Računalna analiza dugih nekodirajućih RNA ogulinske špiljske spužvice (Eunapius subterraneus)

Bodulić, Kristian

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Kristian Bodulić

Računalna analiza dugih nekodirajućih RNA ogulinske špiljske spužvice (*Eunapius subterraneus*)

Diplomski rad

Zagreb, 2020.

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

Zahvaljujem mentoru prof. dr. sc. Kristianu Vlahovičeku na stručnom vodstvu te pruženim savjetima, znanju i vremenu.

Zahvaljujem Grupi za bioinformatiku na stečenom znanju i iskustvu te ugodnim trenutcima provedenim u uredu u posljednje dvije godine.

Posebno zahvaljujem obitelji i prijateljima na velikoj podršci.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

RAČUNALNA ANALIZA DUGIH NEKODIRAJUĆIH RNA OGULINSKE ŠPILJSKE SPUŽVICE (*EUNAPIUS SUBTERRANEUS*)

Kristian Bodulić
Rooseveltov trg 6, 10000 Zagreb. Hrvatska

Pojavom metoda sekvenciranja druge generacije, duge nekodirajuće RNA postale su vrlo zanimljiv predmet bioloških istraživanja. Njihove uloge dokazane su u velikom broju bioloških procesa, od kojih je najvažnije spomenuti regulaciju ekspresije brojnih gena. Ipak, ova skupina RNA još uvijek nije istražena u brojnim koljenima životinja, uključujući i spužve. Radi bolje karakterizacije navedene skupine RNA spužvi, u ovom istraživanju korištene su dvije sekvencirane knjižnice RNA primarnih kultura ogulinske špiljske spužvice. Serijom bioinformatičkih metoda u sastavljenim transkriptomima pronađen je velik broj dugih nekodirajućih RNA. Utvrđeno je postojanje brojnih sličnih svojstava dugih nekodirajućih RNA spužvi i drugih životinjskih koljena, uključujući izrezivanje introna, mogućnost alternativnog prekrajanja i pristranost u sastavu nukleotida. Također, ustanovljena je potencijalna uloga transpozona u nastanku dugih nekodirajućih RNA, kao i moguća uloga navedene skupine RNA u regulaciji ekspresije brojnih protein-kodirajućih gena. Isto tako, opisan je raznolik skup stabilnih sekundarnih struktura dugih nekodirajućih RNA. Na kraju, pronađena je relativno slaba očuvanost dugih nekodirajućih RNA u koljenu spužvi, pri čemu iznimku čini izrazito velika sličnost navedene skupine RNA ogulinske špiljske spužvice i srodne spužve *Ephydatia mulleri*, što baca novo svjetlo na raspravu o mogućnosti pogrešne klasifikacije ogulinske špiljske spužvice.

(59 stranica, 25 slika, 8 tablica, 84 literaturnih navoda, jezik izvornika: hrvatski)
Rad je pohranjen u Središnjoj biološkoj knjižnici
Ključne riječi: nekodirajući dijelovi genoma, koljeno spužvi, transkriptom, transpozoni, očuvanost RNA
Voditelj: Dr. sc. Kristian Vlahoviček, red. prof.
Ocjenitelji: Dr. sc. Kristian Vlahoviček, red. prof.
Dr. sc. Damjan Franjević, izv. prof.
Dr. sc. Duje Lisičić, doc.

Rad prihvaćen: 3.9.2020.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Division of Biology

Graduation Thesis

COMPUTATIONAL ANALYSIS OF LONG NON-CODING RNA IN ENDEMIC CAVE SPONGE (*EUNAPIUS SUBTERRANEUS*)

Kristian Bodulić
Rooseveltov trg 6, 10000 Zagreb. Croatia

With the advances in next-generation sequencing technologies, long non-coding RNA have become one of the focal points of RNA research. Their roles have been demonstrated in a plethora of biological processes, most important being the regulation of gene expression. However, key information surrounding this type of RNA in many phyla is missing, which is why this research focused on characterizing long non-coding RNAs in sponges (Porifera). In this study, I used two RNA libraries isolated from the endemic cave sponge to identify a comprehensive set of highly-confident long non-coding RNAs. This study showed many similarities between long non-coding RNAs of sponges and other animals, including the phenomena of intron excision, alternative splicing, and nucleotide content bias. Furthermore, this research demonstrated the potential role of transposable elements in the origin of sponges' long non-coding RNAs, as well as a possible role of these RNAs in the regulation of expression of many important genes. Finally, this study found a low level of conservation of sponges' long non-coding RNAs, with the exception of the similarity found between endemic cave sponge and a closely related sponge, *Ephydatia mulleri*, which sheds new light on the potential misclassification of the endemic cave sponge.

(59 pages, 25 figures, 8 tables, 84 references, original in: Croatian)
Thesis deposited in the Central Biological library
Key words: non-coding genomic elements, phylum Porifera, transcriptome, transposones, RNA conservation
Supervisor: Dr. Kristian Vlahoviček, Prof.
Reviewers: Dr. Kristian Vlahoviček, Prof.
Dr. Damjan Franjević, Prof.
Dr. Duje Lisičić, Asst. Prof.

Thesis accepted: September 3, 2020.

Sadržaj

Popis kratica

ENCODE - Encyclopedia of DNA Elements

lncRNA – Duge nekodirajuće RNA

pb – Par baza

mRNA – Glasnička RNA

TERRA - Telomeric repeat-containing RNA

Xist - X-inactive specific transcript

circRNA – Circular RNA

LTR – Long terminal repeat

LINE – Long interprespersed nuclear elements

SINE – Short interprespersed nuclear elements

COI1 - Cytochrome C oxidase subunit 1

ITS2 - Internal transcribed spacer

ONT – Oxford Nanopore Technology

Mb - Megabaza

rRNA – Ribosomska RNA

BLAST – Basic Local Alignment Search Tool

DIAMOND - Double Index Alignment of Next-generation sequencing Data

HMM - Hidden Markov model

FEELnc - Flexible Extraction of LncRNAs

GO – Gene Ontology

kb – kilobaza

FPKM - Fragments per kilobase milion

SCI - Structure conservation index

SVM – Support Vector Machines

CM – Covariance model

IR – Interkvartilni rasppon

1. Uvod

1.1. Duge nekodirajuće RNA

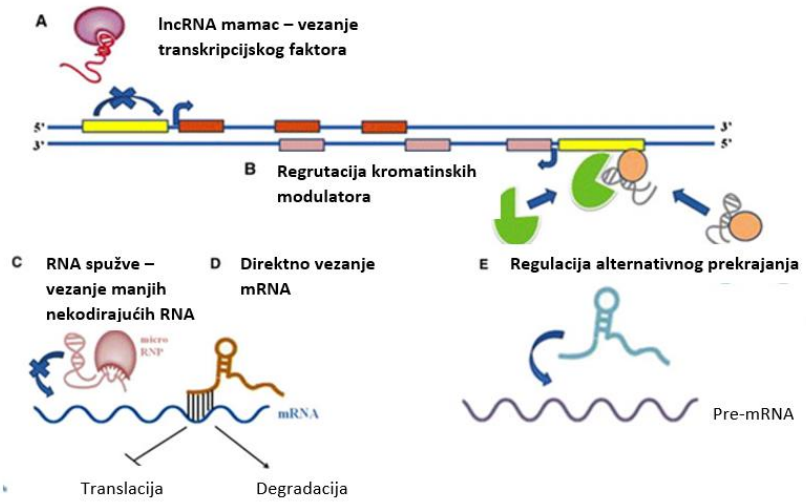
Završetkom projekta sekvenciranja ljudskog genoma ubrzo je postalo jasno da egzoni protein-kodirajućih gena čine tek 1.2% genoma čovjeka (*International Human Genome Sequencing Consortium* 2001). S druge strane, eksperimenti provedeni na gotovo svim ljudskim staničnim linijama u okviru projekta ENCODE (engl. *Encyclopedia of DNA Elements*) dokazali su prisutnost mnogih do tada nepoznatih RNA, pri čemu je procijenjeno da najmanje 93% ljudskog genoma ima sposobnost transkripcije, od čega 54 % obuhvaća gene koji ne kodiraju proteine (Dunham i sur. 2012). Ovakav trend uočen je kod brojnih razreda kralješnjaka, ali i kod ostalih životinja. Spomenut katalog RNA čine brojne vrste nekodirajućih RNA, poput RNA uključenih u proces transkripcije, alternativnog prekrajanja ili regulacije ekspresije gena. Jedna od ključnih vrsta RNA uključenih u regulaciju ekspresije gena, ali i brojne druge stanične procese, su duge nekodirajuće RNA (engl. *long non-coding* RNA, lncRNA). Naime, većina molekula RNA kodiranih u ljudskom genomu, ali i genomu brojnih drugih organizama, pripada skupini molekula lncRNA, zbog čega se u znanstvenoj zajednici postavljaju brojna pitanja o njihovu nastanku, očuvanosti, funkciji, ali i ulozi u brojnim biološkim procesima te u mehanizmima nastanka raznih bolesti (Ponjavic i sur. 2007).

1.1.1. Opće značajke dugih nekodirajućih RNA

Duge nekodirajuće RNA definirane su kao molekule RNA duže od 200 parova baza (pb) koje se ne prevode u polipeptidni lanac. Većina lncRNA-kodirajućih gena pokazuje svojstva slična protein-kodirajućim genima, uključujući prisutnost CpG otoka, prepisivanje od strane RNA polimeraze II, prisutnost egzona i introna te alternativno prekrajanje. Također, velik broj transkripata lncRNA pokazuje slične karakteristike kao i glasničke RNA (engl. *messanger RNA*, mRNA), uključujući modifikaciju 5' kraja RNA dodavanjem 7-metilgvanozinske kape, kao i modifikaciju 3' kraja RNA dodavanjem poliadeninskog repa (Romero-Barrios i sur. 2018). Većina lncRNA-kodirajućih gena čovjeka relativno je slabo eksprimirana, pri čemu se procjenjue da je količina transkripata lncRNA u različitim staničnim linijama čovjeka nekoliko puta manja od količine transkripata mRNA. Isto tako, smatra se da je velika većina transkripata lncRNA specifična za pojedino tkivo čovjeka, pri čemu je najviše različitih transkripata lncRNA otkriveno u testisima i mozgu (Jarroux i sur. 2017). Nadalje, zastupljenost pojedinih transkripata lncRNA uvelike ovisi o razvojnom stadiju, kao i o stanjima uzrokovanih okolišnim uvjetima, poput aktivacije imunološkog odgovora, hipoksije, raznih oblika stresa i slično. Isto tako, mutacije i poremećena ekspresija lncRNA-kodirajućih gena dokazane su u brojim bolestima, poput različitih tumora i Alzheimerove bolesti (Jarroux i sur. 2017). Osim uloge u različitim poremećajima, u prilog važnosti molekula lncRNA govori i njihova brojnost. Primjerice, kod čovjeka je 2019. godine opisano 17960 lncRNA-kodirajućih gena, pri čemu se očekuje da će ovaj broj uskoro nadmašiti broj protein-kodirajućih gena (Frankish i sur. 2019). Slično pravilo pokazuju i brojni drugi organizmi, zbog čega je važno staviti poseban naglasak na efikasnu anotaciju poznatih molekula lncRNA živog svijeta. Taj proces uzeo je maha u zadnjih nekoliko godina, a u njegovoj uspješnosti važnu ulogu ima karakterizacija brojnih obilježja molekula lncRNA o kojima će biti riječ u sljedećem dijelu teksta.

1.1.2. Uloge dugih nekodirajućih RNA

U posljednjih nekoliko godina dokazane su raznovrsne uloge dugih nekodirajućih RNA u različitim staničnim procesima, od kojih je najčešća regulacija ekspresije gena. Smatra se da molekule lncRNA mogu kontrolirati aktivnost gena na različite načine, pri čemu mogu djelovati cis, odnosno na susjedne gene, ili trans, odnosno na udaljene gene (Slika 1).



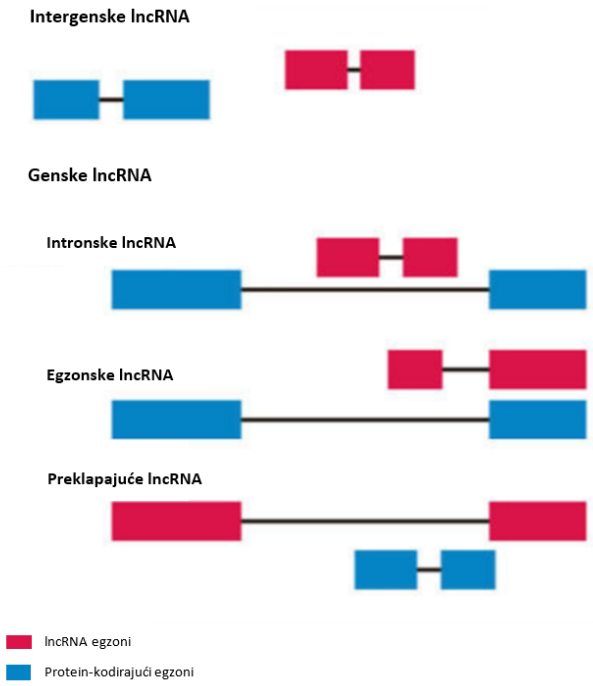
Slika 1. Pregled načina regulacije ekspresije gena od strane dugih nekodirajućih RNA. A) Transkripti lncRNA mogu imati ulogu mamca koji vežu transkripcijske faktore, čime sprječavaju njihovo vezanje za promotor protein-kodirajućeg gena. B) Transkripti lncRNA kontroliraju ekspresiju gena dovodeći modulatore kromatina. C) Transkripti lncRNA mogu imati ulogu spužvi (engl. RNA *sponges*), pri čemu privlače male nekodirajuće RNA, čime sprječavaju njihovu degradaciju transkripata mRNA. D) Transkripti lncRNA imaju sposobnost direktne interakcije s transkriptima mRNA, pri čemu mogu poboljšati efikasnost translacije mRNA ili njihove razgradnje. E) Transkripti lncRNA mogu djelovati na proces alternativnog prekrajanja. Preuzeto i prilagođeno iz Fatima i sur. 2015.

Osim uloge u regulaciji ekspresije gena, transkripti lncRNA sudjeluju u brojnim drugim staničnim procesima. Primjerice, ustanovljeno je da transkripti lncRNA mogu alosterički regulirati brojne enzime i receptore, čime nadziru njihovu funkciju (Fatima i sur. 2015). Također, pronađeni su transkripti lncRNA koji u kompleksu s jezgrinim proteinima reguliraju njezinu arhitekturu (Somarowthu i sur. 2015). Isto tako prisutnost lncRNA-kodirajućih gena dokazana je i u centromernom području, pri čemu nastaju transkripti koji pomažu u vezanju kinetohornih proteina za centromeru kromosoma (Quénet i Dalal 2014). Osim centromernih područja, i telomerna područja imaju mogućnost transkripcije, pri čemu nastaje telomerna lncRNA (engl. *telomeric repeat-containing* RNA, TERRA), koja kod brojnih organizama regulira proces skraćivanja telomera (Feuerhahn i sur. 2010). Navedene uloge molekula lncRNA ukazuju na njihovu izrazitu važnost u mnoštvu staničnih procesa.

1.1.3. Odnos dugih nekodirajućih RNA i protein-kodirajućih gena

Budući da velik broj dugih nekodirajućih RNA djeluje na ekspresiju susjednih protein-kodirajućih gena, česta je podjela lncRNA-kodirajućih gena obzirom na njihov odnos s najbližim protein-kodirajućim genom (Slika 2). Spomenuta podjela korištena je od strane

mnogih velikih projekata anotacije genoma sisavaca, poput projekta ENCODE, kao i od strane brojnih istraživačkih grupa.



Slika 2. Podjela dugih nekodirajućih RNA prema odnosu s najbližim protein-kodirajućim genom. Obzirom na preklapanje s protein-kodirajućim genima, lncRNA-kodirajući geni mogu biti intergenski (ne preklapaju se s protein-kodirajućim genima) i genski (preklapaju se s protein-kodirajućim genima). Genski lncRNA-kodirajući geni podijeljeni su na intronske (nalaze se u intronu protein-kodirajućeg gena), preklapajuće (u njihovu intronu nalazi se protein-kodirajući gen) i egzonske (njihovi egzoni preklapaju se s egzonima protein-kodirajućih gena). Preuzeto i prilagođeno iz Feng i sur. 2014.

Članovi svake od navedenih skupina lncRNA-kodirajućih gena dijele određene karakteristike. Primjerice, intergenske lncRNA čovjeka obično imaju zajedničke histonske oznake (Khalil i sur. 2009). Također, velika većina intergenskih lncRNA-kodirajućih gena pokazuju mogućnost alternativnog prekrajanja i uređivanja RNA 7-metilgvanozinskom kapom te poliadeninskim repom. Velik broj transkripata intergenskih lncRNA pronađen je u jezgri, gdje ima funkciju regulacije ekspresije gena već spomenutim mehanizmima, poput uređivanja kromatina, regulacije alternativnog prekrajanja ili vezanja transkripcijskih faktora . Isto tako, dokazana je uloga intergenskih lncRNA u održavanju strukturne i arhitektonske organizacije jezgre (Jarroux i sur. 2017). O važnosti pojedinih intergenskih lncRNA u prilog govore dokazi o njihovoj relativno visokoj razini ekspresije, ali i relativno visokoj očuvanosti kod većine razreda kralješnjaka (Jarroux i sur. 2017).

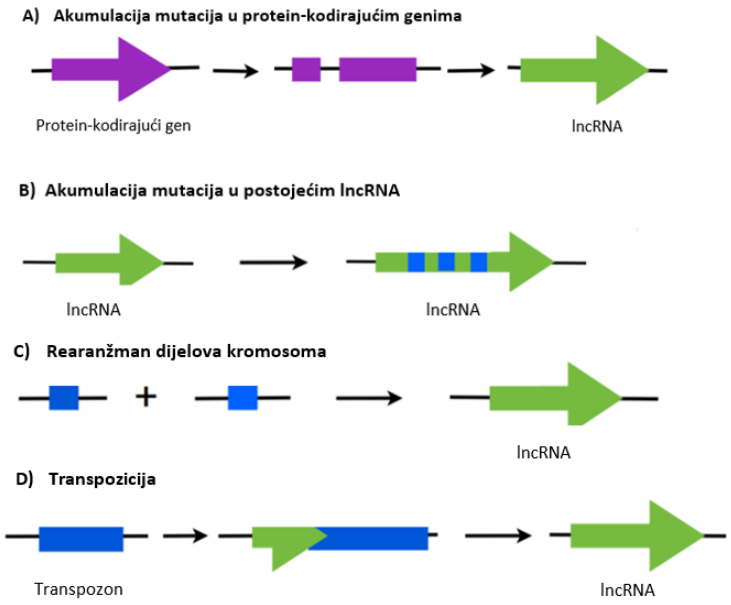
Unutar svake od navedenih vrsta genskih lncRNA postoje članovi koji imaju sposobnost transkripcije s istog (+ smjer), ali i sa suprotnog lanca (- smjer) u odnosu na transkripciju odgovarajuće mRNA. Primjerice, čest je slučaj trannskripcije protein-kodirajućeg gena u - smjeru. Takvi transkripti lncRNA mogu se preklapati s egzonima, netranslatiranim regijama ili promotorima protein-kodirajućih gena, a obično imaju ulogu regulacije ekspresije pripadnog protein-kodirajućeg gena. Primjerice, egzonske lncRNA prepisane u - smjeru, zbog svoje

djelomične ili potpune komplementarnosti egzonima mRNA, često reguliraju ekspresiju gena već spomenutim mehanizmom vezanja odgovarajućeg transkripta mRNA, pri čemu mogu pospješiti efikasnost translacije, ali i degradacije transkripta mRNA. S druge strane, egzonske lncRNA prepisane u + smjeru još uvijek nisu dovoljno istražene zbog poteškoća u razlikovanju takvih transkripata lncRNA i izoformi odgovarajućih transkripata mRNA (Derrien i sur. 2012).

Intronske lncRNA također su relativno slabo istražene. Većina intronskih lncRNA sisavaca nastaju kao samostalni transkripti, dok manji broj nastaje tijekom procesiranja mRNA. Ipak, kod čovjeka su dokazane određene uloge intronskih lncRNA, poput pozitivne regulacije transkripcije ili kontrole alternativnog prekrajanja protein-kodirajućeg gena domaćina (Jarroux i sur. 2017). Slično, uloga većine anotiranih preklapajućih lncRNA čovjeka nije poznata. Međutim, pokazano je da pojedine preklapajuće lncRNA čovjeka imaju važnu ulogu u diferencijaciji, ali i karcinogenezi (Shahryari i sur. 2015). Također, intronske i preklapajuće lncRNA mogu formirati kružne RNA molekule (engl. circular RNA, circRNA), izrazito stabilne RNA s različitim ulogama u regulaciji ekspresije gena. Primjerice, pronađene su citoplazmatske circRNA s ulogom RNA spužvi, koje inhibiraju ulogu manjih nekodirajućih RNA (Hansen i sur. 2013). Iako spomenute četiri klase lncRNA imaju mnogo zajedničkih obilježja, poboljšanje funkcionalne anotacije lncRNA svih organizama u narednom vremenu doprinijet će preciznijem definiranju takvih obilježja.

1.1.4. Nastanak dugih nekodirajućih RNA

Glavni razlog postojanja izuzetno velikog broja različitih dugih nekodirajućih RNA zasigurno je veliki broj različitih mehanizama njihova nastanka (Slika 3).



Slika 3. Mehanizmi nastanka dugih nekodirajućih RNA. A) LncRNA-kodirajući geni često nastaju nakupljanjem mutacija u kopiji protein-kodirajućeg gena. B) LncRNA-kodirajući geni mogu nastati akumulacijom mutacija u kopiji postojećeg lncRNA-kodirajućeg gena. C) LncRNA-kodirajući geni mogu nastati rearanžmanom pojedinih dijelova kromosoma. D) Nastanak lncRNA-kodirajućih gena moguć je i insercijom transpozona u određena mjesta u genomu. Preuzeto i prilagođeno s https://mcmanuslab.ucsf.edu/node/251.

Nakupljanje mutacija u protein-kodirajućim genima jedan je od mehanizama nastanka lncRNA-kodirajućih gena kojima se daje sve više važnosti. Duplikacijom određenog gena i nakupljanjem mutacija u novonastaloj kopiji nastaju pseudogeni, sljedovi bez mogućnosti translacije s relativno velikom sličnošću protein-kodirajućem genu iz kojeg su nastali. Ipak, postoji i relativno mali broj pseudogena koji zbog nakupljanja velikog broja mutacija ne pokazuju sličnost s protein-kodirajućim genima. Iako se pseudogeni ne translatiraju u proteine, postoje brojni dokazi o mogućnosti njihove transkripcije, pri čemu nastaju transkripti lncRNA. Jedan od najpoznatijih primjera takvih molekula lncRNA kod sisavaca je Xist (engl. X-*inactive specific transcript),* koji ima važnu ulogu u regulaciji inaktivacije jedne kopije kromosoma X (Duret i sur. 2006). Svi spomenuti mehanizmi djelovanja transkripata lncRNA pronađeni su kod pseudogena. Primjerice, pseudogeni se mogu transkribirati u + smjeru u odnosu na roditeljski protein-kodirajući gen, pri čemu obično nastaju transkripti lncRNA s ulogom vezanja manjih nekodirajućih RNA. Ipak, češće je prepisivanje u – smjeru, pri čemu nastaju transkripti lncRNA s ulogom vezanja transkripata mRNA i pospješivanja njihove translacije ili degradacije. Također, pronađeni su pseudogeni čiji transkripti lncRNA mogu djelovati kao modulatori kromatina, mamci za privlačenje transkripcijskih faktora i slično (Milligan i Lipovich 2015).

Većina ostalih mehanizama nastanka lncRNA-kodirajućih gena temelji se na rearanžmanima genoma. Naime, smatra se da većina molekula lncRNA ima modularnu organizaciju. Drugim riječima, molekule lncRNA podijeljene su na samostalne domene ključne za njihovu funkciju. Primjerice, domene molekula lncRNA mogu imati specifičnu sekundarnu strukturu koja služi vezanju određenog proteina, ali i primarnu strukturu važnu za vezanje ciljanih nukleinskih kiselina, poput transkripata mRNA ili manjih nekodirajućih RNA. Također, lncRNA-kodirajući geni mogu sadržavati promotorske sljedove, ali i pojačivače te utišivače koji kontroliraju njihovu ekspresiju. Iz navedenog proizlazi da su mnogi lncRNA-kodirajući geni nastali rearanžmanom regulatornih sljedova i strukturnih domena (Zampetari i sur. 2018). Najčešći primjeri genomskih rearanžmana uključuju transpozone, pokretne genomske elemente koji čine relativno velik udio genoma većine životinjskih koljena. Obzirom na mehanizme kretanja po genomu, podijeljeni su u dva razreda. Prvi razred čine retrotranspozoni, pokretni elementi koji se pomiču po genomu stvarajući svoje kopije pomoću RNA intermedijera. Istovremeno, drugi razred transpozona čine DNA transpozoni, pokretni elementi koji obično mijenjaju svoj položaj izrezivanjem i insercijom na drugo mjesto u genomu. Oba razreda transpozona dodatno su podijeljena obzirom na strukturne značajke, pri čemu tri glavne skupine retrotranspozona čine transpozoni s elementima LTR (engl. *long terminal repeats*), transpozoni LINE (engl. *long interprespersed nuclear elements*) i transpozoni SINE (engl. *short interspersed nuclear elements*). (Garica-Perez i sur. 2016). Naime, pronađeno je da su lncRNA-kodirajući geni velikog broja organizama izrazito obogaćeni transpozonima, pri čemu je broj transpozona pronađen u takvim genima znatno veći od broja transpozona u protein-kodirajućim genima. Iz tog razloga, brojne hipoteze o nastanku lncRNA-kodirajućih gena često uključuju inserciju transpozona u nekodirajući slijed u genomu, pri čemu je moguće nastajanje lncRNA-kodirajućeg gena *de novo* ili pridodavanje novih uloga postojećim lncRNA-kodirajućim genima (Johnson i Guigó 2014). Osim mogućnosti insercije ranije spomenutih funkcionalnih domena u nekodirajuća područja genoma, smatra se da transpozoni u egzonima molekula

lncRNA značajno pridonose stabilizaciji njihove sekundarne strukture (Kapusta i sur. 2013). Dakle, može se reći da transpozoni igraju važnu ulogu u nastanku i funkciji brojnih lncRNA-kodirajućih gena različitih organizama te se očekuje da će nastavak istraživanja molekula lncRNA ovu pretpostavku i potvrditi.

1.1.5. Očuvanost dugih nekodirajućih RNA

Duge nekodirajuće RNA pronađene su u svim carstvima živog svijeta, kao i virusima. Također, dokazane su u skoro svim životinjskim koljenima, a najviše pažnje pridano je istraživanjima molekula lncRNA modelnih organizama i čovjeka. Iako se smatra da većina molekula lncRNA imaju određenu funkciju, toj pretpostavci ne ide u prilog relativno mala očuvanost primarne strukture većine molekula lncRNA. Primjerice, više od 50% ljudskih molekula lncRNA slabo je očuvano u redu primata, dok je 20% ljudskih molekula lncRNA specifično za porodicu Hominidae. S druge strane, brojni autori smatraju da relativno slaba očuvanost primarne strukture ne podrazumijeva nedostatak funkcionalnosti molekula lncRNA. Za razliku od proteina, stabilnost molekula lncRNA nije uvjetovana fizikalno-kemijskim svojstvima aminokiselina, što daje veću slobodu promjenama njihove primarne strukture. Također, za funkciju molekula lncRNA najvažnija je njihova tercijarna struktura, zbog čega je upravo to razina na koju primarno djeluje prirodna selekcija. Još jedan argument koji objašnjava relativno slabu očuvanost primarne strukture molekula lncRNA je mehanizam njihova nastanka. Naime, veliki broj molekula lncRNA nastao je promjenom položaja funkcionalnih domena u genomu. Drugim riječima, navedene molekule lncRNA prolaze proces prirodne selekcije tek poslije nastanka, što može objasniti postojanje velikog broja specifičnih molekula lncRNA pojedinih skupina životinja (Lee i sur. 2019). Pretpostavlja se da takve molekule lncRNA predstavljaju važnu značajku biologije pojedinih organizama, zbog čega se sve više istraživanja posvećuje karakterizaciji molekula lncRNA različitih skupina organizama, kao i mogućoj očuvanosti molekula lncRNA u različitim životinjskim koljenima, uključujući i spužve.

1.2. Koljeno Porifera (spužve)

Koljeno spužvi predstavlja skupinu organizama koja se među prvima odvojila od zajedničkog pretka životinja. To su široko rasprostranjeni morski ili slatkovodni organizmi koji izmjenu tvari s okolinom obavljaju preko pora kroz koje teče voda, što je u skladu sa sjedilačkim načinom života odraslih jedinki. Većina spužvi sastoji se od međusobno koordiniranih stanica s različitom morfologijom i funkcijama, pri čemu nije prisutno udruživanje stanica u veće organizacijske razine, poput tkiva ili organa. Spužve su relativno neistraženo koljeno životinja te su brojna pitanja o njihovoj biologiji još uvijek otvorena (Dunn i sur. 2015).

1.2.1. Morfologija spužvi

Skupinu spužvi čine morfološki raznoliki organizmi s radijalno simetričnom ili asimetričnom građom tijela. Morfologija spužvi prilagođena je njihovu mehanizmu izmjene tvari s okolišem. Naime, tijelo većine spužvi građeno je od mnoštva sitnih pora (ostija) kroz koje ulazi voda s otopljenim kisikom i hranjivim tvarima, kao i brojnim mikroorganizmima. Također, tijelo spužve čini središnja šupljina (spongocel), kao i manji broj većih otvora (oskuluma) kroz koje izlazi voda. Na histološkoj razini, tijelo spužve građeno je od raznovrsnih stanica koje tvore tri zasebna sloja. Sloj na površini spužve naziva se pinakoderma, a čine ga pločaste stanice,

pinakocite. Ove stanice analogne su epidermalnim stanicama bilateralno simetričnih životinja, a osim zaštitne uloge, imaju ulogu probavljanja čestica koje ne mogu proći kroz ostije. Središnji sloj spužve naziva se mezofil, a predstavlja želatinozni sloj ispunjen vlaknima kolagena, skeletom i različitim vrstama stanica. Najbrojniju skupinu stanica mezofila čine amebocite, stanice sa sposobnošću rediferencijacije u većinu drugih stanica, što ih čini iznimno važnima za mogućnost regeneracije spužve. Važan dio mezofila čine i sklerocite, stanice koje proizvode spikule, glavne sastavnice skeleta spužve. Ovisno o vrsti spužve, spikule mogu biti građene od silicijeva dioksida, kalcijeva karbonata ili proteina spongina. Unutarnji sloj spužve čine hoanocite, bičaste stanice na stijenkama ostija. Osim uloge pospješivanja strujanja vode kroz tijelo spužve, hoanocite imaju sposobnost fagocitoze mikroorganizama, ali i rediferencijacije u spermalne stanice. Iz navedenog je vidljivo da spužve važne životne funkcije, poput probave, disanja i izlučivanja, obavljaju pomoću specijaliziranih stanica koje predstavljaju jednostavnije oblike životinjskih organskih sustava (Matoničkin i sur. 1998).

1.2.2. Taksonomija i filogenija koljena spužvi

Koljeno spužvi obuhvaća oko 8300 opisanih vrsta, koje su podijeljene u četiri razreda: Demospongiae (kremenorožnjače), Calcarea (vapnenjače), Hexactinellida (staklače) i Homoscleromorpha, pri čemu razred kremenorožnjača čini najviše opisanih spužvi (83%). Taj razred karakteriziraju morske i slatkovodne spužve sa skeletom građenim od silicijeva dioksida ili spongina. Također, ove spužve raznovrnih su oblika, boja i veličina. S druge strane, ostala tri razreda spužvi obuhvaćaju približno jednak broj vrsta. Razred staklača karakterizira sincicijalna građa tijela, kao i spikule građene od silicijeva dioksida. Ove spužve žive isključivo na morskom dnu te većina njih dijeli simetričnu građu tijela s relativno velikim spongocelom. Nasuprot tome, spužve razreda Calcarea imaju skelet građen od kalcijeva karbonata. Karakterizira ih mala veličina tijela te pretežito žive u morskom plićaku. Spužve razreda Homosleromorpha donedavno su smatrane dijelom kremenorožnjača, no filogenetske analize svrstali su ih u poseban razred. Pripadnici ovog razreda obično žive u morskim špiljama, pri čemu je 70% opisanih vrsta pronađeno u Jadranskom moru. Epitel ovih spužvi najsličniji je epitelu ostalih životinjskih koljena zbog prisutnosti bazalne membrane. Klasifikacija spužvi u taksone niže od razreda otežana je zbog njihove relativno jednostavne morfologije te je često predmet modernih taksonomskih i filogenetskih istraživanja. (van Soest i sur. 2012).

Filogenija spužvi još uvijek nije razriješena te nudi brojna otvorena pitanja. Donedavno se smatralo da su spužve prvi organizmi koji su se odvojili od zajedničkog pretka svih životinja (Simion i sur. 2017), a jedan od glavnih dokaza za tu hipotezu bila je iznimna morfološka sličnost spužvinih hoanocita i jednostaničnih hoanoflagelata. Međutim, brojne analize utvrdile su razlike u razvoju, diferencijaciji i funkciji ovih dvaju tipova stanica, što ide u prilog hipotezi koja bič tretira kao homeoplaziju, koja se tijekom evolucije životinja pojavila više puta (Mah i sur. 2014). Također, brojne filogenetske analize koje su uz spužve uključivale i rebraše (Ctenophora) podržavaju hipotezu o rebrašima kao koljenu koje se prvo odvojilo od zajedničkih predaka svih životinja (poput Philippe i sur. 2009). Još jedno od neriješenih pitanja filogenije spužvi su odnosi između njihovih razreda. Primjerice, iako velik broj istraživanja spužve karakterizira kao monofiletsku skupinu, sve veći broj autora izdvaja razred vapnjenjača kao skupinu koja je sličnija skupini Eumetazoa od ostalih razreda spužvi, pri čemu skupini

Eumetazoa pripadaju sve ostale životinje (Borchiellini i sur. 2001). Iz navedenog je vidljivo da filogenija spužvi krije mnoga važna pitanja, a odgovore na njih mogla bi dati buduća molekularna istraživanja većeg broja vrsta spužvi.

1.2.3. Ogulinska špiljska spužvica (*Eunapius subterraneus*)

Ogulinska špiljska spužvica (Slika 4) je endemska spužva opisana od strane Sketa i Velikonje 1984. godine. To je jedina poznata slatkovodna stigobiontska spužva na svijetu, a pronađena je na šest lokaliteta u špiljama blizu Velike Kapele i Ogulina, pri čemu samo do jednog lokaliteta dopire danje svjetlo. Tijelo ogulinske špiljske spužvice može biti tanjurastog ili jajolikog oblika. Bijele je boje, a karakterizira ga iznimna osjetljivost (Bedek i sur. 2008).

Slika 4. Ogulinska špiljska spužvica (*Eunapius subterraneus*) pronađena u špilji Tounjčici. Preuzeto iz Bilandžija i sur. 2007.

Ogulinska špiljska spužvica svrstana je u razred kremenorožnjača, red Spongollida i porodicu Spongollidae (*World Porifera Database* 2020). Ipak, brojna morfološka obilježja, poput građe ličinke gemule ili građe ostija, značajno se razlikuju kod ove vrste i ostalih vrsta roda *Eunapius*. Te razlike mogle bi upućivati na adaptaciju ove vrste na stigobiontski način života, ali i na pogrešnu klasifikaciju vrste u rod Eunapius. Teoriji pogrešne klasifikacije ove vrste u prilog govore i istraživanja provedena na tri molekularna biljega (18S rDNA, molekula COI1 (engl. *cytochrome* C *oxidase subunit* 1) i ITS2 (engl. *internal transcribed spacer*)), koja ovu vrstu smještaju bliže slatkovodnim vrstama rodova *Ephydatia* i *Lubomirskia* nego drugim vrstama unutar roda *Eunapius* (Harcet i sur. 2010).

1.2.4. Sekvencirani genomi koljena spužvi

U posljednih nekoliko godina sastavljen je velik broj transkriptoma, kao i nekoliko genoma spužvi, od kojih je najkvalitetniji genom spužve *Amphimedon queenslandica,* kremenorožnjače pronađene kod Velikog koraljnog grebena. Jedna od glavnih prepreka u sastavljanju kvalitetnih genoma i transkriptoma spužvi su kontaminacije drugim organizmima, najčešće simbiontima i parazitima. Spomenuti organizmi izuzetno su filogenetski, morfološki i biokemijski raznoliki, a većina ih spada u bakterije i arheje. Simbioza ovih organizama sa spužvama obično se temelji na doprinosu brojnim biokemijskim reakcijama, poput fiksacije

dušika ili fotosinteze, kojima navedeni organizmi obogaćuju metabolički potencijal spužvi. Budući da simbionti mogu činiti do 40% biomase spužve (Taylor i sur. 2007), oni predstavljaju dodatni izazov za sastavljanje genoma i transkriptoma spužvi. Iz tog razloga, istraživanja spužvi vrlo često koriste primorfe, primarne kulture dobivene iz pročišćenog uzorka spužve.

Sekvenciranjem genoma spužve *Amphimedon queenslandica* pronađeno je više od 30 000 protein-kodirajućih gena od kojih 18 693 (63%) ima homologe u drugim životinjskim koljenima. Također, od 4670 obitelji protein-kodirajućih gena očuvanih u svim koljenima skupine Metazoa, njih 1286 (27%) pronađeno je kod spužvi. U spomenute obitelji uključeni su brojni transkripcijski faktori i drugi proteini s važnim ulogama u procesima koji predstavljaju odrednice višestaničnosti, poput regulacije staničnog ciklusa i rasta, programirane stanične smrti, međustanične adhezije, signalizacije i genske regulacije tijekom razvoja, staničnog prepoznavanja i imunosti te stanične specijalizacije. Iako kod spužvi nije ustanovljeno postojanje živčanog sustava, u genomu spužve *Amphimedon queenslandica* pronađeni su mnogi protein-kodirajući geni povezani s funkcijama senzornog sustava, poput glutaminskih, dopaminskih i seratonskih receptora. Prisutnost navedenih proteinskih obitelji u genomu spužve *Amphimedon queenslandica* govori u prilog postojanju navedenih funkcija kod zajedničkog pretka svih životinja, što upućuje na njegovu relativno visoku kompleksnost. Ipak, smatra se da se broj članova takvih obitelji kod skupine Eumetazoa višestruko povećao pri čemu su nastali brojni paralozi tih proteina koji su sudjelovali u izgradnji brojnih kompleksnih karakteristika te skupine (Srivastava i sur. 2010). Osim protein-kodirajućih gena, u izgradnji takvih karakteristika sudjelovali su i nekodirajući sljedovi, pri čemu velik broj autora stavlja u korelaciju broj i raznolikost dugih nekodirajućih RNA s kompleksnošću organizma (primjerice, Lee i sur. 2019). U tom kontekstu zanimljiva su istraživanja molekula lncRNA kod spužvi kao organizama koji su se među prvima odvojili od zajedničkog pretka životinja.

1.2.5. Duge nekodirajuće RNA koljena spužvi

Jedino istraživanje koje se bavi karakterizacijom dugih nekodirajućih RNA kod spužvi provedeno je 2015. godine na vrsti *Amphimedon queenslandica*. U četiri razvojna stadija ličinke ove spužve pronađeno je 2935 molekula lncRNA, s predstavnicima u skupinama intronskih lncRNA, egzonskih lncRNA i intergenskih lncRNA. Također, kod ove spužve pronađeni su trendovi u genomskoj organizaciji lncRNA-kodirajućih gena vrlo slični mnogim organizmima skupine Eumetazoa, uključujući i čovjeka. Spomenuti trendovi obuhvaćaju broj i duljina egozna te introna molekula lncRNA, broj izoformi molekula lncRNA i slično. Isto tako, u ovom istraživanju pronađen je velik broj transkripata lncRNA koje pokazuju promjenu ekspresije u različitim stadijima ličinke spužve *Amphimedon queenslandica*, što upućuje na dinamični ekspresijski profil ovih molekula u razvoju spužvi. Nadalje, u navedenom istraživanju utvrđena je relativno slaba očuvanost primarne strukture molekula lncRNA unutar koljena spužvi, pri čemu su pronađene dvije očuvane molekule lncRNA u vrstama *Amphimedon queenslandica* i njezinoj najsrodnijoj vrsti, *Petrosia ficiformis.* Potrebno je naglasiti da ovo istraživanje nije oubhvatilo mnoge molekule lncRNA spužve *Amphimedon queenslandica*, poput molekula lncRNA koje nisu eksprimirane u proučavanim razvojnim stadijima, ali i molekula lncRNA čiji 3' kraj nije poliadeniliran, zbog čega nisu uspješno izolirane metodama korištenim u ovom eksperimentu (Gaiti i sur. 2015). Ipak, brojnost pronađenih

molekula lncRNA ukazuju na njihovu relativnu ranu pojavu u evoluciji skupine Metazoa, zbog čega će buduća istraživanja zasigurno dati odgovore na otvorena pitanja o strukturi, funkciji i brojnosti molekula lncRNA kod drugih spužvi, ali i ostalih koljena koja su se relativno rano odvojila od zajedničkog pretka životinja.

Cilj ovog istraživanja je upotpuniti katalog dugih nekodirajućih RNA spužvi koristeći podatke dobivene sekvenciranjem ogulinske špiljske spužvice. Isto tako, cilj je karakterizirati brojna svojstva pronađenih lncRNA, kao i usporediti navedena svojstva s molekulama lncRNA pronađenim u drugim organizmima, posebnu pažnju pridajući drugim spužvama. Također, prioriteti ovog istraživanja su i analiza sekundarnih struktura pronađenih molekula lncRNA, kao i ispitivanje očuvanosti primarne te sekundarne strukture navedenih molekula lncRNA pronalaskom sličnih molekula unutar i izvan koljena spužvi. Smatram da će ovo istraživanje proširiti znanje o molekulama lncRNA koljena Porifera te dati još bolji uvid u odnos njihovih karakteristika i karakteristika molekula lncRNA drugih organizama.

2. Materijal i metode

2.1. Korištene knjižnice RNA

U ovom istraživanju koristio sam knjižnice RNA izolirane iz ogulinske špiljske spužvice prikupljene u špilji Tounjačica. RNA je izolirana iz primorfa prvi i deseti dan njihova rasta kitom *Rneasy Mini Kit* (Qiagen) te sekvencirana tehnologijom Illumina, pri čemu je korištena metoda sekvenciranja uparenih krajeva bez informacija o lancu (engl. non-stranded *paired-end sequencing*). Osnovne statistički parametri navedenih knjižnica prikazani su u Tablici 1. Za potrebe ovog istraživanja navedene knjižnice zvat ću RNA1 i RNA10, pri čemu brojevi predstavljaju dan rasta primorfa iz kojeg je izolirana RNA.

Tablica 1. Osnovne informacije knjižnica RNA dobivenih sekvenciranjem primorfa ogulinske špiljske spužvice izolirane prvih i deseti dan njihova rasta.

Knjižnica	**Prosječna ocjena kvalitete**	**Prosječni udio GC (%)**	**Broj sljedova**	**Minimalna duljina / pb**	**Prosječna duljina / pb**	**Ukupna duljina / Mb**
RNA1	29.01	48.98	388802520	50	50	19440.13
RNA10	28.02	49.26	361142658	50	50	18057.13

2.2. Korišteni genomi i transkriptomi spužvi

U ovom radu koristio sam genom ogulinske špiljske spužvice sastavljen od strane Grupe za bioinformatiku Zavoda za molekularnu biologiju Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu. Osnovne informaciju o genomu prikazane su u Tablici 2. Genom je sklopljen koristeći hibridni pristup sklapanja pomoću kraćih sljedova dobivenih tehnologijom Illumina, kao i dužih sljedova sekvenciranih tehnologijom ONT (engl. *Oxford Nanopore Technology*). Protein-kodirajući sljedovi genoma anotirani su programom BRAKER2, dok su transpozoni anotirani programima RepeatMasker i RepeatModeler2 od strane Grupe za bioinformatiku.

Tablica 2. Osnovne informacije o korištenom genomu ogulinske špiljske spužvice. Izraz prekinuti sljedovi (engl. *scaffolds*) odnosi se na sljedove dobivene postupkom određivanja relativnih pozicija i udaljenosti neprekinutih sljedova (engl. *contigs*).

Broj prekinutih sljedova	**Broj prekinutih sljedova većih od 50 kb**	**Duljina najduljeg prekinutog slijeda / pb**	**Ukupna duljina genoma / Mb**	**N50 / pb**	**L50**	**GC (%)**
3664	1138	867181	202.2	162773	4567	44.1

Radi određivanja stupnja očuvanosti dugih nekodirajućih RNA u koljenu Porifera, u istraživanju sam koristio genome 6 spužvi, kao i transkriptome 18 spužvi s predstavnicima u sva četiri njihova razreda. Osnovne informacije o korištenim genomima i transkriptomima spužvi, uključujući izvore preuzimanja, dostupne su u Prilogu 1.

2.3. Ostali korišteni podaci

Radi filtriranja molekula rRNA (ribosomska RNA) koristio sam javno dostupnu bazu molekula rRNA Silva (Quast i sur. 2013).

Radi utvrđivanja razine sličnost pronađenih dugih nekodirajućih RNA ogulinske špiljske spužvice s molekulama lncRNA ostalih životinjskih koljena, koristio sam bazu nekodirajućih RNA RNAcentral (*The RNAcentral Consortium* 2019).

Za utvrđivanje sličnosti transkripata sa svim poznatim proteinima koristio sam neredundantnu proteinsku bazu (O'Leary i sur. 2015). Također, za određivanje sličnosti transkripata s očuvanim proteinskim domenama koristio sam bazu domena Pfam (El-Gebali i sur. 2019). Isto tako, za pronalazak potencijalnih homologa molekula lncRNA izvan koljena Porifera koristio sam bazu očuvanih domena RNA Rfam (Kalvari i sur. 2017).

2.4. Obrada sljedova knjižnica RNA1 I RNA10

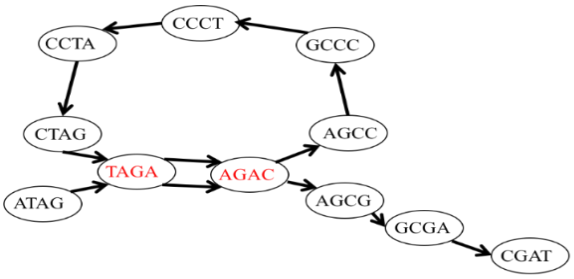
Na početku istraživanja sljedove knjižnica RNA1 i RNA10 obradio sam programom Bbduk, dio paketa BBTools verzije 36.20. (Bushnell 2019). Obrada sljedova uključivala je uklanjanje adaptera korištenih u procesu sekvenciranja i filtriranje sljedova po kriteriju kvalitete. Filtriranje po kvaliteti temeljilo se na uklanjanju baza na krajevima sljedova s kvalitetom manjom od 8 te isključivanjem sljedova s prosječnom kvalitetom manjom od 16. Jedinice kvalitete zadane su bodovima kvalitete (engl. *Phred score*), koji predstavljaju negativni logaritam vjerojatnosti pogrešnog očitanja određene baze pomnožen s 10. Osnovne statističke parametre obrađenih sljedova izračunao sam koristeći program Seqkit verzije 0.12.1. (Shen i sur. 2016), programski jezik R verzije 4.0.2. (R Core Team 2017) i paket data.table verzije 1.13.0. (Dowle i Srinivasan. 2019). U nastavku istraživanja koristio sam programski jezik R navedene verzije. Za paralelizaciju procesa pokrenutih u ljusci Bash u svim dijelovima istraživanja koristio sam program GNU Parallel (verzija 23/ 05/2020, Tange 2020). Programski kod korišten u istraživanju dostupan je na poveznici u Prilogu 2.

2.5. Sastavljanje transkriptoma

Za sastavljanje transkriptoma primorfa ogulinske špiljske spužvice istovremeno sam koristio knjižnice RNA1 i RNA10, pri čemu sam upotrijebio programe Trinity verzije 2.8.6. (Grabherr i sur. 2013) i rnaSPAdes, dio paketa *SPAdes* verzije 3.14. (Bankevich i sur. 2012).

Budući da je sekvenciranje većinom današnjih tehnologija popraćeno fragmentiranjem nukleinskih kiselina, sastavljanje genoma i transkriptoma obično se temelji na pronalaženju sljedova koji se preklapaju te na utvrđivanju odnosa između takvih sljedova, poput njihova rasporeda i udaljenosti. Pri sastavljanju genoma i transkriptoma pomoću kraćih sljedova obično se koriste tzv. de Brujinovi grafovi, odnosno mreže koje se sastoje od vrhova (engl. *nodes*) i poveznica između njih (engl. *edges*). Takav pristup temelji se na podjeli svih sljedova na manje sljedove veličine k (k-meri). Spomenuti k-meri zatim se koriste za sastavljanje grafa, pri čemu vrhovi grafa predstavljaju same k-mere, a poveznice između dva k-mera ukazuju na preklapanje točno k-1 slova između dva k-mera (Slika 5). Na taj način, sekvencirani sljedovi implicitno su sadržani u putu prolaska kroz graf. Gotovo svi programi za sastavljanje genoma

i transkriptoma koriste informacije dostupne u sekvenciranim sljedovima kako bi poboljšali strukturu grafa i uklonili njegove nekonzistentne dijelove. U navedenom pristupu postoje mnoge poteškoće, od kojih valja izdvojiti postojanje višestrukih ponavljanja većih od duljine sekvenciranih sljedova, koji uzrokuju fragmentaciju sastavljenog genoma, odnosno u manjoj mjeri transkriptoma. Isto tako, transkriptome obično karakterizira postojanje mnogih izoformi pojedinih transkripata, čija međusobna sličnost otežava precizno pridodavanje sljedova određenoj izoformi. Također, zbog malog udjela sljedova nastalih iz slabije eksprimiranih gena, njihove transkripte vrlo je teško precizno rekonstruirati (Nagarajan i Pop 2013). Iz tih razloga, dostupan je velik broj različitih programa koji na brojne načine pokušavaju savladati navedene prepreke.



Slika 5. Primjer de Brujinova grafa (k=5) za slijed ATAGACCCTAGACGAT. Crvenom bojom označeno je dvostruko ponavljanje. Preuzeto s http://data-science-sequencing.github.io/Win2018/lectures/lecture7.

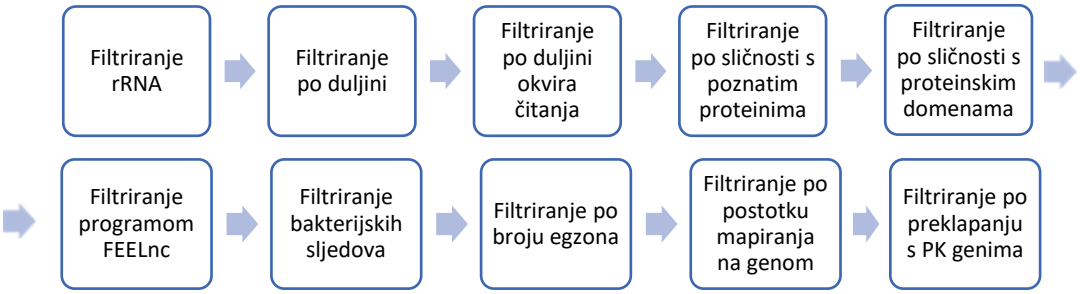
Program Trinity primjenjuje se za *de novo* sastavljanje transkriptoma pomoću kraćih RNA sljedova. Na početku procesa sastavljanja transkriptoma, program Trinity sastavlja listu svih prisutnih k-mera u ulaznim sljedovima, od kojih zatim bira najčešći k-mer pomoću kojeg gradi neprekinuti slijed s drugim k-merima, temeljen na preklapanju u k-1 slova. Proces se zatim ponavlja sa svim ostalim k-merima redoslijedom njihove frekvencije u sekvenciranim sljedovima dok se ne iskoristi cijeli rječnik k-mera. Zatim, dobiveni neprekinuti sljedovi se na temelju preklapanja u točno k-1 slova grupiraju, pri čemu se za svaku grupu gradi de Brujinov graf. Svaka od grupa predstavljena de Brujinovim grafom zapravo odgovara skupu izoformi istog gena kao i transkriptima paralognih gena. Tako nastalim de Brujinovim grafovima pridružuju se sljedovi na temelju njihova preklapanja s najvećim brojem k-mera. Informacije u tim sljedovima na kraju se koriste u rekonstrukciji transkripata, pri čemu se pokušava razriješiti pripadnost sljedova pojedinim izoformama ili paralozima odgovarajućih transkripata (Grabherr i sur. 2013).

Program rnaSPAdes također se koristi za de novo sastavljanje transkriptoma pomoću kraćih RNA sljedova. Za razliku od programa Trinity, program rnaSPAdes sastavljanje transkriptoma započinje izgradnjom de Brujinova grafa iz svih sljedova. Tako izgrađen graf pojednostavljuje se uklanjanjem netočnih poveznica između njegovih vrhova. Primjeri takvih vrhova i poveznica uključuju terminalne vrhove nastale zbog pogrešno pročitanih baza u procesu sekvenciranja, kao i dvije poveznice s istim početnim i krajnjim vrhovima uzrokovane pogrešno očitanim bazama ili alternativnim prekrajanjem. Isto tako, proces pojednostavljivanja grafa obuhvaća i uklanjanje krivo spojenih, odnosno kimernih poveznica grafa (Bankevich i sur. 2012).

Brojne studije koje su se bavile usporedbom preciznosti i kvalitete alata za *de novo* sastavljanje transkriptoma prvo mjesto dale su upravo programima Trinity i rnaSPAdes (poput Hölzer i Marz 2019). Budući da ovi alati sastavljaju transkriptome koristeći relativno različite metode, za sastavljanje transkriptoma ogulinske špiljske spužvice odlučio sam koristiti oba od navedenih programa. Oba programa pokrenuo sam koristeći zadane parametre. Statistički parametri složenih transkriptoma dobiveni su programom Transrate verzije 1.0.3. (Smith-Unna i sur. 2016).

2.6. Pronalazak dugih nekodirajućih RNA

Strategija pronalaska dugih nekodirajućih RNA u transkriptomima primorfa ogulinske špiljske spužvice temeljila se na seriji filtriranja kojima je cilj bio minimizirati broj lažno pozitivnih i lažno negativnih rezultata. Shema postupaka filtriranja prikazana je na Slici 6, dok je postupak detaljnije objašnjen u nastavku teksta.



Slika 6. Shema postupka pronalaska dugih nekodirajućih RNA iz transkriptoma primorfa ogulinske špiljske spužvice. PK geni = protein-kodirajući geni.

2.6.1. Filtriranje rRNA

Iako je RNA iz uzoraka primorfa izolirana kitom koji se temelji na obogaćivanju molekula s 3' poliadeninskim repom, poput mRNA i većine dugih nekodirajućih RNA, moguća je prisutnost i relativno malog udjela molekula rRNA. Za uklanjanje molekula rRNA koristio sam bazu molekula rRNA spužvi Silva i program BLAST (engl. *Basic Local Alignment Search Tool*) verzije 2.2.20. (Altschul i sur. 1990).

Program BLAST jedan je od najpopularnijih programa koji se koristi za heurističko poravnavanje sljedova nukleotida ili aminokiselina, pri čemu se u određenoj bazi sljedova (ciljani sljedovi, engl *target sequences*) traže slični sljedovi jednom ili više ulaznih sljedova (engl. *query sequence*). Pritom, program BLASTn koristi se za pretragu sličnosti između dva slijeda nukleotida. Program BLAST započinje proces pronalaska sličnih sljedova rastavljanjem ulaznog slijeda na manje sljedove fiksne veličine (klice, engl. *seed*). Također, stvara se popis pozicija svih mogućih klica u pretraživanoj bazi sljedova (indeksiranje). Zatim, pronalaze se sve klice iste veličine s ocjenom poravnanja (engl. *alignment score*) koja prelazi određeni bodovni prag, pri čemu se takva ocjena najčešće računa pomoću supstitucijskih matrica. Takve klice se potom poravnavaju s ciljanim sljedovima pri čemu se u obzir uzimaju samo potpuni pogotci. Zatim, program BLAST koristi algoritam lokalnog poravnanja radi produljenja pogotka u oba smjera sve do početka smanjivanja ocjene poravnanja. Na kraju, pogodni susjedni pogotci spajaju se u jedan veći pogodak. Za svaki pogodak računaju se parametri poput postotka

identičnih nukleotida ili aminokiselina, postotka poravnatog ulaznog slijeda i očekivane vrijednosti (engl. *expected value*). Očekivana vrijednost često je korišten statistički parametar pri interpretaciji pogotka programa BLAST, a predstavlja broj takvih pogodaka koji se pretragom baze određene veličine pojavi slučajno (Altschul i sur. 1990).

Za pretragu sljedova rRNA u transkriptomu ogulinske špiljske spužvice koristio sam program BLASTn sa zadanim parametrima. Filtrirao sam sve transkripte s očekivanom vrijednošću manjom od 10^{-20}. Za obradu rezultata koristio sam programski jezik R i paket data.table.

2.6.2. Filtriranje po duljini

Radi filtriranja kraćih nekodirajućih sljedova uklonio sam sve transkripte kraće od 200 parova baza. Duljine transkripata izračunao sam programom Seqkit, a filtriranje sam napravio programskim jezikom R i paketom data.table.

2.6.3. Filtriranje po duljini okvira čitanja

Nakon filtriranja molekula rRNA i molekula RNA kraćih od 200 pb, uklonio sam sve molekule RNA s najdužim okvirom čitanja kraćim od 150 nukleotida, odnosno 50 aminokiselina. Za pronalazak najdužeg okvira čitanja svih transkripata koristio sam program Transdecoder verzije 5.5.0. (Haas i sur. 2013).

Program Transdecoder pronalazi okvire čitanja koristeći Markove modele i homologiju s poznatim proteinima. Prvi korak algoritma koji koristi Transdecoder uključuje pronalazak svih okvira čitanja duljih od određene granične duljine, s kodonom koji kodira metionin na početku okvira čitanja i stop kodonom na kraju okvira čitanja (osim u slučaju okvira čitanja na kraju transkripta). Zatim, na 500 najdužih okvira čitanja trenira se Markov model, koji se potom koristi za potvrdu svih pronađenih okvira čitanja. Također, za potvrdu okvira čitanja mogu se koristiti i rezultati pretrage po sličnosti prevedenih okvira čitanja i svih poznatih proteina (Haas i sur. 2013).

Radi utvrđivanja najdužih okvira čitanja svakog transkripta, pokrenuo sam program Transdecoder sa zadanim parametrima. Transkripte sam zatim filtrirao po duljini dobivenih okvira čitanja koristeći programski jezik R i paket data.table.

2.6.4. Filtriranje transkripata sličnih poznatim proteinima

Sljedeći korak u pronalasku dugih nekodirajućih RNA bio je uklanjanje transkripata koji pokazuju značajnu razinu sličnosti s poznatim proteinima. Programom DIAMOND (engl. *Double Index Alignment of Next-generation sequencing Data*) verzije 0.9.22. (Buchfink i sur. 2015) pretražio sam bazu neredundantnih proteina radi utvrđivanja potencijalnih sličnosti s transkriptima ogulinske špiljske spužvice.

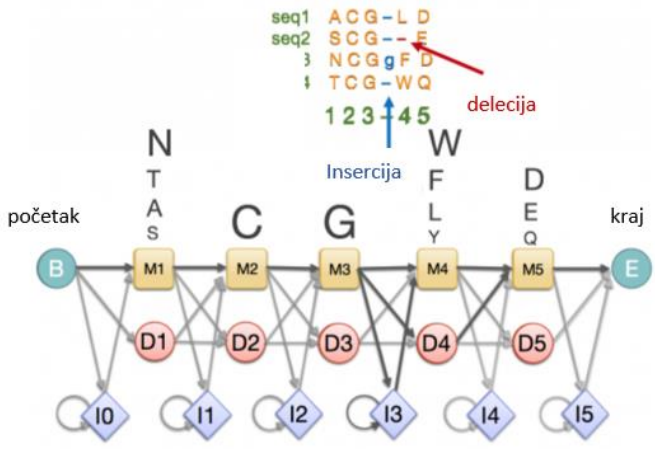
Program DIAMOND još je jedan primjer alata za heurističko poravnavanje proteinskih sljedova, pri čemu pokazuje vrlo sličnu osjetljivost, ali i višestruko povećanje brzine u odnosu na program BLAST. Takvo povećanje brzine omogućilo je nekoliko svojstava ovog programa, među kojima je najvažnije uvođenje tzv. razmaknutih klica (engl. *spaced seeds*), Takve klice duže su od klica korištenih od strane programa BLAST pri čemu se za poravnanje s ciljanim slijedom ne koriste sve pozicije klice. Također, program DIAMOND koristi modificiranu

abecedu aminokiselina, pri čemu su sve međusobno slične aminokiseline označene istim slovom. Još jedna od karakteristika programa DIAMOND koja omogućuje povećanje brzine pretrage je tzv. duplo indeksiranje, pri čemu se stvara sortiran popis klica ciljanih i ulaznih sljedova (Buchfink i sur. 2015).

Program DIAMOND omogućuje poravnanje dvaju proteinskih sljedova, kao i poravnanje slijeda nukleotida s proteinskim slijedom, pri čemu se slijed nukleotida prevodi u svih šest okvira čitanja. Radi utvrđivanja razine sličnosti transkripata sa svim poznatim proteinima, pokrenuo sam program DIAMOND koristeći transkriptome ogulinske špiljske spužvice, neredundantnu proteinsku bazu i zadane parametre. Filtrirao sam transkripte čija je očekivana vrijednost pogotka bilo kojeg proteina bila manja od 10^{-5}. Filtriranje sam proveo koristeći programski jezik R i paket data.table.

2.6.5. Filtriranje transkripata sličnih očuvanim proteinskim domenama

Nakon filtriranja transkripta sa značajnom sličnošću s poznatim proteinima utvrđenom programom DIAMOND, uklonio sam transkripte koji su pokazivali sličnost s očuvanim proteinskim domenama pronađenu alatom hmmscan dostupnim u programskom paketu HMMER (Eddy i sur. 2011). Alat hmmscan za precizno poravnanje koristi profile HMM (engl. *Hidden Markov model*), izgrađene na temelju višestrukog poravnanja proteinskih sljedova određene proteinske obitelji. Takvi profili omogućuju pretvaranje višestrukog poravnanja u sustav bodovanja ovisan o poziciji. Drugim riječima, takvi profili koriste informacije o poziciji pojedinih aminokiselina, kao i informacije o insercijama te delecijama u određenom višestrukom poravnanju. Na temelju takvih profila gradi se skup sljedova koji se zatim koriste u poravnanju s ulazim slijedom (Eddy i sur. 2011).



Slika 7. Shema postupka koji provodi alat hmmscan. Na slici je prikazan profil HMM izgrađen na temelju gornjeg poravnanja. Žutim kvadratima označeni su stupci aminokiselina, crveni krugovi označuju moguće delecije, dok plavi kvadrati predstavljaju moguće insercije. Slova iznad žutih kvadrata predstavljaju moguće aminokiselina na pojedinoj poziciji, pri čemu je veličina slova proporcionalna s vjerojatnošću pronalaska određene aminokiseline na tom mjestu. Ta vjerojatnost određena je prisutnošću pojedine aminokiseline na određenom mjestu u višestrukom poravnanju, frekvencijom pojavnosti te aminokiseline u prirodi te vjerojatnošću supstitucije ostalih aminokiselina u stupcu, koja je određena supstitucijskom matricom. Slično, insercije i delecije označene su strelicama, pri čemu je njihova vjerojatnost određena na temelju višestrukog poravnanja. Preuzeto i prilagođeno s https://www.ebi.ac.uk/training/online/.

Za pretragu transkriptoma ogulinske špiljske spužvice koristio sam alat hmmscan verzije 3.3.1. i skup očuvanih proteinskih domena predstavljenih profilima HMM. Alat sam pokrenuo koristeći translatirane transkriptome i bazu proteinskih domena Pfam sa zadanim parametrima. Programskim jezikom R uklonio sam sve transkripte sa značajnom očekivanom vrijednošću pogotka određene proteinske domene, pri čemu je kao prag značajnosti uzeta vrijednost 10^{-5}. Pritom, koristio sam pakete data.table i rhmmer verzije 0.1.0. (Arendsee 2017).

2.6.6. Pronalazak dugih nekodirajućih RNA programom FEELnc

Program FEELnc (engl. *Flexible Extraction of LncRNAs*) (Wucher i sur. 2017) jedan je od najpoznatijih programa za analizu dugih nekodirajućih RNA. Navedeni program uključuje alat codpot, koji klasificira transkripte obzirom na mogućnost kodiranja proteina. Ovaj alat temelji se na tzv. modelu nasumične šume (engl. *random forest*) koji koristi svojstva poput duljine okvira čitanja i nukleotidnog sastava RNA pomoću kojih se računa potencijal kodiranja svakog transkripta. Model se trenira na poznatim kodirajućim i nekodirajućim sljedovima organizma. Ukoliko poznati nekodirajući sljedovi nisu dostupni, model koristi sljedove dobivene nasumičnim miješanjem nukleotida kodirajućih RNA. Ulazni transkripti se zatim klasificiraju temeljem njihova potencijala kodiranja, kao i potencijala kodiranja kodirajućih i nekodirajućih sljedova iskorištenih u treniranju modela (Wucher i sur. 2017).

Iskoristio sam alat codpot programa FEELnc verzije 0.1.1. radi klasifikacije svih dosad filtriranih transkripata obzirom na njihov sastav nukleotida, koji je korišten za izračun kodirajućeg potencijala svakog transkripta. U nastavku istraživanja koristio sam transkripte koje je ovaj program klasificirao kao transkripte lncRNA.

2.6.7. Mapiranje na genom i filtriranje bakterijskih sljedova

Radi postupaka filtriranja navedenih u nastavku teksta, ali i analize dobivenih dugih nekodirajućih RNA, mapirao sam transkripte dobivene prethodnim filtriranjima na genom ogulinske špiljske spužvice koristeći program Minimap2 verzije 2.16. (Li 2018).

Program Minimap2 jedan je od novijih programa koji omogućava mapiranje različitih vrsta sljedova na referentni genom. Spomenute vrste sljedova uključuju kraće DNA ili RNA sljedove dobivene tehnologijom Illumina, duže sljedove s relativno velikom stopom pogreške, poput sljedova dobivenih tehnologijom ONT, ali i transkripte. Algoritam programa Minimap2 temelji se na indeksiranju referentnog genoma i ulaznih sljedova pri čemu se za svaki prozor veličine w stvara popis tzv. minimizatora (engl. *minimizers*), uređenih trojki minimalne vrijednosti tzv. Hash funkcije, pozicije mapiranja unutar prozora i lanca. Spomenuta Hash funkcija računa vrijednosti za svaku klicu obzirom na njezin sastav nukleotida. Kod mapiranja sljedova na referentni genom dolazi do pogotka u slučaju jednakih vrijednosti minimizatora prozora ulaznih sljedova i referentnog genoma. Navedeni pogotci mogu se spojiti u veći pogodak u slučaju ispunjavanja kriterija udaljenosti i kolinearnosti (Li 2018).

Pri mapiranju analiziranih transkriptoma na genom ogulinske špiljske spužvice koristio sam program Minimap2 s parametrima prilagođenim mapiranju transkripata („-x splice -K 10000M –c -secondary=no“). Rezultate mapiranja u formatu .sam pretvorio sam u format .bam i

sortirao po imenu ulaznih sljedova koristeći program Sambamba verzije 0.6.1. (Tarasov 2015). Zatim, uklonio sam sve sljedove koji su se primarno mapirali na bakterijske prekinute sljedove genoma koristeći programski jezik R i paket data.table, pri čemu sam u R uveo rezultate mapiranja u formatu .paf. Bakterijski sljedovi genoma anotirani su programom MEGAN od strane Grupe za bioinformatiku.

2.6.8. Filtriranje po broju egzona

Sljedeći korak ovog istraživanja bio je uklanjanje svih dugih nekodirajućih RNA s jednim egzonom radi smanjenja broja potencijalno pogrešno složenih transkripata. Koordinate egzona mapiranih na genom izračunao sam koristeći alat mpileup programa Samtools verzije 1.8. (Li i sur. 2009) te programski jezik R s paketima data.table i stringr verzije 1.4. (Vickham 2019). Filtriranje transkripata također sam napravio pomoću programskog jezika R i paketa data.table.

2.6.9. Filtriranje po postotku mapiranja na genom

Radi zadržavanja samo onih transkripata čija je točnost potvrđena genomom, uklonio sam sve transkripte koji su se mapirali na genom s manje od 95% svoje duljine. Transkripte sam uklonio koristeći programski jezik R i paket data.table.

2.6.10. Filtriranje po preklapanju s protein-kodirajućim genima

Kako bih iz transkriptoma uklonio potencijalno pogrešno složene molekule mRNA, filtrirao sam sve preostale transkripte čiji se egzoni preklapaju s egzonima protein-kodirajućih gena. Naime, budući da sekvenciranje knjižnica RNA1 i RNA10 nije uključivao dio protokola koji daje informaciju o prepisanom lancu, ne bih mogao razlikovati transkripte koji se preklapaju s protein-kodirajućim egzonima u + i – smjeru. Iz tog razloga ovo istraživanje zanemarilo je sve egzonske lncRNA.

Koordinate egzona protein-kodirajućih gena izračunate su od strane Grupe za bioinformatiku u sklopu anotacije genoma programom BRAKER2. Transkripte sam filtrirao pomoću programskog jezika R i paketa dana.table, IRanges verzije 2.22.2. (Laurence i sur. 2013) te paketa GenomicRanges verzije 1.40. (Laurence i sur. 2013). Rezultate svih opisanih filtriranja prikazao sam grafički paketom ggplot2 verzije 3.3.2. (Wickham 2016).

2.7. Analiza dugih nekodirajućih RNA

2.7.1. Određivanje konsenzusnog skupa dugih nekodirajućih RNA iz transkriptoma rnaSPAdes i Trinity

U sljedećem dijelu istraživanja odredio sam konsenzusni skup dugih nekodirajućih RNA iz oba korištena transkriptoma, pri čemu je cilj bio izdvojiti sve jedinstvene transkripte lncRNA s minimalnim stupnjem redundantnosti. Postupak određivanja konsenzusa između dva transkriptoma započeo sam pronalaskom koordinata lncRNA-kodirajućih gena. Spomenute koordinate izračunao sam kao početne i krajnje koordinate skupa preklapajućih transkripata pri čemu sam u skup uključivao samo transkripte koji su definirani kao izoforme jednog gena od strane programa rnaSPAdes i Trinity. Na taj način, izbjegao sam moguće združivanje transkripata nastalih prepisivanjem različitih gena. Zatim, odredio sam preklapanja svakog od

parova gena dobivenih iz oba transkriptoma, pri čemu sam prije definiranja preklapajućih parova uklonio gene koje se nalaze unutar dužih gena istog transkriptoma. Konsenzus lncRNA-kodirajućih gena dobio sam uzimanjem gena koji se pojavljuju u samo jednom transkriptomu, kao i uzimanjem dužeg gena od svakog preklapajućeg para čiji je presjek duži od 20% duljine dužeg gena u paru te uzimanjem kraćeg i dužeg gena svakog preklapajućeg para čiji je presjek kraći od 20% duljine dužeg gena u paru. Konsenzus transkripata dobio sam preklapanjem transkripata sa spomenutim konsenzusom gena. Osnovne podatke o broju preklapajućih i jedinstvenih gena te transkripata prikazao sam tablično, dok sam raspodjelu postotka preklapanja parova gena iz oba transkriptoma prikazao grafički. Za ovu analizu koristio sam programski jezik R, kao i pakete data.table, ggplot2, IRanges, GenomicRanges, DescTools verzije 0.99.37. (Signorell i sur. 2020).

2.7.2. Analiza osnovnih svojstava pronađenih dugih nekodirajućih RNA

U sljedećem dijelu analize opisao sam osnovna svojstva pronađenih dugih nekodirajućih RNA, poput broja izoformi, broja i duljine egzona te introna, duljine transkripata, mjesta izrezivanja introna i udjela GC. Broj izoformi odredio sam brojanjem transkripata koji se preklapaju s lncRNA-kodirajućim genima, vodeći računa o brojanju transkripata koji su definirani kao izoforme jednog gena od strane programa rnaSPAdes i Trinity. Također, odredio sam broj izoformi protein-kodirajućih gena koristeći anotaciju programom BRAKER2 te sam dobivene raspodjele udjela gena s pojedinim brojem izoformi usporedio. Isto tako, usporedio sam raspodjelu udjela lnRNA-kodirajućih i protein-kodirajućih gena s određenim brojem egzona. Zatim, pomoću koordinata egzona odredio sam koordinate introna lncRNA-kodirajućih gena i protein-kodirajućih gena te sam Wilcoxonovim testom međusobno usporedio raspodjele duljina gena, transkripata, egzona i introna navedenih gena. Potom, usporedio sam zastupljenost pojedinih mjesta prekrajanja kod lncRNA-kodirajućih gena i protein-kodirajućih gena računajući udjele svih mogućih kombinacija dinukleotida na početku i kraju introna. Nadalje, usporedio sam udio GC transkripata lncRNA i mRNA pomoću t-testa. U svakom dijelu navedene analize koristio sam egzone i introne izoformi s najviše egzona, pri čemu sam u slučaju postojanja više takvih izoformi odabrao dulju. Sve analize opisane u ovom poglavlju popratio sam pripadajućim grafičkim prikazima. Pritom sam koristio programski jezik R i pakete data.table, IRanges, GenomicRanges, ggplot2, Biostrings verzije 4.0. (Pagès i sur. 2020) te paket BSGenome verzije 4.0. (Pagès i sur. 2020).

2.7.3. Analiza odnosa pronađenih dugih nekodirajućih RNA i protein-kodirajućih gena

Budući da većina dugih nekodirajućih RNA ima ulogu u regulaciji ekspresije susjednih protein-kodirajućih gena, sve pronađene lncRNA-kodirajuće gene podijelio sam prema njihovu odnosu s protein-kodirajućim genima, pri čemu je navedena podjela uključivala intergenske, intronske i preklapajuće lncRNA-kodirajuće gene (Poglavlje 1.1.2). Pritom, ako je pojedini gen u svom intronu sadržavao protein-kodirajući gen te istovremeno bio dio introna drugog protein-kodirajućeg gena, klasificirao sam ga kao intronski gen. Također, ukoliko je pojedini lncRNA-kodirajući gen sadržavao veći dio protein-kodirajućeg gena, pri čemu se posljednji intron protein-kodirajućeg gena preklapao s egzonom lncRNA-kodirajućeg gena, dok je posljedni egzon protein-kodirajućeg gena bio izvan lncRNA-kodirajućeg gena, takav gen klasificirao sam kao preklapajući. Izračunao sam broj lncRNA-kodirajućih gena koje pripadaju

svakoj klasi, ukupni broj baza svake od navedenih klasa, kao i prosječni udio intronskih baza svake navedene klase. Također, Kruskal-Wallisovim testom usporedio sam raspodjele duljina gena, transkripata, egzona i introna svih navedenih klasa lncRNA. Isto tako, izračunao sam udaljenosti između svakog intergenskog lncRNA-kodirajućeg gena i najbližeg protein-kodirajućeg gena te sam dobivenu raspodjelu grafički prikazao. Na kraju, dodijelio sam izraze GO (engl. *Gene Ontology*) svakom protein-kodirajućem genu udaljenom manje od 1 kb (kilobaza) od najbližeg intergenskog lncRNA-kodirajućeg gena, kao i protein-kodirajućim genima koji se nalaze u preklapajućim lncRNA-kodirajućim genima te protein-kodirajućim genima koji u svojim intronima sadrže intronske lncRNA-kodirajuće gene. Izrazi GO predstavljaju hijerarhijski vokabular koji se koristi u svrhu funkcionalne anotacije gena. Za određivanje izraza GO koristio sam program Blast2GO verzije 1.3.11 (Götz i sur. 2008), koji izraze GO dodjeljuje na osnovu pronađenih homologa ulaznih protein-kodirajućih gena, pri čemu za pretragu homologa koristi program BLAST i neredundantnu proteinsku bazu. Dobivene izraze GO usporedio sam s izrazima GO dodijeljene svim proteinima ogulinske špiljske spužvice, pri čemu sam obogaćenje pojedinih izraza provjerio Fischerovim testom. Sve analize odnosa lncRNA-kodirajućih gena i protein-kodirajućih gena proveo sam u programskom jeziku R, koristeći pakete data.table, IRanges, GenomicRanges, ggplot2, kao i palete boja za grafičke prikaze iz paketa harrypotter verzije 2.1.1. (Rico 2020).

2.7.4. Analiza odnosa dugih nekodirajućih RNA i transpozona

Za uspoređivanje odnosa transpozona s lncRNA-kodirajućim genima i protein-kodirajućim genima koristio sam anotaciju ponavljajućih sljedova u genomu ogulinske špiljske spužvice. Pritom, izračunao sam broj transpozona koji se cijelom svojom duljinom preklapaju s klasama lncRNA-kodirajućih gena i protein-kodirajućim genima, kao i s 5' regijama navedenih gena, pri čemu je duljina navedene regije bila postavljena na 1 kb. U slučaju postojanja gena čiji se početak nalazi na koordinati genomskog slijeda manjoj od 1000, 5' regija određena je kao raspon između početka genomskog slijeda i početka gena. Također, usporedio sam broj baza pojedinih skupina transpozona koji se preklapaju s 5' regijama, egzonima ili intronima svih klasa lncRNA-kodirajućih gena i protein-kodirajućih gena, pri čemu su klase lncRNA-kodirajućih gena definirane obzirom na njihov odnos s protein-kodirajućim genima. Pritom, dobiveni broj normalizirao sam obzirom na ukupnu duljinu uspoređenih genomskih elemenata. Analizu preklapajućih baza napravio sam i za svaki razred transpozona, pri čemu sam, osim normalizacije po zbroju baza analiziranih genomskih elemenata, primijenio i normalizaciju obzirom na ukupnu duljinu određenog razreda ponavljanja u genomu. Pritom, analizirane skupine transpozona uključivali su transpozone nepoznatog razreda, kao i transpozone razreda LTR, LINE i DNA. Ako se pojedini transpozon preklapao s više od jednim elementom gena, kod obje analize sam u obzir uzeo samo preklapanje s elementom za koji je duljina preklapanja najveća Za ovu analizu koristio sam programski jezik R te pakete data.table, IRanges, GenomicRanges, DescTools, ggplot2 i gameofthrones verzije 1.0.3 (Rico 2020).

2.7.5. Analiza ekspresije dugih nekodirajućih RNA

U ovom dijelu istraživanja analizirao sam relativnu razinu ekspresije dugih nekodirajućih RNA unutar uzorka primorfa izoliranih prvi i deseti dan njihova rasta. U tu svrhu, mapirao sam obrađene sljedove iz knjižnica RNA1 i RNA10 na genom ogulinske špiljske spužvice koristeći program BBmap, dio paketa BBTools verzije 36.20).

Program BBmap još je jedan od programa koji pri mapiranju sljedova na referentni genom koristi strategiju rastavljanja referentnog slijeda, kao i ulaznih sljedova na klice. Nakon mapiranja na genom, klice ulaznih sljedova produljuju se algoritmom lokalnog poravnanja. Ipak, jedna od specifičnosti ovog programa je omogućavanje mapiranja preko relativno velikih insercija u genomu, koje u većini slučajeva predstavljaju introne, što omogućuje mapiranje sljedova koji obuhvaćaju mjesta prekrajanja (Bushnell 2014).

U svrhu mapiranja sljedova knjižnica RNA1 i RNA10 na genom ogulinske špiljske spužvice, pokrenuo sam program BBmap s parametrima „ambiguous=random secondary=f maxindel=100000“. Navedeni parametri omogućavaju maksimalnu veličinu introna od 100 kb te ne dozvoljavaju mapiranja sljedova koji se mapiraju na više mjesta. Ukoliko se slijed mapira podjednako dobro na više mjesta, mjesto mapiranja bira se nasumično. Nakon mapiranja, dobivene datoteke pretvorio sam u format .bam te sortirao programom Sambamba.

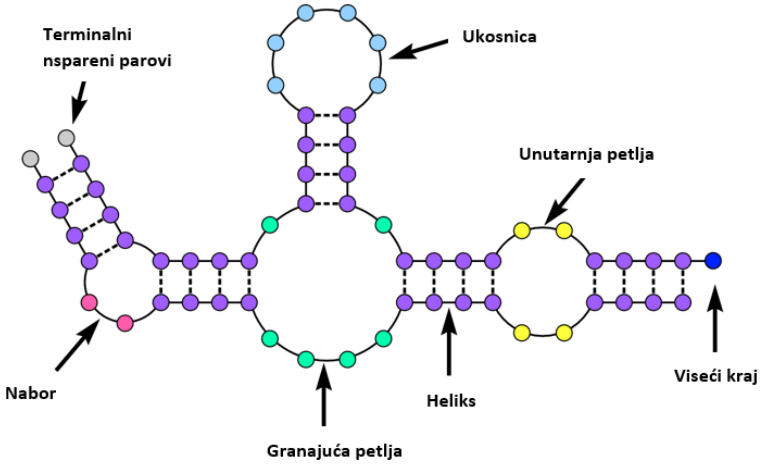
Za brojanje sljedova mapiranih na lncRNA-kodirajuće i protein-kodirajuće gene koristio sam program FeatureCounts verzije 2.0.1. (Liao i sur. 2014). Zadani parametri ovog programa broje svaki ulazni slijed koji se preklapa s određenim genom u jednom ili više parova baza. Također, sljedovi koji se mapiraju na više mjesta se zanemaruju, a od uparenih sljedova broji se samo jedan. Pripremio sam tablicu položaja lncRNA-kodirajućih i protein-kodirajućih gena u formatu .saf te sam pokrenuo program FeatureCounts s opcijom prilagođenom brojanju uparenih sljedova. Zatim, iz broja sljedova mapiranih na određeni gen izračunao sam vrijednost FPKM (engl. *Fragments per kilobase milion*), koja omogućava relativno pouzdanu usporedbu ekspresije gena unutar i između uzoraka, a definirana je kao broj sekvenciranih fragmenata mapiran na kilobazu gena za svaki milijun sekvenciranih sljedova. Drugim riječima, jedinica FPKM služi normalizaciji obzirom na duljinu gena i broj sljedova u knjižnici. Iz daljnjeg tijeka analize izbacio sam lncRNA-kodirajuće gene s FPKM vrijednošću 0. Za pripremu anotacijskih datoteka gena i izračun vrijednosti FPKM koristio sam programski jezik R i paket data.table.

U sljedećem dijelu analize usporedio sam ekspresiju lncRNA-kodirajućih gena prvog i desetog dana rasta primorfa. Također, Tukey-Krameovim testom usporedio sam razine ekspresija klasa lncRNA-kodirajućih gena koristeći logaritamski transformirane vrijednosti FPKM. Isto tako, analizirao sam ekspresije lncRNA-kodirajućih gena s insercijama transpozona u njihovim 5' regijama, egzonima ili intronima. Ekspresiju navedenih gena usporedio sam s ekspresijom lncRNA-kodirajućih gena bez insercija transpozona. Pritom, svakom genu sam pridružio inserciju transpozona koja je bila najduža. Na kraju, Tukey-Kramerovim testom usporedio sam ekspresiju proteina koji u intronu sadrže intronski lncRNA-kodirajući gen, proteina koji se nalaze u intronu preklapajućeg lncRNA-kodirajućeg gena, proteina koji su udaljeni najbližeg od lncRNA-kodirajućieg gena 1 kb ili manje i svih ostalih proteina. Sve navedene usporedbe

popratio sam odgovarajućim grafičkim prikazima. Za ovu analizu koristio sam programski jezik R i pakete data.table, ggplot2 te gameofthrones.

2.7.6. Analiza sekundarnih struktura dugih nekodirajućih RNA

Budući da je struktura dugih nekodirajućih RNA neraskidivo povezana s njihovom funkcijom, u sljedećem dijelu istraživanja odredio sam sekundarne strukture najdužih izoformi s najvećim brojem egzona pronađenih lncRNA-kodirajućih gena pomoću alata RNAfold (Lorenz i sur. 2011). Sekundarna struktura RNA opisana je parovima povezanih nukleotida molekule RNA, pri čemu nukleotidi povezani Watson-Crickovim pravilima tvore helikse, a nespareni nukleotidi tvore različite vrste petlji (Slika 8). Također, česta su i netipična sparivanja nukleotida, poput sparivanja guanina i uracila ili guaninskih tetrapleksa.



Slika 8. Shema motiva koji mogu činiti sekundarnu strukuturu RNA. Nukleotidi povezani vodikovim vezama tvore helikse, dok nespareni nukleotidi tvore petlje. Unutarnja petlja nastaje zbog postojanja nesparenih nukleotida s obje strane heliksa. Granajuća petlja nastaje zbog postojanja nesparenih nukleotida koji spavaju tri ili više različitih heliksa. Ukosnica nastaje zbog postojanja triju ili više nesparenih nukleotida. Nabor nastaje kao posljedica postojanja nesparenih nukletida samo s jedne strane RNA. Osim petlje i heliksa, za strukturu RNA karakteristični su i viseći krajevi, koji predstavljaju nesparen nukleotid na kraju lanca. Također, česti su i terminalni nespareni parovi nukleotida. Preuzeto i prilagođeno s https://rna.urmc.rochester.edu:81/mathews-lab/bootcamp/wikis/RNA-Secondary-Structure.

Programi za predviđanje sekundarnih struktura obično se temelje na algoritmima dinamičkog programiranja, koji implicitno pretražuju prostor svih mogućih sekundarnih struktura određene molekule RNA s ciljem pronalaska strukture s najmanjom slobodnom energijom, pri čemu nije potrebno eksplicitno generiranje svih struktura. Promjena slobodne energije procjenjuje se obzirom na energetski povoljno formiranje heliksa sparivanjem nukleotida i energetski nepovoljno formiranje različitih vrsta petlji. Takve procjene temelje su na eksperimentalno izvedenim promjenama slobodne energije strukture RNA pri postizanju različitih strukturnih motiva. Program RNAfold koristi modifikaciju navedenog algoritma, a pri određivanju sekundarne strukture, osim slobodne energije molekule, računa i parametre poput razine sličnosti između strukture s najmanjom slobodnom energijom i centroida, pri čemu centroid predstavlja strukturu s prosječno najvećom sličnošću svim ostalim mogućim

strukturama (Lorenz i sur. 2011). Isto tako, postoje pomoćni programi za utvrđivanje razine stabilnosti dobivene sekundarne strukture. Jedan od njih je program Randfold (Bonnet i sur. 2004), koji permutacijskim testom procjenjuje stabilnost ulazne strukture. Princip rada ovog programa zasniva se na permutiranju nukleotida ulazne RNA pri čemu nastaju molekule jednake duljine s drugačijim rasporedom nukleotida. Na temelju izračunatih slobodnih energija takvih molekula računa se p-vrijednost, odnosno vjerojatnost dobivanja molekule jednako niske slobodne energije kao i ulazna molekuka RNA uzorkovanjem nasumičnih molekula RNA jednake duljine (Bonnet i sur. 2004).

Iskoristio sam program RNAfold verzije 2.4.14. za određivanje sekundarnih struktura svih najduljih transkripata pronađenih lncRNA-kodirajućih gena i njihovih reverznih komplemenata. Također, programom Randfold verzije 1.0. procijenio sam stabilnost svake od navedenih molekula lncRNA pri čemu sam koristio 100 permutacija. Grafički sam prikazao distribuciju dobivenih p-vrijednosti, pri čemu sam za svaki par molekule lncRNA i njezina reverznog komplementa odabrao onu molekulu s manjom p-vrijednosti. Navedene p-vrijednosti iskoristio sam za usporedbu stabilnosti klasa molekula lncRNA, pri čemu sam Tukey-Kramerovim testom međusobno usporedio logaritamski transformirane p-vrijednosti svake klase. Usporedbu sam napravio koristeći programski jezik R i pakete data.table, ggplot2 te harrypotter.

2.7.7. Analiza očuvanosti dugih nekodirajućih RNA unutar koljena spužvi

Radi pronalaska potencijalnih homologa pronađenih molekula lncRNA unutar koljena Porifera koristio sam zadane parametre programa BLASTn, pri čemu su ulazni sljedovi bili najdulje izoforme s najviše egzona svakog lncRNA-kodirajućeg gena, dok je pretraživana baza uključivala 18 transkriptoma i 6 genoma spužvi navedenih u Prilogu 1. Prije pokretanja programa BLAST, ponavljajuće sljedove u pretraživanoj bazi zamijenio sam slovom N, što odgovara postupku tzv. tvrdog maskiranja sljedova (engl. *hard masking*) radi sprječavanja nespecifičnih pogodaka uzrokovanih ponavljajućim regijama. Navedeno maskiranje napravio sam pomoću programa dustmasker verzije 1.0.0. (Morgulis i sur. 2016) koristeći pažljivo birane parametre „-window 64 -level 18“.

Rezultate dobivene pretragom programa BLASTn filtrirao sam, uklanjajući sva poravnanja s očekivanom vrijednošću većom od 10^{-15}. Potom, za svaki ulazni transkript izračunao sam ukupnu duljinu poravnanja s pogođenim transkriptom, kao i omjer duljine pogotka i ukupne duljine ulaznog transkripta. Također, grafički sam prikazao maksimalni omjer duljine pogotka i ukupne duljine ulaznog transkripta za sve pogođene vrste od strane svakog ulaznog transkripta. Filtrirane rezultate pretrage programa BLASTn iskoristio sam i za određivanje kandidata za pronalazak očuvanih strukturnih motiva molekula lncRNA višestrukim poravnanjem više od dvaju sljedova. Odgovarajuće kandidate definirao sam kao skup ulaznog transkripta i svih pogođenih transkripata određenog ulaznog transkripta u ispravnoj orijentaciji s ukupnom duljinom pogotka većom od 150 pb. Pritom, u obzir sam uzeo sam transkripte s pogotcima u dvije ili više vrsta. Ovu analizu napravio sam koristeći programski jezik R, kao i pakete data.table, ggplot2, IRanges, GenomicRanges te harrypotter.

Dobivene skupove transkripata višestruko sam poravnao programom ClustalW verzije 2.1. (Thompson i sur. 1994), pri čemu sam dobivena poravnanja iskoristio za pronalazak očuvanih motiva molekula lncRNA pomoću programa RNAz verzije 2.0. (Gruber i sur. 2010). Ovaj program temelji se na predviđanju lokalnih konsenzusnih struktura u višestruko poravnatim sljedovima nekodirajućih RNA pri čemu u obzir uzima termodinamičku stabilnost takvih struktura, kao i njihov stupanj očuvanosti. Termodinamička stabilnost procijenjena je z-vrijednošću, koja predstavlja razliku između slobodne energije dobivene strukture i slobodnih energija nasumičnih struktura jednake duljine. S druge strane, procjena očuvanosti strukture temelji se na modelu koji pri računu slobodne energije konsenzusne strukture boduje supstitucije nuklelotida u skladu s njihovim utjecajem na promjenu sekundarne strukture RNA. Time je omogućen račun indeksa očuvanja strukture (engl. *structure conservation index*, SCI), koji predstavlja omjer slobodne energije konsenzusne strukture i prosječne slobodne energije individualnih struktura u poravnanju. Tako dobivene vrijednosti z i SCI kombiniraju se radi klasificiranja poravnanja kao strukturne RNA, pri čemu se za klasifikaciju koristi model SVM (engl. *Support Vector Machines*) istreniran na setu poznatih nekodirajućih RNA. Ovaj model za svaki pronađeni lokalni motiv u poravnanju računa vrijednost vrijednost P, pri čemu se motiv svrstava u klasu strukturnih RNA u slučaju vrijednosti P veće od 0.5 (Gruber i sur. 2010).

2.7.8. Pronalazak potencijalnih homologa pronađenih dugih nekodirajućih RNA izvan koljena spužvi

U posljednjem dijelu istraživanja analizirao sam očuvanost pronađenih dugih nekodirajućih RNA pronalaskom njihovih potencijalnih homologa izvan koljena spužvi. Za pretragu potencijalnih homologa koristio sam program nhmmer (Wheeler i Eddy 2013), varijantu programa HMMER prilagođenu sljedovima nukleotida. Budući da su u ovom slučaju ulazni sljedovi u obliku samostalnih sljedova, a ne poravnanja određene obitelji molekula lncRNA, program nhmmer nije mogao iskoristiti informacije koji bi bile sadržane u takvom poravnanju, zbog čega je njegova osjetljivost smanjena. Ipak, smatra se da je za pronalazak slabo očuvanih molekula lncRNA osjetljivost programa nhmmer malo bolja u odnosu na alate poput programa BLAST (Freyhult i sur. 2007). S druge strane, program nhmmer pokazuje desetak puta sporiju pretragu od programa BLAST, zbog čega ga nisam koristio u analizi očuvanosti molekula lncRNA unutar koljena spužvi.

Za pretragu potencijalnih homologa pronađenih molekula lncRNA izvan koljena spužvi kao ulazne sljedove koristio sam najduže izoforme s najviše egzona pronađenih lncRNA-kodirajućih gena, dok sam kao pretraženu bazu koristio sve poznate RNA prikupljene u bazi RNACentral, pri čemu sam radi smanjena komputacijskog vremena iz te baze uklonio sve molekule rRNA. Također, prije pokretanja pretrage uklonio sam ponavljajuće regije u sljedovima baze RNACentral tzv. tvrdim maskiranjem programom dustmasker (parametri: „window 64 -level 18“). Program nhmmer verzije 3.1.1. pokrenuo sam sa zadanim parametrima.

Rezultate pretrage programom nhmmer filtrirao sam, uklanjajući sva poravnanja s očekivanom vrijednošću većom od 10^{-15}. Pritom, koristio sam programski jezik R i paket

data.table. Za taksonomsku klasifikaciju pogodaka koristio sam paket taxsize verzije 0.9.95. (Chamberlain i Szocs 2013).

Još jedan način provjere očuvanosti pronađenih molekula lncRNA izvan koljena Porifera bila je pretraga sličnosti između pronađenih transkripata lncRNA i obitelji RNA iz baze Rfam. Baza Rfam sadrži očuvane obitelji molekula RNA predstavljene višestrukim poravnanjima i modelom kovarijance (engl. *covariance model*, CM). Za pretragu sličnosti između pronađenih molekula lncRNA i baze Rfam koristio sam alat cmsearch programa Infernal verzije. 1.13 (Nawrocki i Eddy 2013). Ovaj program temelji se na pretrazi sličnosti pomoću modela CM, koji je vrlo sličan već opisanom modelu HMM korištenog od strane programa HMMER. Glavna razlika između spomenutih modela uključuje modeliranje međusobne ovisnosti nukleotida povezanih Watson-Crickovim vezama, što profilima CM daje veću osjetljivost pri pronalasku strukturnih homologa molekula RNA (Nawrocki i Eddy 2013).

Za pretragu sličnosti između pronađenih molekula lncRNA i očuvanih domena RNA u bazi Rfam koristio sam alat cmsearch. Ovaj alat pri pretrazi koristi dostupne informacije iz modela CM kojim su opisana višestruka poravnanja svih očuvanih molekula RNA prisutnih u bazi Rfam. Pri pokretanju alata cmsearch koristio sam parametre „–rfam –cut_ga –nohmmonly –clanin“. Filtrirao sam pogotke s očekivanom vrijednošću većom od 10^{-15}.

3. Rezultati

3.1. Obrada sljedova knjižnica RNA1 i RNA10

Istraživanje sam započeo obradom sljedova knjižnica RNA1 i RNA10, što je uključivalo uklanjanje adaptera, kao i filtriranje te skraćivanje sljedova po kriteriju kvalitete. Osnovni statistički podaci o obrađenim sljedovima prikazani su u Tablici 3.

Tablica 3. Statistički podaci obrađenih sljedova knjižnica RNA1 i RNA10 izoliranih iz primorfa ogulinske špiljske spužvice prvi i deseti dan njihova rasta.

Knjižnica	Prosječna ocjena kvalitete	Prosječni udio GC (%)	Broj sljedova	Minimalna duljina / pb	Prosječna duljina / pb	Ukupna duljina / Mb
RNA1	30.91	48.94	339193192	10	49.74	16871.70
RNA10	29.66	49.12	303798032	10	48.49	14730.84

Iz statističkih podataka obrađenih sljedova knjižnica RNA1 i RNA10 vidljiv je značajan porast prosječne kvalitete sljedova, kao i smanjenje broja te ukupne duljine sljedova u odnosu na neobrađene sljedove. Također, vidljiva je relativno mala promjena u udjelu GC i prosječnoj duljini sljedova. Kao i prije obrade, knjižica RNA1 dulja je od knjižnice RNA10. Isto tako, vidljiva je razlika u prosječnoj kvaliteti sljedova u korist knjižnice RNA1.

3.2. Sastavljanje transkriptoma

Obrađene sljedove knjižnica RNA1 i RNA10 spojio sam u skup sljedova koji sam iskoristio za sastavljanje transkriptoma programima rnaSPAdes i Trinity. Osnovni statistički parametri sklopljenih transkriptoma prikazani su u Tablici 4.

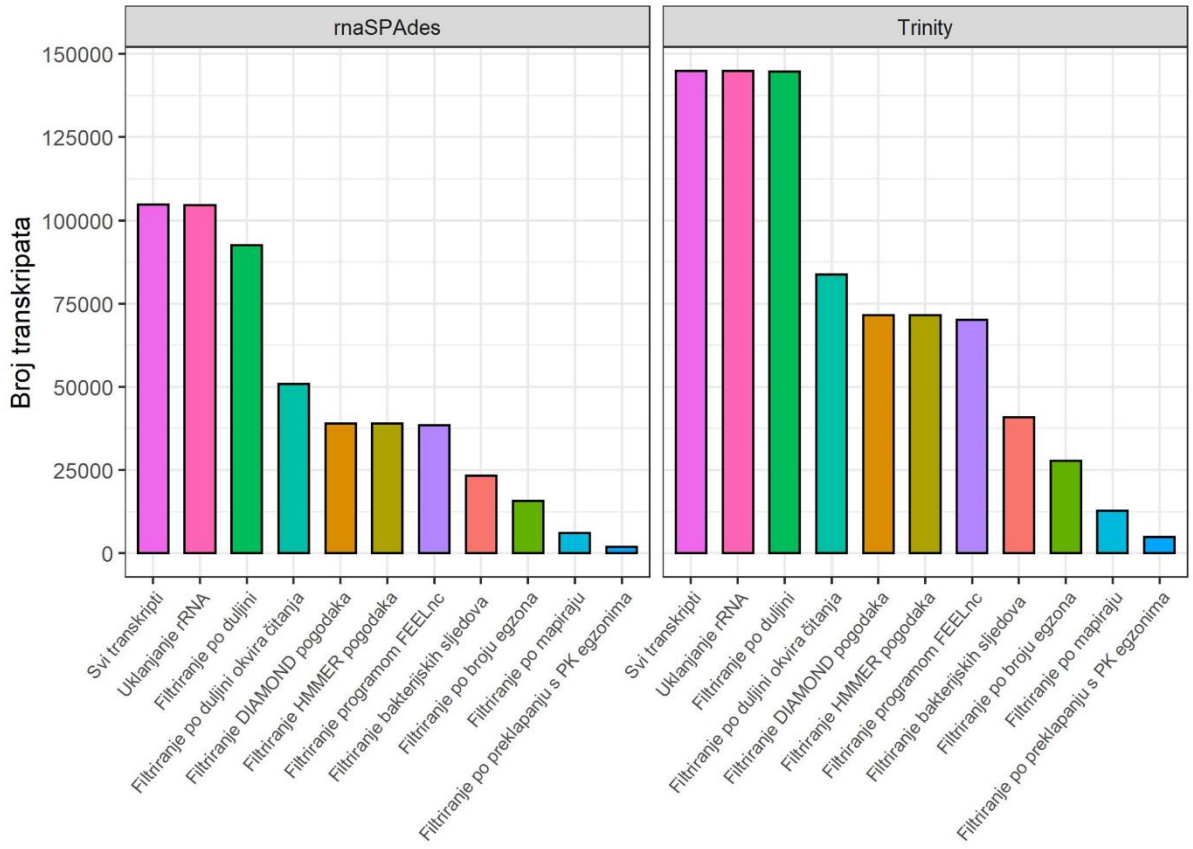
Tablica 4. Statistički parametri transkriptoma sklopljenih pomoću združenog skupa sljedova iz knjižnica RNA1 i RNA10 programima rnaSPAdes i Trinity. Udio mapiranih sljedova RNA izračunat je obzirom na ukupnu duljinu skupa združenih sljedova.

Transkriptom	Broj transkripata	Ukupna duljina / Mb	Prosječna duljina / pb	Minimalna duljina / pb	Maksimalna duljina / pb	N50 / pb	Udio mapiranih sljedova RNA (%)
rnaSPAdes	104671	159.65	1507.09	53	33198	2786	93.54
Trinity	144880	238.95	1649.26	181	144880	3255	92.69

Iz statističkih parametara sklopljenih transkriptoma vidljivo je da transkriptom složen programom Trinity ima značajno veći broj transkripata, koji su prosječno duži od transkripata iz transkriptoma složenim programom rnaSPAdes. Također, kod oba transkriptoma vidljiv je relativno velik udio mapiranih sljedova RNA pomoću kojih su ti transkriptomi složeni.

3.3. Pronalazak dugih nekodirajućih RNA

U sljedećem dijelu istraživanja iskoristio sam transkriptome ogulinske špiljske spužvice sklopljene programima rnaSPAdes i Trinity za pronalazak dugih nekodirajućih RNA. Taj postupak temeljio se na nizu filtriranja kojima je cilj bio pronaći najpouzdaniji mogući skup molekula lncRNA, a uključivao je filtriranje molekula rRNA, uklanjanje transkripata kraćih od 200 pb, uklanjanje transkripata s okvirom čitanja dužim od 150 pb, filtriranje transkripata s pronađenom sličnošću s bilo kojim poznatim proteinom ili proteinskom domenom, uklanjanje transkripata obzirom na njihov sastav programom FEELnc, filtriranje transkripata bakterijskog porijekla, uklanjanje transkripata koji se mapiraju na genom s manje od 95% svoje duljine, filtriranje transkripta s jednim egzonom te uklanjanje transkripata čiji se egzoni preklapaju s egzonima protein-kodirajućih gena. Rezultati navedenih filtriranja prikazani su na Slici 9.



Slika 9. Rezultati serije filtriranja transkripata ogulinske špiljske spužvice sklopljenih programima rnaSPAdes i Trinity u svrhu pronalaska dugih nekodirajućih RNA. Filtriranje DIAMOND pogodaka uključuje uklanjanje svih transkripata koji su pokazali značajnu razinu sličnosti s poznatim proteinima, dok filtriranje HMMER pogodaka označava uklanjanje svih transkripata sa sličnošću visoko očuvanim proteinskim domenama. Također, filtriranje programom FEELnc uključivalo je izdvajanje transkripata lncRNA obzirom na njihov sastav nukleotida. PK = protein-kodirajući geni.

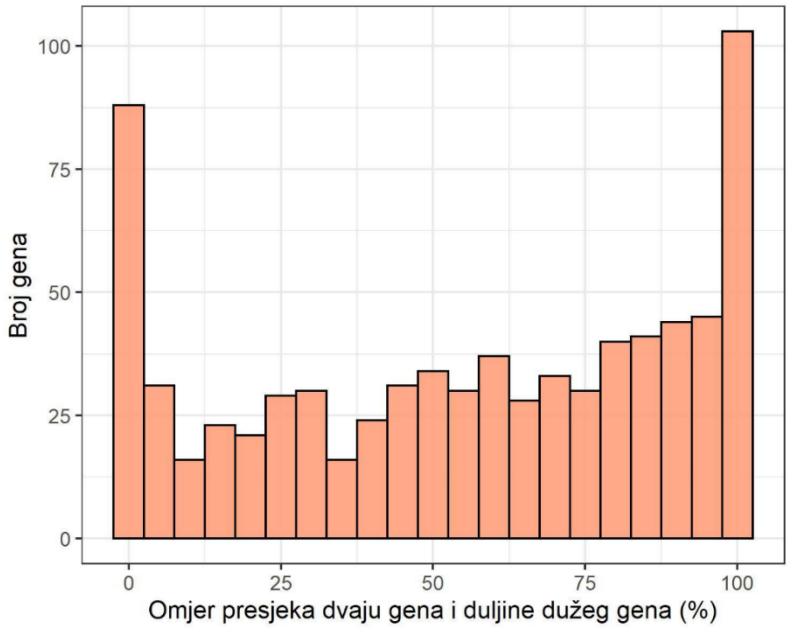
Na slici 9 vidljivo je da je trend smanjenja broja transkripata sličan kod oba transkriptoma, s jedinom bitnijom razlikom u filtriranju po duljini, pri čemu je broj filtriranih transkripata iz transkriptoma rnaSPAdes puno veći od broja filtriranih transkripata iz transkriptoma Trinity. Nadalje, može se uočiti da je ovim redoslijedom filtriranja najviše transkripata uklonjeno

filtracijom duljine okvire čitanja, filtracijom bakterijskih sljedova i filtracijom po postotku mapiranja. S druge strane, najmanje transkripata uklonjeno je filtracijom rRNA, filtracijom transkripata s ustanovljenom sličnošću s poznatim proteinima ili proteinskih domena, kao i filtracijom programom FEELnc. Na kraju navedene serije filtriranja ostalo je 1620 transkripata iz transkriptoma složenog programom rnaSPAdes i 3923 transkripata iz transkriptoma složenog programom Trinity.

3.4. Analiza pronađenih dugih nekodirajućih RNA

3.4.1. Određivanje konsenzusnog skupa dugih nekodirajućih RNA iz transkriptoma rnaSPAdes i Trinity

Nakon pronalaska dugih nekodirajućih RNA iz transkriptoma sklopljenih programima rnaSPAdes i Trinity, izdvojio sam njihov najveći mogući neredundantni skup. Za tu potrebu, iz oba transkriptoma odredio sam sve lncRNA-kodirajuće gene koji se preklapaju, pri čemu su koordinate gena definirane kao početak i kraj skupa preklapajućih transkripata, pri čemu su u navedeni skup uzeti transkripti klasificirani kao izoforme istog gena od strane programa rnaSPAdes i programa Trinity. Raspodjela omjera presjeka dvaju gena i duljine dužeg gena u svakom paru preklapanja prikazana je na Slici 10.



Slika 10. Omjeri presjeka dvaju lncRNA-kodirajućih gena i duljine dužeg gena u paru. Parovi gena predstavljaju preklapajuće lncRNA-kodirajuće gene dobivene iz transkriptoma sklopljenih programima rnaSPAdes i Trinity (N = 1032, raspon intervala = 5).

Konsenzusni skup lncRNA-kodirajućih gena koje sam pronašao u transkriptomima složenim programima rnaSPAdes i Trinity definirao sam kao skup jedinstvenih gena iz oba transkriptoma, dužih gena svakog preklapajućeg para s presjekom većim od 20 %, kao i dužim te kraćim genima svakog preklapajućeg para s presjekom manjim od 20% duljine dužeg gena u paru. S druge strane, konsenzus transkripata predstavlja sve transkripte koji se preklapaju s navedenim genima. U Tablici 5. prikazani su rezultati uzimanja takvog konsenzusnog skupa.

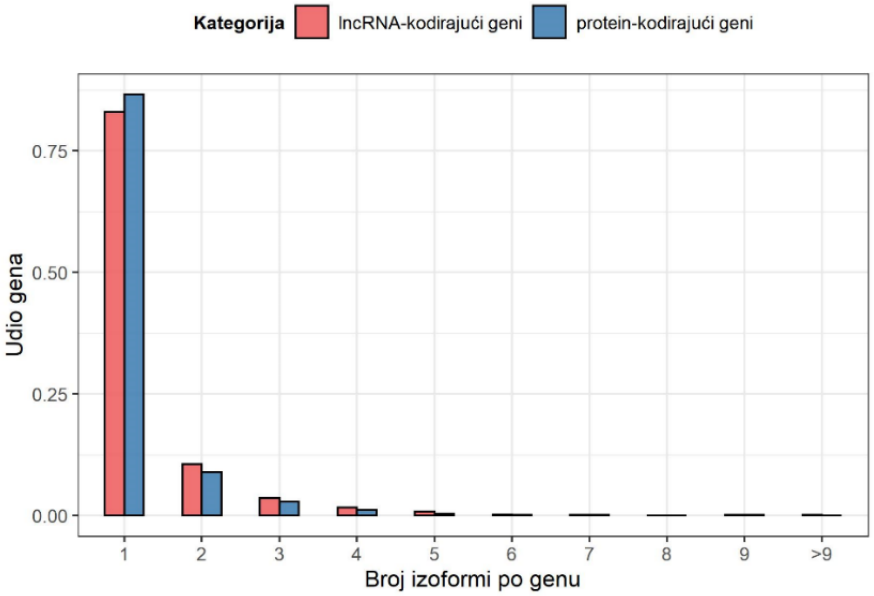
Tablica 5. Rezultati uzimanja konsenzusnog skupa pronađenih dugih nekodirajućih RNA iz transkriptoma sastavljenih programima rnaSPAdes i Trinity

Transkriptom	rnaSPAdes	Trinity
Ukupan broj lncRNA gena	1620	2747
Broj potpuno preklapajućih gena	39	39
Jedinstveni lncRNA geni	942	2100
Duži preklapajući geni parova s presjekom većim od 20% duljine dužeg gena	360	245
Duži preklapajući geni parova s presjekom manjim od 20% duljine dužeg gena	48	63
Kraći preklapajući geni parova s presjekom manjim od 20% duljine dužeg gena	74	78
Konsenzusni geni	1343	2408
Konsenzusni transkripti	1419	3494

Iz Tablice 5 vidljivo je da je u konsenzusnom skupu sadržana većina gena iz transkriptoma sklopljenog programom rnaSPAdes (82.90%), kao i većina gena iz transkriptoma sklopljenog programom Trinity (87.66%). U nastavku istraživanja analizirao sam konsenzusni skup transkripata lncRNA, kojih je ukupno 4913, odnosno 87.59% transkripata (rnaSPAdes) i 89.06% transkripata (Trinity) u odnosu na broj transkripata prije uzimanja konsenzusa.

3.4.2. Analiza osnovnih svojstava pronađenih dugih nekodirajućih RNA

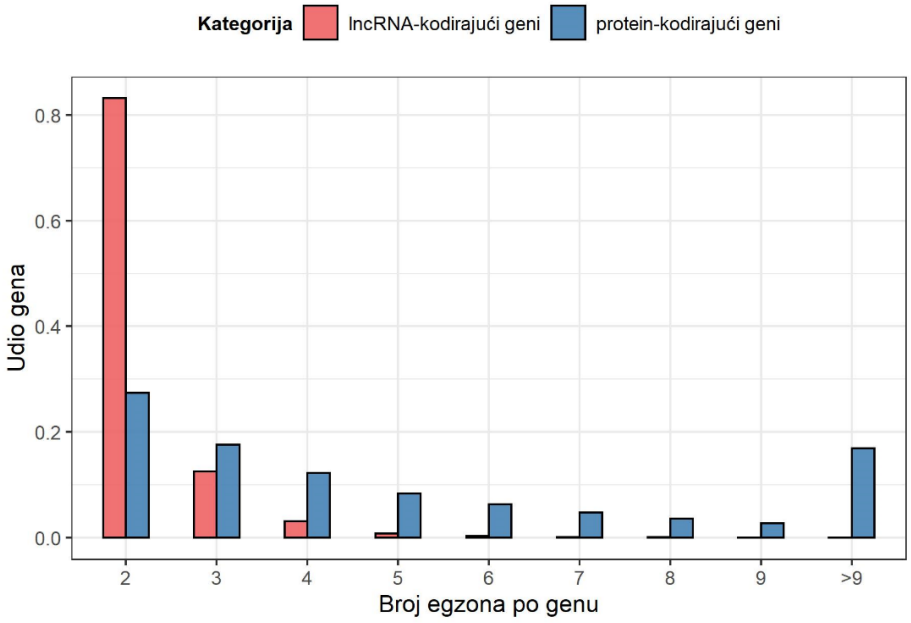
U nastavku istraživanja analizirao sam osnovna svojstva pronađenih dugih nekodirajućih RNA, naglašavajući usporedbu tih svojstava sa svojstvima protein-kodirajućih gena. Prvo od analiziranih svojstava bila je raspodjela broja izoformi po genu lncRNA-kodirajućih i protein-kodirajućih gena. Navedena usporedba prikazana je na Slici 11.



Slika 11. Raspodjela udjela lncRNA-kodirajućih i protein-kodirajućih gena ogulinske špiljske spužvice. s određenim brojem izoformi.

Na slici 11. vidljiv je relativno sličan odnos udjela lncRNA-kodirajućih i protein-kodirajućih gena s određenim brojem izoformi, pri čemu u oba slučaja većina gena sadrži jednu izoformu. Ipak, 16.4 % lncRNA-kodirajućih gena pokazuje dokaze alternativnog prekrajanja, dok taj udio kod protein-kodirajućih gena iznosi 8.7%. Najveći broj izoformi po genu iznosi 13 za lncRNA-kodirajuće gene i 10 za protein-kodirajuće gene.

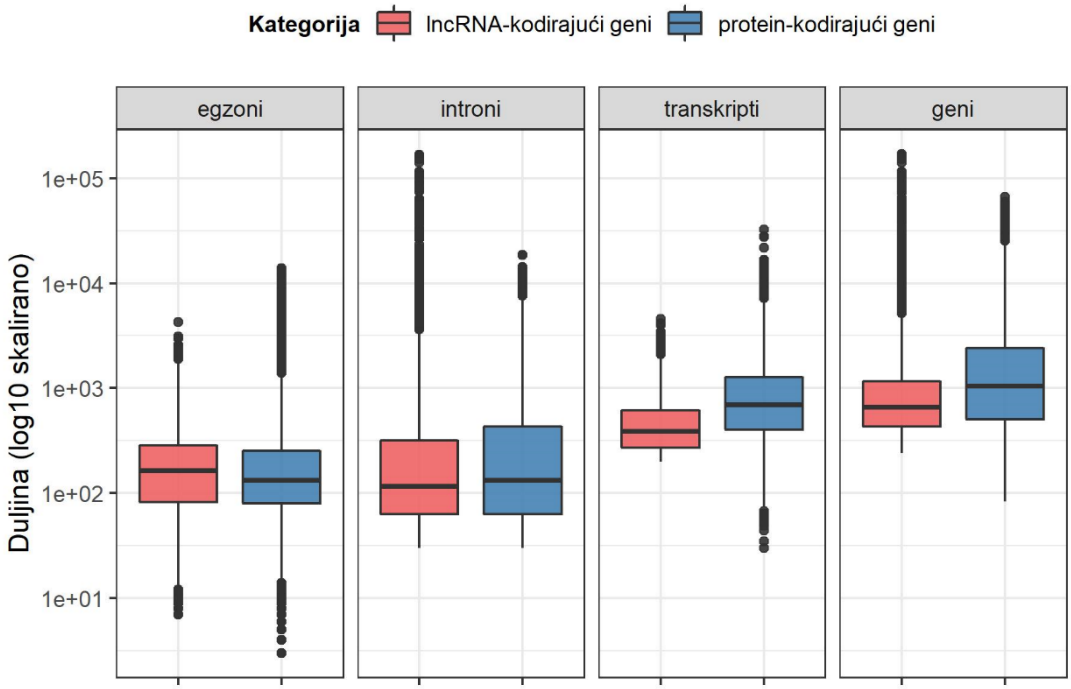
Potom, analizirao sam raspodjelu udjela lncRNA-kodirajućih i protein-kodirajućih gena s određenim brojem egzona. Egzone i introne gena definirao sam kao egzone i introne izoforme gena s najviše egzona, pri čemu sam u slučaju više takvih izoformi izabrao dulju. Dobivene raspodjele prikazane su na Slici 12.



Slika 12. Raspodjela udjela lncRNA-kodirajućih i protein-kodirajućih gena ogulinske špiljske spužvice s određenim brojem egzona. Geni koji sadrže jedan egzon nisu razmatrani.

Iz raspodjele gena s određenim brojem egzona vidljivo je da većina lncRNA-kodirajućih gena sadrži 2 egzona, dok je takva raspodjela kod protein-kodirajućih gena puno ravnomjernija. Najveći broj egzona po lncRNA-kodirajućem genu iznosi 9, dok najveći broj egzona po protein-kodirajućem genu iznosi 110.

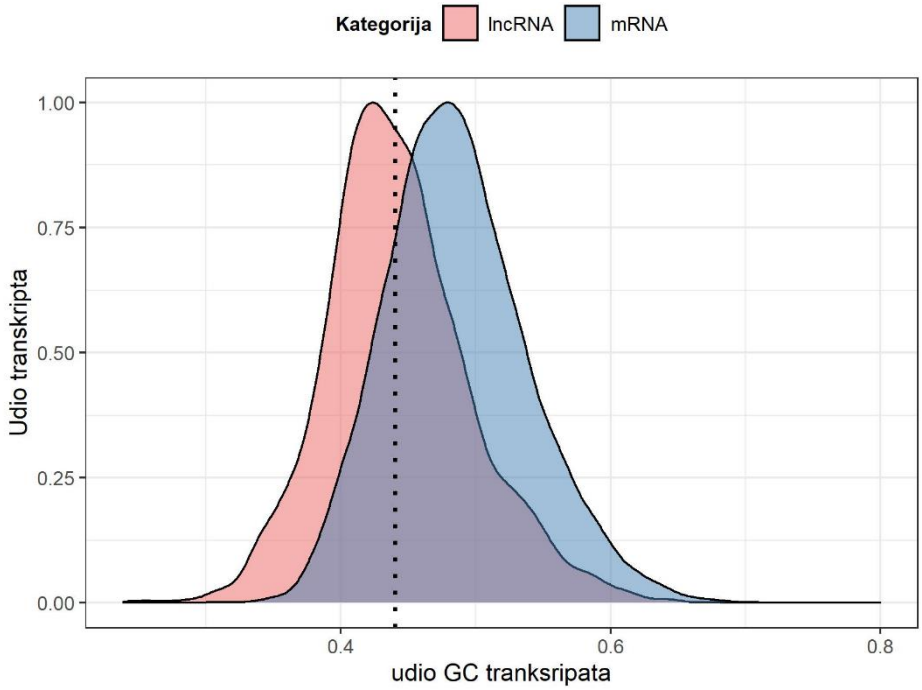
Egzone i introne lncRNA-kodirajućih i protein-kodirajućih gena usporedio sam i po njihovoj duljini. Na Slici 13 prikazana je raspodjela duljina gena, transkripata, egzona i introna lncRNA-kodirajućih i protein-kodirajućih gena, pri čemu su razmatrani samo najduži transkripti s najvećim brojem egzona po genu.



Slika 13. Raspodjela duljina gena, transkripata, introna i egzona lncRNA-kodirajućih i protein-kodirajućih gena ogulinske špiljske spužvice. Horizontalna crna linija predstavlja medijan raspodjele, dok se vertlkalne crne linije protežu od -1.5*IR do 1.5*IR, gdje IR označava interkvartilni raspon raspodjele. Crnim točkama prikazane su ekstremne vrijednosti raspodjele. Odnosi duljina na ovom prikazu su logaritamski skalirani.

Iz raspodjela duljina egzona lncRNA-kodirajućiih i protein-kodirajućih gena vidljivo je da egzoni lncRNA-kodirajućih gena imaju veću srednju vrijednost duljina od egzona protein-kodirajućih gena (Wilcoxonov test, $p < 2.2*10^{-16}$). S druge strane, utvrdio sam veću prosječnu vrijednost duljine introna protein-kodirajućih gena od prosječne vrijednosti duljine introna lncRNA-kodirajućih gena (Wilcoxonov test, $p = 6.893*10^{-12}$). Nadalje, utvrdio sam značajnu razliku između raspodjela duljina transkripata lncRNA-kodirajućih i protein-kodirajućih gena (Wilcoxonov test, $p < 2.2*10^{-16}$), pri čemu je srednja vrijednost duljine transkripata značajno manja kod lncRNA-kodirajućih gena. Sličan trend uočio sam i kod raspodjele duljina lncRNA-kodirajućih gena te protein-kodirajućih gena (Wilxoconov test, $p < 2.2*10^{-16}$), pri čemu je srednja vrijednost duljina lncRNA-kodirajućih gena značajno manja od srednje vrijednosti duljine protein-kodirajućih gena.

U sljedećem dijelu analize karakteristika molekula lncRNA usporedio sam udjele GC molekula lncRNA i molekula mRNA. Pritom sam razmatrao samo najdulje izoforme obje skupine gena s najveći brojem egzona. Raspodjele udjela GC transkripata lncRNA i mRNA prikazane su na Slici 14.



Slika 14. Raspodjele udjela GC transkripata lncRNA i mRNA ogulinske špiljske spužvice. Isprekidanom linijom označena je prosječna vrijednost udjela GC genoma ogulinske špiljske spužvice.

Iz raspodjela udjela GC uočljiva je razlika između transkripata lncRNA i mRNA, pri čemu je prosječna vrijednost udjela GC transkripata lncRNA značajno manja od prosječne vrijednosti udjela GC transkripata mRNA (t-test, $p < 2.2*10^{-16}$, interval pouzdanosti razlike aritmetičkih sredina: <-0.0452, -0.0416>). Također, prosječni udio GC transkripata lncRNA puno je sličniji prosječnom udjelu GC genoma od prosječnog udjela GC transkripata mRNA.

Na kraju analize općenitih svojstava pronađenih molekula lncRNA, usporedio sam udio različitih mjesta prekrajanja lncRNA-kodirajućih i protein-kodirajućih gena, pri čemu sam uzimao u obzir introne najdužih izoformi s najvećim brojem egzona. Većina introna protein-kodirajućih gena pokazuje kanonska mjesta izrezivanja (99%), dok je udio takvih introna lncRNA-kodirajućih gena niži (70.66%).

3.4.3. Analiza odnosa pronađenih dugih nekodirajućih RNA i protein-kodirajućih gena

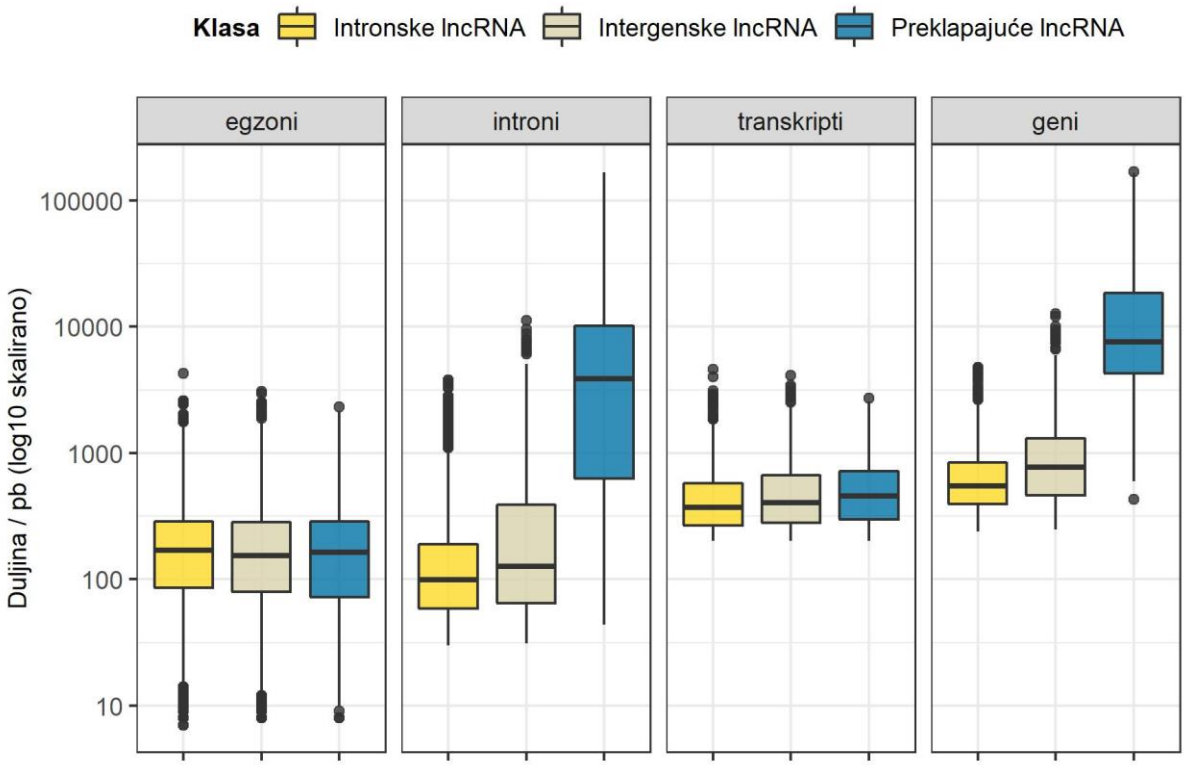
Budući da velik broj dugih nekodirajućih RNA djeluje na ekspresiju susjednih protein-kodirajućih gena, u ovom dijelu istraživanja analizirao sam odnos lncRNA-kodirajućih gena i protein-kodirajućih gena ogulinske špiljske spužvice. U Tablici 6 dani su osnovni podaci o klasama pronađenih lncRNA-kodirajućih gena obzirom na odnos s protein-kodirajućim genima.

Tablica 6. Osnovni podaci o klasama pronađenih lncRNA-kodirajućih gena ogulinske špiljske spužvice obzirom na odnos s protein-kodirajućim genima.

Klasa	**Broj**	**Ukupna duljina / Mb**	**Udio introna u genu (%)**
Intronske lncRNA	1282	1.58	30.77
Intergenske lncRNA	2215	1.44	50.49
Preklapajuće lncRNA	254	4.94	96.91

Iz Tablice 6 vidljivo je da su među pronađenim lncRNA-kodirajajućim genima najčešće intergenske lncRNA, dok najmanji broj pronađenih lncRNA-kodirajućih gena pripada klasi preklapajućih lncRNA

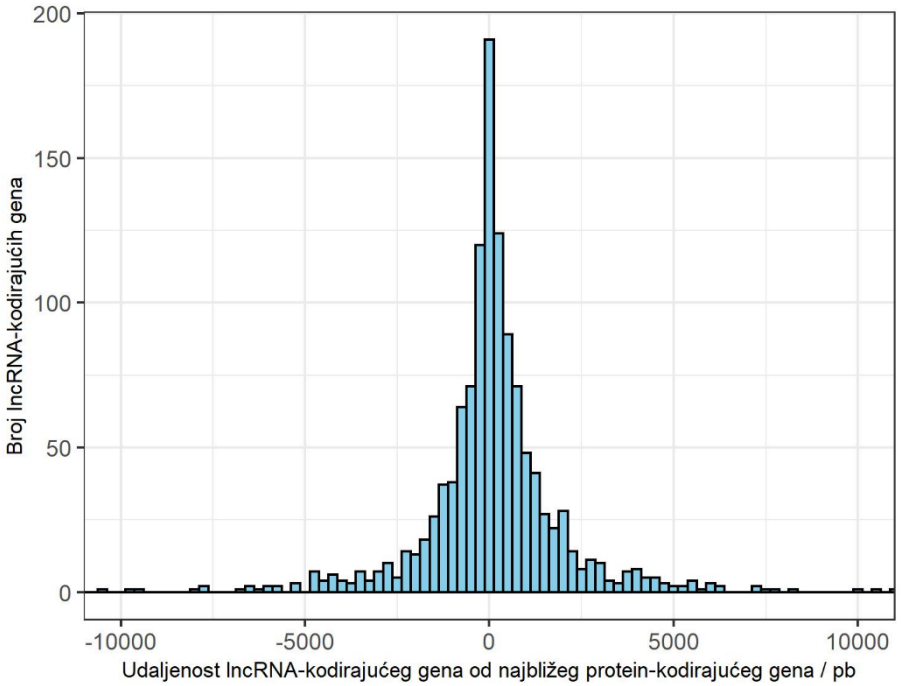
Na Slici 15. prikazane su raspodjele duljina gena, transkripata, egzona i introna klasa lncRNA-kodirajućih gena, pri čemu su klase definirane obzirom na odnos s protein-kodirajućim genima.



Slika 15. Raspodjela duljina gena, transkripata, introna i egzona klasa dugih nekodirajućih RNA ogulinske špiljske spužvice, pri čemu su klase definirane obzirom na odnos s protein-kodirajućim genima. Intronske lncRNA predstavljaju lncRNA-kodirajuće gene koji se nalaze u intronu protein-kodirajućeg gena, dok preklapajuće lncRNA predstavljaju lncRNA-kodirajuće gene koji u svom intronu sadrže protein-kodirajući gen. S druge strane, intergenske lncRNA označuju lncRNA-kodirajuće gene koji se ne preklapaju s protein-kodirajućim genima. Horizontalna crna linija predstavlja medijan raspodjele, dok se vertlkalne crne linije protežu od -1.5*IR do 1.5*IR, gdje IR označava interkvartilni raspon raspodjele. Crnim točkama prikazane su ekstremne vrijednosti raspodjele. Odnosi duljina na ovom prikazu su logaritamski skalirani.

Iz raspodjele duljina klasa lncRNA-kodirajućih gena definiranih obzirom na odnos s protein-kodirajućim genima vidljivo je da ne postoji značajna razlika u raspodjeli duljina egzona navedenih klasa (Kruskal-Wallisov test, p=0.0515). S druge strane, utvrđeno je postojanje razlike u raspodjeli duljina gena (Kruskal-Wallisov test, $p < 2.2*10\text{^}^{-16}$), transkripata (Kruskal-Wallisov test, $p=1.537*10^{-6}$) i introna (Kruskal-Wallisov test, $p < 2.2*10^{-16}$).Važno je naglasiti duljinu introna i gena preklapajućih lncRNA, koje su značajno veće od duljina introna, odnosno gena drugih klasa lncRNA.

U nastavku analize klasa pronađenih molekula lncRNA izračunao sam udaljenosti intergenskih lncRNA-kodirajućih gena od najbližeg protein-kodirajućeg gena. Raspodjela navedenih udaljenosti prikazana je na Slici 16.



Slika 16. Raspodjela udaljenosti intergenskih lncRNA-kodirajućih gena od najbližeg protein-kodirajućeg gena. Negativne udaljenosti dodijeljene su lncRNA-kodirajućim genima koji se nalaze 5' od protein-kodirajućeg gena, dok su pozitivne udaljenosti dodijeljene lncRNA-kodirajućim genima koji se nalaze 3' od protein-kodirajućeg gena. LncRNA-kodirajući geni koji su od najbližeg protein-kodirajućih gena udaljeniji 10 kb, kao i lncRNA-kodirajući geni koji se nalaze na neprekinutim sljedovima koji ne sadrže protein-kodirajućih gena su zanemareni. (N = 1225, raspon intervala = 250).

Radi utvrđivanja potencijalne uloge lncRNA-kodirajućih gena, napravio sam analizu izraza GO skupine gena koja je uključivala protein-kodirajuće gene koji se preklapaju s intronskim ili preklapajućim lncRNA-kodirajućim genima, kao i protein-kodirajuće gena koji su udaljeni manje od 1 kb od najbližeg lncRNA-kodirajućeg gena. Pritom, broj protein-kodirajućih gena s preklapanjem s intronskim lncRNA-kodirajućim genom iznosio je 2234, dok je broj protein-kodirajućih gena u preklapajućim lncRNA-kodirajućim genima iznosio 1026. Isto tako, broj protein-kodirajućih gena udaljenih manje od 1 kb od najbližeg lncRNA-kodirajućeg gena iznosio je 719, dok je broj svih ostalih protein-kodirajućih gena iznosio 29014. Fischerovim testom utvrdio sam značajno veći udio brojnih izraza GO u navedenom skupu protein-kodirajućih gena u odnosu na sve ostale protein-kodirajuće gene ogulinske špiljske spužvice. Neki od obogaćenih izraza GO uključuju vezanje transkcripcijskih faktora (FDR = $6.3*10^{-11}$), vezanje histona (FDR = $1.6*10^{-3}$), proteinsku dimerizaciju (FDR = $7.5*10^{-13}$) i vezanje nukleinskih kiselina (FDR = $1.5*10^{-8}$). Također, u navedenom skupu obogaćeni su i izrazi GO povezani s brojnim signalnim putevima.

3.4.4. Analiza odnosa dugih nekodirajućih RNA i transpozona

U ovom dijelu istraživanja usporedio sam pristunost transpozona u genima dugih nekodirajućih RNA i protein-kodirajućim genima ogulinske špiljske spužvice. Navedena usporedba uključivala je analizu ukupnog broj transpozona koji se cijelom svojom dužinom nalaze unutar lncRNA-kodirajućih gena i protein-kodirajućih gena, kao i unutar 5' regija navedenih gena, pri čemu su 5' regije definirane kao regije duljine 1 kb u 5' smjeru od početka gena. Rezultati navedene analize prikazani su u Tablici 7.

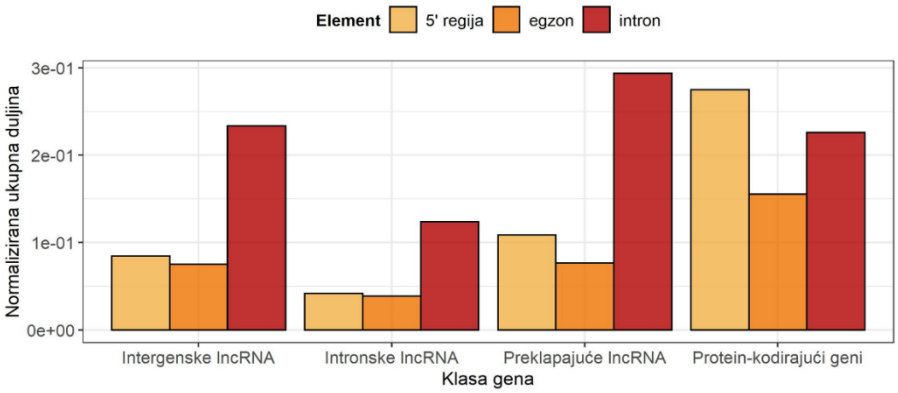
Tablica 7. Broj različitih lncRNA-kodirajućih i protein-kodirajućih gena ogulinske špiljske spužvice koji sadrže jedan ili više transpozona. 5' regije gena definirane su kao regija duljine 1 kb u 5' smjeru od početka gena.

Klasa gena	Intergenske lncRNA	Intronske lncRNA	Preklapajuće lncRNA	lncRNA-kodirajući geni (ukupno)	Protein-kodirajući geni
Broj gena s insercijom transpozona	545	502	238	1285	11323
Udio gena s insercijom transpozona (%)	24.61	39.16	93.70	34.26	35.34
Broj gena s 5' insercijom transpozona (%)	981	668	133	1782	12463
Udio gena s 5' insercijom transpozona	44.27	52.11	52.36	47.51	38.90

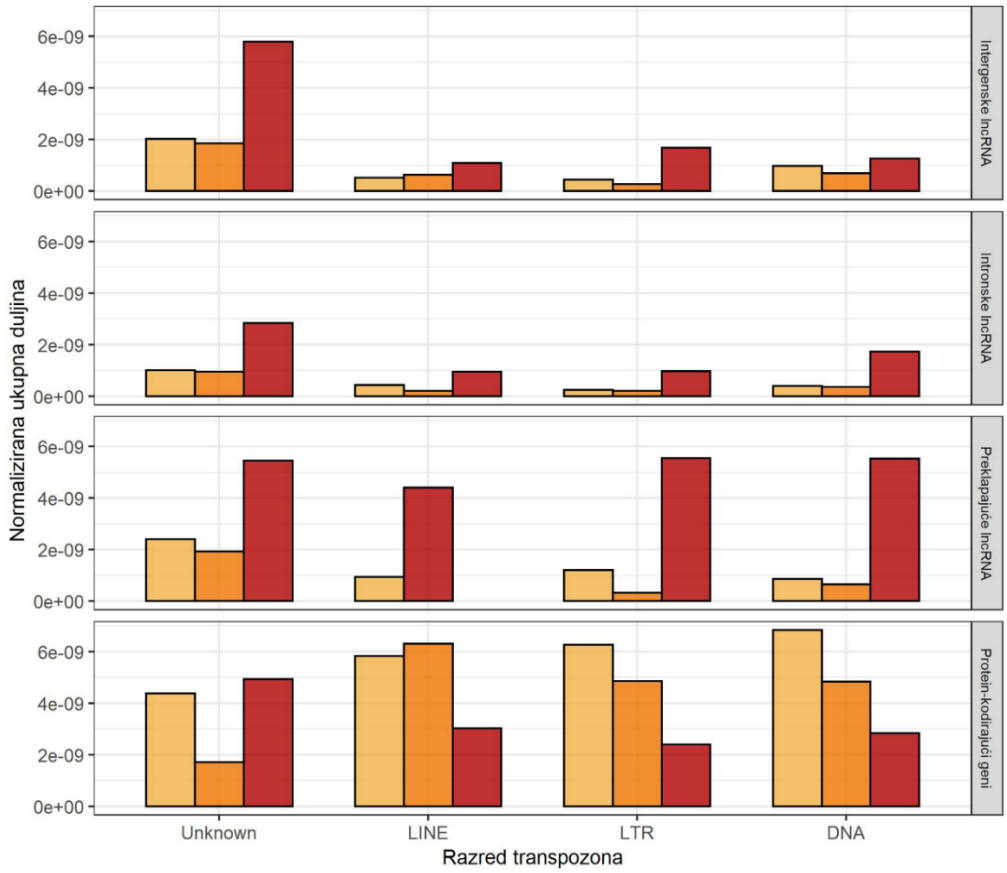
Iz Tablice 7 vidljivo je da u usporedbi s protein-kodirajućim genima, manji broj lncRNA-kodirajućih gena sadrži transpozone, dok za 5' regije navedenih vrsta gena vrijedi obrnuto. Gledajući klase lncRNA-kodirajućih gena, najveći broj gena s insercijom transpozona pokazuju intergenske lncRNA, dok najveći udio gena s insercijom transpozona imaju preklapajuće lncRNA. S druge strane, najveći udio gena koji sadrže inserciju transpozona u svojoj 5' regiji pripada klasi intronskih lncRNA.

Također, u sklopu usporedbe prisutnosti transpozona u lncRNA-kodirajućim i protein-kodirajućim genima, izračunao sam broj preklapajućih baza gena i transpozona za 5' regiju, introne i egzone svake klase lncRNA-kodirajućih gena i protein-kodirajućih gena, pri čemu sam duljinu analizirane 5' regije ponovno postavio na 1 kb. Navedeni broj normalizirao sam obzirom na ukupnu duljinu određenog genomskog elementa (Slika 18 A). Također, izračunao sam ukupni broj preklapajućih baza transpozona svih razreda s odgovarajućim elementom lncRNA-kodirajućih i protein-kodirajućih gena. Pritom, ukupni broj baza normalizirao sam obzirom na ukupnu duljinu određenog genomskog elementa, kao i obzirom na ukupnu duljinu određenog razreda transpozona (Slika 18 B).

A)



B)



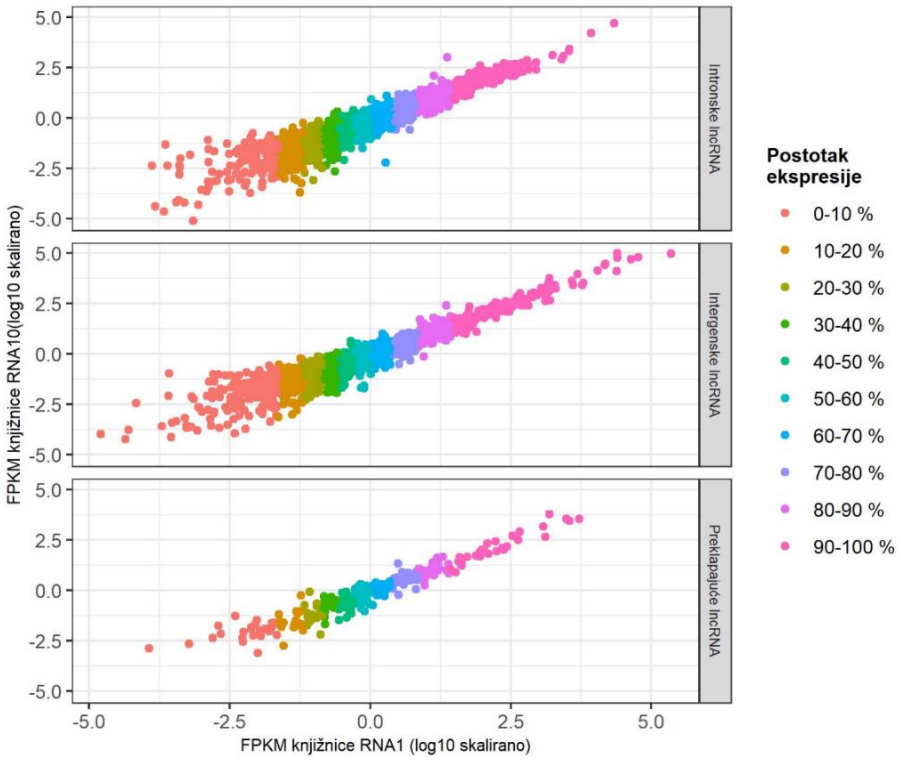
Slika 18. A) Ukupna duljina preklapjućih baza transpozona i elemenata svake klase gena. Navedena duljina normalizirana je obzirom na ukupnu duljinu pojedinog genomskog elementa. B) Ukupna duljina preklapajućih baza transpozona i elemenata svih klasa gena ogulinske špiljske spužvice, Ukupna duljina normalizirana je obzirom na ukupni broj baza određenog razreda transpozona, kao i obzirom na ukupan broj baza pojedinog genomskog elementa.

Na Slici 18 vidljivo je da je normalizirani broj preklapajućih baza protein-kodirajućih gena i svih razreda transpozona veći od normaliziranog broja preklapajućih baza svih klasa lncRNA-kodirajućih gena i svih razreda transpozona. Uzimajući u obzir klase lncRNA-kodirajućih gena definirane obzirom na odnos s protein-kodirajućim genima, najveći normalizirani broj baza transpozona nalazi se u preklapajućim lncRNA-kodirajućim genima, dok se najmanji normalizirani broj baza transpozona nalazi u intronskim lncRNA-kodirajućim genima.

Općenito gledajući, najmanji broj preklapajućih baza transpozona i gena nalazi se u egzonima gena. Isto tako, najveći broj normaliziranih baza svih skupina gena ostvaruju preklapajuće baze transpozona nepoznatog razreda. Nadalje, transpozoni svih ostalih razreda pokazali su značajno veći normalizirani broj preklapajućih baza s protein-kodirajućim genima od normaliziranog broja preklapajućih baza sa svim klasama lncRNA-kodirajućih gena. Iznimka su introni preklapajućih lncRNA, koji pokazuju značajno veći udio preklapajućih baza s transpozonima svih razreda od ostalih klasa gena

3.4.5. Analiza ekspresije dugih nekodirajućih RNA

U nastavku istraživanja analizirao sam relativnu ekspresiju gena dugih nekodirajućih RNA, utjecaj transpozona na ekspresiju lncRNA-kodirajućih gena, kao i utjecaj lncRNA-kodirajućih gena na ekspresiju protein-kodirajućih gena. Na početku analize ekspresije, usporedio sam vrijednosti FPKM klasa lncRNA-kodirajućih gena prvi i deseti dan rasta primorfa ogulinske špiljske spužvice, pri čemu su klase lncRNA definirane obzirom na odnos s protein-kodirajućim genima. Rezultat takve usporedbe prikazan je na Slici 19.



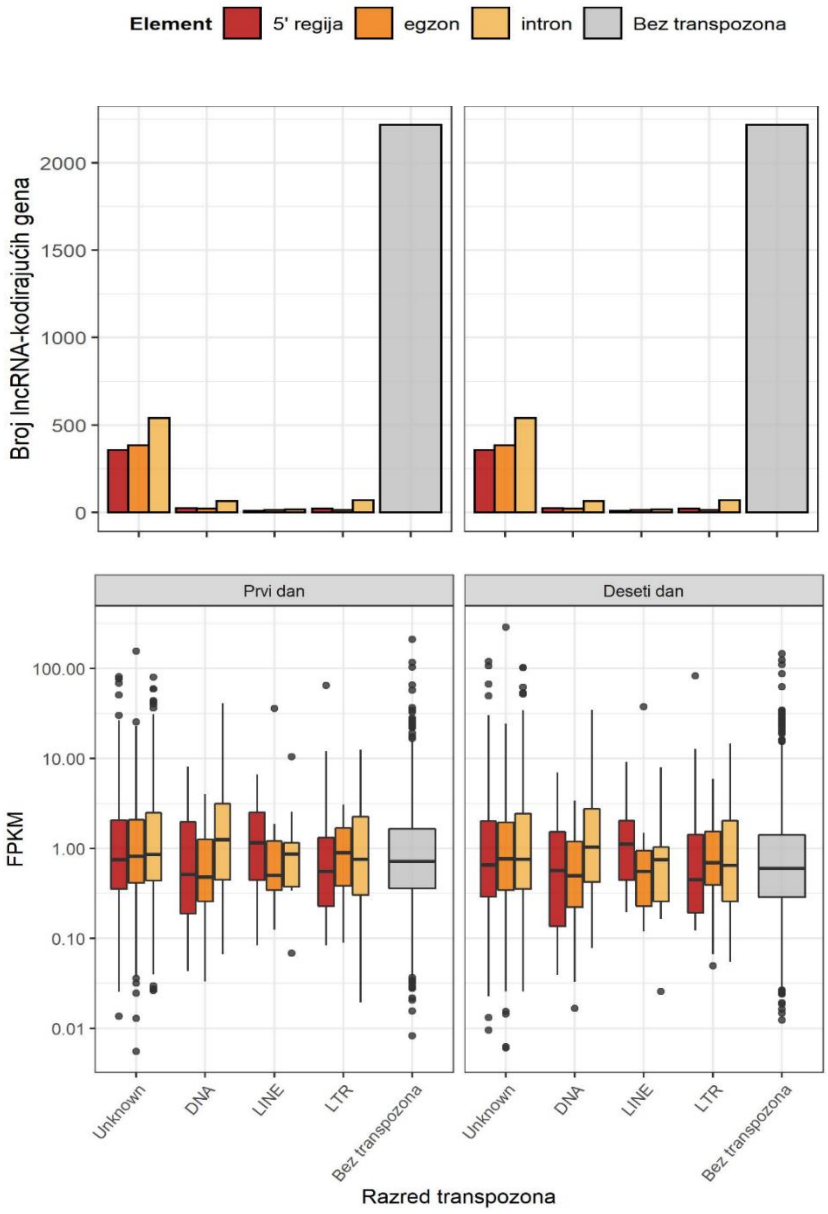
Slika 19. Usporedba ekspresije klasa lncRNA-kodirajućih gena prvi i deseti rast primorfa ogulinske špiljske spužvice, pri čemu su klase definirane obzirom na odnos s protein-kodirajućim genima. Geni su obojani prema postotku ekspresije prvog dana rasta primorfa. Odnosi vrijednosti FPKM su logaritamski skalirani.

Iz Slike 19 vidljivo je da većina lncRNA-kodirajućih gena svih klasa u prvom i desetom danu rasta primorfa pokazuju sličnu razinu ekspresije. Također, utvrđena je značajna razlika u ekspresiji pojedinih klasa lncRNA-kodirajućih gena prvi dan (Kruskal-Wallisov test, p=0.0026, i deseti dan (Kruskal-Wallisov test, $p=1.07*10^{-7}$).

Rad utvrđivanja utjecaja transpozona na ekspresiju lncRNA-kodirajućih gena, u sljedećem dijelu analize usporedio sam ekspresiju lncRNA-kodirajućih gena s insercijom transpozona u 5' regiji, egzonu i intronu gena s ekspresijom lncRNA-kodirajućih gena bez insercija transpozona. Rezultati takve usporedbe prikazani su na Slici 20.

A)

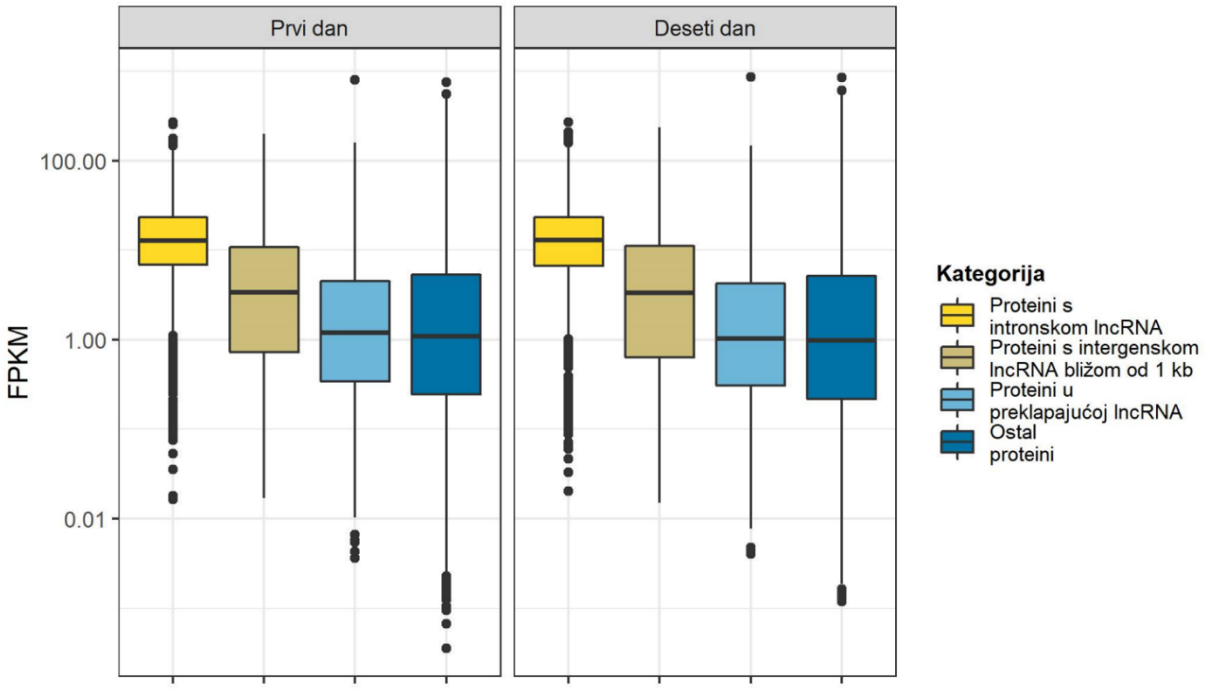
B)



Slika 20. Usporedba ekspresije lncRNA-kodirajućih gena s insercijom transpozona u 5' regiji, egzonu i intronu s ekspresijom gena bez insercije transpozona. A) Broj lncRNA-kodirajućih gena s insercijama pojedinih razreda transpozona. U slučaju preklapanja gena s više transpozona, uzet je transpozon s maksimalnim brojem preklapajućih baza. Radi preglednosti, navedeni graf prikazan je dva puta. B) Raspodjela vrijednosti FPKM lncRNA-kodirajućih gena svih klasa u odnosu na insercije transpozona. Horizontalna crna linija predstavlja medijan raspodjele, dok se vertlkalne crne linije protežu od -1.5*IR do 1.5*IR, gdje IR označava interkvartilni raspon raspodjele. Crnim točkama prikazane su ekstremne vrijednosti raspodjele. Odnosi vrijednosti FPKM na ovom prikazu su logaritamski skalirani.

Na Slici 20 uočljivo je da su odnosi ekspresije lncRNA-kodirajućih gena s insercijom i bez insercija skoro svih razreda transpozona vrlo slični u uzorcima primorfa izoliranih prvog i desetog dana njihova rasta. Svi lncRNA-kodirajući geni s insercijama transpozona u intronima pokazuju veću ekspresiju od lncRNA-kodirajućih gena bez insercije transpozona. S druge strane, lncRNA-kodirajući geni s insercijama transpozona razreda DNA i LINE u egzonima pokazuju manju ekspresiju od lncRNA-kodirajućih gena bez insercija transpozona, dok lncRNA-kodirajući geni s egzonskim insercijama transpozona nepoznatog razreda i transpozona LTR pokazuju veću ekspresiju od lncRNA-kodirajućih gena bez insercija transpozona. Nadalje, lncRNA-kodirajući geni s insercijama transpozona razreda LINE u svojim 5' regijama pokazuju povećanu ekspresiju u odnosu na lncRNA-kodirajuće gene bez insercija transpozona, dok za lncRNA-kodirajući gene s insercijama transpozona svih ostalih razreda vrijedi obrnuto. Ipak, treba uzeti u obzir relativno mali broj gena koji sadrže preklapanja s transpozonima razreda DNA, LINE i LTR.

Na kraju analize ekspresije, istražio sam odnos ekspresije protein-kodirajućih gena koji u svom intronu sadrže intronski lncRNA-kodirajući gen, protein-kodirajućih gena koji se nalaze u intronu preklapajućih lncRNA-kodirajućih gena, protein-kodirajućih gena koji su udaljeni 1 kb ili manje od najbližeg lncRNA-kodirajućeg gena i svih osstalih protein-kodirajućih gena. Navedena usporedba prikazana je na Slici 21.



Slika 21. Ekspresija protein-kodirajućih gena obzirom na njihov odnos s lncRNA-kodirajućim genima ogulinske špiljske spužvice. Horizontalna crna linija predstavlja medijan raspodjele, dok se vertlkalne crne linije protežu od -1.5*IR do 1.5*IR, gdje IR označava interkvartilni raspon raspodjele. Crnim točkama prikazane su ekstremne vrijednosti raspodjele. Odnosi vrijednosti FPKM na ovom prikazu su logaritamski skalirani.

Iz raspodjele vrijednosti FPKM protein-kodirajućih gena podijeljenih obzirom na odnos s lncRNA-kodirajućim genima vidljivo je da proteini s insercijom lncRNA-kodirajućeg gena u svom intronu imaju znatno veću ekspresiju od svih ostalih proteina. Isto tako, proteini čija je

udaljenost od najbližeg lncRNA-kodirajućeg gena manja od 1 kb pokazuju veću razinu ekspresije od protein-kodirajućih gena u preklapajućim lncRNA-kodirajućim genima i protein-kodirajućim genima koji su od najbližeg lncRNA-kodirajućeg gena udaljeniji od 1 kb. Tukey-Kramerovim testom na logaritamski transformiranim vrijednostima FPKM prvog dana usporedio sam ekspresiju svake klase proteina-kodirajućeg gena s proteinim-kodirajućim genima koji nisu u odnosu s lncRNA-kodirajućim genima. Rezultati navedenog testa prikazani su u Tablici 8.

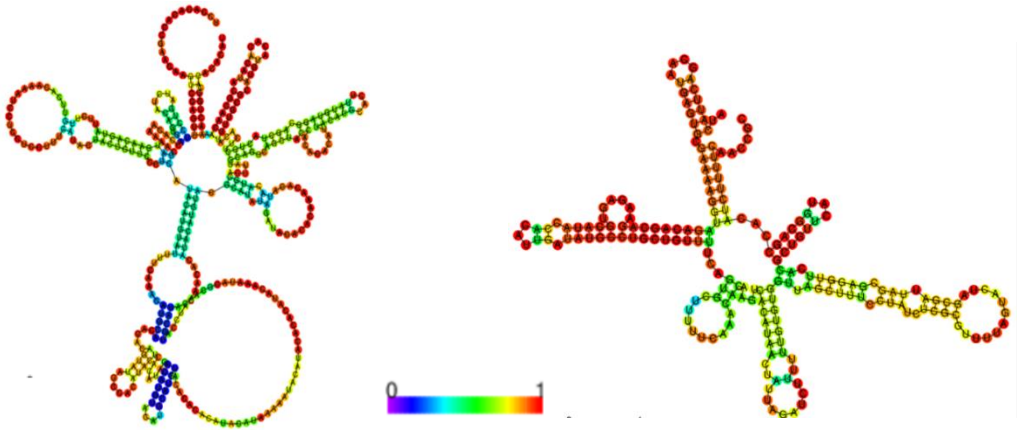
Tablica 8. Rezultati Tukey-Kramerova testa na logaritamski transformiranim vrijednostima FPKM skupina protein-kodirajućih gena, pri čemu su navedene skupine definirane obzirom na odnos s lncRNA-kodirajućim genima. Svaka skupina protein-kodirajućih gena uspoređena je s protein-kodirajućim genima koji su od najbližeg lncRNA-kodirajućeg gena udaljeniji od 1 kb.

Klasa proteina	**Prilagođena p-vrijednost**
Protein s intronskom lncRNA	0.0000
Protein s intergenskom lncRNA bližom od 1 kb	0.0009
Proteini u preklapajućoj lncRNA	0.2875

Navedeni rezultati Tukey-Kramerovih testova sugeriraju statistički značajnu razliku u ekspresiji protein-kodirajućih gena koji sadrže intronsku lncRNA i protein-kodirajućih gena koji su od najbliže lncRNA udaljeniji od 1 kb. Također, uočena je statistički značajna razlika između ekspresije protein-kodirajućih gena koji su od najbližeg lncRNA-kodirajućeg gena udaljeni manje od 1 kb i protein-kodirajućih gena koji su od najbližeg lncRNA-kodirajućeg gena udaljeni više od 1 kb. S druge strane, nije uočena statistički značajna razlika u ekspresiji protein-kodirajućih gena koji se nalaze unutar preklapajućeg lncRNA-kodirajućim genom i protein-kodirajućih gena koji su od najbližeg lncRNA-kodirajućeg gena udaljeni više od 1 kb.

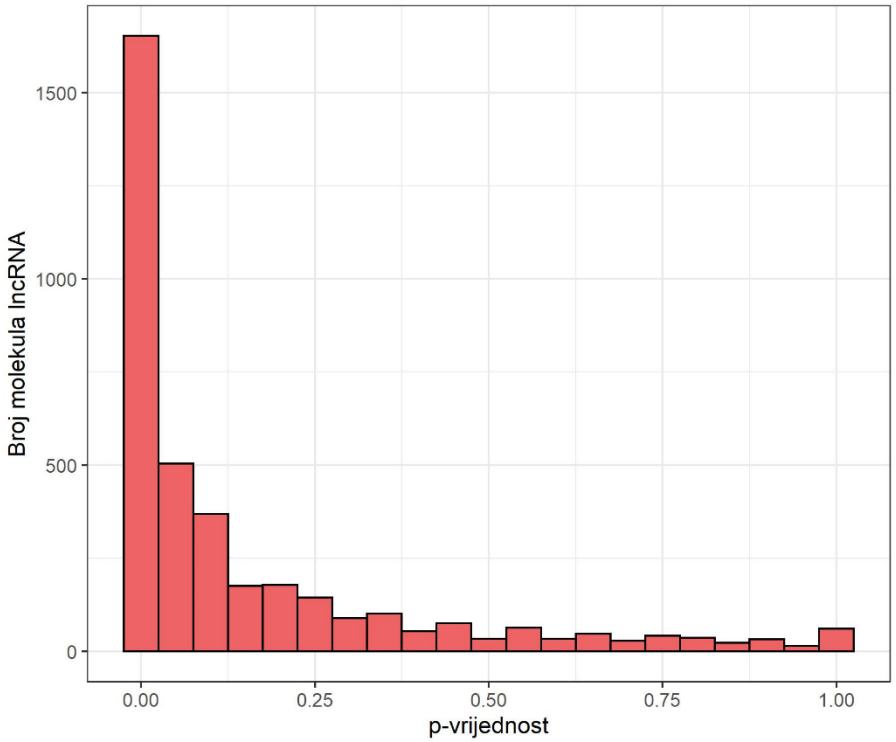
3.4.6. Analiza sekundarnih struktura pronađenih dugih nekodirajućih RNA

U nastavku istraživanja odredio sam sekundarne strukture svim najdužim izoformama s najvećim brojem egzona pronađenih gena dugih nekodirajućih RNA, kao i njihovim reverznim komplementima. Na Slici 21 prikazano je nekoliko primjera određenih sekundarnih struktura.



Slika 21. Primjer dvaju sekundarnih struktura pronađenih dugih nekodirajućih RNA ogulinske špiljske spužvice. U prikazanim strukturama vidljivi su elementi poput heliksa, granajućih petlji i ukosnica. Strukture su obojane obzirom na vjerojatnost sparivanja, odnosno po pouzdanosti njihovih dijelova.

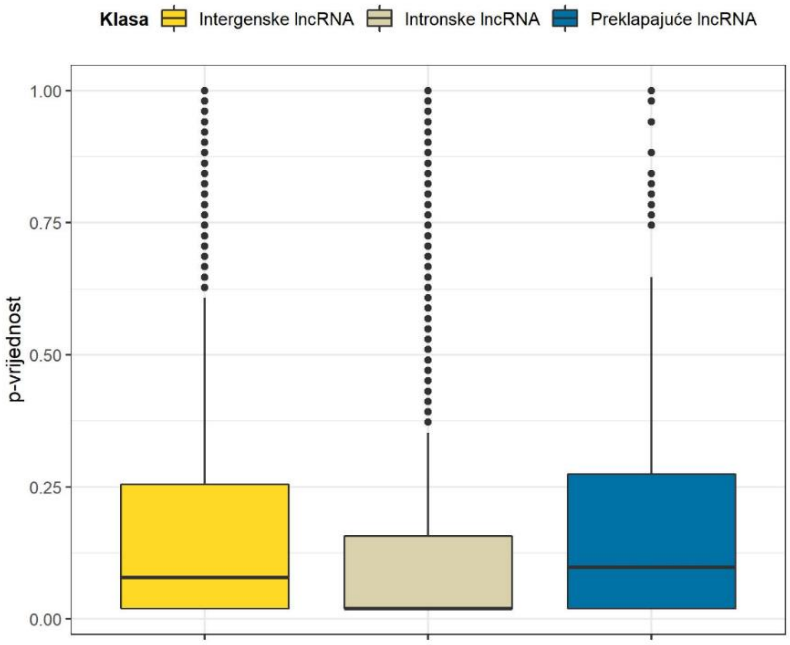
Također, permutacijskim testom usporedio sam stabilnost dobivenih struktura i struktura nasumičnih sljedova jednake duljine. Tim postupkom dobio sam p-vrijednost koja predstavlja vjerojatnost dobivanja strukture jednake stabilnosti iz nasumičnog slijeda. U nastavku analize sam za svaki par molekule lncRNA i njezina reverznog komplementa uzeo onu molekulu koja je imala manju p-vrijednost. Spomenuta raspodjela p-vrijednosti prikazana je na Slici 22.



Sl**ika 22**. Raspodjela p-vrijednosti dobivenih permutacijskim testom stabilnosti sekundarnih struktura najdužih izoformi pronađenih gena dugih nekodirajućih RNA (N=3753, interval = 0.05).

Iz raspodjele p-vrijednosti dobivenih permutacijskim testom stabilnosti sekundarne strukture vidljivo je da većina molekula lncRNA ima relativno niske p-vrijednosti, pri čemu njih 52% ima p-vrijednost nižu od 0.05, a njih 65% ima p-vrijednost nižu od 0.1).

Radi analize stabilnosti pojedinih klasa lncRNA transkripata, pri čemu su klase definirane obzirom na odnos s protein-kodirajućim genima, usporedio sam raspodjele p-vrijednosti svih klasa najdužih izoformi s najvećim brojem egzona lncRNA-kodirajućih gena dobivene permutacijskim testom stabilnosti sekundarnih struktura. Rezultati ove usporedbe prikazani su na Slici 23.

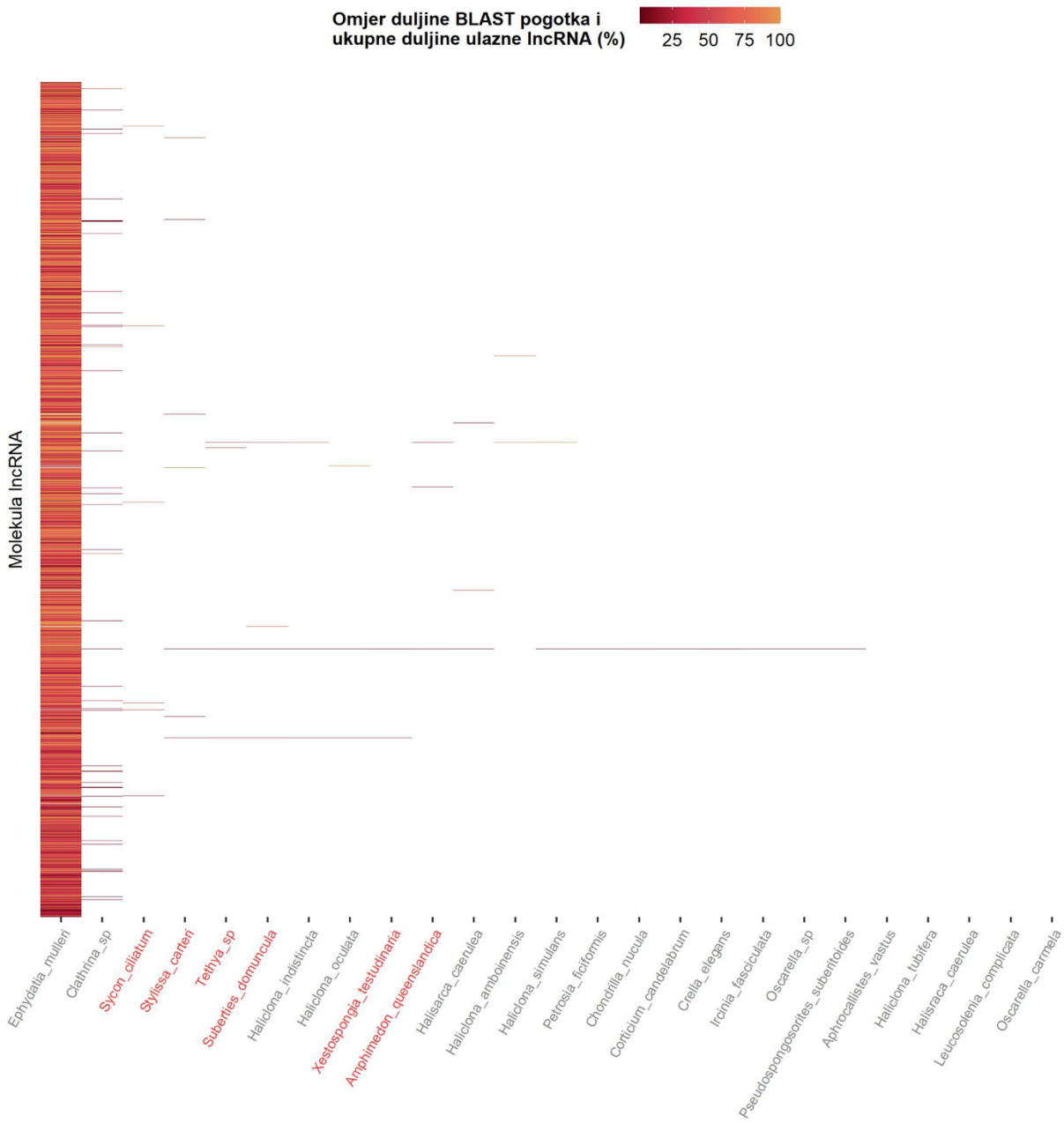


Slika 23. Raspodjela p-vrijednosti dobivenih permutacijskim testom stabilnosti sekundarnih struktura klasa pronađenih dugih nekodirajućih RNA, pri čemu su klase definirane obzirom na odnos s protein-kodirajućim genima. Horizontalna crna linija predstavlja medijan raspodjele, dok se vertlkalne crne linije protežu od -1.5*IR do 1.5*IR, gdje IR označava interkvartilni raspon raspodjele. Crnim točkama prikazane su ekstremne vrijednosti raspodjele.

Iz raspodjele p-vrijednosti klasa lncRNA vidljivo je da su intronske lncRNA najstabilnije. Tukey-Kramerovim testom nad logaritamski transformiranim p-vrijednostima utvrdio sam značaju razliku između p-vrijednosti intronskih i intergenskih lncRNA (p=0.000) te između intronskih i preklapajućih lncRNA (p=0.000). S druge strane, nisam utvrdio statistički značajnu razliku između p-vrijednosti intergenskih i preklapajućih lncRNA (p=0.299).

3.4.7. Analiza očuvanosti dugih nekodirajućih RNA unutar koljena spužvi

Očuvanost dugih nekodirajućih RNA spužvi još uvijek je relativno neistražena tema. Iz tog razloga, u ovom dijelu istraživanja pokušao sam pronaći potencijalne homologe molekula lncRNA ogulinske špiljske spužvice. Potencijalne homologe tražio sam programom BLAST, pri čemu sam kao ulazne sljedove koristio najduže izoforme s najvećim brojem egzona lncRNA-kodirajućih gena, a kao pretraženu bazu 6 dostupnih genoma i 18 dostupnih transkriptoma svih razreda koljena spužvi. Na Slici 24. prikazani su pogotci navedene potrage za svaku od ulaznih molekula lncRNA s očekivanom vrijednošću manjom od 10^{-15}. Pritom, za 2262 ulazna transkripta pronađen je jedan ili više značajan pogodak, dok za 1491 transkript nisu pronađeni značajni pogotci.



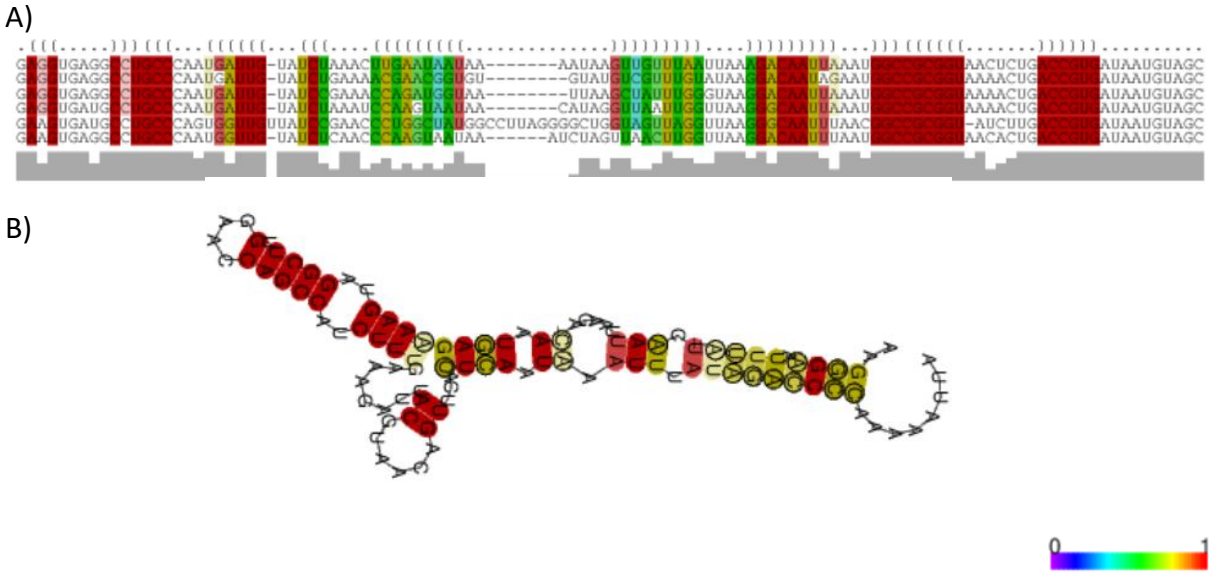
Slika 24. Prikaz pogodaka dugih nekodirajućih RNA ogulinske špiljske spužvice prilikom pretrage svih dostupnih genoma i transkriptoma koljena spužvi. Sivom bojom oznaka x-osi označeni su transkriptomi, dok su crvenom bojom oznaka x-osi označeni genomi. Pogotci su obojani obzirom na omjer duljine BLAST pogotka i ukupne duljine ulaznog transkripta. Prikazani su pogotci s očekivanom vrijednošću manjom od 10^{-15}. Također, transkripti bez pogodaka nisu prikazani.

Iz prikaza pogodaka molekula lncRNA prilikom pretrage genoma i transkriptoma koljena spužvi vidljiva je velika razina sličnosti između pronađenih molekula lncRNA ogulinske špiljske spužvice i transkripata spužve *Ep m,l*

2hydatia mulleri. Od 2148 transkripata kojima je pronađena sličnost u transkriptomu spužve *Ephydatia mulleri*, njih 1298 (60.4%) ima omjer duljine BLAST pogotka i transkripta veći od 50

%, dok kod 12 % transkripata taj omjer iznosi 100%. Također, od 1298 pogodaka s omjerom duljine BLAST pogotka i transkripta većim od 50%, njih 1085 (84%) pogođeno je od strane transkripta spužve *Ephydatia mulleri* čija je duljina veća od 200 pb, a duljina okvira čitanja manja od 150 pb. Slično, od 260 podotaka s omjerom duljine BLAST pogotka i ulaznog transkripta jednakom 100%, 209 (80.4)% pogodeđnih transkripata spužve *Ephidatia mulleri* ima navedene karakteristike. Spužva s drugim najvećim brojem BLAST pogodaka ulaznih molekula lncRNA je *Clathrina sp*, kod koje spomenuti broj pogodaka iznosi 55, od kojih većina obuhvaća mali udio transkripta lncRNA ogulinske špiljske spužvice. S druge strane, pogotci u ostalim transkriptomima i genomima spužvi relativno su rijetki te također većinom obuhvaćaju mali udio transkripta lncRNA.

Također, rezultati pretraživanja sličnosti između transkripata lncRNA ogulinske špiljske spužvice i dostupnih transkriptoma te genoma koljena spužvi ukazuju na postojanje 26 transkripata lncRNA ogulinske špiljske spužvice s pronađenom sličnošću kod dvaju ili više ostalih spužvi. Većina takvih transkripata pokazuje sličnost sa sljedovima dvaju ostalih spužvi, dok najočuvaniji transkript pokazuje sličnost sa sljedovima osam ostalih spužvi. Sve takve transkripte s ukupnom duljinom pogotka većom od 150 pb višestruko sam poravnao sa svim transkriptima s kojima pokazuju sličnost te sam u navedenim poravnanjima programon RNAz analizirao očuvane motive sekundarne strukture RNA. Očuvani elementi sekundarne strukture koje je program RNAz klasificirao kao strukturnu RNA pronađeni su u 15 od 26 poravnanja. Primjer takvog elementa prikazan je na Slici 25.



Slika 25. Primjer očuvanog motiva sekundarne strukture RNA pronađenog u sedam vrsta spužvi. A) Višestruko poravnanje primarnih sljedova navedenog motiva. B) Prikaz konsenzusne sekundarne strukture navedenog motiva. Boje predstavljaju vjerojatnost sparivanja pojedinih parova baza, odnosno pouzdanost određenog dijela konsenzusne strukture.

3.4.8. Pronalazak potencijalnih homologa pronađenih dugih nekodirajućih RNA izvan koljena spužvi

Za pronalazak potencijalnih homologa pronađenih dugih nekodirajućih RNA ogulinske špiljske spužvice izvan koljena Porifera koristio sam bazu nekodirajućih RNA RNACentral i program nhmmer. Pretražio sam sljedove navedene baze radi utvrđivanja sličnih sljedova lncRNA-kodirajućim genima ogulinske špiljske spužvice, pri čemu sam koristio najdužu izoformu s najvećim brojem egzona svakog gena. Pritom, pronađeni su značajni pogotci za 9 ulaznih transkripata lncRNA, od kojih je kod 8 transkripta utvrđena sličnost s transkriptima lncRNA kralješnjaka, dok je kod jednog transkripta utvrđena sličnost s neklasificiranim RNA koljena Protobacteria. Granična očekivana vrijednost pogodaka postavljena je na 10^{-18}, a duljine pronađenih pogodaka kreću se od 225 do 362 nukleotida.

Isto tako, radi utvrđivanja očuvanih domena u transkriptima lncRNA ogulinske špiljske spužvice, programom Infernal pretražio sam bazu Rfam, koja predstavlja skup visoko očuvanih domena RNA. Navedenom pretragom nisam pronašao ni jednu očuvanu funkcionalnu domenu RNA, pri čemu je granična očekivana vrijednost postavljena na 10^{-15}.

4. Rasprava

Duge nekodirajuće RNA donedavno su smatrane molekulama koje predstavljaju nusprodukte transkripcije bez značajne funkcije. Međutim, brojna istraživanja molekula lncRNA sisavaca pokazale su njihovu značajnu ulogu u regulaciji ekspresije brojnih gena, pri čemu na različite načine mogu utjecati na brojne biološke procese, poput razvoja, starenja, ali i nastanka različtiih bolesti (Jarroux i sur. 2017). Ipak, dok su molekule lncRNA čovjeka postale relativno popularna tema istraživanja, molekule lncRNA drugih organizama, posebice beskralješnjaka, još uvijek nisu dovoljno dobro opisane. Spomenuto vrijedi i za koljeno spužvi, skupinu koja se među prvima odvojila od zajedničkog pretka životinja, zbog čega predstavlja zanimljiv predmet istraživanja. Naime, molekule lncRNA do danas su opisane kod samo jedne vrste spužve (Gaiti i sur. 2015), zbog čega će anotacija navedenih molekula kod drugih vrsta zasigurno upotpuniti znanje o važnosti i značaju ovih vrsta RNA u skupini spužvi. Iz tog razloga, cilj ovog rada bio je upotpuniti katalog molekula lncRNA spužvi koristeći sekvencirane sljedove ogulinske špiljske spužvice, ali i opisati te usporediti njihova svojstva s molekulama lncRNA drugih organizama.

U istraživanju sam koristio knjižnice RNA dobivene dubokim sekvenciranjem primorfa ogulinske špiljske spužvice prvog i desetog dana njihova rasta. Na početku analize obradio sam navedene sljedove, pri čemu je došlo do povećanja njihove kvalitete. Takve sljedove iskoristio sam za sastavljanje transkriptoma programima rnaSPAdes i Trinity. Dobiveni statistički parametri navedenih transkriptoma ukazuju na veću fragmentiranost transkriptoma rnaSPAdes. Ipak, vrlo visok postotak mapiranosti sljedova knjižnica RNA1 i RNA10 ukazuju na relativno visoku kvalitetu složenih transkriptoma. Iz dobivenih transkriptoma izdvojio sam molekule lncRNA serijom relativno strogih filtriranja kojima je cilj bio najbolja moguća ravnoteža između lažno pozitivnih i lažno negativnih rezultata. Rezultati filtriranja sugeriraju izrazito mali broj molekula rRNA u transkriptomima, što je rezultat metode izolacije RNA pomoću poliadeniliranih krajeva. S druge strane, filtriranjem je uklonjen velik broj transkripata koji bi mogli predstavljati pogrešno složene sljedove ili kraće nekodirajuće molekule RNA. Također, filtriranjem po duljini okvira čitanja iz transkriptoma sam uklonio većinu molekula mRNA, zbog čega filtriranje obzirom na sličnost s poznatim proteinima i proteinskim domenama nije uklonilo velik broj sljedova. Filtriranje obzirom na sličnost s poznatim proteinima vjerojatno je izbacilo veliki broj molekula lncRNA nastalih iz pseudogena, što predstavlja jedan od češćih načina nastanka novih molekula lncRNA. Ipak, za takvo filtriranje odlučio sam se zbog nemogućnosti razlikovanja potencijalnih pseudogena od pogrešno sklopljenih sljedova mRNA. Najstroži kriteriji filtriranja uključivali su uklanjanje transkripata s jednim egzonom te transkripata koji se mapiraju na genom s manje od 95% svoje duljine, čime je gotovo sigurno uklonjen i velik broj molekula lncRNA. Ipak, smatram da sam navedenim filtriranjem uklonio značajno veći broj lažno pozitivnih transkripata. Za kraj, uklonio sam sve transkripte čiji se egzoni preklapaju s anotiranim egzonima protein-kodirajućih gena, čime sam filtrirao sve pronađene egzonske lncRNA. Naime, transkripti čiji se egzoni preklapaju s anotiranim protein-kodirajućim egzonima osim egzonskih lncRNA mogu predstavljati fragmentirane pogrešno složene mRNA, kao i neanotirane izoforme mRNA. Iz

tog razloga, većina istraživanja koja se bave molekulama lncRNA ne uzimaju u obzir gene čiji se egzoni preklapaju s egzonima protein-kodirajućih gena, pri čemu se transkribiraju u istom smjeru kao i navedeni protein-kodirajući gen (poput Derrien i sur. 2012). Budući da je sekvenciranje RNA sljedova korištenih knjižnica napravljeno bez informacija o prepisanom lancu, nisam mogao karakterizirati molekule lncRNA koje se prepisuju u obrnutom smjeru od prepisivanja odgovarajućeg protein-kodirajućeg gena. Posljedično, nisam mogao razlikovati egzonske lncRNA koje se prepisuju u smjeru – od molekula lncRNA koje se prepisuju u smjeru +, zbog čega sam izbacio sve takve transkripte. Osim filtriranja svih egzonskih RNA, na broj pronađenih molekula lncRNA utjecao je postupak izolacije RNA koji se temeljio na svojstvu poliadenilirajućih krajeva RNA na njihovu 3' kraju, koje znatan broj molekula lncRNA nema.

Obzirom na navedene nedostatke eksperimenta, kao i na relativno strogu seriju filtriranja, smatram da je dobiveni broj konačnih transkripata izrazito visok, pri čemu velika većina pronađenih transkripata predstavlja biološki relevantne molekule lncRNA. Važno je naglasiti da je navedeni broj skoro duplo veći od broja molekula lncRNA pronađenih kod spužve *Amphimedon queenslandica* dobivene iz uzoraka četiri različita razvojna stadija ličinke (Gaiti i sur. 2015). Smatram da tome pridonosi nekoliko razloga, uključujući izrazito veliku dubinu sekvenciranja, što je izrazito važno obzirom na slabu ekspresiju većine molekula lncRNA. Također, u eksperimentu na kojem se temelji ovo istraživanje uzorkovani su primorfi, koji su za razliku od ličinki spužve *Amphimedon queenslandica* vjerojatno sadržali drugačiji repertoar molekula lncRNA, uključujući brojne molekule lncRNA karakteristične za odrasli stadij, ali i regeneraciju te diferencijaciju. Isto tako, jedan od glavnih razloga relativno velikog broja pronađenih molekula lncRNA je korištenje transkriptoma sklopljenih različitim programima. Naime, brojna istraživanja ustanovila su značajnu razliku u sastavu transkriptoma sklopljenih iz istih sljedova RNA, ali različitim programima (poput Nakasugi i sur. 2014). Ipak, u takvim slučajevima treba voditi računa o uzimanju što manje redundantnog skupa sljedova, zbog čega sam analizu pronađenih molekula lncRNA započeo određivanjem njihova neredundantnog konsenzusnog skupa. Uzimajući jedinstvene gene iz oba transkriptoma i duže gene preklapajućih parova osigurao sam visok stupanj raznovrsnosti i neredundantnosti konsensuznog skupa transkripata. S druge strane, smatram da je uzimanje kraćih gena preklapajućih parova s presjekom kraćim od 20% duljine dužeg gena u paru povećalo redundantnost navedenog skupa zbog gena koji se cijelom svojom duljinom nalaze u drugom genu preklapajućeg para. S druge strane, uzimanje navedenih kraćih parova omogućilo je i razmatranje gena slične dužine drugom genu u paru koji čine geni koji se međusobno preklapaju na svojim krajevima, što je uzrokovalo istovremeno povećanje raznovrsnosti transkripata u konsenzusnom skupu. Također, valja naglasiti i mogućnost postojanja bioloških značajnih gena koji se nalaze unutar većih gena (engl. *nested genes*), Uzimanje opisanog konsenzusa omogućilo je razmatranje takvih gena čija duljina iznosi manje od 20% duljine gena u kojem se nalaze. Naime, spomenuta granica od 20% predstavlja ručnu određenu kriterij s ciljem smanjenja redundantnosti gena koji se preklapaju na krajevima i povećanja raznovrsnosti gena koji se nalaze unutar drugih gena. Izrazito velik postotak jedinstvenih gena oba transkriptoma, kao i visok postotak gena te transkripata u dobivenom konsenzusu potvrđuje pretpostavku mnogih autora o raznovrsnosti transkripata dobivenih korištenjem različitih programa za sastavljanje transkriptoma.

Većina trendova pronađenih u analizi općenitih svojstava molekula lncRNA ogulinske špiljske spužvice slični su trendovima pronađenim u drugim organizmima, uključujući i miša te čovjeka. Navedeni trendovi uključuju pravilnosti u broju izoformi po genu, broju egzona po izoformi, duljini transkripata, mjestima prekrajanja i udjelu GC. Od spomenutih karakteristika važno je naglasiti posjedovanje dvaju egzona od strane velike većine transkripata lncRNA, što potvrđuje da molekule lncRNA spužvi prolaze proces izrezivanja introna. Također, iako velika većina molekula lncRNA ima jednu izoformu, postoji značajan broj molekula lncRNA koji prolaze kroz proces alternativnog prekrajanja. Također, relativno visok udio kanonskih mjesta izrezivanja introna pokazuje da se prekrajanje transkripta lncRNA i transkripata mRNA događa sličnim ili istim mehanizmom. Pritom, manji udio kanonskih prekrajajućih mjesta u intronima molekula lncRNA u odnosu na introne molekula mRNA vjerojatno je rezultat nepreciznosti u sastavljanju genoma i transkriptoma, kao i u mapiranju transkripata na genom. Zanimljiva je i razlika u udjelu GC između molekula lncRNA i molekula mRNA ogulinske špiljske spužvice, pri čemu molekule lncRNA pokazuju prosječni udio GC puno sličniji prosječnom udjelu GC genoma od molekula mRNA. Navedeni fenomen može se objasniti selekcijom čije je djelovanje na primarnu strukturu proteina puno značajnije od djelovanja na primarnu strukturu molekula lncRNA.

Analiza odnosa pronađenih lncRNA-kodirajućih gena i protein-kodirajućih gena također je pokazala karakteristike uočene kod brojnih drugih organizama. Navedene karakteristike uključuju brojnost, ali i odnose duljina gena navedenih klasa. Također, utvrđen je velik broj lncRNA-kodirajućih gena u neposrednoj blizini protein-kodirajućih gena. Iako molekule lncRNA mogu djelovati na ekspresiju susjednih i udaljenih gena (Derrien i sur. 2012), vjerojatno je da su navedeni protein-kodirajući geni regulirani od strane intergenskih lncRNA-kodirajućih gena. Također, iako detaljni mehanizmi još uvijek nisu poznati, smatra se da velik dio intronskih te preklapajućih lncRNA djeluje na ekspresiju protein-kodirajućih gena s kojima se preklapaju (Derrien i sur. 2012). Rezultati analize obogaćenih izraza GO protein-kodirajućih gena koji su u neposrednoj blizini intergenskih lncRNA-kodirajućih gena, kao i protein-kodirajućih gena koji se preklapaju s intronskim ili preklapajućim lncRNA-kodirajućim genima pokazuju da bi geni reguliran molekulama lncRNA mogli igrati važnu ulogu u kontroli transkripcije i translacije drugih gena, kao i u brojim signalnim putevima važnih za staničnu homeostazu.

Analiza odnosa pronađenih lncRNA-kodirajućih gena i transpozona pokazuje relativno visok udio lncRNA-kodirajućih gena koji sadrže transpozone, ali i visok udio lncRNA-kodirajućih gena koji u svojoj 5' regiji sadrže transpozone, pri čemu su navedeni udjeli usporedivi s protein-kodirajućim genima. Navedeni rezultati govore u prilog mogućnosti nastajanja ili modifikacije mnogih lncRNA-kodirajućih gena insercijom transpozona u njihove egzone ili introne, pri čemu je broj takvih gena usporediv s brojem protein-kodirajućim genima koji su modificirani na isti način. Također, navedeni rezultati sugeriraju i mogućnost regulacije ekspresije mnogih lncRNA-kodirajućih gena od strane transpozona insertiranih u njihovu 5' regiju, pri čemu je udio takvih gena također usporediv s udjelom protein-kodirajućih gena koji pokazuju navedena svojstva. S druge strane, analiza ukupnog broja preklapajućih baza svih klasa lncRNA-kodirajućih gena i protein-kodirajućih gena te transpozona svih razreda pokazuje da

protein-kodirajući geni posjeduju veće obogaćenje bazama nastalih transpozicijom od lncRNA-kodirajućih gena, što se razlikuje od trendova opisanih kod drugih vrsta (Kapusta i sur. 2013). To bi mogla biti posljedica nepreciznosti u anotaciji lncRNA-kodirajućih gena, protein-kodirajućih gena ili transpozona, ali i netipičnog broja protein-kodirajućih gena ogulinske špiljske spužvice nastalih ili modificiranih insercijama transpozona. Od klasa lncRNA-kodirajućih gena, preklapajući i intergenski lncRNA-kodirajući geni pokazuju izrazito visoko obogaćenje transpozona u svojim intronima, pri čemu velik broj navedenih baza preklapajućih lncRNA vjerojatno odgovara transpozonima insertiranim u 5' regije ili unutarnje elemente protein-kodirajućih gena koji se nalaze u intronu navedene preklapajuće lncRNA. Također, dok protein-kodirajući geni pokazuju obogaćenost svim razredima transpozona, daleko najbrojniji transpozoni lncRNA-kodirajućih gena pripadaju nepoznatom razredu. Iznimka tome su introni preklapajućih lncRNA-kodirajućih gena, koji su obogaćeni svim razredima transpozona, što sugerira mogućnost istovremene regulacije preklapajućeg lncRNA-kodirajućeg gena i unutarnjeg protein-kodirajućeg gena istim transpozonom. Relativno visoka obogaćenost lncRNA-kodirajućih gena transpozonima nepoznatih razreda upućuju na mogućnost relativno velike uloge još uvijek nepoznatih transpozona u nastanku i evoluciji molekula lncRNA spužvi.

O važnosti transpozona u pronađenim lncRNA-kodirajućim genima govori i dokazana razlika u ekspresiji gena s insercijom transpozona svih razreda i gena bez takvih insercija, Pronađeni obrasci ekspresije lncRNA-kodirajućih gena s insercijama transpozona u svojim 5' regijama slažu se s potencijalnom ulogom dijelova transpozona kao promotora, pojačivača ili represora ekspresije lncRNA-kodirajućih gena. Također, relativno visoka ekspresija lncRNA-kodirajućih gena s intronskim insercijama mogla bi biti uzrokovana brojnim razlozima, uključujući uvođenje novih mjesta prekrajanja ili intronskih pojačivača od strane transpozona. Isto tako, relativno visoka ekspresija lncRNA-kodirajućih gena u čijem se egzonu nalaze transpozoni razreda LINE mogla bi biti uzrokovana pozitivnim utjecajem navedenog transpozona na stabilnost transkripta lncRNA. Ova mogućnost potkrijepljena je dokazima o povećanoj stabilnosti ljudskih transkripata lncRNA s insercijama transpozona (Kapusta i sur. 2013).

Promatrajući razlike u ekspresiji lncRNA-kodirajućih gena iz uzoraka primorfa izoliranih prvi i deseti dan njihova rasta, kod većine gena nisam uočio značajnije razlike. Navedeno bi moglo biti posljedica uloga velikog broja lncRNA-kodirajućih gena koji su važni tijekom cijelog procesa rasta primorfa te manjeg broja lncRNA-kodirajućih gena koji su tijekom rasta primorfa dinamično eksprimirani. Također, promatrao sam ekspresiju proteina povezanih s lncRNA-kodirajućim genima, od kojih su sve skupine takvih proteina bile jače eksprimirane od protein-kodirajućih gena koji su udaljeniji od lncRNA-kodirajućih gena. Iznimno visoka prosječna ekspresija pronađena je kod protein-kodirajućih gena u čijim se intronima nalazi molekula lncRNA, što bi moglo potvrditi pretpostavljenu ulogu intronskih lncRNA kao RNA koje inhibiraju ulogu represivnih kraćih nekodirajućih RNA (Hansen i sur. 2013). S relativno visokom ekspresijom navedenih protein-kodirajućih gena slaže se i pretpostavljena funkcija tih gena u regulaciji ekspresije drugih gena i važnim signalnim putevima. Ipak, za donošenje pouzdanih zaključaka o navedenim odnosima ekspresije potrebno je provesti sekvenciranje na više bioloških replikata te napraviti kontrolirane eksperimente drugačijeg tipa.

Analiza sekundarnih struktura pronađenih molekula lncRNA pokazala je veliku raznolikost pronađenih struktura, uključujući sve poznate strukturne motive RNA. Permutacijski test pokazao je veliku stabilnost većine struktura što potvrđuje njihovu ulogu u biološkim procesima.

Analiza očuvanosti pronađenih transkripata lncRNA pokazala je izrazito veliku sličnost molekula lncRNA ogulinske špiljske spužvice i molekula lncRNA spužve *Ephydatia mulleri*. Budući da su molekule lncRNA relativno slabo očuvane na razini primarne strukture, navedena sličnost vrlo je iznenađujuća te bi mogla ukazivati na točnost pretpostavke o pripadnosti ovih dvaju spužvi u isti rod (Harcet i sur. 2010). U prilog tom zaključku ide i činjenica da je u drugim vrstama spužvi pronađen relativno mali broj očuvanih molekula lncRNA. Također, valja istaknuti da sam zbog navedene sličnosti između ogulinske špiljske spužvice i spužve *Ephidatia mulleri*, kao i zbog relativno velikog broja pronađenih molekula lncRNA, pronašao puno više očuvanih struktura od autora koji su karakterizirali molekule lncRNA spužve *Amphimedon queenslandica*, Pritom, pojedine strukture očuvane su u više od dva organizma, što ukazuje na potencijalno važnu ulogu određenih molekula lncRNA kod više vrsta spužvi. S druge strane, pronašao sam mali broj potencijalnih homologa molekula lncRNA izvan koljena Porifera, pri čemu su pronađeni homolozi pokazivali relativno malu sličnost. Kao što je već spomenuto, molekule lncRNA nisu očuvane po kriteriju primarne, već tercijarne strukture, koja uvjetuje njihovu specifičnu funkciju. Također, molekule lncRNA većine nižih taksonomskih kategorija karakteristične su upravo za te kategorije (Zampetaki i sur. 2018), zbog čega navedeni rezultat ne iznenađuje. Ipak, nalaz očuvanosti molekula lncRNA unutar koljena Porifera baca novo svjetlo na mogući odnos između ogulinske špiljske spužvice i spužve *Ephydatia mulleri*, kao i na potencijalno važnu ulogu homolognih lncRNA u sličnim biološkim procesima različitih vrsta spužvi.

5. Zaključak

Duge nekodirajuće RNA predstavljaju iznimno heterogenu skupinu molekula s ulogom regulacije ekspresije brojnih gena, čime utječu na mnoge biološke procese. Ipak, molekule lncRNA relativno su dobro istražene samo u skupini sisavaca, pri čemu su druga životinjska koljena zanemarena. Jedno od takvih koljena su spužve, skupina koja se među prvima odvojila od zajedničkog pretka životinja. Naime, molekule lncRNA do danas su opisane kod samo jedne vrste spužve, što ostavlja velik potencijal za daljnja istraživanja. Iz tog razloga, ovaj rad bavio se pronalaskom i analizom molekula lncRNA ogulinske špiljske spužvice, hrvatskog endema i jedine slatkovodne stigobiontske spužve na svijetu. Iz sljedova RNA izoliranih prvog i desetog dana rasta primorfa sklopljeni su transkriptomi iz kojih je serijom relativno strogih filtriranja pronađen velik broj pouzdanih molekula lncRNA. Također, pronađene su brojne sličnosti navedenih molekula lncRNA i molekula lncRNA organizama unutar i izvan koljena spužvi. Neka od sličnih svojstava uključuju raspodjele duljina gena, transkripata i egzona molekula lncRNA, izrezivanje introna i mogućnost alternativnog prekrajanja, prevladavanje kanonskih mjesta izrezivanja introna te pristranost u udjelu GC. Isto tako, analizirani su odnosi pronađenih molekula lncRNA i protein-kodirajućih gena, pri čemu je utvrđena mogućnost regulacije mnogih transkripcijskih faktora, kao i proteina važnih u brojnim signalnim putevima od strane molekula lncRNA. Nadalje, analiziran je odnos molekula lncRNA i transpozona, pri čemu je utvrđen manji doprinos transpozona strukturi lncRNA-kodirajućih gena od doprinosa transpozona strukturi protein-kodirajućih gena. Odnos molekula lncRNA i transpozona analiziran je i ekspresijskom analizom molekula lncRNA s insercijom i bez insercija transpozona, pri čemu je utvrđena korelacija između ekspresije lncRNA-kodirajućeg gena i mjesta insercije transpozona, što sugerira potencijalnu ulogu transpozona kao pojačivača ili utišivača transkripcije lncRNA-kodirajućih gena spužvi. Analiza ekspresije otkrila je i vrlo visoku razinu ekspresije protein-kodirajućih gena koji u svojim intronima sadrže lncRNA-kodirajući gen, što upućuje na potencijalnu važnost takvih proteina u rastu primorfa ogulinske špiljske spužvice. Još jedno od važnih svojstava karakteriziranih u ovom radu su sekudnarne strukture pronađenih molekula lncRNA, koje su se pokazale iznimno raznovrsnima i stabilnima. Na kraju, analiza očuvanosti pronađenih molekula lncRNA bacila je novo svjetlo na potencijalno pogrešnu klasifikaciju ogulinske špiljske spužvice. Naime, pronađen je iznenađujuće velik broj sličnih molekula lncRNA ogulinske špiljske spužvice i spužve *Ephydatia mulleri*, oko čije se srodnosti u znanstvenom svijetu već dugo vodi rasprava. Obzirom na činjenicu da primarna struktura molekula lncRNA obično nije dobro očuvana, ali i na postojanje relativno efikasnih mogućnosti postanka novih molekula lncRNA, niže taksonomske kategorije obično sadrže jedinstven set takvih molekula, što ovo otkriće čini još zanimljivijim. Ipak, unatoč slaboj očuvanosti primarne strukture, molekule lncRNA dijele brojne strukturne motive važne za njihovu funkciju, ali i svojstva koja su očuvana u širokom rasponu životinjskih koljena, što je u ovom radu i pokazano.

6. Literatura

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. i Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.

Arendsee, Z. (2017). rhmmer: Utilities Parsing 'HMMER' Results. R package version 0.1.0. (https://CRAN.R-project.org/package=rhmmer).

Bedek, J., Bilandžija H. i Jadžić, B. (2008). Ogulinska špiljska spužvica Eunapius subterraneus Sket et Velikonja. 1984, rasprostranjenost i ekologija vrste i staništa. Modruški zbornik 2, 103-130.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. i Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455–477.

Bonnet, E., Wuyts, J., Rouzé, P. i Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20(17), 2911-2917.

Buchfink, B., Xie, C. i Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59–60.

Bushnell, B. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. United States. (https://www.osti.gov/servlets/purl/1241166).

Aguilar-Camacho, J. M., Doonan, L. i McCormack, G. P. (2019). Evolution of the main skeleton-forming genes in sponges (phylum Porifera) with special focus on the marine Haplosclerida (class Demospongiae). Molecular Phylogenetics and Evolution, 131, 245–253.

Chamberlain, S. i Scozs, E. (2013). taxize - taxonomic search and retrieval in R. F1000Research, 2:191 (http://f1000research.com/articles/2-191/v2).

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research, 22(9), 1775–1789.

Dowle, M. i Srinivasan, A. (2020). data.table: Extension of `data.frame`. R package version 1.13.0. (https://CRAN.R-project.org/package=data.table).

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74.

Dunn, C. W., Leys, S. P. i Haddock, S. H. D. (2015). The hidden biology of sponges and ctenophores. Trends in Ecology and Evolution, 30(5), 282–291.

Duret, L., Chureau, C., Samain, S., Weissanbach, J. i Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science, 312(5780), 1653–1655.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. i Finn, R. D. (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432.

Fatima, R., Akhade, V. S., Pal, D. i Rao, S. M. (2015). Long noncoding RNAs in development and cancer: potential biomarkers and therapeutic targets. Molecular and Cellular Therapies, 3(1).

Feng, Y., Hu, X., Zhang, Y., Zhang, D., Li, C. i Zhang, L. (2014). Methods for the Study of Long Noncoding RNA in Cancer Cell Signaling. Methods in Molecular Biology 1165, 115-143.

Feuerhahn, S., Iglesias, N., Panza, A., Porro, A. i Lingner, J. (2010). TERRA biogenesis, turnover and implications for function. FEBS Letters, 584(17), 3812–3818.

Fortunato, S. A. V., Adamski, M., Ramos, O. M., Leininger, S., Liu, J., Ferrier, D. E. K. i Adamska, M. (2014). Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature, 514(7524), 620–623.

Francis, W., Eitel, M., Vargas, S., Adamski, M., Haddock, S., Krebs, S., Blum, H., Erpenbeck, D. i Wörheide, G. (2017). The genome of the contractile demosponge Tethya wilhelma and the evolution of metazoan neural signalling pathways. BioRxiv, 120998.

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research, 47(D1), D766–D773.

Freyhult, E. K., Bollback, J. P. i Gardner, P. P. (2007). Exploring genomic dark matter : A critical assessment of the performance of homology search methods on noncoding RNA. 117–125.

Gaiti, F., Fernandez-Valverde, S. L., Nakanishi, N., Calcino, A. D., Yanai, I., Tanurdzic, M. i Degnan, B. M. (2015). Dynamic and widespread lncRNA expression in a sponge and the origin of animal complexity. Molecular Biology and Evolution, 32(9), 2367–2382.

Garcia-Perez, J. L., Widmann, T. J. i Adams, I. R. (2016). The impact of transposable elements on mammalian development. Development (Cambridge), 143(22), 4101–4114.

Götz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M, J., Robles, M., Talon, M., Dopazo, J. i Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-3435.

Grabherr, M. G. ., Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N. i Friedman, A.

R. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology, 29(7), 644–652.

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L. i Stadler, P. F. (2010). RNAZ 2.0: Improved noncoding RNA detection. Pacific Symposium on Biocomputing 2010, PSB 2010, June 2014, 69–79.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Philip, D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., … Regev, A. (2013). De novo transcript sequence recostruction from RNA-Seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8(8), 1494-1512.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K. i Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature, 495(7441), 384–388.

Harcet, M., Bilandžija, H., Bruvo-Madarić, B. i Ćetković, H. (2010). Taxonomic position of Eunapius subterraneus (Porifera, Spongillidae) inferred from molecular data - A revised classification needed? Molecular Phylogenetics and Evolution, 54(3), 1021–1027.

Hemmrich, G. i Bosch, T, C, G. (2010). BoschCompagen, a comparative genomics platform for early branching metazoan animals reveals early origins of of genes regulating stem cell differentiation. BioEssays 20(10), 1010-1018.

Hölzer, M. i Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience, 8(5), 1–16.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Jarroux, J., Morillon, A. i Pinskaya, M. (2017). Long Non Coding RNA Biology. In Advances in experimental medicine and biology. Advances in Experimental Medicine and Biology, 1008, 1-46.

Johnson, R. i Guigó, R. (2014). The RIDL hypothesis: Transposable elements as functional domains of long noncoding RNAs. Rna, 20(7), 959–976.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D. i Petrov, A. I. (2018). Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research, 46(D1), D335–D342.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L. A., Bourque, G., Yandell, M. i Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genetics, 9(4).

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, R. D., Thomas, K., Presser, A., Bernstein, E. B., Oudenaarden, A., Regev, A., Lander, S. E. i Rinn, L. J. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect

gene expression. Proceedings of the National Academy of Sciences of the United States of America 106(28), 11667-11672.

Lawrence, M., Huber, W., Pag`es, H., Aboyoun, P., … Carlson, M. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8).

Lee, H., Zhang, Z. i Krause, H. M. (2019). Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? Trends in Genetics, 35(12), 892–902.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. i Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079.

Liao, Y., Smyth, G. K. i Shi, W. (2014). Subread manual 1.5.0-p1. Bioinformatics, 30(7), 923–930.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. i Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6(1), 1–14.

Mah, J. L., Christensen-Dalsgaard, K. K. i Leys, S. P. (2014). Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. Evolution and Development, 16(1), 25–37.

Matoničkin, I., Habdija, I i Primc-Habdija, B. (1998). Beskralješnjaci: biologija nižih avertebrata. Školska knjiga, Zagreb, 190-205.

Milligan, M. J. i Lipovich, L. (2015). Pseudogene-derived lncRNAs: Emerging regulators of gene expression. Frontiers in Genetics, 6(FEB), 1–7.

Morgulis, A., Gertz, E. M., Schäffer, A. A. i Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. Journal of Computational Biology, 13(5), 1028–1040.

Nagarajan, N. i Pop, M. (2013). Sequence assembly demystified. Nature Reviews Genetics, 14(3), 157–167.

Nakasugi, K., Crowhurst, R., Bally, J. i Waterhouse, P. (2014). Combining transcriptome assemblies from multiple de novoassemblers in the allo-tetraploid plant nicotiana benthamiana. PLoS ONE, 9(3).

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745.

Ottstein, S. A. G. (2007). The morphological variability, distribution patterns and endangerment in the ogulin cave sponge Eunapius subterraneus Sket & Velikonja, 1984 (Demospongiae), 16(1), 1–17.

Pages, H., Aboyoun, P., Gentleman, R. i DebRoy, S. (2020). Biostrings: Efficient manipulation of biological strings. R package version 2.56.0.

Pages, H. (2020). BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. R package version 1.56.0.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. i Manuel, M. (2009). Phylogenomics Revives Traditional Views on Deep Animal Relationships. Current Biology, 19(8), 706–712.

Ponjavic, J., Ponting, C. P. i Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Research, 17(5), 556–565.

Pérez-Porro, A. R., Navarro-Gómez, D., Uriz, M. J. i Giribet, G. (2013). A NGS approach to the encrusting Mediterranean sponge Crella elegans (Porifera, Demospongiae, Poecilosclerida): Transcriptome sequencing, characterization and overview of the gene expression along three life cycle stages. Molecular Ecology Resources, 13(3), 494–509.

Quénet, D. i Dalal, Y. (2014). A long non-coding RNA is required for targeting centromeric protein A to the human centromere. eLife 12;3:e03254.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (URL https://www.R-project.org/).

Rico, A. J. (2020). gameofthrones: Palettes Inspired in the TV Show "Game of Thrones". R package version 1.0.2. (https://CRAN.R-project.org/package=gameofthrones).

Rico, A. J. (2020). harrypotter: Palettes Generated from All "Harry Potter" Movies. R package version 2.1.1. (https://CRAN.R-project.org/package=harrypotter).

Riesgo, A., Farrar, N., Windsor, P. J., Giribet, G. i Leys, S. P. (2014). The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. Molecular Biology and Evolution, 31(5), 1102–1120.

Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F. i Crespi, M. (2018). Splicing regulation by long noncoding RNAs. Nucleic Acids Research 46(5), 2169-2184.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. i Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research, 41(D1), 590–596.

Salinas-jazmín, N., Hisaki-itaya, E. i Velasco-velázquez, M. A. (2014). Chapter 16 A Flow Cytometry-Based Assay for the Evaluation (ADCC) in Cancer Cells. Cancer Cell Signaling, 1165, 241–252.

Shahryari, A., Jazi, M. S., Samaei, N. M. i Mowla, S. J. (2015). Long non-coding RNA SOX2OT: Expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis. Frontiers in Genetics, 6(JUN), 1–9.

Shen, W., Le, S., Li, Y. i Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLOS ONE 11(10).

Signorell, A (2020). DescTools: Tools for descriptive statistics. R package version 0.99.37. (https://cran.r-project.org/package=DescTools).

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G. i Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Current Biology, 27(7), 958–967.

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M. i Kelly, S. (2016). TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Research 26(8), 1134-1144.

Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F. i Pyle A. M. (2015). HOTAIR forms an intricate and modular secondary structure. Molecular Cell 58(2), 353-361.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. A., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N. H., Stanke, M., Adamska, M., Darling, A., Degnan, S. M., Oakley, T. H., Plachetzki, D. C., Zhai, Y., … Rokhsar, D. S. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. Nature, 466(7307), 720–726.

Tange, O. (2020). GNU Parallel 20200522 ('Kraftwerk'). Zenodo.

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. i Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. Bioinformatics, 31(12), 2032–2034.

Taylor, M. W., Radax, R., Steger, D. i Wagner, M. (2007). Sponge-associated microorganisms: evolution, ecology and biotechnological potential. Microbiology and Molecular Biology Reviews 71, 295-347.

Thompson, J. D., Higgins, D. G. i Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22(22), 4673–4680.

van Soest, R. W. M., Boury-Esnault, N., Vacelet, J., Dohrmann, M., Erpenbeck, D., de Voogd, N. J., Santodomingo, N., Vanhoorne, B., Kelly, M. i Hooper, J. N. A. (2012). Global diversity of sponges (Porifera). PLoS ONE, 7(4).

van Soest, R. W. M., Boury-Esnault, N., Hooper, J. N. A., Rützler, K., de Voogd, N. J., Alvarez, B., Hajdu, E., Pisera, A. B., Manconi, R., Schönberg, C., Klautau, M., Kelly, M., Vacelet, J., Dohrmann, M., Díaz, M. C., Cárdenas, P., Carballo, J. L., Ríos, P., Downey, R. i Morrow, C. C. (2020). World Porifera Database (http://www.marinespecies.org/porifera).

Wheeler, T. J. i Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. Bioinformatics, 29(19), 2487–2489.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. (https://ggplot2.tidyverse.org).

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. (https://CRAN.R-project.org/package=stringr).

Wickham, H. i Lionel, H. (2020). tidyr: Tidy Messy Data. R package version 1.1.1. (https://CRAN.R-project.org/package=tidyr).

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P. A. J., Le Béguec, C., Fieten, H., Johnson, J., Alföldi, J., André, C., … Derrien, T. (2017). FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research, 45(8), 1–12.

Zampetaki, A., Albrecht, A. i Steinhofel, K. (2018). Long non-coding RNA structure and function: Is there a link? Frontiers in Physiology, 9(AUG), 1–8.

Web-stranice

https://mcmanuslab.ucsf.edu/node/251 Posjećeno 1.8.2020.

http://data-science-sequencing.github.io/Win2018/lec tures/lecture7 Posjećeno 2.8.2020.

https://rna.urmc.rochester.edu:81/mathews-lab/bootcamp/wikis/RNA-Secondary-Structure Posjećeno 3.8.2020.

7. Prilozi

Prilog 1 – Popis korištenih genoma i transkriptoma spužvi

Prilog 2 – Programski kod korišten u istraživanju

Prilog 1. Popis korištenih genoma i transkriptoma spužvi

Spužva	Tip podataka	Izvor
Amphimedon queenslandica	Genom	Srivastava i sur. 2010
Aphrocallistes vastus	Transkriptom	Hemmrich i Bosch 2010
Chondrilla nucula	Transkriptom	Riesgo i sur. 2014
Clathrina sp	Transkriptom	Grupa za bioinformatiku (PMF BO, Zagreb)
Cortisium candelabrum	Transkriptom	Riesgo i sur. 2014
Crella elegans	Transkriptom	Porro i sur. 2013
Ephydatia mulleri	Transkriptom	Grupa za bioinformatiku (PMF BO, Zagreb)
Haliclona amboinensis	Transkriptom	Hemmrich i Bosch 2010
Haliclona indistincta	Transkriptom	Camacho i sur. 2019
Haliclona oculata	Transkriptom	Camacho i sur. 2019
Haliclona simulans	Transkriptom	Camacho i sur. 2019
Haliclona tubifera	Transkriptom	Hemmrich i Bosch 2010
Halisarca caerulea	Transkriptom	Hemmrich i Bosch 2010
Ircinia fasciculata	Transkriptom	Riesgo i sur. 2014
Leucosolenia complicata	Transkriptom	Hemmrich i Bosch 2010
Oscarella carmela	Transkriptom	Hemmrich i Bosch 2010
Oscarella sp.	Transkriptom	Hemmrich i Bosch 2010
Petrosia ficiformis	Transkriptom	Riesgo i sur. 2014
Pseudospongosorites subertoides	Transkriptom	Riesgo i sur. 2014
Suberites domuncula	Genom	Grupa za bioinformatiku (PMF BO, Zagreb)
Sycon cilliatum	Genom	Fortunato i sur. 2014
Sycon cilliatum	Transkriptom	Hemmrich i Bosch 2010
Stylissa carteri	Genom	Hemmrich i Bosch 2010
Tethya sp.	Genom	Francis i sur. 2017
Xestospongia testudinaria	Genom	Hemmrich i Bosch 2010

Prilog 2. Programski kod korišten u istraživanju

Programski kod korišten u istraživanju dostupan je na GitHub poveznici:

https://github.com/kbodulic/eunapius_lncRNA?fbclid=IwAR1pH_Lu1Vq00_3CVe9kP45K8NN9ApC-sd806J-j-CpFmgbdnGDf_JikQHs

Životopis

Osnovne informacije:

Ime I prezime: Kristian Bodulić

Adresa: Vrhovec 142, 10000 Zagreb

Datum rođenja: 10.3.1997.

Mjesto rođenja: Erba (Como), Republika Italija

Spol: Muški

Državljanstvo: Hrvatsko

Obrazovanje

2011-2015 Prirodoslovna škola Vladimira Preloga, smjer Prirodoslovna gimnazija

2015-2018 Preddiplomski studij molekularne biologje, Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu

2018-2020 Diplomski studij molekularne biologije, Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu

Radno iskustvo

2018 – 2020 Stručna praksa u Grupi za bioinformatiku Zavoda za molekularnu biologiju Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, pod mentorstvom prof. dr. sc Kristiana Vlahovičeka

Nagrade

2018. Dekanova nagrada za najboljeg studenta Preddiplomskog studija molekularne biologije

2019. Rektorova nagrada za samostalni istraživački rad naslova “Računalna analiza sljedova ogulinske špiljske spužvice (*Eunapius subterraneus* Sket & Velikonja 1984) prikupljenih tehnologijom sekvenciranja nanoporama”, pod mentorstvom prof. dr. sc. Kristiana Vlahovičeka

Ostale informacije

2017. Demonstrator na kolegiju “Osnove fizikalne kemije” na Kemijskom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, prof. dr. sc. Davor Kovačević

2018-2019. Demonstrator na kolegiju “Bioinformatika” na Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, prof. dr. sc. Kristian Vlahoviček

2018-2020 Voditelj Sekcije za bioinformatiku Udruge studenata biologije BIUS. Održavanje brojnih predavanja i radionica s raznovrsnim temama iz područja računalne biologije