SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

BIOLOŠKI ODSJEK

Sara Sumić

Računalna analiza nekodirajućeg dijela transkriptoma u modelu embrionalnog razvoja miša (*Mus Musculus*)

Zagreb, 2011.

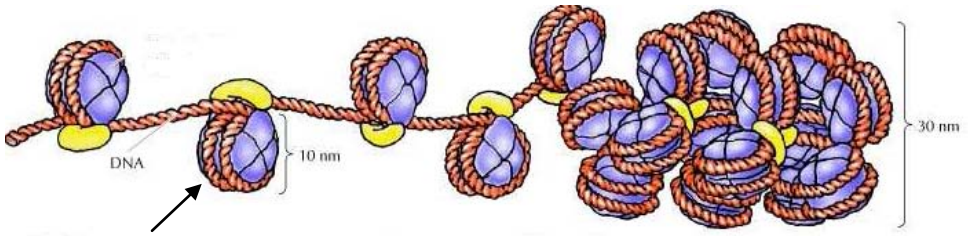
Ovaj rad izrađen je u sklopu Grupe za bioinformatiku pri Zavodu za molekularnu biologiju Biološkog odsjeka Prirodoslovno-matematičkog fakulteta u Zagrebu, pod vodstvom prof.dr.sc. Kristiana Vlahovičeka i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2010./ 2011.

Sadržaj

1. UVOD

1.1. Organizacija i struktura eukariotskog genoma

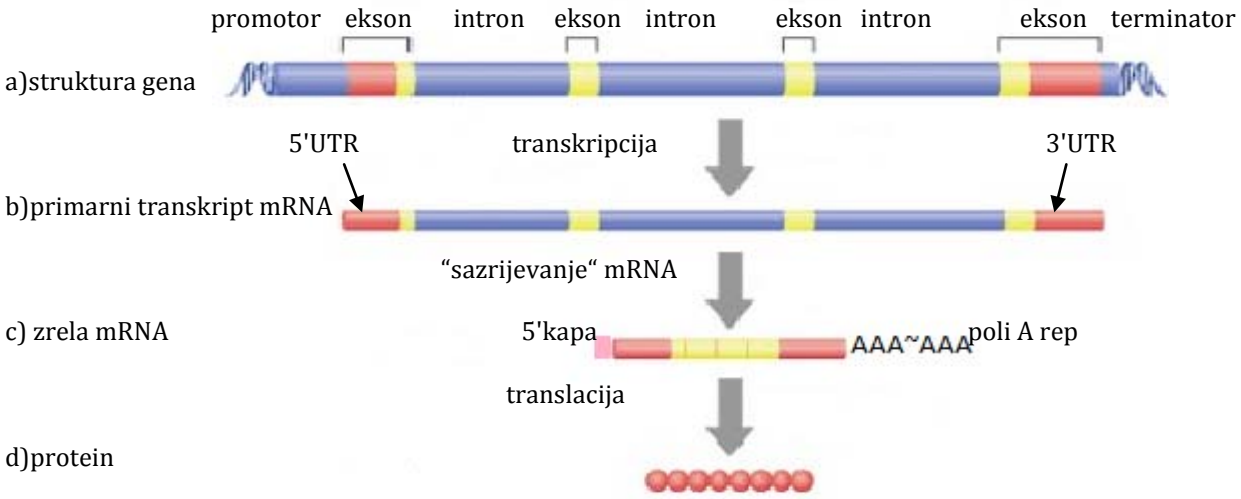
Eukariotski genom čini skup linearnih dvolančanih molekula DNA složenih u strukturu kromatina, koji je smješten u poseban odijeljak stanice – jezgru. Kromatin uz DNA sadrži i proteine (histonske i nehistonske) te omogućava kompaktno pakiranje DNA u jezgri, ali i određenu razinu regulacije jezgrinih procesa poput transkripcije, replikacije i rekombinacije. Osnovna jedinica kromatina je nukleosom, a čine ga DNA omotana oko oktamerne jezgre proteina histona (Slika 1). Nukleosomi su raspoređeni duž genoma tvoreći nukleofilament koji daljnjim smatanjem u strukture višeg reda (kondenzacijom) tvori vrlo gusto složeni metafazni kromosom (Babu i Verma 1987). Tijekom interfaze kromatin je u manje gustom stanju i organiziran u funkcionalne domene: eukromatin i heterokromatin. Eukromatin se dekondenzira tijekom interfaze i stoga većinom nije gusto pakiran, dok je heterokromatin gusto pakiran i uglavnom ne mijenja svoju kondenzaciju tijekom staničnog ciklusa.



Slika 1. Shema kromatina. Sastavna jedinica kromatina je nukleosom (označen strelicom), kojeg čini DNA omotana oko oktamerne jezgre proteina histona (promjera ~10 nm)

Geni su odsječci molekule DNA koji sadrže informaciju o određenom genskom produktu, nositelju stanične funkcije – proteinu ili molekuli RNA. Struktura eukariotskih gena (shematski prikazana na Slici 2.a) je znatno složenija od gena prokariota (Lynch 2006). Čine je promotor, eksoni i introni, te terminator. Na promotoru započinje proces transkripcije, a na terminatoru završava. Postupkom transkripcije se s molekule DNA prepisuje uputa te nastaje molekula mRNA (glasnička RNA, od engl. *messenger RNA*). Primarni transkript sadrži eksone i introne, kao i dijelove na 3' i 5' kraju transkripta koji se ne prevode u protein (UTR, od engl. *untranslated regions*). Nakon dorade primarnog transkripta mRNA (Slika 2.b) procesima izrezivanja introna, dodavanja 5' kape te poliadenilacije (Slika 2.c), nastaje zrela mRNA koja po izlasku iz jezgre na ribosomu kao kalup diktira redoslijed aminokiselina u rastućem polipeptidnom lancu za vrijeme biosinteze proteina. Taj proces nazivamo prevođenje (translacija) određenog redoslijeda nukleotida mRNA, u slijed

aminokiselina proteina (Slika 2.d). Cjelokupni navedeni slijed događaja nazivamo ekspresijom gena (Latchman 1992).



Slika 2. Struktura eukariotskog gena i slijed genske ekspresije. a) Struktura gena: promotor, eksoni i introni, terminator. b) Transkripcija započinje na promotoru, a završava na terminatoru. Produkt transkripcije je primarni transkript mRNA, a sastoji se od eksona, introna, 5' UTR i 3'UTR (od engl. *untranslated regions*; dijelovi koji se ne prevode u protein). c) Doradom primarnog transkripta procesima izrezivanja introna, dodatku 5'kape i poliadenilacije, nastaje zrela mRNA d) Nakon izlaska iz jezgre, mRNA na ribosomu u procesu translacije kao daje uputu za biosintezu proteina.

Genom sisavaca se prepisuje s oba lanca DNA, odnosno u oba smjera (orijentacije) (Katayama, Tomaru i sur. 2005), a tako nastale transkripte možemo općenito podijeliti u dvije kategorije: kodirajući i nekodirajući. Prvi se odnose na molekule mRNA čiji nukleotidni slijed nosi uputu za biosintezu proteina (kodiraju za proteine), a potonji su funkcionalne molekule RNA koje se ne prevode u proteine (Chen 2011). Nekodirajućim transkriptima pripadaju veoma važne RNA molekule kao što su transfer RNA (tRNA), ribosomalna RNA (rRNA), male jezgrine RNA (snRNA), male jezgricine RNA (snoRNA), mikro RNA (miRNA) i druge (Jacquier 2009). Dok kod prokariota većina gena u genomu kodira za protein (ili tRNA i rRNA), uz izrazito maleni udio nekodirajuće DNA, iznenađujuća je činjenica da je kod eukariota većina DNA nekodirajuća, bilo da se radi o nekodirajućim transkriptima ili nekodirajućim regijama genoma koje se ne transkribiraju, poput ponavljajućih ili *cis*-regulatornih regija. Pretpostavljeno je da iako samo 2% genoma sisavaca sadrži kodirajuće regije, čak 50% sudjeluje u nekom obliku transkripcije (Baker 2011).

1.2. Regulacija ekspresije gena u eukariota

Svaka stanica u višestaničnom organizmu sadrži cjelokupni genom u svojoj jezgri. Visoko specijalizirane stanice, poput živčanih i mišićnih, eksprimiraju različite udjele cjelokupnog skupa gena u određenom trenutku. Takvu diferencijalnu ekspresiju gena stanica postiže koristeći odgovarajuće regulacijske mehanizme, i upravo je diferencijalna ekspresija gena ono što čini stanice različitih tkiva morfološki i funkcionalno različitima. Stoga je jedna od najvažnijih zadaća u razvoju i održanju višestaničnog organizma regulacija ekspresije gena (Latchman 1992).

Svaki korak u procesu genske ekspresije predstavlja potencijalnu kontrolnu točku u kojoj se njegova ekspresija može uključiti ili isključiti, kao i potaknuti ili smanjiti. Te kontrolne točke, odnosno razine regulacije uključuju: pakiranje kromatina i kromatinske modifikacije, proces transkripcije, procesiranje (dorada) RNA, translaciju, te post-translacijske modifikacije proteina. Stanica u svakom ovom koraku mora uložiti energiju za ekspresiju potrebnog gena. Stoga je regulacija same transkripcije kao jedne od prvih kontrolnih točaka ključna u definiranju skupa gena eksprimiranih u određenom trenutku (Latchman 1992; Jaenisch i Bird 2003). Kromatin svojim stupnjem kondenzacije regulira dostupnost za transkripciju. Visoko kondenzirani heterokromatin je onemogućuje (najvjerojatnije zbog toga što transkripcijski proteini ne mogu pristupiti DNA prekrivenoj histonima), dok su područja relativno razmotanog eukromatina obično mjesta aktivne transkripcije (Jaenisch i Bird 2003). Pozicioniranjem nukleosoma, odnosno omatanjem određenih dijelova DNA oko histonske jezgre regulira se aktivacija gena. S obzirom da se nukleosomi pozicioniraju na promotorskim regijama gena koji su inaktivni, aktivacija gena zahtijeva premještanje tih nukleosoma i oslobađanje promotorskih regija za dostupnost inicijacijskom kompleksu transkripcije (Shaffer, Wallrath i sur. 1993). Modifikacija histona (metiliranje, acetiliranje, fosforiliranje i dr.) uvelike utječe na strukturu kromatina, a time i na regulaciju transkripcije. Modifikaciju pojedinih aminokiselinskih ostataka na histonima obavljaju specifični enzimi: acetiltransferaze, kinaze i metiltransferaze histona (Cedar i Bergman 2009).

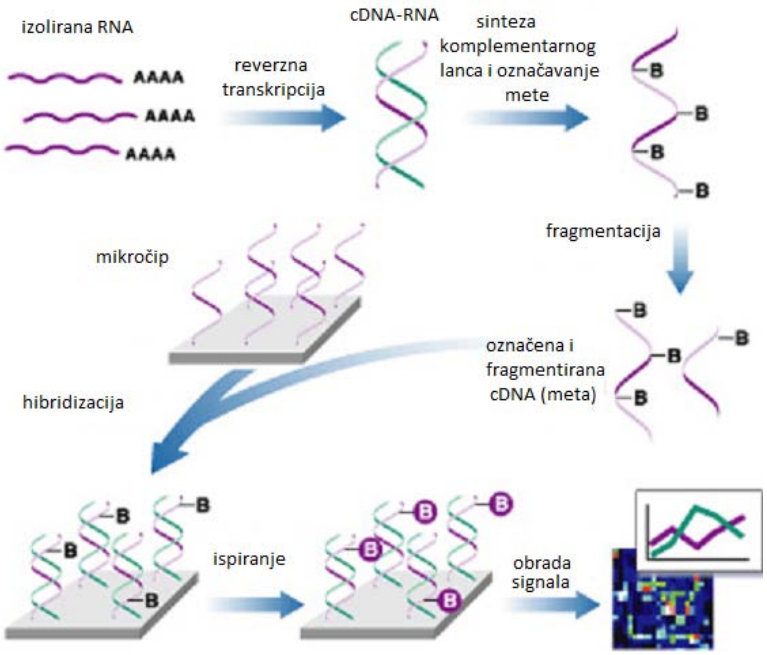
Tek nakon što je omogućena dostupnost za transkripciju, daljnja modulacija razine ekspresije postiže se interakcijom transkripcijskih faktora sa regulatornim sekvencama DNA, tzv. pojačivačima (od engl. *enhancers*) koji kontroliraju ekspresiju ciljnog gena putem pokretanja (inicijacije) transkripcije. Transkripcijski inicijacijski kompleks se nakon toga može formirati na promotoru ciljnog gena čija vrlo važna komponenta, RNA polimeraza, zatim obavlja transkripciju samog gena, stvarajući jednolančanu molekulu mRNA (Calkhoven i Ab 1996).

1.3. Mikročip tehnologija: mjerenje razine ekspresije cjelokupnog skupa gena

Cjelokupni skup molekula RNA u stanici nastalih procesom transkripcije nazivamo transkriptom, a sadržaj transkriptoma i razina pojedinih transkripata predstavljaju primarne odrednice fenotipa i funkcije stanice. Za razliku od postojanog genoma, transkriptom se tijekom staničnog ciklusa mijenja vrlo dinamično kao odgovor na okolišne uvjete ili sukladno trenutnim procesima koji se odvijaju u stanici (Jacquier 2009).

Danas su razvijene visokoprotočne metode kojima možemo ispitivati razinu ekspresije cijelog skupa staničnih gena (cjelokupnog transkriptoma) u određenim uvjetima, i koji predstavlja značajan tehnološki odmak od proučavanja ograničenog broja gena od uskog interesa. Jedna od najvažnijih eksperimentalnih metoda je zasnovana na tehnologiji DNA mikročipova (ili DNAčip, engl. *DNA microarrays*) kojom možemo kvantificirati količinu mRNA gotovo svih gena u željenim staničnim linijama u danim uvjetima. Glavna pretpostavka ove tehnologije jest da je količina mRNA transkripata u korelaciji sa količinom proteina za kojeg kodira, odnosno proporcionalna je razini genske ekspresije i možemo ju koristiti kao pouzdanu mjeru ekspresije gena (Pollack 2009).

Mikročip se sastoji od nekoliko stotina tisuća aktivnih polja koncentriranih na površini od npr. stakla ili silikona. Za njihovu je proizvodnju potrebna specijalna robotizirana oprema (Hwang 1997; Glazer, Fidanza i sur. 2006). Svako polje sadrži određenu poznatu tzv. *probu* ili *sondu* – jednolančane fragmente DNA ili oligonukleotide. S obzirom da su jednolančane, takve probe će na sebe vezati DNA ili RNA koje sadrže slijed komplementarnih baza, s kojima će tvoriti dvolančani lanac DNA u procesu hibridizacije (Slika 3.).



Slika 3. Shematski prikaz mikročip tehnologije. Izolirana RNA služi kao kalup za sintezu komplemetarnog lanca DNA u procesu reverzne transkripcije, rezultat čega je hibrid cDNA-RNA. Zatim se degradira lanac mRNA i na kalupu novosintetiziranog lanca DNA sintetizira njemu komplementaran. Nastaje dvolančana molekula DNA od koje se samo jedan lanac (meta; po svojoj orijentaciji odgovara početnoj mRNA) označava fluorescentnim dodatkom (preuzeto s: http://www.dkfz.de/gpcf/24.html)

Kad se na mikročip nanese stanični izolat koji sadrži nepoznate mete (nukleotidne sekvence, bilo jednolančanu DNA ili RNA), hibridizacija će se dogoditi samo na mjestima koja sadrže komplementarnu probu. Detekcija ovako hibridiziranih polja (na čvrstoj podlozi) omogućuje identifikaciju mete a ujedno i procjenu njene količine u stanici. Trenutno se detekcija hibridizacije na podlozi provodi korištenjem markera (fluorescentne boje) kojima su prethodno označene sve mete (Glazer, Fidanza i sur. 2006). Postupak ide na slijedeći način: RNA se prvo izolira iz željenog tkiva ili stanične kulture, potom provede reverzna transkripcija u DNA. Reverznu transkripciju obavlja enzim reverzna transkriptaza koji na kalupu mRNA sintetizira komplementaran lanac DNA. Zatim se degradira lanac mRNA i na kalupu novosintetiziranog lanca DNA sintetizira njemu komplementaran. Rezultat je dvolančana molekula DNA od koje se samo jedan lanac (meta; po svojoj orijentaciji odgovara početnoj mRNA) označava fluorescentnim (ili biotinskim) dodatkom (Pollack 2009). Hibridizacija tehnologijom mikročipova se stoga može koristiti za utvrđivanje razine ekspresije tisuće različitih molekula mRNA u samo jednom eksperimentu, što nam daje potpunu sliku stanja trenutnog transkriptoma u stanici (Eschrich i Yeatman 2004).

Nukleotidni slijed svake probe dizajniran je tako da odgovara nukleotidnom slijedu određenog kratkog dijela genoma, te im se pridružuju pripadne koordinate položaja u genomu (u postupku

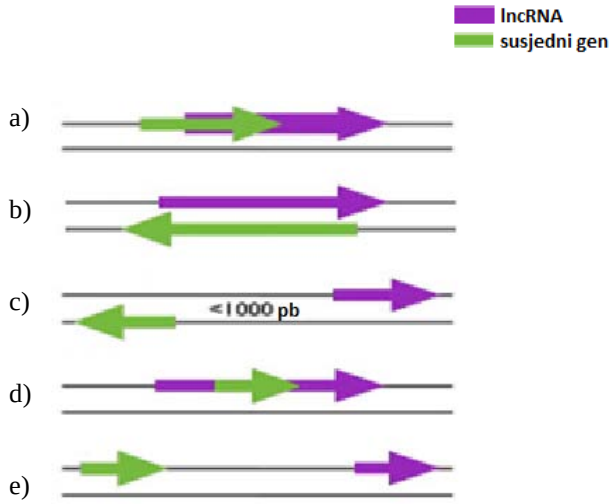
zvanom mapiranje i anotacija). Nukleotidni slijed skupa proba kojom detektiramo prisutnost određenih meta je osmišljen tako da probe međusobno mapiraju na isti gen, obično jedna iza druge. Skup proba sadrži između 10 i 20 proba, a najčešće 11 (Liu, Loraine i sur. 2003; Cheng, Sun i sur. 2004) . Kako bi iskoristili najveću prednost mikročip tehnologije, a to je opsežnost broja gena koje možemo ispitivati na samo jednom čipu, probe nisu dizajnirane samo na temelju dobro poznatih i anotiranih gena, već i na temelju manje poznatih sekvenci iz baze UniGene (Affymetrix 2010). UniGene je projekt izrade baza podataka koja sadrži sljedove EST (od engl. *expressed sequence tags*), odnosno kratke fragmenti transkribiranih RNA, složene u klastere koji odgovaraju nekom genu. Baza sadrži neredundantni skup sekvenci koje mapiraju na određene regije u genomu, bilo da se radi o anotiranim (opisanim) ili neanotiranim genima (Pontius, Wagner i sur. 2003). Dizajniranje proba na temelju svih sekvenci iz UniGene baze znatno povećava mogućnost detekcije eksprimiranih gena, bez obzira jesu li kodirajući ili nekodirajući.

1.4. Ekspresija nekodirajućih regija genoma

Kodirajući transkripti mRNA nisu jedina vrsta transkripata koji se mogu detektirati visokoprotočnim tehnologijama. Dok su neki od njih poput tRNA, rRNA i dr. relativno dobro opisani u literaturi, postoji još jedna vrlo zastupljena, no prilično slabo opisana klasa nekodirajućih transkripata. To su dugačke nekodirajuće RNA (lncRNA, od engl. *long noncoding RNA*) čiji su transkripti dulji od 200 nukleotida (Mercer, Dinger i sur. 2009), na temelju čega ih prvenstveno možemo razlikovati od različitih kratkih nekodirajućih transkripata poput snRNA, snoRNA, miRNA i dr. Transkripti lncRNA se, poput kodirajućih transkripata mRNA, dorađuju i sazrijevaju u procesu izrezivanja introna, dodavanja 5' kape i poliadenilacije (Hung i Chang 2010). Postojanje nekih lncRNA kao što su *H19* i *Xist* poznato je još od 1980-ih (Brannan, Dees i sur. 1990); otkrivene su tradicionalnim laboratorijskim metodama i pokazano je da su nekodirajuće tek analizom samih sekvenci kojima nedostaje otvoreni okvir čitanja (ORF, od engl. *open reading frame*) za translaciju u protein. U prilog nekodirajućim osobinama takvih transkripata govori i činjenica da se ne prenose iz jezgre u citoplazmu, nego se većinom nakupljaju u jezgri (Chen 2011). Unatoč tome, sve do nedavnog razvoja i porasta upotrebe visokoprotočnih metoda, smatralo se da je prisustvo molekula lncRNA u transkriptomu posljedica eksperimentalne pogreške.

Trenutno se smatra da je broj lncRNA u transkriptomu sisavaca znatno veći od broja kodirajućih transkripata (Carninci, Kasukawa i sur. 2005; Chodroff, Goodstadt i sur. 2010;

Jia, Osak i sur. 2010). Primjerice, procijenjeno je da mišji transkriptom sadrži otprilike 180 000 transkripata organiziranih u 44 000 transkripcijskih klastera, iako se samo za njih ~20 000 smatra da su kodirajući, dok je ostatak nekodirajuća RNA (Carninci, Kasukawa i sur. 2005). Unatoč tolikom broju nekodirajućih transkripata, dosad je kod miša anotirano samo ~2000 lncRNA transkripata (Flicek, Amode i sur. 2011). Shematski prikaz nekih od mogućih položaja lncRNA u genomu prikazani su na Slici 4.



Slika 4. Prikaz poznatih položaja lncRNA u genomu u odnosu na susjedni gen (obično kodirajući). a) transkript lncRNA ima jednaku orijentaciju kao i preklapajući gen na istovjetnom položaju. b) lncRNA je suprotne orijentacije od preklapajućeg gena, ali također na istovjetnom položaju u genomu. c) lncRNA je suprotne orijentacije od susjednog gena, ali nisu na istovjetnom genomskom položaju. d) lncRNA se nalazi unutar introna drugog gena. e) lncRNA se nalazi između drugih gena (intergenska lncRNA) (preuzeto s: http://mcmanuslab.ucsf.edu/node/251).

U svjetlu novih spoznaja o ulozi lncRNA u staničnim procesima otkriveno je da lncRNA naziva *HOTAIR* sudjeluje u privlačenju proteinskog kompleksa PRC2 na više od 800 mjesta u genomu, što dovodi do trimetilacije proteina histona a zatim i promjenu u ekspresiji (obično utišavanju) čitavog niza gena (Gupta, Shah i sur. 2010). Također je pokazano da lncRNA *HOTAIR* pokreće metastaze kod nekih tumora u mišu, te se njena ekspresija u raku dojke kod ljudi pokazala individualnim prognostičkim markerom za vrijeme preživljavanja i metastazu (Gupta, Shah i sur. 2010).

1.5. Miš (*Mus Musculus*) kao model embrionalnog razvoja sisavaca

Miš je modelni organizam kojeg je najlakše uzgojiti i manipulirati u svrhu eksperimenata (Spencer 2007), a ujedno je po svojoj fiziologiji dovoljno blizak čovjeku. Genom miša dijeli 85% sličnosti sekvence sa genomom čovjeka i približno je iste duljine. Također, mnogi geni su evolucijski

očuvani između ta dva organizma (Spencer 2007). Miš je posebno značajan u proučavanju embrionalnog razvoja zbog velike sličnosti sa čovječjim, tijekom kojeg se odvija vrlo striktna regulacija ekspresije gena, a posebice ranog embrija do stanja blastociste. Ono što omogućuje pravilan razvoj ranog embrija je iznimno aktivno tzv. genetičko reprogramiranje koje rezultira kaskadom ekspresije niza gena tijekom jednostaničnog i posebno dvostaničnog stadija ranog embrija (Li, Zheng i sur. 2010).

2. OPĆI I SPECIFIČNI CILJEVI RADA

2.1. Opći ciljevi rada

Standardna mikročip tehnologija jednog od vodećih proizvođača naziva *Affymetrix* iskazuje ekspresiju gena mjerenjem intenziteta hibridizacije ciljnih fragmenata DNA (meta) na skupove proba – u pravilu 11 proba duljine 25 nukleotida. Za svaki skup proba se potom računa srednja vrijednost intenziteta hibridizacijskog signala na svakoj od pojedinih proba (Dalma-Weiszhausz, Warrington i sur. 2006). No, u mnogim slučajevima, intenzitet signala nije u dobroj korelaciji sa razinom ekspresije odgovarajućeg gena. Razlozi tome mogu biti hibridizacija probe na transkript neodgovarajućeg gena, alternativno izrezivanje transkripata, probe koje mapiraju na više od jednog položaja u genomu ili imaju neodgovarajuću orijentaciju u odnosu na gen kojeg predstavljaju, mapiranje probe na introne odgovarajućeg gena i dr. (Nurtdinov, Vasiliev i sur. 2010). Prvi opis (anotaciju) proba *Affymetrix* priložen je u bazi NetAffx (Liu, Loraine i sur. 2003). Iako se nukleotidni slijed probi ovog proizvođača nije promijenio niz godina, naše znanje o genomu i transkriptomu kojeg predstavljaju raste veoma brzo. Stoga ne čudi da mnoge probe u anotaciji *Affymetrix*-a imaju jedinstveni položaj na genomu, no neovisnim analizama je određeno da ipak mapiraju na više položaja u genomu (npr. prema genomskom pregledniku Ensembl, www.ensembl.org). Takve probe mogu hibridizirati s bilo kojim od transkripata na nejednoznačno određenim položajima i nemoguće je odrediti o kojem se transkriptu radi. Također, ukoliko neka proba ili skup proba mapira na intron gena, razina intenziteta hibridizacije može biti znatno niža nego ukupna razina odgovarajućeg gena, jer se introni u najvećem broju slučajeva izrezuju prilikom dorade mRNA, dakle prije mjerenja. Isto se odnosi i na probe koje imaju neodgovarajuću orijentaciju obzirom na gen kojeg predstavljaju na mikročipu, jer nisu komplementarne sa njegovim transkriptom, te ne može doći do hibridizacije. Heterogeni skupovi proba se, s obzirom na sve gore navedene karakteristike neodgovarajućih proba sastoje od odgovarajućih probi u odnosu na gen kojeg predstavljaju i onih koji to nisu. Neki autori su izrazili zabrinutost u vezi same anotacije proba koja zastarijeva vrlo brzo, te predložili da autori prilikom objave rada na temelju korištenja mikročip podataka, objave i anotaciju koju su pritom koristili (Perez-Iratxeta i Andrade 2005).

Prilikom standardnog protokola za analizu podataka iz pokusa na mikročipovima uvijek se računa srednji intenzitet skupa proba kojeg predstavljaju sve probe u skupu, bez obzira da li se radi o heterogenim skupovima probi ili ne. Statistički je značajnije ukoliko imamo više proba koje predstavljaju razinu ekspresije nekog gena, te taj dio protokola nema potrebe mijenjati. Ono što je moguće napraviti jest „maskirati“, odnosno izostaviti neželjene probe iz heterogenih skupova probi

pri izračunu srednjeg intenziteta, kako ne bi imale utjecaj na jačinu signala hibridizacije dotičnog skupa probi i time doprinijele pogrešnoj procjeni razine ekspresije odgovarajućeg gena.

Stoga mi je glavni cilj u ovom radu razviti protokol za postupanje i računalnu metodu kojom bih najprije klasificirala sve probe obzirom na njihov položaj u genomu i orijentaciju (kao i one sa više položaja), te potom definirala vlastite skupove proba ostavljajući u originalnim skupovima samo one koje uistinu predstavljaju odgovarajuće gene. Metodom selektivnog odabira proba se također iz heterogenih skupova mogu izolirati samo one probe koje odgovaraju kodirajućim dijelovima (npr. eksonima), kao i nekodirajućim (npr. intronima), čime nam srednji intenzitet takvog novodefiniranog skupa može u dvije analize dati dvije zasebne informacije. Imajući na umu da se veliki dio lncRNA nalazi upravo u intronima kodirajućih gena (Nordstrom, Mirza i sur. 2009), mogućnost korištenja heterogenih skupova u analizi intenziteta nekodirajućih regija znatno proširuje naše mogućnosti detekcije signala za takvu klasu transkripata. Nadalje, ovom metodom je moguće najprije odrediti samo dio transkriptoma kojeg želimo ispitivati (primjerice kodirajući ili nekodirajući), te potom definirati i zasebno analizirati samo one skupove proba koji predstavljaju te transkripte, odnosno gene, na mikročipu. Prilikom obje analize moguće je i „maskirati“ probe koje mapiraju na više položaja u genomu, budući da unose grešku u dobiveni signal.

Danas je u nizu bioloških baza dostupna velika količina neobrađenih podataka dobivenih visokoprotočnim tehnologijama. U vrijeme kada su podaci dobiveni, autori su ih obradili i iskoristili u svrhu postavljene hipoteze, obično se pritom usredotočivši na kodirajuće transkripte i objavom rada u znanstvenom časopisu omogućili su pristup svojim neobrađenim podacima u biološkim bazama podataka (Edgar i Barrett 2006; Barrett, Troup i sur. 2009). Pojavom novih znanstvenih hipoteza postavlja se potreba za prikupljanjem novih eksperimentalnih podataka kako bi pronašli odgovore. Primjena računalnih metoda u kvalitetnijoj i ciljanoj obradi već postojećih podataka nam pruža mogućnost dobivanja novih informacija iz već dostupnih neobrađenih podataka.

Svrha ovog rada je stoga razviti računalnu metodu koja bi omogućila zasebnu analizu skupova probi koje odgovaraju nekodirajućim regijama u genomu u nastojanju rasvjetljavanja uloge dugačkih nekodirajućih transkripata (lncRNA) u stanici. U ovom radu to ću učiniti na primjeru podataka mikročip tehnologije koji predstavljaju stadije ranog embrija miša kao najvažnijeg modela embrionalnog razvoja sisavaca.

2.2. Specifični ciljevi rada

- Klasificirati sve probe sa željenih mikročipova obzirom na lokaciju u genomu prema najnovijoj anotaciji sa baze podataka i genomskog preglednika Ensembl, i to na: eksonske, intronske, intergenske te ih podijeliti i u skupine s obzirom na orijentaciju gena (smjer transkripcije).

- Definirati vlastite skupove proba za izračun srednjeg intenziteta s obzirom na gore navedenu podjelu . Posebni osvrt posvetit ću skupovima proba u kojima probe mapiraju isključivo na transkribirane nekodirajuće regije u genomu (introni, intergenske regije, te probe sa neodgovarajućom orijentacijom obzirom na pripadni ekson) .

- Anotirati navedene skupova proba usporedbom s dosad anotiranim nekodirajućim RNA (lncRNA) iz baze podataka Ensembl i FANTOM3 koje sadrže predviđene transkripte lncRNA

- Provesti zasebnu analizu skupova nekodirajućih proba i ispitati njihovu diferencijalnu ekspresiju između različitih stadija embrija miša: stadij jajne stanice, jednostanični, te dvostanični stadij ranog embrija.

3. MATERIJALI I METODE

3.1. Biološke baze podataka

3.1.1. Baza podataka i genomski preglednik Ensembl

Ensembl je zajednički projekt dvaju instituta: EMBL-EBI i The Welcome Trust Sanger Institute, sa svrhom anotacije niza eukariotskih genoma (Hubbard, Barker i sur. 2002; Spudich, Fernandez-Suarez i sur. 2007). Ensembl (http://www.ensembl.org/) sadrži široku lepezu informacija koja uključuje ne samo anotirane gene, transkripte gena, eksone, introne te predviđene gene određene računalnim metodama, već i poveznice prema mnogim drugim bazama podataka. Ensembl također omogućava pristup i slobodno preuzimanje svih navedenih podataka, kao i njihovu vizualizaciju putem web sučelja (Spudich, Fernandez-Suarez i sur. 2007; Sand, Thomas-Chollier i sur. 2009).

Svi podaci baze Ensembl pohranjeni su u MySQL bazama. MySQL je besplatan poslužitelj baza podataka (od engl. *database server*). Ensembl pruža mogućnost programskog pristupa podacima putem Perl API modula. Perl API (od engl. *Application Programme Interface*) je skup modula-skripti napisanih u Perl programskom jeziku koji se mogu koristiti za pristupanje i preuzimanje željenih podataka iz baze Ensembl. U svom radu sam putem Perl API odgovarajućih modula-skripti preuzela i koristila slijedeće podatke:

- genomske koordinate (položaj u genomu) proba koje pripadaju vrsti mikročipova MOE430A i MOE430B proizvođača *Affymetrix*
- genomske koordinate svih poznatih gena, te pripadnih eksona i introna za vrstu *Mus Musculus* (verzija genoma mm9)
- anotaciju navedenih gena prema njihovoj funkciji u stanici (na one koji kodiraju za proteine i nekodirajuće)

3.1.2. FANTOM3

FANTOM3 (od engl. *Functional Annotation of the Mouse*) je projekt instituta RIKEN, Japan (http://fantom3.gsc.riken.jp/), sa ciljem anotacije kompletnog mišjeg transkriptoma (Okazaki, Furuno i sur. 2002). Tijekom projekta FANTOM3 identificirano je 34,030 nekodirajućih transkripata sa oko10,000 različitih genomskih regija. Navedeni transkripti imaju mnoga obilježja molekula mRNA kao što je dodavanje 5' kape, izrezivanje, poliadenilacija, no nedostaje im ORF (Carninci, Kasukawa i sur. 2005), te su svrstani u klasu transkripata lncRNA. Uz prikupljanje iznimne količine eksperimentalnih podataka, u okviru projekta FANTOM3 razvijene su i metode računalne analize kojima je omogućeno bolje raspoznavanje kodirajućih od nekodirajućih transkripata (Maeda,

Kasukawa i sur. 2006). Stoga sam genomske pofložaje takvih nekodirajućih transkripata preuzela iz baze podataka FANTOM3 (Pang, Stephen i sur. 2005) te ih koristila za kasniju anotaciju pretpostavljenih nekodirajućih regija u vlastitoj računalnoj analizi.

3.1.3. GEO

Gene Expression Omnibus (GEO) je javna baza koja pohranjuje i besplatno distribuira podatke dobivene visokoprotočnim tehnologijama poput mikročip i *RNAseq* (Edgar, Domrachev i sur. 2002). Zbog potrebe da se podatci iz visokoprotočnih pokusa učine dostupnima cjelokupnoj znanstvenoj zajednici, upravo na način kako je to učinjeno za nukleotidne sljedove, svi znanstveni časopisi pred autore postavljaju zahtjev da neobrađene podatke na temelju kojih su publicirali rad prilože u navedenu bazu (Edgar i Barrett 2006).

Sve neobrađene podatke pokusa na mikročipovima u formatu zapisa .CEL sam prikupila iz baze podataka GEO (http://www.ncbi.nlm.nih.gov/geo). Prikupljeni podaci imaju pristupni broj GSE1749, a preuzela sam sljedeće podatke:

- stadij jajne stanice: OO1_MOE430A.CEL, OO2_MOE430A.CEL , OO3_MOE430A.CEL , OO4_MOE430A.CEL, OO1_MOE430B.CEL, OO2_MOE430B.CEL, OO3_MOE430B.CEL, OO4_MOE430B.CEL
- stadij jednostaničnog embrija: 1Cell1_MOE430A.CEL, 1Cell2_MOE430A.CEL, 1Cell3_MOE430A.CEL, 1Cell4_MOE430A.CEL, 1Cell1_MOE430B.CEL, 1Cell2_MOE430B.CEL, 1Cell3_MOE430B.CEL, 1Cell4_MOE430B.CEL
- stadij dvostaničnog embrija miša: 2Cell1_MOE430A.CEL, 2Cell2_MOE430A.CEL, 2Cell3_MOE430A.CEL, 2Cell4_MOE430A.CEL, 2Cell1_MOE430B.CEL, 2Cell2_MOE430B.CEL, 2Cell3_MOE430B.CEL, 2Cell4_MOE430B.CEL

3.1.4. UCSC Genome Browser

Baza podataka i genomski preglednik *UCSC Genome Browser* (dostupan na adresi http://genome.ucsc.edu/) sadrži potpune genomske zapise različitih organizama koji se mogu pretraživati i pregledavati na brojne načine. Pomoću ovog genomskog preglednika moguće je pronaći određeni gen, locirati ga na kromosomu i vizualizirati usporedno s mnoštvom drugih dostupnih informacija o željenoj regiji u genomu, kao što su anotacijske staze sa poznatim ili predviđenim genima, mnRNA, CpG otocima i sl. Korisnicima je na raspolaganju i mogućnost učitavanja vlastitih anotacijskih staza i na taj način vizualizirati i uspoređivati sa postojećim pohranjenim podacima

(Karolchik, Baertsch i sur. 2003). Uz genomski preglednik na sveučilištu Santa Cruz, SAD, ovaj preglednik je moguće instalirati i podesiti na lokalnom poslužitelju, za dodatnu fleksibilnost u radu s anotacijskim stazama.

Uz pomoć *UCSC Genome Browser*-a vizualizirala sam vlastite anotacijske staze na lokalno instaliranom pregledniku na poslužitelju grupe za bioinformatiku (format zapisa: .bedGraph) sa obrađenim podacima za svaki od stadija mišjeg embrionalnog razvoja.

3.2. Računalni programi i programski jezici

3.2.1. Perl

Perl je besplatan programski jezik koji je po svojoj arhitekturi tzv. tumačeni (od engl. *interpreted*) programski jezik,koji se sastoji od interpretera i programa-skripte. Interpreter čita i izvršava liniju po liniju programskog koda (instrukcija) . Perl programski jezik se primarno koristi na Unix računalima, no postoje i binarne inačice interpretera za Windows okruženje (Tisdall 2001).

U Perl programskom jeziku sam napisala program-skriptu za preuzimanje željenih podataka putem Perl API s baze Ensembl.

3.2.2. R

R je besplatan programski sustav za statističku obradu i grafičku vizualizaciju podataka (http://www.r-project.org/). R je dijalekt programskog jezika S (S-PLUS). Može se koristiti na različitim platformama poput Unixa,Windowsa i MacOS. Ukupno postoji 2725 tematskih statističkih paketa koji se lako instaliraju te dalje implementiraju. Paket je skup generičkih funkcija koje služe za niz standardnih statičkih obrada podataka. Neki od njih su automatski instalirani, te se mogu odmah koristiti na sučelju R-a (Pevsner 2009). Funkcije koje sam koristila u svojim programima dolaze iz takvih osnovnih paketa, kao i iz onih specijaliziranijih i opsežnijih koje je potrebno instalirati, te ću ih u daljnjem tekstu pobliže opisati.

3.2.3. Bioconductor

Bioconductor je projekt u sklopu kojeg nalazimo niz besplatnih programskih alata za analizu podataka dobivenih visokoprotočnim metodama na genomima različitih vrsta (Reimers i Carey 2006). Primarno koristi R programski jezik te se većina paketa, od njih trenutno 460 koje nudi Bioconductor, distribuira u obliku paketa za kroštenje u R-u. Osim paketa za analizu navedenih podataka, postoje i

paketi za anotaciju genoma različitih organizama. U svojoj analizi koristila sam slijedeće pakete: *IRanges, affy, gcrma i limma*.

3.3. Obrada mikročip podataka i statistička analiza

3.3.1. Opis (anotacija) probi i njihova klasifikacija s obzirom na položaj u genomu

Uz pomoć paketa *IRanges* u R-u izradila sam skriptu koja provjerava preklapanje genomskih koordinata svake probe mikročipova MOE430A i MOE430B s koordinatama svih poznatih gena miša, a posebno kodirajućih transkripata (pripadnih eksona i introna). Na temelju toga sam odredila odgovarajuće probe za svaki gen. Također sam ih klasificirala na eksonske, intronske te intergenske (koordinate proba koje se ne preklapaju sa koordinatama kodirajućih transkripata, već odgovaraju nekodirajućim transkriptima ili neanotiranim regijama između gena), te na temelju toga definirala kolekciju proba koje mapiraju na nekodirajuće regije (intergenske i intronske regije). Nakon učitavanja podataka uz pomoć *affy* paketa, provela sam „maskiranje“ neželjenih proba (u ovom slučaju probe koje mapiraju na kodirajuće regije) programom koji omogućuje da se one izostave, tj. „maskiraju“ prije daljnje analize (Gregory Alvord, Roayaei i sur. 2007). Ukoliko cijeli skupovi proba mapiraju na kodirajuće regije, maskiraju se čitavi takvi skupovi.

3.3.2. Obrada mikročip podataka metodom GCRMA

Normalizacija je proces korekcije intenziteta potrebnog prije usporedbe među mikročipovima. GCRMA (od engl. *GC Robust Multiarray Analysis*) je jedna od metoda za korekciju pozadinskog šuma u intenzitetima, normalizaciju te izračun srednjih intenziteta skupova probi (Pevsner 2009). Njome se provodi transformacija intenziteta kako bi korigirali nedosljednosti u samom eksperimentu (različite učinkovitosti fluorescencijskog označavanja meta, hibridizacije, različite količine početne RNA) kao i sistemske greške za vrijeme samog mjerenja intenziteta fluorescencije prilikom hibridizacije (Quackenbush 2002).

Kvartilna normalizacija mikročip podataka rezultira jednakom razdiobom visina intenziteta dobivenih na svim mikročipovima (Bolstad, Irizarry i sur. 2003). Provodi se tako da se vrijednosti intenziteta svakog skupa probi dodijeli kvartil kojem pripada obzirom na raspon svih vrijednosti. Zatim se za istovjetne skupove proba sa svih mikročipova izračuna njihova srednja vrijednost te im se dodijele nove vrijednosti intenziteta koje odgovaraju kvartilu izračunate srednje vrijednosti (Pevsner 2009). Navedeno je uključeno u metodu GCRMA, kao i korekcija za nespecifičnu hibridizaciju

obzirom na GC sastav probi, zbog čega se navedena metoda smatra vrlo točnom i preciznom. Tek nakon normalizacije intenziteta možemo međusobno uspoređivati intenzitete odgovarajućih skupova probi među mikročipovima te na temelju njihove razlike zaključiti o relativnoj promjeni razine ekspresije gena (Pevsner 2009).

Sve korištene podatke u .CEL formatu zapisa obradila sam metodom GCRMA uz pomoć paketa *gcrma* u R-u. Koristila sam normalizirane intenzitete u daljnoj obradi podataka te izradi histograma, *MA* grafova, mapi intenziteta, datoteka za vizualizaciju na *UCSC Genome Browser*-u (formati zapisa .bedGraph), kao i u priloženoj Tablici 2. Također sam definirane skupove probi mikročipova MOE430A i MOE430B objedinila u jedinstvenu tablicu, s obzirom da predstavljaju komplementarne skupove proba prema njihovom položaju u genomu.

3.3.3. Analiza diferencijalne ekspresije nekodirajućih regija genoma

Nakon koraka normalizacije slijedi analiza diferencijalne ekspresije gena – određivanje koji skupovi proba imaju statistički različite intenzitete među stadijima (predstavljeni odgovarajućim mikročipovima), što upućuje na različitu ekspresiju pripadnih gena. Navedeno je moguće provesti primjenom t-testa: izračun razlike u srednjoj vrijednosti intenziteta svakog od skupova probi podijeljenih sa procjenom varijance skupa. S obzirom da su kod ove analize svaki od stadija predstavljeni s 4 mikročipa , koristila sam Bayes-ov pristup koji je puno prikladniji od t-testa kada koristimo relativno malo tzv. bioloških replikata (Smyth 2004).

Bayes-ovim pristupom se umjesto s procjenom varijance specifične za svaki gen (tj. skup probi) razlika u srednjim vrijednostima intenziteta dijeli s modificiranom procjenom varijance, dobivene na temelju kombinacije navedene procjene varijance i procjene cjelokupne varijance, kako bi se dobili pouzdaniji rezultati (Baldi i Long 2001). Ovaj test implementiran je u *limma* paketu, te sam uz pomoć njega izradila Tablicu 2.

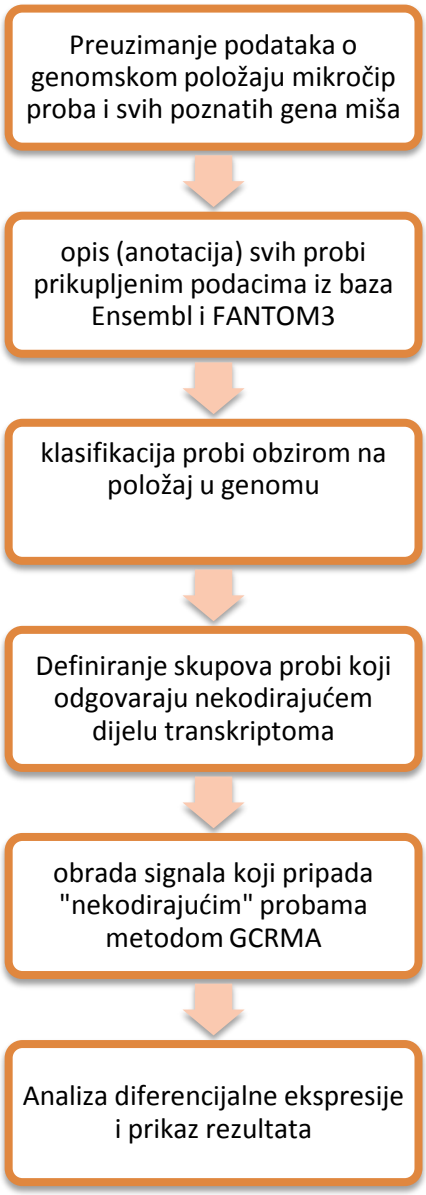
P vrijednost je mjera statističke značajnosti rezultata (razlike u visini intenziteta skupa probi između stadija kojeg predstavljaju mikročipovi). Niža p vrijednost ukazuje na statistički značajniji rezultat (granična vrijednost se često kreće od 0.01-0.05 za testiranje jedne vrijednosti rezultata). Kada istovremeno testiramo niz vrijednosti, granična p vrijednost od ~0.01 nam daje cijeli skup značajnih rezultata, a što je ona viša, odredit ćemo više vrijednosti rezultata kao značajne, uključujući i dio lažnih pozitiva. Potrebno je stoga prilagoditi dobivene p vrijednosti za višestruko testiranje. Koristila sam metodu po Benjamini-Hochbergu (Pevsner 2009) prilikm izrade mapi intenziteta i Tablice 2.

Mape intenziteta izradila sam uz pomoć paketa *gplots*, a prije vizualnog prikaza obavlja se i grupiranje gena po sličnosti razine ekspresije, kao i istovjetnih mikročipova (predstavljaju istovjetni biološki stadij). Za njihovu izradu koristila sam podskup gena sa različitom ekspresijom među stadijima, koje sam odredila na temelju dodijeljene im p vrijednosti, prilagođene za višestruko testiranje (uzeta granična prilagođena p vrijednost: 0.05).

4. REZULTATI

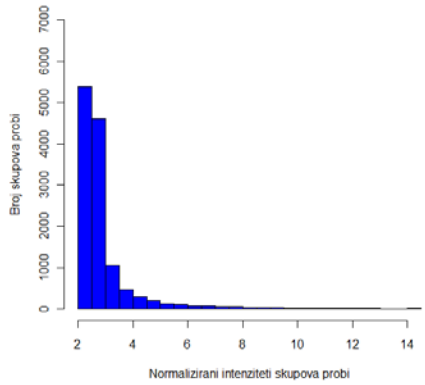
4.1. Protokol za analizu podataka iz pokusa na mikročipovima

Razvijeni protokol za postupanje napisala sam uz pomoć Perl i R programskog jezika. Izvorni kod skripti priložen je u Dodatku, a slijed rada shematski prikazan na Slici 5.

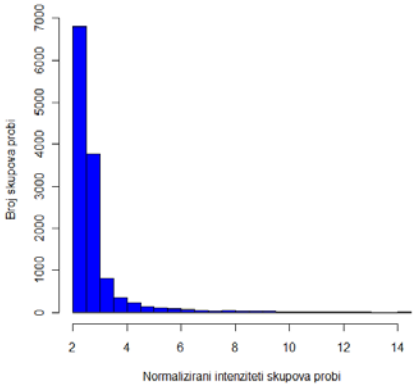


Slika 5. Shematski prikaz koraka korištenih u protokolu analize nekodirajućeg dijela transkriptoma

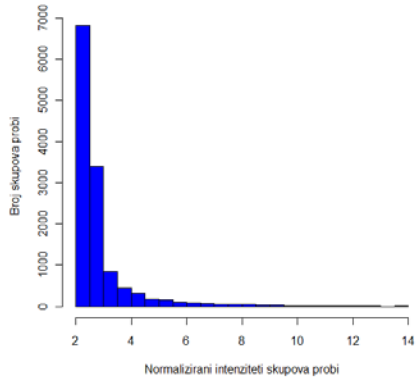
4.2. Histogrami razdiobe normaliziranih intenziteta



Histogram A.



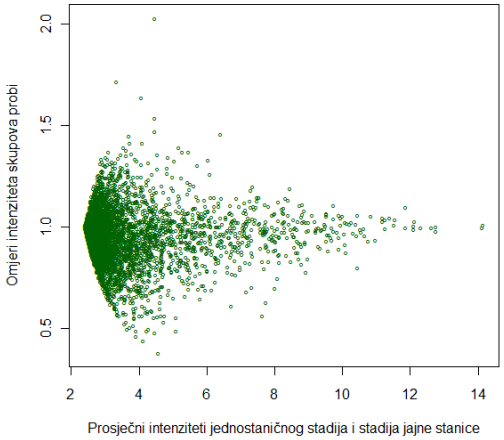
Histogram B.



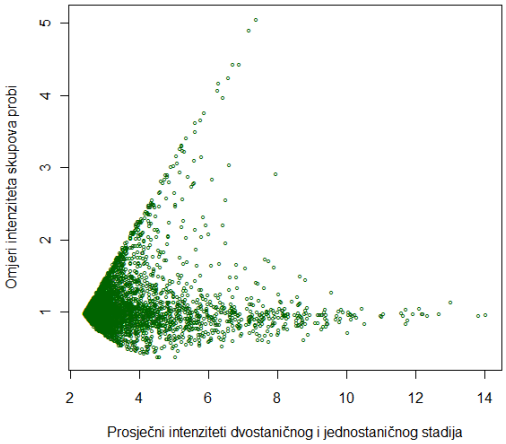
Histogram C.

Slika 6. Histogrami razdiobe normaliziranih intenziteta srednjih vrijednosti skupova probi sa istovjetnih mikročipova. Histogram A prikazuje stadij jajne stanice, histogram B jednostanični stadij, a histogram C dvostanični stadij embrija miša. Razdioba intenziteta ne pokazuje značajnu razliku između pojedinih stadija.

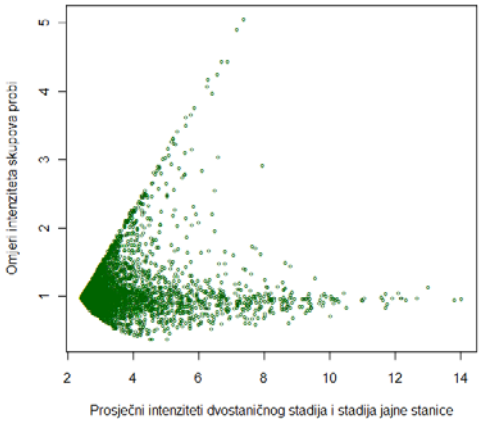
4.3. *MA* grafovi



MA graf A.



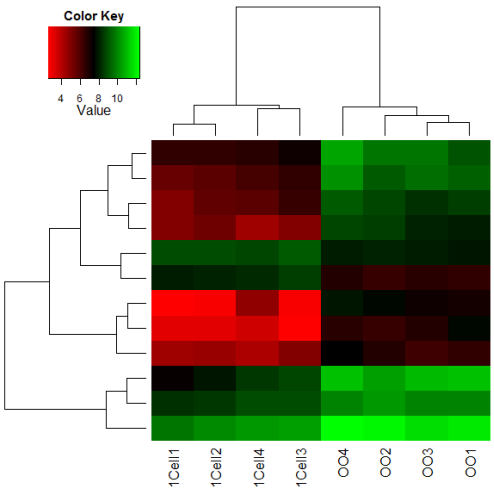
MA graf B.



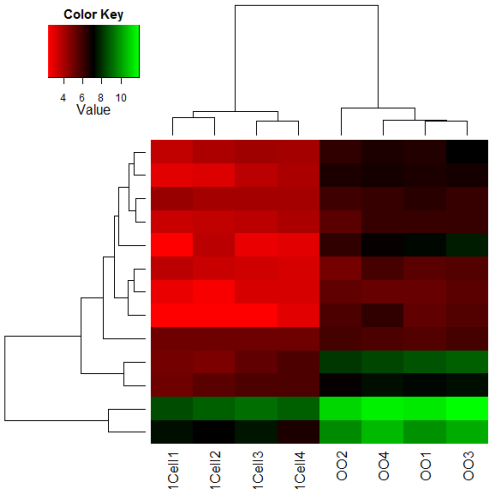
MA graf C.

Slika 7. *MA* grafovi prikazuju omjere intenziteta skupova probi (na ordinati) u odnosu na prosječne intenzitete skupova probi svih mikročipova u usporedbi (apscisa). *MA* graf A prikazuje takav omjer jednostaničnog stadija mišjeg embrija i stadija jajne stanice, *MA* graf B dvostaničnog stadija u odnosu na jednostanični stadij embrija miša, a *MA* graf C dvostaničnog stadija u odnosu na stadij jajne stanice. *MA* graf A ne pokazuje znatne razlike u razini ekspresije gena između jednostaničnog stadija i jajne stanice, dok *MA* grafovi B i C pokazuju povišenu razinu ekspresije podskupa gena u dvostaničnom stadiju u odnosu na jednostanični, odnosno stadij jajne stanice.

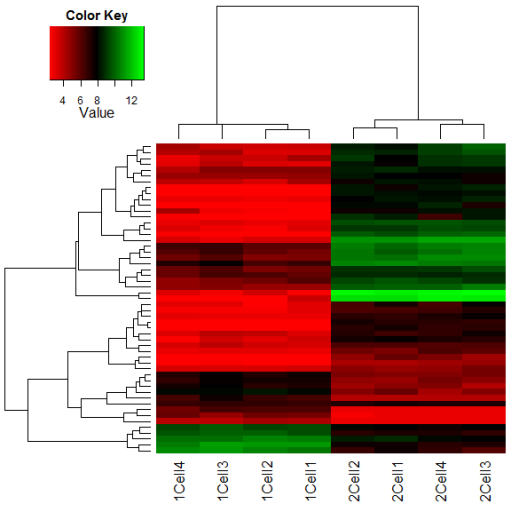
4.4. Mape intenziteta



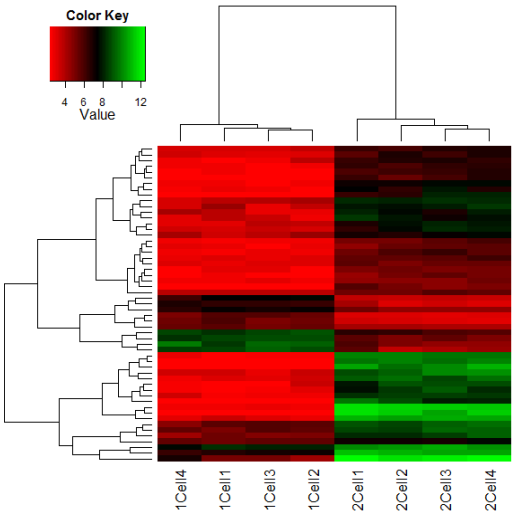
Mapa intenziteta 1A.



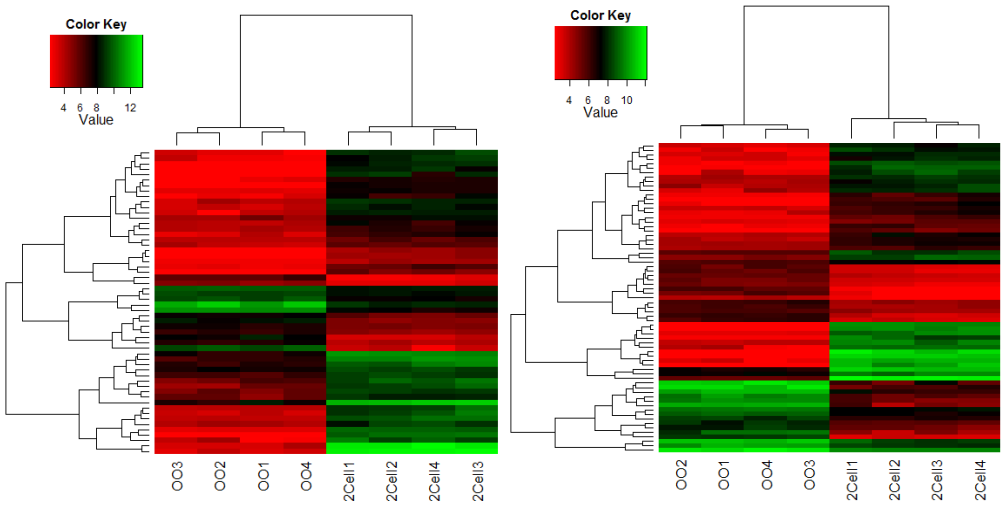
Mapa intenziteta 1B.



Mapa intenziteta 2A.



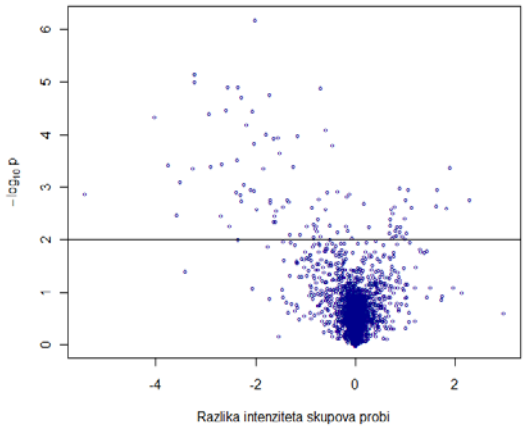
Mapa intenziteta 2B.



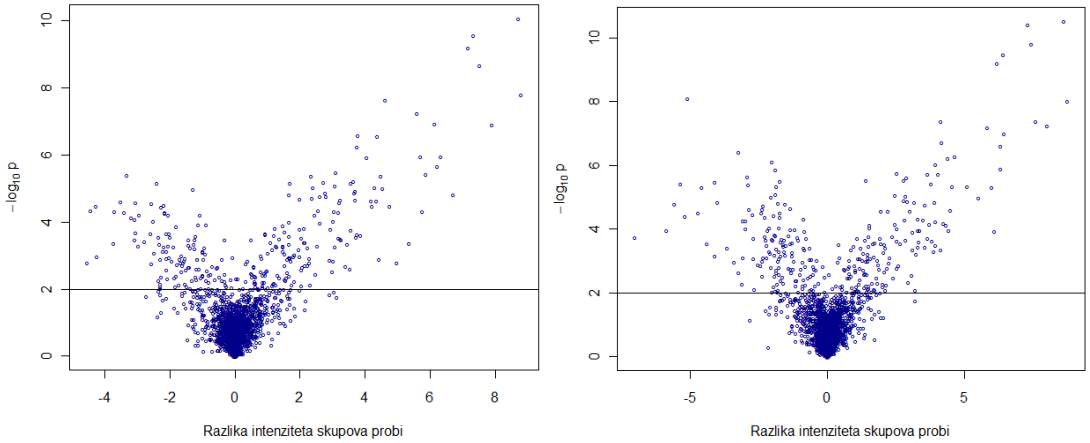
Mapa intenziteta 3A. Mapa intenziteta 3B.

Slika 8. Mape intenziteta skupova probi koji odgovaraju razinama ekspresije nekodirajućih transkripata. Mape intenziteta 1A i 1B odgovaraju usporedbi jednostaničnog stadija i stadija jajne stanice na mikročipu MOE430A, odnosno MOE430B. Mape intenziteta 2A i 2B odgovaraju usporedbi dvostaničnog u odnosu na jednostanični stadij na mikročipu MOE430A, odnosno MOE430B, dok mape intenziteta 3A i 3B odgovaraju usporedbi dvostaničnog stadija i stadija jajne stanice na mikročipu MOE430A, odnosno MOE430B. Stadij jajne stanice: OO1-4, jednostanični stadij: 1Cell1-4, dvostanični stadij: 2Cell1-4. Svaki redak mape svrstan je u pojedinu grupu (lijevo na mapama) i predstavlja jedan skup probi, dok je bojom prikazana visina njegova intenziteta u svakom od naznačenih istovjetnih mikročipova (od najniže vrijednosti ka najvišim: intenzivno crvena→tamno crvena→ zelena). Gore na prikazima naznačene su grupe u koje spadaju pojedini mikročipovi. Mape intenziteta 1A i 1B pokazuju uglavnom više razine ekspresije kod stadija jajne stanice, dok kod mapa intenziteta 2A,2B, te 3A i 3B vidimo naznake povišene razine nekodirajućih transkripata kod dvostaničnog stadija, uz prisustvo skupine transkripata sa relativno povišenom razinom ekspresije u stadiju jajne stanice.

4.5. *Volcano* grafovi



Volcano graf A.



Volcano graf B. *Volcano* graf C.

Slika 9. *Volcano* grafovi prikazuju negativni logaritam p vrijednosti (ordinata) u odnosu na razliku intenziteta skupova probi (apscisa), koja predstavlja razliku u razini ekspresije nekodirajućih transkripata među mikročipovima koji odgovaraju različitim stadijima mišjeg embrija. *Volcano* graf A prikazuje razliku razina ekspresije između jednostaničnog stadija i stadija jajne stanice, *Volcano* graf B razliku razina ekspresije između dvostaničnog i jednostaničnog stadija, a *Volcano* graf C razliku razina ekspresije između dvostaničnog stadija mišjeg embrija i stadija jajne stanice. Uočavamo porast razine nekodirajućih transkripata kod usporedbi dvostaničnog stadija s jednostaničnim i jajnom stanicom (*Volcano* graf B i C), dok na *Volcano* grafu A ne uočavamo porast ekspresije znatnog broja gena kod jednostaničnog stadija.

4.6. Primjer anotacije skupova proba i analize nekodirajućeg dijela transkriptoma

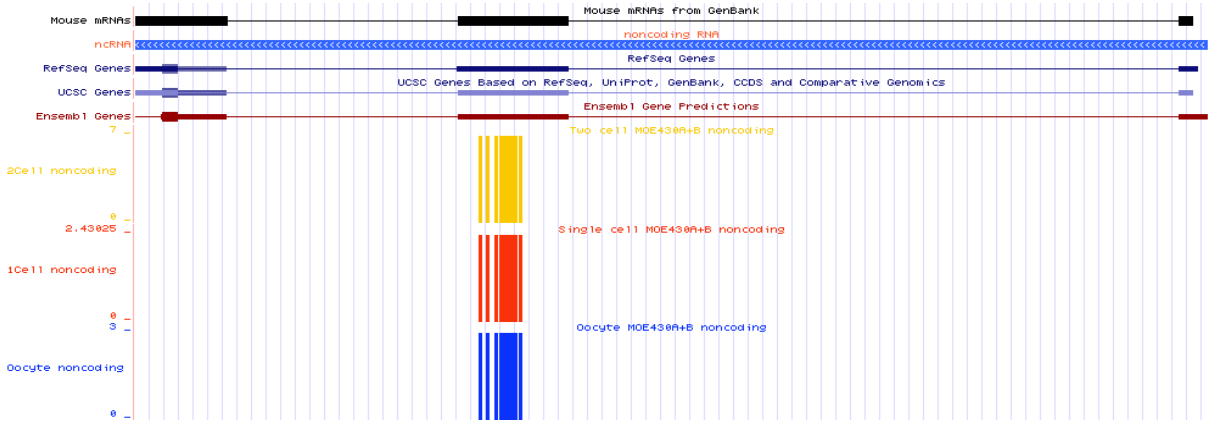
Tablica 1. Primjer anotacije skupova probi usporedbom s dostupnim podacima o nekodirajućim transkriptima lncRNA. Uz svaki naznačeni skup probi, prikazan je i njegov položaj u genomu (kromosom,početak,kraj), kao i orijentacija na genomu. Klasa označava klasifikaciju skupa obzirom na položaj u genomu (u odnosu na kodirajuće gene). Anotacija koja započinje s ENS predstavlja anotaciju s baze Ensembl, dok anotacija FT pripada bazi FANTOM3.

skup probi	kromosom	početak	kraj	orijentacija	klasa	anotacija
1452239_at	6	113020792	113020816	-	intergenska	ENSMUSG00000086429
1424516_at	17	45578926	45578950	+	intergenska	ENSMUSG00000073393
1424515_at	17	45578064	45578088	+	intergenska	ENSMUSG00000073393
1433789_at	4	131908205	131908229	-	intronska	ENSMUSG00000091021
1428529_at	9	88415938	88415962	-	intergenska	ENSMUSG00000032424
1437658_a_at	19	8800368	8800392	+	intergenska	ENSMUSG00000090867
1460103_at	1	138518887	138518911	-	intergenska	ENSMUSG00000091667
1455904_at	1	162968466	162968490	+	intergenska	ENSMUSG00000053332
1456179_at	4	146634201	146634225	+	intergenska	ENSMUSG00000084976
1457302_at	8	23593466	23593490	+	intronska	ENSMUSG00000078859
1425082_s_at	5	66014191	66014215	-	intronska	FT310321
1428288_at	19	23242243	23242267	+	intergenska	FT33636
1436522_at	11	106017895	106017919	+	intergenska	FT328010
1419641_at	11	6373510	6373534	-	intergenska	FT326175
1460065_at	2	151715382	151715406	-	intergenska	FT318307
1435524_at	3	123210520	123210544	-	intergenska	FT33211
1454955_at	7	117199105	117199129	+	intergenska	FT332009
… još 12342 reda						

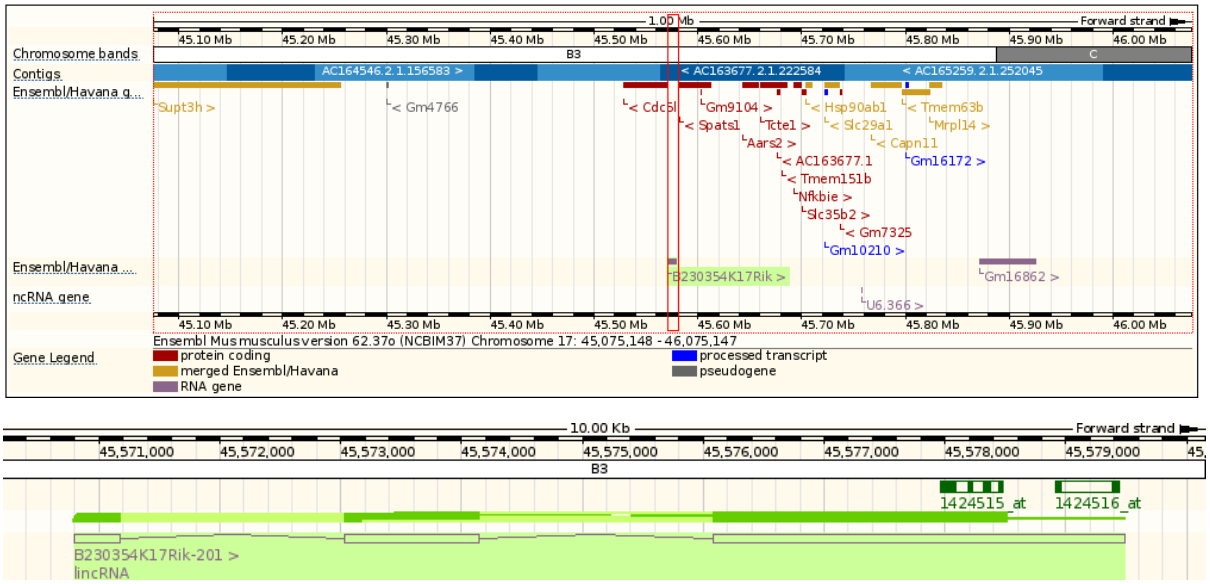
Tablica 2. Primjer analize diferencijalne ekspresije gena. Na temelju novodefiniranih skupova probi analiza diferencijalne ekspresije nekodirajućih transkripata Bayes-ovim pristupom. Razlika intenziteta prikazuje razliku u srednjim vrijednostima dvostaničnog u odnosu na jednostanični stadij embrija miša. Prilagođena p vrijednost po Benjamini-Hochberg metodi.

skup probi	razlika intenziteta	p_vrijednost	prilagođena_p_vrijednost(BH)
1422768_at	6.65	1e-10	1e-07
1437952_at	10.16	1e-09	5e-07
1447427_at	7.40	1e-09	5e-07
1455988_a_at	8.00	1e-09	5e-07
1430978_at	6.32	2e-08	4e-06
1456365_at	10.04	3e-08	5e-06
1436522_at	5.43	3e-08	6e-06
1425082_s_at	5.84	7e-08	1e-05
1451602_at	5.29	7e-08	1e-05
1450690_at	4.29	2e-07	3e-05
1418189_s_at	3.24	3e-07	4e-05
… još 12573 reda			

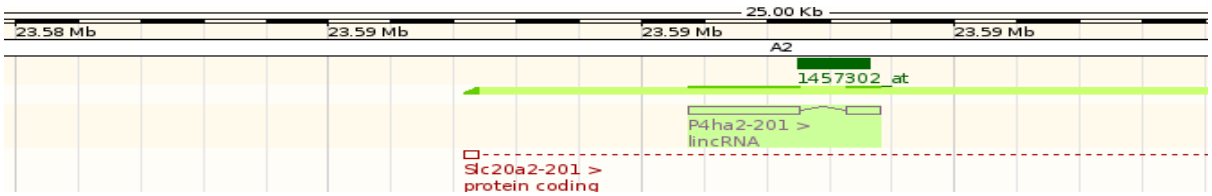
4.7. Vizualizacija genomskim preglednicima UCSC Genome Browser i Ensembl



Slika 10. Prikaz ekspresije gena *gene trap ROSA 26* (transkript lncRNA) na web sučelju UCSC Genome Browser-a u tri stadija embrionalnog razvoja miša. Prikazani stupci u žutoj (dvostanični stadij), narančastoj (jednostanični stadij) i plavoj (stadij jajne stanice) boji označavaju visinu intenziteta skupa probi koji odgovara navedenom genu. Prikazan je i iznos intenziteta s lijeve strane, te možemo uočiti da je kod dvostaničnog stadija jednak 7, dok je kod ostala dva ~3, što znači da ovaj gen ima dvostruko veću razinu ekspresije kod dvostaničnog stadija mišjeg embrija u odnosu na jednostanični stadij i stadij jajne stanice.



Slika 11. Prikaz nekodirajućeg transkripta lncRNA B230354K17Rik (ENSMUSG00000073393) na genomskom pregledniku Ensembl. Nalazi se između dva kodirajuća gena, a pri dnu slike možemo vidjeti i strukturu navedenog gena (eksoni i introni), kao i položaj dvaju skupova probi (1424515_at i 1424516_at) koje mapiraju upravo na ovaj transkript lncRNA.



Slika 12. Prikaz nekodirajućeg transkripta lncRNA koji se nalazi u intronu kodirajućeg gena (koji kodira za protein *solute carrier family 20, member 2*). Naznačeni skup probi (145730_at) nalazi se upravo u intronu kodirajućeg gena, te predstavlja navedeni kodirajući gen na mikročipu u anotaciji *Affymetrix-a.*

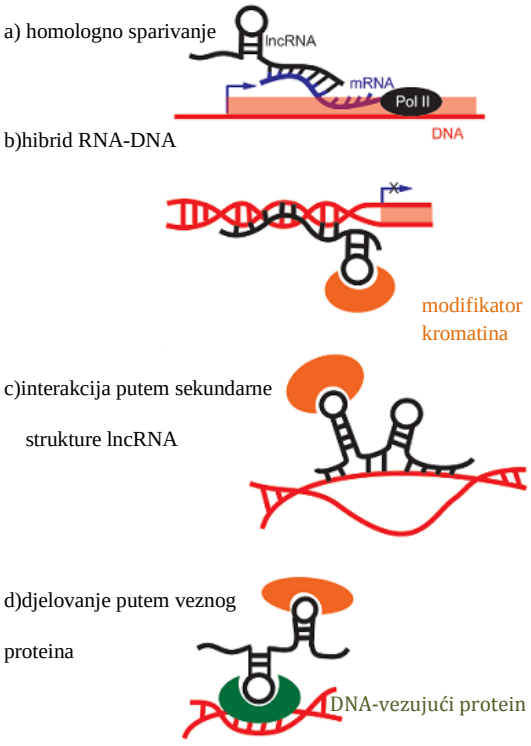
5. RASPRAVA

Dugogodišnji pristup istraživanju staničnih procesa bio je usredotočen na razmjerno mali broj gena u pojedinačnim eksperimentima, i to uglavnom onih koji kodiraju za proteine (u skladu s centralnom dogmom molekularne biologije). Iako je takav pristup omogućio dobro poznavanje funkcije pojedinih proteina u stanici, ostavio je neistraženim ostatak genoma koji ne sadrži kodirajući potencijal uz pretpostavku da nema funkciju (tzv. *junk* DNA). Sekvenciranje cjelokupnih genoma raznih organizama (posebice viših eukariota) ostvarilo je velik pomak ka razumijevanju i ostalih dijelova genoma koje nije bilo moguće istraživati na razini proteinskog produkta. Otkriće skupine znanstvenika (Bejerano, Pheasant i sur. 2004) da su neke nekodirajuće regije u eukariotskom genomu evolucijski daleko očuvanije nego što bi se to očekivalo da ne obavljaju nikakvu funkciju, pokrenulo je lavinu istraživanja upravo na nekodirajućim dijelovima genoma. U narednom razdoblju pronađene su i opisane brojne *cis*-regulatorne regije koje imaju ulogu u regulaciji transkripcije, interakcijom sa nizom pripadnih transkripcijskih faktora.

Usporedno sa saznanjima o funkciji regija genoma koje se ne transkribiraju, znanstvenici su došli do otkrića o postojanju niza novih klasa nekodirajućih transkripata koje nastaju transkripcijom određenih genomskih regija. Primjer su kratke molekule RNA poput male jezgrine RNA(snRNA), male jezgricine RNA(snoRNA), mikro RNA(miRNA), piko RNA (piRNA) i mnoge druge. Također se u posljednjih nekoliko godina pronalazi sve više transkripata koji su, za razliku od navedenih, strukturno vrlo slični kodirajućim molekulama mRNA (s obzirom na izrezivanje introna, poliadenilaciju, dodatak 5' kape), a ujedno su i prisutni u velikom broju u transkriptomima kralješnjaka (Chodroff, Goodstadt i sur. 2010). Takvi transkripti nazvani su dugačke nekodirajuće molekule RNA (lncRNA), a jedna od najistaknutijih osobina im je nedostatak otvorenog okvira čitanja duljeg od 100 nukleotida, kao i činjenica da se mnoge zadržavaju u jezgri (Chen 2011). Unatoč tome što se sve više transkripata pripisuje upravo navedenoj klasi, njihova uloga u stanici još uvijek je u velikoj mjeri nepoznata. Biološka uloga lncRNA bila bi možda upitna u slučaju da su one pronađene samo kod uske skupine organizama, no njihova očuvanost i sveprisutnost u Amniota, očuvanost strukture eksona, kao i sekundarne strukture takvih molekula RNA upućuje na to su molekule lncRNA biološki važni čimbenici. lncRNA su prvenstveno uključene u razvojne procese eukariota, a posebno sisavaca, kroz njihovu ulogu u kontroli epigenetičke regulacije (modeliranjem strukture kromatina) genske ekspresije (Amaral i Mattick 2008), kao i u razvoju mozga kralješnjaka (Chodroff, Goodstadt i sur. 2010). Danas se također pretpostavlja regulacijska uloga lncRNA i u drugim važnim biološkim procesima, kao što su inaktivacija X kromosoma (Lee 2009), imprinting (Brannan, Dees i sur. 1990; Pauler, Koerner i sur. 2007; Royo i Cavaille 2008), diferencijacija i razvoj (Guttman, Amit i sur. 2009), očuvanje integriteta staničnih odjeljaka (Sasaki, Ideue i sur. 2009), procesiranje kratkih nekodirajućih molekula RNA

(Tam, Aravin i sur. 2008; Wilusz, Freier i sur. 2008). Uz to, pronađena je i povezanost lncRNA s nizom bolesti u čovjeka (Moseley, Zu i sur. 2006; Prasanth i Spector 2007; Faghihi, Modarresi i sur. 2008). Upravo zbog toga postoji potreba za razumijevanjem mehanizama kojim lncRNA djeluju i utječu na svoje ciljne gene u stanici.

Najčešće opisani mehanizam djelovanja lncRNA uključuje privlačenje proteinskih kompleksa za preinaku strukture kromatina na određenim mjestima u genomu (jezgri). Takvi proteinski kompleksi najčešće sadrže proteine grupe Polycomb, koji imaju ulogu utišavanja gena, odnosno znatno smanjenje genske ekspresije (Hung i Chang 2010). Predloženi mehanizmi djelovanja uključuju: homologno sparivanje, stvaranje hibrida DNA-RNA, interakciju sa proteinima putem sekundarne strukture lncRNA, i djelovanje putem veznog proteina, a prikazani su na Slici 13.



Slika 13. Predloženi mehanizmi djelovanja nekodirajućih transkripata lncRNA. a) Mehanizam homolognog sparivanja uključuje transkripte prepisane sa istovjetnog položaja u genomu, ali u suprotnoj orijentaciji, što ih čini komplementarnima za hibridizaciju (i potencijalnu regulaciju količine slobodnog kodirajućeg transkripta) b) Komplementarnost sekvenci za hibridizaciju također može poslužiti u formiranju kompleksnijih struktura poput RNA-DNA hibridnih dupleksa i tripleksa, te privući proteinske komplekse za preinaku strukture kromatina c) Mogućnost formiranja brojnih sekundarnih struktura kod jednolančanih molekula RNA omogućuje vezanje takvih specifičnih struktura za proteinske komplekse, s posljedicom modulacije genske ekspresije d) Mehanizam privlačenja lncRNA putem proteina koji posjeduju domen za vezanje nukleinskih kiselina na ciljna mjesta u genomu, gdje obavljaju svoju funkciju (Hung i Chang 2010).

Prikupljanjem i obradom velike količine podataka o tkivno specifičnoj ekspresiji transkripata lncRNA možemo dobiti izravne indicije o njihovoj funkciji i usmjeriti daljnja istraživanja o mehanizmu njihova djelovanja. Tkivno specifična ekspresija jedan je od prvih pokazatelja funkcionalnosti molekula, jer ukazuje na to da su molekule lncRNA prisutne upravo onda kada su potrebne u stanici. Metodom opisanom u ovom radu proširujemo moguće izvore eksperimentalnih

podataka za otkrivanje znanja o lncRNA. Skupom računalnih alata i protokolom za računalnu analizu, uspješno sam izolirala signal iz dostupnih podataka koji pripada ekspresiji nekodirajućih RNA, odredila različito eksprimirane nekodirajuće regije genoma među embrionalnim stadijima miša, te ih anotirala kombinacijom dostupnih podataka iz bioloških baza.

Definirala sam 56% intronskih, 39% intergenskih , te 5% eksonskih skupova proba neodgovarajuće orijentacije u ukupno 12587 skupova probi kombinacijom mikročipova MOE430A i MOE430B proizvođača *Affymetrix*. Anotacija navedenih skupova probi usporedbom s dostupnim podacima iz baza Ensembl i FANTOM3 uspješno je provedena za intronske i intergenske skupove probi. Broj „nekodirajućih" skupova proba je manji od ukupnog (29% od ukupnog) upravo zbog toga što često cijeli skupovi proba pripadaju kodirajućim transkriptima. Unatoč tome, još uvijek možemo definirati popriličan broj skupova probi koje daju signal nekodirajućih transkripata, što je iznimno vrijedna informacija obzirom na veliku količnu danas dostupnih mikročip podataka u bazama poput GEO (Barrett, Troup i sur. 2009; Barrett, Troup i sur. 2011). Rezultati dobiveni opisanom metodom ukazuju na promjenu ekspresije nekodirajućih transkripata u svim kategorijama „nekodirajućih" skupova probi (prikazane one anotirane u primjeru Tablice 1.), dobivenih međusobnom usporedbom triju stadija ranog embrionalnog razvoja miša (stadij jajne stanice, jednostanični i dvostanični stadija ranog embrija nastalog nakon oplodnje jajne stanice).

Histogrami razdiobe normaliziranih intenziteta na Slici 6. pokazuju da su za sva tri razvojna stadija rasponi vrijednosti (~2 do ~14 na logaritamskoj skali) i izgled same distribucije približno jednaki. Najviše skupova probi na svakom od tri histograma pokazuje relativno niske vrijednosti intenziteta (vrijednosti od ~2 do ~3), što znači da većina ispitanih gena (koji nose uputu za nekodirajuće transkripte) ima bazalnu ekspresiju , sa vrlo malim udjelima onih koji pokazuju povišene razine ekspresije (vrijednosti intenziteta ~7 do ~14).

Skupovi proba prikazani su kao točke na *MA* grafovima (Slika 7.), na kojem su na ordinati omjeri vrijednosti intenziteta svakog skupa probi između dva stadija. Na *MA* grafu A uočavamo da se većina točaka nalazi pri omjeru koji je jednak 1 (središnja linija), koje predstavljaju gene sa relativno nepromijenjenom razinom ekspresije među ispitivanim stadijima. One točke koje odstupaju od središnje linije predstavljaju gene koji su više izraženi u jednom od stadija. Na *MA* grafu A to je omjer između jednostaničnog stadija i stadija jajne stanice, te možemo uočiti desetak gena iz dvije skupine: onih koji odstupaju prema višim vrijednostima ordinate (viša razina ekspresije u jednostaničnom stadiju), i prema nižim (viša razina ekspresije u stadiju jajne stanice). Pet gena je 1.5-2 puta izraženije u jednostaničnom stadiju, kao i pet onih koji imaju dva puta veću razinu ekspresije u jajnoj stanici

(vrijednost 0.5 na ordinati). Kod usporedbe dvostaničnog stadija u odnosu na jednostanični, prikazano na *MA* grafu B, a posebice jajnu stanicu (*MA* graf C), vidimo da znatan broj gena (~100) ima višu razinu ekspresije u dvostaničnom stadiju, stoga na *MA* grafovima B i C postoje točke na vrijednostima omjera između 4 i 5.

Mape intenziteta na Slici 8. slikovito prikazuju aktivaciju i utišavanje ekspresije pojedinih gena. Svaki redak na mapi predstavlja skup proba kojem je pridijeljena boja u skladu sa pripadnom visinom intenziteta na pojedinim mikročipovima (stupci na mapi). Razlike u bojama skupova probi među stupcima mape ukazuju na promjenu ekspresije odgovarajućeg gena među uspoređivanim stadijima (predstavljeni odgovarajućim mikročipovima). Na mapama intenziteta 1A i 1B vidljiv je mali broj skupova probi vrlo različitih boja među stadijima, većinom uočavamo zelenu i tamnocrvenu boju (više vrijednosti intenziteta) kod istovjetnih mikročipova koji predstavljaju stadij jajne stanice, dok je intenzivno crvena boja (manje vrijednosti intenziteta) prisutna kod mikročipova koji predstavljaju jednostanični stadij. To upućuje na smanjenje razine ekspresije odgovarajućih gena u jednostaničnom stadiju, odnosno degradaciju dijela transkripata prisutnih u jajnoj stanici (tzv. „majčini“ transkripti). Nasuprot tome, na mapama intenziteta 2A, 2B, kao i 3A i 3B (usporedba dvostaničnog stadija u odnosu na jednostanični, odnosno jajnu stanicu) uočavamo nekoliko desetaka izrazito eksprimiranih gena u odnosu na stadij u usporedbi, ali i dijelove koji također ukazuju na degradaciju majčinih transkripata (crvena boja dvostaničnog stadija u usporedbi sa zelenom i tamno crvenom stadija jajne stanice na mapama intenziteta 3A i 3B). Na svim mapama intenziteta (Slika 8.) vidljivo je da istovjetni mikročipovi (tzv. biološki replikati) grupiranjem po razinama ekspresije gena uistinu pripadaju zasebnim grupama (ujednačeni su) i da različiti stadiji pokazuju veće razlike u međusobnoj usporedbi. To nam dodatno potvrđuje pouzdanost razlike u ekspresiji nekog promatranog podskupa gena između različitih razvojnih stadija mišjeg embrija.

Volcano grafovi (Slika 9.) na apscisi pokazuju razliku intenziteta pojedinih skupova probi među stadijima, dok je na ordinati negativni logaritam p vrijednosti (pouzdanosti procjene razlike intenziteta). Horizontalna linija povučena je pri vrijednosti 2, što odgovara p=0.01. Više vrijednosti na ordinati ukazuju na statistički značajnije rezultate (točke koje predstavljaju gene sa ujednačeno povišenom razinom ekspresije u jednom od stadija). Posebno su zanimljivi geni sa velikom razlikom u srednjoj vrijednosti među stadijima (ekstremne vrijednosti na apscisi) koji upućuju gene sa znakovito različitom razinom ekspresije među stadijima. Na *Volcano* grafu A, u usporedbi jednostaničnog stadija i jajne stanice, vidljivo je nekoliko desetaka točaka na vrijednostima manjim od -2, što opet ukazuje na smanjenje količine „majčinih“ transkripata u jednostaničnom stadiju, dok svega 4 gena pokazuju

povišenu razinu u jednostaničnom stadiju. Kao i kod mapi intenziteta na Slici 8., na *Volcano* grafovima B i C (usporedba dvostaničnog stadija u odnosu na jednostanični, odnosno jajnu stanicu) vidljiva je aktivacija gena u dvostaničnom stadiju, također uz smanjenje razine ekspresije gena koji su povišeni u jajnoj stanici (*Volcano* graf C).

Na temelju navedenih prikaza rezultata možemo vidjeti jasnu pojavu ekspresije niza gena kod dvostaničnog stadija, dok je kod jednostaničnog genska ekspresija znatno slabije izražena u cjelokupnom uzorku. Također, možemo u jednostaničnom i dvostaničnom stadiju uočiti i nestanak određenog dijela transkriptoma u odnosu na stadij jajne stanice. Svi nam navedeni rezultati ukazuju na to da ne samo da postoji ekspresija nekodirajućih transkripata tijekom sva 3 ključna stadija embrionalnog razvoja, već i da postoje značajne razlike u samoj razini ekspresije. Ovi rezultati govore u prilog hipotezi da nekodirajući transkripti imaju važnu ulogu u staničnim procesima tijekom embrionalnog stadija. Nagli porast ekspresije također je pokazan za kodirajuće transkripte za vrijeme dvostaničnog stadija, dok kod jednostaničnog, a posebice stadija jajne stanice, transkriptom uvelike čine majčini transkripti za koje je pokazano da potiču ekspresiju točno određene skupine gena potrebnih za daljnji razvoj i diferencijaciju embrija (Li, Zheng i sur. 2010).

Tablica 1. prikazuje podskup svih anotiranih skupova proba i pokazana je kao primjer podataka koje sam prikupila u analizi. Na primjer, skup proba 1452239_at odgovara *Gt(ROSA)26Sor* lncRNA. Iz grafičkog prikaza razine ekspresije u genomskom pregledniku (Slika 10.) vidljivo je da navedena lncRNA ima povišenu razinu ekspresije u dvostaničnom stadiju u odnosu na prethodne stadije razvoja mišjeg embrija (razlika intenziteta ~4). Slijedeća dva po redu skupa probi, 1424515_at i 1424516_at, anotacijom uz pomoć opisanog protokola odgovaraju slabo opisanoj lncRNA, RP23-20A6.1-001 unutar gena ENSMUSG00000073393 (Slika 11.). Ekspresija ove lncRNA je također veća u dvostaničnom stadiju (razlika srednjih vrijednosti ~4). Skup probi 145730_at protokolom je klasificiran kao intronski (nalazi se u intronu kodirajućeg gena *solute carrier family 20, member 2*), te dodijeljen nekodirajućem transkriptu lncRNA (*P4h2-201*). Iz Slike 12. vidljivo je da navedeni skup proba mapira na lncRNA koja se nalazi u intronu navedenog gena. Unatoč tome, u izvornoj anotaciji *Affymetrix-a* taj skup proba opisan je kao da predstavlja gen koji kodira za *solute carrier family 20, member 2*, a ne odgovarajuću lncRNA.

Prilagodbom dosadašnjih eksperimentalnih visokoprotočnih tehnologija (npr. mikročip čije su probe dizajnirane isključivo za nekodirajuće transkripte), kao i primjenom novih eksperimentalnih pristupa istraživanja cjelokupnog transkriptoma kao što je npr. *RNAseq* tehnologija (sekvenciranje čitavih transkriptoma metodama sekvenciranja treće generacije), može se provesti prikupljanje

podataka o razini ekspresije nekodirajućih transkripata u različitim vrstama tkiva i stanica (staničnih linja). Kao konačnu potvrdu transkripcije moguće je u konačnici provesti i ciljane *in vitro* pokuse transkripcije na kraćim genomskim odsječcima koji sadrže potencijalno mjesto transkripcije. Dosadašnji podaci dobiveni tehnologijom mikročipova također pokrivaju razmjerno velik dio genoma (čak i oni mikročipovi koji nisu dizajnirani da prekriju cjelokupni genom, tzv. *whole genome tiling micorarrays*). Unatoč činjenici da je njihova obrada bila usredotočena većinom samo na kodirajuće transkripte, oni sadrže i signale nekodirajućih transkripata. Stoga protokol razvijen u ovom radu omogućuje pristup i dobivanje vrijednih informacija iz takvih podataka koji su javno dostupni i sadrže velik broj obavljenih pokusa. S obzirom da se za svaku novu tehnologiju moraju razviti prikladne statističke metode obrade podataka, prilikom utvrđivanja novih metoda uvijek postoje poteškoće prilikom utvrđivanja pouzdanosti mjerenja i statističke utemeljenosti analize. Tehnologija kojom se koriste mikročipovi je ovaj razvojni put završila i ušla je u stabilno i pouzdano analitičko razdoblje – poznata su eksperimentalna i statistička ograničenja same tehnologije. Stoga se podaci dobiveni mikročip tehnologijom mogu iskoristiti prilikom potvrđivanja statističke utemeljenosti rezultata dobivenih novim metodama. Biološki je svakako pouzdaniji onaj podatak ili rezultat kojeg možemo potvrditi u dva ili više pokusa koristeći različite metode i izvore podataka. Prema tome, važnost ovog rada proizlazi i iz povećanja mogućnosti da se već postojeći eksperimenti dodatno iskoriste.

Protokol razvijen u ovom radu također omogućuje standardnu obradu cjelokupnog skupa proba, ali i kontroliranu analizu uz mogućnosti vlastite anotacije skupova probi prema najnovijim i osvježenim izvorima iz baze podataka kao što je Ensembl (Flicek, Amode i sur. 2011). S obzirom na količinu pristupnih podataka u biološkim bazama poput GEO koji često sadrže rezultate pokusa izrađenih u vrhunskim institutima čije grupe znanstvenika rade na različitim, često delikatnim staničnim linijama kao što su embrionalne stanice, uz prikladne računalne metode poput ove, moguće je takve podatke dodatno iskoristiti u svrhu stvaranja novih hipoteza i pritom izbjeći potrebu za ponavljanjem skupih pokusa.

6. ZAKLJUČCI

- Brojni nekodirajući RNA transkripti, koji u svom nukleotidnom slijedu ne nose informaciju za biosintezu proteina, prisutni su u transkriptomima eukariota. Značajnu klasu takvih transkripata čine dugačke nekodirajuće RNA molekule (lncRNA).

- Skup računalnih alata i protokol za analizu razvijen u ovom radu omogućuje obradu podataka prikupljenih mikročipovima, u svrhu analize i karakterizacije ekspresije nekodirajućih transkripata, izolacijom signala koji specifično pripada nekodirajućem dijelu transkriptoma, što nije moguće provesti postojećom standardnom analizom.

- Protokol omogućuje dodatni izvor informacija o ekspresiji nekodirajućih transkripata obradom već dostupnih podataka iz mikročipova, kao i provođenje standardne analize cjelokupnog transkriptoma, ali uz najnoviji opis proba mikročipa i popravak grešaka vezanih uz neprikladne probe mikročipova mišjih transkripata.

- Rezultati analize podataka iz pokusa na mikročipovima mjerenjem razine ekspresije transkriptoma stadija ranog embrionalnog razvoja miša (jajna stanica, jednostanični i dvostanični stadij nakon oplodnje) ukazuju na različite razine ekspresije nekodirajućih transkripata lncRNA između stadija. Najveća razlika u profilima ekspresije primjetna je u usporedbi dvostaničnog stadija s jednostaničnim i jajnom stanicom.

- Tjekom dvostaničnog stadija odvija se aktivna ekspresija nekodirajućih transkripata i genetičko reprogramiranje potrebno za embrionalni razvoj.

7. ZAHVALE

Veliko hvala prof.dr.sc. Kristianu Vlahovičeku na mentorstvu nad ovim radom, stručnim savjetima i pomoći u svakom trenutku, kao i inspiraciji za rad. Dipl.ing. Vedranu Frankeu zahvaljujem na korisnim savjetima i čitanju rada.

Hvala mojim roditeljima na bezuvjetnoj podršci i razumijevanju.

8. LITERATURA

Amaral, P. P. i J. S. Mattick (2008). "Noncoding RNA in development." Mamm Genome **19**(7-8): 454-492.
Babu, A. i R. S. Verma (1987). "Chromosome structure: euchromatin and heterochromatin." Int Rev Cytol **108**: 1-60.
Baker, M. (2011). "Long noncoding RNAs: the search for function." Nat Meth **8**(5): 379-383.
Baldi, P. i A. D. Long (2001). "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes." Bioinformatics **17**(6): 509-519.
Barrett, T., D. B. Troup, i sur. (2011). "NCBI GEO: archive for functional genomics data sets--10 years on." Nucleic Acids Res **39**(Database issue): D1005-1010.
Barrett, T., D. B. Troup, i sur. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Res **37**(Database issue): D885-890.
Bejerano, G., M. Pheasant, i sur. (2004). "Ultraconserved elements in the human genome." Science **304**(5675): 1321-1325.
Bolstad, B. M., R. A. Irizarry, i sur. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.
Brannan, C. I., E. C. Dees, i sur. (1990). "The product of the H19 gene may function as an RNA." Mol Cell Biol **10**(1): 28-36.
Calkhoven, C. F. i G. Ab (1996). "Multiple steps in the regulation of transcription-factor level and activity." Biochem J **317 (Pt 2)**: 329-342.
Carninci, P., T. Kasukawa, i sur. (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563.
Cedar, H. i Y. Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms." Nat Rev Genet **10**(5): 295-304.
Chen, L. L. (2011). "Long noncoding RNAs in mammalian cells: what, where,and why?" Wiley Interdisciplinary Reviews-RNA **1**(1).
Cheng, J., S. Sun, i sur. (2004). "NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis." Bioinformatics **20**(9): 1462-1463.
Chodroff, R. A., L. Goodstadt, i sur. (2010). "Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes." Genome Biol **11**(7): R72.
Dalma-Weiszhausz, D. D., J. Warrington, i sur. (2006). "The affymetrix GeneChip platform: an overview." Methods Enzymol **410**: 3-28.
Edgar, R. i T. Barrett (2006). "NCBI GEO standards and services for microarray data." Nat Biotechnol **24**(12): 1471-1472.
Edgar, R., M. Domrachev, i sur. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.
Eschrich, S. i T. J. Yeatman (2004). "DNA microarrays and data analysis: an overview." Surgery **136**(3): 500-503.
Faghihi, M. A., F. Modarresi, i sur. (2008). "Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase." Nat Med **14**(7): 723-730.
Flicek, P., M. R. Amode, i sur. (2011). "Ensembl 2011." Nucleic Acids Res **39**(Database issue): D800-806.
Glazer, M., J. A. Fidanza, i sur. (2006). "Kinetics of oligonucleotide hybridization to photolithographically patterned DNA arrays." Anal Biochem **358**(2): 225-238.
Gregory Alvord, W., J. A. Roayaei, i sur. (2007). "A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R." Briefings in Bioinformatics **8**(6): 415-431.
Gupta, R. A., N. Shah, i sur. (2010). "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." Nature **464**(7291): 1071-1076.

Guttman, M., I. Amit, i sur. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.
Hubbard, T., D. Barker, i sur. (2002). "The Ensembl genome database project." Nucleic Acids Res **30**(1): 38-41.
Hung, T. i H. Y. Chang (2010). "Long noncoding RNA in genome regulation: Prospects and mechanisms." RNA Biol **7**(5): 582-585.
Hwang, S. Y. (1997). "Whole genome analysis using DNA chips." Korean Soc. Med. Biochem. Mol. Biol. News **4**: 6.
Jacquier, A. (2009). "The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs." Nat Rev Genet **10**(12): 833-844.
Jaenisch, R. i A. Bird (2003). "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals." Nat Genet.
Jia, H., M. Osak, i sur. (2010). "Genome-wide computational identification and manual annotation of human long noncoding RNA genes." RNA **16**(8): 1478-1487.
Karolchik, D., R. Baertsch, i sur. (2003). "The UCSC Genome Browser Database." Nucleic Acids Res **31**(1): 51-54.
Katayama, S., Y. Tomaru, i sur. (2005). "Antisense transcription in the mammalian transcriptome." Science **309**(5740): 1564-1566.
Latchman, D. S. (1992). "Gene regulation." BMJ **304**(6834): 1103-1105.
Lee, J. T. (2009). "Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome." Genes Dev **23**(16): 1831-1842.
Li, L., P. Zheng, i sur. (2010). "Maternal control of early mouse development." Development **137**(6): 859-870.
Liu, G., A. E. Loraine, i sur. (2003). "NetAffx: Affymetrix probesets and annotations." Nucleic Acids Res **31**(1): 82-86.
Lynch, M. (2006). "The origins of eukaryotic gene structure." Mol Biol Evol **23**(2): 450-468.
Maeda, N., T. Kasukawa, i sur. (2006). "Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs." PLoS Genet **2**(4): e62.
Mercer, T. R., M. E. Dinger, i sur. (2009). "Long non-coding RNAs: insights into functions." Nat Rev Genet **10**(3): 155-159.
Moseley, M. L., T. Zu, i sur. (2006). "Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8." Nat Genet **38**(7): 758-769.
Nordstrom, K. J., M. A. Mirza, i sur. (2009). "Critical evaluation of the FANTOM3 non-coding RNA transcripts." Genomics **94**(3): 169-176.
Nurtdinov, R. N., M. O. Vasiliev, i sur. (2010). "PLANdbAffy: probe-level annotation database for Affymetrix expression microarrays." Nucleic Acids Res **38**(Database issue): D726-730.
Okazaki, Y., M. Furuno, i sur. (2002). "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs." Nature **420**(6915): 563-573.
Pang, K. C., S. Stephen, i sur. (2005). "RNAdb--a comprehensive mammalian noncoding RNA database." Nucleic Acids Res **33**(Database issue): D125-130.
Pauler, F. M., M. V. Koerner, i sur. (2007). "Silencing by imprinted noncoding RNAs: is transcription the answer?" Trends Genet **23**(6): 284-292.
Perez-Iratxeta, C. i M. A. Andrade (2005). "Inconsistencies over time in 5% of NetAffx probe-to-gene annotations." BMC Bioinformatics **6**: 183.
Pevsner, J. (2009). "Bioinformatics and Functional Genomcs."
Pollack, J. R. (2009). "DNA microarray technology. Introduction." Methods Mol Biol **556**: 1-6.

Pontius, U. J., L. Wagner, i sur. (2003). "UniGene: a unified view of the transcriptome."
Prasanth, K. V. i D. L. Spector (2007). "Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum." Genes Dev **21**(1): 11-42.
Quackenbush, J. (2002). "Microarray data normalization and transformation." Nat Genet **32 Suppl**: 496-501.
Reimers, M. i V. J. Carey (2006). "Bioconductor: an open source framework for bioinformatics and computational biology." Methods Enzymol **411**: 119-134.
Royo, H. i J. Cavaille (2008). "Non-coding RNAs in imprinted gene clusters." Biol Cell **100**(3): 149-166.
Sand, O., M. Thomas-Chollier, i sur. (2009). "Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl." Bioinformatics **25**(20): 2739-2740.
Sasaki, Y. T., T. Ideue, i sur. (2009). "MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles." Proc Natl Acad Sci U S A **106**(8): 2525-2530.
Shaffer, C. D., L. L. Wallrath, i sur. (1993). "Regulating genes by packaging domains: bits of heterochromatin in euchromatin?" Trends Genet **9**(2): 35-37.
Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.
Spencer, G. (2007). "Background on Mouse as a Model Organism in NationalHuman Genome Research Institute.".
Spudich, G., X. M. Fernandez-Suarez, i sur. (2007). "Genome browsing with Ensembl: a practical overview." Brief Funct Genomic Proteomic **6**(3): 202-219.
Tam, O. H., A. A. Aravin, i sur. (2008). "Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes." Nature **453**(7194): 534-538.
Tisdall, J. D. (2001). "Beginning perl for bioinformatics."
Wilusz, J. E., S. M. Freier, i sur. (2008). "3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA." Cell **135**(5): 919-932.

9. SAŽETAK

Računalna analiza nekodirajućeg dijela transkriptoma u modelu embrionalnog razvoja miša (*Mus Musculus*)

Sara Sumić

Nove spoznaje o molekularnoj regulaciji staničnih procesa u eukariota sve više značaja pridaju molekulama pod nazivom nekodirajući transkripti. To su regije genoma koje se u procesu transkripcije prepisuju s DNA ali čiji nukleotidni slijed ne nosi uputu za biosintezu proteina. Vrlo značajnu klasu nekodirajućih transkripata, koji su visoko zastupljeni u stanici, čine dugačke nekodirajuće molekule RNA (lncRNA, od engl. *long noncoding RNA*) sa potencijalnom ulogom u regulaciji ekspresije svih staničnih gena, posebice tijekom diferencijacije i razvoja. Opis i analiza njihova tkivno specifičnog izražaja omogućit će nam rasvjetljavanje mehanizama kojima ostvaruju svoju ulogu u stanici. Visokoprotočne tehnologije poput DNA mikročipa (engl. *microarrays*) pružaju veliku količinu informacija o ekspresiji cjelokupnog skupa gena u stanici, no standardna obrada podataka usmjerena je većinom na kodirajuće transkripte, unatoč činjenici da mikročipovi sadrže i određenu količinu informacija o nekodirajućim transkriptima. Ovaj rad uvodi skup računalnih alata i metodu za obradu mikročip podataka koja omogućuje izolaciju i analizu signala koji pripada nekodirajućem dijelu transkriptoma. Ova metoda pruža mogućnost izolacije i obrade određene klase signala iz podataka prikupljenih mikročipovima, kao i kontroliranu analizu na standardan način zbog potpunijeg i novijeg opisa svih nukleotidnih proba pojedinog mikročipa. Uz razvoj metode, provela sam i analizu podataka koji su prikupljeni iz pokusa na mikročipovima na ranim stadijma razvoja mišjeg embrija, usporedivši razinu izražaja nekodirajućih transkripata među stadijima. Rezultati pokazuju značajan porast količine nekodirajućih transkripata kod dvostaničnog stadija embrija u odnosu na oplođenu i neoplođenu jajnu stanicu. Također, iz rezultata je vidljiva postupna razgradnja majčinih transkripata u jajnoj stanici nakon oplodnje.

Ključne riječi: Nekodirajuća RNA, transkriptom, mikročip tehnologija, računalna analiza, bioinformatika

10. SUMMARY

Computational analysis of the early mouse embryo noncoding portion of the transcriptome

Sara Sumić

Recent advances in research of molecular regulatory mechanisms in eukaryotic cells have led to discovery that noncoding genome regions play a pivotal role in cellular processes. Especially important regulatory molecules were found among the RNA molecules termed *noncoding transcripts*, i.e. genomic regions that are transcribed but lack the coding potential that directs protein biosynthesis. An increasingly high amount of eukaryotic cell transcriptome (collection of all transcripts in the cell) belongs to the class of long noncoding RNA transcripts (*lncRNA*), with potentially essential role in gene expression regulation, in particular during the course of cell differentiation and early development. Characterisation of their spatio-temporal expression patterns in different tissues will enable better understanding of their precise role in cellular processes. *High-throughput* technologies, like microarrays, result in a large collection of experimental data that describe the molecular state of a given cell, for example microarrays can measure expression patterns of all genes in the cell. However, standardized protocols for bioinformatic analysis of microarray raw data is focused mainly on protein-coding portion of transcriptome, despite the fact that raw data contains considerable amount of signal representing noncoding transcripts. This work introduces a collection of computational tools and defines a method for treating signal from microarray datasets that is pertinent to noncoding transcripts. This method enables user to focus specifically on the noncoding signal in the collection of micorarray data, and perform a controlled and standardized analysis with more complete and newly annotated microarray probes. I have applied this analysis method to study microarray data from three different stages of early mouse embryo, by comparing the expression levels of noncoding transcripts between developmental stages. Results indicate significant increase of expression levels in the two-cell embryo compared to both one-cell stage and mature oocyte. I have also found evidence of gradual degradation of maternal transcripts after the oocyte fertilization and in the subsequent stages of mouse embryonic development.

Keywords: Noncoding RNA, transcriptome, microarray technology, computational analysis, bioinformatics

11. ŽIVOTOPIS

Rođena sam 12. studenog 1987. godine u Makarskoj, gdje sam završila osnovnu školu kao i opću gimnaziju. Preddiplomski studij biologije (smjer: molekularna biologija) na Prirodoslovno-matematičkom fakultetu upisala sam 2005. godine. Za vrijeme 3. godine prediplomskog studija, 2009. godine, obavljam laboratorijsku praksu na Zavodu za Biokemiju, Kemijski Odsjek, PMF, pod vodstvom doc.dr.sc. Ite Gruić-Sovulj, u trajanju od dva mjeseca. Završavam navedeni studij 2009. godine, stjecanjem titule prvostupnika molekularne biologije (*univ.bacc.biol.mol.)* . Iste godine upisujem diplomski studij biologije (smjer: molekularna biologija). 2010. godine obavljam laboratorijsku praksu u Laboratoriju za celularnu biokemiju, Zavod za Organsku Kemiju i Biokemiju, Institut „Ruđer Bošković“, u trajanju od četiri mjeseca, pod vodstvom dr.sc. Marije Abramić. Iste godine primljena sam na edukacijski rad u sklopu programa za međunarodne studente diplomskih studija pri Institutu SARS u Bergenu, Norveška, u trajanju od tri mjeseca, pod vodstvom dr.sc. Fabiana Rentzscha. Studentica sam 2.godine diplomskog studija.

12. DODATAK

12.1. Prikupljanje genomskih koordinata proba

```
#!/usr/bin/perl
use strict;
use warnings;
use Bio::EnsEMBL::Registry;
use Data::Dumper;

my $registry = 'Bio::EnsEMBL::Registry';

my $outpath =
'/common/ProgrammingBiology10/HomeWork/ssumic/Project/lncRNA/moe430a_moe430b.txt';
open OUT, '>', $outpath;

$registry->load_registry_from_db(
    -host => 'ensembldb.ensembl.org',
    -user => 'anonymous'
);

my @arrays=('MOE430A','MOE430B');

foreach my $arr(@arrays){
       my $array = $oaa->fetch_by_name_vendor($arr, 'AFFY');

       $probes= $array->get_all_Features();

       foreach my $probe(@probes){
              my @pfeatures= @{$pfeature_adaptor->fetch_all_by_Probe($probe)};
              print OUT "@pfeatures\n";
       }

}
```

12.2. Prikupljanje genomskih koordinata gena, eksona i introna

```
#!/usr/bin/perl
use strict;
use warnings;
use Bio::EnsEMBL::Registry;
use Data::Dumper;

my $registry = 'Bio::EnsEMBL::Registry';

my $outpath =
'/common/ProgrammingBiology10/HomeWork/ssumic/Project/lncRNA/genes/transcripts.txt';
open OUT, '>', $outpath;

my $outpath = '/common/ProgrammingBiology10/HomeWork/ssumic/Project/lncRNA/gene61.txt';
open OUT, '>', $outpath;

$registry->load_registry_from_db(
    -host => 'ensembldb.ensembl.org',
    -user => 'anonymous'
);

my $gene_adaptor  = $registry->get_adaptor( 'Mouse', 'Core', 'Gene' );
my $slice_adaptor = $registry->get_adaptor( 'Mouse', 'Core', 'Slice' );
my $transcript_adaptor = $registry->get_adaptor( 'Mouse', 'Core', 'Transcript' );

my($genes,$slice,$transcripts,$tr,$start,$end,$strand,$stable_id,$gene,$gene_id,$transcript,$t
ranscript_id);

my @chromosomes = qw(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y);

foreach my $chr(@chromosomes){
```

```
        $slice = $slice_adaptor->fetch_by_region('chromosome', $chr);
        $genes = $slice->get_all_Genes();

        foreach my $gene (@genes) {
                $dbID = $gene->dbID();
                $start = $gene->start();
                $end = $gene->end();
                $strand = $gene->strand();
                $stable_id = $gene->stable_id();

                print "Gene $stable_id $dbID $chr $start-$end($strand)\n";
                print OUT1 "$stable_id\t$dbID\t$chr\t$start\t$end\t$strand\n";
        }
        while( $gene = shift @{$genes}){
                $transcripts = $gene->get_all_Transcripts();
                $gene_id=$gene->stable_id();

                while ($transcript = shift @{$transcripts}) {
                   $transcript_id=$transcript->stable_id();

                   foreach my $intron ( @{ $transcript->get_all_Exons() } ) {
                       #foreach my $intron ( @{ $transcript->get_all_Introns() } ) {
                               $stable_id  = $intron->stable_id();
                               $start = $intron->start();
                               $end = $intron->end();
                               $strand = $intron->strand();
                               print OUT2
"$gene_id\t$transcript_id\t$chr\t$start\t$end\t$strand\n";
                       }
                }
        }
}
```

12.3. Klasifikacija i anotacija proba

```
library(IRanges)

datadir<-"/common/ProgrammingBiology10/HomeWork/ssumic/Project/lncRNA/genes/"
probedir<-"/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/unique_mapping_probes/"
outdir<-"/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/protein_probes/"

##nekodirajuci transkripti:
ncRNA<-
read.table("/common/BAZE/lncRNA/mm9.ncRNA.txt",header=T,colClasses=c("character","numeric","nu
meric","character","character","character"))

##koordinate gena koji kodiraju za proteine:
gene<-
read.table(paste(datadir,"protein_coding_61_known.txt",sep=""),header=T,colClasses=c("characte
r","character","character","numeric","numeric","numeric","character"))
gene_put<-read.table(paste(datadir, "protein_coding_61_genscan.txt", sep=""),header=T,
colClasses=c("character","character","numeric","numeric","numeric","character"))
gene<-rbind(gene,gene_put)

##introni:
intron<-
read.table(paste(datadir,"transcripts.txt",sep=""),header=T,colClasses=c("character","numeric"
,"numeric","numeric"))

##eksoni:
exon<-
read.table(paste(datadir,"exons.txt",sep=""),header=T,colClasses=c("character","character","ch
aracter","character","numeric","numeric","numeric"))

##proteinski eksoni:
```

```
pr_exons<-exon[exon$Gene %in% gene$Gene,]

##eksoni nekodirajucih transkripata:
noncoding_exon<-exon[-which(exon$Gene %in% gene$Gene),]

##probe sa jedinstvenim položajem na genomu:
probemaps<-read.table(paste(probedir,"unique_moe430_mapped_probes.txt",sep=""),header=T,
colClasses=c("character","character","character","character",
"numeric","numeric","numeric","character", "character","character"))
query<-query[order(query$chr),]
subject<-subject[order(subject$chr),]
gir<-IRanges(query$start,query$end)
generange<-RangedData(gir,space=query$chr)
pir<-IRanges(probemaps_n$start,probemaps_n$end)
proberange<-RangedData(pir,space=probemaps_n$chr)
overlaps<-findOverlaps(generange,proberange)
m<-as.matrix(overlaps)
final<-cbind(query[m[,1],],probemaps_n[m[,2],])
strands<-final[,7] + final[,13]
ncRNA_put<-final[strands ==0,]
remain<-final[strands == 2| strands == -2,]
int<-intersect(ncRNA_put$probes,remain$probes)
ncRNA_strands<-ncRNA_put[-which(ncRNA_put$probes %in% int),]
```

12.4. „Maskiranje“ proba

```
library(affy)
arrays<-ReadAffy()
cleancdf <- cleancdfname(arrays@cdfName,addcdf=FALSE)

##položaji svih proba:
probemaps<-
read.table("/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/unique_mapping_probes/uni
que_moe430a_mapped_probes.txt",header=T,colClasses=c("character","character","character",
"character", "numeric","numeric","numeric","character", "character","character"))

##“nekodirajuce“ probe:
ncRNA_prb<-
read.table("/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/protein_probes/moe430a_pu
tative_ncRNA_probes_new.txt",header=T,colClasses=c("character", "character",
"numeric","numeric","numeric","character","character","character","character","numeric","numer
ic","numeric","character"))

##probe s više položaja na genomu:
mult_loc<-
read.table("/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/multi_location_probes/moe
430a_multi_location_probes.txt",header=T,colClasses=c("character","character","character",
"character", "numeric","numeric","numeric","character", "numeric","character"))

##probe koje mapiraju na kodirajuce gene:
all_protein_prb<-
read.table("/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/protein_probes/moe430a_to
_protein_genes_new.txt",header=T,colClasses=c("character", "character",
"numeric","numeric","numeric","character","character","character","character","numeric","numer
ic","numeric","character"))

##intronske probe:
introns<-
read.table("/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/protein_probes/moe430a_to
_introns_clean.txt",header=T,colClasses=c("character","numeric","numeric","numeric",
"character","character","numeric","numeric","numeric","character"))

mult_loc_prb<-as.vector(unique(mult_loc$probes))
mult_loc_set<-as.vector(unique(mult_loc$probe_id))

intron<-as.vector(unique(introns$probes))
```

```
all_probes<-rownames(probes(arrays))
all_set<-sub('\\d+$','',all_probes,perl=T)
all_protein_set<-as.vector(unique(all_protein_prb$probe_id))
all_protein_set<-unique(all_set[which(all_set %in% all_protein_set)])
protein_prb<-as.vector(unique(all_protein_prb$probes))
all_protein_probes<-all_probes[which(all_set %in% all_protein_set)]
outside<-setdiff(all_protein_probes,protein_prb)
ncRNA_probes<-as.vector(unique(ncRNA_prb$probes))
ncRNA_probes<-all_probes[which(all_probes %in% ncRNA_probes)]
ncRNA_set<-as.vector(unique(ncRNA_prb$probe_id))
ncRNA_set<-unique(all_set[which(all_set %in% ncRNA_set)])
all_ncRNA<-c(ncRNA_probes,intron)
all_ncRNA_set<-unique(sub('\\d+$','',all_ncRNA,perl=T))
pure_protein_prb<-all_protein_probes[-which(all_protein_probes %in% all_ncRNA)]
all_pr_set<-unique(sub('\\d+$','',pure_protein_prb,perl=T))
overlap_sets<-intersect(all_pr_set,all_ncRNA_set)
maskprobes<-unique(c(pure_protein_prb,mult_loc_prb,outside,diff))
clear<-sub('\\d+$','',maskprobes,perl=T)
sets<-unique(clear)
cnt<-sapply(1:length(sets),
function(i)length(clear[clear == sets[i]]))
m_set<-sets[cnt >10]
maskprobes<-maskprobes[-which(clear %in% m_set)]
maskset<-all_pr_set[-which(all_pr_set %in% overlap_sets)]
maskset<-unique(c(maskset,m_set,i))

##maskiranje:
RemoveProbes(listOutProbes=maskprobes,
listOutProbeSets=maskset,cleancdf)
ResetEnvir(cleancdf)
```

12.5. Normalizacija intenziteta i analiza diferencijalne ekspresije

```
library(gcrma)
library(limma)
dir="/common/ProgrammingBiology10/HomeWork/ssumic/Project/lncRNA/"
probemaps<-
read.table("/common/BAZE/lncRNA/mapped_probes/numbered_mapped_probes/unique_mapping_probes/uni
que_moe430_mapped_probes.txt",header=T,colClasses=c("character","character","character",
"character", "numeric","numeric","numeric","character", "character","character"))

frame<-data.frame(first=c("one_cell","two_cell","two_cell"),
second=c("oocyte","one_cell","oocyte"))

for(i in 1:length(frame[,1])){
       setwd(paste(dir,paste(frame[i,2],"/",sep=""),sep=""))
       arrays1<-ReadAffy()
       first_id<-sampleNames(arrays1)
       setwd(paste(dir,paste(frame[i,1],"/",sep=""),sep=""))
       arrays2<-ReadAffy()
       sec_id<-sampleNames(arrays2)
       #data frame with annotation:
       a<-rep(c("first","second"),c(length(first_id),length(sec_id)))
       SampleID<-c(first_id,sec_id)
       desc<-data.frame(SampleID,a)
       print(desc)
       rownames(desc)=desc$SampleID
       arrays<-merge(arrays1,arrays2)
       mt<-match(rownames(desc),sampleNames(arrays))
       vmd = data.frame(labelDescription = c("Sample ID",
       "status: second or first"))
       phenoData(arrays) = new("AnnotatedDataFrame",
       data = desc[mt, ], varMetadata = vmd)
       arraysRMA= gcrma(arrays)
       des<-as.factor(a)
```

```
        design = model.matrix(~des)
        print(design)
        arrayslim = lmFit(arraysRMA, design)
        arraysEB = eBayes(arrayslim)
        genes<-arraysEB$genes
        p<-arraysEB$p.value[,2]
        FC<-arraysEB$coefficients[,2]
        sig<-data.frame(genes,FC,p)
        sub<-sig[order(sig$FC,decreasing=T),]
        mat<-match(all$probe_id,sub$ID)
        all<-cbind(all,fold_ch=sub[mat,2])
}
```

12.6. Izrada prikaza rezultata

```
##Histogrami razdiobe:
hist(data$mean,col="blue", xlab="Normalizirani intenziteti skupova probi", ylab="Broj skupova
probi",breaks=40)

##MA grafovi:
plot(ratio~mean,col="dark green",pch='°',xlab="Prosjecni intenziteti dvostanicnog stadija i
stadija jajne stanice",ylab="Omjeri intenziteta skupova probi")

##Mape intenziteta:
selected  <- p.adjust(arraysEB$p.value[, 2])<0.05
esetSel <- arraysRMA[selected, ]
heatmap.2(exprs(esetSel),scale="none",col=redgreen(75),cexRow=0.7,trace="none",
density.info="none")

##Volcano grafovi:
lod = -log10(arraysEB$p.value[,2])
plot(FC, lod, pch="°", xlab="Razlika intenziteta skupova probi",col="dark blue",
ylab=expression(-log[10]~p))
abline(h=2)
```