Univerza v Mariboru

Fakulteta za kemijo
in kemijsko tehnologijo

**Katja Repnik, Uroš Potočnik**

# BIOINFORMATIKA in GENOMIKA

Navodila za računalniške vaje z nalogami

Maribor, 2013/2014

# KAZALO

# 1 ISKANJE PODATKOV V GENETIKI

Ko načrtujemo genetsko raziskavo npr. asociacijsko analizo, se pri tem, tako kot tudi pri ostalih znanostih, naslanjamo na znanje predhodnih študij. Danes so najpomembnejši vir informacij strokovni članki objavljeni v znanstvenih revijah in prispevkih na konferencah, saj nam le ti podajajo podatke o najnovejših odkritjih, ki jih v učbenikih še ni moč zaslediti. Ker pa se vsak dan objavi ogromna količina člankov, je brez bioinformatskih orodij in spletnih podatkovnih zbirk praktično nemogoče izmed vseh dostopnih informacij izbrati tiste, ki jih potrebujemo.

Pri preiskovanju genov oz. polimorfizmov, ki so vključeni v kompleksne bolezni, so se izkazale kot zelo uspešna metoda asociacijske študije. Z njimi skušamo najti povezavo med polimorfizmi in boleznijo pri posameznikih, ki niso v sorodu, tako da primerjamo frekvence alelov med skupino bolnikov in kontrolno skupino zdravih posameznikov (t.j. študija primeri:kontrole). V primerjavi z analizo genetske vezave, ki lahko zajema velike genomske regije, so bile asociacijske študije doslej praviloma omejene le na proučevanje nekaj kandidatnih genov. Napredek tehnologije v zadnjih letih omogoča hkratno proučevanje velikega števila SNP-jev oziroma asociacijske študije na celotnem genomu (genome wide associacion – GWA). Pri GWA študijah hkrati analiziramo povezavo določene kompleksne bolezni z nekaj deset tisoč polimorfizmi. Vendar so ugotovljeni označevalci za posamezno bolezen pogosto samo v neravnotežju vezave z odseki genoma, ki so dejansko etiološko vpleteni v nastanek bolezni. Za natančnejšo identifikacijo gena in polimorfizma, ki je vpleten v patogenezo posamezne kompleksne bolezni, so še vedno potrebne asociacijske študije kandidatnih genov, ki pa so s pomočjo GWA študij tudi lahko usmerjene v ožja področja genoma.

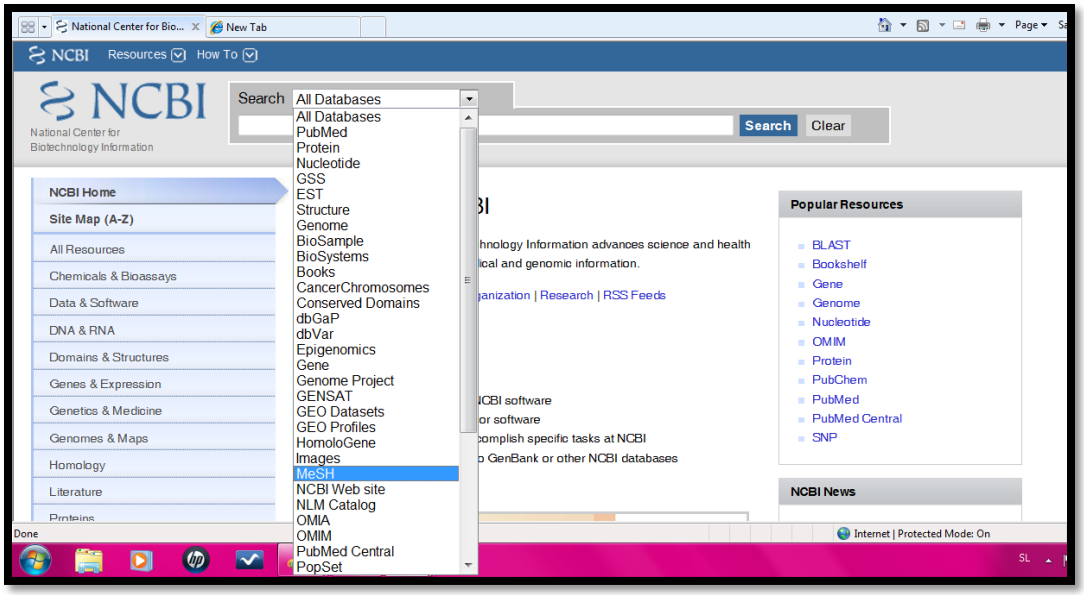
## 1.1 Medical Subject Headings

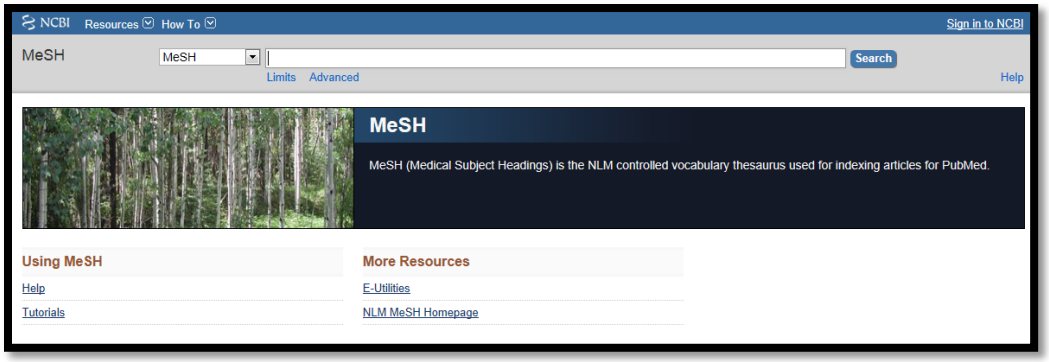
Kadar želimo iskati informacije v medicinskih znanostih je najuporabnejša uporaba tezavra MeSH (Medical Subject Hedings).

**Tezaver** imenujemo seznam vsebinskih konceptov in navodil za njihovo uporabo. Koncepti v tezavru so povezani z relacijami.

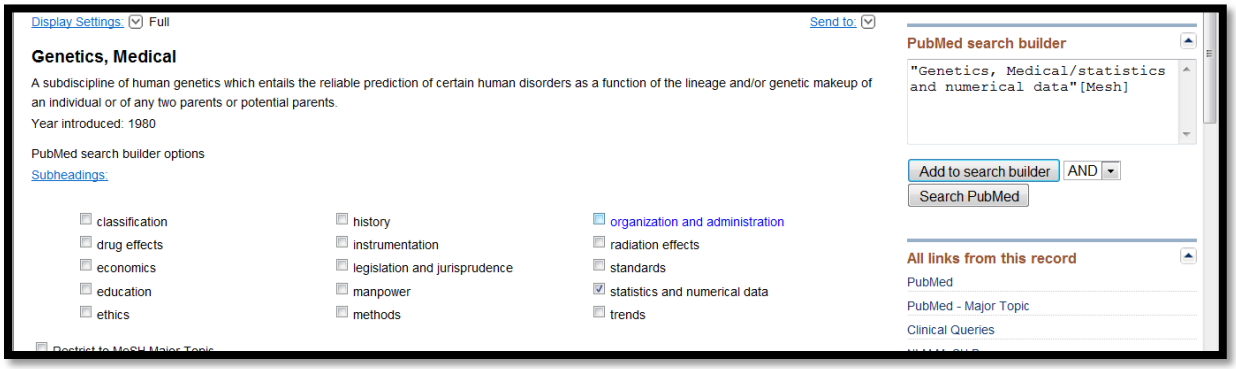
Vsebina dokumenta je opisana z **deskriptorji** – to so izrazi, ki jih določi indekser in pri tem skuša uganiti izraze, ki bi jih uporabil iskalec, če bi hotel poiskati dani dokument. **Kvalifikatorji** pa podrobneje omejijo vsebinski obseg deskriptorja.

Možnih je več različnih pristopov k uporabi tezavra MeSH. Do njega lahko dostopamo na strani Nacional center for Biotehnology informations – NCBI:

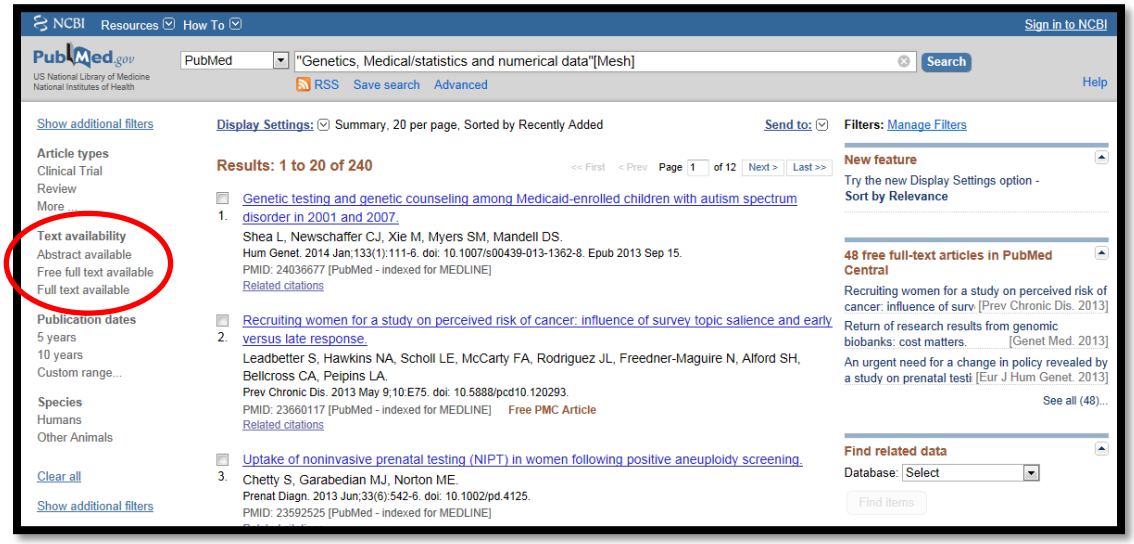



Najprej z uporabo podatkovne baze MeSH poiščemo deskriptor »Genetics, Medical«, nato pa izberemo še kvalifikator, ki nas zanima oz. omejuje področje iskanja npr. »statistical and numerical data«. S klikom na gumb »Add to search builder« dodamo izbrane deskriptorje in kvalifikatorje ter izberemo »Search PubMed«.



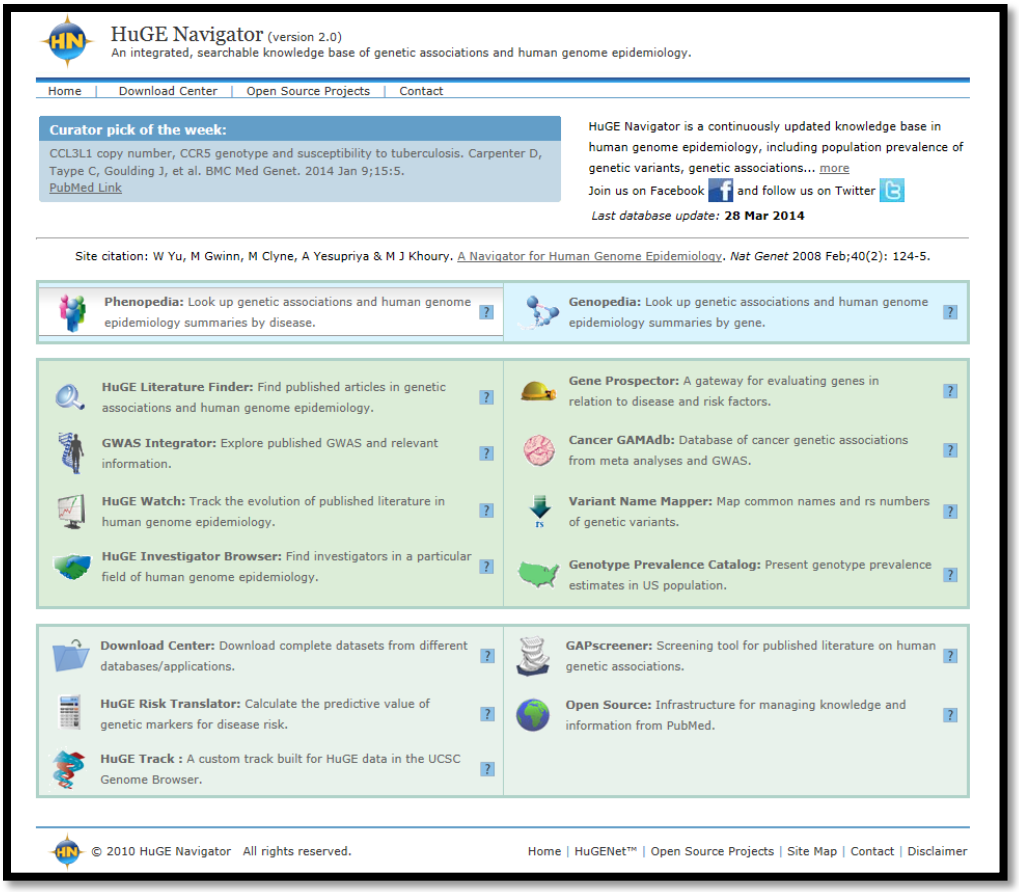
V tem primeru dobimo kot rezultat 240 člankov (april 2014 – z dodajanjem člankov v bazo pub med se to število seveda spreminja). levi strani lahko izberemo prikaz samo prosto dostopnih člankov ali prikaz preglednih člankov.




Iz tako velikega števila člankov je seveda zelo težko izbrati tiste, ki bodo za naše študije najuporabnejši. Zato lahko za iskanje uporabimo več deskriptorjev, ki jih povežemo z logičnimi operatorji AND, OR oz. NOT.

## 1.2 Huge Navigator

Medtem, ko lahko z uporabo tezavra MeSH učinkovito iščemo informacije in študije o točno določeni temi, ki nas zanima, pa obstajajo tudi drugi iskalniki, ki so primerni za iskanje informacij na ožjem področju. Tako je npr. Huge Navigator (http://hugenavigator.net) vmesnik, ki podatke črpa iz podatkovne zbirke PubMed in je namenjen zgolj iskanju asociacijskih študij. Kot izhodišče lahko uporabi gene (Genopedia) ali bolezen (Phenopedia).



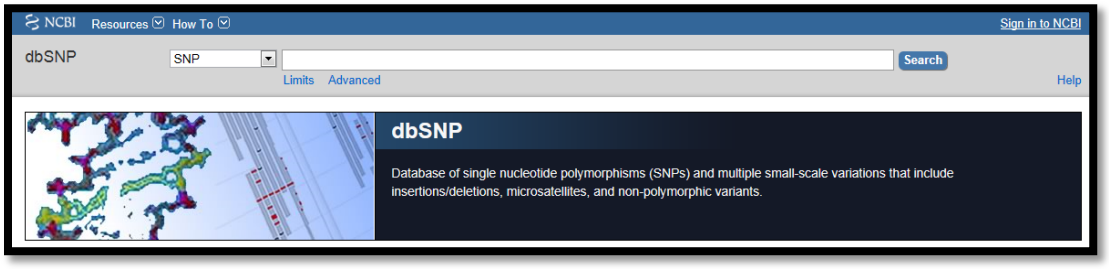

Iskalnik je pregleden in uporabniku prijazen. Z uporabo HugeNavigator z lahkoto poiščemo vse študije, ki so preučevale povezavo določenega gena in bolezni. Ugotovimo, koliko genov je do sedaj že bilo povezanih z določeno boleznijo, koliko je bilo narejenih GWA študij, koliko metaanaliz, s katerimi boleznimi so naši predhodniki že povezali določen gen...

***Naloga 1:*** *S pomočjo Huge-navigator-ja poišči, kateri geni se največkrat omenjajo v študijah glede bolezni, ki si jo raziskoval v nalogi 1. Uporabi Phenopedio! Izpiši 3 gene, ki se ponovijo v največ študijah in število le teh. Napiši tudi podatek, koliko GWA študij je bilo narejenih v zvezi z izbrano boleznijo! Nato izberi en gen in poišči 3 SNP-je, za katere je nekdo že potrdil povezavo. Zapiši tudi referenco. Za ta isti gen na Genopedii poišči bolezn,i s katerim so ga poleg izbrane še povezali.*

## 1.3 National Center for Biotechnology Information

NCBI (National Ceneter for Biotechnology Information) je ena najpomembnejših podatkovnih zbirk in iskalnikov po človeškem genomu in genomu modelnih organizmov. Podobna iskalnika sta tudi Ensebmel, ki je skupen projekt Evropskega bioinformacijskega inštituta in iskalnik Univerze v Californiji.

### 1.3.1 NCBI SNP



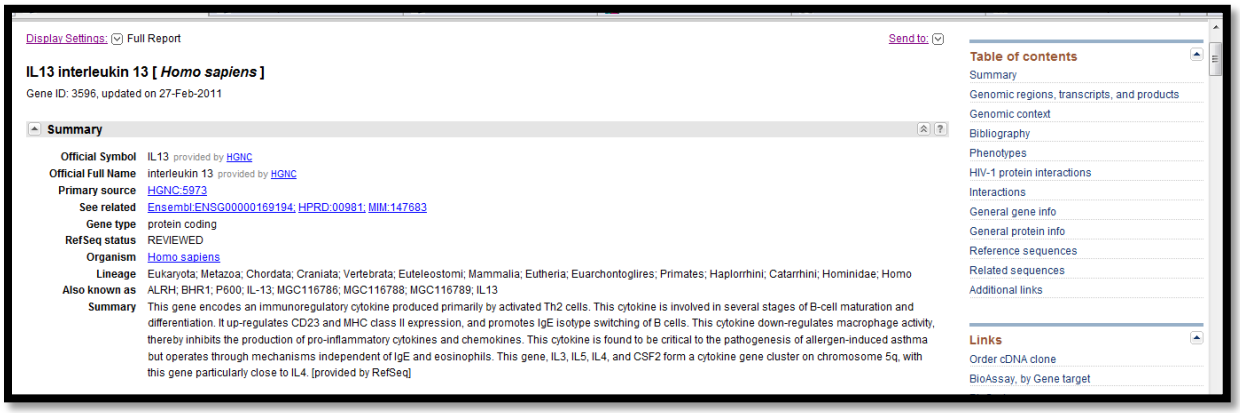

Spremembe v sekvenci so prisotne na določenih položajih znotraj genoma in so odgovorne za posamezne fenotipske lastnosti, vključno z nagnjenostjo posameznika k razvoju kompleksnih bolezni, kot so rak in bolezni srca. Podatkovna zbirka dbSNP je javna domena s široko zbirko preprostih genetskih polimorfizmov. Ta zbirka polimorfizmov vključuje polimorfizme posameznega nukleotida (SNP ang. za single nucleotide polymorphism), večbazne delecije ali insercije, retropozone ter mikrosatelitne ponovitve. V bazi najdemo podatke o sekvenci okoli polimorfizma, frekvenco polimorfizma v različnih populacijah ter lokacijo za vse vnešene variatne. Variatne so podane za različne organizme iz kateregakoli dela genoma.

***Naloga 3:*** *Za vse 3 izbrane SNP-je iz prejšnjih nalog v SNP database poišči fasta sekvence. Za vsak SNP izpiši alelne in genotipske frekvence v evropski populaciji. Izpiši tudi lokacijo. Ali se SNP nahaja na genu ali izven njega? Če se nahaja na genu izpiši na katerem in poišči ali se nahaja v intronu ali eksonu?*

***Naloga 4:*** *Za enega izmed izbranih SNP-jev prikaži toliko obsežno sliko (bp), da bodo na njej vidni 3 geni pred in trije za izbranim polimorfizmom. Odčitaj lokacijo polimorfizma iz slike. Pod sliko klikni na povezavo – Table wiev (Genes on sequence) in izpiši lokacije genov – začetek in konec. Izračunaj oddaljenost SNPja od vsakega gena.*

### 1.3.2 NCBI Gene


Baza Gene daje podrobne informacije za znane in predvidene gene, ki so definirani z nukleotidno sekvenco ali položajem. Baza vsebuje več kot 14 milijonov vpisov in vključuje podatke od vseh večjih taksonomskih skupin. Vsak zapis v bazi ustreza enemu genu. Zapis o posameznem genu vključuje nomenklaturo, referenčno sekvenco (RefSeq), lokacijo, poti, variacije, fenotipe in povezave do genotip-fenotip in lokus-specifične vire po vsem svetu.




***Naloga 5:*** *Za gen, na katerem se SNP nahaja oz. za najbližji gen izpiši podatke, ki jih najdeš v podatkovni zbirki NCBI Gene (lokacijo, funkcijo, procese v katerih nastopa, izooblike, fasta sekvenco).*

# 2 ZEMLJEVID KROMOSOMSKEGA PODROČJA

Izdelaj zemljevid kromosomskega področja (tabelarično in lahko tudi grafični prikaz v obliki premice), kjer so prikazani najpomembnejši lokusi iz GWA študij, povezani z izbrano boleznijo. Vključi vse SNP-je s $p<10^{-8}$. Vključi vse gene, ki se nahajajo na razdalji 250 kbp od posameznega SNPja.

Izračunaj in grafično prikaži LD za posamezen lokus.

Tabela (premica) naj ima sledeče stolpce:

- Ime SNPja
- Ime gena
- Kromosom na katerem leži
- Začetna lokacija
- Končna lokacija
- Lokacija SNPja glede na gen (ekson, intron, izven gena)
- Polno ime gena
- Funkcija gena
- Število študij, ki so preučevale ta gen z isto boleznijo
- Ostale bolezni, povezane z istim genom
- Število GWA študij
- Število meta-analiz

Premico pripravi v obliki excelove datoteke. Poimenuj jo: Priimek_premica_bolezen_2012

Excelovo tabelo (premico) in tudi povzetek študije pošlji na moj e-mail: katja.repnik@um.si.

# 3 HAPMAP, HAPLOVIEW


Prelomnica v razvoju GWA študij je bil projekt HapMap. V projekt HapMap je vključena genotipizacija 1 milijona SNP-jev človeškega genoma v prvi fazi in 3,1 milijona SNP-jev v drugi fazi projekta. Na spletni strani HapMap-a (www.hapmap.org) lahko dostopamo do podatkov o frekvencah SNP-jev in njihovih korelacij skozi vezavno neravnovesje (LD).

Cilj mednarodnega projekta HapMap je razviti haplotipski zemljevid človeškega genoma. HapMap opisuje skupne vzorce variacij človeške DNK sekvence. HapMap pričakuje, da bo ključni vir za raziskovalce, ki podatke projekta uporabljajo za raziskovanje genov, ki vplivajo na zdravje in bolezen, odzive na zdravila ter dejavnike okolja. Informacije, ki jih nudi projekt HapMap, so na voljo brezplačno na njihovi spletni strani (www.hapmap.org) (International HapMap Project, 2010).

**HapMap faze:**

Faze HapMap projekta:
• Faza 1 – sekvenčni podatki 4 populacij CEU (S Evropejci), YRI (Yoruban, Nigerija), CHB (Kitajska, Beijing), JPT (Tokijo, Japan). 269 posameznikov.
#SNP = 1 milijon, prib. 1 na 5 kb preko genoma, MAF > 0.05
• Faza 2 – več sekvenciranja na zgoraj omenjenih populacijah.
#SNP = 3.1 milijon več, prib. 1 kb preko genoma MAF > 0.05
• Faza 3 – več sekvenciranja na zgoraj omenjenih populacijah in sekvenciranje na 7 novih populacijah
#SNP = 1.6 miljonov in več

**Populacije v projektu HapMap:**

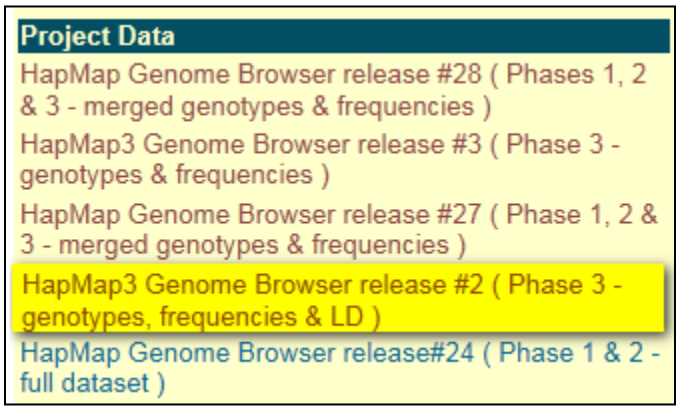
- ASW* Potomci afričanov v SZ ZDA
- CEU* Prebivalci Utaha, potomci S in Z evropejcev
- CHB Han Kitajci Beijing, Kitajska
- CHD Kitajci v Denverju, Colorado
- GIH Gujarati Indijanci v Houston, Texas
- JPT Japonci v Tokiju, Japonska
- LWK Luhya v Webuya, Kenija
- MEX* Mehikanski predniki v Los Angeles, Kalifornia

- MKK* Maasai v Kinyawa, Kenija
- TSI  Toscanci v Italiji
- YRI* Yoruba v Ibadan, Nigerija

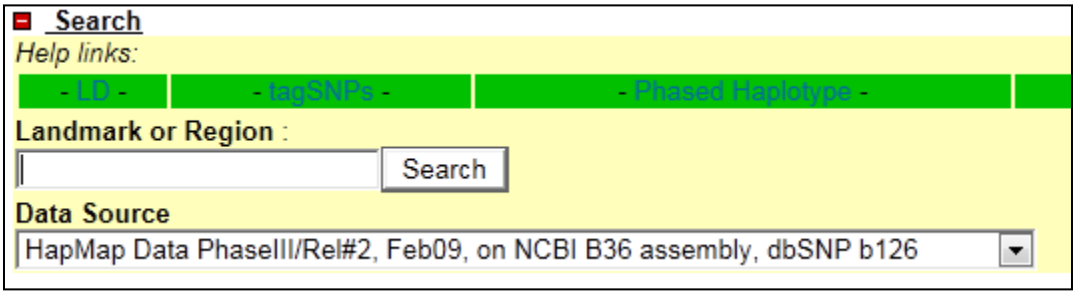
** populacija je iz družinskih trojk (m, ž, otrok)*

**Primer:** Išči podatke za SNP: rs2211792. Imenuj alele za ta SNP v glavni verigi glede na referenčno zaporedje. V katerem genu se SNP nahaja? Ali je polimorfen v vseh populacijah? Ali spada med redke v CEU populaciji? ***Če je alelna frekvenca > 0.20, se klasificira kot splošni, drugače kot redek.***

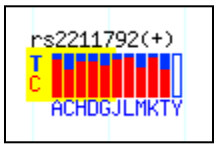
V brskalniku odpri spletno stran HapMap URL: http://hapmap.ncbi.nlm.nih.gov/index.html.en;
Kliknemo na podatkovno zbirko: HapMap3 Genome Browser r#2;

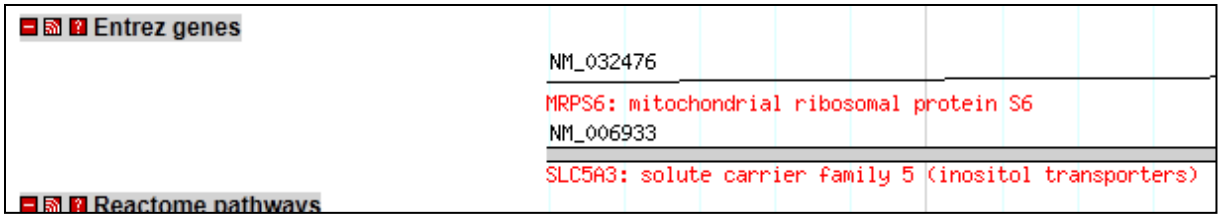



V iskalno okence vpiši ID SNP-ja in poženi iskalnik;




Z drsnikom se pomakni navzdol in si izpiši podatke, ki jih zahteva naloga 1;






Ponovno kliknemo na grafikon alelnih frekvenc (slika zgoraj) populacij in navigiramo na alelne in genotipske frekvence (slika spodaj):

| Population | Genotype frequencies | | | | | | | | | | Allele frequencies | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | genotype | freq | count | genotype | freq | count | genotype | freq | count | Total | Ref-allele allele | Ref-allele freq | Ref-allele count | Other-allele allele | Other-allele freq | Other-allele count | Total |
| ASW (A) | T/T | 0 | 0 | C/T | 0.200 | 10 | C/C | 0.800 | 40 | 50 | T | 0.100 | 10 | C | 0.900 | 90 | 100 |
| CEU (C) | T/T | 0.071 | 8 | C/T | 0.536 | 60 | C/C | 0.393 | 44 | 112 | T | 0.339 | 76 | C | 0.661 | 148 | 224 |
| CHB (H) | T/T | 0 | 0 | C/T | 0.325 | 27 | C/C | 0.675 | 56 | 83 | T | 0.163 | 27 | C | 0.837 | 139 | 166 |
| CHD (D) | T/T | 0.048 | 4 | C/T | 0.262 | 22 | C/C | 0.690 | 58 | 84 | T | 0.179 | 30 | C | 0.821 | 138 | 168 |
| GIH (G) | T/T | 0.080 | 7 | C/T | 0.322 | 28 | C/C | 0.598 | 52 | 87 | T | 0.241 | 42 | C | 0.759 | 132 | 174 |
| JPT (J) | T/T | 0.024 | 2 | C/T | 0.181 | 15 | C/C | 0.795 | 66 | 83 | T | 0.114 | 19 | C | 0.886 | 147 | 166 |
| LWK (L) | T/T | 0 | 0 | C/T | 0.057 | 5 | C/C | 0.943 | 82 | 87 | T | 0.029 | 5 | C | 0.971 | 169 | 174 |
| MEX (M) | T/T | 0.062 | 3 | C/T | 0.542 | 26 | C/C | 0.396 | 19 | 48 | T | 0.333 | 32 | C | 0.667 | 64 | 96 |
| MKK (K) | T/T | 0.007 | 1 | C/T | 0.130 | 18 | C/C | 0.862 | 119 | 138 | T | 0.072 | 20 | C | 0.928 | 256 | 276 |
| TSI (T) | T/T | 0.125 | 11 | C/T | 0.420 | 37 | C/C | 0.455 | 40 | 88 | T | 0.335 | 59 | C | 0.665 | 117 | 176 |
| YRI (Y) | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |

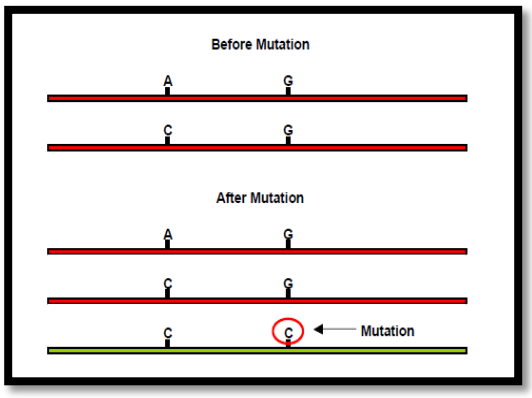
Note: the 'reference' allele is the base observed in the reference genome sequence at this location

***Naloga 6:*** *Išči podatke za SNP: rs12777960 in en poljuben SNP. Imenuj alele za ta SNP v glavni verigi glede na referenčno zaporedje. V katerem genu se SNP nahaja? Ali je polimorfen v vseh populacijah? Ali spada med redke v JPT populaciji?*

## 3.1 Linkage disequilibrium – vezavno neravnovesje

- **Vezavno ravnovesje** (ang. linkage equlibrium): genotipi na določenem lokusu (ali za določen SNP) se pojavljajo neodvisno od genotipov na drugem lokusu (za drug SNP)
- **Vezavno neravnovesje** (ang. linkage disequlibrium): genotipi na dveh lokusih se ne pojavljajo neodvisno drug od drugega

Polimorfizmi, ki danes obstajajo, so nastali z naključnimi mutacijami. Najprej se pojavi eden (slika levo), nato pa še drugi (slika desno):

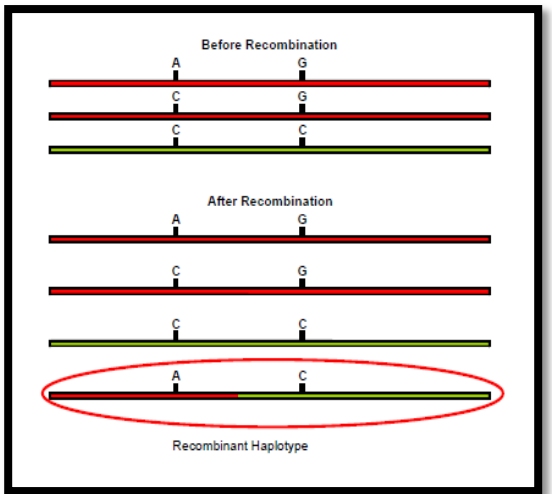




Kombinacijo dveh ali več SNP-jev imenujemo haplotip.

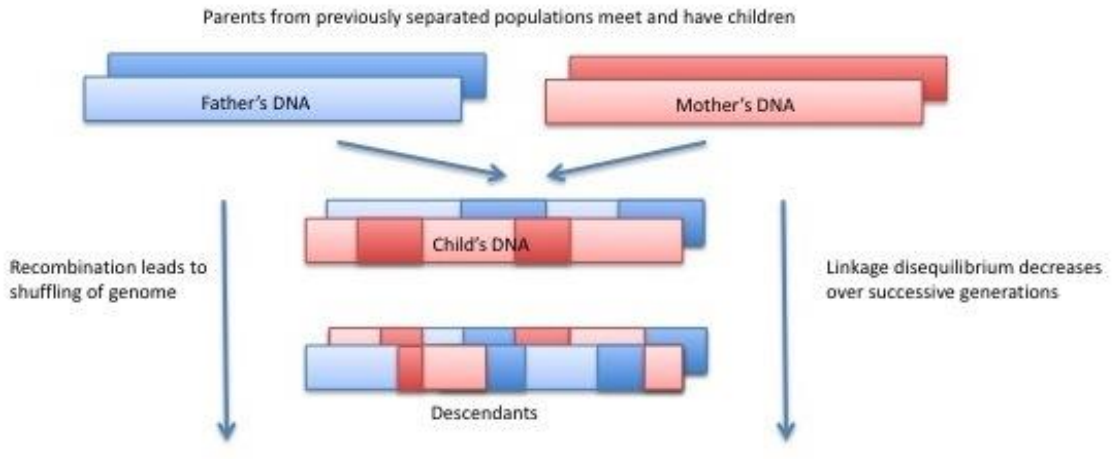
Recumo da je s prvo mutacijo nastal SNP1, ki ima možna alela A in C, z drugo pa SNP2, ki ima možna alela G in C. Druga mutacija se je zgodila na eni molekuli DNA, ki je za SNP1 vsebovala alel C. Zato so po tej mutaciji, gledano na kombinacijo teh dveh SNP-jev v naravi mogoči trije različni haplotipi.

→ Haplotip 1 (SNP1 = A; SNP2 = G)
→ Haplotip 2 (SNP1 = C; SNP2 = G)
→ Haplotip 3 (SNP1 = C; SNP2 = C)

Vendar pa se med celično delitvijo kromosomi oz. obe molekuli DNA v celici prekrižajo (rekombinacija) – novonastala rekombinantna DNA lahko vsebuje novo kombinacijo alelov – v tem primeru lahko dobimo → Haplotip 4 (SNP1 = A; SNP2 = C).



Če sta dva SNP-ja na molekuli DNA zelo oddaljena, je možnost, da se po rekombinaciji še vedno nahajata na isti molekuli 50 %. Kadar pa sta locirana v bližini, je zelo verjetno, da do rekombinacije med njima ne bo prišlo in se bosta skupaj dedovala. Temu pravimo, da sta v vezavnem neravnovesju (LD).




Vezavno neravnovesje (LD) se meri z različnimi parametri, kot sta Lewontinov D' in indeks $r^2$. Lewontinov D' se imenuje tudi asociacijska verjetnost in predstavlja pomembno mero za identifikacijo regij, v katerih je bilo malo rekombinacij. Kadar je vrednost D' = 1, lahko govorimo o popolnem neravnovesju. Indeks $r^2$ nam pokaže moč posrednih asociacij in lahko varira, kljub temu da je D' = 1. Vrednost indeksa $r^2$ je povezana z alelnimi frekvencami in s pozicijo korespondenčne mutacije.

D’ lahko zavzema vrednosti med 1 in -1, pomembno |D'|;

- D’ = 1 pomeni dogodek **brez** rekombinacije;
- vrednost < 1 pomeni rekombinacijo;
- vmesne vrednosti za D’ je težko interpretirati (vrednosti se dvignejo, ko je vzorec majhen ali alelne frekvence nizke);

Uporabljamo tudi kvadrirani korelacijski koeficient $\mathbf{r^2}$. Dobimo ga z deljenjem **D’** z alelnimi frekvencami na obeh lokusih [**D’/(p1p2qp1q2)**]:

- $\mathbf{r^2=1}$ – ni rekombinacije med označevalci, alelne frekvence so enake.

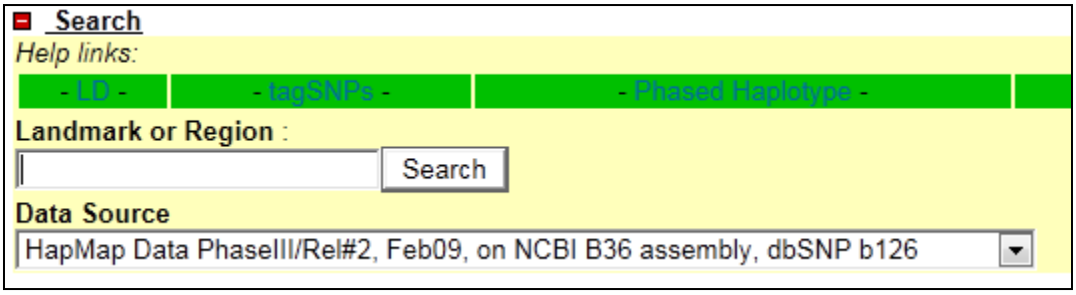
***Alelne frekvence SNP-jev se lahko razlikujejo med različnimi populacijami. LD se lahko razlikuje med populacijami. Rezultati genotipizacije ene populacije (ne) predstavljajo druge.***

## 3.2 Programsko orodje HaploView

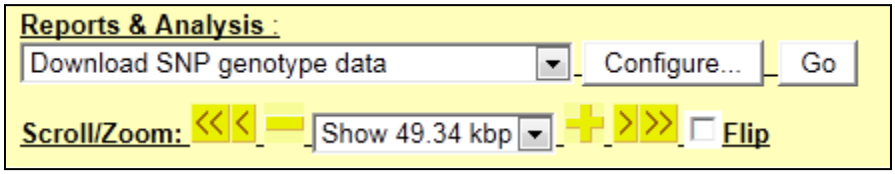
Programsko orodje HaploView uporabljamo za računanje vezavnega neravnovesja med posameznimi SNP-ji. Orodje je prosto dostopno na URL:
http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview, BROAD inštituta.

**Primer:** Polimorfizmi v LCT genu so povezani z motnjami presnavljanja laktoze. Za načrtovanje SNP testov v CEU populaciji najdi najmanjše število SNP-jev, ki bodo pokrili vse haplotipe. ***SNP-ji, ki pokrivajo haplotipske bloke imenujemo TAG SNP-ji.***

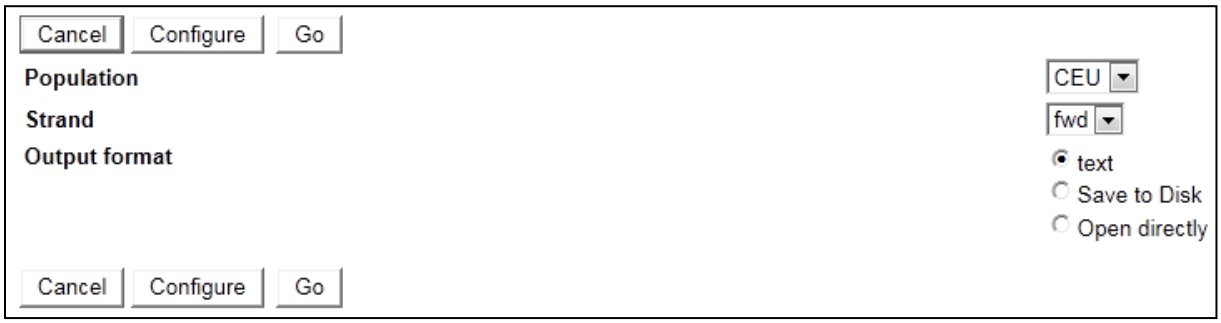
Ponovno v iskalno okence vpišemo ime ali oznako gena in poženemo iskalnik;




Pomaknemo se desno in v drsnem okencu nastavimo Download SNP gfenotype data in kliknemo configure;



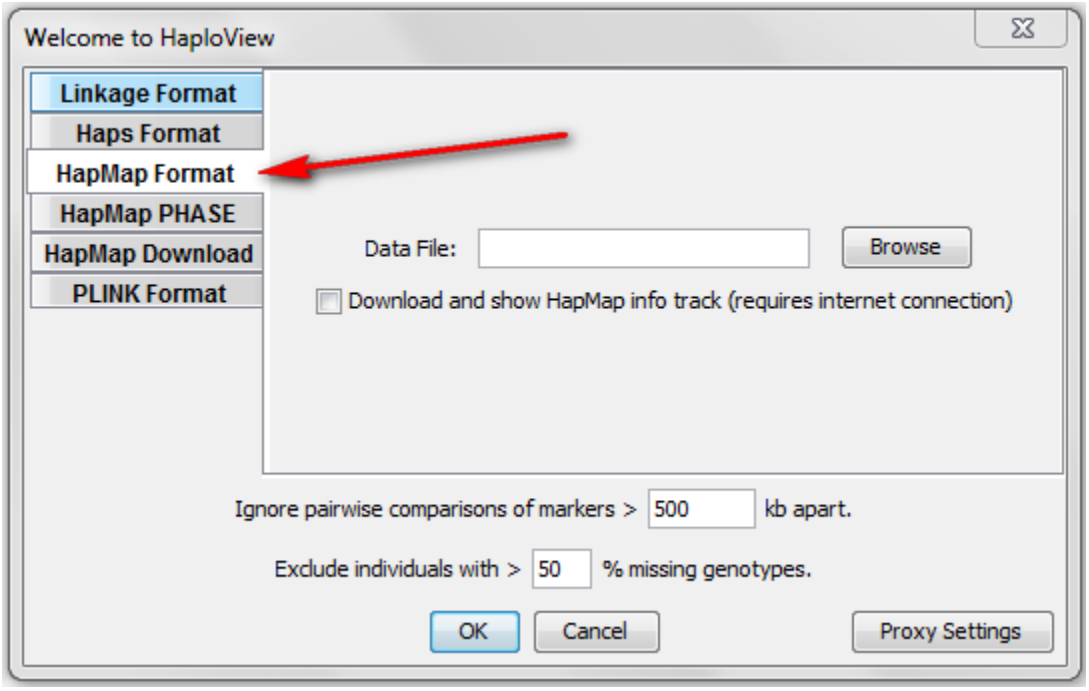

V opciji konfiguracije izberemo želeno populacijo, nastavimo Save to DISK in s klikom na GO datoteko shranimo na disk;



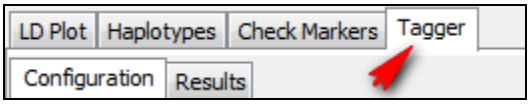

***HapMap privzeto shrani datoteko pod imenom dumped_region.***

V naslednjih korakih poženemo program HaploView, ki je pred-inštaliran v programskih datotekah na računalniku.

Izberemo tabulator HapMap FORMAT in s pomočjo funkcije BROWSE poiščemo svojo dumped_region datoteko.

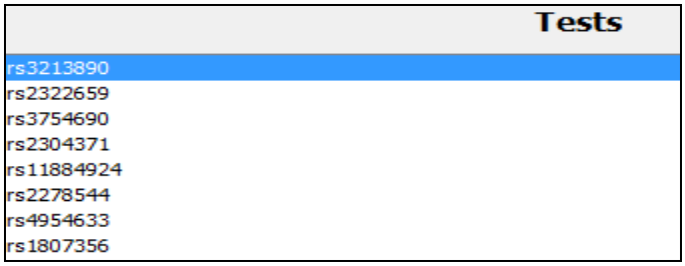



Ko je datoteka uspešno naložena v program, se pomaknemo na tabulator TAGGER.




Izberemo pairwise tagging only, preverimo $r^2$ mejno vrednost ter poženemo tagger.

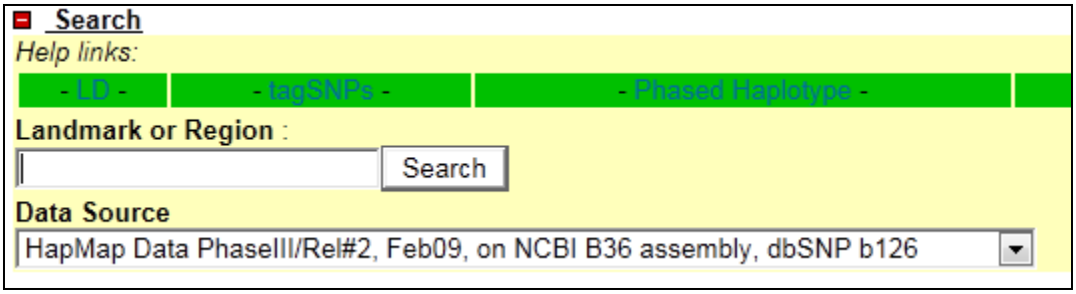

Kot rezultat dobimo TAG SNP-je v okencu TESTS.



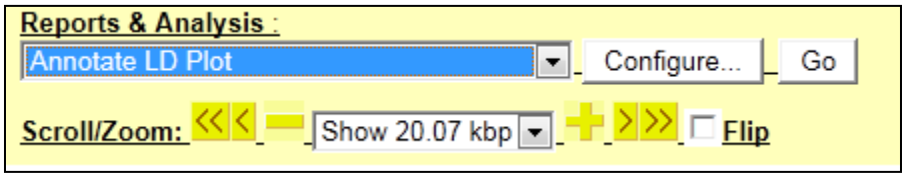

***Naloga 7:*** *Za gen IL16 in en poljuben gen v populaciju CEU in še posebej v eni poljubni populaciji najdi najmanjše število SNP-jev, ki bodo pokrili vse halplotipe.*

***Primer:*** Išči podatke za ABO gen v Rel#2 (NCBI B36 baza podatkov). Koliko haplotipskih blokov lahko vizualno določiš?

V iskalno okence vpišemo ime ali ID gena:

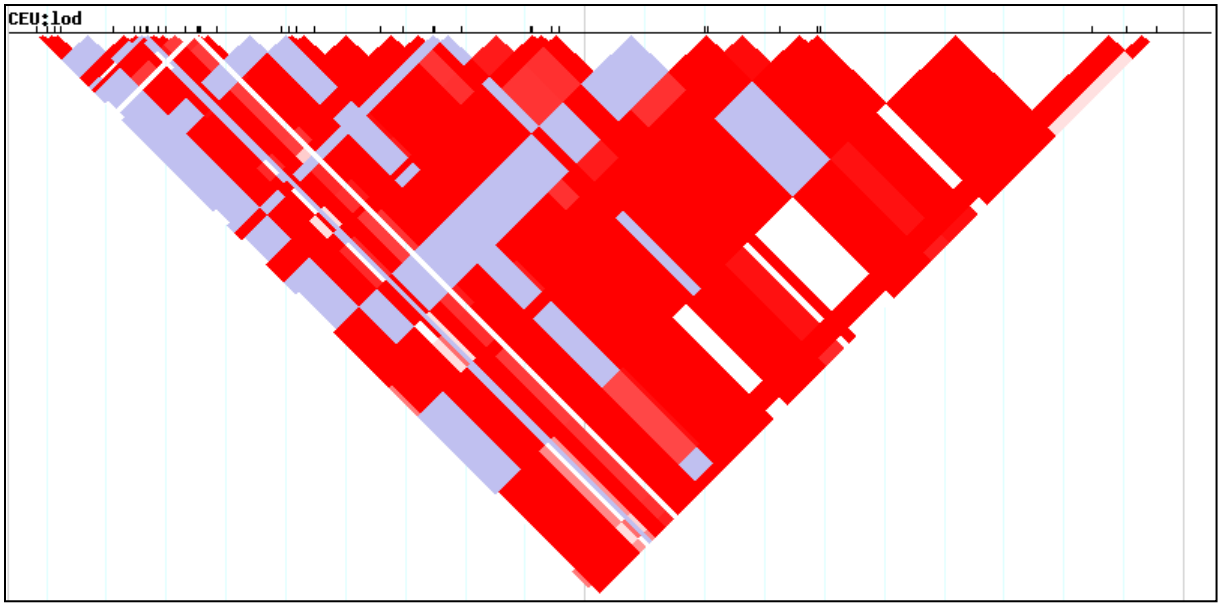



V drsnem oknu na levi pa izberemo Annotate LD plot;




V meniju spodaj obkljukamo Recombination rate in kliknemo UPDATE IMAGE:


Kot rezultat dobimo LD graf, na katerem lahko odčitamo 2 haplotipska bloka:




***Naloga 8:*** *Išči podatke za SEPHS1 gen v Rel#2 (NCBI B36 baza podatkov). Vključi stopnjo rekombinacije. Koliko haplotipskih blokov lahko vizualno določiš?*

## 3.3 Konstrukcija LD grafa v programu HaploView

**Primer:** Poišči vrednosti LD v .txt formatu in načrtaj LD graf za sledeče SNP-je: rs4584192, rs4742222, rs1885923 (izpiši lokacije SNP-jev).

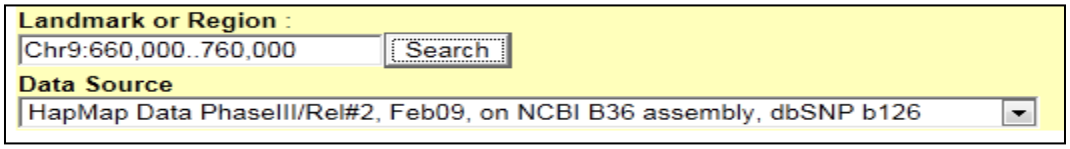
***Zgoraj navedenega primera se lahko lotimo na več načinov.***

**Način 1:**

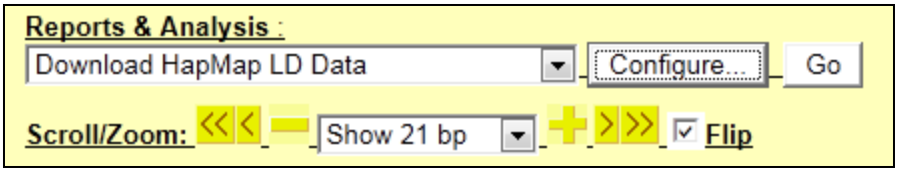
V podatkovni bazi dbSNP poiščemo fizično lokacijo SNP-jev na kromosomu in v iskalno okence vnesemo sledeč zapis – interval (gledamo build 36.3 reference):

***Chr(navedemo kromosom):lokacija1...lokacija2***

***Chr10:12,252,213...12,500,000***




V desnem okencu nastavimo download HapMap LD data, pazimo, da pod konfiguracijo nastavimo OUTPUT FORMAT TEXT:

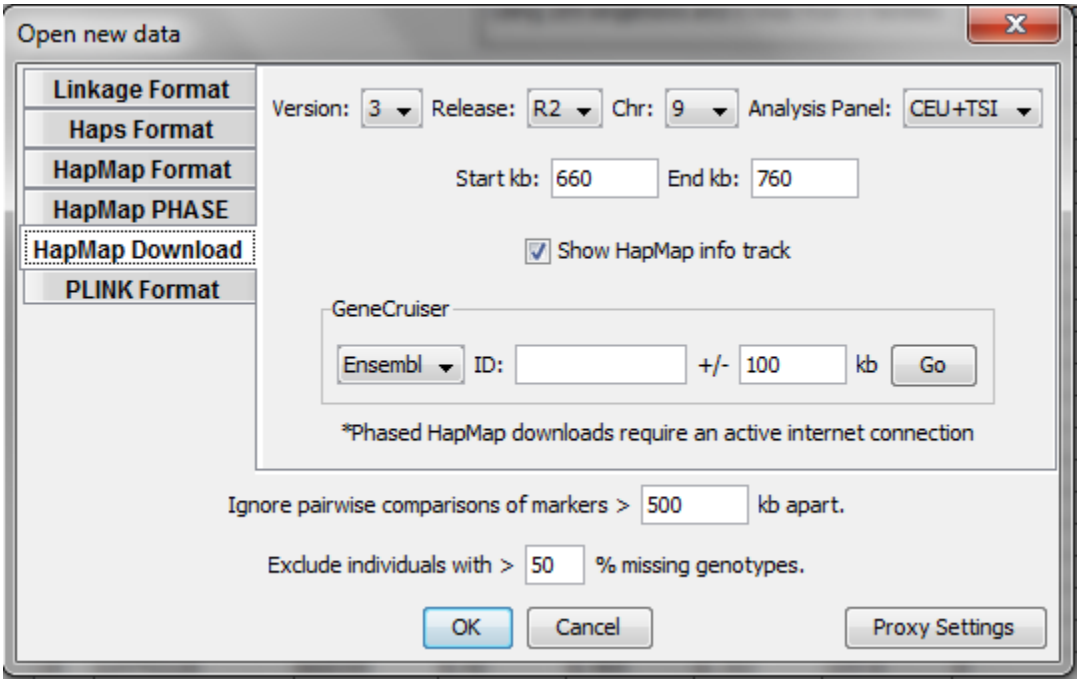



Rezultate dobimo podane v tekstovni obliki ter poiščemo tarčne SNP-je (uporabi funkcijo Ctrl+F):

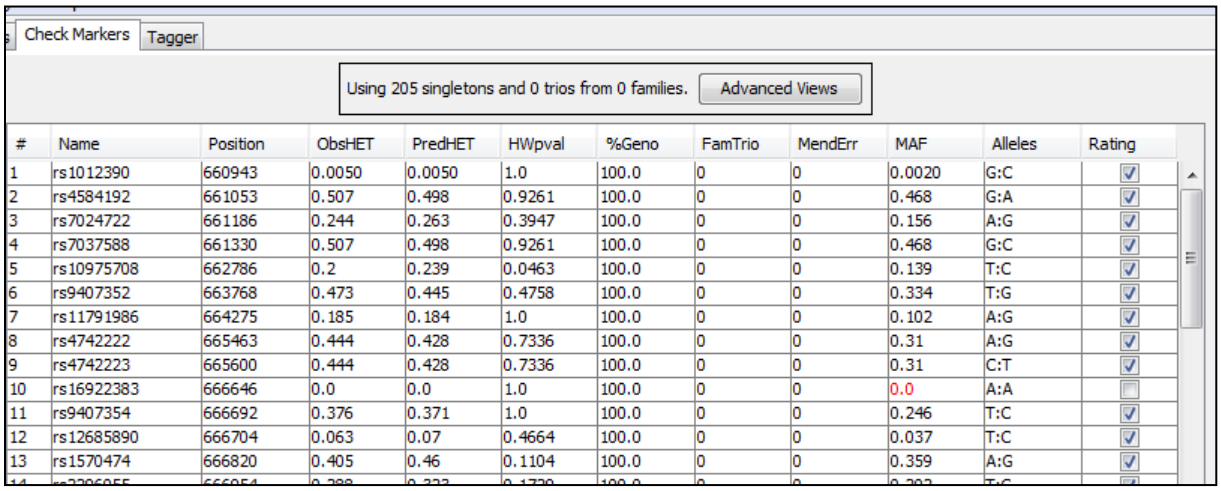
```
#Thu Mar 10 09:26:33 2011: HapMap LD data dump, 142 S
#pos1 pos2 pop  marker1  marker2 D' r^2 LOD
660943 661053 CEU rs1012390 rs4584192 1 0.008 0.09
660943 661186 CEU rs1012390 rs7024722 1 0.001 0.06
660943 661330 CEU rs1012390 rs7037588 1 0.008 0.09
660943 662786 CEU rs1012390 rs10975708 1 0.001 0.07
660943 663768 CEU rs1012390 rs9407352 1 0.013 0.23
660943 664275 CEU rs1012390 rs11791986 1 0 0.04
660943 664428 CEU rs1012390 rs4742217 1 0.007 0.05
660943 665463 CEU rs1012390 rs4742222 1 0.015 0.27
660943 665600 CEU rs1012390 rs4742223 1 0.015 0.27
```

**Način 2:**

Zaženemo HaploView in izberemo tabulator HapMap Download ter nastavimo ustrezne parametre:

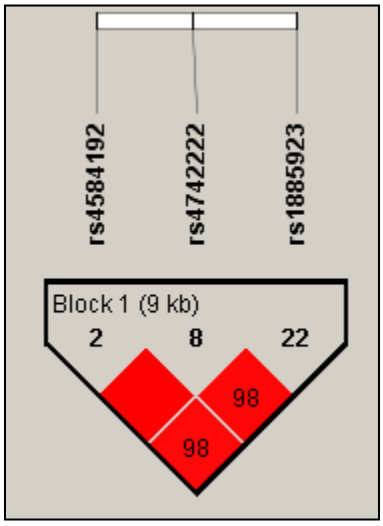


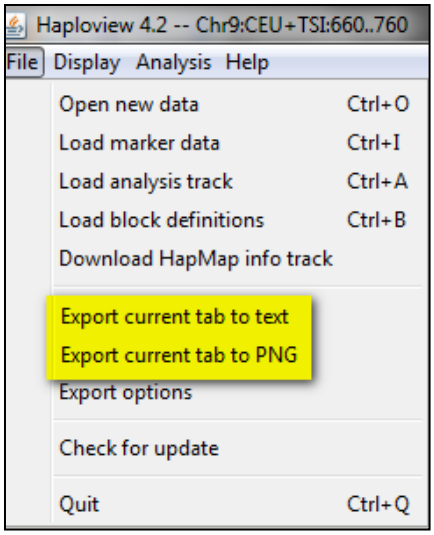
Ko program naloži ustrezne podatke izberemo tabulator CHECK MARKERS in pustimo obkljukane samo tarčne markerje;




| # | Name | Position | ObsHET | PredHET | HWpval | %Geno | FamTrio | MendErr | MAF | Alleles | Rating |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs1012390 | 660943 | 0.0050 | 0.0050 | 1.0 | 100.0 | 0 | 0 | 0.0020 | G:C | ☑ |
| 2 | rs4584192 | 661053 | 0.507 | 0.498 | 0.9261 | 100.0 | 0 | 0 | 0.468 | G:A | ☑ |
| 3 | rs7024722 | 661186 | 0.244 | 0.263 | 0.3947 | 100.0 | 0 | 0 | 0.156 | A:G | ☑ |
| 4 | rs7037588 | 661330 | 0.507 | 0.498 | 0.9261 | 100.0 | 0 | 0 | 0.468 | G:C | ☑ |
| 5 | rs10975708 | 662786 | 0.2 | 0.239 | 0.0463 | 100.0 | 0 | 0 | 0.139 | T:C | ☑ |
| 6 | rs9407352 | 663768 | 0.473 | 0.445 | 0.4758 | 100.0 | 0 | 0 | 0.334 | T:G | ☑ |
| 7 | rs11791986 | 664275 | 0.185 | 0.184 | 1.0 | 100.0 | 0 | 0 | 0.102 | A:G | ☑ |
| 8 | rs4742222 | 665463 | 0.444 | 0.428 | 0.7336 | 100.0 | 0 | 0 | 0.31 | A:G | ☑ |
| 9 | rs4742223 | 665600 | 0.444 | 0.428 | 0.7336 | 100.0 | 0 | 0 | 0.31 | C:T | ☑ |
| 10 | rs16922383 | 666646 | 0.0 | 0.0 | 1.0 | 100.0 | 0 | 0 | 0.0 | A:A | ☐ |
| 11 | rs9407354 | 666692 | 0.376 | 0.371 | 1.0 | 100.0 | 0 | 0 | 0.246 | T:C | ☑ |
| 12 | rs12685890 | 666704 | 0.063 | 0.07 | 0.4664 | 100.0 | 0 | 0 | 0.037 | T:C | ☑ |
| 13 | rs1570474 | 666820 | 0.405 | 0.46 | 0.1104 | 100.0 | 0 | 0 | 0.359 | A:G | ☑ |

Nato se pomaknemo na tabulator LD plot, kjer lahko izvozimo graf v obliki .png ali pa tekstovno obliko;



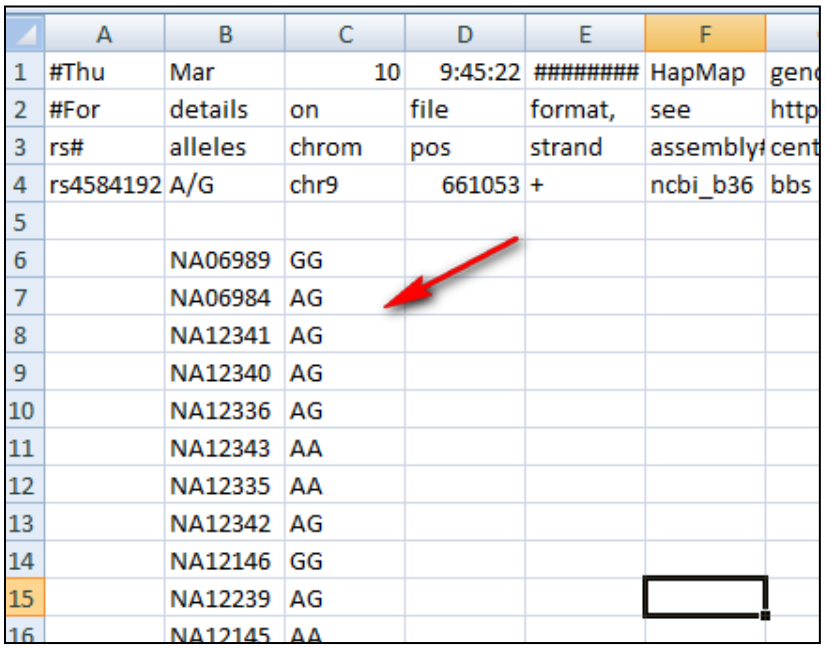



**Način 3:**

Ta način lahko uporabimo tudi, kadar imamo svoje podatke o genotipih, npr. rezultati genotipizacij po testu RFLP,...

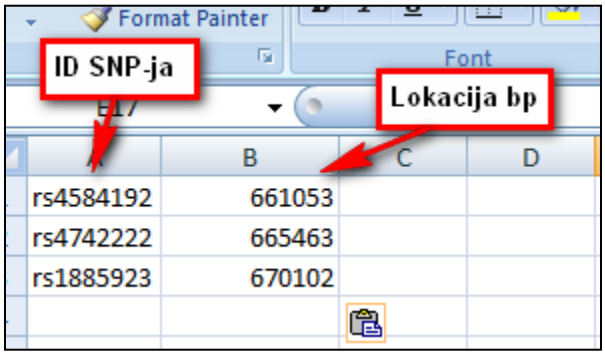
Da bi pridobili rezultate genotipov tarčnih SNP-jev, bomo uporabili HapMap. V iskalno okence bomo vpisali ID SNP-ja in v desnem okencu nastavili Download HapMap genotype data, pod konfiguracijo pa izbrali text format.

Ko bo rezultat podan, ga bomo kopirali v beležnico in shranili na DISK. Datoteko.txt bomo odprli s programom MS Excel in transponirali podatke o genotipih tako, da bodo razporejeni navpično in ponovno shranili Excelove datoteke:

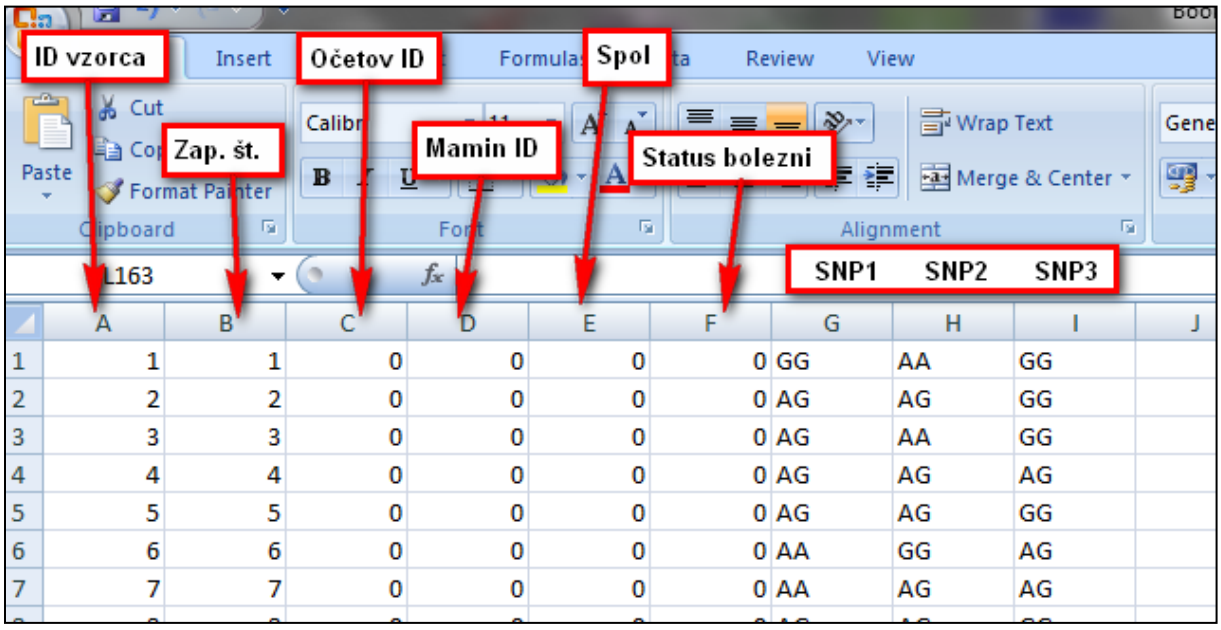


| | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| 1 | #Thu | Mar | 10 | 9:45:22 | ######## | HapMap | genc |
| 2 | #For | details | on | file | format, | see | http |
| 3 | rs# | alleles | chrom | pos | strand | assembly# | cent |
| 4 | rs4584192 | A/G | chr9 | 661053 | + | ncbi_b36 | bbs |
| 5 | | | | | | | |
| 6 | | NA06989 | GG | | | | |
| 7 | | NA06984 | AG | | | | |
| 8 | | NA12341 | AG | | | | |
| 9 | | NA12340 | AG | | | | |
| 10 | | NA12336 | AG | | | | |
| 11 | | NA12343 | AA | | | | |
| 12 | | NA12335 | AA | | | | |
| 13 | | NA12342 | AG | | | | |
| 14 | | NA12146 | GG | | | | |
| 15 | | NA12239 | AG | | | | |
| 16 | | NA12145 | AA | | | | |

Ko končamo z urejanjem, si pripravimo novo Excelovo datoteko, kjer bomo na prvi list zapisali ime INFO, na drugega pa PED. V bazi dbSNP poiščemo lokacije (36.3 reference) SNP-jev in jih vnesemo na prvi list (INFO);




Na list PED pa v enakem zaporedju vnesemo genotipe SNP-jev. Torej, če je prvi SNP na listu INFO rs4584192, potem bo prvi stolpec genotipov pripadal točno temu SNP-ju:




Ker nimamo podanih drugih podatkov (za izračun LD jih niti ne potrebujemo) v stolpec ID vzorca in Zap. Št. Vnesemo zaporedne številke od 1 do n. V stolpce ID očeta in matere ter spola in statusa bolezni pa vnesemo kar vrednosti 0.

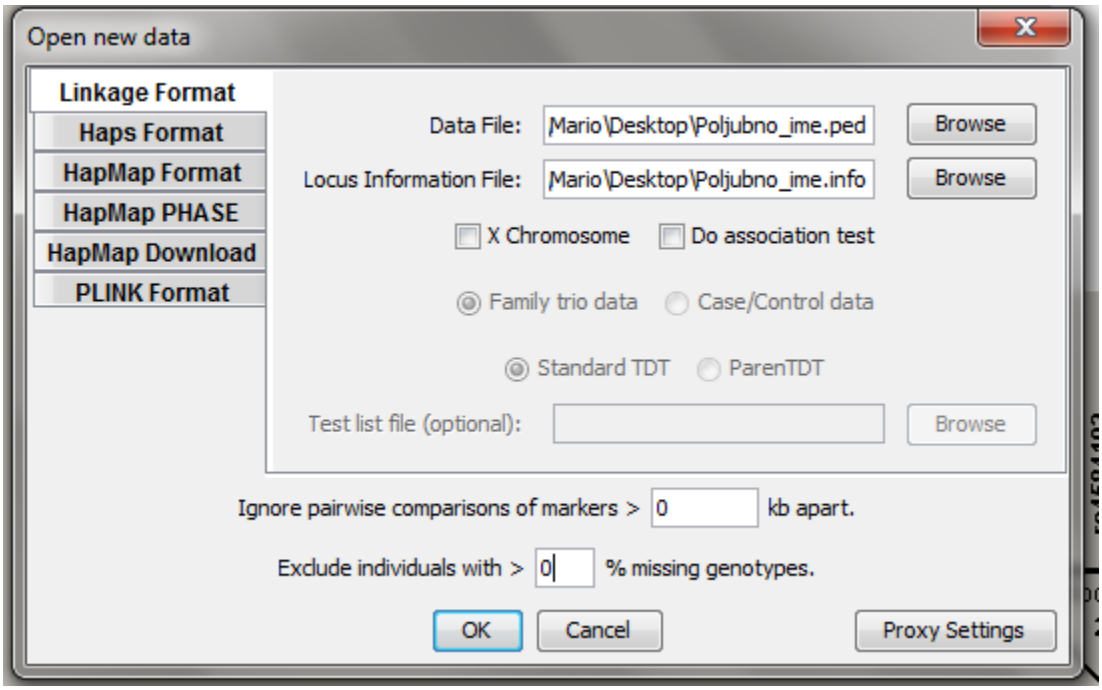
V naslednjem koraku moramo zamenjati genotipe v numerične vrednosti:
A→1
C→2
G→3
T→4
N→0

Genotip AA → 1- -1 (- - pomeni, da moramo med 11 dodati še dva presledka), pomagamo si s funkcijo Ctrl+F ter REPLACE. V kolikor se nam zgodi, da za kak vzorec nimamo genotipa (oznaka NN) ga na enak način zamenjamo NN →0- -0;

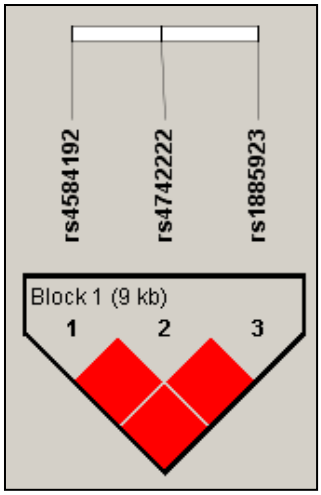
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 3 3 | 1 1 | 3 3 |
| 2 | 2 | 2 | 0 | 0 | 0 | 0 | 1 3 | 1 3 | 3 3 |
| 3 | 3 | 3 | 0 | 0 | 0 | 0 | 1 3 | 1 1 | 3 3 |
| 4 | 4 | 4 | 0 | 0 | 0 | 0 | 1 3 | 1 3 | 1 3 |
| 5 | 5 | 5 | 0 | 0 | 0 | 0 | 1 3 | 1 3 | 3 3 |
| 6 | 6 | 6 | 0 | 0 | 0 | 0 | 1 1 | 3 3 | 1 3 |
| 7 | 7 | 7 | 0 | 0 | 0 | 0 | 1 1 | 1 3 | 1 3 |

Po urejanju Excelove datoteke, odpremo Beležnico in v njo skopiramo list INFO. Beležnico shranimo kot poljubno_ime.info. Enako storimo z listom PED ter shranimo Beležnico kot poljubno_ime.ped.

Zaženemo program HaploView ter izberemo tabulator Linkage Format. S pomočjo brskalnika naložimo obe datoteki ter vrednosti Ignore pairwise in Exclude nastavimo na 0:




Rezultat je LD graf, katerega lahko izvozimo v poljubnem formatu;

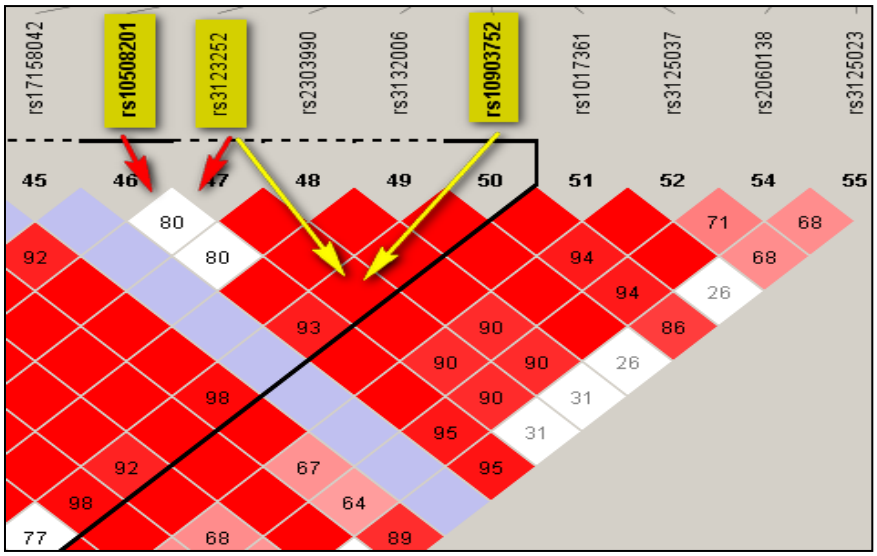



## 3.4 Interpretacija LD grafa

| | D' < 1 | D' = 1 |
|---|---|---|
| LOD < 2 | BELA | MODRA |
| LOD ≥ 2 | ROŽNATA | RDEČA |

***Če se na grafu postavimo na kvadratek in kliknemo z desno miškino tipko, se nam izpišejo vrednosti.***

**Kako graf beremo:**




Iz slike je razvidno, da SNP rs10508201 in SNP rs3123252 nista v visokem vezavnem neravnovesju (beli kvadrat), kar pomeni, da med seboj nista povezana, medtem ko pa sta SNP rs3123252 in SNP rs10903752 v visokem vezavnem neravnovesju (rdeči kvadrat). Barvna shema

kvadratov nam prikazuje vrednosti D' in LOD. V primeru, da je LOD < 2 in D' < 1, je barva kvadrata bela, v primeru, da je LOD < 2 in D' = 1, je barva kvadrata modra, v primeru, da je LOD ≥ 2 in D' je < 1, je barva kvadrata rožnata in v primeru, da je LOD ≥ 2 in D' = 1, potem je barva kvadrata rdeča, ki nam predstavlja najvišjo vezavno neravnovesje. Številke v kvadratih pa nam predstavljajo vrednosti $r^2$.

***Naloga 9:*** *Skonstruiraj LD graf 5 SNP-jev, ki ležijo na istem kromosomu, ter izpiši SNP-je v vezavnem neravnovesju vključno z vrednostmi vseh treh parametrov.*

# 4 1000 genomes

En glavnih ciljev biologije in medicine je razumeti odnos med genotipom in fenotipom. Rezultat projekta humani genom je referenčna genomska sekvenca, ki predstavlja osnovo za študij humane genetike. Kljub temu pa ne omogoča študija variacije v genomu, saj je za to potrebno natančno poznavanje sekvence DNA. Leta 2008 je izšel katalog variabilnih mest človeškega genoma, ki je vseboval 11 milijonov SNP in 3 milijone kratkih insercij in delecij. Mednarodni projekt HapMap je zbral vse alelne frekvence in korelacijske vzorce sosednjih variant (fenomen, poznan kot vezavno ravnovesje, ang. LD »linkage disequlibium«) večih populacij za 3,5 milijonov SNP-ov.
Omenjene raziskave so peljale odkrivanje genov, povezanih z boleznimi v smer asociacijskih študij na celotnem genomu (ang. GWAS »genome-wide association studies«), v katerih so primerjali genotipe posameznikov na več sto tisočih variabilnih mestih, v kombinaciji z znano LD strukturo. Tako so odkrili asociacije med posameznimi SNP in pojavom bolezni. V zadnjih petih letih so GWA študije identificirale več kot 1000 genomskih regij povezanih z verjetnostjo za razvoj posamezne bolezni. Kljub tem odkritjem, je za razumevanje prispevka posameznega gena k razvoju bolezni potrebno vložiti še veliko dela. Ko enkrat odkrijemo da posamezna regija na DNA vsebuje rizičen lokus, ki prispeva k razvoju bolezni, je potrebno preučiti vse genetske variante v tem lokusu. Ko odkrijemo vzročno varianto oz. variante, odgovorne za razvoj bolezni, določimo njihovo vlogo pri razvoju bolezni in pojasnimo njihovo vlogo v funkcionalnih bioloških poteh.

Variante z nizko frekvenco (definirane kot tiste s frekvenco manj prisotnega alela med 0,5 in 5%) so bolj pogoste kot variante z veliko frekvenco v populaciji in prav tako signifikantno prispevajo k genetski arhitekturi bolezni, vendar jih do sedaj še ni bilo mogoče sistemsko preučevati. Da bi lahko popolnoma razumeli vlogo pogostih in nizko-frekvenčnih variacij, je bil potreben popoln katalog človeške gentske varibilnosti.

V ta namen se je leta 2008 razvil projekt »1000 Genomes«. Pri slednjem so želeli odkriti, genotipizirati in zagotoviti točno informacijo o haplotipih vseh DNA polimorfizmov različnih človeških populacij. Cilj projekta je bil določiti 95% vseh variant v genomu, ki so dostopne za sekveniranje z visokotehnološkimi metodami in imajo alelno frekvenco večjo ali enako 1% (kar je tudi klasična definicija SNP) v vsaki od petih glavnih populacijskih skupin (populacije iz Evrope, vzhodne Azije, južne Azije, zahodne Afrike in Amerike).

V projektu je sodelovalo devet sekvenčnih centrov, ki so skupaj sekvenirali 4,9 terabaz DNA sekvence iz DNA, pridobljene iz nesmrtnih limfoblastnih celičnih linij. Sama genotipizacija sekvenc je potekala v okviru treh glavnih projektov.

V okviru Trio projekta je potekalo sekveniranje z metodo, ki omogoča visoko pokrivnost (ang. »high-coverage«, npr. 42-kratna pokrivnost pomeni, da je bilo pri sekvenci dolgi 3 milijone baznih parov določenih 126 milionov baznih parov). Sekvenirali so DNA dveh družin nigerijske in

evropske populacije, sestavljenih iz obeh staršev in hčere. V okviru projekta z nizko pokrivnostjo (2-6 -kratna pokrivnost) so sekvenirali DNA 179 posameznikov različnih populacij. V okviru Exon projekta pa so najprej iz 697 posameznikov različnih populacij ciljno zajeli 8.140 eksonov iz 906 genov (skupno 1.4 Mb), katere so nato sekvenirali z visoko pokrivnostjo.

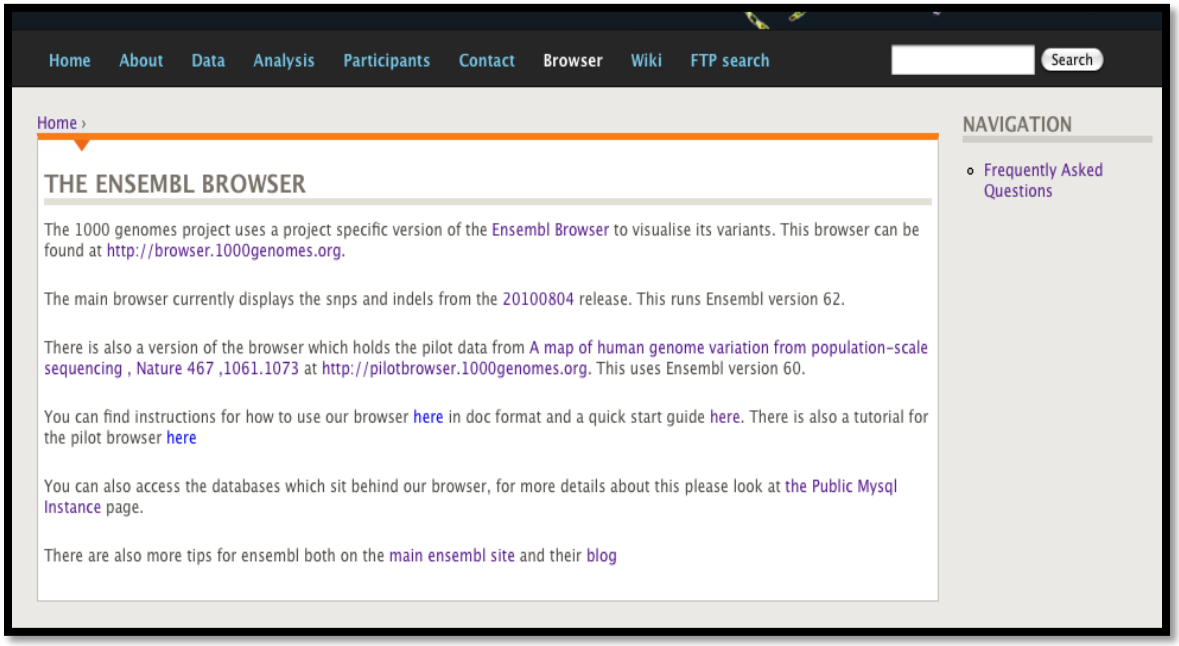
Vsem projektom so bile skupne štiri lastnosti: odkritje, prečiščenje, genotipizacija in potrditev. Pri odkritju so znanstveniki sekvenirane sekvence porazdelili glede na referenčno sekvenco in identificirali kandidatne regije, ker so se posamezniki razlikovali od referenčne sekvence. Sledilo je prečiščenje s kontrolami kvalitete, s katerimi so odstranili lažno pozitivna variabilna mesta. Genotipizacija je omogočila določitev alelov posameznika na posameznem variabilnem mestu. Zadnji korak je bila potrditev novo odkritih variant z neodvisno tehnologijo, kar je omogočilo določitev stopnje lažnih odkritij (ang. »false discovery rate«). Eksperimentalni projekti so se med sabo razlikovali v sposobnosti pridobitve podatkov in frekvenc posameznih variacij ter v analitičnih metodah, uporabljenih za sklepanje o posameznikovem genotipu.

Popoln katalog genetske variacije omogoča identifikacjo signalov, ki so bili prej v študijah zgrešeni in tako opazno poveča število kandidatnih funkcijskih alelov na posameznem lokusu. Podatki iz kataloga so že bili uporabljeni pri GWA študija različnih lastnosti in bolezni, od kajenja do multiple skleroze. Prav tako so bili uporabljeni kot izključitveni faktor pri študijah Mendlovih bolezni (monogenske bolezni) in študijah tumorigeneze. Informacije iz projekta se uporabljajo tudi pri dizajniranju testov za genotipizacijo nove generacije in kot vzorec za načrtovanje prihodnjih GWA študij na večjih vzorcih.[19] V prilogi 1 je prikazana uporaba spletnega brskalnika »1000 genomes«.

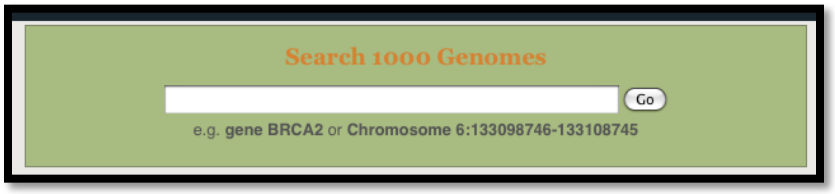
Prihodnji plan za projekt je sekveniranje dodatnih vzorcev, za dosego končnega cilja 2.500 sekveniranih vzorcev pri 4-kratni pokritosti. Zadnji vzorci naj bi bili sekvenirani konec leta 2011.

Merjenje variabilnosti človeške DNA je glavni predpogoj za raziskovanje humane genetike. Projekt »1000 Genomes« predstavlja korak naprej k popolnem opisu človeških DNA polimorfizmov. Večja količina podatkov, ki jo ponuja projekt »1000 Genomes« bo omogočila bolj natančno vključitev variant v GWA študije, kar bo omogočilo boljšo lokalizacijo z boleznijo povezanih področij DNA. Aplikaciji podatkov iz projekta in metod uporabljenih v projektu bosta prispevali k bolj doumljivemu razumevanju vloge podedovane variabilnosti DNA v človeški evoluciji in razvoju bolezni.

Pojdi na: http://www.1000genomes.org/ensembl-browser




Izberi link »Browser« (http://browser.1000genomes.org/), kjer najdeš iskalno okence na vrhu strani na levi strani.



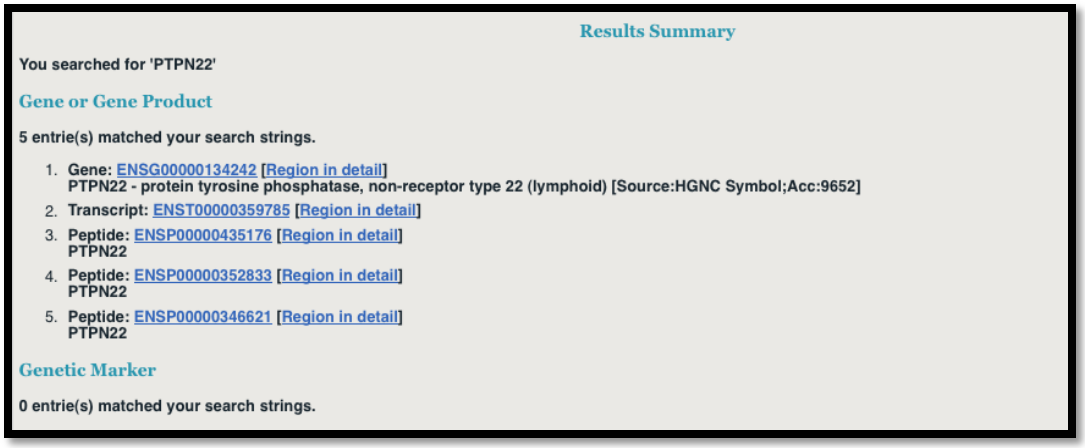

V iskalno okence vpiši ime gena, simbol, kromosom... in klikni »Go«.

Za primer v skripti je opisan gen PTPN22.

Kadar iščemo regijo na kromosomu, se odpre stran »chromosomal location«, ki zajema del kromosoma v okolici navedene regije.

Kadar iščemo gen ali genski produkt, spada rezultat iskanje v eno izmed naslednjih kategorij:

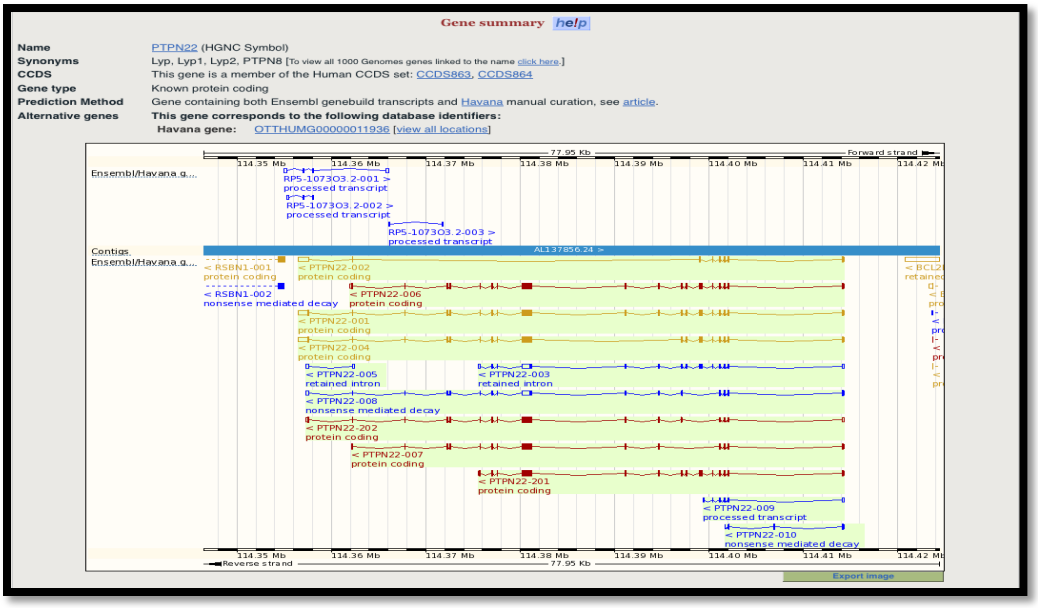
- »Gene« - gen
- »Transcript« - transkript
- »Peptide« - protein




- S klikom na »gene« v rezultatu iskanja odpremo »gene page« - stran s podatki o genu
- S klikom na »transcript« v rezultatu iskanja odpremo »transcript page« - stran s podatki o genskem transkriptu (RNA)
- S klikom na »peptide« v rezultatu iskanja odpremo »peptide page« - stran s podatki o proteinu
- S klikom na »region in detail« odpremo »LocationView page« - stran s podatki o regije, kjer s ta zadetek nahaja

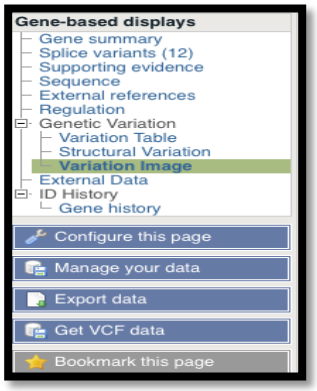
## 4.1 Dostop do podatkov o variacija na strani »gene«

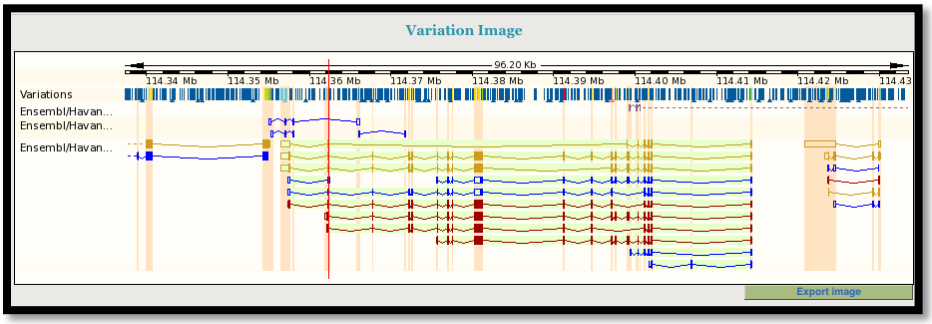
- S klikom na »gene« v rezultatu iskanja odpremo »gene page« - stran s podatki o genu
- S klikom na »transcript« v rezultatu iskanja odpremo »transcript page« - stran s podatki o genskem transkriptu (RNA)

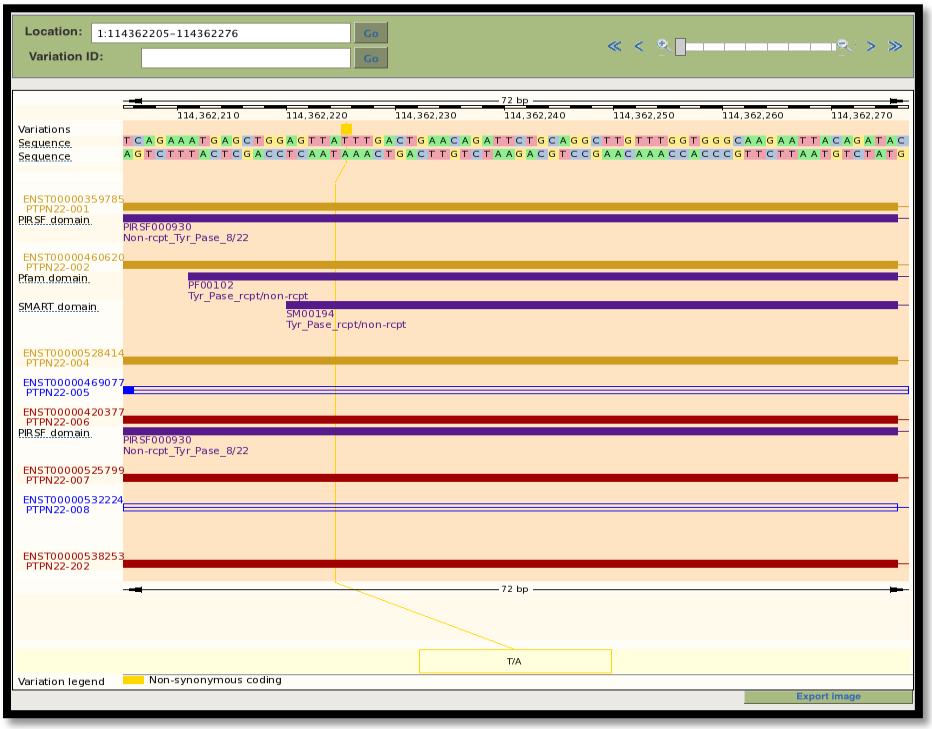


Stran vsebuje celoten razpon gena, prikazane so vse kopiji gena.

- Klikni na povezavo »Variation Image« pod »Genetic Variation«

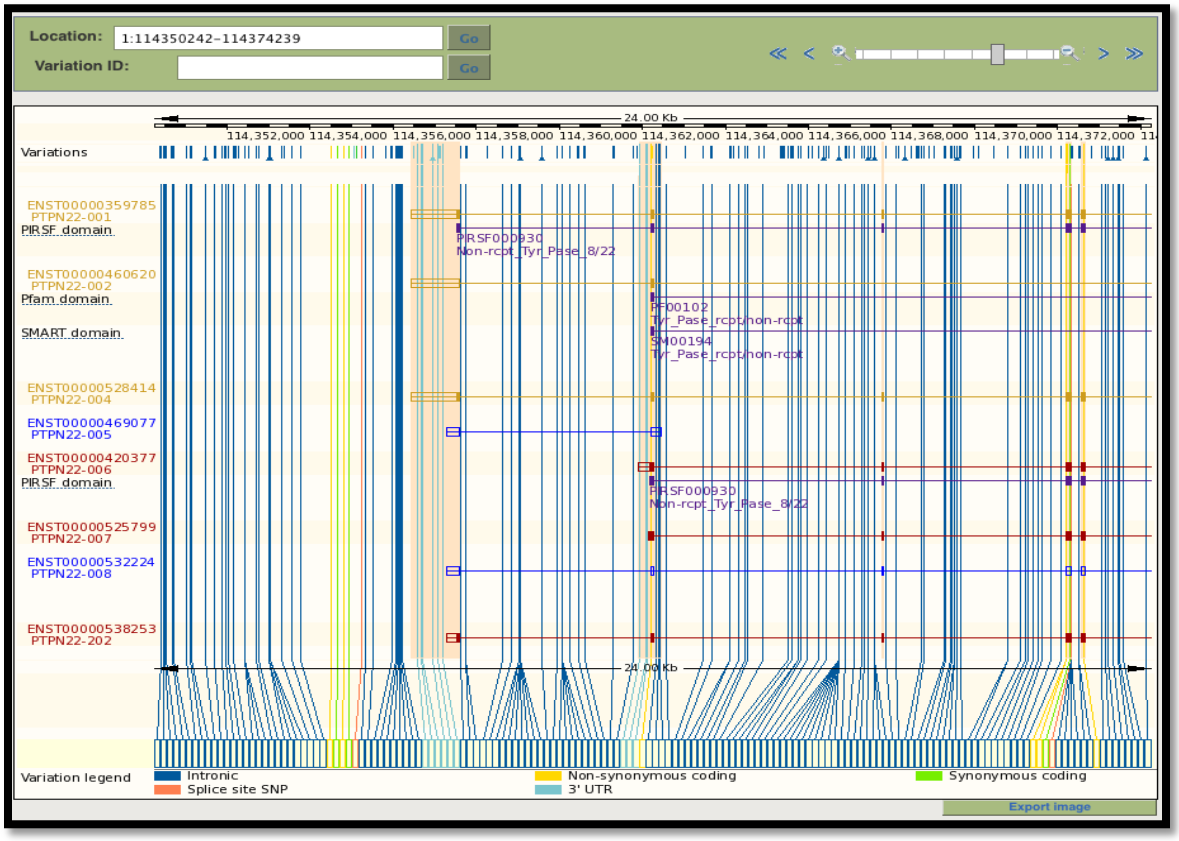


Zgornja slika predstavlja celoten razpon gena in spremljevalnie regije. Spodnja slika predstavlja povečan pogled na prvi ekson.

V okencu »Location« lahko spremenite razpon in vidite širše (ali ožje) področje, kot je prikazano spodaj:

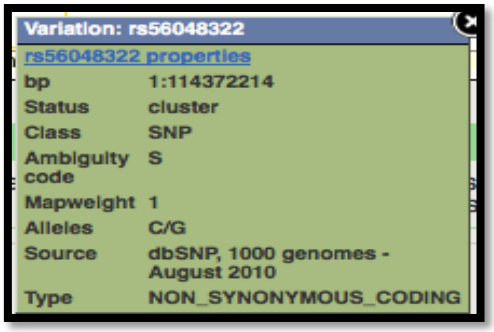



Slika zajema le del variacij. Povezava do popolne slike je sledeča:

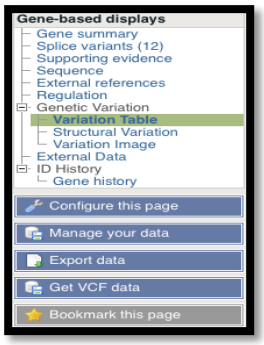
http://browser.1000genomes.org/Homo_sapiens/Gene/Variation_Gene/Image?db=core;g=ENSG00000134242;r=1:114356433-114414381

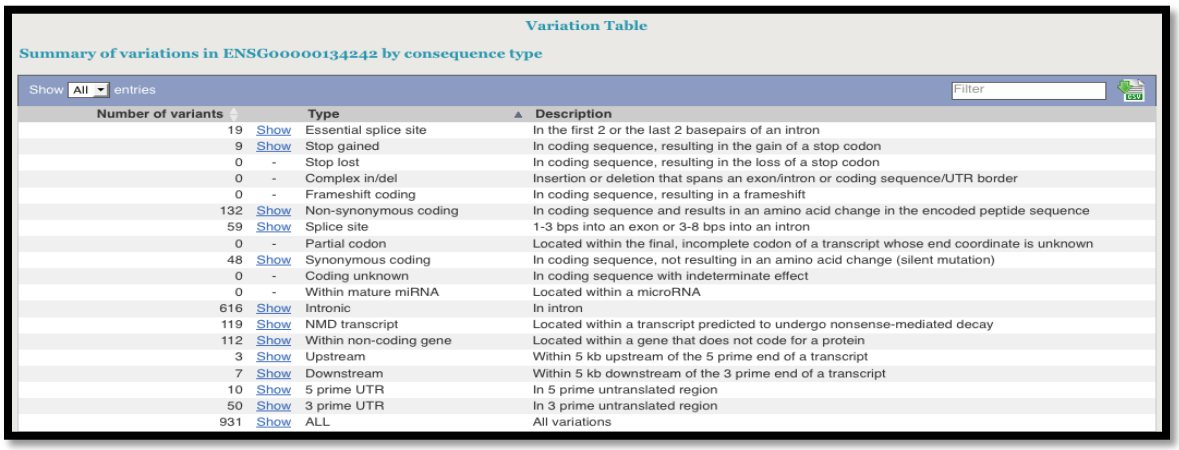
SNP-i so prikazani v različnih barvah, ki pomenijo različne funkcije SNP-ov (intronski, sinonimen, nesinonimen). Nesinonimni SNP-i so še dodatno označeni.

- klikni na kateri koli SNP in odpre se okno z informacijami o njem




- S klikom na "Variation Table" pod "Genetic Variation" v levem meniju dobimo sezma, SNP-ov v določeni regiji gena.



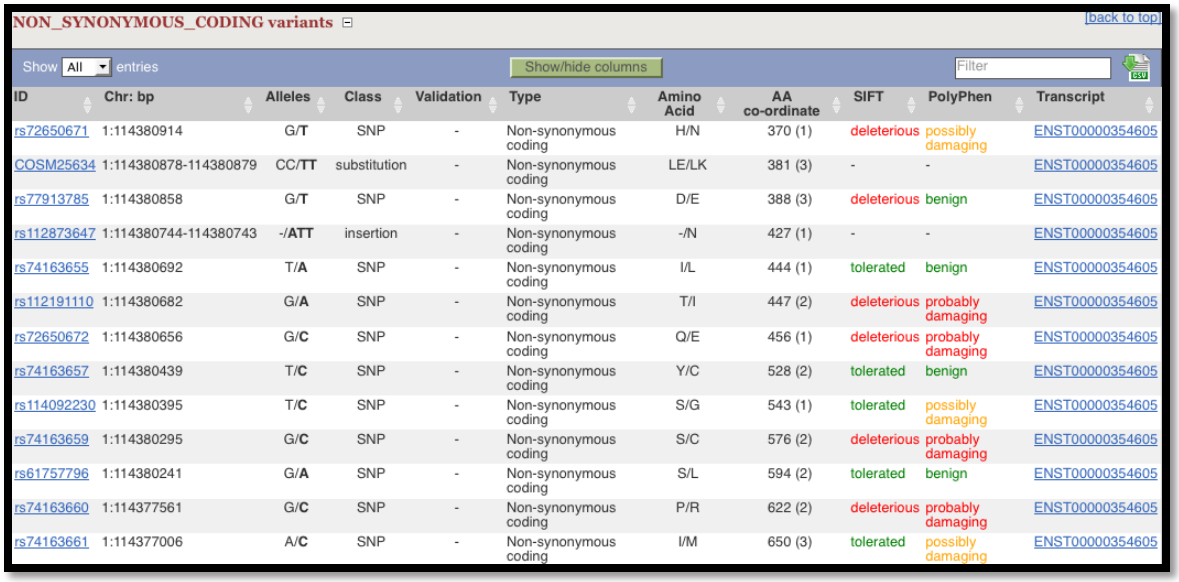
Variation Table

Summary of variations in ENSG00000134242 by consequence type

Show All entries Filter

| Number of variants | | Type | Description |
|---|---|---|---|
| 19 | Show | Essential splice site | In the first 2 or the last 2 basepairs of an intron |
| 9 | Show | Stop gained | In coding sequence, resulting in the gain of a stop codon |
| 0 | - | Stop lost | In coding sequence, resulting in the loss of a stop codon |
| 0 | - | Complex in/del | Insertion or deletion that spans an exon/intron or coding sequence/UTR border |
| 0 | - | Frameshift coding | In coding sequence, resulting in a frameshift |
| 132 | Show | Non-synonymous coding | In coding sequence and results in an amino acid change in the encoded peptide sequence |
| 59 | Show | Splice site | 1-3 bps into an exon or 3-8 bps into an intron |
| 0 | - | Partial codon | Located within the final, incomplete codon of a transcript whose end coordinate is unknown |
| 48 | Show | Synonymous coding | In coding sequence, not resulting in an amino acid change (silent mutation) |
| 0 | - | Coding unknown | In coding sequence with indeterminate effect |
| 0 | - | Within mature miRNA | Located within a microRNA |
| 616 | Show | Intronic | In intron |
| 119 | Show | NMD transcript | Located within a transcript predicted to undergo nonsense-mediated decay |
| 112 | Show | Within non-coding gene | Located within a gene that does not code for a protein |
| 3 | Show | Upstream | Within 5 kb upstream of the 5 prime end of a transcript |
| 7 | Show | Downstream | Within 5 kb downstream of the 3 prime end of a transcript |
| 10 | Show | 5 prime UTR | In 5 prime untranslated region |
| 50 | Show | 3 prime UTR | In 3 prime untranslated region |
| 931 | Show | ALL | All variations |

V okencu »Show« lahko izberemo prikaz samo določene kategorije SNP-ov.



NON_SYNONYMOUS_CODING variants

back to top

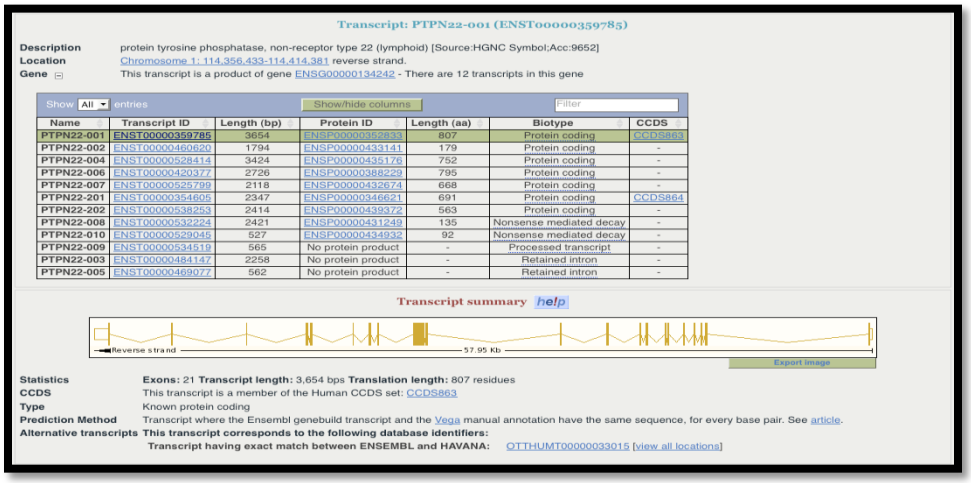
Show All entries | Show/hide columns | Filter

| ID | Chr: bp | Alleles | Class | Validation | Type | Amino Acid | AA co-ordinate | SIFT | PolyPhen | Transcript |
|---|---|---|---|---|---|---|---|---|---|---|
| rs72650671 | 1:114380914 | G/T | SNP | - | Non-synonymous coding | H/N | 370 (1) | deleterious | possibly damaging | ENST00000354605 |
| COSM25634 | 1:114380878-114380879 | CC/TT | substitution | - | Non-synonymous coding | LE/LK | 381 (3) | - | - | ENST00000354605 |
| rs77913785 | 1:114380858 | G/T | SNP | - | Non-synonymous coding | D/E | 388 (3) | deleterious | benign | ENST00000354605 |
| rs112873647 | 1:114380744-114380743 | -/ATT | insertion | - | Non-synonymous coding | -/N | 427 (1) | - | - | ENST00000354605 |
| rs74163655 | 1:114380692 | T/A | SNP | - | Non-synonymous coding | I/L | 444 (1) | tolerated | benign | ENST00000354605 |
| rs112191110 | 1:114380682 | G/A | SNP | - | Non-synonymous coding | T/I | 447 (2) | deleterious | probably damaging | ENST00000354605 |
| rs72650672 | 1:114380656 | G/C | SNP | - | Non-synonymous coding | Q/E | 456 (1) | deleterious | probably damaging | ENST00000354605 |
| rs74163657 | 1:114380439 | T/C | SNP | - | Non-synonymous coding | Y/C | 528 (2) | tolerated | benign | ENST00000354605 |
| rs114092230 | 1:114380395 | T/C | SNP | - | Non-synonymous coding | S/G | 543 (1) | tolerated | possibly damaging | ENST00000354605 |
| rs74163659 | 1:114380295 | G/C | SNP | - | Non-synonymous coding | S/C | 576 (2) | deleterious | probably damaging | ENST00000354605 |
| rs61757796 | 1:114380241 | G/A | SNP | - | Non-synonymous coding | S/L | 594 (2) | tolerated | benign | ENST00000354605 |
| rs74163660 | 1:114377561 | G/C | SNP | - | Non-synonymous coding | P/R | 622 (2) | deleterious | probably damaging | ENST00000354605 |
| rs74163661 | 1:114377006 | A/C | SNP | - | Non-synonymous coding | I/M | 650 (3) | tolerated | possibly damaging | ENST00000354605 |

Veliko tabel v brskalniku omogoča, da filtrirate vsebino oz. da uredite tabelo na podlagi različnih stolpcev ter tudi odstranite posamezne stolpce iz tabele.

## 4.2 Dostop do podatkov o variacijah na strani »transcript«

- Na strani rezultatov iskanja s klikom na »transcript« izberi »transcript page«.



Transcript: PTPN22-001 (ENST00000359785)

**Description** protein tyrosine phosphatase, non-receptor type 22 (lymphoid) [Source:HGNC Symbol;Acc:9652]
**Location** Chromosome 1: 114,356,433-114,414,381 reverse strand.
**Gene** This transcript is a product of gene ENSG00000134242 - There are 12 transcripts in this gene

Show All entries | Show/hide columns | Filter

| Name | Transcript ID | Length (bp) | Protein ID | Length (aa) | Biotype | CCDS |
|---|---|---|---|---|---|---|
| PTPN22-001 | ENST00000359785 | 3654 | ENSP00000352833 | 807 | Protein coding | CCDS863 |
| PTPN22-002 | ENST00000460620 | 1794 | ENSP00000433141 | 179 | Protein coding | - |
| PTPN22-004 | ENST00000528414 | 3424 | ENSP00000435176 | 752 | Protein coding | - |
| PTPN22-006 | ENST00000420377 | 2726 | ENSP00000388229 | 795 | Protein coding | - |
| PTPN22-007 | ENST00000525799 | 2118 | ENSP00000432674 | 668 | Protein coding | - |
| PTPN22-201 | ENST00000354605 | 2347 | ENSP00000346621 | 691 | Protein coding | CCDS864 |
| PTPN22-202 | ENST00000538253 | 2414 | ENSP00000439372 | 563 | Protein coding | - |
| PTPN22-008 | ENST00000532224 | 2421 | ENSP00000431249 | 135 | Nonsense mediated decay | - |
| PTPN22-010 | ENST00000529045 | 527 | ENSP00000434932 | 92 | Nonsense mediated decay | - |
| PTPN22-009 | ENST00000534519 | 565 | No protein product | - | Processed transcript | - |
| PTPN22-003 | ENST00000484147 | 2258 | No protein product | - | Retained intron | - |
| PTPN22-005 | ENST00000469077 | 562 | No protein product | - | Retained intron | - |

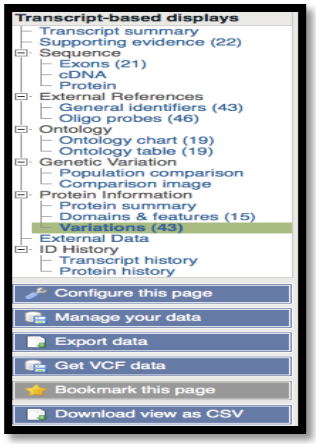
Transcript summary help

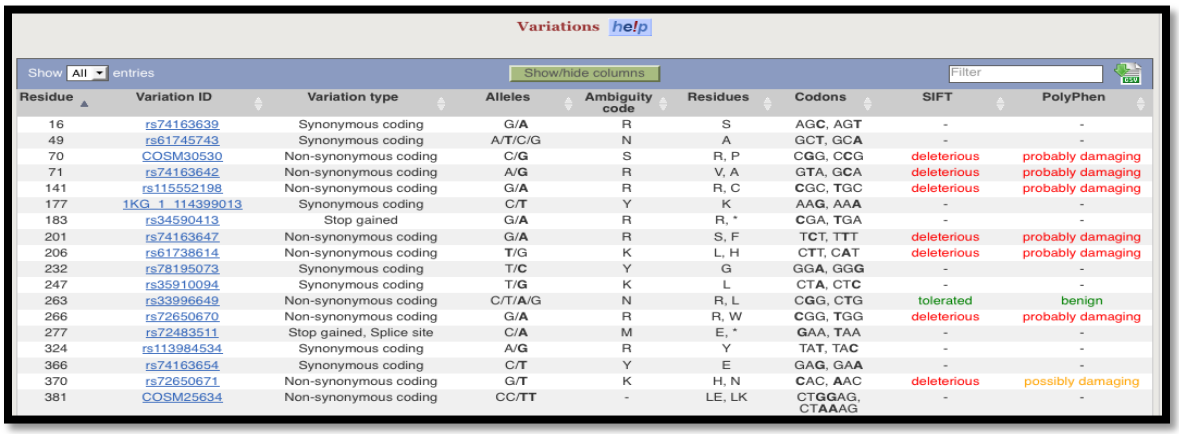
Reverse strand — 57.95 Kb

Export image

**Statistics** **Exons:** 21 **Transcript length:** 3,654 bps **Translation length:** 807 residues
**CCDS** This transcript is a member of the Human CCDS set: CCDS863
**Type** Known protein coding
**Prediction Method** Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.
**Alternative transcripts** **This transcript corresponds to the following database identifiers:**
**Transcript having exact match between ENSEMBL and HAVANA:** OTTHUMT00000033015 [view all locations]

S klikom na “Variations” pod “Protein Informationna levi, dobimo funkcije SNP-ov v regiji.



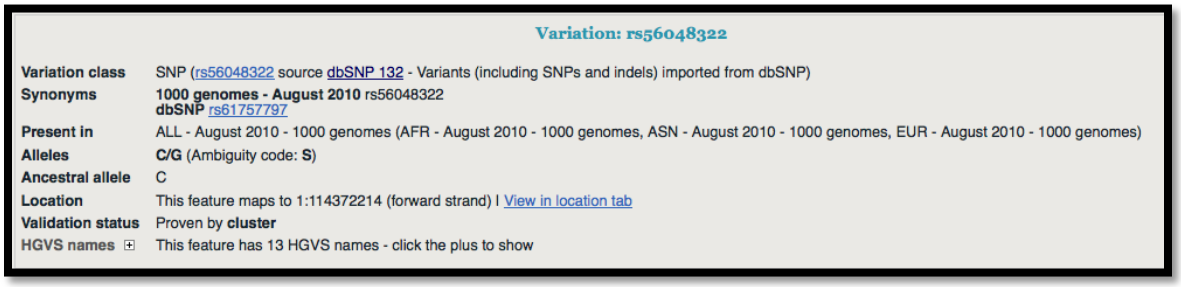
Variations help

Show All entries | Show/hide columns | Filter

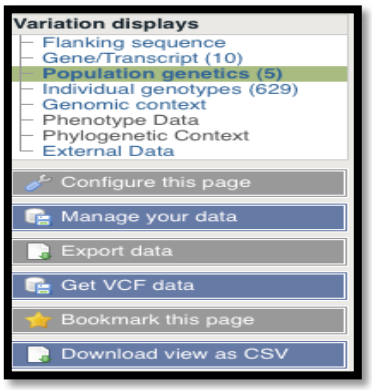
| Residue | Variation ID | Variation type | Alleles | Ambiguity code | Residues | Codons | SIFT | PolyPhen |
|---|---|---|---|---|---|---|---|---|
| 16 | rs74163639 | Synonymous coding | G/**A** | R | S | AG**C**, AG**T** | - | - |
| 49 | rs61745743 | Synonymous coding | A/**T**/C/G | N | A | GC**T**, GC**A** | - | - |
| 70 | COSM30530 | Non-synonymous coding | C/**G** | S | R, P | C**G**G, C**C**G | deleterious | probably damaging |
| 71 | rs74163642 | Non-synonymous coding | A/**G** | R | V, A | G**T**A, G**C**A | deleterious | probably damaging |
| 141 | rs115552198 | Non-synonymous coding | G/**A** | R | R, C | **C**GC, **T**GC | deleterious | probably damaging |
| 177 | 1KG_1_114399013 | Synonymous coding | C/**T** | Y | K | AA**G**, AA**A** | - | - |
| 183 | rs34590413 | Stop gained | G/**A** | R | R, * | **C**GA, **T**GA | - | - |
| 201 | rs74163647 | Non-synonymous coding | G/**A** | R | S, F | **T**C**T**, **T**T**T** | deleterious | probably damaging |
| 206 | rs61738614 | Non-synonymous coding | **T**/G | K | L, H | C**T**T, C**A**T | deleterious | probably damaging |
| 232 | rs78195073 | Synonymous coding | T/**C** | Y | G | GG**A**, GG**G** | - | - |
| 247 | rs35910094 | Synonymous coding | T/**G** | K | L | CT**A**, CT**C** | - | - |
| 263 | rs33996649 | Non-synonymous coding | C/T/**A**/G | N | R, L | C**G**G, C**T**G | tolerated | benign |
| 266 | rs72650670 | Non-synonymous coding | G/**A** | R | R, W | **C**GG, **T**GG | deleterious | probably damaging |
| 277 | rs72483511 | Stop gained, Splice site | C/**A** | M | E, * | **G**AA, **T**AA | - | - |
| 324 | rs113984534 | Synonymous coding | A/**G** | R | Y | TA**T**, TA**C** | - | - |
| 366 | rs74163654 | Synonymous coding | C/**T** | Y | E | GA**G**, GA**A** | - | - |
| 370 | rs72650671 | Non-synonymous coding | G/**T** | K | H, N | **C**AC, **A**AC | deleterious | possibly damaging |
| 381 | COSM25634 | Non-synonymous coding | CC/**TT** | - | LE, LK | CT**GG**AG, CT**AA**AG | - | - |

Podobno kot prej, lahko tabelo filtriramo, remikamo stolpce...

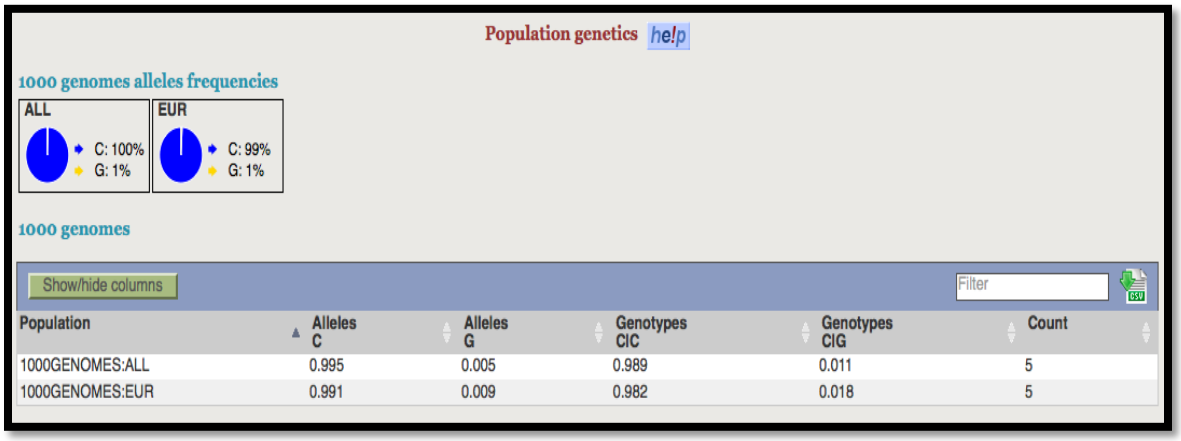
S kom na ID variacije se odpre »Variation tab«



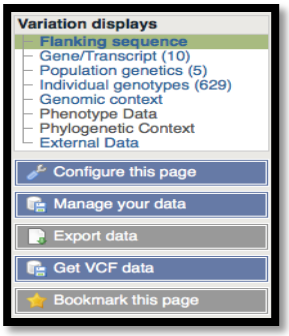

S klikom na “Population genetics” v levem meniju, dobimo podatke o alelnih frekvencah.




Frekvence za različne populacije, so tudi grafično prikazane.

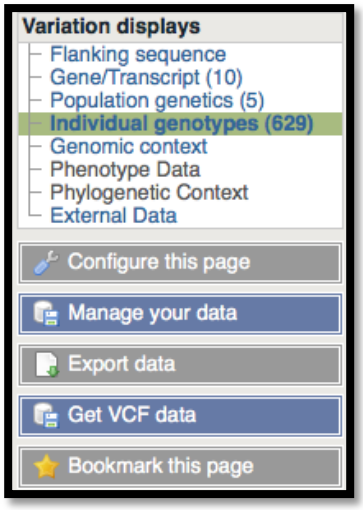


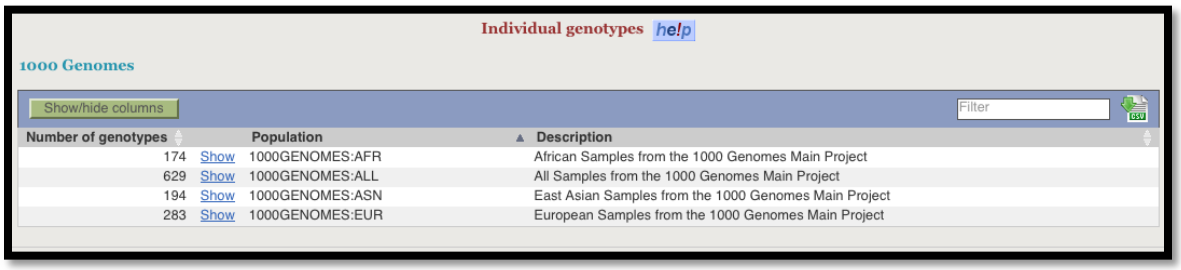
S klikom na "Flanking sequence" dobimo sekvenco okoli SNP-a:



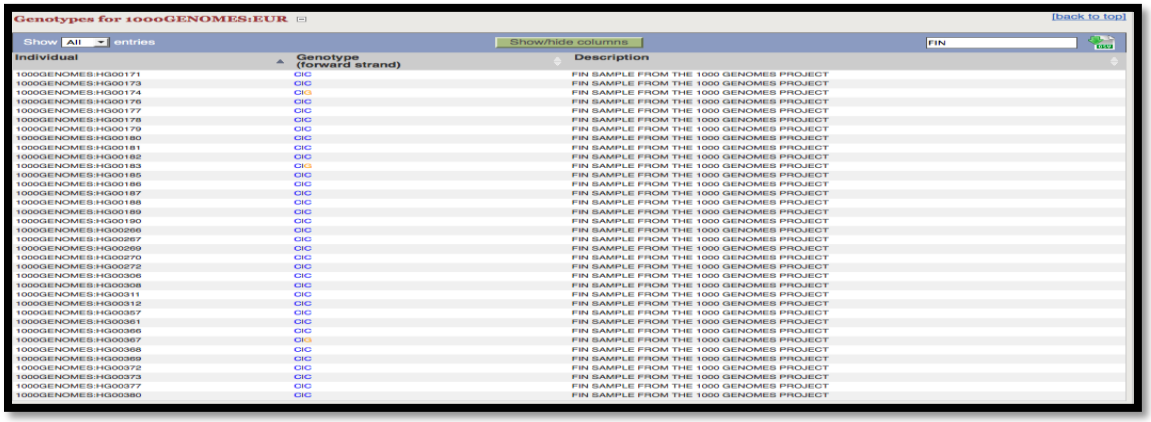


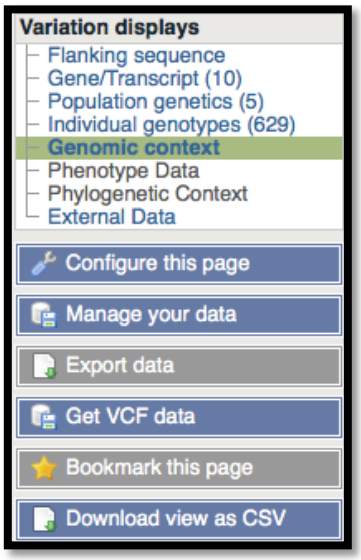
Klikni na "Individual genotypes" v levem meniju, da dobiš podatke o genotipih



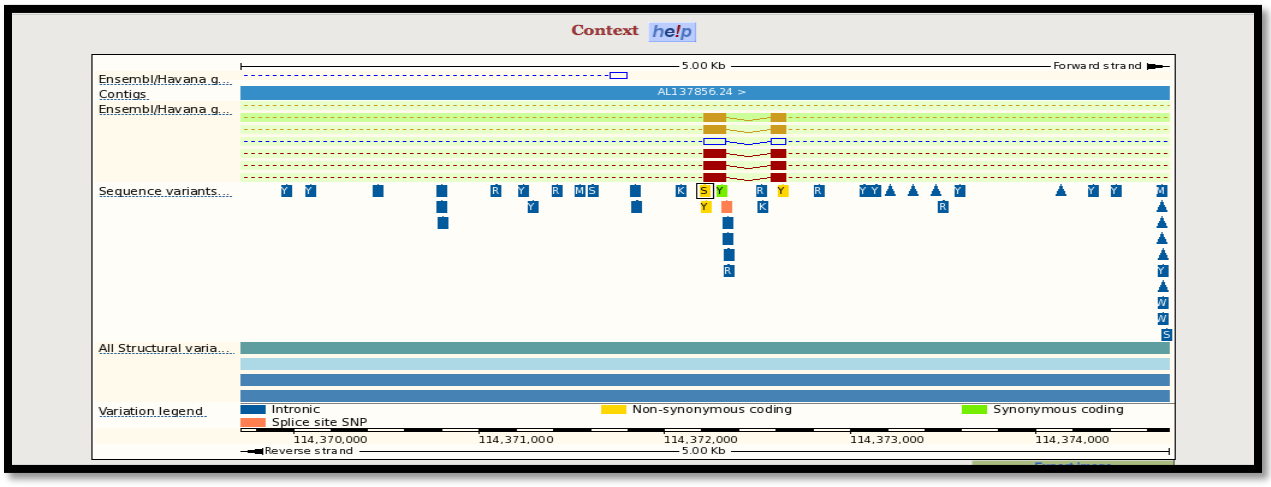



Klikni na “Show” da vidiš genotipe za vsak vzorec izbrane populacije:



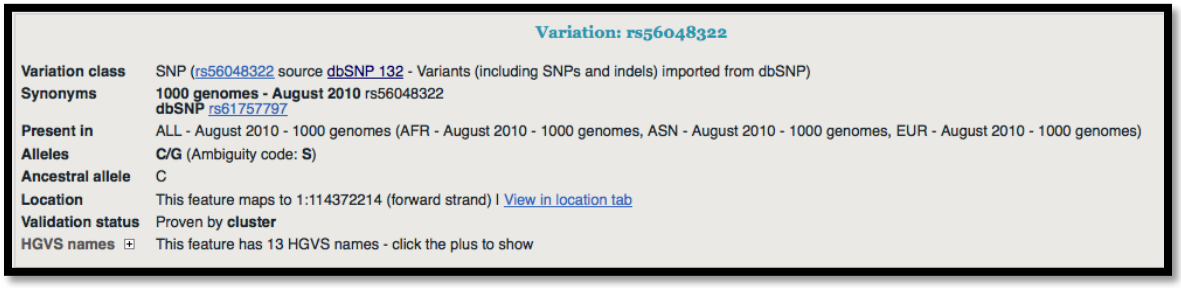




Klikni na “Genomic context” da vidiš variacije in gene v okolici izbrane variacije.

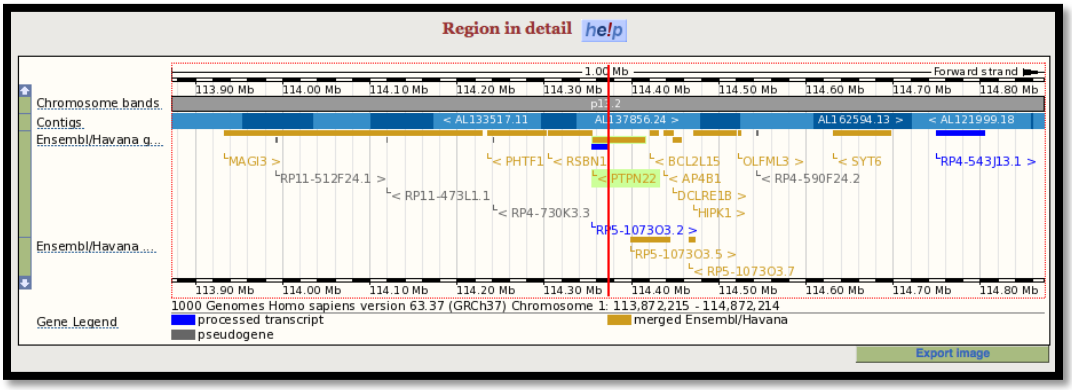


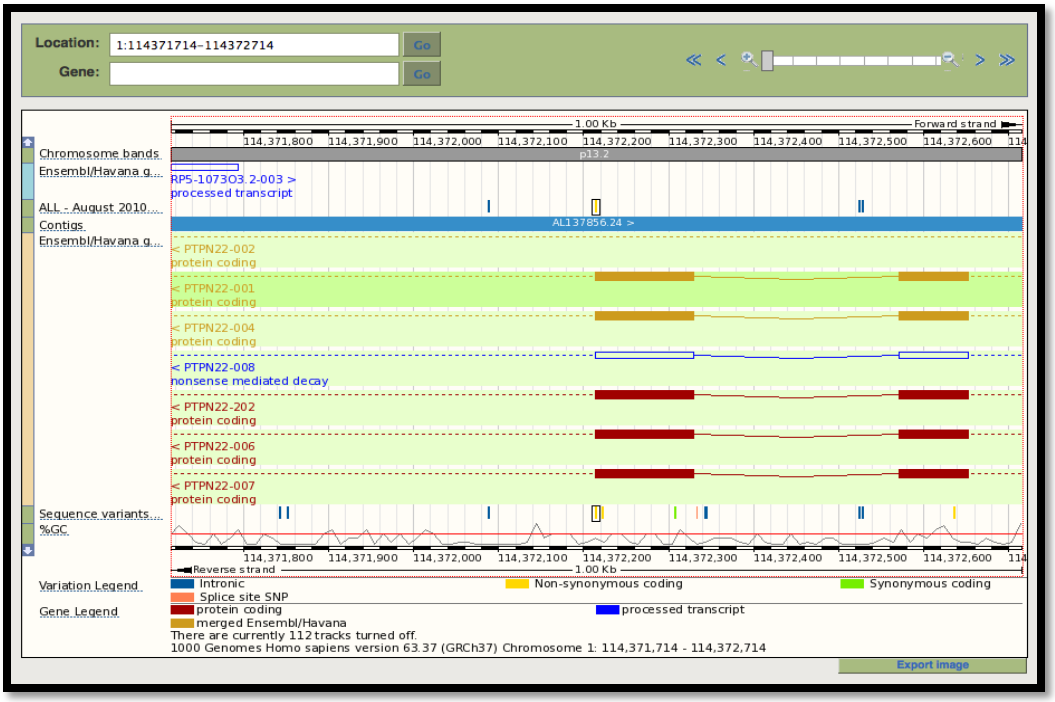
## 4.3 Dostop do podatkov o variacija na strani »Location View«

- V »variation summary« klikni na »view in location tab«

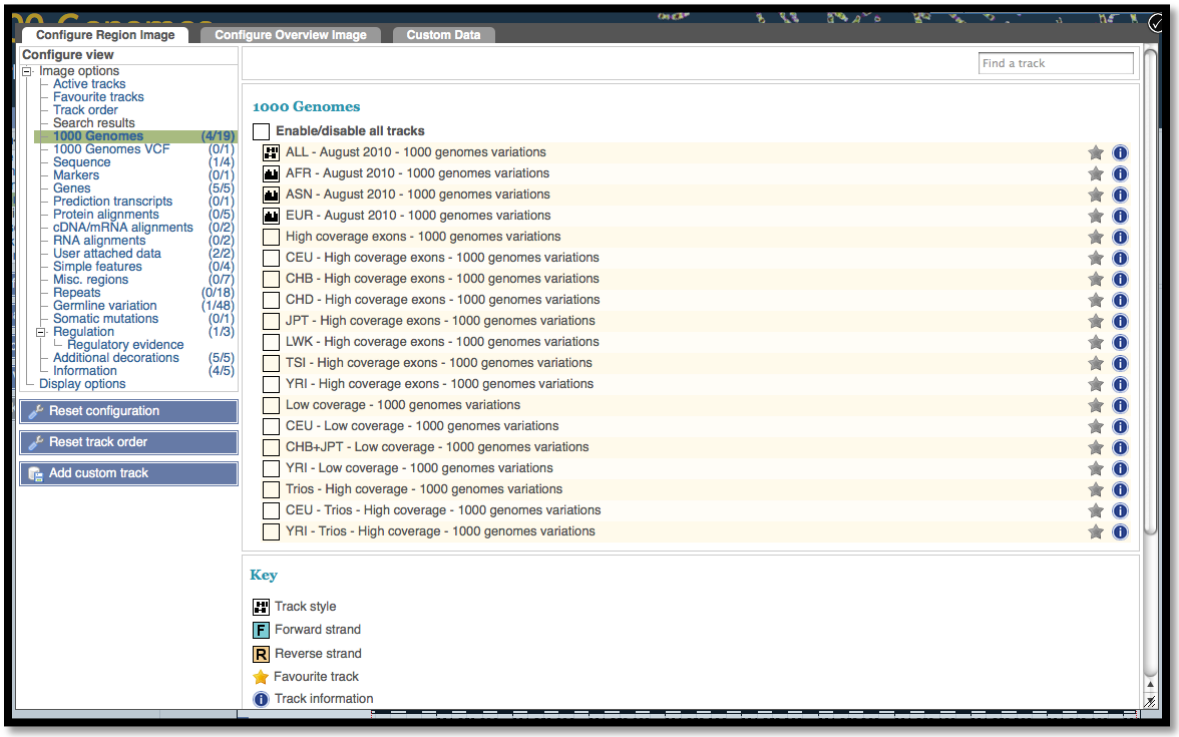



Odpre se okno »Location View« za regijo okoli variacije – izbrana variacija pa je označena z rdečim okvirjem.



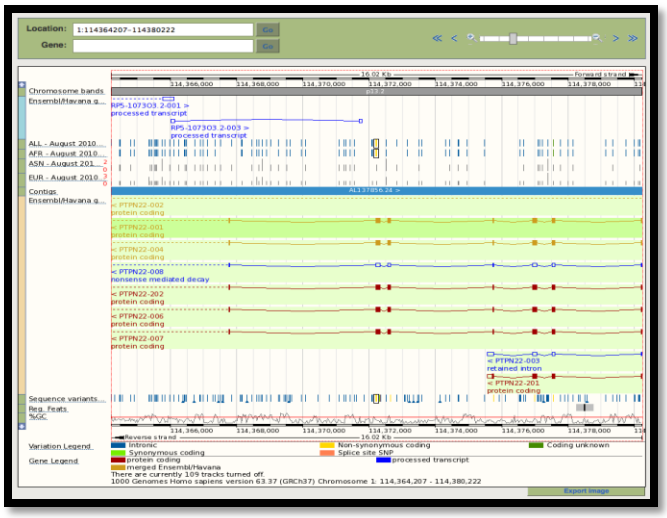



- Klikni na "Configure this page" v levem meniju in nastavi kaj vse želiš prikazati na sliki.



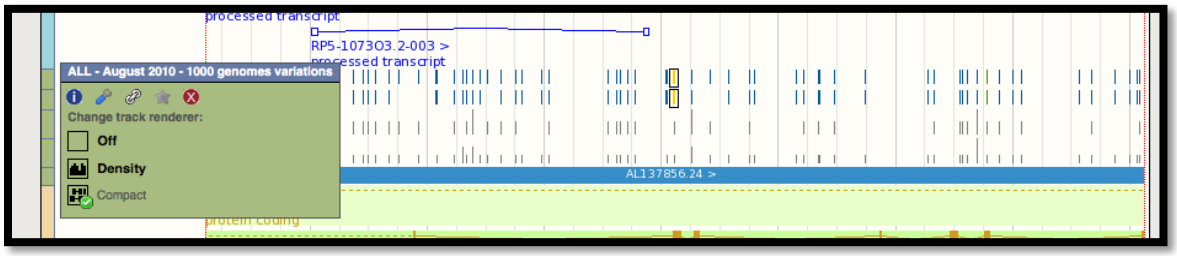

- V levem meniju izberi »1000 Genoms« in izberi nekaj možnosti (npr. EUR, ASN in AFR populacije)

Če želimo prikazati histogram (gostota SNP-ov) moramo izbrati še »density plot« in kliknite na kljukico v zgornjem desnem kotu.

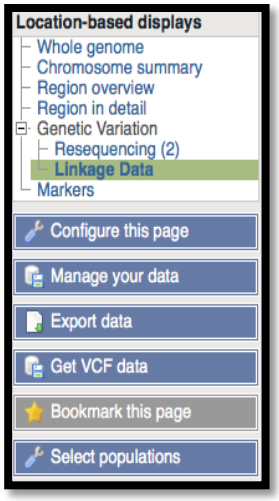


V »Location view« lahko prav tako vklopimo in izklopimo prikaz različnih podatkov s klikom na »Configuration page«

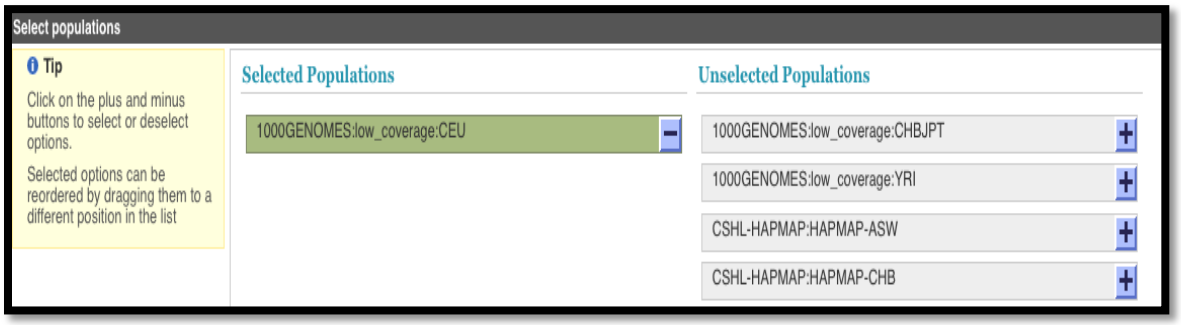
- Pojdi z miško imenom podatkov, in odpre se okno za spreminjanje nastavitev. V spodnjem primeru so "1000 genomes – August 2010 variations " lahko preklapljate med dvema vrstama pogleda.



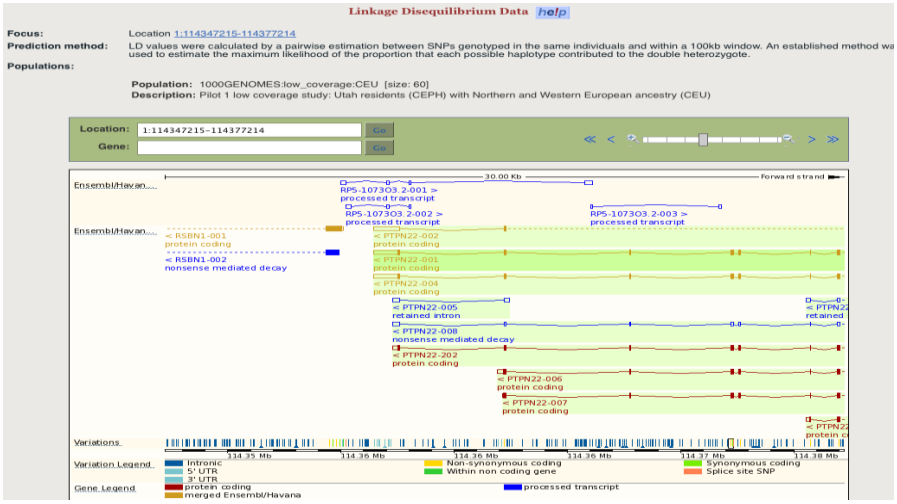
- Klikni na »Linkage Data« pod »Genetic Variation« v levem meniju




- Izberi populacijo s klikom na “Select Populations”




Podatki o vezavnem neravnovsju so prikazani na podlagi r2 in d’




Celozaslonski prikaz najdete na slednji povezavi:

http://browser.1000genomes.org/Homo_sapiens/Location/LD?vf=9395694&db=core&vdb=variation&v=rs56048322&pop1=34338&t=ENST00000359785&g=ENSG00000134242&r=1%3A114347215-114377214

***Naloga 10:*** Za CDKN2B in še en poljuben gen odgovori na sledeča vprašanja:

a. Koliko je transkriptov za ta gen? Koliko je velik daljši transkript? Koliko AK kodira?
b. Koliko je intronov? Koliko eksonov? Napiši prvih pet nukleotidov vsakega od njih.
c. Ali vsi transkripti kodirajo protein? Ali so kodirani na glavni verigi? Napiši CCDS kode za protein? Kakšna je NCBI oznaka tega proteina? Katere proteinske domene vsebuje?
d. Napiši lokacijo gena?
e. Izberi SNP lociran v eksonu gena ter izpiši alelne frekvence po posameznih populacijah.
f. Kaj lahko izveš o funkciji gena in z njim povezanih fenotipih?

# 5 VERIŽNA REAKCIJA S POLIMERAZO (PCR) IN POLIMORFIZEM DOLŽIN RESTRIKCIJSKIH FRAGMENTOV

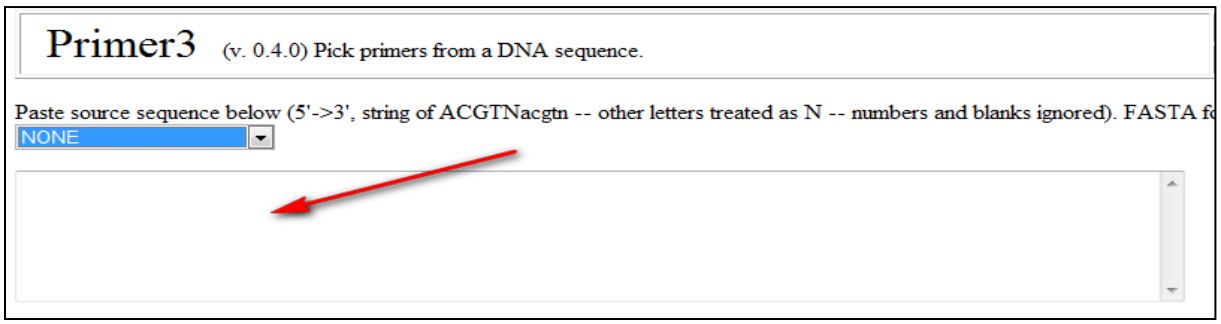
## 5.1 Načrtovanje začetnih oligonukleotidov

Začetne oligonukleotide za reakcijo PCR moramo skrbno načrtovati. Najpomembnejše točke in pravila za načrtovanje začetnih oligonukleotidov (ang. PRIMER):

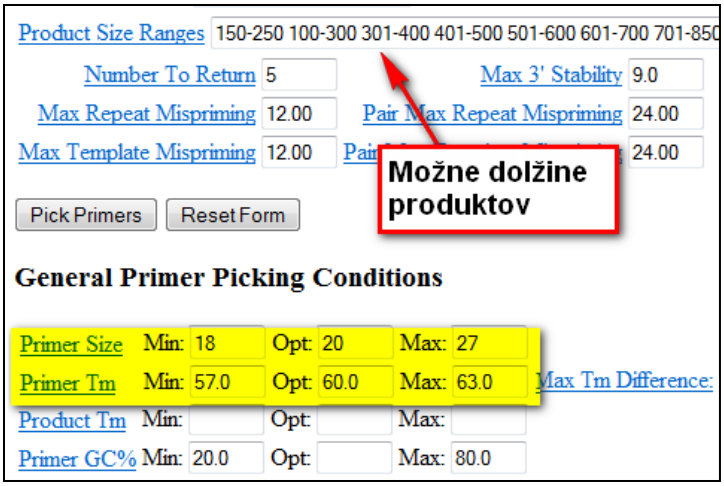
- Razlika pripenjale temperature med paroma primerjev ne sme biti večja kot 1°C;
- Primerja naj bosta dolga od 18 do 24 bp;
- Temperatura pripenjanja naj bo med 54 in 64°C;
- Delež baz G in C naj bo med 50 in 60%;
- Na 5' koncu naj primerja vsebujeta baze A in T;
- Na 3' koncu naj primerja vsebujeta vsaj 2-3 bazi G in C;
- Primerja ne smeta tvoriti sekundarnih struktur pri temperaturi višji od 35°C;
- Par primerjev ne sme biti komplementaren med seboj.

### 5.1.1 PRIMER3: spletno orodje za načrtovanje začetnih oligonukleotidov

V okence za sekvenco prilepimo sekvenco v FASTA formatu, ki smo jo dobili iz dbSNP ali iz katere druge podatkovne zbirke. Z oglatimi oklepaji pa v sekvenci določimo regijo, katero morata para primerjev zaobjemati oz. katero mora vsebovati produkt:




V nadaljnji fazi izberemo in nastavimo želene parametre, kot so temperatura pripenjanja, dolžina produkta, itd.:



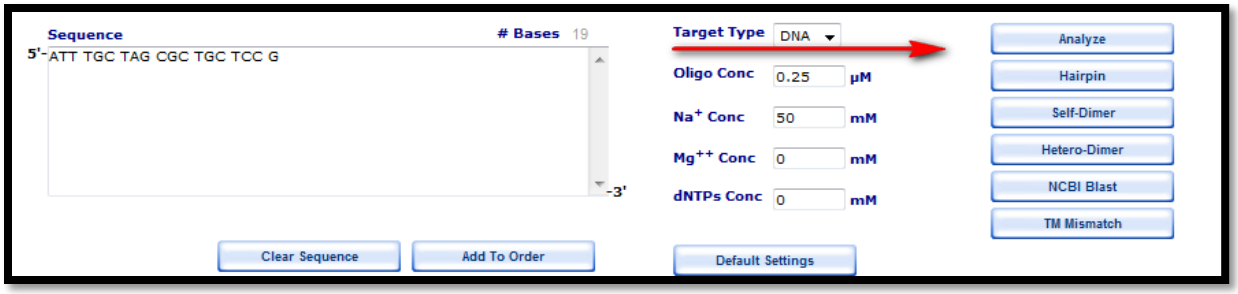

Po končani nastavitvi želenih parametrov kliknemo gumb pick primers in dobimo rezultate. Orodje primer 3 nam izbere več parov začetnih oligonukleotidov izmed katerih si sami izberemo najbolj ustrezen par:

```
ADDITIONAL OLIGOS
                      start  len      tm      gc%   any   3'  seq

 1 LEFT PRIMER          327   21   58.95    52.38  3.00  2.00 TTCAGACACCTACAGCCCTCT
   RIGHT PRIMER         546   20   60.59    40.00  3.00  0.00 aaatgccctcaagcaatgaa
   PRODUCT SIZE: 220, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00

 2 LEFT PRIMER          467   22   59.36    45.45  2.00  2.00 tcacacaagatactgctggaca
   RIGHT PRIMER         692   20   59.88    50.00  6.00  2.00 ccggccaggatattctcata
   PRODUCT SIZE: 226, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

 3 LEFT PRIMER          326   22   60.30    50.00  3.00  2.00 TTTCAGACACCTACAGCCCTCT
   RIGHT PRIMER         545   20   59.53    40.00  3.00  2.00 aatgccctcaagcaatgaat
   PRODUCT SIZE: 220, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00

 4 LEFT PRIMER          327   21   58.95    52.38  3.00  2.00 TTCAGACACCTACAGCCCTCT
   RIGHT PRIMER         547   20   59.26    40.00  3.00  2.00 taaatgccctcaagcaatga
   PRODUCT SIZE: 221, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00
```

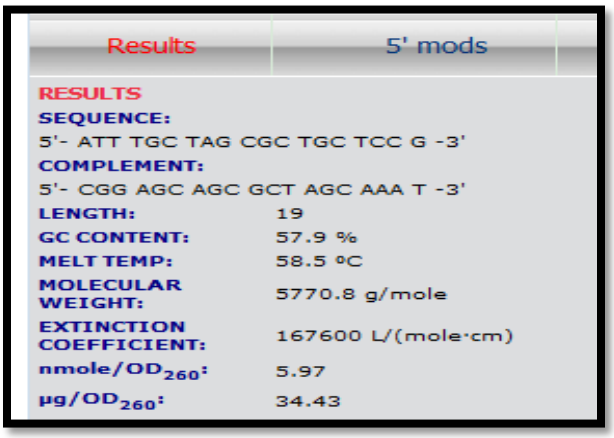
### 5.1.2 IDT Oligo analyzer: Spletno orodje za načrtovanje in analizo začetnih oligonukleotidov

S tem orodjem preverjamo začetne oligonukleotide, ki smo jih sami izbrali (izbrali ročno). Držati se moramo navodil za načrtovanje začetnih oligonukleotidov.

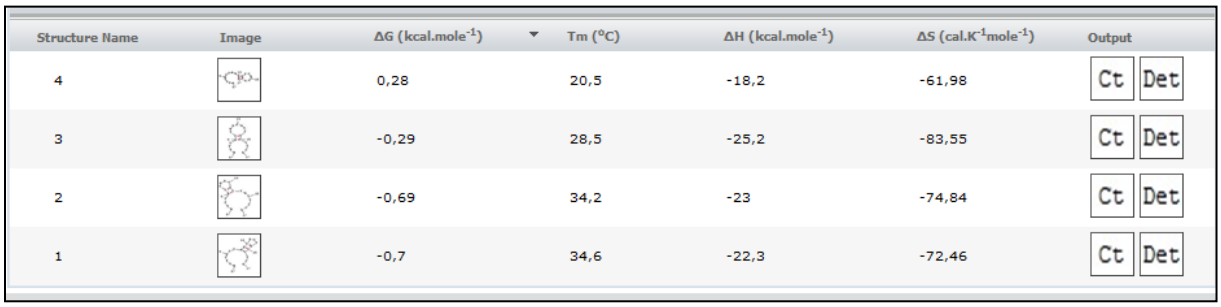
Del sekvence, ki smo jo izbrali za FW začetni oligonukleotid prilepimo v okence za analizo in kliknemo na analize:




Kot rezultat dobimo naslednje parametre:



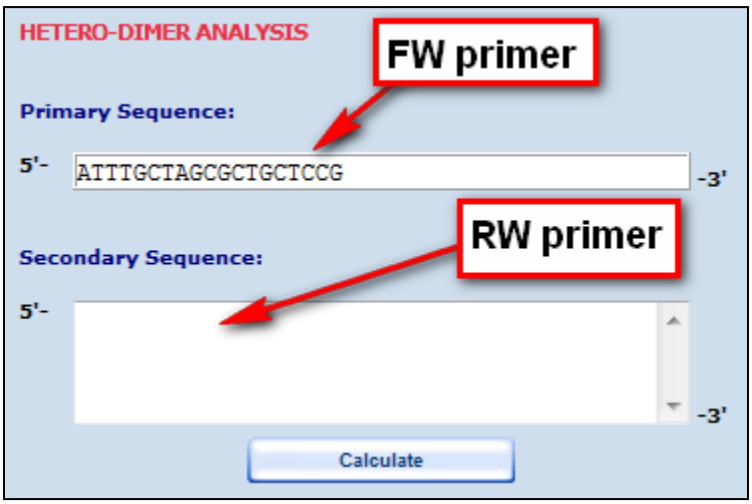

V naslednjem koraku preverimo tvorbo sekundarnih struktur, kliknemo na Hairpin in preverimo temperature pri katerih nastajajo sekundarne strukture:



| Structure Name | Image | ΔG (kcal.mole$^{-1}$) | Tm (°C) | ΔH (kcal.mole$^{-1}$) | ΔS (cal.K$^{-1}$mole$^{-1}$) | Output |
|---|---|---|---|---|---|---|
| 4 | | 0,28 | 20,5 | -18,2 | -61,98 | Ct Det |
| 3 | | -0,29 | 28,5 | -25,2 | -83,55 | Ct Det |
| 2 | | -0,69 | 34,2 | -23 | -74,84 | Ct Det |
| 1 | | -0,7 | 34,6 | -22,3 | -72,46 | Ct Det |

V kolikor nobena temperature ne presega 35°C se lotimo preverjanja Self Dimer. Tukaj preverimo, kako je primer komplementaren sam s seboj. Paziti moramo, da se tvori čim manj vezi in, da primerski homo dimer ni komplementaren na 3' koncih, saj ga v tem primeru polimeraza lahko podaljšuje:

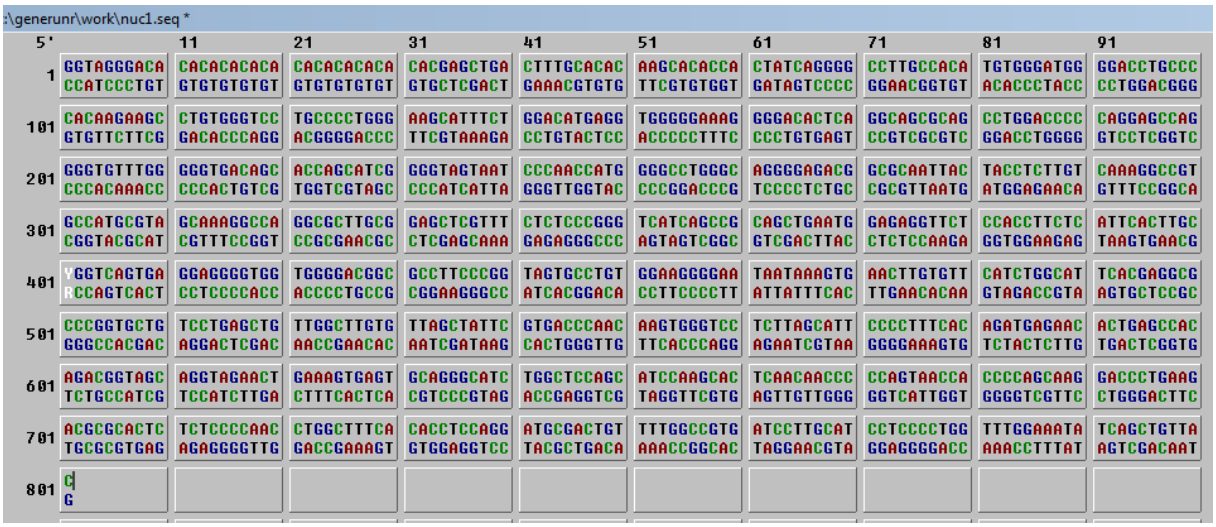

V testu za hetero dimer pa preverimo enako lastnost, vendar jo preverjamo z drugim primerjem iz našega primerskega para:



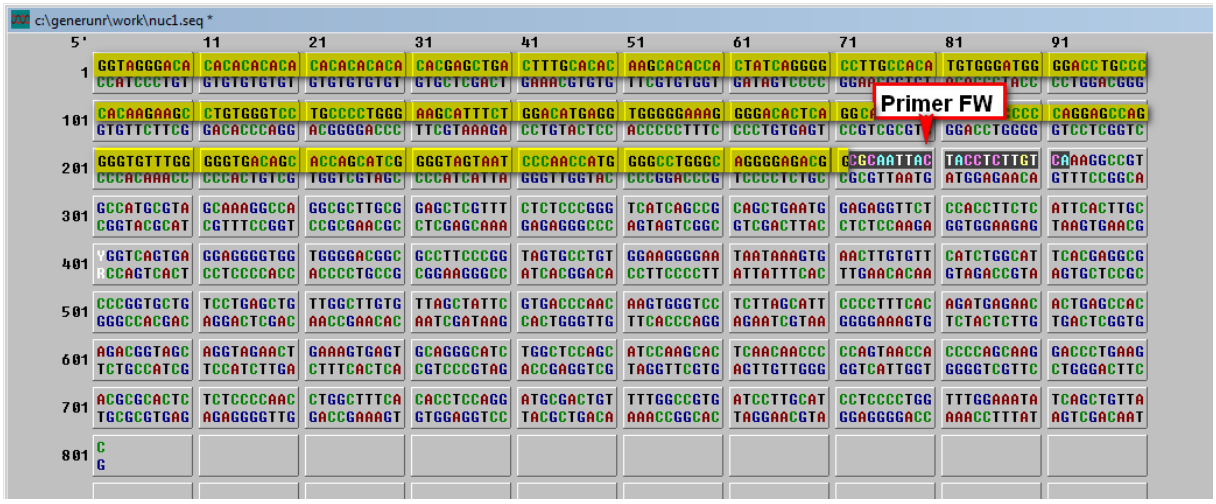

## 5.2 RFLP – restriction fragment length polymorphisms

### 5.2.1 GeneRunner

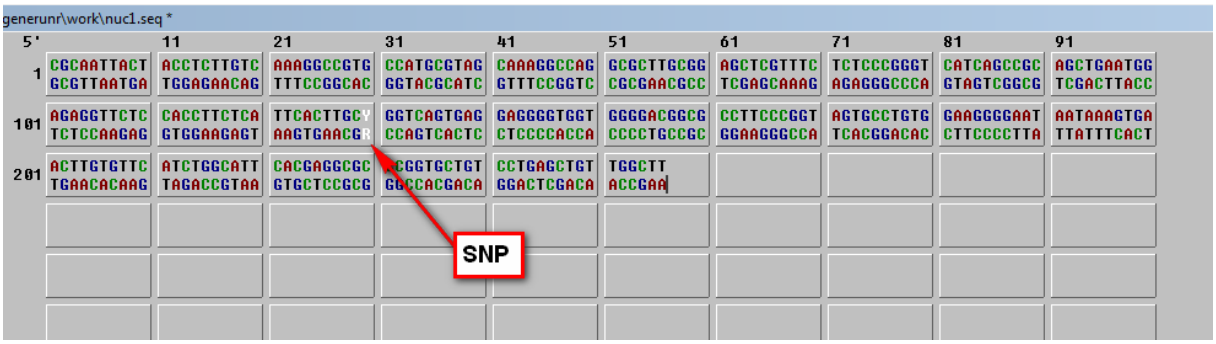
V prejšnji nalogi smo že načrtali začetne oligonukleotide za rs1248696. Sedaj bomo uporabili orodje GeneRunner in bomo vanj prilepili sekvenco SNP-ja rs1248696 ter jo shranili kot CONTEX sekvenco, prav tako pa jo bomo shranili tudi v Excelovo RFLP datoteko pod stolpec Contex sequence.



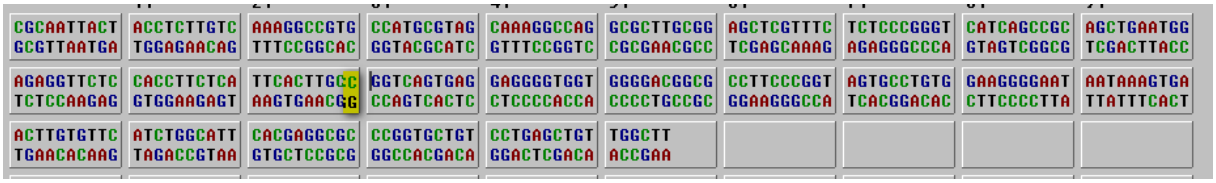

V naslednjem koraku bomo s funkcijo išči F5 poiskali naše začrtane začetne oligonukleotide. Ko najdemo pozicijo primerja FW, bomo izbrisali (rumeno označeno na sliki) ostalo sekvenco v smeri 5'. Enako bomo storili s primerjem RW, vendar bomo tukaj izbrisali sekvenco v smeri 3' ter dobljeno skrajšano sekvenco shranili kot amplicon sequence (to je hkrati naš produkt).



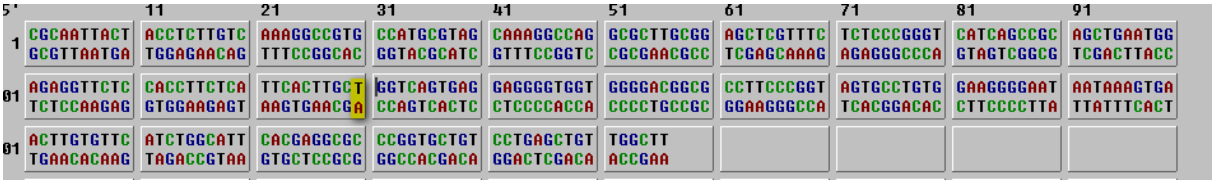

V amplikonski sekvenci bomo nato zamenjali Y → C, jo shranili pod imenom CCC ter ponovno odprli našo amplicon sekvenco in zamenjali Y → T ter ponovno shranili sekvenco pod imenom TTT.



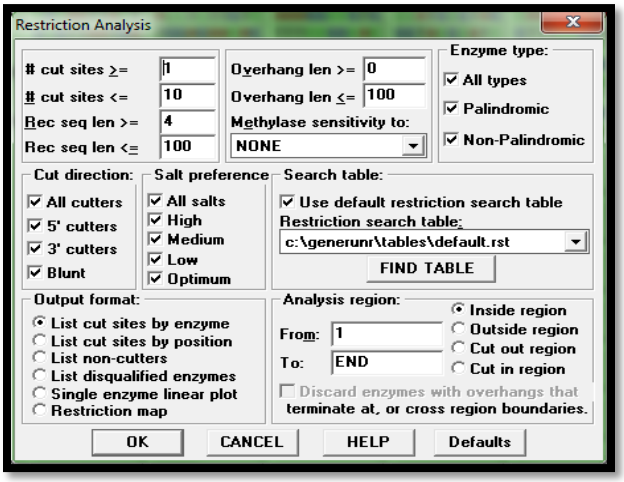

Y → C



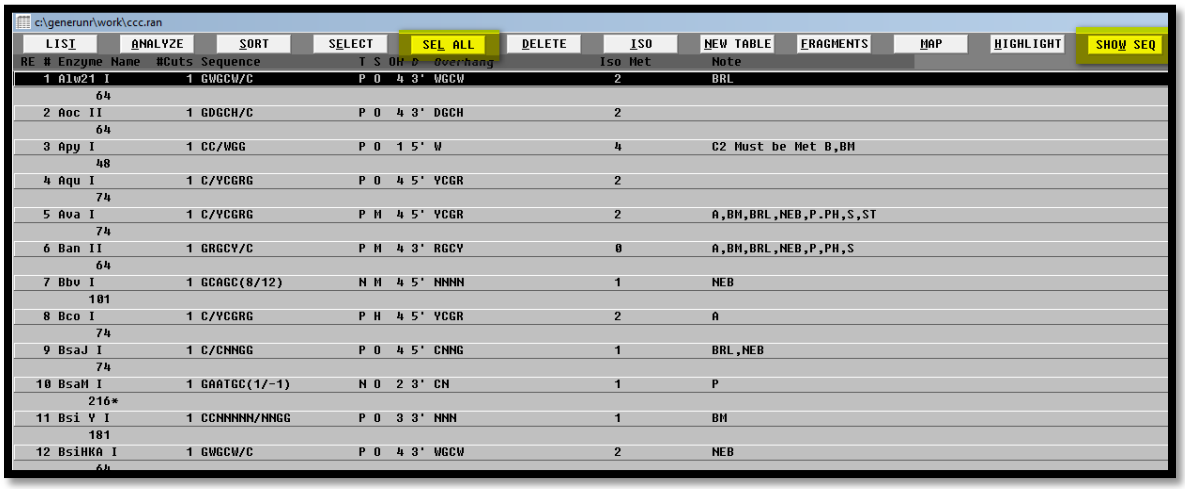
Y → T



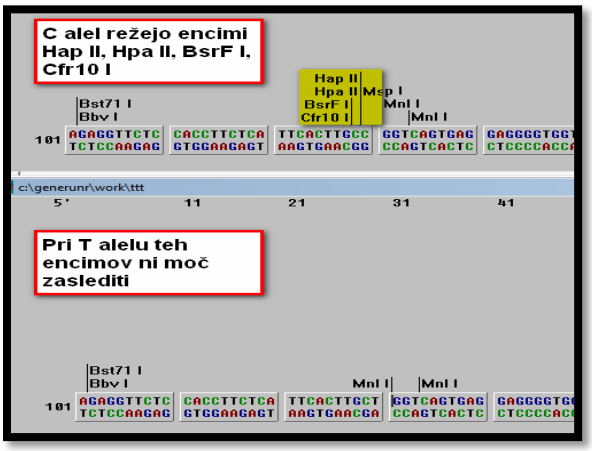
Odprli bomo ponovno sekvenco CCC. Izbrali bomo ukaz ANALYSIS → NUCLEIC ACID → RESTRICTION SITES in kliknili OK:



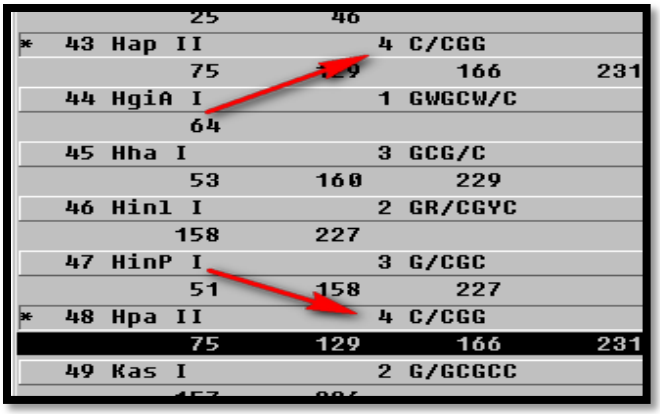
Prikazalo se nam bo informacijsko okno na katerem bomo spet potrdili izbiro s klikom na OK. V naslednjem meniju bomo izbrali SELECT ALL ter nato SHOW SEQUENCE. Enako bomo ponovili še za sekvenco TTT ter ju primerjali med seboj:



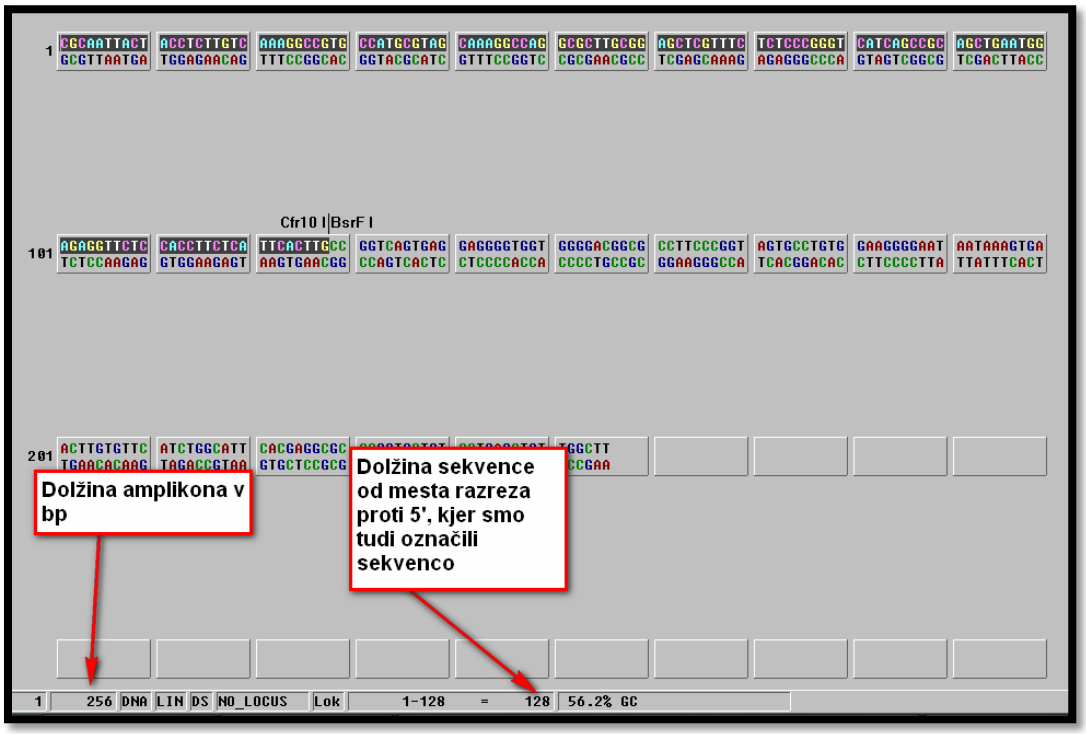

Iskali bomo na območju, kjer se nahaja SNP. Iščemo različne restrikcijske encime (RE), tako, da lahko s pomočjo restrikcijskega encima ločimo genotipsko različne sekvence med seboj:




Smotrno je tudi preveriti, kolikokrat kateri encim reže sekvenco. To lahko preverimo v meniju, kjer smo označili vse RE in pritisnili SHOW SEQUENCE. V našem primeru encima Hap II, Hpa II, Msp I režejo 4x, Cfr10 I ter BsrF I pa režete sekvenco 1x:

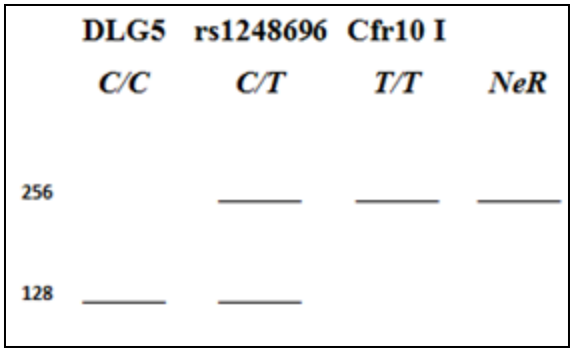



Na podlagi teh podatkov smo se odločili, da izberemo slednja dva encima. V enakem oknu nato označimo samo ta dva encima in ponovno kliknemo na SHOW SEQUENCE, kjer pogledamo na kakšne dolžine fragmentov nam ta RE razrežeta našo sekvenco:




Iz teh podatkov dobimo rezultat, ki nam pove, da sekvenco z alelom C RE razreže na fragmenta dolžine 128bp in 128bp, sekvenco z alelom T pa ta RE ne režeta.

Po končanem načrtovanju RFLP testa za ta SNP, vnesemo podatke v datoteko RFLP_VAJE. Rezultat prikažemo tudi grafično:

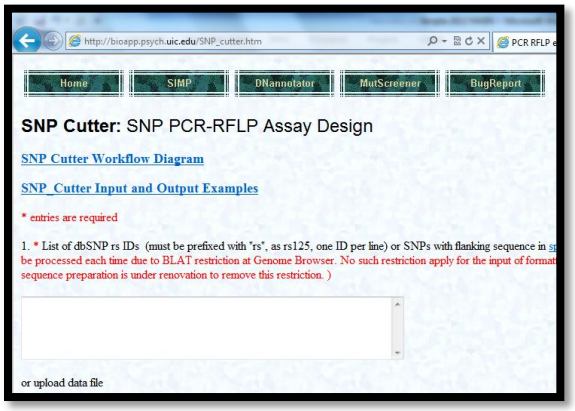


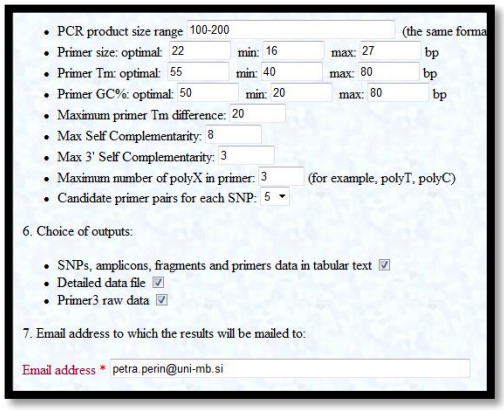

***Najmanjša dovoljena razlika med produkti je 20bp, saj v nasprotnem primeru ni mogoče ločiti fragmentov na agaroznem gelu.***

### 5.2.2 SNP Cutter

Tako začetne oligonukleotide, kot tudi izbira restrikcijskega encima je mogoča tudi z uporabo spletne aplikacije SNP Cutter, dostopne na naslovu: http://bioapp.psych.uic.edu/SNP_cutter.htm

V okence vpišemo ID SNP-a (rs številko), določimo parametre oz. pogoje za končni produkt ter vpišemo naš e-naslov, na katerega v nekaj minutah prejmemo sporočilo z rezultati.




***Naloga 11:*** Z uporabo spletnih aplikacij Primer3, SNP Cutter in GeneRunner dizajniraj RFLP teste za dodeljene SNP-e iz GWA študij. Za vsak par primer-jev preveri, če ustreza krterijem – uporabi program IDT oligo analyzer. Rezultate prikaži v tabeli, ki naj vsebuje naslednje stolpce:

- ID SNP-a (rs številka)
- Vrsta SNP-a (intronski, eksonski, UTR, izvengenski)
- ID Gena (na katerem SNP leži oz. v primeru izvengenskega SNP-a ID najbljižnega gena).
- FASTA sekvenca
- Forward primer sekvenca
- Reverse primer sekvenca
- Predvidena temperatura pripenjanja primerjev
- Sekvenca PCR produkta

- Velikost PCR produkta
- Restrikcijski encim
- Velikosti fragmentov po restrikciji
- Alelne frekvence HapMap-CEU

# 6 ASOCIACIJSKA ŠTUDIJA S PROGRAMSKIM PAKETOM SPSS

Pri asociacijskih študijah primerjamo genotipe oz. genotipske in alelne frekvence med dvema skupinama, najpogosteje so to bolniki in kontrolna skupina zdravih posameznikov. Pri izbiri vzorca je pomembno, da je fenotip posameznikov dobro določen, kar pomeni, da morajo zdravi posamezniki resnično biti zdravi ter obratno.

***Ključnega pomena je tudi skrbno urejena baza podatkov!***

Pri primerjavi genotipskih in alelnih frekvenc uporabljamo Fisherjev natančni test, ki je na razpolago v statističnem programskem orodju SPSS.

***Primer:*** *S programom Excel odpri datoteko Primer_za_vaje. Izračunaj genotipske in alelne frekvence bolnikov, in kontrolne skupine. Pripravi tabelo za statistično analizo s programom SPSS, ter rezultate genotipizacij statistično analiziraj.*

Tabela je sestavljena iz naslednjih stolpcev:

- ID pacienta;
- Diagnoza (1 = zdrav, 2 = bolan);
- Genotip/SNP;
- Model (recesiven za alel 1 in dominanten za alel 2) → 11/12+22;
- Model (recesiven za alel 2 in dominanten za alel 1) → 11+12/22;
- ID pacienta (aleli) (za primerjavo alelnih frekvenc);
- Diagnoza (aleli) (za primerjavo alelnih frekvenc);
- Alel → 1/2
- Pomožni stolpci za spremembo genotipa v alele.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID pacienta | diagnoza | Genotip rs3087243 | AA/AG+GG | AA+AG/GG | ID pacienta (aleli) | Diagnoza (aleli) | Alel rs3087243 | 1. alel | 2.alel |
| 2 | ASTMA 1 | 2 | AG | | | | | | | |
| 3 | ASTMA 2 | 2 | AG | | | | | | | |
| 4 | ASTMA 3 | 2 | GG | | | | | | | |
| 5 | ASTMA 4 | 2 | AG | | | | | | | |
| 6 | ASTMA 5 | 2 | AA | | | | | | | |
| 7 | ASTMA 6 | 2 | AG | | | | | | | |
| 8 | ASTMA 7 | 2 | AG | | | | | | | |
| 9 | ASTMA 8 | 2 | GG | | | | | | | |

## 6.1 Genotipske in alelne frekvence

Genotipske in alelne frekvence lahko določimo kar v Excelu. Z uporabo Excelovih formul izračunamo odstotek posameznega genotipa oz. posameznega alela v skupini bolnih, zdravih in v vseh vzorcih. Podatke zapišemo v tabeli:

| | ASTMA | KONTROLA | SKUPAJ |
|---|---|---|---|
| AA (n) | | | |
| AG (n) | | | |
| GG (n) | | | |
| skupaj (n) | | | |
| AA (%) | | | |
| AG (%) | | | |
| GG (%) | | | |
| A (%) | | | |
| G (%) | | | |

## 6.2 Točnost rezultatov: Hardy-Weinberg-ov zakon

Ko imamo genotipske in alelne frekvence našega vzorca, najprej preverimo, če se ujemajo s frekvencami v podatkovnih zbirkah (Npr. dbSNP). Vedno upoštevamo tiste frekvence, ki se nanašajo na evropsko populacijo (npr. HapMap CEU). Če se frekvence zelo razlikujejo od dobljenih pri našem eksperimentu, potem lahko sumimo, da smo delali nekaj narobe.

Frekvence genotipov v naravi niso naključne, temveč so genotipi v Hardy-Weinberg-ovem ravnovesju, kar v našem primeru pomeni, da če je frekvenca alela A = q in frekvenca alela G = p, potem je frekvenca genotipa AA = $q^2$ ; genotipa AG = 2pq ; in genotipa GG = $p^2$.

Z Excelovimi formulami izračunamo, koliko posameznikov v našem vzorcu bi naj imelo določen genotip ter izračunane podatke primerjamo z eksperimentalno dobljenimi. Za primerjavo lahko uporabimo katero izmed mnogih spletnih aplikacij, npr. http://quantpsy.org/chisq/chisq.htm, kamor vpišemo števila (n) posameznikov z določenim genotipom. P vrednost mora biti v tem primeru čim višja (med 0,5 in 1).

| | Gp 1 | Gp 2 | Gp 3 | Gp 4 | Gp 5 | Gp 6 | Gp 7 | Gp 8 | Gp 9 | Gp 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cond. 1: | | | | | | | | | | | |
| Cond. 2: | | | | | | | | | | | |
| Cond. 3: | | | | | | | | | | | |
| Cond. 4: | | | | | | | | | | | |
| Cond. 5: | | | | | | | | | | | |
| Cond. 6: | | | | | | | | | | | |

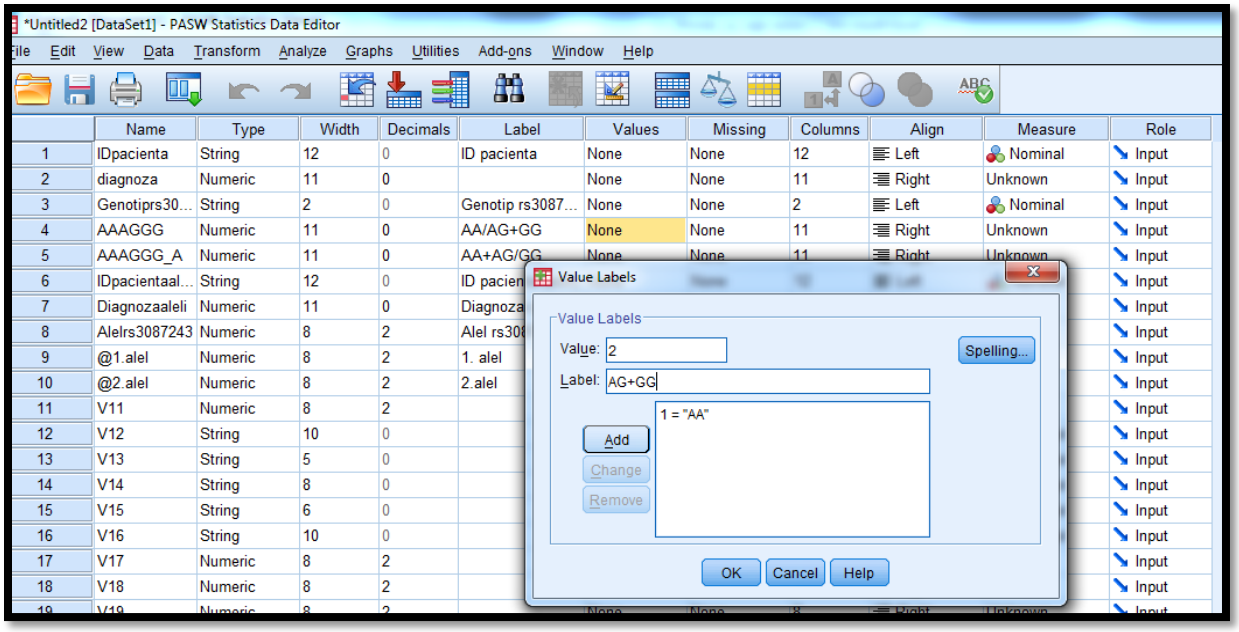
## 6.3 Primerjava genotipskih frekvenc

Za rs3087243 sta možna alela A in G. Statistično bomo genotipe primerjali na tri načine. Najprej bomo privzeli, da je model G dominanten, zato bomo med skupinama (bolani in zdravi) primerjali frekvence genotipa AA proti AG in GG (AA/AG+GG). Genotipu AA bomo priredili vrednost 1, AG in GG pa vrednost 2. V drugem modelu bomo privzeli, da je dominanten alel A (AA+AG/GG), zato bomo genotipoma AA in AG priredili vrednost 1 ter genotipu GG vrednost 2. Za pretvorbo si pomagaj s funkcijo IF v Excelu.

=IF(C2="AA";1;IF(C2="AG";2;IF(C2="GG";2)))

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | ID pacienta | diagnoza | Genotip rs3087243 | AA/AG+GG | AA+AG/GG |
| 2 | ASTMA 1 | 2 | AG | 2 | 1 |
| 3 | ASTMA 2 | 2 | AG | 2 | 1 |
| 4 | ASTMA 3 | 2 | GG | 2 | 2 |
| 5 | ASTMA 4 | 2 | AG | 2 | 1 |
| 6 | ASTMA 5 | 2 | AA | 1 | 1 |
| 7 | ASTMA 6 | 2 | AG | 2 | 1 |
| 8 | ASTMA 7 | 2 | AG | 2 | 1 |
| 9 | ASTMA 8 | 2 | GG | 2 | 2 |

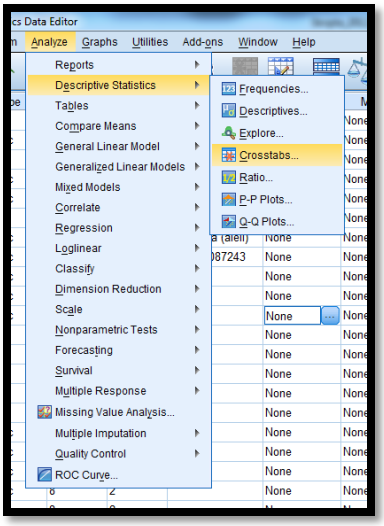
Tako pripravljeno tabelo odpremo v programu SPSS.

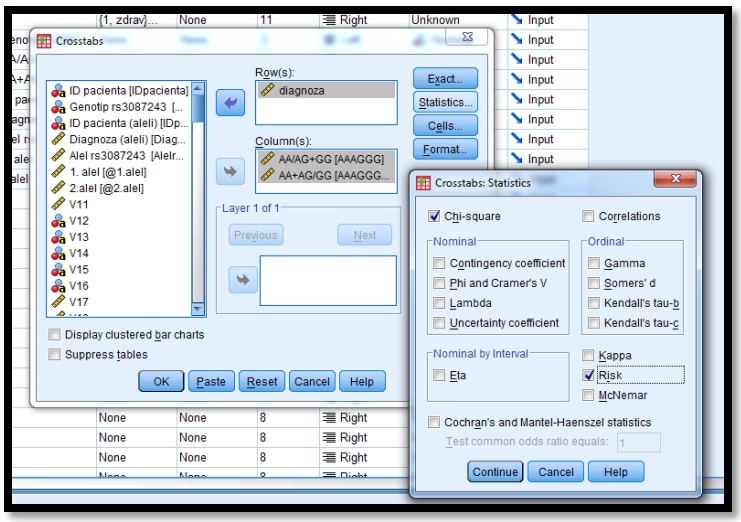



| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IDpacienta | String | 12 | 0 | ID pacienta | None | None | 12 | Left | Nominal | Input |
| 2 | diagnoza | Numeric | 11 | 0 | | None | None | 11 | Right | Unknown | Input |
| 3 | Genotiprs30... | String | 2 | 0 | Genotip rs3087... | None | None | 2 | Left | Nominal | Input |
| 4 | AAAGGG | Numeric | 11 | 0 | AA/AG+GG | None | None | 11 | Right | Unknown | Input |
| 5 | AAAGGG_A | Numeric | 11 | 0 | AA+AG/GG | None | None | 11 | Right | Unknown | Input |
| 6 | IDpacientaal... | String | 12 | 0 | ID pacien | | | | | | Input |
| 7 | Diagnozaaleli | Numeric | 11 | 0 | Diagnoza | | | | | | Input |
| 8 | Alelrs3087243 | Numeric | 8 | 2 | Alel rs30 | | | | | | Input |
| 9 | @1.alel | Numeric | 8 | 2 | 1. alel | | | | | | Input |
| 10 | @2.alel | Numeric | 8 | 2 | 2.alel | | | | | | Input |
| 11 | V11 | Numeric | 8 | 2 | | | | | | | Input |
| 12 | V12 | String | 10 | 0 | | | | | | | Input |
| 13 | V13 | String | 5 | 0 | | | | | | | Input |
| 14 | V14 | String | 8 | 0 | | | | | | | Input |
| 15 | V15 | String | 6 | 0 | | | | | | | Input |
| 16 | V16 | String | 10 | 0 | | | | | | | Input |
| 17 | V17 | Numeric | 8 | 2 | | | | | | | Input |
| 18 | V18 | Numeric | 8 | 2 | | | | | | | Input |

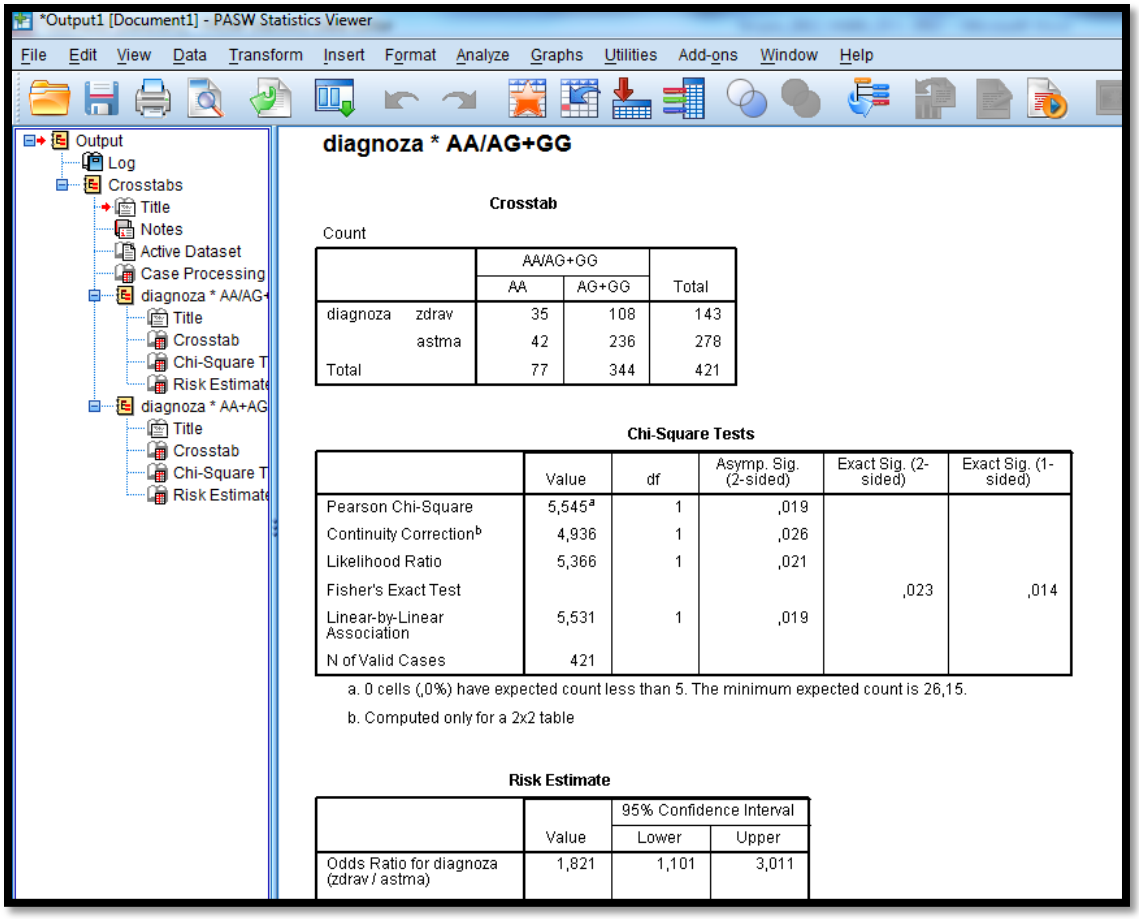
Za boljši pregled vnesemo kaj predstavljajo numerične vrednosti v posameznih stolpcih.

Nato genotipe med zdravimi in bolnimi primerjamo s Chi-squere test-om.




Odčitamo rezultat:




Ugotovili smo, da je SNP rs3087243 povezan z astmo in sicer je vrednost p = 0,023, kar pomeni statistično značilno razliko. To razliko smo ugotovili pri primerjavi genotipov po dominantnem modelu za alel G.

Iz prve tabele (Crosstab), ter še bolj iz izračunanih genotipskih frekvenc, je razvidno, da je tistih, ki imajo genotip AG ali GG, značilno višji delež v skupini astmatikov (236 od 278 = 85%) v primerjavi s kontrolno skupino (108 od 143 = 76%).

OR vrednost (Odds Ratio) znaša OR=1,821, in nam pove, kolikšen je vpliv oz. vloga SNP-a pri preučevani bolezni.

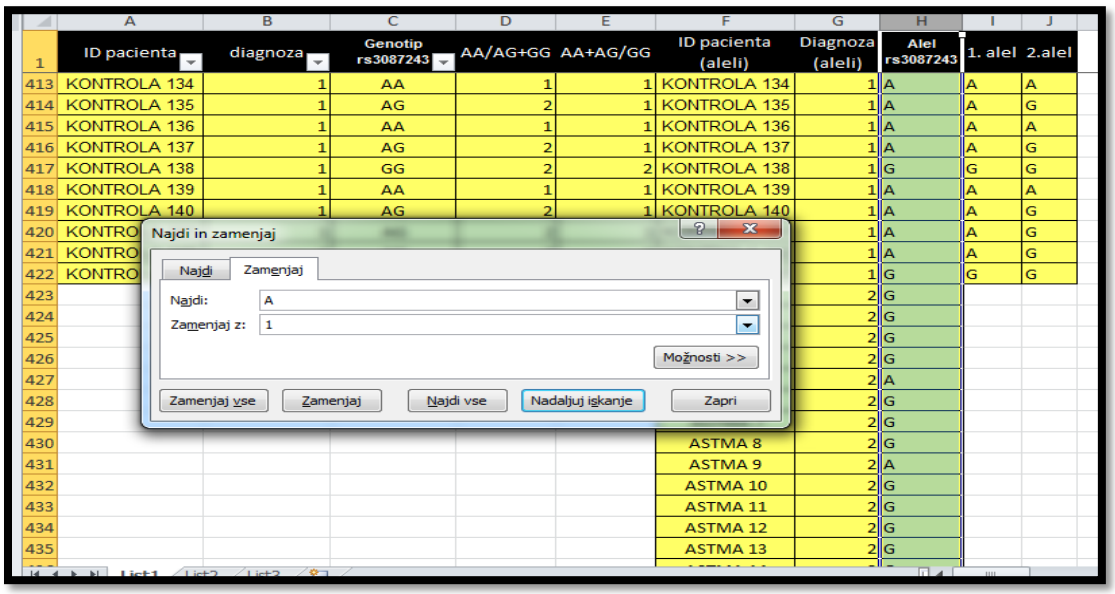
## 6.4 Primerjava alelnih frekvenc

Ker alela pogosto delujeta ko-dominantno, je nujna tudi primerjava med aleli oz. alelnimi frekvencami obeh skupin. V tem primeru predpostavimo, da imata pri heterozigotih oba alela enako velik vpliv.

Najprej moramo v Excelu pripraviti ustrezno tabelo. Podvojimo vrednosti v stolpcu diagnoze (zdrav/bolan oz. 1/2), zaradi večje preglednosti pa tudi ID vzorcev. Prvi stolpec zato dvakrat prekopiramo ter prilepimo enega pod drugim v stoplca »ID pacienta aleli« in »Alel rs…«.

Nato moramo vsakemu vzorcu dopisati alel in sicer dvakrat – k eni kopiji prvega in k drugi kopiji drugega. To lahko naredimo na več načinov. Lahko naredimo dva dodatna stolpca in v prvem navedemo 1. Alel ter v drugem 2. Alel. Tako dobimo en genotip zapisan v dveh stolpcih, ki ju prekopiramo enega pod drugega, tako kot smo to predhodno naredili s stolpcema »ID-pacienat aleli« in »Diagnoza-aleli«.

| GENOTIP | 1. Alel | 2. Alel |
|---|---|---|
| AA | A | A |
| AG | A | G |
| GG | G | G |




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID pacienta | diagnoza | Genotip rs3087243 | AA/AG+GG | AA+AG/GG | ID pacienta (aleli) | Diagnoza (aleli) | Alel rs3087243 | 1. alel | 2.alel |
| 413 | KONTROLA 134 | 1 | AA | 1 | 1 | KONTROLA 134 | 1 | A | A | A |
| 414 | KONTROLA 135 | 1 | AG | 2 | 1 | KONTROLA 135 | 1 | A | A | G |
| 415 | KONTROLA 136 | 1 | AA | 1 | 1 | KONTROLA 136 | 1 | A | A | A |
| 416 | KONTROLA 137 | 1 | AG | 2 | 1 | KONTROLA 137 | 1 | A | A | G |
| 417 | KONTROLA 138 | 1 | GG | 2 | 2 | KONTROLA 138 | 1 | G | G | G |
| 418 | KONTROLA 139 | 1 | AA | 1 | 1 | KONTROLA 139 | 1 | A | A | A |
| 419 | KONTROLA 140 | 1 | AG | 2 | 1 | KONTROLA 140 | 1 | A | A | G |
| 420 | KONTRO | | | | | | 1 | A | A | G |
| 421 | KONTRO | | | | | | 1 | A | A | G |
| 422 | KONTRO | | | | | | 1 | G | G | G |
| 423 | | | | | | | 2 | G | | |
| 424 | | | | | | | 2 | G | | |
| 425 | | | | | | | 2 | G | | |
| 426 | | | | | | | 2 | G | | |
| 427 | | | | | | | 2 | A | | |
| 428 | | | | | | | 2 | G | | |
| 429 | | | | | | | 2 | G | | |
| 430 | | | | | | ASTMA 8 | 2 | G | | |
| 431 | | | | | | ASTMA 9 | 2 | A | | |
| 432 | | | | | | ASTMA 10 | 2 | G | | |
| 433 | | | | | | ASTMA 11 | 2 | G | | |
| 434 | | | | | | ASTMA 12 | 2 | G | | |
| 435 | | | | | | ASTMA 13 | 2 | G | | |

Oznake alelov nato še pretvorimo v numerične vrednosti (npr. A=1 / G=2) ter tabelo odpremo v SPSSu, kjer podatke analiziramo na enak način kot predhodno genotipe. Vse podatke nato zberemo v tabeli.

***Naloga 12:*** *Kronično vnetno črevesno bolezen (KVČB) delimo na dva podtipa – ulcerozni kulitis (UC) in Crohnovo bolezen (CB). Gre za kompleksno bolezen, k nastanku katere prispeva delovanje mnogih genov oz. polimorfizmov. Nekateri kandidatni polimorfizmi so skupni obema podtipoma, drugi pa vplivajo na razvoj le ene izmed obeh oblik.*

*Da bi potrdili povezavo nekaterih kandidatnih genov, za katere je bilo predhodno ugotovljeno, da vplivajo na nastanek KVČB ali katerega od pod-tipov, smo izvedli analizo polimorfizmo v teh genih v slkupini bolnikov in skupini zravih posameznikov.*

*V dodeljeni excel-ovi tabeli se nahajajo podatki o genotipih. Izberi si 2 različna polimorfizma in preveri, če lahko povezavo s KVČB potrdiš? Najprej preveri, genotipske frekvence v Hardly-Weinbergovem ravnovesju (p>0,5)! Nato izvedi asociasijsko študijo tako, da primerjaš genotipske in alelne frekvence med bolniki in zdravimi posamezniki? Izračunaj p vrednosti z uporabo Chi-square testa!*

*Analizo ponovi za oba pod-tipa bolezni (z zdravimi posamezniki primerjaj vsako skupino posebej)!*

*Rezultate podaj v tabeli:*

| | Genotipi [%] | AA/AG+GG | AA+AG/GG | Aleli (%) | A/G |
|---|---|---|---|---|---|
| KVČB (n=__) | AA: __,_<br>AG: __,_<br>GG: __,_ | p= _,___ | p= _,___ | A: __,_<br>G: __,_ | p= _,___ |
| CB (n=__) | AA: __,_<br>AG: __,_<br>GG: __,_ | p= _,___ | p= _,___ | A: __,_<br>G: __,_ | p= _,___ |
| UC (n=__) | AA: __,_<br>AG: __,_<br>GG: __,_ | p= _,___ | p= _,___ | A: __,_<br>G: __,_ | p= _,___ |
| Zdravi (n=__) | AA: __,_<br>AG: __,_<br>GG: __,_ | | | A: __,_<br>G: __,_ | |

*Rezultate komentiraj.*

# 7 LITERATURA


1. Arthur M. Lesk: Introduction to Bioinformatics, 2. izdaja, 2005, Oxford University Press
2. Jin Xiong: Essential Bioinformatics, 2009, Cambridge University Press
3. Darren George, Paul Mallery: IBM SPSS Statistics 19 Step by Step: A Simple Guide and Reference, 2011, Pearson Education (US)
4. http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview