona odgovara kolonama iz ulazne datoteke. Svaka od narednih kolona odgovora protokolu (jednom ili više njih) navedenom u komandnoj liniji prilikom poziva programa. Kolone `Func.refGene`, `Gene.refGene`, `GeneDetail.refGene`, `ExonicFunc.refGene` i `AAChange.refGene` sadrže informacije kako mutacija utiče na gensku strukturu. Jedna od kolona u nastavku je `ExAC*` i ona predstavlja učestalost alela u uzorcima. Ostale kolone sadrže skorove predviđanja za nesinonimne varijacije.

U slučaju kada je ulazna datoteka *.vcf* tipa, program *Annovar*, između ostalog, generiše i novu izlaznu *.vcf* datoteku sa dodatnim poljem *Info* koje je popunjeno informacijama vezanim za anotacije. Vrednost ovog polja na početku sadrži *Annovar_Date* i završava se sa *Allele_End*. U slučaju više alela na istom lokusu, polje *Info* će imati više ovakvih segmenata, po jedan za svaki alel.

Primeri izlaznih datoteka u *.txt* i *.vcf* formatu dati su na slikama 4.2 i 4.3.

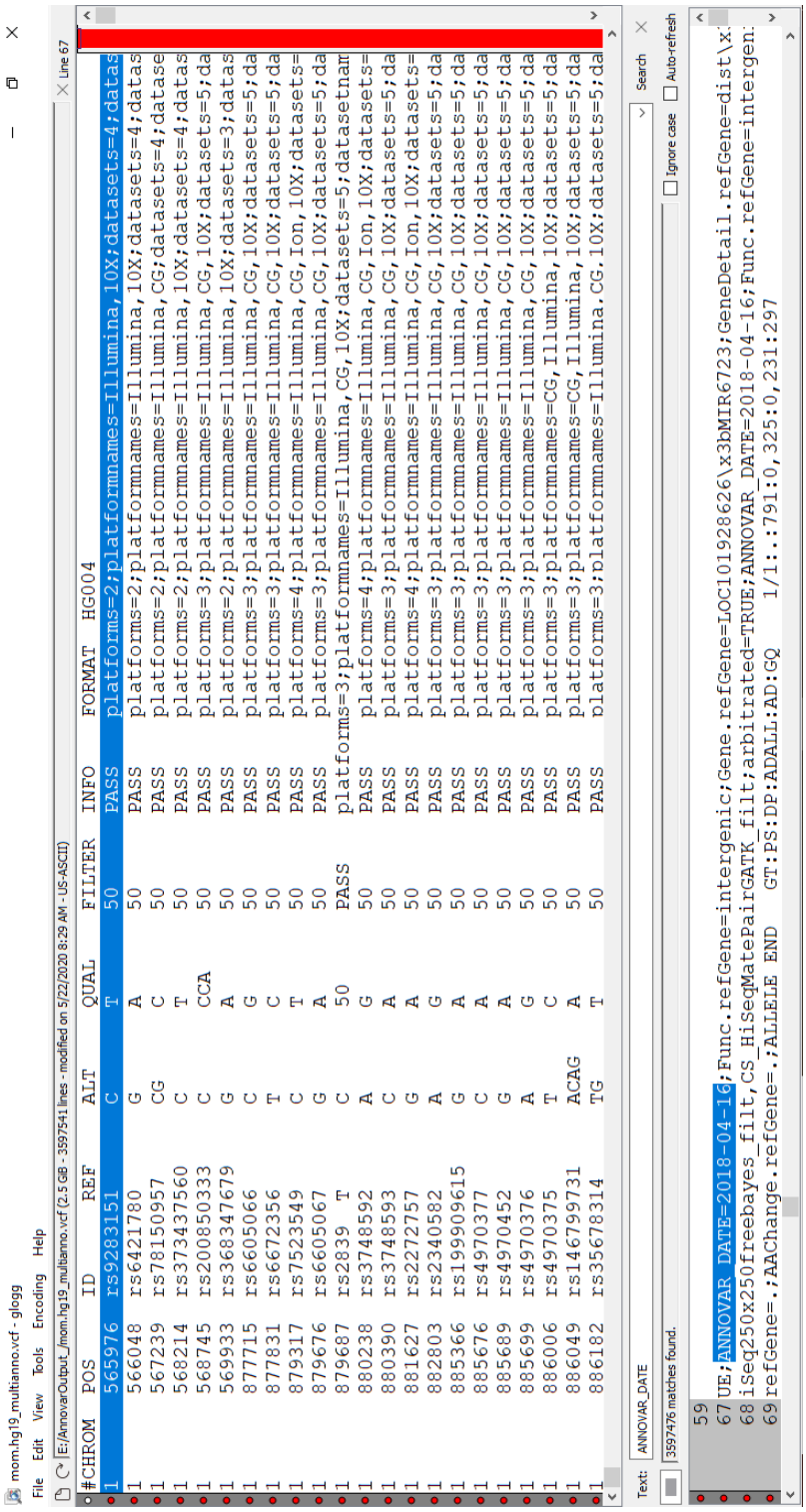
```
##FORMAT=<ID=PS,Number=1,Type=String,Description="Phase set in which this variant falls">
#CHROM  POS     ID        REF     ALT     QUAL    FILTER  INFO    FORMAT  HG004
1       565976  rs9283151 C       T       50      PASS    platforms=2;platformnames=Illumina,10X;datasets=4;datas
1       566048  rs6421780 G       A       50      PASS    platforms=2;platformnames=Illumina,10X;datasets=4;datas
1       567239  rs78150957        CG      C       50      PASS    platforms=2;platformnames=Illumina,CG;datasets=4;datase
1       568214  rs373437560       C       T       50      PASS    platforms=2;platformnames=Illumina,10X;datasets=4;datas
1       568745  rs200850333       C       CCA     50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       569933  rs368347679       G       A       50      PASS    platforms=2;platformnames=Illumina,10X;datasets=3;datas
1       877715  rs6605066 C       G       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       877831  rs6672356 T       C       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       879317  rs7523549 C       T       50      PASS    platforms=4;platformnames=Illumina,CG,Ion,10X;datasets=
1       879676  rs6605067 G       A       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       879687  rs2839  T       C       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;datasetnam
1       880238  rs3748592 A       G       50      PASS    platforms=4;platformnames=Illumina,CG,Ion,10X;datasets=
1       880390  rs3748593 C       A       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       881627  rs2272757 G       A       50      PASS    platforms=4;platformnames=Illumina,CG,Ion,10X;datasets=
1       882803  rs2340582 A       G       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       885366  rs199909615       G       A       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       885676  rs4970377 C       A       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       885689  rs4970452 G       A       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       885699  rs4970376 A       G       50      PASS    platforms=3;platformnames=Illumina,CG,10X;datasets=5;da
1       886006  rs4970375 T       C       50      PASS    platforms=3;platformnames=CG,Illumina,10X;datasets=5;da
1       886049  rs146799731       ACAG    A       50      PASS    platforms=3;platformnames=CG.Illumina.10X;datasets=5;da
```

Slika 4.1: Primer ulazne datoteke - *.vcf* datoteka za Aškenazi majku



Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene
1	565976	565976	C	T	intergenic	LOC101928626;MIR6723	dist=1587;dist=1729	.
1	566048	566048	G	A	intergenic	LOC101928626;MIR6723	dist=1659;dist=1657	.
1	567240	567240	G	-	downstream	MIR6723 dist=465	. . 1	50
1	568214	568214	C	T	upstream	MIR6723 dist=421	. . 0.5	50
1	568745	568745	-	CA	upstream	MIR6723 dist=952	. . 0.5	50
1	569933	569933	G	A	intergenic	MIR6723;OR4F16 dist=2140;dist=51163	. .	0.5
1	877715	877715	C	G	intronic	SAMD11 . .	1 50	732
1	877831	877831	T	C	exonic SAMD11	. nonsynonymous SNV	SAMD11:NM_152486:exon10:c.T1027	
1	879317	879317	C	T	exonic SAMD11	. synonymous SNV	SAMD11:NM_152486:exon14:c.C1830T:p.Y610	
1	879676	879676	G	A	UTR3 NOC2L;SAMD11	NM_015658:c.*398C>T;NM_152486:c.*143G>A	.	.
1	879687	879687	T	C	UTR3 NOC2L;SAMD11	NM_015658:c.*387A>G;NM_152486:c.*154T>C	.	.
1	880238	880238	A	G	intronic	NOC2L . .	1 50	1417
1	880390	880390	C	A	intronic	NOC2L . .	0.5 50	835
1	881627	881627	G	A	exonic NOC2L	. synonymous SNV	NOC2L:NM_015658:exon16:c.C1843T:p.L615L	
1	882803	882803	A	G	intronic	NOC2L . .	1 50	800
1	885366	885366	G	A	intronic	NOC2L . .	0.5 50	925
1	885676	885676	C	A	intronic	NOC2L . .	1 50	795
1	885689	885689	G	A	intronic	NOC2L . .	1 50	822
1	885699	885699	A	G	intronic	NOC2L . .	1 50	815
1	886006	886006	T	C	intronic	NOC2L . .	1 50	781
1	886050	886052	CAG	-	intronic	NOC2L . .	1 50	780
1	886183	886183	G	-	intronic	NOC2L . .	1 50	685

Slika 4.2: Primer izlazne datoteke - *.txt* datoteka za Aškenazi majku



Slika 4.3: Primer izlazne datoteke - *.vcf* datoteka za Aškenazi majku

Glava 5

Arhitekura i implementacija dodatka *AnnoPI*

Praktični deo ovog istraživanja je podrazumevao implementaciju dodatka za softver *Annovar* koji za rezultat ima pregledni prikaz iz *Annovar*-a obogaćenog dodatnim korisnim informacijama. Ulazni podaci su datoteke u *.vcf* formatu koje sadrže podatke o varijacijama na humanom genomu, a izlazni podatak su *.html* stranice koje prikazuju objedinjene informacije - deo podataka iz izlaznih datoteka *Annovar*-a, kao i dodatne informacije o funkciji i fenotipu gena koje se mogu naći u javno dostupnim ontologijama - *Gene Ontology* i *Human Phenotype Ontology*. Kako se one redovno ažuriraju, omogućeno je preuzimanje njihovih najsvežijih verzija zadavanjem odgovarajućih opcija pri pozivu aplikacije.

5.1 Datoteke

Ulazne *.vcf* datoteke koje su korišćene za potrebe rada se nalaze na stranici projekta za obezbeđivanje referentnih uzoraka i podataka "Genom u boci"[1]. Predmet analize su podaci Aškenazi otac-majka-sin trija iz projekta "Lični genom". Jedinstvene šifre podataka su sledeće:

- HG002-NA24385-huAA53E0 (sin)
- HG003-NA24149-hu6E4515 (otac)
- HG004-NA24143-hu8E87A9 (majka)

[1] https://jimb.stanford.edu/giab

Ulazne datoteke u formatu *.vcf* su preuzete sa naredne lokacije: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio. Za Aškenazi trio postoje različite verzije genoma, a u ovom istraživanju korišćena je verzija *NIST v3.3.2/GRCh37* koja je bila najsvežija na početku ovog istraživanja. Datoteke koje su korišćene su:

- HG002 https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-22_v.3.3.2_highconf_triophased.vcf)
- HG003 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-22_v.3.3.2_highconf_triophased.vcf
- HG004 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-22_v.3.3.2_highconf_triophased.vcf

Datoteke sa opisima ontologija se preuzimaju sa zvaničnih stranica *GO*/*HPO* ontologija:

- http://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/goa_human.gaf.gz
- https://ci.monarchinitiative.org/view/hpo/job/hpo.annotations/lastSuccessfulBuild/artifact/rare-diseases/util/annotation/genes_to_phenotype.txt

5.1.1 Struktura izlazne datoteke

Zadatak *AnnoPI* aplikacije je da na osnovu ulazne *.vcf* datoteke kreira izlaznu *.html* datoteku sa tabelom koja za svaku varijaciju iz ulazne datoteke sadrži red sa pojedinim informacijama iz ulazne datoteke i sa dodatnim, pridruženim informacijama o genu na kom se varijacija nalazi. Preciznije, izlazna *.html* datoteka sadrži sledeće kolone:

- `Chr`
  - hromozom

  - odogovara koloni `Chr` preuzetoj iz *Annovar*-a

- `Start`
  - početna pozicija varijacije na hromozomu
  - odgovara koloni `Start` preuzetoj iz *Annovar*-a
- `End`
  - krajnja pozicija varijacije na hromozomu
  - odgovara koloni `End` preuzetoj iz *Annovar*-a
- `Ref`
  - referentni nukleotid
  - odgovara koloni `Ref` preuzetoj iz *Annovar*-a
- `Alt`
  - posmatrani nukleotid
  - odgovara koloni `Alt` preuzetoj iz *Annovar*-a
- `Exonic Function`
  - vrsta varijacije u odnosu na to da li menja aminokiselinu ili ne (*synonymous SNV* ili *nonsynonymous SNV*)
  - odgovara koloni `ExonicFunc.refGene` preuzetoj iz *Annovar*-a
- `Gene`
  - naziv gena
  - odgovara koloni `Gene.refGene` preuzetoj iz *Annovar*-a
- `GO Associations`
  - prikazuje identifikatore *GO* funkcija pridruženih genu na kom se data varijacija nalazi
  - svaki *GO* identifikator predstavlja link na stranicu odgovarajuće funkcije[2]

[2]Na primer, za identifikator *GO:1990782* to je stranica https://www.ebi.ac.uk/QuickGO/term/GO:1990782

- `HPO Annotations`
  - prikazuje identifikatore *HPO* fenotipa pridruženih genu na kom se data varijacija nalazi
  - svaki *HPO* identifikator predstavlja link na stranicu odgovarajućeg fenotipa[3]
- `All GO`
  - sadrži logo ontologije *GO* i predstavlja link do stranice[4] koja pruža informaciju o svim *GO* funkcijama odgovarajućeg gena
- `All HPO`
  - sadrži logo ontologije *HPO* i predstavlja link do stranice[5] koja pruža informaciju o svim *HPO* fenotipovima odgovarajućeg gena na osnovu identifikatora gena

5.1.2 Dodatne informacije

Radi preglednosti sa jedne strane, a prikaza što više važnih informacija sa druge strane, obezbeđeno je da postavljanjem kursora na ime gena, *GO* ili *HPO* identifikator budu prikazane dodatne informacije (eng. *tooltips*) o njima. Informacije uključuju:

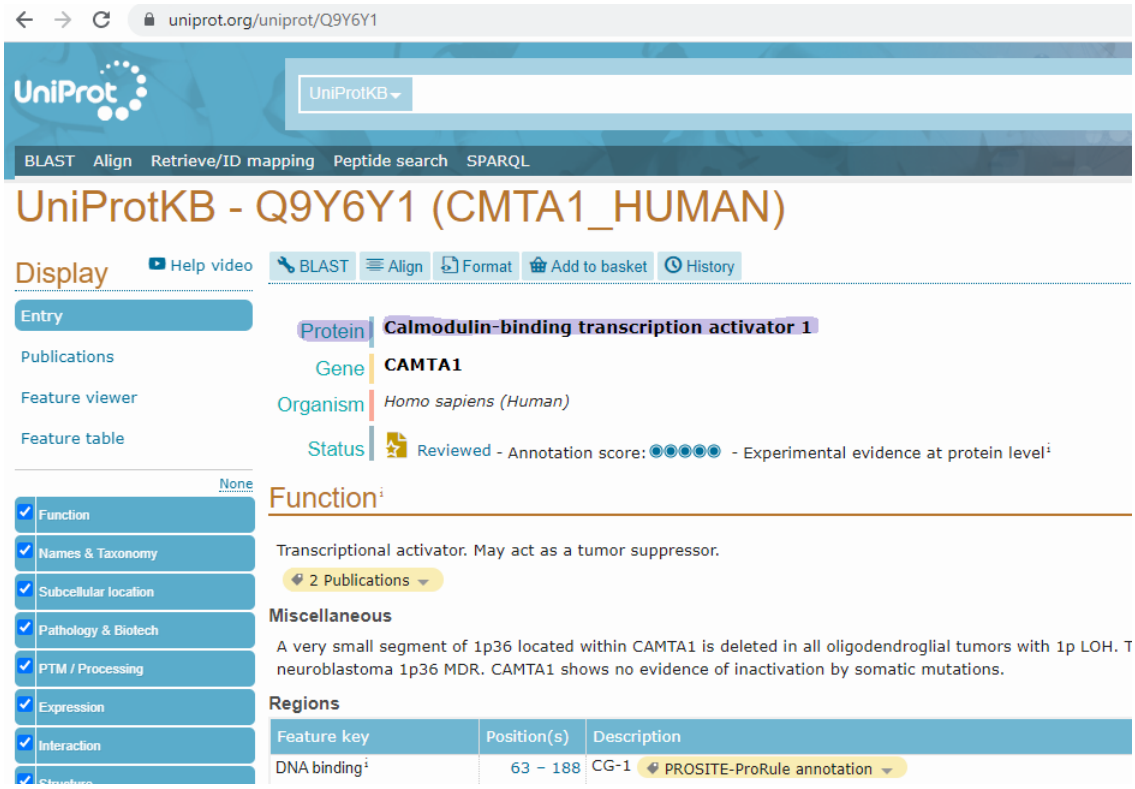
- opis gena koji se preuzima iz *Uniprot* baze na osnovu identifikatora gena koji nalazimo u datoteci sa opisom ontologije *GO*
- naziv i opis *GO* funkcije koji se nalaze na stranici na koju vodi link postavljen na identifikator *GO* funkcije
- naziv i opis *HPO* fenotipa koji se nalaze na stranici na koju vodi link postavljen na identifikator *HPO* fenotipa

[3]Na primer, za identifikator *HP:0100540* to je stranica https://hpo.jax.org/app/browse/term/HP:0100540

[4]Na primer, za gen *Q96NU1* to je stranica https://www.ebi.ac.uk/QuickGO/annotations?geneProductId=Q96NU1

[5]Na primer, gen *ISG15* ima identifikator *9636* (uparivanje naziva gena *ISG15* i identifikatora gena *9636* vrši se preko datoteke sa opisom ontologije *HPO*) i odgovarajuća stranica je https://hpo.jax.org/app/browse/gene/9636

Informacije potrebne za *tooltip*-ove za gen preuzimane su sa *Uniprot* stranice gena[6]. Informacije potrebne za *tooltip*-ove *GO* funkcija i *HPO* fenotipa dobijane su iz *JSON* reprezentacija njihovih veb stranica[7]. Nazivu *GO* funkcije odnosno *HPO* fenotipa odgovara vrednost polja *name*, dok se opis dobija na osnovu vrednosti polja *definition*. Na slici 5.1 možemo videti primer opisa gena, a na slikama 5.2 i 5.3 možemo videti primere pomenutih *JSON* reprezentacija kao i označena pomenuta polja i njihove vrednosti.



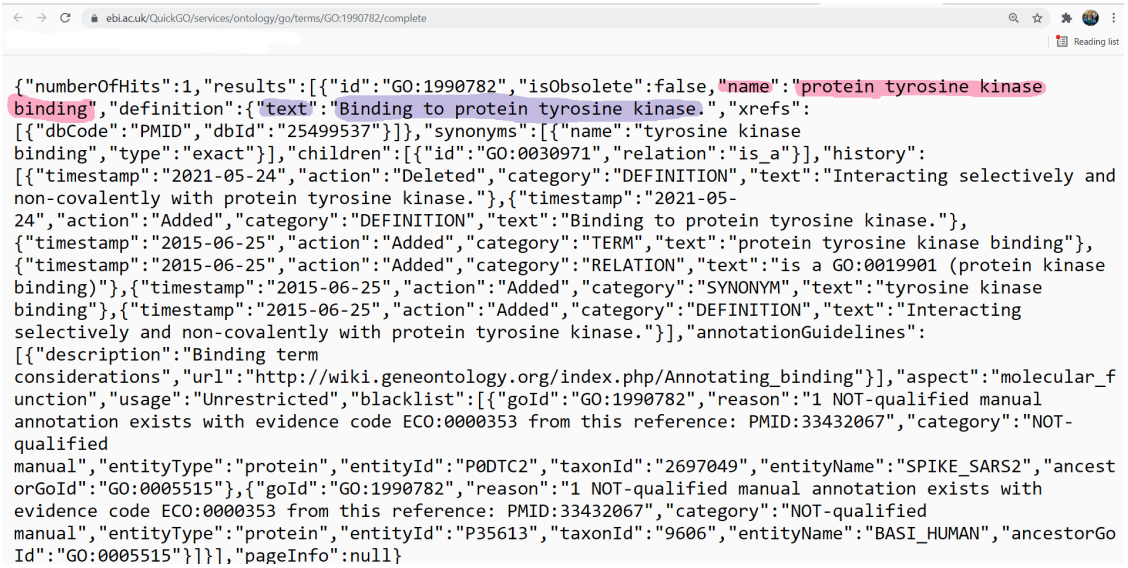
Slika 5.1: *Uniprot* stranica za gen *CAMTA1*

Prilikom testiranja, za svaki gen iz datoteke sa opisom ontologije *GO* opis gena sa *Uniprot* stranice smešten je u pomoćnu tekstualnu datoteku *Output-Gene*. U ovu datoteku smešteni su i opisi gena koji se pojavljuju u anotacijskoj datoteci, a za koje nemamo odgovarajući identifikator na osnovu koga preuzimamo opis sa *Uniprot* stranice. Za njihovu obradu iskorišćen je *Uniprot*-ov *ID Mapper*[8] koji je na osnovu

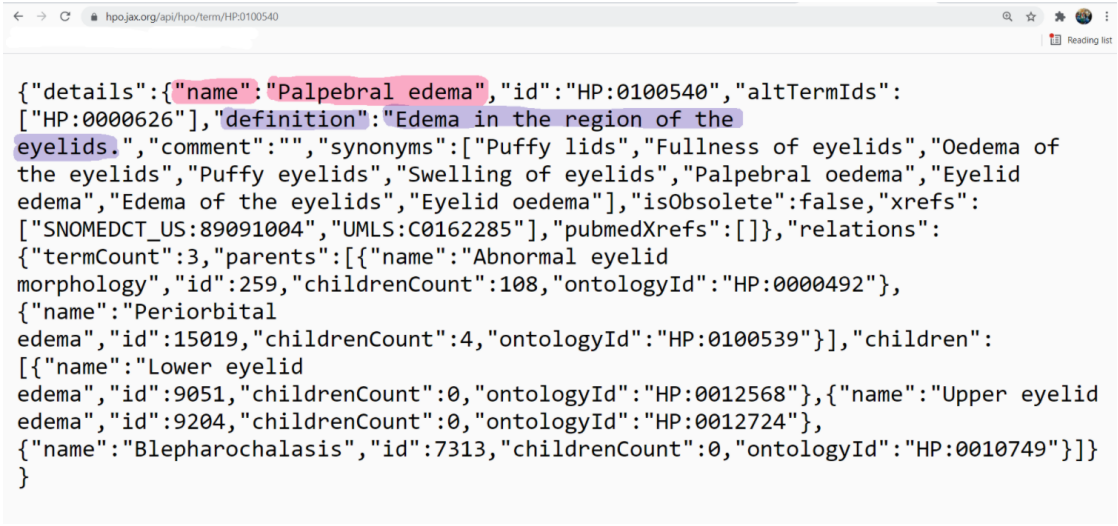
[6]Na primer, gen *CAMTA1* ima identifikator *Q9Y6Y1* i odgovarajuća stranica je https://www.uniprot.org/uniprot/Q9Y6Y1

[7]Podsetimo se za *GO* funkciju sa identifikatorom *GO:0005634*, stranica je https://www.ebi.ac.uk/QuickGO/term/GO:0005634, dok je za *HPO* fenotip sa identifikatorom *HP:0100540* stranica https://hpo.jax.org/app/browse/term/HP:0100540

[8]https://www.uniprot.org/uploadlists/



Slika 5.2: *JSON* reprezentacija za funkciju *GO:1990782*



Slika 5.3: *JSON* reprezentacija za fenotip *HP:0100540*

naziva gena dao informacije o identifikatorima gena pomoću kojih su dobijeni odgovarajući opisi. Za svaku *GO* funkciju, odnosno *HPO* fenotip iz polja kolona `GO Associations` i `HPO Annotations` (npr. *GO:005634*, odnosno *HP:0100540*) odgovarajuća *JSON* datoteka isparsirana i naziv i opis *GO* funkcije, odnosno *HPO* fenotipa su smešteni u pomoćnu tekstualnu datoteku *Output-GO Data*, odnosno *Output-HPO Data*. Nakon učitavanja podataka u ove datoteke, podaci se iz njih smeštaju u rečnik gde je ključ naziv gena odnosno *GO* ili *HPO* identifikator, a vrednost opis

gena odnosno naziv i opis funkcije ili fenotipa razdvojeni simbolom „|". Ove datoteke čuvaju svoj sadržaj između dva pokretanja programa, pa se tako prilikom novog pokretanja programa najpre kreira rečnik na osnovu postojećih pomoćnih datoteka[9], za svaki gen/*GO*/*HPO* identifikator iz datoteka sa opisima ontologija se proverava da li postoji kao ključ u rečniku, a ukoliko ne postoji čitaju se podaci sa odgovarajuće *Uniprot* stranice gena, odnosno parsira se odgovarajuća *JSON* datoteka i ažurira sadržaj datoteka i rečnika. U tabeli 5.1 možemo videti broj ključeva u rečniku, odnosno broj gena, *GO* i *HPO* identifikatora (kolone `Broj različitih gena`, `Broj različitih GO funkcija` i `Broj različitih HPO fenotipa`). Ukupan broj pojavljivanja svih identifikatora (sa ponavljanjima) u polaznim datotekama možemo videti u kolonama: `Ukupan broj gena (redova u .html tabeli)`, `Ukupan broj GO funkcija` i `Ukupan broj HPO fenotipa`.

Tabela 5.1: Informacije o ukupnom broju redova, *GO* funkcija, *HPO* fenotipa, broju različitih gena, *GO* funkcija i *HPO* fenotipa

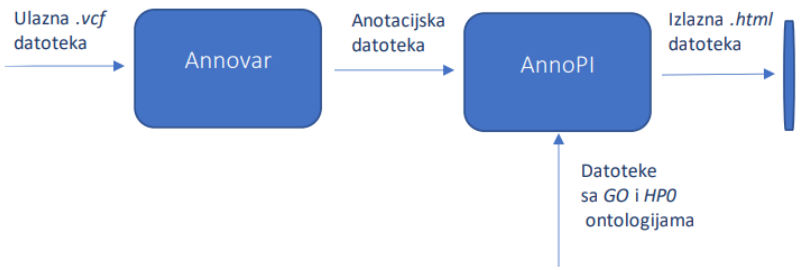
Član trija	Broj različitih gena	Ukupan broj gena (redova u *.html* tabeli)	Broj različitih *GO* funkcija	Ukupan broj *GO* funkcija	Broj različitih *HPO* fenotipa	Ukupan broj *HPO* fenotipa
son	8880	21042	127182	299308	6893	229285
mom	7877	17345	116245	255114	6717	199269
dad	7774	17342	115560	257247	6734	199572

[9]pomoćne datoteke su uvedene sa ciljem smanjenja ogromnog broja posećivanja pomenutih veb stranica

5.2 Princip rada aplikacije

Aplikacija je razvijana u programskom jeziku *Python* i pokreće se iz komandne linije. Primer jednog pokretanja programa je: `python master.py -g goUrl -h hpoUrl -d vcfInputFileName` gde `-g` predstavlja opcioni parametar koji zadajemo ako želimo da preuzmemo najnoviju datoteku sa opisom ontologije *GO* sa *URL* adrese `goUrl`, dok opcioni parametar `-h` zadajemo ukoliko želimo da preuzmemo najnoviju verziju ontologije *HPO* sa *URL* adrese `hpoUrl`. Parametar `-d` je obavezan, nakon koga se zadaje ime datoteke `vcfInputFileName` koja predstavlja ulaznu datoteku za softver *Annovar*. Aplikaciju je moguće pokrenuti i u *offline* režimu ne zadavši joj parametre i tada će aplikacija koristiti postojeće lokalne verzije datoteka sa ontologijama.

Grafički prikaz koraka aplikacije dat je na slici 5.4. Aplikacija se sastoji od sledećih koraka:



Slika 5.4: Grafički prikaz koraka aplikacije

1. Ukoliko su zadati opcioni parametri (`-g, -h`), aplikacija preuzima najsvežije datoteke sa opisima ontologija sa zadatih lokacija, dok u suprotnom koristi lokalne verzije datoteka. Datoteka koja predstavlja ontologiju *GO* se čuva pod nazivom *goa.human.gaf.gz*. Datoteka ontologije *HPO* se čuva pod nazivom *genes_to_phenotype.txt*

2. Vrši se obrada datoteka sa opisima ontologija (*goa.human.gaf* i *genes_to_phenotype.txt*) i podaci iz njih se smeštaju u rečnik gde je ključ ime gena, a vrednost lista uređenih torki *(goId, function, geneId)*, odnosno *(hpoId, phenotype)*.

3. Poziva se *Annovar* za *.vcf* datoteku čije ime unosi korisnik kao argument komandne linije. Ukoliko se radi o humanom genomu, *Annovar* se poziva sa opcijom `hg19`. Dobijena izlazna datoteka predstavlja anotacijsku *.txt* datoteku u kojoj svaki red sadrži informacije o jednoj genetskoj varijaciji odgovarajućeg genoma. Anotacijska datoteka predstavlja ulaznu datoteku za *AnnoPI*. Detaljan opis ove datoteke dat je u poglavlju 4.

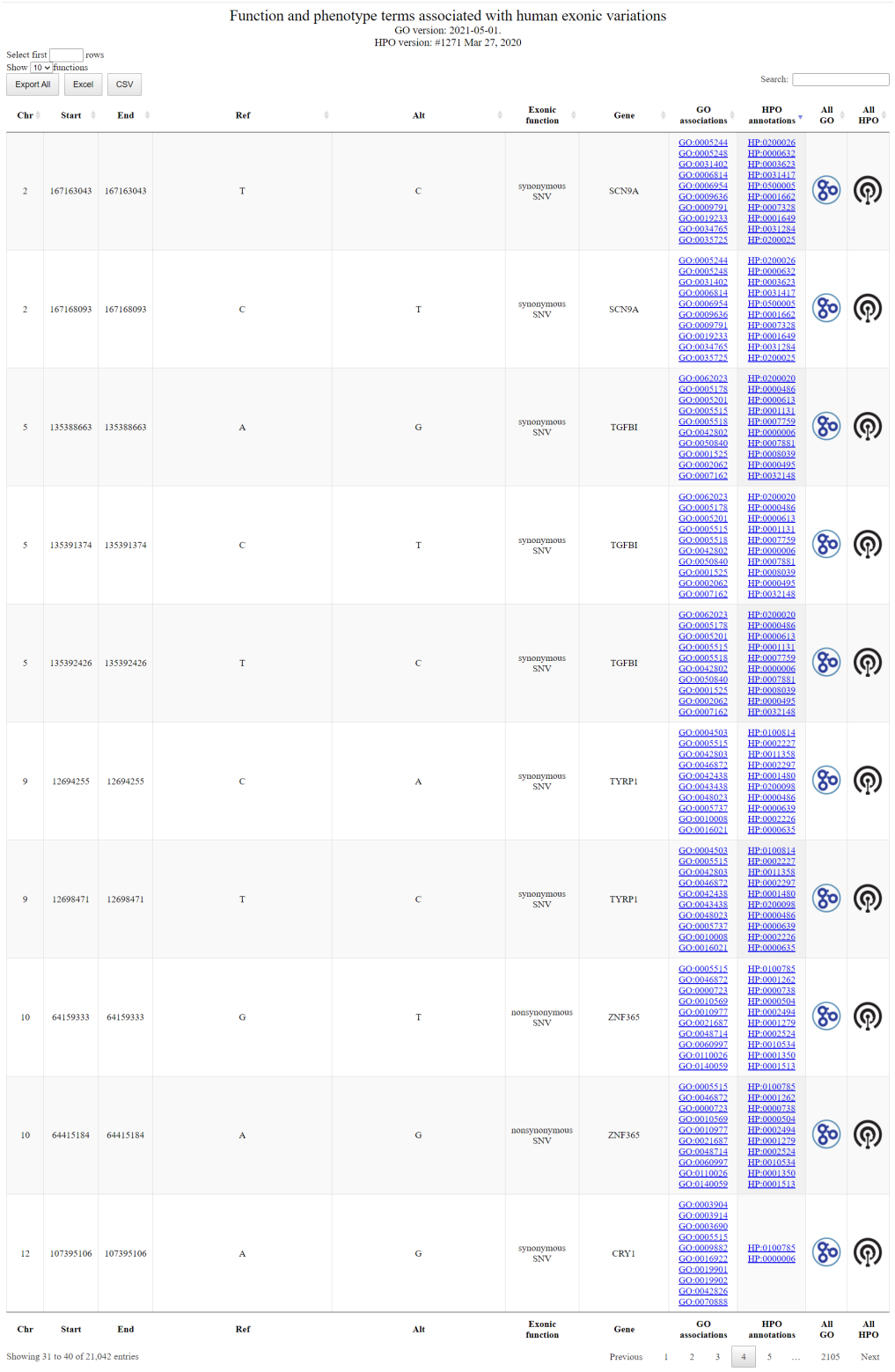
4. Iz anotacijske datoteke uzima se identifikator gena na kom se nalazi odgovarajuća varijacija. Na osnovu identifikatora gena moguće je doći do informacija o njegovoj funkciji i fenotipu koje *Annovar* ne obezbeđuje (kolone `GO Associations`, `HPO Annotations`, `All GO` i `All HPO`) preko veb stranica navedenih u opisu kolona na početku ovog potpoglavlja.

5. Iz anotacijske datoteke se uzimaju u obzir samo informacije o varijacijama koje se nalaze u egzonima, onih kod kojih je sadržaj kolone `Func.refGene` *Exonic*

6. Generiše se *.html* kod koji sadrži tabelu gde se u svakom redu nalaze informacije o varijacijama dobijene iz *Annovar*-a (kolone `Chr`, `Start`, `End`, `Ref`, `Alt`, `Exonic Function` i `Gene`) kao i pridružene informacije o funkciji i fenotipu gena (kolone `GO Associations`, `HPO Annotations`, `All GO` i `All HPO`).

Rad *AnnoPI* aplikacije je testiran za *.vcf* datoteke Aškenazi trija sin-majka-otac. Na slikama 5.5, 5.6 i 5.7 prikazane su izlazne *.html* datoteke za svakog člana trija. Dobijene tabele su veoma glomazne - broj različitih varijacija, a time i redova u tabeli, iznosi par desetina hiljada, preciznije za Aškenazi sina 21042, za Aškenazi majku 17345, a za Aškenazi oca 17342.

Aplikacija je pokretana i za *srb-hwe.vcf* datoteku koja je služila kao kontrolni mehanizam za proveru ispravnosti aplikacije. Time je pokazano da aplikaciju moguće koristiti za proizvoljnu *.vcf* datoteku. Na slici 5.8 dat je prikaz *.html* stranice koja je rezultat pokretanja aplikacije za *srb-hwe.vcf* datoteku. Prikaz sadržaja *tooltip*-a za *GO* funkciju dat je na slici 5.9.



Function and phenotype terms associated with human exonic variations
GO version: 2021-05-01.
HPO version: #1271 Mar 27, 2020

Select first rows
Show 10 functions
Export All | Excel | CSV
Search:

Chr	Start	End	Ref	Alt	Exonic function	Gene	GO associations	HPO annotations	All GO	All HPO
2	167163043	167163043	T	C	synonymous SNV	SCN9A	GO:0005244 GO:0005248 GO:0031402 GO:0006814 GO:0006954 GO:0009636 GO:0009791 GO:0019233 GO:0034765 GO:0035725	HP:0200026 HP:0000632 HP:0003623 HP:0031417 HP:0500005 HP:0001662 HP:0007328 HP:0001649 HP:0031284 HP:0200025		
2	167168093	167168093	C	T	synonymous SNV	SCN9A	GO:0005244 GO:0005248 GO:0031402 GO:0006814 GO:0006954 GO:0009636 GO:0009791 GO:0019233 GO:0034765 GO:0035725	HP:0200026 HP:0000632 HP:0003623 HP:0031417 HP:0500005 HP:0001662 HP:0007328 HP:0001649 HP:0031284 HP:0200025		
5	135388663	135388663	A	G	synonymous SNV	TGFBI	GO:0062023 GO:0005178 GO:0005201 GO:0005515 GO:0005518 GO:0042802 GO:0050840 GO:0001525 GO:0002062 GO:0007162	HP:0200020 HP:0000486 HP:0000613 HP:0001131 HP:0007759 HP:0000006 HP:0007881 HP:0008039 HP:0000495 HP:0032148		
5	135391374	135391374	C	T	synonymous SNV	TGFBI	GO:0062023 GO:0005178 GO:0005201 GO:0005515 GO:0005518 GO:0042802 GO:0050840 GO:0001525 GO:0002062 GO:0007162	HP:0200020 HP:0000486 HP:0000613 HP:0001131 HP:0007759 HP:0000006 HP:0007881 HP:0008039 HP:0000495 HP:0032148		
5	135392426	135392426	T	C	synonymous SNV	TGFBI	GO:0062023 GO:0005178 GO:0005201 GO:0005515 GO:0005518 GO:0042802 GO:0050840 GO:0001525 GO:0002062 GO:0007162	HP:0200020 HP:0000486 HP:0000613 HP:0001131 HP:0007759 HP:0000006 HP:0007881 HP:0008039 HP:0000495 HP:0032148		
9	12694255	12694255	C	A	synonymous SNV	TYRP1	GO:0004503 GO:0005515 GO:0042803 GO:0046872 GO:0042438 GO:0043438 GO:0048023 GO:0005737 GO:0010008 GO:0016021	HP:0100814 HP:0002227 HP:0011358 HP:0002297 HP:0001480 HP:0200098 HP:0000486 HP:0000639 HP:0002226 HP:0000635		
9	12698471	12698471	T	C	synonymous SNV	TYRP1	GO:0004503 GO:0005515 GO:0042803 GO:0046872 GO:0042438 GO:0043438 GO:0048023 GO:0005737 GO:0010008 GO:0016021	HP:0100814 HP:0002227 HP:0011358 HP:0002297 HP:0001480 HP:0200098 HP:0000486 HP:0000639 HP:0002226 HP:0000635		
10	64159333	64159333	G	T	nonsynonymous SNV	ZNF365	GO:0005515 GO:0046872 GO:0000723 GO:0010569 GO:0010977 GO:0021687 GO:0048714 GO:0060997 GO:0110026 GO:0140059	HP:0100785 HP:0001262 HP:0000738 HP:0000504 HP:0002494 HP:0001279 HP:0002524 HP:0010534 HP:0001350 HP:0001513		
10	64415184	64415184	A	G	nonsynonymous SNV	ZNF365	GO:0005515 GO:0046872 GO:0000723 GO:0010569 GO:0010977 GO:0021687 GO:0048714 GO:0060997 GO:0110026 GO:0140059	HP:0100785 HP:0001262 HP:0000738 HP:0000504 HP:0002494 HP:0001279 HP:0002524 HP:0010534 HP:0001350 HP:0001513		
12	107395106	107395106	A	G	synonymous SNV	CRY1	GO:0003904 GO:0003914 GO:0003690 GO:0005515 GO:0009882 GO:0016922 GO:0019901 GO:0019902 GO:0042826 GO:0070888	HP:0100785 HP:0000006		

Showing 31 to 40 of 21,042 entries
Previous 1 2 3 4 5 … 2105 Next

Slika 5.5: Prikaz *.html* stranice za Aškenazi sina

Function and phenotype terms associated with human exonic variations

GO version: 2021-05-01.
HPO version: #1271 Mar 27, 2020

Select first rows
Show 10 functions
Export All | Excel | CSV
Search:

Chr	Start	End	Ref	Alt	Exonic function	Gene	GO associations	HPO annotations	All GO	All HPO
12	40742254	40742254	G	A	synonymous SNV	LRRK2	GO:0034260 GO:0000149 GO:0000287 GO:0003779 GO:0003924 GO:0004672 GO:0004674 GO:0004708 GO:0005096 GO:0005515	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
12	40758652	40758652	T	C	nonsynonymous SNV	LRRK2	GO:0034260 GO:0000149 GO:0000287 GO:0003779 GO:0003924 GO:0004672 GO:0004674 GO:0004708 GO:0005096 GO:0005515	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
3	184037533	184037533	A	G	nonsynonymous SNV	EIF4G1	GO:0003723 GO:0003729 GO:0003743 GO:0005515 GO:0005524 GO:0008135 GO:0008190 GO:0031369 GO:0042802 GO:0060090	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
3	184039666	184039666	A	G	nonsynonymous SNV	EIF4G1	GO:0003723 GO:0003729 GO:0003743 GO:0005515 GO:0005524 GO:0008135 GO:0008190 GO:0031369 GO:0042802 GO:0060090	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
16	46696284	46696284	G	A	synonymous SNV	VPS35	GO:0005739 GO:0005515 GO:0031748 GO:0010628 GO:0010821 GO:0016055 GO:0016241 GO:0031647 GO:0032268 GO:0032456	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
2	233708806	233708806	A	G	synonymous SNV	GIGYF2	GO:0017148 GO:0003723 GO:0005515 GO:0045296 GO:0070064 GO:0007631 GO:0008344 GO:0009791 GO:0016441 GO:0021522	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
2	233712227	233712229	ACA	-	nonframeshift deletion	GIGYF2	GO:0017148 GO:0003723 GO:0005515 GO:0045296 GO:0070064 GO:0007631 GO:0008344 GO:0009791 GO:0016441 GO:0021522	HP:0100660 HP:0004926 HP:0100315 HP:0004409 HP:0012450 HP:0000338 HP:0002172 HP:0100753 HP:0100710 HP:0002120		
4	177605082	177605084	TCA	-	nonframeshift deletion	VEGFC	GO:0005515 GO:0042056 GO:0043185 GO:0002052 GO:0002576 GO:0006929 GO:0007165 GO:0008284 GO:0009887 GO:0016331	HP:0100658 HP:0003828 HP:0100797 HP:0000006 HP:0000962 HP:0000034 HP:0001004 HP:0000286 HP:0002619 HP:0200058		
10	70641860	70641860	T	C	nonsynonymous SNV	STOX1	GO:0000977 GO:0005515 GO:0007049 GO:0008284 GO:0010468 GO:0010628 GO:0010629 GO:0010800 GO:0010821 GO:0010971	HP:0100602 HP:0010982 HP:0100601 HP:0003259 HP:0002027 HP:0002910 HP:0001511 HP:0001873 HP:0002315 HP:0005117		
6	167549775	167549775	T	C	synonymous SNV	CCR6	GO:0004950 GO:0005515 GO:0016493 GO:0019957 GO:0038023 GO:0002407 GO:0002523 GO:0006935 GO:0006959 GO:0006968	HP:0100579 HP:0002960 HP:0100958 HP:0001053 HP:0002017 HP:0002206 HP:0009473 HP:0200042 HP:0002015 HP:0002020		

Showing 61 to 70 of 17,345 entries

Previous 1 … 6 7 8 … 1735 Next

Slika 5.6: Prikaz *.html* stranice za Aškenazi majku

Function and phenotype terms associated with human exonic variations

GO version: 2021-05-01.
HPO version: #1271 Mar 27, 2020

Select first rows
Show 10 functions
Export All | Excel | CSV
Search:

Chr	Start	End	Ref	Alt	Exonic function	Gene	GO associations	HPO annotations	All GO	All HPO
17	56435885	56435885	G	T	nonsynonymous SNV	RNF43	GO:0004842 GO:0005109 GO:0005515 GO:0046872 GO:0061630 GO:0006511 GO:0007275 GO:0016055 GO:0016567 GO:0030178	HP:0032222 HP:0000006 HP:0100808 HP:0002862 HP:0100574 HP:0005227 HP:0012189 HP:0003002 HP:0012125 HP:0100615		
17	56436109	56436109	C	T	nonsynonymous SNV	RNF43	GO:0004842 GO:0005109 GO:0005515 GO:0046872 GO:0061630 GO:0006511 GO:0007275 GO:0016055 GO:0016567 GO:0030178	HP:0032222 HP:0000006 HP:0100808 HP:0002862 HP:0100574 HP:0005227 HP:0012189 HP:0003002 HP:0012125 HP:0100615		
17	56448297	56448297	C	T	nonsynonymous SNV	RNF43	GO:0004842 GO:0005109 GO:0005515 GO:0046872 GO:0061630 GO:0006511 GO:0007275 GO:0016055 GO:0016567 GO:0030178	HP:0032222 HP:0000006 HP:0100808 HP:0002862 HP:0100574 HP:0005227 HP:0012189 HP:0003002 HP:0012125 HP:0100615		
17	56492800	56492800	T	C	nonsynonymous SNV	RNF43	GO:0004842 GO:0005109 GO:0005515 GO:0046872 GO:0061630 GO:0006511 GO:0007275 GO:0016055 GO:0016567 GO:0030178	HP:0032222 HP:0000006 HP:0100808 HP:0002862 HP:0100574 HP:0005227 HP:0012189 HP:0003002 HP:0012125 HP:0100615		
22	42159229	42159229	G	T	synonymous SNV	MEI1	GO:0007127	HP:0032192 HP:0000007 HP:0000789		
6	74072937	74072937	G	C	nonsynonymous SNV	KHDC3L	GO:0003723 GO:0005515 GO:0007015 GO:0008150 GO:0031297 GO:0032880 GO:0040019 GO:0043066 GO:0050769 GO:0051656	HP:0032192 HP:0000007		
6	74073531	74073531	C	G	nonsynonymous SNV	KHDC3L	GO:0003723 GO:0005515 GO:0007015 GO:0008150 GO:0031297 GO:0032880 GO:0040019 GO:0043066 GO:0050769 GO:0051656	HP:0032192 HP:0000007		
18	34310668	34310668	C	T	synonymous SNV	FHOD3	GO:0005515 GO:0030837 GO:0030018 GO:0051015 GO:0005737 GO:0051639 GO:0030866 GO:0005856 GO:0055003 GO:0045214	HP:0032092 HP:0012664 HP:0001645 HP:0005157 HP:0003581 HP:0000006 HP:0001297 HP:0005110 HP:0031656 HP:0031295		
2	207631461	207631461	G	A	nonsynonymous SNV	FASTKD2	GO:0003723 GO:0005515 GO:0019843 GO:0006396 GO:0006915 GO:0032543 GO:0044528 GO:0070131 GO:1902775 GO:0005739	HP:0031936 HP:0001250 HP:0001290 HP:0001350 HP:0001639 HP:0002059 HP:0000007 HP:0100660 HP:0002151 HP:0000639		
9	139904037	139904037	A	G	synonymous SNV	ABCA2	GO:0045540 GO:0010872 GO:0032384 GO:0032805 GO:0090370 GO:0150110 GO:1901873 GO:1902004 GO:1902993 GO:1905598	HP:0031936 HP:0000718 HP:0000007 HP:0001324 HP:0002286 HP:0001250 HP:0003141 HP:0002066 HP:0001260 HP:0002311		

Showing 131 to 140 of 17,342 entries

Previous 1 … 13 14 15 … 1735 Next

Slika 5.7: Prikaz *.html* stranice za Aškenazi oca

Function and phenotype terms associated with human exonic variations

GO version: 2021-05-01.
HPO version: #1271 Mar 27, 2020

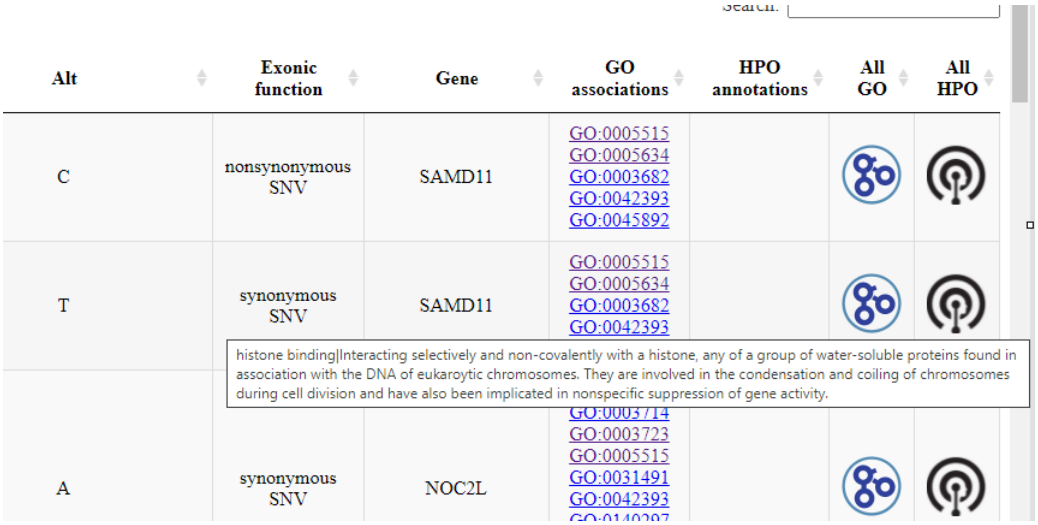
Select first rows
Show 10 functions
Export All | Excel | CSV
Search:

Chr	Start	End	Ref	Alt	Exonic function	Gene	GO associations	HPO annotations	All GO	All HPO
chr1	40307503	40307503	G	C	nonsynonymous SNV	TRIT1	GO:0003676 GO:0005524 GO:0008270 GO:0052381 GO:0070900 GO:0005739 GO:0005759 GO:0006400	HP:0001298 HP:0001332 HP:0003593 HP:0000545 HP:0003828 HP:0000252 HP:0002059 HP:0001257 HP:0001290 HP:0001344		
chr1	40310254	40310254	A	G	nonsynonymous SNV	TRIT1	GO:0003676 GO:0005524 GO:0008270 GO:0052381 GO:0070900 GO:0005739 GO:0005759 GO:0006400	HP:0001298 HP:0001332 HP:0003593 HP:0000545 HP:0003828 HP:0000252 HP:0002059 HP:0001257 HP:0001290 HP:0001344		
chr1	40310265	40310265	G	A	nonsynonymous SNV	TRIT1	GO:0003676 GO:0005524 GO:0008270 GO:0052381 GO:0070900 GO:0005739 GO:0005759 GO:0006400	HP:0001298 HP:0001332 HP:0003593 HP:0000545 HP:0003828 HP:0000252 HP:0002059 HP:0001257 HP:0001290 HP:0001344		
chr1	46818662	46818662	G	A	synonymous SNV	NSUN4	GO:0005515 GO:0008168 GO:0009383 GO:0019843 GO:0031167 GO:0005759 GO:0005762 GO:0001510			
chr1	55059646	55059648	AGG	-	nonframeshift deletion	ACOT11	GO:0008289 GO:0052689 GO:0102991 GO:0006631 GO:0006637 GO:0009266 GO:0009409 GO:0035556			
chr1	55075501	55075501	A	G	synonymous SNV	FAM151A	GO:0003674 GO:0008150 GO:0016020 GO:0016021 GO:0070062			
chr1	67206980	67206980	T	C	synonymous SNV	SGIP1	GO:0005515 GO:0005543 GO:0008017 GO:0017124 GO:0002021 GO:0048260 GO:0061024 GO:0097009 GO:2000253 GO:0005737			
chr1	94001891	94001891	G	A	synonymous SNV	FNBP1L	GO:0060271 GO:0005515 GO:0008289 GO:0045296 GO:0051020 GO:0006900 GO:0006914 GO:0007165 GO:0010324 GO:0016050			
chr1	94037295	94037295	C	T	nonsynonymous SNV	BCAR3	GO:0001784 GO:0005085 GO:0005515 GO:0019900 GO:0002089 GO:0007165 GO:0007264 GO:0008286 GO:0042493 GO:0043410			
chr1	94079421	94079421	A	T	nonsynonymous SNV	BCAR3	GO:0001784 GO:0005085 GO:0005515 GO:0019900 GO:0002089 GO:0007165 GO:0007264 GO:0008286 GO:0042493 GO:0043410			

Showing 31 to 40 of 727 entries

Previous 1 2 3 4 5 … 73 Next

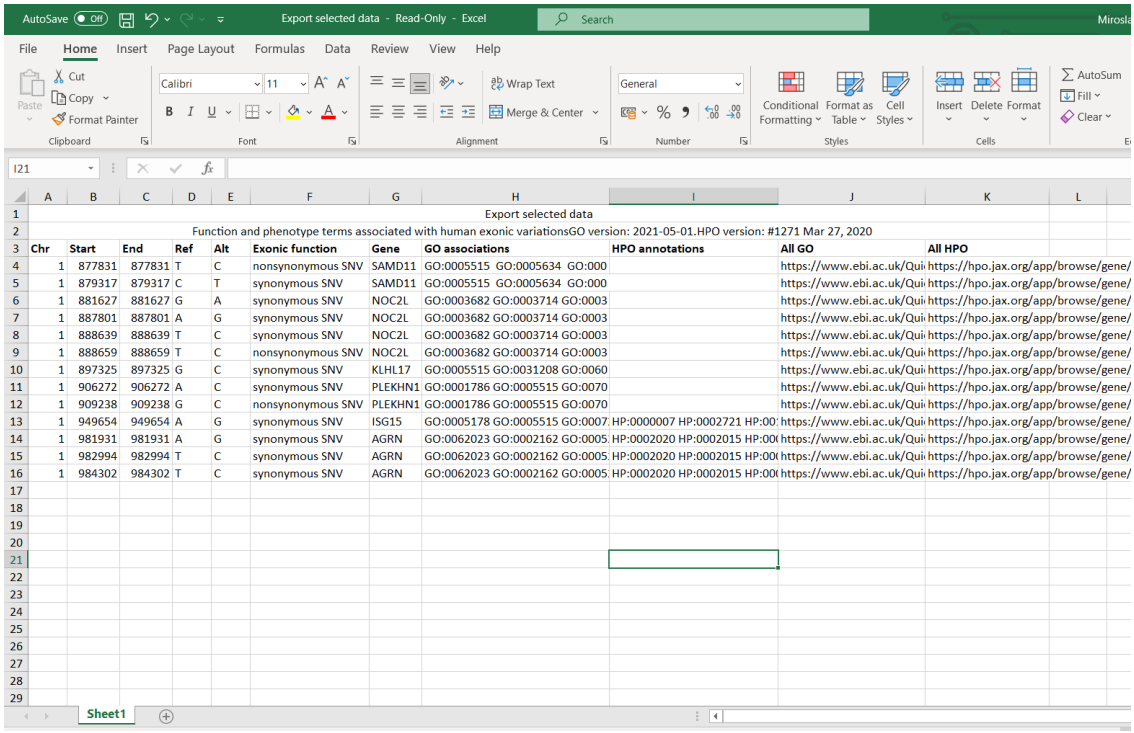
Slika 5.8: Prikaz *.html* stranice za *srb-hwe.vcf* datoteku



Slika 5.9: *GO tooltip* za funkciju *GO:0042393*

5.2.1 Funkcionalnosti izlazne *.html* datoteke

Iz izlazne *.html* datoteke je moguće izvesti podatke u *.xlsx* i *.csv* formatu. Omogućen je izvoz kompletne tabele kao i selektovanih redova. Izvezene datoteke imaju strukturu kao i tabele *.html* datoteke. Kolone `Chr`, `Start`, `End`, `Ref`, `Alt`, `Exonic function` i `Gene` imaju sadržaj isti kao i na odgovarajućoj *.html* stranici. Vrednosti kolona `GO associations` i `HPO annotations` razdvojene su razmakom. Kolone `All GO` i `All HPO` umesto ikonica sadrže linkove ka odgovarajućim stranicama koje pružaju informaciju o svim *GO* funkcijama, odnosno *HPO* fenotipu gena tekuće varijacije. Izvezena datoteka takođe sadrži i informaciju o verziji datoteka koje predstavljaju ontologije *GO* i *HPO*. Na slici 5.10 dat je primer izvezene datoteke u *.xlsx* formatu, dok je na slici 5.11 dat primer izvezene datoteke u *.csv* formatu.

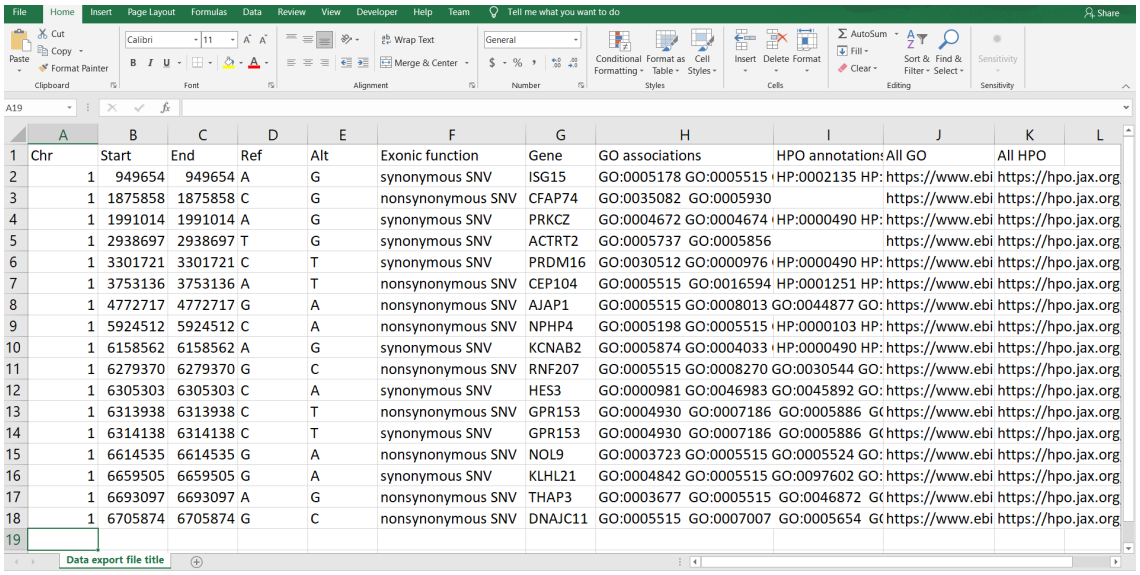


Export selected data

Function and phenotype terms associated with human exonic variationsGO version: 2021-05-01.HPO version: #1271 Mar 27, 2020

Chr	Start	End	Ref	Alt	Exonic function	Gene	GO associations	HPO annotations	All GO	All HPO
1	877831	877831	T	C	nonsynonymous SNV	SAMD11	GO:0005515 GO:0005634 GO:000		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	879317	879317	C	T	synonymous SNV	SAMD11	GO:0005515 GO:0005634 GO:000		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	881627	881627	G	A	synonymous SNV	NOC2L	GO:0003682 GO:0003714 GO:0003		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	887801	887801	A	G	synonymous SNV	NOC2L	GO:0003682 GO:0003714 GO:0003		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	888639	888639	T	C	synonymous SNV	NOC2L	GO:0003682 GO:0003714 GO:0003		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	888659	888659	T	C	nonsynonymous SNV	NOC2L	GO:0003682 GO:0003714 GO:0003		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	897325	897325	G	C	synonymous SNV	KLHL17	GO:0005515 GO:0031208 GO:0060		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	906272	906272	A	C	synonymous SNV	PLEKHN1	GO:0001786 GO:0005515 GO:0070		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	909238	909238	G	C	nonsynonymous SNV	PLEKHN1	GO:0001786 GO:0005515 GO:0070		https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	949654	949654	A	G	synonymous SNV	ISG15	GO:0005178 GO:0005515 GO:0007	HP:0000007 HP:0002721 HP:00	https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	981931	981931	A	G	synonymous SNV	AGRN	GO:0062023 GO:0002162 GO:0005	HP:0002020 HP:0002015 HP:00	https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	982994	982994	T	C	synonymous SNV	AGRN	GO:0062023 GO:0002162 GO:0005	HP:0002020 HP:0002015 HP:00	https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/
1	984302	984302	T	C	synonymous SNV	AGRN	GO:0062023 GO:0002162 GO:0005	HP:0002020 HP:0002015 HP:00	https://www.ebi.ac.uk/Qui	https://hpo.jax.org/app/browse/gene/

Slika 5.10: Prikaz *.xlsx* datoteke za Aškenazi sina



Chr	Start	End	Ref	Alt	Exonic function	Gene	GO associations	HPO annotations	All GO	All HPO
1	949654	949654	A	G	synonymous SNV	ISG15	GO:0005178 GO:0005515	HP:0002135 HP:	https://www.ebi	https://hpo.jax.org
1	1875858	1875858	C	G	nonsynonymous SNV	CFAP74	GO:0035082 GO:0005930		https://www.ebi	https://hpo.jax.org
1	1991014	1991014	A	G	synonymous SNV	PRKCZ	GO:0004672 GO:0004674	HP:0000490 HP:	https://www.ebi	https://hpo.jax.org
1	2938697	2938697	T	G	synonymous SNV	ACTRT2	GO:0005737 GO:0005856		https://www.ebi	https://hpo.jax.org
1	3301721	3301721	C	T	synonymous SNV	PRDM16	GO:0030512 GO:0000976	HP:0000490 HP:	https://www.ebi	https://hpo.jax.org
1	3753136	3753136	A	T	nonsynonymous SNV	CEP104	GO:0005515 GO:0016594	HP:0001251 HP:	https://www.ebi	https://hpo.jax.org
1	4772717	4772717	G	A	nonsynonymous SNV	AJAP1	GO:0005515 GO:0008013 GO:0044877	GO:	https://www.ebi	https://hpo.jax.org
1	5924512	5924512	C	A	nonsynonymous SNV	NPHP4	GO:0005198 GO:0005515	HP:0000103 HP:	https://www.ebi	https://hpo.jax.org
1	6158562	6158562	A	G	synonymous SNV	KCNAB2	GO:0005874 GO:0004033	HP:0000490 HP:	https://www.ebi	https://hpo.jax.org
1	6279370	6279370	G	C	nonsynonymous SNV	RNF207	GO:0005515 GO:0008270 GO:0030544	GO:	https://www.ebi	https://hpo.jax.org
1	6305303	6305303	C	A	synonymous SNV	HES3	GO:0000981 GO:0046983 GO:0045892	GO:	https://www.ebi	https://hpo.jax.org
1	6313938	6313938	C	T	nonsynonymous SNV	GPR153	GO:0004930 GO:0007186 GO:0005886	G(	https://www.ebi	https://hpo.jax.org
1	6314138	6314138	C	T	synonymous SNV	GPR153	GO:0004930 GO:0007186 GO:0005886	G(	https://www.ebi	https://hpo.jax.org
1	6614535	6614535	G	A	nonsynonymous SNV	NOL9	GO:0003723 GO:0005515 GO:0005524	GO:	https://www.ebi	https://hpo.jax.org
1	6659505	6659505	G	A	synonymous SNV	KLHL21	GO:0004842 GO:0005515 GO:0097602	GO:	https://www.ebi	https://hpo.jax.org
1	6693097	6693097	A	G	nonsynonymous SNV	THAP3	GO:0003677 GO:0005515 GO:0046872	G(	https://www.ebi	https://hpo.jax.org
1	6705874	6705874	G	C	nonsynonymous SNV	DNAJC11	GO:0005515 GO:0007007 GO:0005654	G(	https://www.ebi	https://hpo.jax.org

Slika 5.11: Prikaz *.csv* datoteke za Aškenazi oca

U kolonama `GO associations` i `HPO annotations` *.html* datoteka inicijalno je prikazano najviše pet funkcija. Za prikaz ostalih koristimo funkcionalnost *Show XY functions*. Odabirom vrednosti 10, 20, 30 ili *all* iz padajuće liste, određujemo koliko će se funkcija prikazati.

Stranica nudi pretraživanje tabele po ključnim rečima koje se navode u polju za pretragu. Valja pomenuti da pretraga uzima u obzir cele reči te za pojavljivanje reda sa *GO* funkcijom sa identifikatorom 0000007 moramo u polju za pretragu navesti *GO:0000007*. Kako sadržaj kolone `Exonic function` može biti *synonymous SNV* ili *nonsynonymous SNV*, bilo je potrebno uvesti ovakav princip pretrage zbog razlikovanja redova sa *synonymous SNV* i *nonsynonymous SNV* sadržajem. Na taj način, ukoliko bismo u polje za pretragu uneli *syn*, pojavili bi se samo oni redovi koji se odnose na *synonyomus SNV*.

Tabela pored zaglavlja ima i podnožje (eng. *footer*) koje odgovara kolonama iz zaglavlja. Ugrađena je mogućnost sortiranja po kolonama kao i straničenje. Inicijalno je postavljeno deset redova po strani. Korisnik ima mogućnost da odabere stranu unevši broj u polje u donjem desnom uglu.

5.2.2 Implementacija

Izlazna *.html* stranica je implementirana u *Python*-u, a podršku za interakciju sa korisnikom (pretraživanje, izvoz podataka, sortiranje, selektovanje, itd) je pružio *JavaScript*. Tabela predstavlja instancu *DataTable* objekta čiju podršku pruža *jQuery*[10] biblioteka *DataTables*[11]. Podršku za mogućnost selekcije redova pružila je biblioteka *Select*[12], a za izvoz podataka biblioteka *Buttons*[13].

Prikazivanje *tooltip*-ova u okviru izlazne *.html* stranice se pokazalo kao svojevrstan izazov. Naime, izlazne *.html* stranice za Aškenazi trio su i bez ovih dodatnih informacija bile glomazne (za Aškenazi sina/majku/oca, redom 66.9 MB/57.3 MB/57.6 MB), a njihovim pridruživanjem su još narasle i njihovo otvaranje u veb pregledačima je bilo značajno usporeno. Pristup ovakvim *.html* datotekama je testiran sa dva pregledača - *Google Chrome* i *Mozilla Firefox*. *Google Chrome* nije uspevao da otvori generisane datoteke, dok *Mozilla Firefox* jeste. U daljem testiranju stoga je korišćen samo *Mozilla Firefox*. U tabeli 5.2 prikazane su veličine

[10] https://ajax.googleapis.com/ajax/libs/jquery/3.4.1/jquery.min.js
[11] https://cdn.datatables.net/1.10.20/js/jquery.dataTables.js
[12] https://cdn.datatables.net/select/1.3.1/js/dataTables.select.min.js
[13] https://cdn.datatables.net/buttons/1.7.0/js/dataTables.buttons.min.js

izlaznih *.html* datoteka Aškenazi trija i vremena potrebna za njihovo otvaranje u različitim verzijama aplikacije *AnnoPI* koje su opisane u narednim pasusima.

U prvoj verziji aplikacije, tekst za *tooltip*-ove je pridružen svakom identifikatoru (i za gen, i za *GO* funkciju i za *HPO* fenotip) posebno, na nivou polja tabele pod tagom `td`, u okviru atributa *data-tippy-content*. Za njihovo prikazivanje korišćena je *JavaScript* biblioteka *Tippy*[14]. Dobijena izlazna *.html* datoteka je tada sadržala veliku količinu redundantnih informacija. Prilikom testiranja za Aškenazi trio, izlazne datoteke su bile veličina 179.5 MB, 154.1 MB i 154.8 MB (tabela 5.2, verzija 1). Korišćenje ove biblioteke rezultiralo je velikom količinom vremena pri otvaranju stranice, pa i nemogućnošću njenog otvaranja na računaru slabije konfiguracije.

U drugoj verziji aplikacije, umesto upisivanja teksta za *tooltip*-ove direktno u *.html* datoteku, ovaj tekst se upisuje u *JavaScript* datoteke *GeneJSMap.js*, *GOtooltipsJS.js* i *HPOtooltipsJS.js* (svaka od datoteka je veličine par megabajta). Za prikazivanje *tooltip*-ova i u ovoj verziji koristimo biblioteku *Tippy* pri čemu se vrednost *data-tippy-content* atributa postavlja u *scriptFile.js* datoteci čitanjem podataka iz odgovarajućih *.js* datoteka. Svi *tooltip*-ovi smeštaju se u rečnik, konkretno u odgovarajuću instancu *Map* objekta (*geneMap, goMap* i *hpoMap*). Izlazne *.html* stranice su u drugoj verziji manje nego u prvoj (66.5 MB, 56.7 MB, 57.3 MB, tabela 5.2, verzija 2) ali učitavanje i dalje traje dugo. Ono što je prednost ovog pristupa je izostanak velikog broja ponavljanja istih funkcija odnosno fenotipskih osobina.

U trećoj verziji, *data-tippy-content* u okviru taga *td* zamenjen je *HTML* atributom *title* u okviru istog taga. Njegove vrednosti se postavljaju u datoteci *scriptFile.js* čitanjem iz datoteka *GeneJSMap.js*, *GOtooltipsJS.js* i *HPOtooltipsJS.js* pri čemu nema potrebe za korišćenjem *JavaScript* biblioteke *Tippy*.

Treća verzija je rezultovala najmanjim datotekama koje mogu da se otvore i na računarima slabije konfiguracije. Postoje male razlike u veličini u odnosu na drugu verziju, ali su razlike u vremenu prikazivanja velike. Razlog za to leži u činjenici da se u trećoj verziji koristi *title* atribut koji nije neophodno inicijalizovati u *.html* datoteci za razliku od atributa *data-tippy-content* koji se mora inicijalizovati. Dodatno, *tooltip* prikazan pomoću atributa *data-tippy-content* zahtevniji je jer korišćenje *Tippy* biblioteke znači i kreiranje instance objekta u *.js* datoteci. Atribut *title* je ugrađeni *HTML* atribut i za njegovo korišćenje dovoljno je samo u *.js* datoteci postaviti vrednost odgovarajućem elementu.

[14]Za njeno korišćenje potrebno je uključiti `https://unpkg.com/@popperjs/core@2` i `https://unpkg.com/tippy.js@6`

Tabela 5.2: Vremena otvaranja *.html* datoteka u različitim verzijama aplikacije na računaru slabije konfiguracije (4 GB RAM memorije, Intel(R) Core(TM) i3-5005U @ 2.00 GHz procesor) i na računaru jače konfiguracije (16 GB RAM memorije, Intel(R) Core(TM) i7-9750H @2.60 GHz procesorom)

Verzija	Ime	Veličina datoteke	Vreme otvaranja na računaru slabije konfiguracije	Vreme otvaranja na računaru jače konfiguracije
1	son	179 MB	neuspešno otvaranje	97 s
	mom	155 MB	neuspešno otvaranje	91 s
	dad	154 MB	neuspešno otvaranje	90 s
2	son	66.5 MB	neuspešno otvaranje	98 s
	mom	56.7 MB	neuspešno otvaranje	91 s
	dad	57.3 MB	neuspešno otvaranje	91 s
3	son	65.4 MB	122 s	34 s
	mom	55.8 MB	120 s	32 s
	dad	56.2 MB	120 s	32 s

AnnoPI je softver koji se jednostavno instalira i koristi. Pored softvera *AnnoPI*, neophodno je da korisnik na lokalnom računaru ima softver *Annovar* kao i interpretere za *Perl* i *Python*. Pokretanje softvera *Annovar* i generisanje anotacijskih datoteka (*.txt* datoteka koje predstavljaju ulaz za *AnnoPI*) može da potraje na slabijim računarima, do sat vremena (4GB RAM memorije). Nakon generisanja anotacijskih datoteka, softver *AnnoPI* kreira izlaznu datoteku za svega nekoliko minuta na računaru slabije konfiguracije. Pretpostavka je da bi izvršavanje bilo značajno brže na računarima jače konfiguracije što zbog nedostatka vremena nije provereno.

Glava 6

Zaključak

U savremenim bioinformatičkim istraživanjima, sekvencioniranje nove generacije dovelo je do svakodnevnog priliva sekvencioniranih genoma i velikih količina podataka o njima. Posebno su značajni podaci o genetskim varijacima pojedinačnih genoma jer genetske varijacije mogu biti uzrok mnogih anomalija i kod coveka i drugih vrsta.

Podaci o genetskim varijacijama se nalaze rasuti po raznim javno dostupnim bazama podataka. Softver *Annovar* automatski prikuplja određene podatke iz pomenutih baza i time olakšava i ubrzava dolazak do njih. Cilj ovog rada je izrada dodatka *AnnoPI* za softver *Annovar* koji sakuplja dodatne podatke o genetskim varijacijama, preciznije podatke o funkciji i fenotipu gena na kojima se genetske varijacije nalaze. *AnnoPI* preuzima ove podatke iz javno dostupnih baza podataka i zajedno sa podacima koje *Annovar* prikuplja pregledno ih prikazuje u okviru *.html* datoteke sa različitim funkcionalnostima (sortiranje, selekcija, izvoz podataka, pretraga). Softver *AnnoPI* je javno dostupan i može da pruži podršku istraživačima različitih profila u genomici kao i u raznim bioinformatičkim disciplinama pri pretraživanju informacija o genetskim varijacijama.

Jedan pravac unapređenja softvera *AnnoPI* bi bila serverska obrada podataka što bi se ogledalo u učitavanju podataka *.html* datoteke „strana-po-strana", kao i elegantnijem prikazu *tooltipa*-ova. Uz to, s obzirom da se softver pokreće iz komandne linije, implementacija grafičkog interfejsa bi doprinela njegovoj široj upotrebi među istraživačima bez informatičke ekspertize.

Literatura

[1] Céline Brouard. „Inférence de réseaux d'interaction protéine-protéine par apprentissage statistique". Doktorska teza. Feb. 2013.

[2] *GitHub repozitorijum projekta AnnoPI.* URL: https://github.com/AndjelaMijailovic/AnnoPI.

[3] Xiaofeng Gong et al. "A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology". In: *The Sixteenth Asia Pacific Bioinformatics Conference* 15 (2018), pp. 1–2.

[4] Radivoje Papović i saradnici Medicinskog fakulteta u Beogradu. *Humana genetika.* Medicinski fakultet, Beograd, 2007.

[5] Board on Environmental Studies National Research Council Commission on Life Sciences i Committee on Developmental Toxicology Toxicology. *Scientific Frontiers in Developmental Toxicology and Risk Assessment.* National Academies Press (US), 2000.

[6] Philip Compeau Pavel A. Pevzner. *Bioinformatics algorithms: An Active Learning Approach.* English. 2015.

[7] Wing-Kin Sung. *Algorithms for Next-Generation Sequencing.* CRC Press, 2017.

[8] *Uputstvo za korišćenje softvera Annovar.* URL: https://annovar.openbioinformatics.org/en/latest/user-guide/startup/.

[9] *Zvanična prezentacija Nacionalnog istraživačkog instituta za humani genom (National Human Genome Research Institute, NHGRI).* URL: https://www.genome.gov/.

[10] *Zvanična stranica Švajcarskog bioinformatičkog instituta.* URL: https://www.sib.swiss/.

[11] *Zvanična stranica Evropskog bioinformatičkog instituta*. URL: https://www.ebi.ac.uk/.

[12] *Zvanična stranica Gene Ontology*. URL: http://geneontology.org/docs/ontology-documentation/.

[13] *Zvanična stranica GO projekta*. URL: http://current.geneontology.org/annotations/index.html.

[14] *Zvanična stranica Human Phenotype Ontology*. URL: https://hpo.jax.org/app/help/introduction.

[15] *Zvanična stranica Kjoto enciklopedije gena i genoma*. URL: https://www.genome.jp/kegg/.

[16] *Zvanična stranica konzorcijuma Genom u boci*. URL: https://www.nist.gov/programs-projects/genome-bottle/.

[17] *Zvanična stranica Nacionalnog centra za biološke informacije*. URL: https://www.ncbi.nlm.nih.gov/.

[18] *Zvanična stranica softvera Annovar*.
URL: https://annovar.openbioinformatics.org/en/latest/.

Biografija autora

Anđela Mijailović (*Kraljevo, 21. oktobar 1994.*) je završila Računarstvo i informatiku na Matematičkom fakultetu 2013. godine. Iste godine nastavlja master studije...