Računalna analiza metagenoma probavnog sustava bolesnika s cirozom jetre

Fabijanić, Maja

Sveučilište u Zagrebu
Prirodoslovno - matematički fakultet
Biološki odsjek

Maja Fabijanić

Računalna analiza metagenoma probavnog sustava bolesnika s cirozom jetre

Diplomski rad

Zagreb, 2015.

Ovaj rad, izrađen pri Zavodu za molekularnu biologiju, pod vodstvom prof. dr. sc. Kristiana Vlahovičeka, predan je na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistra molekularne biologije.

Zahvaljujem profesoru Kristianu Vlahovičeku na mentorstvu, utrošenom vremenu i savjetima pri izradi rada.Hvala svim asistentima bioinfo grupe na pristupačnosti, a posebno hvala Maši na pomoći s prvim koracima bez kojih ne bi bilo ovog rada.

Zahvaljujem roditeljima i sestri na strpljenju i beskonačnoj potpori.

Hvala Borisu što mi uljepšava život!

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek
Diplomski rad

RAČUNALNA ANALIZA METAGENOMA PROBAVNOG SUSTAVA BOLESNIKA S CIROZOM JETRE

Maja Fabijanić
Zavod za molekularnu biologiju, Horvatovac 102a, 10000 Zagreb, Hrvatska

Prokarioti zauzimaju dvije od tri domene života, a samo se 1-5% bakterija može uzgojiti u laboratorijskim uvjetima. Metabolizam čovjeka ne bi bio potpun bez metabolizma mikroba koji ga nastanjuju, a oni imaju utjecaja i u raznim bolestima kao što su pretilost, upalna bolest crijeva, simptomatična ateroskleroza i bolesti jetre. Metagenomika je pristup koji nam omogućava proučavanje mikrobnih genoma uzorkovanjem izravno iz okoliša, nakom čega analizom obrazaca upotrebe sinonimnih kodona možemo odrediti gene optimirane za translaciju što je u korelaciji s ekspresijom tih gena. Na ovaj način odredila sam razinu translacijske optimizacije gena u crijevnim metagenomima bolesnika s cirozom jetre i usporedila ju sa onom zdravih pojedinaca. Gene sam prema razini translacijske optimizacije i ishodišnom uzorku klasificirala u skupine povezane sa zdravim i bolesnim fenotipom, a zatim i razvrstala u pripadne metaboličke putove. Nejednaka zastupljenost translacijski optimiranih gena u različitim metaboličkim putovima crijevnih mikrobnih zajednica zdravih i bolesnih pojedinaca otvara mogućnost za olakšanu djagnostiku ciroze jetre ali i mehanistički uvid u interakciju između mikrobnog i ljudskog metabolizma u razvoju bolesti.

(41 stranica, 13 slika, 4 tablice, 42literaturnih navoda, jezik izvornika: hrvatski)

Rad je pohranjen u Središnjoj biološkoj knjižnici

Ključne riječi: mikrobne zajednice, metagenom, optimizacija translacije, upotreba kodona, MILC, MELP

Voditelj: Prof. dr. sc. Kristian Vlahoviček

Ocjenitelji: Prof. dr. sc. Kristian Vlahoviček
Doc. dr. sc. Damjan Franjević
Izv. prof. dr. sc. Dijana Škorić

Zamjena: Izv. prof. dr. sc. Dunja Leljak-Levanić
Rad prihvaćen: 16. lipnja, 2015.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Division of Biology
Graduation Thesis

COMPUTATIONAL ANALYSIS OF METAGENOMES OF THE INTESTINAL TRACT IN PATIENTS WITH LIVER CIRRHOSIS

Maja Fabijanić
Department of Molecular Biology, Horvatovac 102a, 10000 Zagreb, Croatia

Prokaryotes occupy two of the three domains of life, and only 1-5% of bacteria can be grown in laboratory conditions. Human metabolism would not be complete without metabolism of microbes that inhabit it, and they have an impact in various diseases such as obesity, inflammatory bowel disease, symptomatic atherosclerosis and liver disease. Metagenomics is an approach that allows us to study the microbial genome by sampling directly from the environment, followed by analyzing patterns of use of synonymous codons. In this way I determined the level of translational optimization of genes in intestinal metagenomes of cirrhotic patients and compared it with that of healthy individuals. Based on levels of translational optimization and initial sample, I have classified genes into groups associated with healthy and diseased phenotype, and then sorted them in their corresponding metabolic pathways. Unequal representation of translationally optimized genes in different metabolic pathways of intestinal microbial communities of healthy and sick individuals opens up a possibility for easier diagnosis of liver cirrhosis as well as mechanistic insight into the interaction between microbial and human metabolism in development of this disease.

(41 pages, 13 figures, 4 tables, 42 references, original in: Croatian)

Thesis deposited in the Central Biological Library

Key words: microbes, optimization of translation, codon usage, MILC,MELP

Supervisor: Professor Kristian Vlahoviček, PhD

Reviewers: Professor Kristian Vlahoviček, PhD
Asst. Prof Damjan Franjević, PhD
Assoc. Prof Dijana Škorić, PhD

Substitution: Assoc. Prof Dunja Leljak-Levanić, PhD

Thesis accepted: June 16, 2015.

Sadržaj

1. UVOD

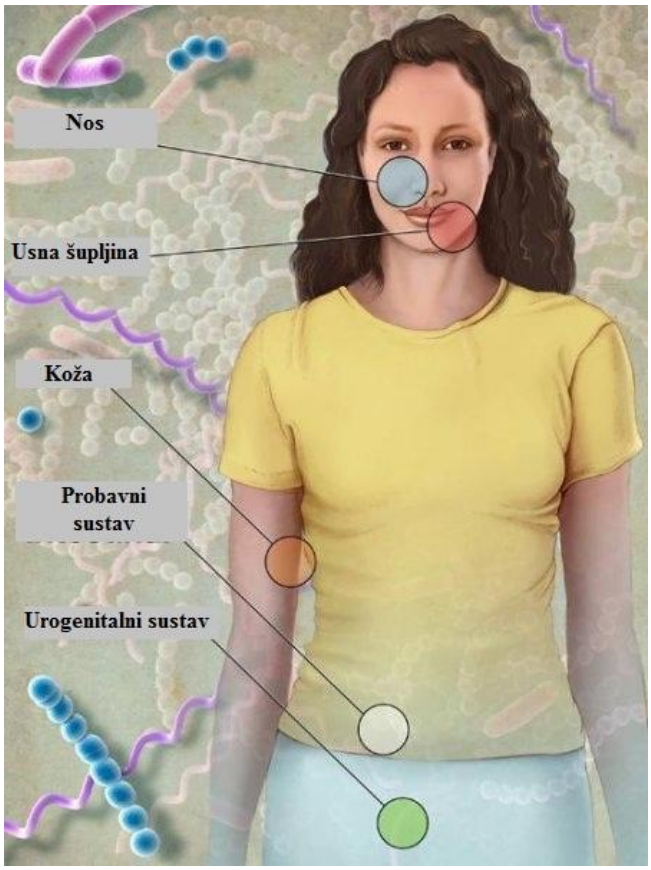
1.1. Metagenomika

Prokarioti zauzimaju dvije od tri domene života, a samo se 1-5% bakterija iz okoliša može uzgojiti u laboratorijskim uvjetima. Ova je pojava poznata pod nazivom "*the great plate count anomaly*" (Staley i Konopka, 1985). Čiste mikrobne kulture sadrže klonove jednog organizma dok u prirodi mikrobi žive u zajednicama i horizontalnim prijenosom razmjenjuju genetički materijal.To doprinosi iznimno velikoj genetičkoj raznolikosti koju je nemoguće istražiti uzgojem u laboratoriju. Ovaj problem je uspješno riješen korištenjem tehnika visokoprotočnog sekvenciranja uz prikupljanje DNA izravno iz uzoraka okoliša, bez prethodnog kultiviranja pojedinih vrsta. Takav se pristup zove metagenomika, a podrazumijeva skupljanje uzoraka iz okoliša, sekvenciranje i analizu dobivenih sljedova. Ovisno o načinu sekvenciranja dobivaju se sljedovi različitih veličina, od 20 do 1000 baza, koje je potrebno računalnim programima složiti u dulje sljedove kako bi dobili vrijednije informacije – djelomične ili čitave gene koje možemo analizirati, ili čak djelomične genome pojedinačnih organizama (Hess i sur., 2011). Analiza metagenoma usmjerena je u dva pravca. Prvi nam daje odgovor na pitanje koje se sve vrste nalaze u nekom okolišu, odnosno daje procjenu filetičke distribucije mikroba prema pretragama po sličnosti u bazama poznatih mikrobnih vrsta. Drugi pravac pokušava klasificirati funkcije opaženih gena ili okvira čitanja u kategorije određene bazama podataka genskih ortologa kao što su KEGG („*Kyoto Encyclopedia of Genes and Genomes"*) (Kanehisa i sur., 2014) ili eggNOG (Powell i sur., 2014) te ocjeniti važnost funkcije prema brojnosti gena u okolišu.

Metagenomskim istraživanjima otkrivena je do tada nepoznata raznolikost mikrobnog sastava. Rappé i Giovannoni (2003) u svom su radu podijelili domenu bakterija na 52 koljena, od kojih je samo 26 tada imalo predstavnike uzgojene u laboratorijskim uvjetima. Jedan od prvih opsežnih metagenomskih projekata bila je karakterizacija mikrobne populacije Sargaškog mora (Venter i sur., 2004), gdje je sekvencirano preko milijardu parova baza DNA i otkriveno preko 1.2 milijuna novih gena. Također, moguća je i identifikacija novih gena iz određene genske obitelji izravno iz metagenoma korištenjem hibridizacijskih proba DNA ili pomoću lančane reakcije polimerazom. Neki od mikroba koje žive u okolišu potencijalno proizvode sekundarne metabolite koji bi mogli služiti kao novi antibiotici. Feng i sur. (2011)

napravili su biblioteku DNA okolišnih uzoraka i pretražili ju degeneriranim početnicama koje su ciljale poliketid sintazu tipa II. Funkcionalnom analizom pronađenih klonova pronađeni su poznati, ali i neki nepoznati spojevi s antimikrobnim djelovanjem, od kojih su neki pokazali jaku aktivnost protiv meticilin-otporne *Staphylococcus aureus* (MRSA, „*Methicillin-resistant Staphylococcus aureus*") (Rosenbach, 1884) i vankomicin-otpornih *Enterococcus sp*. (VRE, „*Vancomycin-Resistant Enterococcus*").

1.1.1. Ljudski mikrobiom



Slika 1: Lokacije uzimanja uzoraka s ljudskog tijela za HMP, preuzeto i prilagođeno s http://hmpdacc.org/micro_analysis/microbiome_analyses.php

Po završetku sekvenciranja genoma čovjeka, mnogi su očekivali kako će naš genom sadržavati oko 100 000 gena koji kodiraju za proteine, te su ostali iznenađeni brojkom od samo 20 000. Ako malo proširimo svoja razmišljanja i u procjenu uključimo i gene mikrobnih vrsta koje nastanjuju ljudski organizam, ispada da je 100 000 gena zapravo preniska procjena. Naime, ljudsko tijelo dom je za 10 puta više mikrobnih stanica nego što je ukupno ljudskih stanica, a ljudski mikrobiom (skup svih mikrobnih staništa u i na čovjeku) sadrži barem 100 puta više gena nego sam genom čovjeka. Mikrobiom čovjeka ima značajno izražen metabolizam glikana, aminokiselina i ksenobiotika, i pomaže nam u proizvodnji vitamina i izoprenoida. Zbog toga ljude možemo zapravo smatrati superorganizmima čiji metabolizam ne bi bio potpun bez mikroba koji ih nastanjuju (Gill i sur., 2006). Projekt genomskog istraživanja ljudskog mikrobioma (HMP; „*Human microbiome project*") zbog toga je logični nastavak sekvenciranja ljudskog genoma. HMP je skup projekata

kojima su glavni ciljevi napraviti referentni skup genoma mikroba koji nastanjuju naše tijelo, okarakterizirati metagenom u zdravim ljudima, naći vezu između bolesti i promjena u mikrobiomu čovjeka, te razviti nove tehnologije i alate za računalnu analizu (NIH HMP Working Group i sur., 2009). Do sada je stvoren veliki katalog gena s različitih anatomskih lokacija na 242 zdrava pojedinca. Lokacije su uključivale 5 područja tijela (slika 1) : kožu, usnu šupljinu, nos, urogenitalni i probavni sustav (Turnbaugh i sur., 2007).

1.1.2. Bolesti povezane s izmijenjenim mikrobnim sastavom

Druga faza HMP bavi se istraživanjem uloge ljudskog mikrobioma u bolestima preko istraživanja tri stanja: trudnoće, bolesti probavnog sustava koristeći upalnu bolest crijeva kao model i respiratorne virusne infekcije uz pojavu dijabetesa tipa 2 *("The Integrative Human Microbiome Project"* 2014). Najviše radova usmjereno je na proučavanje mikrobioma probavnog sustava jer u debelom crijevu nalazimo najveći broj mikroba. U sklopu projekta MetaHIT, sekvenciranjem sadržaja probavnog sustava 124 europska pojedinca, određen je skup gena koji možemo naći u probavnom sustavu ljudi. On premašuje naš skup gena za 150 puta, a 99 % tih gena pripada bakterijama. Pretpostavlja se da svaki pojedinac nosi najmanje 160 različitih vrsta bakterija od kojih je većina zajednička svim ljudima (Qin i sur., 2010). Postoji povezanost između mikrobioma probavnog sustava i nekih kroničnih bolesti kao što su pretilost (Turnbaugh i Gordon, 2009), upalna bolest crijeva (Garrett i sur., 2010), simptomatična ateroskleroza (Karlsson i sur., 2012) i bolesti masne jetre čiji uzrok nije alkohol (Yan i sur., 2011). Studije na pacijentima s cirozom jetre također su pokazale promijenjeni mikrobiom probavnog sustava.

1.2. Metaproteomika

Najbolji način za određivanje genske ekspresijeje promatrati završni proizvod – proteine. Metaproteom je pojam koji je uveo Rodriguez-Valera (2004)kako bi opisao gene i/ili proteine koji su prisutni u okolišnim uzorcima u najvišim koncentracijama. Wilmes i Bond (2004) metaproteomikom su nazvali opsežnu karakterizaciju svih proteina u određenom vremenu u uzorku okolišnih mikroba. Iako bi određivanje proteina izravno iz uzorka bilo najizravniji i najtočniji način predviđanja ekspresije gena u određenom vremenu, zbog složenih metoda, teže pripreme uzoraka i veće

cijene u odnosu na količinu dobivenih rezultata, metaproteomskih studija je još uvijek malo (Keller i Hettich, 2009).

1.3. Upotreba sinonimnih kodona

Aminokiseline su osnovne građevne jedinice proteina. Svaka je aminokiselina u kodirajućoj DNA zapisana kroz kombinaciju tri nukleotida, nazvanom kodon. Molekula DNA građena je od 4 različite vrste nukleotida. Iz toga bi se moglo zaključiti da postoji 4*4*4=64 različite aminokiseline, no to nije slučaj. Naime, većinu aminokiselina možemo dobiti na više različitih načina, te se pripadni kodoni koji kodiraju za istu aminokiselinu zovu sinonimni kodoni. Tablica 1 prikazuje standardni genetički kod(Nirenberg i sur., 1965).

Tablica 1: Standardni genetički kod. 1. baza, 2. baza i 3. baza redom označavaju mjesto u kodonu. Prikazani su svi kodoni i pripadne aminokiseline koje kodiraju.

1. baza	2. baza								3. baza
	U		C		A		G		
U	UUU	(Phe/F) Fenilalanin	UCU	(Ser/S) Serin	UAU	(Tyr/Y) Tirozin	UGU	(Cys/C) Cistein	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucin	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	(Trp/W) Triptofan	G
C	CUU		CCU	(Pro/P) Prolin	CAU	(His/H) Histidin	CGU	(Arg/R) Arginin	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamin	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Izoleucin	ACU	(Thr/T) Treonin	AAU	(Asn/N) Asparagin	AGU	(Ser/S) Serin	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lizin	AGA	(Arg/R) Arginin	A
	AUG[A]	(Met/M) Metionin	ACG		AAG		AGG		G
G	GUU	(Val/V) Valin	GCU	(Ala/A) Alanin	GAU	(Asp/D) Aspartaginska kiselina	GGU	(Gly/G) Glicin	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutaminska kiselina	GGA		A
	GUG		GCG		GAG		GGG		G

Učestalost kojom se sinonimni kodoni koriste u kodirajućoj DNA nije uvijek jednaka unutar genoma. Neki od kodona brže su translatirani od drugih, te se nazivaju optimiranima za translaciju u tom organizmu. Upotreba optimiranih kodona pozitivno korelira s ekspresijom gena koji koriste takve kodone (Ikemura, 1981). Visokoeksprimirani geni su najčešće ribosomalni geni, šaperoni i transkripcijski faktori. Razlog selekcije optimalnih kodona u genima za ove proteine je povećana učinkovitost translacije. Naime, u brzo rastućoj populaciji *Escherichia coli* (Escherich, 1885), ribosomi čine dvije trećine proteinskog sastava bakterije (Pedersen i sur., 1978), stoga je količina ribosoma ograničavajući faktor brzine rasta. Optimalni kodoni brže se translatiraju od ne-optimalnih (Sørensen i Pedersen, 1991) zbog veće dostupnosti odgovarajućih molekula tRNA - ribosomi brže klize molekulom mRNA i brže se otpuštaju što u konačnici ubrzava translaciju molekula mRNA. Zbog toga korištenje optimiranih kodona u visokoeksprimiranim genima omogućava brži rast takvih bakterija što je očita selektivna prednost, pogotovo u brzo rastućim organizmima (Sharp i sur., 2010). Pregledom 461 mikrobnih genoma utvrđeno je kako je postojanje selekcije na razini kodona u prokariotima univerzalna pojava; u svakoj od vrsta (osim jedne) koristeći klasifikaciju pomoću slučajnih šuma („*Random forest*“) nađen je podskup gena koji pokazuje veću sličnost s ribosomalnim genima. Taj podskup gena sačinjava oko 5% do 33% genoma i eksperimentalnim metodama je potvrđena veća količina molekula mRNA eksprimiranih s tih gena (Supek i sur., 2010). Slični obrasci upotrebe sinonimnih kodona vrijede ne samo unutar iste vrste, već i unutar iste zajednice organizama. Roller i sur. (2013) pokazali su kako neovisno o filogeniji mikrobi unutar iste ekološke niše dijele sklonost k upotrebi sličnih sinonimnih kodona te predložili kako se upotreba kodona u metagenomima može koristiti za predviđanje ekspresije gena na isti način kako se predviđaju optimirani visoko eksprimirani geni u pojedinačnim vrstama. S druge strane, obrasci upotrebe sinonimnih kodona razlikuju se između različitih ekoloških niša, čak i između istih vrsta u različitim nišama. Prema tome, možemo pretpostaviti da postoji razlika u okolišu koji stvara stanje ciroze jetre u probavnom sustavu čovjeka i onom koje postoji u zdravom probavnom sustavu, te prema određivanju važnosti gena u metagenomima tih stanja analizom upotrebe sinonimnih kodona mogli bismo odrediti bitne biomarkere ili diskriminirati bolesne od zdravih metagenoma.

1.4. Određivanje gena optimiranih za translaciju

Kako bismo odredili gene koji su optimirani za translaciju moramo moći procijeniti upotrebu optimiranih kodona u svakom genu. Većina studija se slaže kako su visokoeksprimirani geni povezani s korištenjem najčešćih kodona zbog čega su mnoge mjere koje pronalaze optimirane gene utemeljene na određivanju frekvencija korištenja kodona (Cannarozzi i Schneider, 2012). Neke od tih mjera koriste usporedbu skupa gena od interesa sa referentnim skupom gena željenih kvaliteta, npr. kao skup referentnih gena koristi se skup poznatih visokoeksprimiranih gena, jer su oni pod većom selekcijom. Ovakva mjera je npr. indeks adaptacije kodona (CAI; „*Codon Adaptation Index*“) (Sharp i Li, 1987). Druge skupine mjera temelje se na određivanju odstupanja u korištenju kodona od procjenjene distribucije, ili na računanju interakcija kodona s antikodonima tRNA kao ograničavajućem faktoru translacije. Ostale se mjere ne mogu svrstati u niti jednu od do sada pobrojanih skupina. Takva je npr.često korištena Nc-efektivni broj kodona (Wright, 1990). Zbog varijabilnosti u duljini sekvenci i/ili nukleotidnom sastavu na koju su neke od ovih mjera osjetljive, one nisu posve primjerene.

1.4.1. MILC – mjera neovisna o duljini i nukleotidnom sastavu

MILC („*Measure Independent of Length and Composition*“) (Supek i Vlahovicek, 2005) je mjera koja uspoređuje udaljenost u raspodjeli korištenja kodona u zadanom otvorenom okviru čitanja s očekivanom distribucijom kodona. Temelji se na određivanju prikladnosti modela („*goodness of fit*“) te je neovisna o duljini i nukleotidnom sastavu zbog čega je korištena u ovom radu.

2. CILJ ISTRAŽIVANJA

Qin i sur.(2014) okarakterizirali su mikrobiom probavnog sustava kod 98 bolesnika oboljelih od ciroze jetre usporedbom s 83 zdravih pojedinaca. Izgradili su referentni skup gena povezanih s cirozom jetre od čega su izdvojili gene koji se razlikuju u brojnosti između zdravih i bolesnih ljudi, te predložili 15 biomarkera na temelju kojih je mogućevrlo točno razlikovati ljude oboljele od ciroze jetre i zdrave pojedince.

U ovom radu analizirala sam metagenome 30 osoba oboljelih od ciroze jetre i 30 kontrolnih pojedinaca nasumično odabranih iz prethodno spomenutog istraživanja. Cilj je provesti iste postupke kao u istraživanju Qin i sur., ali za svaki uzorak napraviti procjenu genske ekspresije na temelju korištenja optimalnih kodona. S tako dobivenim podacima cilj mi je klasificirati gene na zdrave i bolesne metodom strojnog učenja slučajnim šumama, kao i odrediti fenotip uzoraka na temelju predviđenih klasa gena. Također, cilj je izdvojiti gene za koje je na ovakav način nađena razlika u predviđenoj ekspresiji između zdravih pojedinaca i bolesnih osoba, i usporediti dobivene rezultate s prethodno spomenutim istraživanjem.

3. MATERIJAL I METODE

3.1. Materijali

3.1.1. Izvor korištenih sekvenci

DNA sljedovi korišteni u ovom istraživanju dio su sljedova dobivenih sekvenciranjem ekstrakata DNA iz uzoraka stolica bolesnika s cirozom jetre i zdravih pojedinaca kineskog podrijetla u sklopu istraživanja koje su proveli Qin i sur. (2014). Sljedovi su pohranjeni u europskoj knjižnici nukleotida (ENA; „*European Nucleotide Archive*“) pod pristupnim brojem ERP005860. ID oznake korištenih uzoraka nalaze se u tablici 2.

3.1.2. Baza podataka KEGG

KEGG („*Kyoto Encyclopedia of Genes and Genomes“*) je skup 15 baza podataka podijeljenih u 4 kategorije: sistematske informacije, genomske informacije, kemijske informacije, te informacije povezane sa zdravljem. KEGG je osmišljen kao baza znanja o metabolizmu i staničnim procesima – alat koji bi omogućio zaključivanje o biološkim funkcijama na temelju genomskih sekvenci. Baza podataka „*KEGG pathway*“ sadrži ručno sastavljene informacije u obliku metaboličkih mapa koje predstavljaju eksperimentalno prikupljena znanja o metabolizmu. Svaka mapa sadrži mrežu molekularnih interakcija i reakcija te je dizajnirana s ciljem povezivanja gena i genskih produkata. Svaki unos u mapi obilježen je K brojem koji označava pojedine skupine ortologa. Koristila sam verziju od 07.07.2010.

3.2. Metode

3.2.1. Sravnjenje sljedova

Sljedovi dobiveni sekvenciranjem nove generacije duljine su 30 – 700 parova baza, ovisno o korištenoj metodi. Illumina HiSeq sekvence korištene u ovom radu duljine su 100 parova baza. Kako bi našla i odstranila iz uzoraka sekvence koje pripadaju genomu čovjekaa nisu dio njihovih mikrobioma, sljedove dobivene sekvenciranjem usporedila sam s ljudskim genomom. Postupak uspoređivanja dvaju sljedova s ciljem pronalaska podsekvence većeg sljeda koja odgovara manjem sljedu s dopuštanjem određenog broja pogreški naziva se sravnjenje sljedova (eng. „*sequence alignment*“). Duljina haploidnog genoma čovjeka je preko 3 milijarde parova baza, te za mnoge pročitane sljedove duljine 100 nukleotida postoji više od jednog mjesta na kojima se mogu nalaziti unutar genoma. Također, prilikom sekvenciranja su moguće pogreške i poznato je da kvaliteta dobivenog slijeda pada prema 3' kraju. Ovo su sve razlozi iz kojih je problem sravnjenja sljedova računalno vrlo zahtjevan, a ovisno o problemu na koji se primjenjuje postaje i zahtjevniji, no kako je postupak našao iznimno široku primjenu u biologiji razvijeni su razni algoritmi kojima se uspješno i u realnom vremenu pronalazi optimalno rješenje. Neki od njih opisani su u ovom poglavlju.

3.2.1.1. BWA

Burrows-Wheeler alat za poravnanja (eng. „*Burrows-Wheeler Alignment tool*“) je programski paket za mapiranje DNA sljedova na velike referentne genome kao što je genom čovjeka. Temeljen je na Burrows-Wheeler transformaciji te omogućava efikasna sravnjenja kratkih sljedova uz dopuštanje postojanja grešaka i praznina pri poravnanjima. Kao rezultat sravnjenja daje SAM format koji je moguće obrađivati pomoću programskog paketa SAMtools.Koristila sam verziju bwa 0.7.5. (Li i Durbin, 2009).

3.2.1.2. BLAST

BLAST („*Basic Local Alignment Search Tool*“) je algoritam za uspoređivanje nukleotidnih ili aminokiselinskih sljedova prema sličnosti. Omogućuje pretraživanje baze gena ili proteina, te kao rezultat daje sve sekvence koje s našom sekvencom od

interesa imaju sličnost jednaku ili veću od određenog praga. Postoji nekoliko inačica programa ovisno o tipu sekvence koju uspoređujemo i tipu baze s kojom se uspoređuje. U radu sam koristila BLASTX. Kao sekvenca dan je nukleotidni slijed predviđenog gena, blastX prevodi taj slijed u aminokiselinski slijed po svih 6 mogućih okvira čitanja i uspoređuje s proteinskom bazom podataka. Koristila sam verziju 2.2.28 (Altschul i sur., 1990).

3.2.2. SAM format i SAMtools

SAM („*Sequence Alignment/Map*“) je tekstualni format za spremanje podataka dobivenih sravnjenjem kratkih sekvenci na referentni genom. Svaka se tekstualna datoteka u SAM formatu sastoji od opcionalnog zaglavlja koje, ako postoji, počinje sa znakom „@“ , te nužnog dijela koji se sastoji od 11 polja i daje potpune informacije o poravnanju i položaju kratkog slijeda na referentnom genomu. Svaki zapis u SAM datoteci predstavlja poravnanje jednog slijeda nukleotida s referentnim genomom. SAMtools je skup alata koji omogućava manipulaciju zapisima u datoteci SAM tipa, uključujući sortiranje, spajanje više datoteka, pretvaranje u kompaktniji BAM (binarni SAM) ili često korišteni FASTQ oblik, izoliranje samo onih zapisa koji imaju poravnanje s genomom, i ostalo(H. Li i sur., 2009).

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

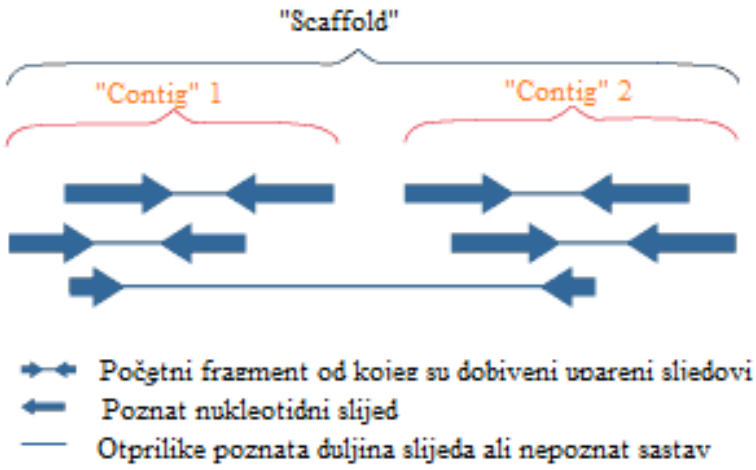
Slika 2: Primjer zapisa u SAM formatu

3.2.3. R

R je slobodno dostupno programsko okruženje koje omogućuje obradu velike količine podataka, a osobito je pogodan za statističku obradu bioloških podataka zbog velikog broja također slobodno dostupnih programskih paketa s tom namjenom. Zbog iznimno velike količine podataka s kojom sam radila R je korišten na Linux okruženju. Koristila sam verziju 3.1.1 (*R Core Team, 2014*).

3.2.4. *De novo*sastavljanje genoma

Iz uzoraka stolica zdravih i bolesnih osoba izolirana je DNA i sekvencirana na Illumina HiSeq 2000 sekvenatoru. Pri tome su dobiveni upareni sljedovi sekvencirane DNA duljine 100 parova baza. Kako bi iz ovakvih sljedova mogli dobiti relevantne informacije oni moraju biti složeni u dulje neprekinute sljedove („*contigs*“) temeljem preklapanja. Takvi dulji neprekinuti sljedovi slažu se dodatno korištenjem informacija o međusobnoj udaljenosti uparenih sekvenciranih početnih sljedova u isprekidane složene dijelove genoma nazvane eng. „*scaffolds*“ (slika 2).



Slika 3: Slaganje genoma, preuzeto i prilagođeno s: http://genome.jgi.doe.gov/help/scaffolds.html

Ovakav postupak zove se sastavljanje genoma*de novo* („*de novoassembly*“) i matematički spada u skupinu NP potpunih problema te nema efikasno računalno rješenje, što znači da zahtjeva veliku količinu memorije i vremena za rješavanje. Ovaj problem je dodatno otežan zbog relativno male duljine početnih sekvenciranih sljedova, te činjenice da ne dolaze svi sljedovi iz jednog genoma nego iz njih nepoznato mnogo (jer smo sekvencirali DNA čitave mikrobne zajednice).

3.2.4.1. SOAPdenovo

SOAPdenovo je računalni program koji služi za sastavljanje kratkih sljedova DNA dobivenih sekvenciranjem pomoću sekvenatora nove generacije, posebno je dizajniran za sastavljanje sekvenci dobivenih na IlluminaGenome Analyzer sekvenatoru (R. Li i sur., 2009). Isti kemijski postupak pri sekvenciranju koristi Illumina HiSeq 2000 model sekvenatora koji je korišten u istraživanju (Qin i sur,. 2014), a iz kojeg su preuzeti podaci. Koristila sam verziju SOAPdenovo-V1.05.

Algoritam koji se koristi radi na način da prvo odredi sve moguće riječi duljine k (za određeni k=podsekvence duljine k) iz svih sekvenci. Nakon toga za svaku riječ duljine k napravi (k-1)-mere tj. prefiks i sufiks i složi ih u čvorove De Brujin grafa, te na rubove ispiše sekvencu ako postoji riječ koja sadrži prefiks i sufiks kao podsekvence. Nakon toga nalazi se Eulerov put kroz graf koji odgovara složenom genomu. Riječi s niskom pojavnosti mogu značiti greške u sekvenciranju koje znaju zakomplicirati izračun te se posebno razmatraju.

3.2.5. MetaGeneMark

Računalni algoritam za pronalazak prokariotskih gena. Temeljen je na frekvencijama parova kodona procjenjenih preko sadržajaGC i ostalih mjera za dani slijed. Na umjetnim podacima dobivenim sekvenciranjem kratkim sljedovima postigao je osjetljivost od 95% i specifičnost od 90%. Predviđa gene za bakterije i arheje, te je prikladan za korištenje u metagenomskim studijama. Koristila sam prokariotski GeneMark.hmm, verzija 2.8 (Noguchi i sur., 2006).

3.2.6. MILC i MELP

MILC je mjera neovisna o duljini i nukleotidnom sastavu, te je definirana kao

$$\mathrm{MILC} = \frac{\sum_{\mathrm{a}} \mathrm{M_a}}{L} - C$$

gdje je M_a pojedinačni doprinos po svakoj aminokiselini, L broj kodona otvorenog okvira čitanja na kojem radimo statistiku, a C faktor korekcije zbog precjenjivanja ukupne mjere u kraćim sekvencama. Iz ORFa se isključuju stop kodoni jer često pristrano koriste jedan kodon a nisu dio kodirajuće sekvence.

M_a je definiran kao

$$M_a = 2\sum_c O_c ln\frac{O_c}{E_c} = 2\sum_c O_c ln\frac{f_c}{g_c}$$

gdje je O_c opaženi broj kodona c, E_c očekivani broj istog kodona. Ista formula vrijedi i u slučaju zamjene broja kodona s njegovom frekvencijom, odnosno f_c označava opaženu frekvenciju kodona c, a g_c očekivanu

Faktor korekcije C definiran je formulom:

$$C = \frac{\sum_a (r_a - 1)}{L} - 0.5$$

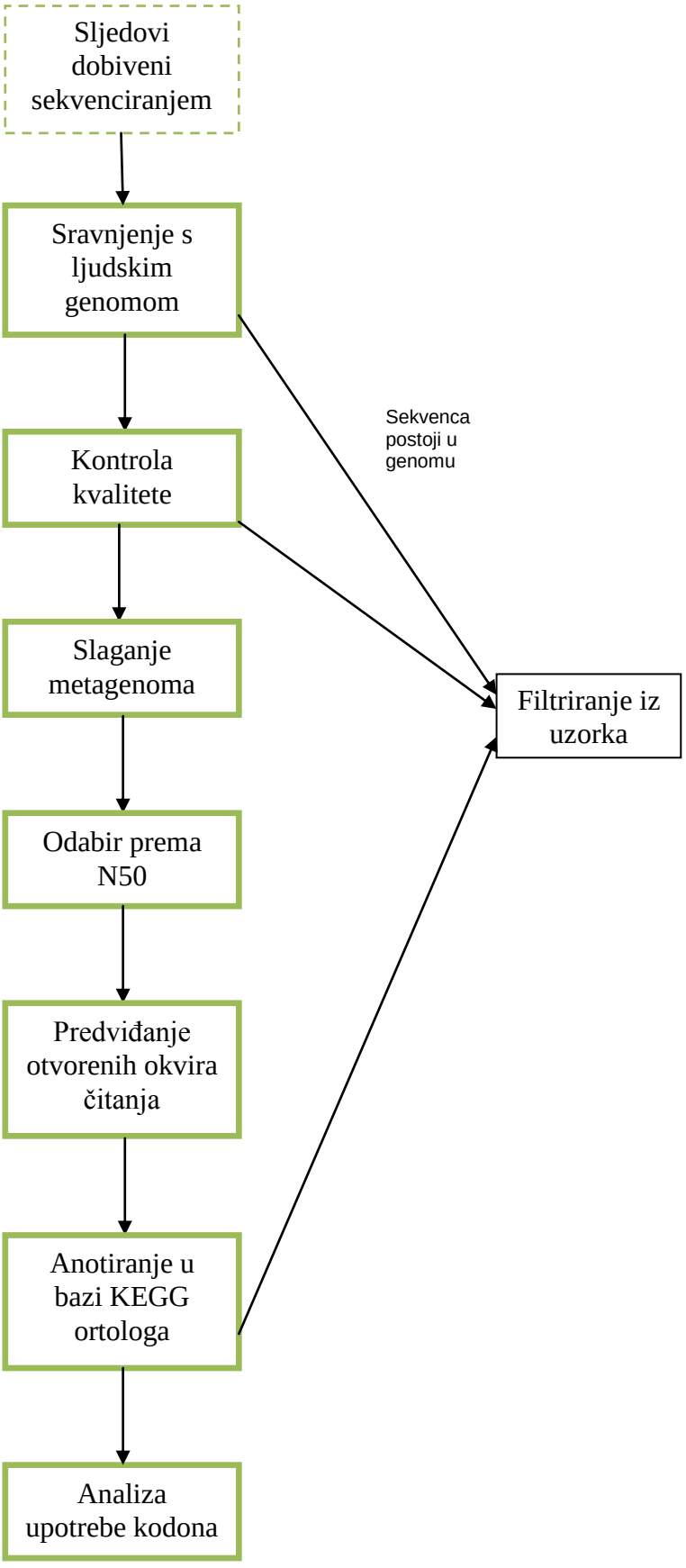
u kojoj je broj mogućih kodona koji kodiraju za aminokiselinu a označen s r_a .

MILC vrijednost se koristi za predviđanje razine ekspresije gena usporedbom s MILC vrijednosti referentnog skupa visoko eksprimiranih gena. Ovakva statistika naziva se MELP (MILC-based Expression Level Predictor). Definirana je kao:

$$MELP = \frac{MILC_{genom}}{MILC_{referentniskup}}$$

$MILC_{genom}$ predstavlja udaljenost obrasca upotrebe kodona u nekom genu od raspodjele upotrebe kodona u čitavom genomu, a $MILC_{\text{referentni skup}}$ udaljenost upotrebe kodona u genu od interesa od raspodjele upotrebe kodona u zadanom referentnom skupu. Ako želimo odrediti koliko je visoko naš gen eksprimiran, kao referentni skup uzimamo neke visoko eksprimirane gene a takvi su geni za ribosomalne proteine.

3.3. Postupak



Slika 4: Shema provedenih postupaka

3.3.1. Sravnjenje s genomom čovjeka

Uzorci sekvenciranja DNA iz stolice30 bolesnika s cirozom jetre i 30 zdravih pojedinaca preuzeti su iz podataka istraživanja Qin i sur. (2014) u fastq obliku. Svaki uzorak sekvenciran je u uparenim sljedovima („*Paired-end*“). Sravnjenjem svih DNA sljedova svakog uzorka s genomom čovjeka pomoću programa BWA s parametrom “-n 0.2“ odredila sam sekvence koje su porijeklom iz ljudskom genoma. Oba zapisa (po jedan za svaki iz para sljedova) po svakom uzorku kombinirana su u isti iz kojeg sam odstranila sve sekvence koje su imale porijeklo u genomu čovjeka ili im je par moguće naći u ljudskom genomu pomoću samtools s parametrom „–f 12“.

3.3.2. Kontrola kvalitete

Napravila sam kontrolu kvalitete u računalnom programu R na sljedeći način: (1) uklonila sam sve sljedove koji su imali 3 ili više neodređenih baza (N); (2) iz uzoraka sam uklonila sve sljedove koji su sadržavali 50 ili više baza čija je kvaliteta 2; (3) svi su sljedovi skraćivani od 3' kraja do minimalne duljine 90 nukleotida ako je kvaliteta baza bila 2. Sve sekvence iz jednog uzorka nakon skraćivanja i filtriranja razdvojila sam prema postojanju uparenog slijeda i poziciji u paru u 4 različita zapisa.

3.3.3. Sastavljanje genoma

Sve preostale sljedove sastavila sam u odgovarajuće veće sekvence pomoću programa SOAPdenovo-63mer v1.05 s parametrom „–d –M3“. Iz dobivenih podataka uklonila sam nepoznate nukleotide i odstranila sekvence duljine manje od 500 parova baza. Testirala sam sve vrijednosti k-mera od 31 do 59 te izabrala onaj skup podataka koji je imao najveći N50 nakon filtriranja.

3.3.4. Predviđanje otvorenih okvira čitanja

Za svaki od 30 zdravih i 30 bolesnih uzoraka od preostalih sekvenci predviđeni su otvoreni okviri čitanja (ORF„„*Open Reading Frame*“) pomoću programa MetaGeneMark. Dobivene otvorene okvire čitanja usporedilasam s bazom KEGG ortologa pomoću programa BLASTX s parametrom „–evalue 1e-5“.

3.3.5. Filtriranje otvorenih okvira čitanja

Skriptom vlastite izrade u programu R odredila sam sve dijelovi svih sekvenci koje sam mogla jednoznačno identificirati u KEGG bazi temeljem 3 najbolja rezultata ukoliko su njihove bit vrijednosti bile iznad 60 i e vrijednost ispod 10^{-5}, te sam na njima radila analizu.

3.4. Statistička analiza podataka

Svaki uzorak sam posebno obrađivala u računalnom programu R na sljedeći način: na svim sekvencama unutar istog uzorka odredila sam MILC i MELP vrijednost prema prethodno opisanom postupku (Supek and Vlahovicek, 2005). U daljnju obradu nisam uzimala gene čija je duljina manja od 30 kodona što je preporučeno pri analizi upotrebe kodona. Postoje tri razine ontologije na kojima sam radila statistiku: B – vrlo široka, C –razina metaboličkih puteva, i KO – razina ortologa. Na svakoj razini pobrojala sam ortologe/metaboličke puteve kojima melp vrijednosti spadaju u 95-ti, 90-ti, 85-ti, 70-ti i 50-ti kvantil. Za svaki ortolog/put odredila sam obogaćenje (eng. „*enrichment*“) kao:

$$Enr = 100\frac{Top_s - All_s}{All_s}$$

Gdje je Top_s broj ortologa/puteva koji imaju melp vrijednosti iznad spomenutih određenih granica, a All_s broj svih ortologa/puteva koji očekujemo za svaki kvantil, uz dodatak pseudovrijednosti 1. n_{all} označava ukupni broj ortologa/puteva, a n_{top} ukupno zapaženih za svaki kvantil.

$$All_s = All\frac{n_{all}}{n_{top}} + 1$$

Izračunate su i M i A vrijednosti:

$$M = \log_2 \frac{Top_s}{All_s}$$

$$\mathrm{A} = \frac{\log_2 All_s Top_s}{2}$$

Za svaki ortolog/put napravila sam binomni test i odredila pripadnu p vrijednost. Korekciju p vrijednosti napravila samprema Benjamini & Hochberg metodi za kontrolu lažnih pozitivnih rezultata.

3.5. Metode strojnog učenja

Metodom slučajnih šuma (eng. „*Random forest*“) napravila sam klasifikaciju svih gena prema pripadnim statističkim vrijednostima na „zdrave“ i „bolesne“ gene. Za validaciju sam koristila unakrsnu validaciju s particijom svih gena na 10 podskupina. (eng. „*10 fold cross validation*“) u R programskom okruženju.

3.5.1. Slučajne šume

Stablo odlučivanja je vizualni prikaz klasifikacijske odluke ili regresije. U strojnom učenju koristi se za predviđanja ponašanja podataka na temelju ponašanja skupa podataka na kojem se obavlja trening (nadzirano strojno učenje). Iako su stabla odlučivanja neosjetljiva na skaliranje i neke ostale transformacije na podacima, dobrau filtriranju nebitnih varijablii vrlo lako razumljiva, rijetko daju zadovoljavajuće rezultate zbog problema s nedovoljno ili previše podešenim modelom (eng. „*Low bias, high variance*“). Slučajne šume rješavaju probleme koje imaju stabla odlučivanja tako što istovremeno rade velik broj modela izgradnjom različitih stabala odlučivanja te uzimaju prosjek kao odluku. Koristila sam R paket randomForest 4.6-10.

3.5.2. Unakrsna validacija

Unakrsna validacija je pristup koji služi za procjenu greške pri predviđanju testnih podataka u metodama strojnog učenja. Postupak koji primjenjujemo sastoji se od podijele skupa podataka na k podjednakih dijelova, te uzimanje k-1 dijela kao skupa na kojem treniramo metodu. Ostatak je testni skup koji služi za validaciju, te na njemu izračunamo pogrešku klasifikacije odnosno predviđanja, Err_1. U drugom koraku za validacijski skup podataka izaberemo drugi podskup, a treniramo metodu na ostatku, te opet izračunamo grešku, Err_2. Nakon što izračunamo greške na svih k dijelova, ukupnu grešku dobijemo kao prosjek:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i$$

Gdje je $Err_i = I(y_i \neq \widehat{y_i})$,$\widehat{y_i}$je predviđena vrijednost, a y_i stvarna. Ako skupovi nisu jednaki po veličini uzima se težinski prosjek.

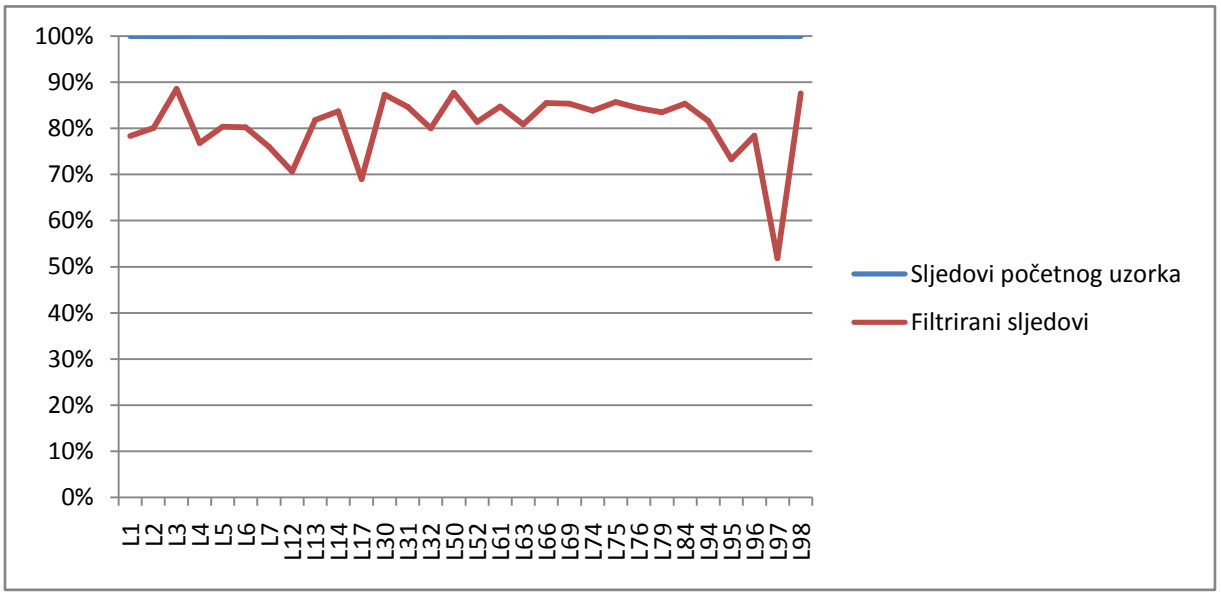
4. REZULTATI

4.1. Početnaobrada podataka

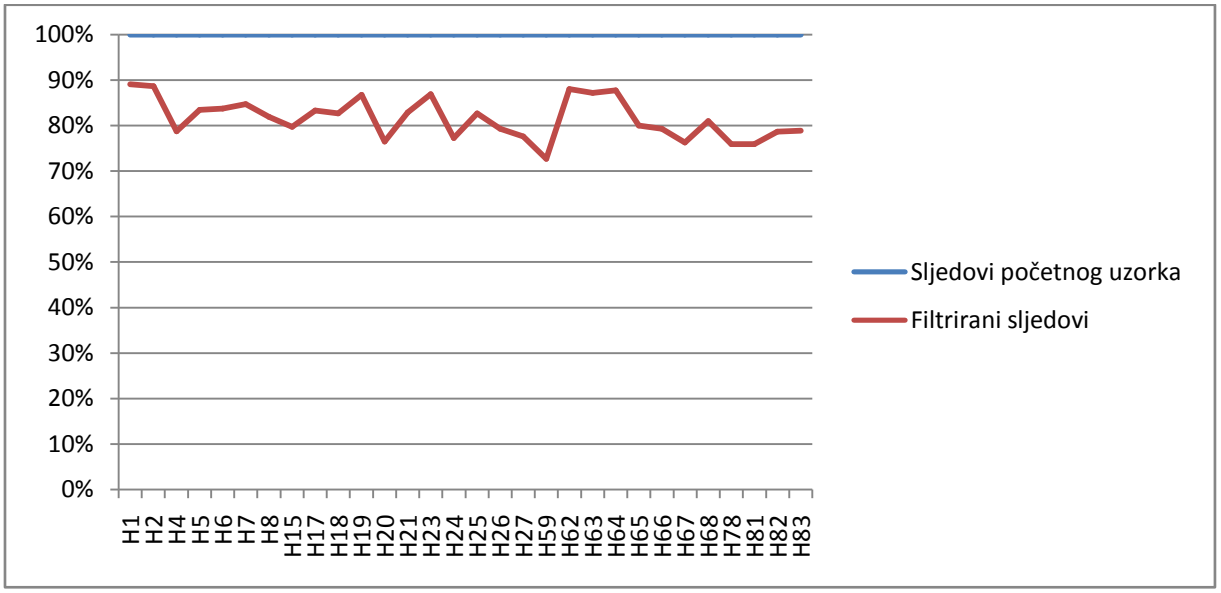
Prikazan je broj sljedova po uzorcima nakon sekvenciranja, postotak DNA koja je moguće mapirati na genom čovjeka sravnjenjem te broj preostalih sljedova nakon uklanjanja svih mapiranih sekvenci (Tablica 2). Slika 5 i 6 prikazuju udio podataka preostalih nakon filtriranja bolesnih i zdravih uzoraka.

Tablica 2: Broj sljedova u uzorcima nakon filtriranja i kontrole kvalitete

	Bolesni uzorci				Zdravi uzorci			
Redni broj uzorka	**ID uzorka**	**Broj sljedova početnog uzorka**	**Postotak ljudske DNA**	**Broj sljedova nakon filtriranja**	**ID uzorka**	**Broj sljedova početnog uzorka**	**Postotak ljudske DNA**	**Broj sljedova nakon filtriranja**
1	L1	33.454.542	0,03	26.197.227	H1	30.363.084	0,01	27.045.480
2	L2	35.053.782	0,01	28.064.408	H2	29.340.068	0,01	26.003.415
3	L3	63.060.022	0,00	55.839.171	H4	51.629.990	0,02	40.649.724
4	L4	42.936.780	0,06	32.972.723	H5	31.658.258	0,01	26.411.159
5	L5	40.642.762	0,00	32.655.622	H6	71.449.782	0,06	59.825.207
6	L6	106.332.924	0,01	85.299.246	H7	54.778.534	0,02	46.398.814
7	L7	66.765.480	0,01	50.732.270	H8	98.148.348	0,20	80.372.583
8	L12	44.243.110	0,01	31.244.665	H15	83.452.056	0,06	66.510.552
9	L13	63.748.028	0,01	52.130.334	H17	42.639.042	0,01	35.508.986
10	L14	46.862.302	0,00	39.203.907	H18	38.306.954	1,37	31.655.924
11	L17	99.623.118	0,13	68.663.694	H19	53.123.420	0,01	46.091.858
12	L30	91.888.542	0,01	80.207.816	H20	56.486.348	2,36	43.216.043
13	L31	29.028.876	0,02	24.574.165	H21	51.623.474	0,03	42.784.109
14	L32	39.166.804	0,18	31.336.370	H23	52.184.770	0,12	45.350.518
15	L50	49.082.910	0,02	43.024.705	H24	34.499.188	1,24	26.641.150
16	L52	55.423.012	0,64	45.103.843	H25	36.928.836	0,02	30.523.458
17	L61	49.256.794	0,05	41.706.663	H26	39.296.662	0,03	31.151.768
18	L63	35.835.300	3,78	28.963.315	H27	50.308.570	0,14	39.051.528
19	L66	44.884.242	0,04	38.356.566	H59	49.695.512	2,73	36.131.313
20	L69	88.651.320	0,07	75.668.698	H62	55.613.486	0,01	48.949.320
21	L74	71.905.768	1,45	60.249.518	H63	45.238.358	0,10	39.432.943
22	L75	162.386.110	0,16	139.114.242	H64	31.361.754	0,25	27.510.551
23	L76	132.805.776	0,35	111.996.216	H65	61.195.352	0,03	48.937.661
24	L79	60.845.942	0,09	50.788.798	H66	44.515.410	0,05	35.284.845
25	L84	89.490.776	0,01	76.375.718	H67	52.399.336	0,02	39.976.099
26	L94	61.492.856	0,07	50.169.777	H68	52.091.294	0,03	42.192.567
27	L95	107.866.604	0,01	79.049.687	H78	34.865.102	0,01	26.457.857
28	L96	44.261.288	0,23	34.707.843	H81	48.191.744	0,10	36.592.607
29	L97	187.822.526	0,05	97.284.211	H82	35.655.800	1,21	28.042.040
30	L98	45.479.164	0,29	39.819.698	H83	47.589.374	0,03	37.531.966

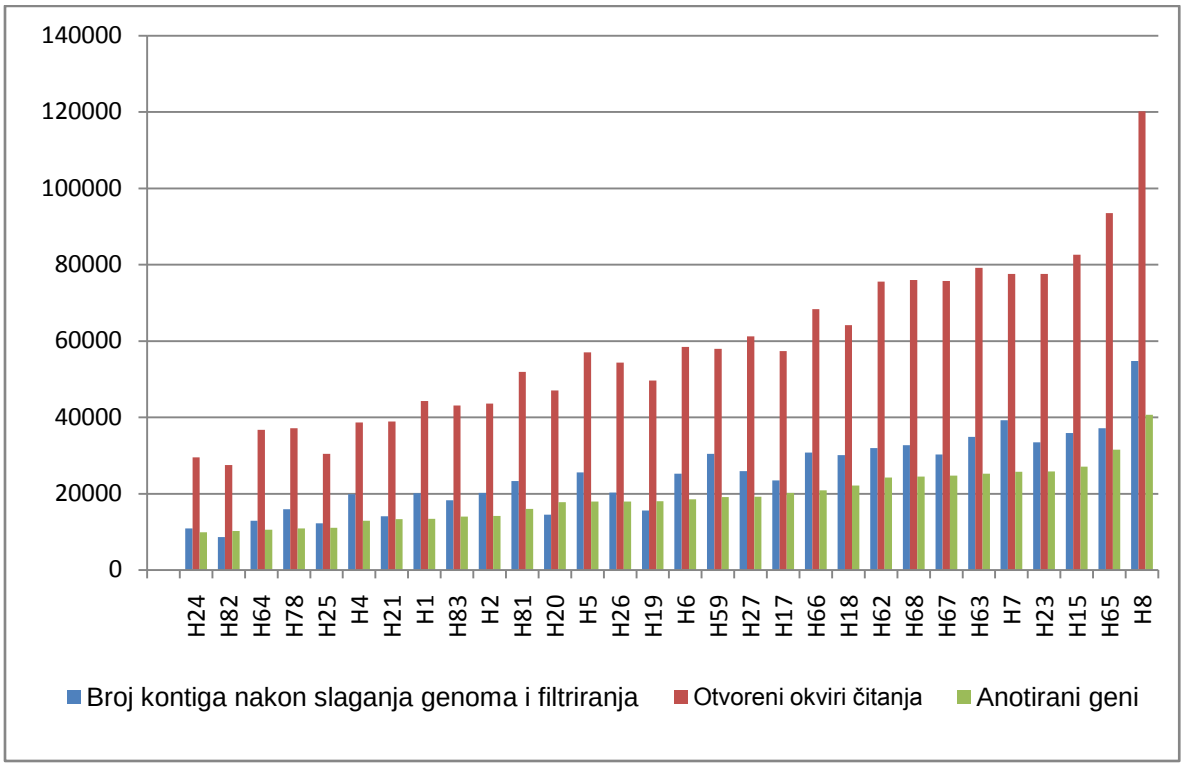


Slika 5: Udio filtriranih sljedova u bolesnim uzorcima nakon sravnjenja s genomom čovjeka

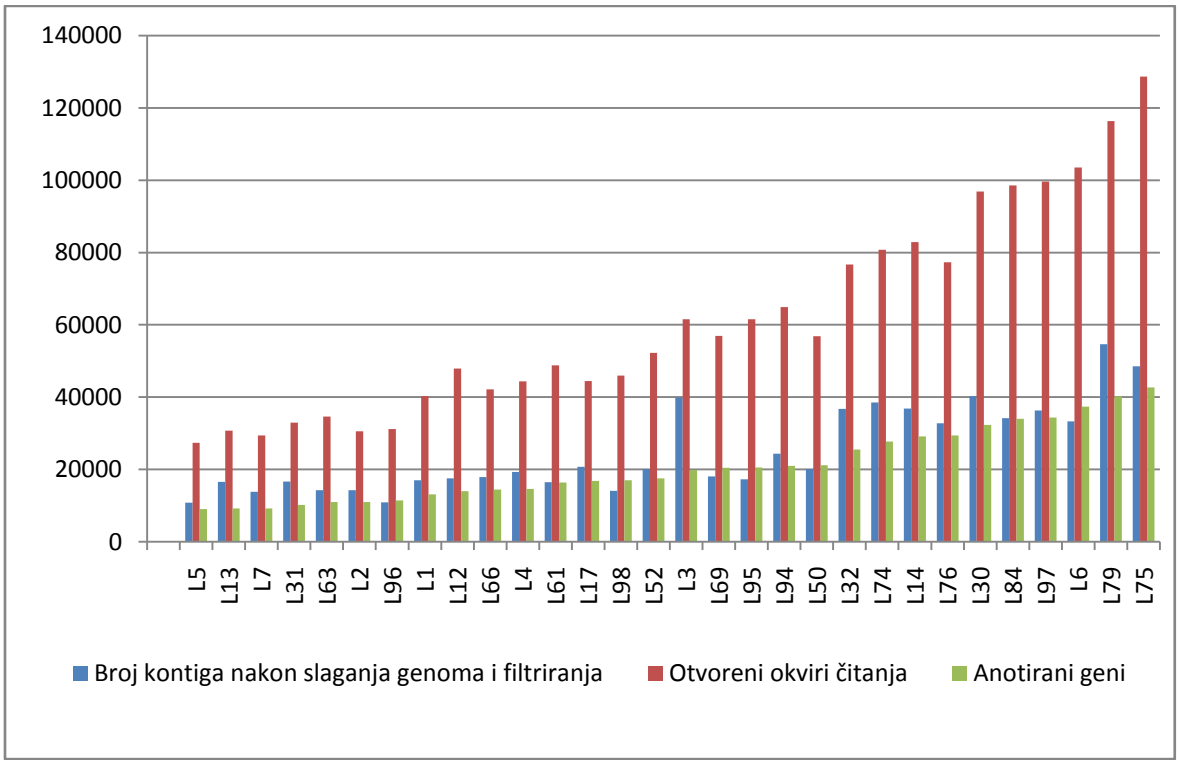


Slika 6: Udio filtriranih sljedova u zdravim uzorcima sravnjenja s genomom čovjeka

4.2. Sastavljanje metagenoma i predviđanje otvorenih okvira čitanja



Slika 7: Usporedba rezultata slaganja metagenoma, predviđanja otvorenih okvira čitanja i anotacije gena zdravih pojedinaca.



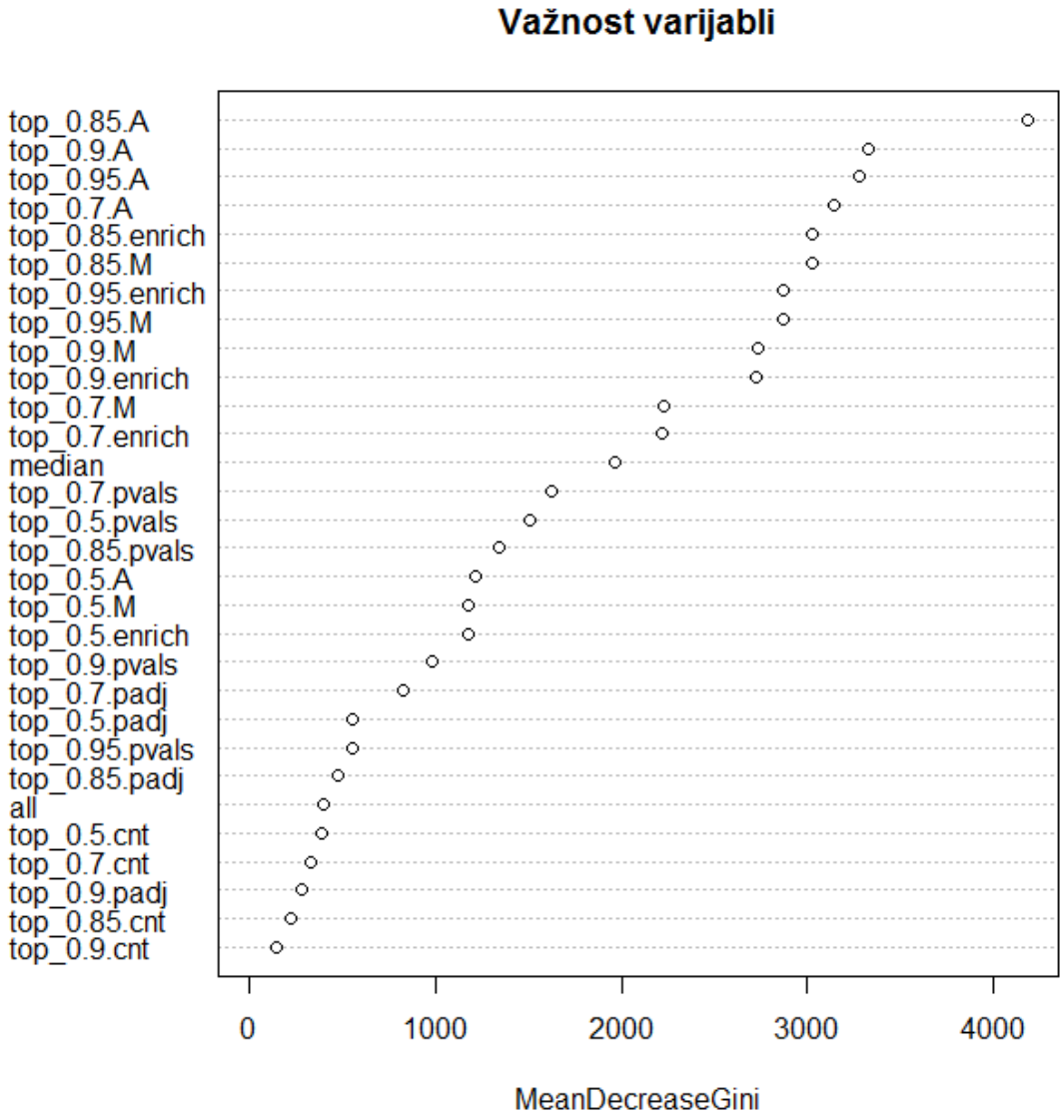
Slika 8:Usporedba rezultata slaganja metagenoma, predviđanja otvorenih okvira čitanja i anotacije gena kod bolesnih poojedinaca.

Slike 7 i 8 prikazuju broj sekvenci nakon slaganja metagenoma u zdravim i bolesnim uzorcima, po svakom uzorku, nakon odabira metagenoma složenih programom SOAPdenovo te izborom onih s najboljim N50 vrijednostima (plava boja). Prikazan je i broj predviđenih gena MetaGeneMark programom (narančasta boja), te broj predviđenih gena kojima su nađeni homolozi u KEGG bazi (zelena boja).Ukupno sam odredila 751 935 kontiga nakon slaganja genoma u bolesnim uzorcima i 749 247 u zdravim. Na temelju njih ukupno je predviđeno 1 845 896 gena u 30 bolesnih uzoraka, od čega je za 630 730 nađen homolog u KEGG bazi, te 1 755 338 u 30 zdravih uzoraka, od čega je anotirano 578 064 koje sam koristila u daljoj analizi. U prosjeku je 33,0% predviđenih gena uspješno anotirano u zdravim osobama, te 33,9% u bolesnim.

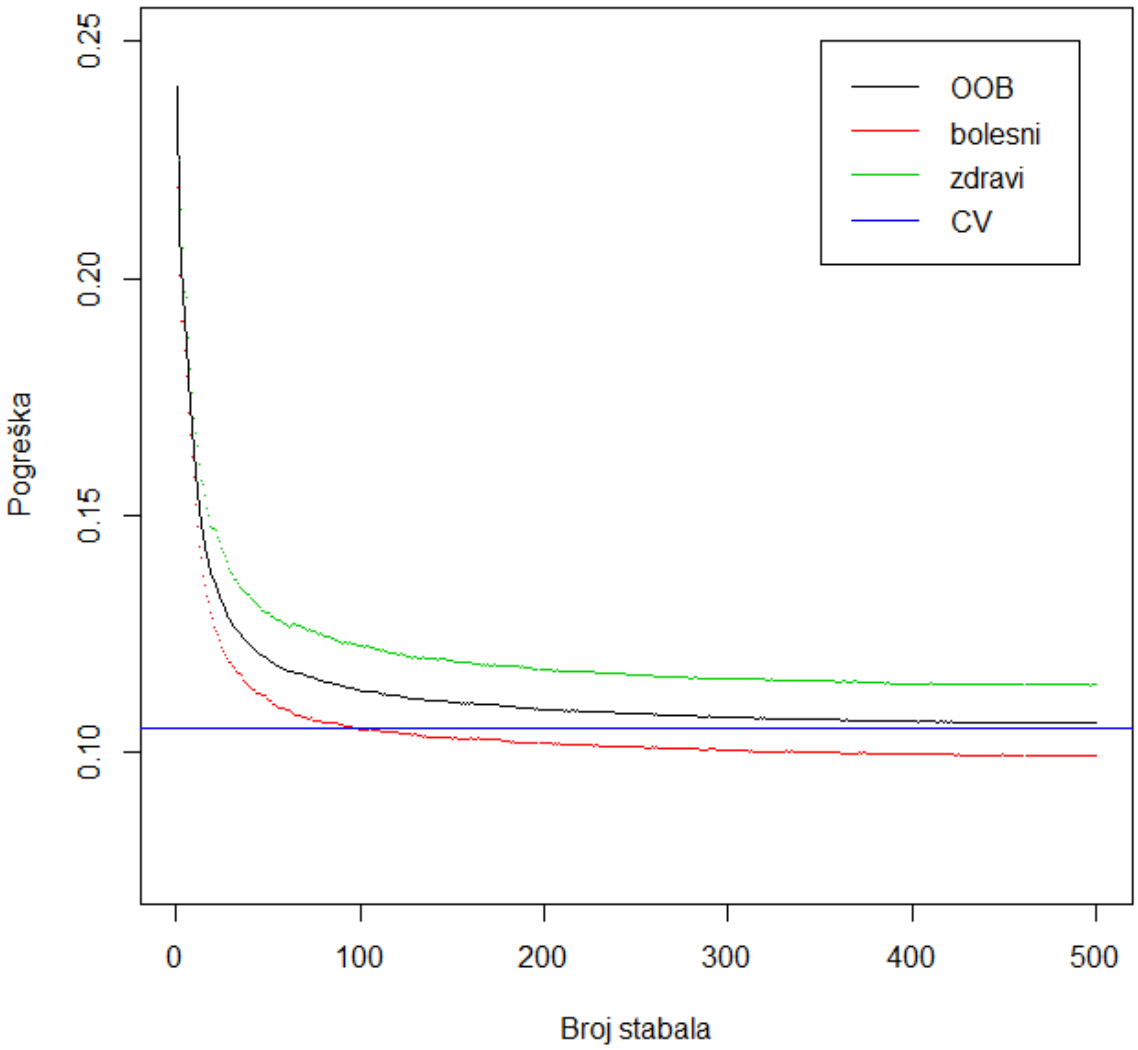
4.3. Rezultati klasifikacije gena korištenjem slučajnih šuma

Geni predviđeni programom MetaGeneMark za koje sam našla anotaciju u KEGG bazi korišteni su u daljnjoj analizi. Napravila sam klasifikaciju gena korištenjem slučajnih šuma i odredila klasifikacijsku pogrešku (slika 10) ovisno o broju izgrađenih stabala. Pogreška procijenjena unakrsnom validacijom iznosi 10.498% što se smatra vrlo niskom vrijednošću odnosno vrlo dobrom klasifikacijom.

Odredila sam važnost varijabli pri izradi slučajnih šuma (slika 9) kao prosjek važnosti varijabli u unakrsnoj validaciji. Najveći utjecaj ima varijabla top_0.85A koja označava A vrijednost izračunatu za 85-i kvantil vrijednosti MELP svih gena. Nakon toga slijede A vrijednosti 90-og, 95-og i 70-og kvantila, te M vrijednost 85-og kvantila i pripadna vrijednost povećanja ekspresije.



Slika 9: Važnost varijabli u izračunu slučajnih šuma prema Gini indeksu. Prikazane vrijednosti dobivene su kao srednja vrijednost važnosti u unakrsnoj validaciji.

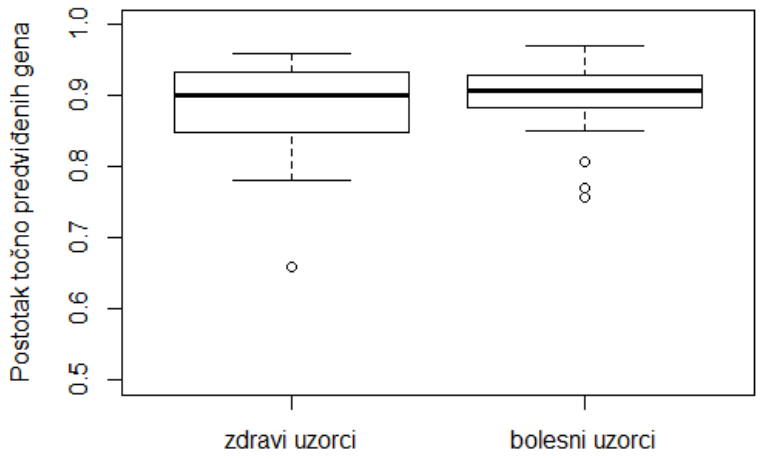


Slika 10: Pogreške u klasifikaciji. Izračunate je OOB (eng. „*Out of bag*“) procjena pogreške (crna linija), pogrešno klasificirani bolesni geni, te pogrešno klasificirani zdravi geni (crvena i zelena linija), te CV procjena pogreške (plava linija).

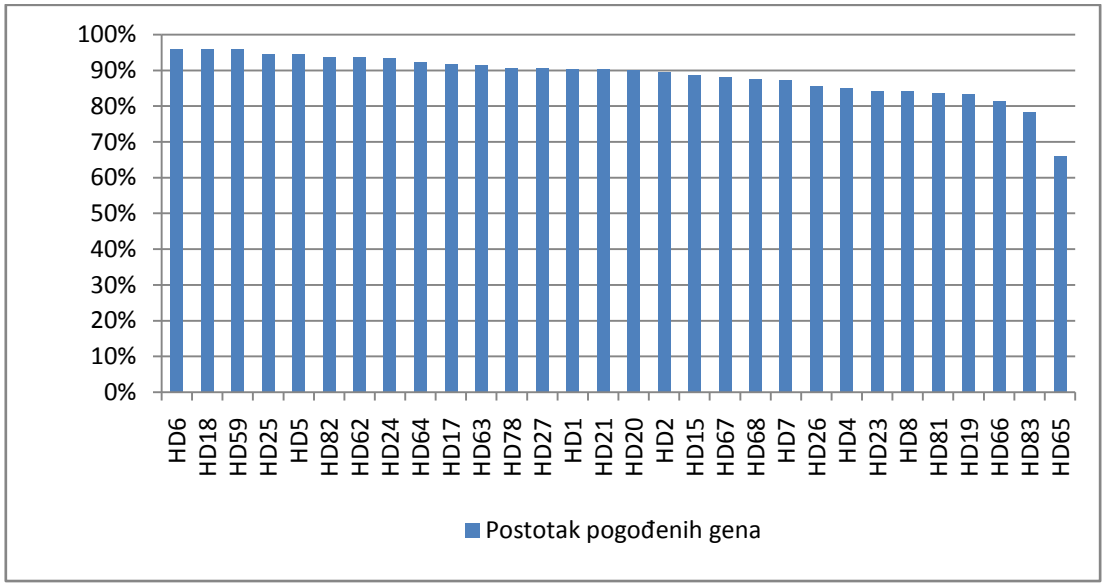
4.4. Rezultati klasifikacije uzoraka prema predviđenim genima

Na slici 11 prikazana je točnost predviđenih gena po zdravim i bolesnim uzorcima u obliku dijagrama pravokutnika. Srednja vrijednost točnosti predviđenih gena u zdravim uzorcima iznosi 88,54%, a u bolesnim 90,02%.

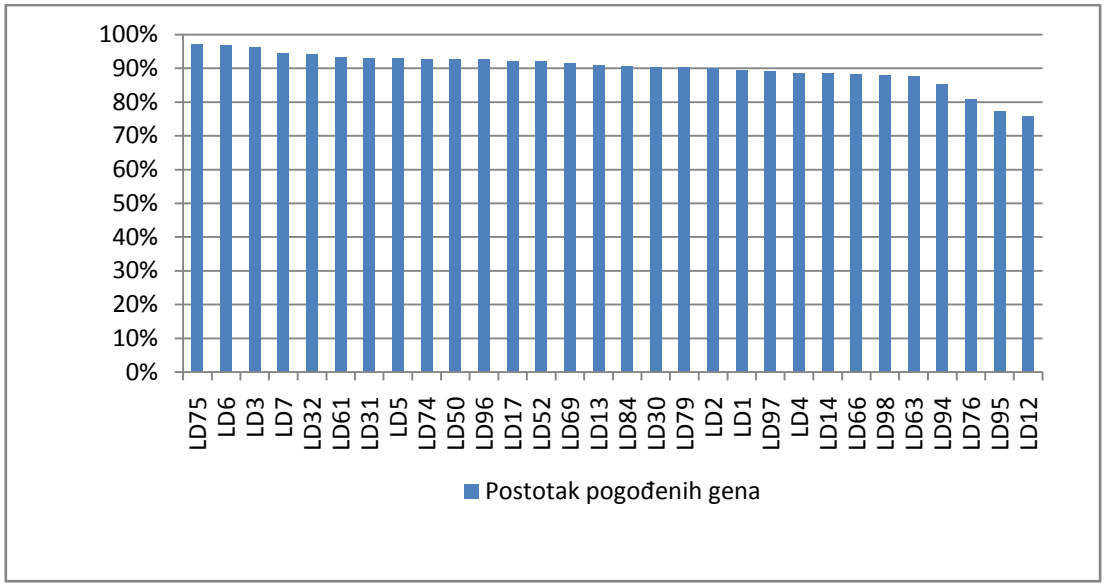
Na slikama 12 i 13 prikazani su rezultati točnosti predviđanja gena metodom slučajnih šuma po svakom uzorku, za zdrave i bolesne uzorke.



Slika 11: Dijagram pravokutnika točnosti predviđanja gena po zdravim i bolesnim uzorcima. Označena je srednja vrijednost, te raspon od prvog do trećeg kvartila u obliku pravokutnika, ostale vrijednosti obuhvaćene su linijama, dok su iznimke prikazane kao krugovi



Slika 12: Postotak pogođenih gena po svakom uzorku iz skupine zdravih uzoraka



Slika 13: Postotak pogođenih gena po svakom uzorku iz skupine bolesnih uzoraka

Prilikom predviđanja stanja osobe s obzirom na pogođene gene, korištenjem razboritog praga od 75% (ako je više od 75% gena pogođeno kao „bolesni“ odnosno „zdravi“ osoba se proglašava „bolesnom“ odnosno „zdravom“, inače se stanje proglašava nepoznatim), točno je predviđeno stanje 29 od 30 zdravih osoba i 30/30 bolesnih (Tablica 3).

Tablica 3: Stvarna (redovi) i pogođena (stupci) stanja ljudipri različitim pragovima odlučivanja o stanju osobe na temelju predviđenih gena

			Predviđeno stanje		
		prag	bolesni	nepoznato	zdravi
Stvarno stanje	bolesni	0.5	30	0	0
	zdravi		0	0	30
	bolesni	0.75	30	0	0
	zdravi		0	1	29
	bolesni	0.85	27	3	0
	zdravi		0	8	22

4.5. Predviđanje translacijski optimiranih gena u uzorcima

4.5.1. Analiza na razini gena

Pregledom najviših 30% eksprimiranih gena čija je M vrijednost bila najmanje 1, što odgovara dva puta većem zabilježenom broju gena u uzorku od očekivanog, s prilagođenom p vrijednosti 0,05 u zdravim i bolesnim uzorcima odredila sam translacijski optimirane gene. Najveća je razlika u predviđenoj ekspresiji u ovakvim uvjetima pronađena u genu kduD (nađen u 1 zdravom i 8 bolesnih uzoraka), rplY i rpmF (1 zdravi i 7 bolesnih), oadB (1 zdravi i 6 bolesnih), i kdgK (2 zdrava i 10 bolesnih). Geni koji su bili više eksprimirani u zdravim osobama su mcp (6 zdravih i 2 bolesna), glnB (5 zdrava i 2 bolesna) i dctM (4 zdrava i 2 bolesna). Potpuni dobiveni rezultati nalaze se u tablici 2 priloga.

4.5.2. Analiza na razini metaboličkih puteva

Tablica 4 prikazuje predviđeno obogaćenje na razini metaboličkih puteva po zdravim i bolesnim uzorcima.

Tablica 4: Razlike u predviđenom obogaćenju u ekspresiji na razini metaboličkih puteva. Prikazan je broj zdravih i bolesnih uzoraka u kojima je nađena M vrijednost (pridružena svakom metaboličkom putu) veća od 0,5. Ova vrijednost odgovara 1,41 puta većem od očekivanog zapaženom broju ortologa koji pripadaju istom metaboličkom putu, prilikom analize gena čija je predviđena ekspresija bila u gornjih 30%, uz prilagođenu p vrijednost manju ili jednaku 0,05. Prikazani su samo oni metabolički putevi za koje je omjer nađenih zdravih i bolesnih uzoraka veći od 1,5.

Opis	Zdravi	Bolesni	Omjer
Sinteza i razgradnja ketonskih tijela	6	1	6,00
Bakterijski proteini pokretljivosti	12	3	4,00
Sklapanje bičeva	15	4	3,75
Signalni put oksitocina	4	12	3,00
Bakterijska kemotaksija	14	5	2,80
Proteasom	6	14	2,33
Biosinteza valina, leucina i izoleucina	13	6	2,17
GABAergična sinapsa	4	8	2,00
Glutamanergična sinapsa	2	4	2,00
Metabolizam dušika	5	10	2,00
Egzosom	10	17	1,70
Put pentoza fosfata	6	10	1,67
Eksport proteina	15	23	1,53

5. RASPRAVA

Prvi korak ovog istraživanja bio je za svaki od 30 nasumično odabranih zdravih i bolesnih uzoraka preuzetih iz istraživanja Qin i sur. (2014) odrediti pripadne metagenome. Dobiveni rezultati u skladu su sa spomenutim istraživanjem. Prilikom slaganja pojedinačnih metagenoma dobila sam u prosjeku 6% manji broj neprekinutih sljedova DNA po uzorku. Uzrok razlici u dobivenim rezultatima može biti pretjerano filtriranje podataka prilikom kontrole kvalitete sljedova ili ne odstranjivanje k-mera koji se pojavljuju samo jednom, prilikom slaganja genoma programom SOAPdenovo. Ovakav rezultat u daljnjem postupku utjecao je na 20% manji ukupni broj predviđenih gena, od kojih je prosječno trećina uspješno identificirana u KEGG bazi ortologa, što je očekivani postotak.

Drugi korak je bio odrediti bitne karakteristike metagenoma (odnosno procijeniti kako su geni eksprimirani) analizom obrazaca upotrebe sinonimnih kodona korištenjem MILC vrijednosti i MELP statistike, i klasificirati gene u skupinu „bolesnih“ i „zdravih“ gena korištenjem slučajnih šuma kao metode strojnog učenja. Ovu sam metodu izabrala jer je prethodno s uspjehom korištena u analizama metagenoma (Supek i sur., 2010). Optimizacija translacije postoji na razini čitavog metagenoma, a izražene funkcije unutar metagenoma neovisne su o vrstama koje nalazimo u njemu (Burke i sur., 2011). Ovakav pristup omogućava traženje dominantnih funkcionalnih karakteristika koje nisu nužno povezane s brojnosti gena nego su odraz više organizacijske razine te daju „karakteristični funkcionalni otisak“ mikrobnog ekosustava (Roller i sur., 2013).

Osim srednje vrijednosti MELP statistike za svaki gen po uzorku, pri klasifikaciji sam koristila i ostale izvedenice dobivene iz ove mjere. Najveću važnost pri odluci o klasi u koju spada neki gen pokazale su „A“ vrijednosti. U budućim istraživanjima bilo bi dobro istražiti i druge mogućnosti transformacija podataka za koje vrijedi neovisnost varijance o srednjoj vrijednosti. Ovaj nedostatak nije imao utjecaja na metodu klasifikacije slučajnim šumama, no utjecao bi na analizu glavnih komponenti koju bi također bilo zanimljivo napraviti. Prilikom klasifikacije slučajnim šumama, predviđena „*out of bag*“ pogreška dobro je aproksimirala pogrešku izmjerenu unakrsnom validacijom. Ukupna točnost predviđanja bila je podjednaka za zdrave i bolesne osobe i u prosjeku je iznosila visokih 89%, no predviđanje je napravljeno na nasumično odabranom podskupu svih gena. U budućim istraživanjima bilo bi dobro kao skup na kojem se trenira metoda strojnog učenja uzeti čitave uzorke, a validaciju

raditi na nekoliko izuzetih uzoraka. Takva bi situacija bila sličnija realnoj u kojoj bi za metagenom dobiven iz osobe nepoznatog stanja trebali predvidjeti je li osoba zdrava ili bolesna. Ovakav pristup ovisi o varijabilnosti MELP vrijednosti gena unutar svakog uzorka, kao i varijabilnosti između uzoraka, što bi se također trebalo u budućnosti istražiti.

Iako je pristup slučajnim šumama dao dobar rezultat, predlažem isprobavanje „*boosting*“ metode koja obično daje još manju pogrešku u predviđanju od nasumičnih šuma (Ogutu i sur., 2011). Zanimljivo bi bilo istražiti razliku u predviđenoj ekspresiji svakog pojedinog gena izmađu zdravih i bolesnih osoba neparametarskim neuparenim statističkim testom, i za gene koji se razlikuju najviše napraviti klasifikaciju – drugim riječima naći gene markere za cirozu jetre.

Uz pomoć predviđenih gena za svaki uzorak odlučeno je li „zdrav“ ili „bolestan“. Za visoki prag odluke (od čak 85%) niti jedan uzorak nije pogrešno klasificiran, a točno ih je klasificirano 27 zdravih odnosno 22 bolesna, dok je za ostale proglašeno „nepoznato“ stanje. Pronalaskom markera ciroze jetre u budućim istraživanjima ovakva odluka bit će još točnija.

Posljednji korak u analizi metagenoma probavnog sustava kod bolesnika s cirozom jetre u ovom radu bio je predvidjeti koji su geni i metabolički putevi procijenjeni kao najvažniji prema zastupljenosti gena s visokom MELP vrijednosti. Kako bi ovo odredila, pregledala sam one gene čija je predviđena ekspresija u najviših 30%. Od njih sam izdvojila zdrave od bolesnih, uzela one koji su pokazivali dva puta veću brojnost od očekivane s prilagođenom p vrijednosti od 0.05 i izdvojila sve za koje je omjer pronađenih bolesnih i zdravih uzoraka bio veći od 2, odnosno 1.5 za metaboličke puteve. Ovakav izbor parametara djelomično je proizvoljan te bi bilo zanimljivo istražiti detaljnije koji postotak gena bi bilo najbolje proučavati. U svom istraživanju uzela sam prag od 30% najviše eksprimiranih jer sam za taj postotak, uz fiksirane ostale parametre, pronašla najveći broj uzoraka s „obogaćenim“ genima.

Najupečatljivija razlika u predviđenoj ekspresiji čitavih metaboličkih puteva vidi se u sintezi i razgradnji ketonskih tijela, koja je 6 puta više zastupljena u zdravim nego u bolesnim uzorcima. Ovakav rezultat vrlo dobro opisuje kliničku sliku jer je arterijski omjer ketonskih tijela smanjen u bolesnicima s cirozom jetre (Yamaoka i sur., 1998),

pa možemo pretpostaviti da se zbog smanjene količine ketonskih tijela u probavnom sustavu bolesnika s cirozom jetre ovaj metabolički put nema potrebu optimirati. Također sam uočila razliku u metaboličkim putevima vezanim za pokretljivost bakterija. U zdravim uzorcima češće su jače eksprimirani putevi za sklapanje bičeva, bakterijske proteine pokretljivosti i kemotaksiju, što se može objasniti nedostatkom izobilja hrane u zdravom probavnom traktu čovjeka, dok se kod ljudi s cirozom jetre često razvija dijabetes (neosjetljivost na inzulin) zbog povećane koncentracije glukoze u krvi (Nolte i sur., 1995). Zanimljiv je i pronalazak više zdravih uzoraka s akivnijom ekspresijom metaboličkog puta sinteze valina, leuicina i izoleucina nego bolesnih, ako znamo da nedostatak valina+leucina+izoleucina/fenilalanina+tirozina u plazmi korelira s težinom oštećenja jetre, i da se smatra kako je smanjenje posljedica bolesti jetre (Morgan et al., 1978), (Kawaguchi et al., 2011).

Markeri pronađeni na razini metaboličkih puteva u istraživanju Qin i sur. djelomično su potvrđeni i u ovom istraživanju. GABAergična sinapsa, glutamanergična sinapsa i metabolizam dušika obogaćen je u bolesnicima, no nije potvrđeno obogaćenje u ostalim navedenim metaboličkim putevima kod zdravih osoba. U budućim istraživanjima bilo bi zanimljivo detaljno proučiti uloge svih gena za koje je predviđena visoka ekspresija, i usporediti ih s dosadašnjim saznanjima.

6.ZAKLJUČAK

- Provela sam analizu metagenoma probavnog sustava 30 bolesnika s cirozom jetre i 30 zdravih pojedinaca
- Sastavljeni su genomi i predviđeni otvoreni okviri čitanja kojima sam našla genske ortologe usporedbom s KEGG bazom
- Na anotiranim predviđenim genima analizirala sam optimizaciju translacije MILC vrijednostima i MELP statistikom
- Korištenjem strojnog učenja, metodom slučajnih šuma klasificirala sam gene na zdrave i bolesne s greškom od 10.5%
- Prema predviđenim genima klasificirala sam uzorke na zdrave i bolesne s visokom specifičnošću
- Usporedbom MELP vrijednosti pronašla sam KEGG ortologe čija se ekspresija razlikuje u bolesnim osobama i zdravim pojedincima
- Pronašla sam metaboličke puteve koji se razlikuju u ekspresiji između zdravih i bolesnih osoba
- Analize optimizacije translacije u metagenomima u budućim istraživanjima mogli bi doprinjeti razumijevanju mehanizma interakcije mikrobnog metabolizma s metabolizmom čovjeka a time i nastanka različitih fizioloških stanja

7. LITERATURA

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., Thomas, T., 2011. Bacterial community assembly based on functional genes rather than species. Proc. Natl. Acad. Sci. U. S. A. 108, 14288–14293. doi:10.1073/pnas.1101591108

Cannarozzi, G.M., Schneider, A., 2012. Codon Evolution: Mechanisms and Models. Oxford University Press.

Feng, Z., Kallifidas, D., Brady, S.F., 2011. Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. Proc. Natl. Acad. Sci. U. S. A. 108, 12629–12634. doi:10.1073/pnas.1103921108

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., Nelson, K.E., 2006. Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312, 1355–1359. doi:10.1126/science.1124234

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M., 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331, 463–467. doi:10.1126/science.1200387

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389–409.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–205. doi:10.1093/nar/gkt1076

Karlsson, F.H., Fåk, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., Bäckhed, F., Nielsen, J., 2012. Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat. Commun. 3, 1245. doi:10.1038/ncomms2266

Kawaguchi, T., Izumi, N., Charlton, M.R., Sata, M., 2011. Branched-chain amino acids as pharmacological nutrients in chronic liver disease. Hepatol. Baltim. Md 54, 1063–1070. doi:10.1002/hep.24412

Keller, M., Hettich, R., 2009. Environmental Proteomics: a Paradigm Shift in Characterizing Microbial Activities at the Molecular Level. Microbiol. Mol. Biol. Rev. MMBR 73, 62–70. doi:10.1128/MMBR.00028-08

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinforma. Oxf. Engl. 25, 1966–1967. doi:10.1093/bioinformatics/btp336

Morgan, M.Y., Milsom, J.P., Sherlock, S., 1978. Plasma ratio of valine, leucine and isoleucine to phenylalanine and tyrosine in liver disease. Gut 19, 1068–1073.

NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., Baker, C.C., Di Francesco, V., Howcroft, T.K., Karp, R.W., Lunsford, R.D., Wellington, C.R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A.R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M.H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., Guyer, M., 2009. The NIH Human Microbiome Project. Genome Res. 19, 2317–2323. doi:10.1101/gr.096651.109

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., O'Neal, C., 1965. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc. Natl. Acad. Sci. U. S. A. 53, 1161–1168.

Noguchi, H., Park, J., Takagi, T., 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 34, 5623–5630. doi:10.1093/nar/gkl723

Nolte, W., Hartmann, H., Ramadori, G., 1995. Glucose metabolism and liver cirrhosis. Exp. Clin. Endocrinol. Diabetes Off. J. Ger. Soc. Endocrinol. Ger. Diabetes Assoc. 103, 63–74. doi:10.1055/s-0029-1211331

Ogutu, J.O., Piepho, H.-P., Schulz-Streeck, T., 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 5 Suppl 3, S11. doi:10.1186/1753-6561-5-S3-S11

Pedersen, S., Bloch, P.L., Reeh, S., Neidhardt, F.C., 1978. Patterns of protein synthesis in E. coli: a catalog of the amount of 140 individual proteins at different growth rates. Cell 14, 179–190.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L.J., Mering, C. von, Bork, P., 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. 42, D231–D239. doi:10.1093/nar/gkt1253

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich, S.D., Wang, J., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65. doi:10.1038/nature08821

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., Zhou, J., Ni, S., Liu, L., Pons, N., Batto, J.M., Kennedy, S.P., Leonard, P., Yuan, C., Ding, W., Chen, Y., Hu, X., Zheng, B., Qian, G., Xu, W., Ehrlich, S.D., Zheng, S., Li, L., 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature 513, 59–64. doi:10.1038/nature13568

Rappé, M.S., Giovannoni, S.J., 2003. The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394. doi:10.1146/annurev.micro.57.030502.090759

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/., n.d.

Rodríguez-Valera, F., 2004. Environmental genomics, the big picture? FEMS Microbiol. Lett. 231, 153–158.

Roller, M., Lucić, V., Nagy, I., Perica, T., Vlahovicek, K., 2013. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res. 41, 8842–8852. doi:10.1093/nar/gkt673

Sharp, P.M., Emery, L.R., Zeng, K., 2010. Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. B Biol. Sci. 365, 1203–1212. doi:10.1098/rstb.2009.0305

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Sørensen, M.A., Pedersen, S., 1991. Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. J. Mol. Biol. 222, 265–280.

Staley, J.T., Konopka, A., 1985. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annu. Rev. Microbiol. 39, 321–346. doi:10.1146/annurev.mi.39.100185.001541

Supek, F., Škunca, N., Repar, J., Vlahoviček, K., Šmuc, T., 2010. Translational Selection Is Ubiquitous in Prokaryotes. PLoS Genet 6, e1001004. doi:10.1371/journal.pgen.1001004

Supek, F., Vlahovicek, K., 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6, 182. doi:10.1186/1471-2105-6-182

The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease, 2014. . Cell Host Microbe 16, 276–289. doi:10.1016/j.chom.2014.08.014

Turnbaugh, P.J., Gordon, J.I., 2009. The core gut microbiome, energy balance and obesity. J. Physiol. 587, 4153–4158. doi:10.1113/jphysiol.2009.174136

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C., Knight, R., Gordon, J.I., 2007. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449, 804–810. doi:10.1038/nature06244

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74. doi:10.1126/science.1093857

Wilmes, P., Bond, P.L., 2004. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ. Microbiol. 6, 911–920. doi:10.1111/j.1462-2920.2004.00687.x

Wright, F., 1990. The “effective number of codons” used in a gene. Gene 87, 23–29.

Yamaoka, K., Kanayama, M., Tajiri, K., Yamane, M., Marumo, F., Sato, C., 1998. Clinical significance of arterial ketone body ratio in chronic liver disease. Digestion 59, 360–363.

Yan, A.W., Fouts, D.E., Brandl, J., Stärkel, P., Torralba, M., Schott, E., Tsukamoto, H., Nelson, K.E., Brenner, D.A., Schnabl, B., 2011. Enteric dysbiosis associated with a mouse model of alcoholic liver disease. Hepatol. Baltim. Md 53, 96–105. doi:10.1002/hep.24018

8. PRILOZI

Ostali rezultati

Tablica 1: Prilog: Rezultati slaganja metagenoma i predviđanja otvorenih okvira čitanja kod bolesnika i u zdravim uzorcima

Bolesni uzorci					Zdravi uzorci				
ID	**Broj kontiga nakon slaganja genoma i filtriranja**	**Broj predviđenih gena**	**Broj anotiranih gena**	**Postotak anotiranih gena**	**ID**	**Broj kontiga nakon slaganja genoma i filtriranja**	**Broj predviđenih gena**	**Broj anotiranih gena**	**Postotak anotiranih gena**
L1	17.007	40.313	13.100	32,5%	H1	20.148	44.249	13.428	30,3%
L2	14.235	30.565	10.997	36,0%	H2	20.251	43.579	14.218	32,6%
L3	39.889	61.505	19.796	32,2%	H4	19.877	38.637	12.939	33,5%
L4	19.313	44.325	14.641	33,0%	H5	25.549	57.021	17.909	31,4%
L5	10.774	27.383	9.070	33,1%	H6	25.287	58.439	18.510	31,7%
L6	33.303	103.532	37.391	36,1%	H7	39.289	77.564	25.777	33,2%
L7	13.799	29.378	9.218	31,4%	H8	54.783	120.174	40.714	33,9%
L12	17.537	47.871	13.981	29,2%	H15	35.881	82.601	27.061	32,8%
L13	16.552	30.763	9.216	30,0%	H17	23.513	57.359	20.244	35,3%
L14	36.794	82.928	29.145	35,1%	H18	30.154	64.200	22.117	34,5%
L17	20.753	44.444	16.800	37,8%	H19	15.600	49.699	18.032	36,3%
L30	40.312	96.839	32.315	33,4%	H20	14.543	47.096	17.761	37,7%
L31	16.604	32.933	10.183	30,9%	H21	14.107	38.885	13.327	34,3%
L32	36.740	76.699	25.472	33,2%	H23	33.490	77.571	25.803	33,3%
L50	20.111	56.838	21.146	37,2%	H24	10.917	29.507	9.892	33,5%
L52	20.107	52.217	17.564	33,6%	H25	12.275	30.448	11.110	36,5%
L61	16.461	48.812	16.395	33,6%	H26	20.292	54.378	17.912	32,9%
L63	14.287	34.587	10.960	31,7%	H27	25.905	61.199	19.201	31,4%
L66	17.895	42.168	14.397	34,1%	H59	30.475	57.971	19.161	33,1%
L69	18.088	56.947	20.443	35,9%	H62	31.954	75.552	24.211	32,0%
L74	38.485	80.739	27.700	34,3%	H63	34.888	79.211	25.243	31,9%
L75	48.485	128.660	42.726	33,2%	H64	12.891	36.709	10.593	28,9%
L76	32.722	77.278	29.418	38,1%	H65	37.192	93.504	31.558	33,8%
L79	54.622	116.392	40.240	34,6%	H66	30.788	68.332	20.878	30,6%
L84	34.206	98.542	34.045	34,5%	H67	30.322	75.716	24.728	32,7%
L94	24.345	64.903	20.995	32,3%	H68	32.704	76.036	24.481	32,2%
L95	17.284	61.570	20.585	33,4%	H78	15.912	37.127	10.944	29,5%
L96	10.869	31.183	11.421	36,6%	H81	23.307	51.947	16.032	30,9%
L97	36.306	99.639	34.333	34,5%	H82	8.662	27.477	10.252	37,3%
L98	14.050	45.943	17.037	37,1%	H83	18.291	43.150	14.028	32,5%

Tablica 2: Prilog: Potpuni rezultati predviđanja translacijski optimiranih gena kod bolesnih i zdravih osoba, za gene čija je MELP vrijednost po uzorku u top 30%, s M vrijednosti većom od 1 što odgovara dva puta većem zapaženom broju od očekivanogm, uz prilagođenu p vrijednost od 0.05. Prikazani su samo oni geni čiji je omjer zdravih i bolesnih uzoraka u kojima su nađeni veći od 1.5, i koji su nađeni u više od 4 uzorka.

Kratica/kratice gena	zdravi	bolesni	omjer
kduD	1	8	8,00
RP-L25, rplY	1	7	7,00
RP-L32, MRPL32, rpmF	1	7	7,00
oadB	1	6	6,00
kdgK	2	10	5,00
ftsZ	1	5	5,00
rho	1	5	5,00
RP-S14, MRPS14, rpsN	1	5	5,00
EEF2	2	9	4,50
mdh	2	9	4,50
fusA, GFM, EFG	3	12	4,00
ATPF0C, atpE	1	4	4,00
ATPVK, ntpK, atpK	1	4	4,00
cspA	1	4	4,00
infC, MTIF3	1	4	4,00
lacE, araN	1	4	4,00
PTS-Fru-EIIA, fruB	1	4	4,00
E2.4.1.5	0	4	4,00
galK	0	4	4,00
TPI, tpiA	5	17	3,40
mcp	6	2	3,00
purA, ADSS	4	12	3,00
yajC	4	12	3,00
kduI	2	6	3,00
RP-L34, MRPL34, rpmH	2	6	3,00
ACADS, bcd	1	3	3,00
E4.1.1.15, gadB, gadA, GAD	1	3	3,00
E4.3.1.19, ilvA, tdcB	1	3	3,00
infA	1	3	3,00
purH	1	3	3,00
purM	1	3	3,00
E2.7.1.90, pfk	4	11	2,75
glnB	5	2	2,50
ATPF1A, atpA	2	5	2,50
FARSB, pheT	2	5	2,50
gyrA	2	5	2,50
PTS-HPR	2	5	2,50
purC	7	17	2,43
RP-S20, rpsT	6	14	2,33
NDUFAB1	3	7	2,33
recA	3	7	2,33
VARS, valS	5	11	2,20
porA	6	12	2,00
htpG, HSP90A	5	10	2,00
htpG, HSP90A citopl.	5	10	2,00

Kratica/kratice gena	zdravi	bolesni	omjer
MARS, metG	4	8	2,00
metQ	4	8	2,00
PARS, proS	4	8	2,00
uxaB	4	8	2,00
dctM	4	2	2,00
ATPF1B, atpD	3	6	2,00
eda	3	6	2,00
nagB, GNPDA	3	6	2,00
cheY	2	4	2,00
gyrB	2	4	2,00
pyrE	2	4	2,00
QARS, glnS	2	4	2,00
adk, AK	8	15	1,88
E2.3.1.8, pta	8	15	1,88
MUT	8	15	1,88
RP-S19, rpsS	6	11	1,83
uxaC	6	11	1,83
RP-S1, rpsA	11	20	1,82
PCCB, pccB	5	9	1,80
KARS, lysS	8	14	1,75
secA	8	14	1,75
livK	7	4	1,75
SARS, serS	7	12	1,71
RP-L13, MRPL13, rplM	9	15	1,67
fucO	5	3	1,67
asd	3	5	1,67
ENO, eno	3	5	1,67
guaB	3	5	1,67
metK	3	5	1,67
RP-L29, rpmC	3	5	1,67
RP-L35, MRPL35, rpmI	3	5	1,67
RP-L18, MRPL18, rplR	8	13	1,63
ABC.PE.S	5	8	1,60
PRPS, prsA	12	19	1,58
GPI, pgi	9	14	1,56
RP-L7, MRPL12, rplL	9	14	1,56
RP-L31, rpmE	15	10	1,50
GARS, glyS1	12	18	1,50
RP-S7, MRPS7, rpsG	12	18	1,50
infB, MTIF2	10	15	1,50
secDF	10	15	1,50
LARS, leuS	8	12	1,50
E2.2.1.1, tktA, tktB	6	9	1,50
manB	6	9	1,50
cysI	6	4	1,50

R skripte

Kontrola kvalitete nemapiranih sljedova:

```
library("ShortRead")
args<- commandArgs(T)

setwd(args[3])
path <- args[1]
outpath <- args[2]
doit<-function(path,outpath){
f <- FastqStreamer(path,100000)
while(length(reads <- yield(f))){
  reads <- reads[nFilter(3)(reads)]
  reads <- reads[!(alphabetFrequency(quality(reads),as.prob=F)[,c("#")]>50)]
  seqs <- sread(reads)# get sequence list
  qual <- PhredQuality(quality(quality(reads)))# get quality score list as
PhredQuality
  myqual_mat <- matrix(charToRaw(as.character(unlist(qual))),nrow=length(qual),
byrow=TRUE)# convert quality score to matrix
  at <- myqual_mat == "23"# find positions of low quality
  letter_subject <- DNAString(paste(rep.int("N", width(seqs)[1]),collapse=""))#
create a matrix of Ns
  letter <- as(Views(letter_subject,start=1,end=rowSums(at)),"DNAStringSet")# trim
to length needed for each read
  injectedseqs <- replaceLetterAt(seqs, at, letter)# inject Ns at low quality
positions
  adapter <- paste(rep("N",max(width(injectedseqs))), sep="",collapse="")
  mismatchVector <- c(rep(0,width(adapter)))# allow no mismatches at each adapter
offset
  trimCoords <- trimLRPatterns(Rpattern=adapter, subject=injectedseqs,
max.Rmismatch=mismatchVector, ranges=T)
# Trim sequences looking for a right end pattern (polyN in this case)
# Gets IRanges object with trimmed coordinates
  kraj<-end(trimCoords)
  kraj[kraj<90]<-90
  seqs <- DNAStringSet(seqs,start=start(trimCoords),end=kraj)
  qual <- BStringSet(qual,start=start(trimCoords),end=kraj)
  qual <- SFastqQuality(qual)# reapply quality score type
  trimmed <- ShortReadQ(sread=seqs, quality=qual, id=id(reads))
  writeFastq(trimmed,outpath,"a",full=F,compress=F)
}
close(f)
```

Filtriranje rezultata i odabir najboljeg nakon slaganja metagenoma:

```
library(Biostrings)
args<- commandArgs(TRUE)
setwd(paste("/common/WORK/mfabijanic/assembly/array/",args[1],sep=""))
red <- vector()
j<-0
iis <- c(seq(from=31,to=59,by=2))
for(i in iis){
j<-j+1
path<-paste(c(args[1],".",i,".scafSeq"),collapse="")
x<-readDNAStringSet(path)
lista<-strsplit(as.character(x),"N")
y<-unlist(lista)
novi<-y[nchar(y)>499]
nn<-DNAStringSet(novi)
red[j]<-N50(width(nn))
if(red[j]==max(red)){
        writeXStringSet(nn,paste(c("Cut_",args[1],".fa"),collapse=""))
}
write(c("N50:",path,N50(nchar(novi)),"Ukupna duljina
(sum(nchar),length):",sum(nchar(nn
)),length(nn)),paste(c(args[1],"_info"),collapse=""),append=T)
}
```

Priprema ORFova nakon blastX s KEGG bazom:

```
library(Biostrings)
library(IRanges)
args<- commandArgs(T)

setwd(args[1])

zero_hits_found <- function(path,infileID,kegseq,ids){
# path - gdje se nalazi file sa svim IDevima (grep "# Query:" blastXrezultat.txt)
# kegseq - DNA string set sa svim sekvencama
  x_id <- scan(paste(path,infileID,sep=""),what="c")
  x_id <- x_id[grep("gene_id_*",x_id)]
  none_hits <- x_id[!(x_id %in% ids)]
  none_hits %in% names(kegseq)
  indexes <- which(names(kegseq)%in%none_hits)
return(indexes)
}
priprema_ORF <-
```

```
function(path,infile,infileID,infile_KEGG,infile_fasta,outfile,outfile_biljeska,ou
tfile_zero_hits,outfile_reversed,out
file_frameshifted,outfile_beztriista){
x <- read.table(paste(path,infile,sep=""),stringsAsFactors = F)
colnames(x)<-
c("id","KEGG","perc","length","mismatches","gap_opens","qfrom","qto","sfrom","sto"
,"eval","bitscore")
KEGG <-
read.table(infile_KEGG,stringsAsFactors=F,col.names=c("id","ko","ko2","ko3","ko4",
"ko5"),fill=T)
KEGG[,2]<- substring(KEGG[,2],4)
ids <- unique(x$id)
kegseq<-readDNAStringSet(infile_fasta)
frameshifted <- rep(0,length(ids))
imena <- vector()
sekvence<-vector()
kojenisuimaletrihita <- vector()
k<-0
krivoime<-vector()
reversed <- vector()
kkkk <-vector(length=length(ids))
for(i in1:length(ids)){
  kkkk[i]<-width(kegseq[ids[i]==names(kegseq)])
}
for(i in1:length(ids))#po unique id-evima od 1 do length(ids)
{
if(sum(!(x[x$id==ids[i],"qfrom"]<x[x$id==ids[i],"qto"]))==0)
{     ID <- ids[i]
        y <- x[x$id==ID,c("id","KEGG","qfrom","qto","bitscore")]
        y <- y[y$bitscore>60,c("id","KEGG","qfrom","qto")]
if(nrow(y)>2)
{
if(!length((grep(":",y[3,"KEGG"])))||(!length(grep(":",y[2,"KEGG"])))||
(!length(grep(":",y[1,"KEGG"]))))
{
          krivoime<-c(krivoime,i)
          istiid<-0
}else{
            ime <- KEGG[y[1,"KEGG"]==KEGG$id,2]
if((ime==(KEGG[y[3,"KEGG"]==KEGG$id,2]))&&(ime==(KEGG[y[2,"KEGG"]==KEGG$id,2])))#a
ko su prva tri retultata u blastu isti
KEGG id
{
```

```
                istiid <- 1
                okvir <- (y$qfrom-1)%%3==0#okvir citanja ako T
                frameshifted[i]<- frameshifted[i] + sum(!okvir)#koliko rezultata pod
id i nije u okviru
                temp <- logical()
                temp <- rep(F,kkkk[i])
for(j in1:nrow(y))
if(okvir[j])#samo za one koji su u okviru citanja
                temp <- ((1:kkkk[i]) %in%(y$qfrom[j]:y$qto[j])) | temp #radim
coverage
}else{
                istiid <- 0
                kojenisuimaletrihita <- c(kojenisuimaletrihita,i)#ako prva tri
rezultata nisu isti id ne uzimam ih
}
}
#(min(y$qfrom):max(y$qto))[temp]
if(istiid)
{
        stemp <-
subset(unlist(strsplit(toString(kegseq[ID==names(kegseq)]),"")),temp)
        sekvenca<- paste(stemp,collapse="")
print(i)
                k <- k+1#koliko se sekvenci rezalo
        imena[k]<-ime
                sekvence[k]<-sekvenca
}
}else{
        kojenisuimaletrihita <- c(kojenisuimaletrihita,i)#ako ne postoji 3
rezultata blasta ne uzimam ih
}
}else{
    reversed <- c(reversed,i)#ako je from>to ne uzimam ih
}
}

frameshifted_indexesinkeg<-
which(names(kegseq)%in%unique(x$id)[which(frameshifted!=0)])
writeLines(as.vector(t(cbind(paste(">",imena,sep=""),sekvence))),outfile)

writeXStringSet(kegseq[which(names(kegseq)%in%ids[reversed])],filepath=outfile_rev
ersed)
writeXStringSet(kegseq[frameshifted_indexesinkeg],filepath=outfile_frameshifted)
```

```
writeXStringSet(kegseq[zero_hits_found(path,infileID,kegseq,ids)],filepath=outfile
_zero_hits)
writeXStringSet(kegseq[which(names(kegseq)%in%ids[kojenisuimaletrihita])],filepath
=outfile_beztriista)

write(c("Frameshifted: ",unique(x$id)[which(frameshifted!=0)],
"No blast hits: ",names(kegseq[zero_hits_found(path,infileID,kegseq,ids)]),
"qfrom>qto: ",which(names(kegseq)%in%ids[reversed]),
"Krivo ime:", krivoime
), outfile_biljeska)
}
priprema_ORF(path = args[1],
             infile = args[2],
             infile_KEGG=args[3],
             infileID = args[4],
             infile_fasta = args[5],
             outfile= args[6],
             outfile_biljeska = args[7],
             outfile_zero_hits = args[8],
             outfile_reversed = args[9],
             outfile_frameshifted = args[10],
             outfile_beztriista = args[11]
)
```

Preimenovanje pripremljenih orfova:

```
library(Biostrings)
library(IRanges)
args<- commandArgs(T)

path <- args[4]
setwd(path)
infile_kegg_super <- args[1]#metabolism kegg super v2
file<- args[2]#fasta s KO
outfile <- args[3]#gdje se spremi
super <- read.table(infile_kegg_super,sep="\t",header=T,fill=T,stringsAsFactors =
F,quote = "",nrow=12177)
id <- unique(super$KEGG)
imena<-vector(length=length(id))
brojilo<-vector(length=length(id))
lista<-list()
for(i in1:length(id))
{
```

```
#brojilo[i] <- length(unique(super[super[,1]==id[i],2]))
#sum(brojilo!=1) daje 0
  imena[i]<- unique(super[super[,1]==id[i],2])
  lista[[i]]<- super[super[,1]==id[i],4]

}
pripadnost <- vector(length=length(id))

for(i in1:length(id)){
    pripadnost[i]<- paste(lista[[i]],collapse=" ")
}
kljuc <-
cbind(id,paste(paste(id,imena,sep="|definition|"),pripadnost,sep="|group|"))
prip <- readDNAStringSet(file)
pripid<- names(prip)
names<- vector(length=length(pripid))
sekvence <- vector(length=length(pripid))
for(i in1:length(pripid)){
  index<- which(pripid[i]==kljuc[,1])
names[i]<- kljuc[index,2]
  sekvence[i]<- as.character(prip[[i]])
}
writeLines(as.vector(t(cbind(paste(">",names,sep=""),sekvence))),outfile)
```

Računanje svih MELP vrijednosti pomoću calculateMelp funkcije

```
melp_all <- function(dir){
#currwd <- getwd()
#setwd(dir)
    uzorak <- list()
for(fileinlist.files(dir))
{
        uzorak[[file]]<- calculateMelp(file,RPKOs)
}
    separate <- function(l){
#l - list containing all melp datasets (called by melp_all(dir) function)
#returnes a list with names healthy and "disease"

        healthy <- list()
        disease <- list()
for(name innames(l))
{
if(length(grep("H",name))==1){
                one <- l[[name]]
```

```
                one$name <-
substr(name,gregexpr("_",name)[[1]][1]+1,gregexpr("\\.",name)[[1]][1]-1)
                healthy <- rbind(healthy,one)

}else{
                one <- l[[name]]
                one$name <-
substr(name,gregexpr("_",name)[[1]][1]+1,gregexpr("\\.",name)[[1]][1]-1)
                disease <- rbind(disease,one)
}
}
        healthy$healthy <- T
        disease$healthy <- F
data<- rbind(healthy,disease)
return(data)
}
data<- separate(uzorak)
return(data)
#setwd(currwd)
}
```

Slučajne šume i klasifikacija:

```
library(randomForest)
setwd(path)
load(file="enrichment_ALLseparately.Robj")
svi <- enrichment$KO
svi <- na.exclude(svi)
svi$sick <- as.factor(svi$sick)
set.seed(1755)
sampled <- sample(1:nrow(svi))

borders <- seq(from=1,by=11472,length.out=11)
borders[11]<- 114715
cv.randomforests <- list()
pred <- list()
means <- vector()
for(i in1:10){
  training_set <- svi[sampled[!((1:nrow(svi))%in%(borders[i]:(borders[i+1]-1)))],]
  test_set <- svi[sampled[((1:nrow(svi))%in%(borders[i]:(borders[i+1]-1)))],]
attach(training_set)
  rf.svi <- randomForest(sick~.-sick-desc-KO-name,data=training_set)
  cv.randomforests[[i]]<- rf.svi
```

```
  predicted <- predict(rf.svi,test_set)
  pred[[i]]<- predicted
table(test_set$sick,predicted)
  means[i]<-mean(test_set$sick!=predicted)
detach(training_set)
print(i)
}
save(file="rf.cv_randomgenes_svi.Robj",cv.randomforests)
save(file="means_randomgenes_svi.Robj",means)
```

Rezultati CV:

```
library(randomForest)

setwd(path)
load(file="rf.cv_randomgenes_all.Robj")
varImpPlot(cv.randomforests[[2]])
plot(cv.randomforests[[3]])
cv.randomforests[[3]]
cv.all <- sapply(cv.randomforests,function(x){
                        x$confusion
})
cv.confmeans <- apply(cv.all,1,function(x){sum(x)/10})
cv.confmeans_matrix <- matrix(cv.confmeans,nrow=2,ncol=3)
mean(cv.confmeans_matrix[,3])
plot(cv.confmeans_matrix,log="y")
cv.randomforests[[1]]$err.rate
cv.error <- sapply(cv.randomforests,function(x){
            x$err.rate
})
load(file="means_randomgenes_all.Robj")
cv.mean.error <- matrix(c(OOB <- apply(cv.error[1:500,],1,mean),
          bolesni <- apply(cv.error[501:1000,],1,mean),
          zdravi <- apply(cv.error[1001:1500,],1,mean)),
          byrow = F,ncol=3)
plot(1:500,cv.mean.error[,1],type="l",ylim=c(0.075,0.25),pch= ".",ylab
="Pogreška",xlab="Broj stabala")
points(1:500,cv.mean.error[,2],col=2,pch=".")
points(1:500,cv.mean.error[,3],col=3,pch=".")
legend(350,0.25,c("OOB","bolesni","zdravi","CV"),col=c(1,2,3,4),lty=1)
abline(a=mean(means),b=0,col=4)
plot(sort(cv.randomforests[[1]]$importance[,i], dec = TRUE),type = "h")
```

```
importance_all<- sapply(cv.randomforests,function(x){
    x$importance
})
rownames(importance_all)<-rownames(cv.randomforests[[2]]$importance)
importance_all
mean_imp <-apply(importance_all,1,mean)
#mean_imp<-sort(mean_imp,decreasing = T)
cv.rf_mean <- cv.randomforests[[1]]
mat <- as.matrix(mean_imp)
colnames(mat)<- "MeanDecreaseGini"
cv.rf_mean$importance <- mat
varImpPlot(cv.rf_mean,sort = T,main="Važnost varijabli")
```

Obrada podataka dobivenih predviđanjem slučajnim šumama:

```
pred <- list()
means <- vector()
test_setsick <- list()
test_setnames <- list()
for(i in1:10){
 test_set <- svi[sampled[((1:nrow(svi))%in%(borders[i]:(borders[i+1]-1)))],]
      test_setsick[[i]]<- test_set$sick
      test_setnames[[i]]<- test_set$name
      rf.svi <- cv.randomforests[[i]]#<- randomForest(sick~.-sick-desc-KO-
name,data=training_set)
      predicted <- predict(rf.svi,test_set)
      pred[[i]]<- predicted
      table(test_set$sick,predicted)
      print(i)
}
matrica <- matrix(ncol=3)
matrica_a <- matrix(ncol=3)
for(i in1:10){
  test<-ifelse(test_setsick[[i]]=="zdravi","zdravi","bolesni")
  predicted <- ifelse(pred[[i]]=="zdravi","zdravi","bolesni")
  matrica_a <- matrix(c(test,predicted,test_setnames[[i]]),ncol=3)
  matrica <- rbind(matrica,matrica_a)
}
matricasorted <- matrica[order(matrica[,3]),]
ve<- ifelse(matricasorted[,1]==matricasorted[,2],1,0)
matricasorted <- cbind(matricasorted,ve)
colnames(matricasorted)<- c("test","predicted","name","same")
```

```
matricasorted<-na.exclude(matricasorted)
v <- vector()
for(i inunique(matricasorted[,3]))
{
  v[i]<-
sum(as.numeric(matricasorted[matricasorted[,3]==i,4]))/length(matricasorted[matric
asorted[,3]==i,4])
}
plot(1:30,sort(rezultat[1:30,]),type="b",col=3,ylab="Udio pogođenih
gena",xlab="Uzorci",ylim=c(0,1))
lines(1:30,sort(rezultat[31:60,]),type="b",col=2)
abline(a=0.80,b=0,col=1)
summary(rezultat[1:30,])
summary(rezultat[31:60,])
axis(labels=c("zdravi uzorci","bolesni uzorci"),side=1,at=c(1,2))
odluka <- function(prag){
  z <- ifelse(rezultat[1:30,1]>prag,"zdravi","nepoznato")
  b <- ifelse(rezultat[31:60,1]>prag,"bolesni","nepoznato")
table(c(rep("zdravi",30),rep("bolesni",30)),c(z,b))
}
odluka(0.5)
odluka(0.75)
odluka(0.85)
Analiza diferencijalno eksprimiranih gena/puteva

setwd("C:/Users/maja_2/Desktop//Diplomski rad")
load("enrichment_ALLseparately.Robj")
getall <- function(z,b){
  this <- merge.data.frame(z,b,by.x="Group.1",by.y="Group.1",all=T)
  omjer <- ifelse(this$x.x>this$x.y,this$x.x/this$x.y,this$x.y/this$x.x)
  this <- cbind(this,omjer)
  this$omjer[is.na(this$x.x)]<- this$x.y[is.na(this$x.x)]
  this$omjer[is.na(this$x.y)]<- this$x.x[is.na(this$x.y)]
  this <- this[order(this$x.y,decreasing=T),]
  this[is.na(this)]<-0.01

par(mfrow=c(2,1))
colnames(this)<- c("description","zdravi","bolesni","omjer")
plot(1:length(this$bolesni),this$bolesni,type="l",main="bolesni/zdravi",col=2)
lines(1:length(this$bolesni),this$zdravi,col=1)
  this <- this[order(this$zdravi,decreasing=T),]

plot(1:length(this$zdravi),this$zdravi,type="l",main="zdravi/bolesni")
```

```
lines(1:length(this$zdravi),this$bolesni,col=2)
par(mfrow=c(1,1))
  this <- this[order(this$omjer,decreasing=T),]
return(this)
}
getALL <- function(x,pval,enrval,level,omjer=1){
# x - percentil na kojem gledamo enrichment
# pval - pripadna padj vrijednost
# enrval - koliki enrichment zelimo
# level - "KO" ili "C"
# omjer - koji omjer će biti prikazan? geni koje nalazimo u x bolesnih i y
zdravih,
#x/y odnosno y/x je omjer koji dajemo

  svi <- enrichment[[level]]
  svi <- na.exclude(svi)
  svi$sick <- as.factor(ifelse(grepl("L",svi$name),"bolesni","zdravi"))
  svi <- svi[svi[,paste(c("top_",x,".padj"),collapse ="")]<=pval,]
  zdravi<-svi[svi$sick=="zdravi",]
  bolesni<-svi[svi$sick=="bolesni",]
#zdravienr<- zdravi
#bolesnienr<- bolesni
  zdravienr<-zdravi[zdravi[,paste(c("top_",x,".M"),collapse ="")]>enrval,]
  bolesnienr<-bolesni[bolesni[,paste(c("top_",x,".M"),collapse ="")]>enrval,]
  bolesnienr <- cbind(bolesnienr[,paste(c("top_",x,".M"),collapse
="")],bolesnienr$desc)
  zdravienr <- cbind(zdravienr[,paste(c("top_",x,".M"),collapse
="")],zdravienr$desc)

  agrbolesnienr <-aggregate.data.frame(bolesnienr[,2],list(bolesnienr[,2]),length)
  agrzdravienr <- aggregate.data.frame(zdravienr[,2],list(zdravienr[,2]),length)

  zdravi_KO<- agrzdravienr[order(agrzdravienr[,2],decreasing=T),]
  bolesni_KO <- agrbolesnienr<-
agrbolesnienr[order(agrbolesnienr[,2],decreasing=T),]
  x<-getall(zdravi_KO,bolesni_KO)
return(x[x$omjer>=omjer,])
}

lapply(c(0.95,0.9,0.85,0.7),function(x){sum(c(getALL(x,0.05,1,"KO",2)$zdravi,getAL
L(x,0.05,1,"KO",2)$bolesni))})
lapply(c(0.95,0.9,0.85,0.7,0.5),function(x){sum(c(getALL(x,0.05,0.5,"C",1.5)$zdrav
i,getALL(x,0.05,0.5,"C",1.5)$bolesni))})
```

```
# najbolji KO je 0.7
# najbolji C je 0.7

write.table(file="rezultatiKO.txt",getALL(0.7,0.05,1,"KO",1.5),sep="\t")
write.table(file="rezultatiC.txt",getALL(0.7,0.05,0.5,"C",1.5),sep="\t")
```

ŽIVOTOPIS

Maja Fabijanić
Cuglini 9, 10040 Zagreb (Croatia)
+385922615204
maja.fabijanic@gmail.com
Skype maja.fabijanic3
Datum rođenja 05/04/1989

OBRAZOVANJE

2013–2015
Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Biološki odsjek, diplomski studij molekularne biologije

2010–2013 **Prvostupnik molekularne biologije**

Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Biološki odsjek, preddiplomski studij molekularne biologije

2007–2010
Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, preddiplomski studij matematike

RADNO ISKUSTVO

2013–2014 Paul Hartmann d.o.o.
Pomoćnik – student u odjelu „Customer service“

VJEŠTINE

Jezici: Engleski jezik

Rad na računalu: Linux
R
C++
MS Office

Hobiji: Sportski ples (SPK Petrinia – Petrinja)
Ronjenje (Društvo istraživača mora „20000 milja“)