Sveučilište Josipa Jurja Strossmayera u Osijeku,

Sveučilište u Dubrovniku,
Institut Ruđer Bošković

Doktorski studij Molekularne bioznanosti

Ivan Pokrovac

Utjecaj okolišnih čimbenika na strukturne varijacije genoma

Doktorski rad

Osijek, 2025

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište Josipa Jurja Strossmayera u Osijeku **Doktorski rad**
Sveučilište u Dubrovniku
Institut Ruđer Bošković
Doktorski studij Molekularne bioznanosti

Znanstveno područje: Interdisciplinarno područje znanosti
Znanstvena polja: biologija

Utjecaj okolišnih čimbenika na strukturne varijacije genoma

Ivan Pokrovac

Doktorski rad je izrađen u: Laboratoriju za evolucijsku genetiku, Zavod za molekularnu biologiju, Institut Ruđer Bošković

Mentor/i: dr.sc. Željka Pezer Sakač

Kratki sažetak doktorskog rada:

Strukturne varijante (SVs) su oblik genetičke varijacije koja utječe na fenotip, pridonoseći biološkoj raznolikosti i bolestima. Ova disertacija istražuje utjecaj okoliša na SVs unutar jedne generacije na mišjem modelu u pokusu prehrane bogate mastima i na mikroevolucijskoj razini kod prirodnih populacija meksičke tetre (*Astyanax mexicanus*) uslijed prilagodbe na život u podzemlju. Genomske analize podataka optičkog mapiranja genoma i sekvenciranja sljedeće generacije otkrile su da okolišni čimbenici pojačavaju postojeću strukturnu varijaciju u genomu, dok selekcija oblikuje funkcionalne prilagodbe, pretežito kroz povećanje broja kopija.

Broj stranica: 178
Broj slika: 64
Broj tablica: 21
Broj literaturnih navoda: 118
Jezik izvornika: hrvatski
Ključne riječi: strukturne varijante, optičko mapiranje genoma, telomere, ekološka prilagodba

Datum javne obrane:

Povjerenstvo za javnu obranu:
1.
2.
3.
4. (zamjena)

Doktorski rad je pohranjen u: Nacionalnoj i sveučilišnoj knjižnici Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb;Gradskoj i sveučilišnoj knjižnici Osijek, Europska avenija 24, Osijek; Sveučilištu Josipa Jurja Strossmayera u Osijeku, Trg sv. Trojstva 3, Osijek

BASIC DOCUMENTATION CARD

Josip Juraj Strossmayer University of Osijek **PhD thesis**
University of Dubrovnik
Ruđer Bošković Institute
Doctoral Study of Molecular biosciences

Scientific Area: Interdisciplinary area of science
Scientific Fields: biology

Impact of environmental factors on genomic structural variation

Ivan Pokrovac

Thesis performed at: Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute

Supervisor/s: dr.sc. Željka Pezer Sakač

Short abstract:

Structural variants (SVs) are a form of genetic variation that affects phenotype, contributing to biological diversity and disease. This dissertation explores environmental impacts on SVs within a single generation in mice fed high-fat diet and, on a microevolutionary scale, in natural populations of the Mexican tetra (*Astyanax mexicanus*) upon adaptation to subterranean environment. Genomic analyses of genome optical mapping and next generation sequencing data revealed that environmental factors amplify variation at sites of existing SVs, and that selection shapes functional adaptations, mainly through copy-number increase.

Number of pages: 178
Number of figures: 64
Number of tables: 21
Number of references: 118
Original in: Croatian
Key words: structural variants, genome optical mapping, telomeres, ecological adaptation

Date of the thesis defense:

Reviewers:
1.
2.
3.
4. (substitute)

Thesis deposited in: National and University Library in Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb; City and University Library of Osijek, Europska avenija 24, Osijek; Josip Juraj Strossmayer University of Osijek, Trg sv. Trojstva 3, Osijek

Doktorska disertacija izrađena je u Laboratoriju za evolucijsku genetiku, Zavoda za molekularnu biologiju, Instituta Ruđer Bošković, pod vodstvom dr. sc. Željke Pezer Sakač u sklopu projekta „Varijacije u broju kopija uzrokovane okolišem u mišjim spermijima“ (HRZZ-UIP-2019-04-7898)

Zahvale

Želim izraziti svoju najdublju zahvalnost svojoj mentorici, dr.sc. Željki Pezer Sakač, na njezinom neprocjenjivom vodstvu u znanosti i mudrim savjetima tijekom ovog rada.

Također, želim zahvaliti svim kolegama iz Laboratorija za evolucijsku genetiku na suradnji i podršci tijekom istraživanja.

Posebno se zahvaljujem Kristini na nezamjenjivoj pomoći pri eksperimentalnom radu s miševima te djelatnicima Pogona laboratorijskih životinja Instituta Ruđer Bošković (IRB) na njihovoj podršci u istraživanju.

Zahvaljujem i kolegama iz Laboratorija za nekodirajuću DNA na velikodušnoj posudbi mnogobrojnih kemikalija i reagensa.

Ovaj rad posvećujem svojoj kćeri Niji čija je pomoć prilikom računalnih analiza (vidi sliku ispod) bila neprocjenjiva, a posebno zahvaljujem svojoj ženi Idi, te puncu Branku i punici Vlasti na njihovoj nesebičnoj podršci.

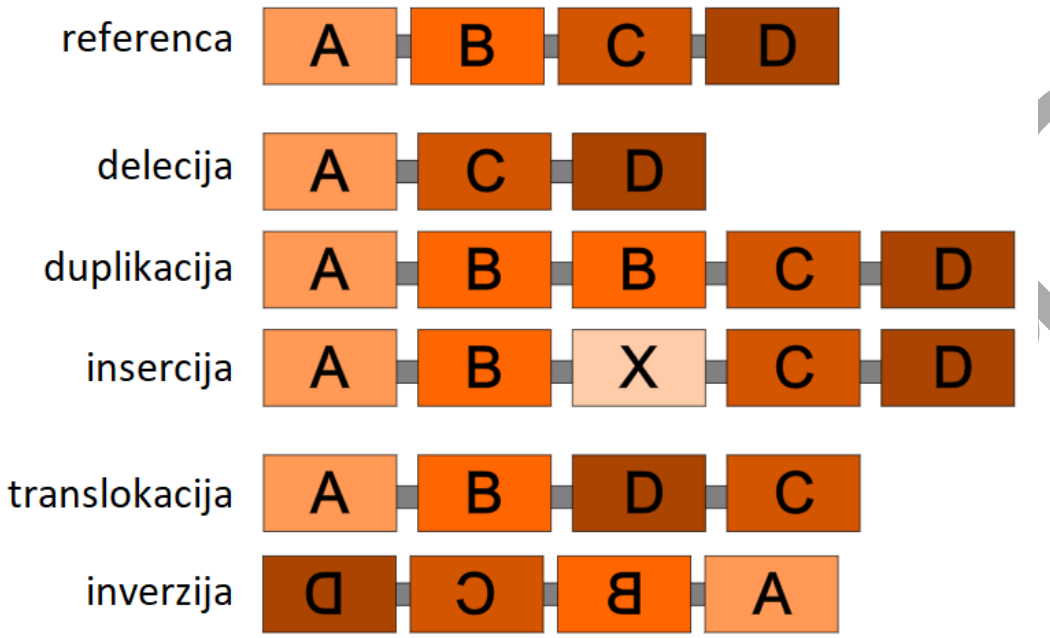
SADRŽAJ

1. UVOD

1.1. Strukturne varijante

Strukturne varijante (eng. *structural variants*, SVs) odnose se na inter-individualne razlike u linearnoj strukturi genoma koje su veće od 50 baznih parova (Alkan i sur., 2011). Iako su SVs manje česti od varijanta jednog nukleotida (eng. *single nucleotide variants,* SNVs), zbog svoje veličine oni utječu na znatno veći udio genoma te doprinose većini genetskih razlika između ljudi u pogledu ukupne količine genetske sekvence. Primjerice, procjenjuje se da je otprilike 13 % ljudskog genoma podložno nekom obliku strukturne varijacije (Sudmant i sur., 2015.), što je skoro dva reda veličine više od količine genoma koju SNVs zauzimaju (Haraksingh i Snyder, 2013.).

SVs se mogu klasificirati na temelju postojanja neto promjene genetičkog materijala na uravnotežene (eng. *balanced*) i neuravnotežene (eng. *unbalanced*) SVs (Slika 1). Uravnotežene SVs su inverzije u kojima dolazi do promjene smjera sekvence, te translokacije, u kojima dolazi do premještanja segmenta genoma na drugu genomsku lokaciju unutar istog (intrakromosomalna translokacija) ili drugog kromosoma (interkromosomalna translokacija). Ove promjene mogu poremetiti funkciju ili regulaciju gena, ali ukupna količina DNA ostaje konstantna. Neuravnotežene SVs rezultiraju ili neto dobitkom ili neto gubitkom genetskog materijala, te se kolektivno nazivaju varijantama broja kopija (eng. *copy-number variants,* CNVs). Odnose se na insercije i duplikacije (pod kojima se često u literaturi podrazumijevaju i višestruke amplifikacije), te delecije genetičkog materijala.



Slika 1. Shematski prikaz glavnih tipova strukturnih varijanti. Strukturne varijante se određuju na osnovi usporedbe s referentnim genomom (referenca).

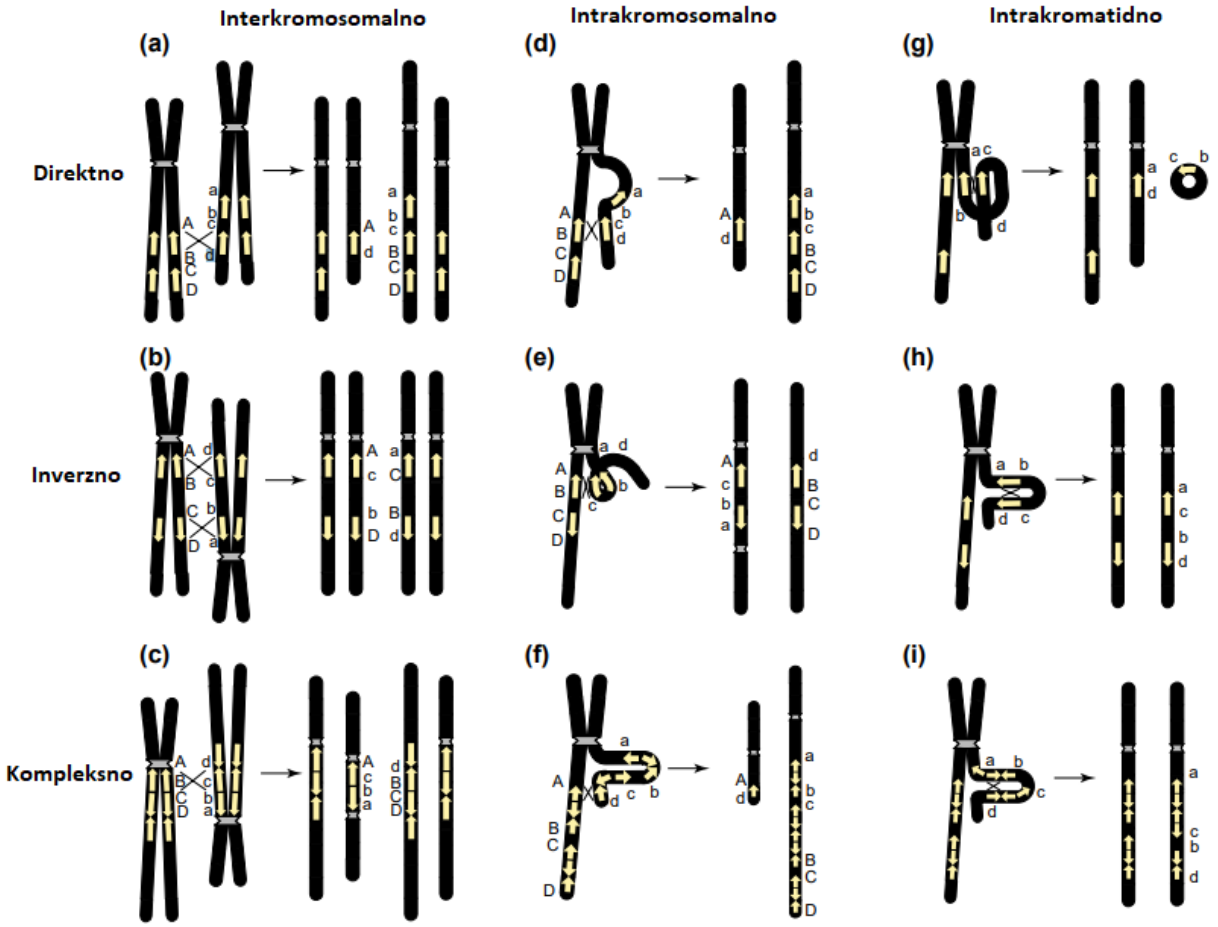
S obzirom na definiciju, SVs obuhvaćaju vrlo široki spektar genetičkih varijacija, od najmanjih kao što su mikrosateliti ili transpozoni (eng. *transposable element,* TE), do velikih genetskih događaja kao što su aneuploidije ili aneusomije koje uključuju promjene u broju kromosoma. U širem smislu, i specijalizirana područja genoma kao što su telomere i centromere se tehnički smatraju strukturnim varijantama zbog varijabilnog broja kopija osnovnih jedinica ponavljanja (Pokrovac i Pezer, 2022).

Ako je određeni SV prisutan u više od 1 % populacije, govorimo o polimorfizmu. Varijanta prisutna samo na jednom od dva homologna kromosoma smatra se heterozigotnom, dok se varijanta prisutna na oba homologna kromosoma smatra homozigotnom (Auton i sur., 2015).

1.2. Mehanizmi nastajanja strukturnih varijanti

Različiti mutacijski mehanizmi mogu dovesti do nastanka strukturne varijante, uključujući procese vezane uz rekombinaciju, replikaciju i popravak DNA (Carvalho i Lupski, 2016). Najčešći mehanizam nastanka SVs je ne-alelna homologna rekombinacija (eng. *non-allelic homologous recombination*, NAHR) koja se javlja kada se dva vrlo slična (homologna), ali ne-alelna, DNA segmenta pogrešno spare tijekom stanične diobe. Unakrsna izmjena genetskog materijala (eng. *cross-over*) koja slijedi između ta dva segmenta rezultira delecijom, duplikacijom ili inverzijom sekvence omeđene homolognim segmentima (Slika 2). Posrednici NAHR su elementi koje postoje u genomu u varijabilnom broju kopija kao što su ponavljanja niskog broja kopija (eng. *low copy repeats,* LCRs). Segmentalne duplikacije (eng. *segmental duplications*, SDs) su česti tip LCRs u ljudskom genomu a definiraju se kao sekvence od otprilike 10 tisuća do 300 tisuća parova baza (eng. *base pairs*, bp) u dužini koje se nalaze na barem dva mjesta u genomu i imaju međusobno visoku stopu (>95 %) sličnosti nukleotidnog slijeda (Sharp i sur., 2005.). Kod ljudi, SDs čine otprilike 5 % ukupne DNA (Eichler, 2001.), dok u mišjem genomu čine oko 1 % (Cheung i sur., 2003.). NAHR posredovana SDs i ostalim LCRs sekvencama je zaslužna za većinu ponavljajućih (eng. *recurrent*) SVs. To su SVs koji se unutar populacije pojavljuju često, na istim genomskim lokusima s identičnim ili gotovo identičnim točkama prekida (eng. *breakpoints*) - njihova stopa mutacije premašuje stopu nukleotidnih supstitucija jedne baze (SNVs) za dva do tri reda veličine (Zhang i sur., 2009).

Ostali mehanizmi nastanka SVs su vezani uz procese popravka dvolančanih lomova i replikaciju: ne-homologno sparivanje krajeva (eng. *non-homologous end joining,* NHEJ), mikrohomologijom posredovana replikacija inducirana prekidom (eng. *microhomology-mediated break-induced replication*, MMBIR) i zaustavljanje replikacijske vilice i promjena predloška (eng. *fork stalling and template switching*, FoSTeS) (Carvalho i Lupski, 2016). Ovi mehanizmi ne zahtijevaju značajnu homologiju susjednih sekvenci te prvenstveno stvaraju ne-rekurentne (eng. *non-recurrent*) SVs s jedinstvenim točkama prekida, i tipično znatno manjom stopom učestalosti u populaciji.



Slika 2. Shema prikaza mehanizama koji dovode do genomskih rearanžmana a temelje se na ponavljanjima niskog broja kopija i ne-alelnoj homolognoj rekombinaciji (LCR/NAHR). Kromosomi su prikazani crnom bojom, dok je centromera prikazana sivo. Žute strelice prikazuju LCRs. Slika prikazuje LCRs raspoređene horizontalno prema orijentaciji i strukturi (direktno, inverzno, kompleksno). Kromosomske pregradnje i predviđeni produkti rekombinacije navedeni su vertikalno prema mehanizmima (interkromosomalno, intrakromosomalno i intrakromatidno). Pogrešno sparivanje kromosoma dovodi do delecije i duplikacije (direktno orijentirani LCR-ovi) (a) i inverzije (inverzna ponavljanja) (b). Intrakromatidna petlja inverznih ponavljanja rezultira inverzijom (h). Nejednaka izmjena za vrijeme interkromosomalnih (c) ili intrakromosomalnih (e) pregradnji može dovesti do invertiranih duplikacija. Intrakromosomalno pogrešno sparivanje direktnih ponavljanja rezultira delecijom i duplikacijom (d). Intrakromatidno pogrešno poravnanje direktnih ponavljanja (g) može rezultirati delecijom i acentričnim fragmentom (segmentom kromosoma bez centromere). Kompleksne LCR mogu posredovati delecijama i duplikacijama (f) ili inverzijama (i). Preuzeto iz Stankiewicz i Lupski, 2002. i prilagođeno.

1.3. Funkcionalne posljedice strukturnih varijanti

Sama priroda varijacije unutar linearne strukture genoma može imati funkcionalne posljedice po organizam kroz utjecaj na gensku funkciju, regulaciju, te stabilnost genoma. Duplikacije ili delecije mogu promijeniti broj kopija gena ili regulatornih elemenata te tako imati izravan utjecaj na ekspresiju gena (Pokrovac i Pezer, 2022). SVs unutar kodirajućih regija mogu promijeniti unutarnju strukturu gena ili spojiti različite gene (eng. *gene fusion).* Nadalje, mogu utjecati na regulaciju gena kroz preuređivanje regulatornih elemenata (Weischenfeldt i sur., 2013). Unatoč tome, brojni polimorfni SVs mogu se pronaći u genomima zdravih pojedinaca a njihovi genomski profili uglavnom odražavaju demografsku sliku populacije, sugerirajući da većina SVs evoluira neutralno (Iskow i sur., 2012.).

U usporedbi sa SVs koje se nalaze u nekodirajućim regijama, SVs koji pogađaju gene su značajno manji i rjeđi u populaciji (Iskow i sur., 2012.; Pezer i sur., 2015.; Hämälä i sur., 2021.). Ovo sugerira da su SVs u funkcionalnim regijama genoma češće pod utjecajem negativne selekcije, kao što su primjerice SVs povezane s bolestima i poremećajima.

1.3.1. Maladaptivne SVs vezane uz bolest

Funkcionalni utjecaj SVs je najviše proučavan u kontekstu ljudskih bolesti i poremećaja. Neki od najpoznatijih primjera vezani su uz velike kromosomalne aberacije, primjerice Downov sindrom, kojeg u većini slučajeva uzrokuje trisomija 21. kromosoma (Weischenfeldt i sur., 2013). Kronična mijeloična leukemija je poznati primjer genetske bolesti uzrokovane recipročnom translokacijom, odnosno razmjenom genetskog materijala između kromosoma 9 i 22. Ova translokacija dovodi do fuzije gena *BCR* i *ABL1* čime nastaje onkogen uzročnik leukemije (Advani i Pendergast, 2002). SVs mogu biti vezane uz monogenske bolesti kao što su Potocki-Lupski sindrom i sindrom duplikacije 7q11.23 ili mogu biti vezane uz kompleksna svojstva kao što su sklonost autizmu, shizofreniji, Parkinsonovoj i Alzheimerovoj bolesti (Zhang i sur., 2009). Većina SVs vezanih uz bolesti su rijetke unutar populacija što sugerira da su pod negativnim evolucijskim pritiskom (Itsara i sur., 2010.).

1.3.2. Adaptivne SVs

Iako neke SVs predstavljaju benignu pozadinsku varijaciju a druge pak mogu imati negativne fenotipske posljedice, pronađeno je da neke SVs imaju potencijalnu korist za organizam.

Primjerice, gen *AMY1*, čiji je produkt amilaza u slini, nalazi se u većem broju kopija kod ljudskih populacija čija je dijeta bogata škrobom (Perry i sur., 2007). Kako je količina amilaze u slini proporcionalna broju kopija gena *AMY1*, smatra se da duplikacije ovog gena predstavljaju prilagodbu na dijete bogate škrobom (Perry i sur., 2007). Delecija u promotorskoj regiji gena *ENO*, čiji je produkt transkripcijski faktor koji regulira veličinu biljnog meristema, dovodi do većeg ploda rajčice, a povećanje ploda jedna je od bitnijih promjena vezanih uz domesticiranje rajčice (Yuste-Lisbona i sur., 2020).

Unatoč raširenosti unutar genoma i utjecaju na fenotip, nejasno je u kojoj mjeri SVs pridonose adaptaciji. Jedna od poteškoća proučavanja utjecaja SVs na adaptaciju i njihove posljedične pozitivne selekcije je to što tradicionalne metrike selekcije (kao što su substitucije amino-kiselina ili testovi temeljeni na učestalosti alela) definirane prema SNP-ovima nisu primjenjive na sve SVs. Razlog tome je što su SVs, za razliku od SNPs, oblik genetičke varijacije koja je unutar sebe vrlo heterogena, s obzirom na veličinu, tip, stopu mutacije, genomsko okruženje i mehanizam nastanka. Stoga većina alata koji se standardno upotrebljavaju za evolucijske analize iz SNP podataka (primjerice filogenetske analize) nije prikladna za SVs jer koristi statističke modele evolucije koji nisu primjenjivi na SVs. Nadalje, dok jednostavne (uglavnom jednostanične) organizme možemo istraživati s pomoću eksperimentalne evolucije, utjecaj selekcije unutar kompleksnih višestaničnih organizama se uglavnom istražuje posredno, prvenstveno na temelju diferencijacije populacija (Iskow i sur., 2012). Stoga prirodni evolucijski eksperimenti, odnosno slučajevi paralelne evolucije predstavljaju sustav u kojem se uloga SVs u adaptaciji može utvrditi s većom pouzdanošću, po principu reproducibilnosti: ako se neki fenotip i njegova genetička osnova ponavljaju u neovisnim populacijama koje su pod istim selekcijski pritiskom, veća je vjerojatnost da je takav fenotip rezultat te prirodne selekcije a ne nasumičnih procesa.

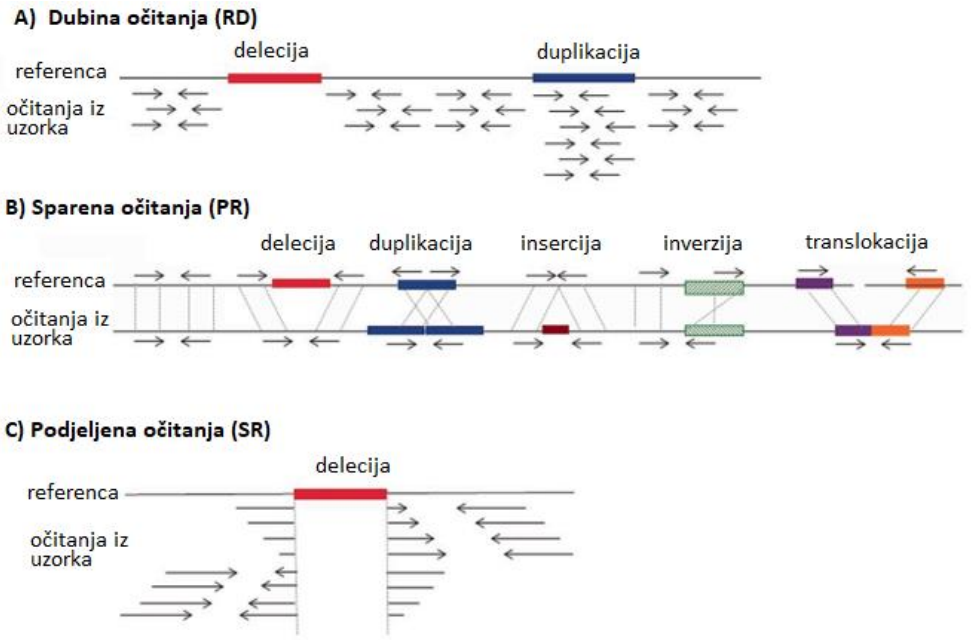
Unatoč očitom potencijalu, SVs su slabo proučeni u kontekstu paralelne evolucije i opetovanih adaptacija različitih populacija na slične uvjete. Jedan sustav koji je iznimka je paralelna evolucija kod koljuške, morske ribe, čije populacije su mnogo puta neovisno kolonizirale slatkovodna staništa. Diferencijacija broja kopija gena vezanih uz imunološki sustav između riječnog i jezerskog ekotipa koljuški doprinosi diferencijaciji ekspresije tih gena između ta dva ekotipa, te je potencijalan izvor genetičke varijacije koja pridonosi opetovanoj adaptaciji na novi okoliš (Huang i sur., 2019). Osim toga, pronađeno je da CNVs pozitivno utječu na kolonizaciju slatkovodnih staništa od strane koljuški koje su slanovodne vrste – smatra se da povećanje broja kopija gena koji sudjeluju u metabolizmu masnih kiselina i hormona štitnjače te gena koji moduliraju imunološki sustav izravno doprinosi prilagodbi životu u slatkoj vodi kao sustavu koji je siromašan nutrijentima i ima drugačiji mikrobiološki profil (Ishikawa i sur., 2019; Ishikawa i sur., 2022).

1.4. Detekcije strukturnih varijanta iz genomskih podataka

Moderne metode za detekciju SVs temelje se na metodama sekvenciranja cijelog genoma (eng. *whole genome sequencing,* WGS) gdje se određuje slijed nukleotida u genomu. WGS se dijeli na tehnologiju sekvenciranja kratkih (eng. *short read sequencing*, SRS) i dugih (eng. *long read sequencing*, LRS) očitanja (eng. *reads*) (De Coster i Van Broeckhoven, 2019). Alternativna metoda za detekciju SVs iz genomskih podataka bazirana je na optičkom mapiranju genoma (eng. *optical genome mapping,* OGM*).*

SRS tehnologija se sastoji od fragmentiranja genoma u 500 do 800 baznih parova duge segmente te paralelno sekvenciranja njihovih krajeva u rasponu od 100 do 250 baznih parova koristeći tehnologije sekvenciranja sparenog kraja. Očitanja se potom mapiraju na referentni genom a prisutnost abnormalnog poravnanja sugerira prisutnost SV. Postoje tri temeljna pristupa za detekciju SV iz rezultata SRS (Slika 3) (Pabinger i sur., 2013):

1. Strategije dubine očitanja (eng. *read-depth*, RD) podrazumijevaju nasumičnu distribuciju očitanja, te regije koje su duplicirane pokazuju veći broj (dubinu) očitanja, a obrnuto vrijedi za delecije.
2. Strategije sparenih očitanja (eng. *paired-reads*, PR) temeljene su na orijentaciji i dužini sparenih očitanja. Razlike u poravnanju parova očitanja s referentnim genomom, poput neočekivanih udaljenosti između parova, nepravilne orijentacije ili mapiranja na različite kromosome, ukazuju na prisutnost određenog tipa SVs u uzorku.
3. Strategije podijeljenog očitanja (eng. *split-read*, SR) detektiraju prisutnost SV iz isprekidanog poravnanja naspram referentnog genoma. Procjep u uzorku znači deleciju, dok procjep u referentnom genomu znači inserciju.



Slika 3. Shematski prikaz 3 glavne strategije za detekciju strukturnih varijanti iz podataka dobivenih sekvenciranjem kratkih očitanja. Preuzeto iz Escaramís i sur., 2015. i prilagođeno.

SRS tehnologiju odlikuje visoka protočnost i niska cijena, ali i nekoliko nedostataka. Prvo, ni jedna od tri strategija detekcija ne može detektirati sve tipove i veličine SVs (Slika 3), te većina algoritama i njihovih implementacija funkcioniraju najbolje za specifične tipove ili čak za određeni raspon veličina SVs (Kosugi i sur., 2019.). Nadalje, stopa stvarno pozitivnih rezultata kod ovih metoda je generalno niska, dok stopa lažno pozitivnih rezultata može biti drastično visoka te su obje mjere značajno ovisne o veličini i tipu SVs (Mahmoud i sur., 2019.). S obzirom na to da su SRS tehnologije temeljene na fragmentima kratke, ali uniformne, duljine, suočavaju se s problemima pri detekciji varijanti u dugačkim repetitivnim regijama. Kratka očitanja iz repetitivnih regija mogu se mapirati na više lokacija, što uzrokuje nepouzdanost rezultata u tim regijama (Mahmoud i sur., 2019.). Nadalje, SRS tehnologije temeljene su na fragmentaciji DNA molekula te amplifikaciji putem metode lančane reakcije polimerazom (eng. *polymerase chain-reaction,* PCR). Fragmentacija narušava genomski integritet, a PCR je inherentno pristrana metoda koja može diferencijalno amplificirati fragmente DNA na temelju njihovog nukleotidnog sastava (Aird i sur., 2011.).

S druge stane, LRS tehnologije generiraju očitanja drastično veće duljine u rasponu nekoliko tisuća parova baza, čuvajući pritom integritet genomskih regija, a veća duljina očitanja dovoljna je da premosti repetitivne elemente u nekim genomskim regijama i time točno identificira SVs unutar ili između repetitivnih regija (Mahmoud i sur., 2019). Mane LRS tehnologije su što su skuplje, zahtijevaju više DNA, te imaju nižu protočnost od SRS metoda. Nadalje, ove tehnologije imaju praktičnu granicu duljine očitanja od otprilike 20 tisuća parova baza (Jain i sur., 2018.) što je i dalje nedovoljno da premosti određene duge repetitivne regije kao što su pericentromerne, centromerne, i akrocentrične krakove kromosoma koji sadrže ponavljanja duljine između 3 i 10 milijuna parova baza (Wevrick i Willard, 1989.; Eichler i sur., 2004.).

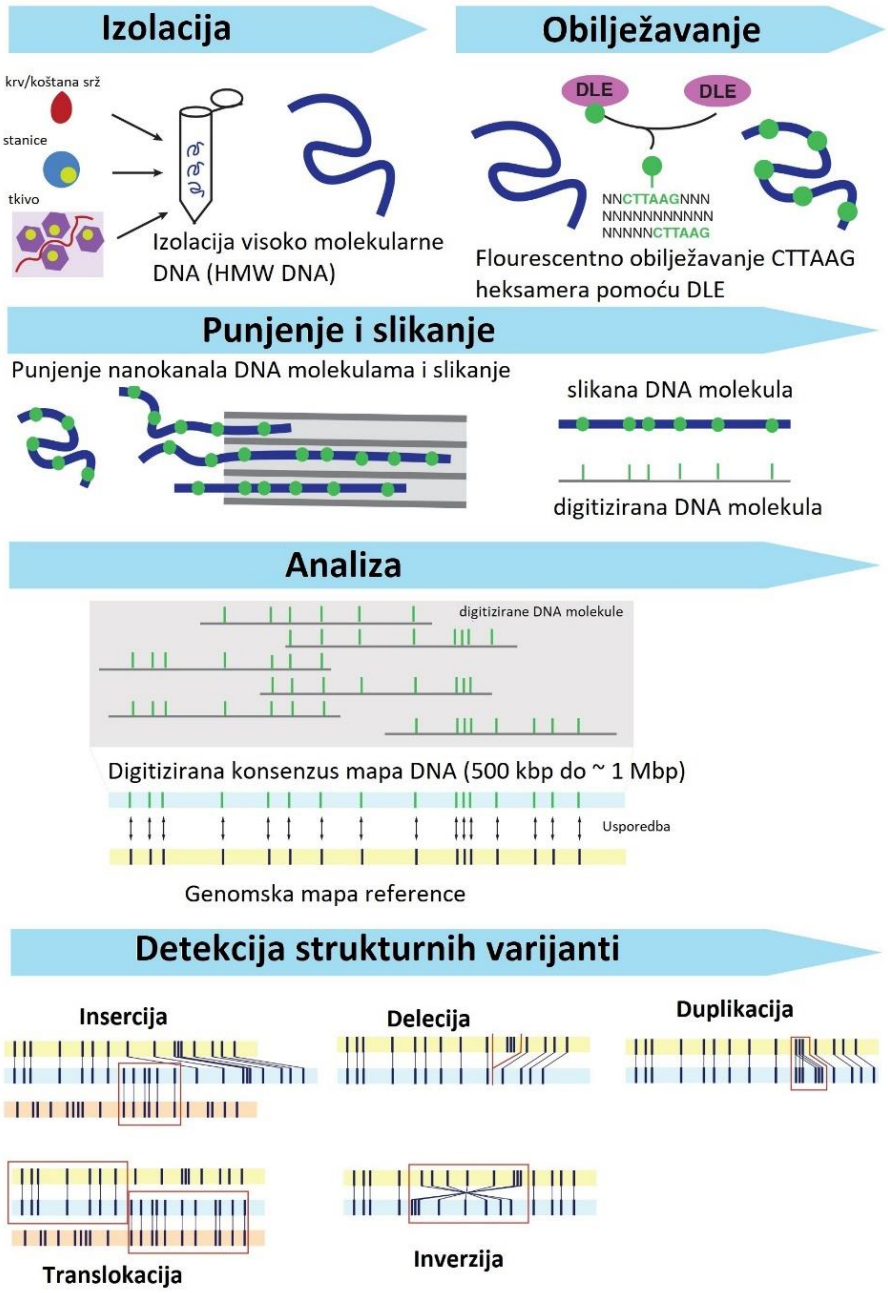
1.4.1. Optičko mapiranje genoma

Optičko mapiranje genoma je alternativna visokoprotočna tehnologija analize genoma koja ne rezultira određivanjem slijeda nukleotida nego genomskom mapom. Visokomolekularna DNA (eng. *high molecular weight* DNA*,* HMW DNA) fluorescencijski se obilježava pomoću enzima DLE-1 na motivu CTTAAG, koji se u genomu ponavlja u prosjeku svakih 10-15 tisuća parova baza (eng. *kilobase pairs*, kbp). Duge molekule DNA rastežu se i lineariziraju unutar nanokanala s pomoću električnog polja te se Slikaju svjetlosnim mikroskopom visoke razlučivosti pri čemu se detektiraju fluorescentni signali na obilježenim (eng. *label*) mjestima (Yuan i sur., 2020).

Slikane molekule se digitaliziraju, te međusobnim usporedbama između različitih molekula stvaraju se detaljne „konsenzus karte“ (eng. *consensus maps*, CMAP) genomskih regija koje se potom uspoređuju s *in-silico* obrađenim referentnim genomom, te se postojanje strukturne varijacije određuje na temelju neslaganja unutar usporedbi s navedenim „kartama“ i referencom (Slika 4).

Molekule DNA koje sačinjavaju složene karte (mape) genoma imaju prosječnu duljinu od oko 200 kbp, što je red veličine više od tehnologija LRS (Dremsek i sur., 2021) te se genomske mape nastale slaganjem ovih molekula mogu prostirati cijelim kromosomom, što omogućava razrješavanje SV unutar repetitivnih i drugih kompleksnih regija (Lam i sur., 2012). Nadalje, OGM ima sposobnost slaganja genoma koristeći informacije o haplotipu (eng. *haplotype-aware assembly*), proceduru koja eksplicitno uzima u obzir činjenicu da većina diploidnih organizama imaju dvije kopije (haplotipa) svakog kromosoma. Ovo omogućava veću rezoluciju u identifikaciji SVs jer ih povezuje sa specifičnim haplotipom, time određujući je li određeni SVs homozigot ili heterozigot.

OGM spada u tehnologije jedne molekule (eng. *single-molecule technologies*) koje omogućavaju direktno proučavanje pojedinačnih DNA molekula bez potrebe za fragmentacijom ili amplifikacijom što smanjuje potencijalne greške vezane uz ove postupke. Unatoč tome, OGM ne nudi informacije o nukleotidnom sastavu, te je zbog prirode svjetlosne mikroskopije ograničen na detekciju SVs iznad granice od ~1500 bp, a svi detektirani događaji ispod te granice smatraju se slabo pouzdanim (Yuan i sur., 2020).



Slika 4. Shematski prikaz optičkog mapiranja genoma. Preuzeto iz Brody i sur., 2023 i prilagođeno. Visoko molekularna DNA se izolira, te se fluorescentno obilježava s pomoću enzima DLE-1 (eng. *Direct Label Enzyme*) na specifičnom heksameru CTTAAG. Obilježena DNA nanosi se na mikročipove koji se sastoje od nanokanala. Primjenom električnog polja, DNA se linearizira i prolazi kroz nanokanale, pri čemu se fotografira. Ovako fotografirana molekula se digitalizira, te se više digitaliziranih molekula slaže u konsenzus mapu. Usporedbom s genomskom mapom reference detektiraju se strukturne varijante.

1.5. Istraživanje utjecaja okolišnih čimbenika

Nekolicina istraživanja direktno povezuju određene okolišne stresore s mutacijama SV u višestaničnim organizmima. Primjerice, zagađenja zraka i dim duhana povećavaju stope mutacije mikrosatelitnog lokusa kod miševa (Marchetti i sur., 2011; Somers i sur., 2002), a ozračivanje muških miševa tijekom spermatogeneze dovodi do veće frekvencije SVs i indela u njihovom potomstvu (Adewoye i sur., 2015).

Adaptivne promjene se također mogu dogoditi, iako se dokazi zasad svode na jednostavne organizme. Primjerice, kolonije pivskog kvasca izložene visokoj razini bakra razvijaju veći broj kopija gena *CUP1* koji je odgovoran za rezistenciju na bakar (Hull i sur., 2017). Do ove varijacije broja kopija dolazi zbog nestabilnosti genoma na mjestu transkripcije uslijed pojačane ekspresije *CUP1* gena, odnosno zbog kolizije procesa transkripcije i replikacije. Iako sličan oblik usmjerene mutageneze zasad još nije dokazan kod kompleksnijih organizama, sve molekularne komponente ovog adaptivnog mehanizma su evolucijski sačuvane i u višestaničnim eukariotima (Hull i sur., 2017).

Utjecaj okolišnih čimbenika na formiranje i učestalost SVs se uglavnom istražuje posredno, s pomoću komparativne i populacijske genomike, te s pomoću uvida dobivenih iz obiteljskih studija (Pokrovac i Pezer, 2022). U prvom slučaju, istraživanja na makro- i mikro-evolucijskim razinama su ograničena na varijante koje su fiksirane ili su postigle relativno visoku učestalost u populaciji, čime se indirektno zaključuje o učinku okoliša. Obiteljske studije pak, iako pružaju mogućnost istraživanja rijetkih i *de novo* varijanti, zahtijevaju veći broj jedinki odnosno obitelji, i prvenstveno se provode u kontekstu bolesti, te su stoga ograničene na proučavanje maladaptivnih SVs. U oba slučaja, analize se provode nad diploidnim genetičkim materijalom, pri čemu informacije o pojedinačnim alelima često ostaju nepoznate.

1.6. Cilj rada

Cilj ovog rada je istražiti utjecaj okolišnih čimbenika na učestalost i pojavnost strukturnih varijanti na dvije vremenske razine:

1. Unutar jedne generacije, koristeći srođeni laboratorijski soj C57BL/6 kućnog miša (*Mus musculus*) u pokusu prehrane bogate mastima, kao primjer okolišnog čimbenika koji uzrokuje izraženu promjenu fenotipa.
2. Na mikroevolucijskoj razini, koristeći genomske podatke prirodnih populacija meksičke tetre (*Astyanax mexicanus*), kao model paralelne prilagodbe na život u špiljama gdje su selekcijski pritisci snažni i dobro poznati.

1.6.1. Hipoteze

1. Prehrana bogata masnoćama utječe na strukturne varijacije u genomu miša unutar jedne generacije.
2. Razlika u okolišnim uvjetima između površinskih i špiljskih staništa odražava se u razlikama u učestalosti varijanti broja kopija kod ribe *Astyanax mexicanus*, te varijante broja kopija doprinose prilagodbi na podzemne uvjete života.

2. MATERIJALI I METODE

2.1. Istraživanje unutar jedne generacije na modelu miša (*Mus musculus*)

U ovom istraživanju, koristili smo mišji model za istraživanje izravnog utjecaja okolišnih čimbenika unutar jedne generacije. Kućni miš (*Mus musculus*) jedan je od najkorištenijih znanstvenih i medicinskih modela. Laboratorijski soj C57BL/6 korišten u ovom istraživanju je srođeni soj što znači da je očekivana homozigotnost na razini jedinke visoka, te je genetska varijacija između različitih jedinki minimalna zbog čega očekujemo veću moć detekcije utjecaja okoliša. Nadalje, referentni genom miša koji je korišten u ovom istraživanju temelji se upravo na ovom soju, a odlikuje ga visoki kontinuitet i dobra anotiranost, što su osnovni preduvjeti za kvalitetne genomske analize visoke protočnosti i rezolucije.

U svrhu ovog istraživanja, usredotočili smo se na genome spermija, kao prikladnom tkivu za istraživanje utjecaja okoliša na genetičke varijacije. Kao prvo, spermiji sadrže genetski materijal koji sljedeća generacija izravno nasljeđuje. Drugo, spermiji su haploidne stanice što znači da je u populaciji stanica moguće izravno odrediti frekvenciju pojedinog alela (jedan alel po stanici). Treće, zreli spermiji nastaju procesom spermatogeneze iz diploidnih spermatogonija, zametnih stanica muških gameta. Ovaj proces uključuje niz staničnih dioba i rekombinacijske događaje - pokretače nastanka strukturnih varijacija.

Nadalje, genomi spermija reaktivni su na okoliš, pri čemu epigenetičke promjene uzrokovane životnim čimbenicima poput prehrane, stresa, i konzumacije alkohola utječu na fenotip potomstva. Primjerice, pokazano je da prehrana bogata mastima kod štakora dovodi do reprogramiranja epigenoma spermija i promjena u metabolizmu potomaka (de Castro Barbosa i sur., 2016.). Kod miševa konzumacija alkohola mijenja brojnost i sastav malih RNA u spermijima (Rompala i sur., 2018.) te utječe na stres, osjetljivost, i sklonost potomstva prema alkoholu (Finegersh i Homanics, 2014.), a olfaktorno iskustvo očeva utječe na ponašanje i neuronsku strukturu u sljedećim generacijama (Dias i Ressler, 2013.).

Tkivo bubrega korišteno je kao primjer somatskog tkiva, kako bismo usporedbom unutar istog miša mogli detektirati promjene u strukturnim varijacijama u stanicama spermija.

2.1.1. Pokus prehrane bogate mastima

Obavljanje pokusa na pokusnim životinjama odobreno je u sklopu znanstveno-istraživačkog projekta „Varijacije u broju kopija uzrokovane okolišem u mišjim spermijima“ rješenjem dobivenim od Uprave za veterinarstvo Ministarstva poljoprivrede, šumarstva, i vodnog gospodarstva Republike Hrvatske (Klasa: UP/I-322-01/19-01/59, Urbroj: 525-10/0543-20-4). Provedeni pokusi bili su u skladu s bioetičkim standardima o provođenju pokusa na pokusnim životinjama te u skladu s osnovama 3R principa.

Miševi su do pokusa održavani u pogonu laboratorijskih životinja Instituta Ruđer Bošković pod standardnim laboratorijskim uvjetima – 12 satni ciklus svjetlo-tama, temperature od 22 °C te vlažnosti zraka do 60 %. Miševi su držani u kavezima u grupama od tri miša. Promjena kaveza („mijenjanje stelje“) obavljana je jednom tjedno.

Mužjaci C57BL/6 soja starosti 4-5 tjedana (što odgovara pretpubertetnoj fazi) nasumično su odabrani za pokus. Šest miševa je u kontrolnoj skupini hranjeno standardnom laboratorijskom prehranom, a šest miševa u pokusnoj skupini hranjeni su tzv. zapadnjačkom hranom (eng. *Western Diet*, WD) koja se odlikuje visokim udjelom masti, šećera i kolesterola (Tablica 1). Pokus je trajao 80 dana što je vrijeme trajanja jednog do dva potpuna ciklusa spermatogeneze (Oakberg, 1957.). Hrana i voda bili su dostupni miševima *ad libitum* u obje skupine.

Tablica 1. Udio pojedinih komponenti u sastavu kontrolne i zapadnjačke hrane.

	KONTROLNA HRANA (4RF21 MUCEDOLA)	**WESTERN DIET (EF TD88137)**
PROTEINI	18.5 %	17.3 %
MASTI	3.0 %	21.1 %
VLAKNA	6.0 %	5.0 %
PEPEO	7.0 %	4.2 %
ŠKROB	42.6 %	14.4 %
SAHAROZA	3.7 %	34.3 %
KOLESTEROL	-	0.15 %

2.1.2. Izolacija spermija i bubrega

Nakon završenog pokusa u trajanju od 80 dana, miševi su eutanazirani cervikalnom dislokacijom. Iz miševa su izolirani bubrezi, te rep epididimisa (*cauda epididymis*). Bubrezi su segmentirani u 2-3 komada što odgovara ~30 mg po segmentu te smrznuti direktnim uranjanjem u tekući dušik, te su pohranjeni na -80 °C do trenutka analize. Rep epididimisa prebačen je u Eppendorf tubicu sa 150 µL DPBS pufera.

Rep epididimisa postavljen je na čistu podlogu parafilma, te očišćen od adipoznog tkiva. Epididimis je probušen oko 10 puta iglom (16-21G), i nakon što je viskozni sadržaj iscurio, prebačen je natrag u istu Eppendorf tubicu.

Tuba je dopunjena s 850 µL DPBS-a do ukupnog volumena ~1000 µL DPBS-a, horizontalno položena u hibridizacijsku pećnicu te inkubirana 1 sat pri 37 °C na 6 rotacija po minuti (RPM).

Nakon inkubacije, suspenzija je filtrirana kroz filter veličine 40 µm, te je tuba isprana s još 500 µL DPBS-a koji je potom ponovno filtriran. Filtrati su objedinjeni i sakupljeni.

10 µL filtrata prebačeno je u Neubarovu komoru u svrhu kvantifikacije spermija. Koncentracija spermija po mikrolitru suspenzije određena je prema formuli:

$$Koncentracija\ spermija\ \left(\frac{br.st.}{\mu L}\right) = \frac{Broj\ izbrojanih\ stanica\ u\ kvadratu\ \times 10^3}{4}$$

Gdje se kao kvadrat uzima jedan od 25 kvadrata koji čine središnji kvadrat Neubarove komore. Konačna koncentracija spermija izračunata je kao prosječna vrijednost koncentracije spermija izračunata iz pet nasumično odabranih kvadrata.

Filtrat je centrifugiran 10 minuta pri 1000 g na 4 °C, te je supernatant odbačen. Preostali talog u Eppendorf tubici uronjen je izravno u tekući dušik te pohranjen na -80 °C do trenutka analize.

2.1.3. Izolacija visoko-molekularne DNA (HMW DNA) iz spermija

Budući da je komercijalni protokol izolacije genetičkog materijala za OGM specifično namijenjen za somatsko tkivo, bilo je potrebno prilagoditi postupak kako bi se mogao primijeniti na stanice spermija. Ključni dio prilagođenog protokola je dodatak ditiotreitola (DTT) koji djeluje kao reducens. Reducens je potreban za razbijanje disulfidne veze u protaminima koji su specifično prisutni u kromatinu spermija (Balhorn, 2007.). Nakon izolacije visoko-molekularne DNA (eng. *high-molecular weight DNA*, HMW DNA), dio izolata se razgrađuje restrikcijskim enzimom, kako bi se ustanovila dostupnost izolirane DNA, odnosno učinkovitost uklanjanja protamina. Uspješnost razgradnje DNA i provjera kvalitete izolirane DNA vrši se gel-elektroforezom u pulsirajućem polju (eng. *pulsed-field gel electrophoresis*, PFGE).

2.1.3.1. Priprema agaroznog bloka

Agaroza niske točke tališta (LMP agaroza, eng. *low melting point agarose*) je odvagana te suspendirana u DPBS puferu do konačne masene (w/v) koncentracije od 2 % (2 mg LMP agaroze po 100 µL DPBS). Agaroza je otopljena u kipućoj vodenoj kupelji te prebačena u vodenu kupelj na temperaturu od 45 °C.

Talozi stanica spermija suspendirani su u količini DPBS-a tako da je u 40 µL DPBS-a između 3 i 4 milijuna stanica, te je suspenzija prebačena u vodenu kupelj na temperaturu od 45 °C.

Nakon kratkog vremena, suspenzija spermija i otopina LMP agaroze je pomiješana, resuspendirana nekoliko puta pipetom u svrhu ravnomjernog miješanja, te je 80 µL alikvotirano u BioRad kalup(e) – ukupna količina stanica spermija po agaroznom bloku iznosila je između 3 i 4 milijuna stanica. Kalupi su prebačeni u hladnjak na temperaturu 2-8 °C te ostavljeni 20-30 minuta da se agarozni blokovi stvrdnu.

2.1.3.2. Razgradnja stanica u agaroznom bloku

Nakon što su se agarozni blokovi skrutili, prebačeni su u Eppendorf tubice (jedan agarozni blok po tubi) te je dodano 500 µL liznog pufera (Tablica 6) što odgovara ~10 puta većem volumenu bloka.

Eppendorf tubice s liznim puferom postavljene su horizontalno u hibridizacijsku pećnicu, te su inkubirane 16 sati na 50 °C (radna temperatura proteinaze K) na 3 rotacije po minuti (RPM).

Nakon 16-satne inkubacije, dodano je još 50 µL proteinaze K (koncentracije 20 mg/mL, odnosno 1 mg po bloku), te se smjesa inkubirala dodatnih sat vremena u istim uvjetima.

Agarozni blokovi ispirali su se s ~10 puta njihovog volumena 30 minuta na ledu sa svakim od navedenih pufera, koristeći minimalnu agitaciju na orbitalnoj tresilici (100 okretaja po minuti):

- Jedan puta ledenom 50 mM EDTA (pH 8).
- Jedan puta ledenim TE 1 X (pH 8).
- Tri puta ledenim TE 1 X (pH 8) 0.1 mM PMSF (1 µL 100mM PMSF otopine u izopropanolu po 1mL pufera TE 1 X).
- Tri puta ledenim TE 1 X (pH 8).

Nakon zadnjeg ispiranja, agarozni blokovi su prebačeni u „Wash Buffer“ (Tablica 6) i čuvani na temperaturi 2-8 °C.

2.1.3.3. Restrikcijska razgradnja

Polovica agaroznog bloka (~40 µL volumena) je prebačena u Eppendorf tubu koja sadrži 100 µL 1 X pufera za restrikcijsku digestiju (bez enzima), te inkubirana 15 minuta pri sobnoj temperaturi. Agarozni blok je zatim uklonjen iz tubice i prebačen u novu Eppendorf tubu koja sadrži 100 µL 1 X pufera za restrikcijsku digestiju s 20 U restrikcijskog enzima EcoRI te je smjesa inkubirana 3-4 sata na 37° C.

2.1.3.4. Gel-elektroforeza u pulsirajućem polju

Pripremljen je 1 % agarozni gel u 0.5 X TBE puferu suspendiranjem 1.5 g agaroze po 150mL u Erlenmayerovoj tikvici. Tikvica je prekrivena komadom papira, te grijana u mikrovalnoj pećnici do točke vrenja, nakon čega je izvađena iz mikrovalne pećnice te nježno promiješana. Postupak je ponovljen sve dok otopina ne postane prozirna. Ovakva otopina agaroze stavljena je u vodenu kupelj zagrijanu na 50 °C 15 do 20 minuta. Otopina agaroze je zatim izlivena na BioRad kalup, te ostavljena na sobnoj temperaturi 30 minuta dok se ne stvrdne.

Agarozni blokovi uklonjeni su iz pufera, postavljeni na čist komad parafilma, a suvišak pufera je uklonjen pipetiranjem. Agarozni blokovi umetnuti su u jažice agaroznog gela. Kao marker veličine DNA korišten je lambda marker. Ispunjene jažice zatvorene su s 1 % otopinom LMP agaroze ohlađenom na 50 °C, te se gel hladio 10 do 15 minuta.

CHEF (eng. *clamped homogeneous electric field*) sustav za PFGE ispran je s 2 L destilirane vode. Dodano je ~2.5 L 0.5 X TBE pufera te je pokrenuta cirkulacija pufera sa sustavom hlađenja. Kada je temperatura dostigla 14 °C, gel s uzorcima je postavljen u centralnu komoru CHEF sustava. Elektroforeza je pokrenuta prema postavkama u Tablici 2.

Tablica 2. Korištene postavke PFGE na CHEF III uređaju.

Switch time	5/50 s
Run time	16 h
Volts/cm	6
Included angle	120°
Temperature	14 °C

Gel je obojen uranjanjem u ~0.5 L 0.5X TBE pufera s 1 µg/mL etidij bromida te je nakon otprilike pola sata izvađen iz otopine te vizualiziran na SynGene Transiluminatoru.

2.1.4. Optičko mapiranje genoma i bioinformatičke analize

Uzorci tkiva bubrega i HMW DNA iz spermija poslani su na Institut za primijenjene biotehnologije (Institute of Applied Biotechnologies, IAB, Češka Republika) koji je provodio OGM na platformi Saphyr (Bionano Genomics, BNG). Otprilike 30 mg bubrežnog tkiva alikvotirano je u zasebne kriotube, te poslano na suhom ledu. HMW DNA je izolirana iz otprilike 3 milijuna stanica spermija, te je agarozni blok s HMW DNA umetnut u Eppendorf tubu od 5mL. Eppendorf tuba je dopunjena do vrha Wash Bufferom, te je zatim začepljena i zabrtvljena parafilmom. Eppenderf tube s HMW DNA iz spermija poslane su na sobnoj temperaturi.

Optičko mapiranje genoma provedeno je na ukupno devet miševa: pet iz pokusne skupine te četiri miša iz kontrolne skupine. Iz svakog od ovih miševa OGM je provedeno na HMW DNA izoliranoj iz spermija i tkiva bubrega.

Izolacija HMW DNA iz uzoraka bubrega vršena je od strane IAB prema BNG protokolu za izolaciju iz tkiva i tumora („*SP Tissue and Tumor DNA Isolation Protocol*“ pod brojem dokumenta 30339 na mrežnim stranicama kompanije Bionano Genomics). Enzimsko obilježavanje HMW DNA iz bubrega i spermija vršeno je u IAB prema BNG protokolu opisanom u dokumentu „*Bionano Prep Direct Label and Stain (DLS) Protocol*“ pod brojem dokumenta 30206.

Kao mjera početne kvalitete DNA uzeta je koncentracija DNA nakon DLS obilježavanja i koeficijent varijacije koncentracije DNA računat iz tri mjerenja.

Inicijalnu obradu podataka, uključujući *de novo* slaganje genoma i detekciju SVs provodio je IAB koristeći skup bioinformatičkih alata BioNano Solve, prema BNG smjernicama „*Guidelines for Running Bionano Solve on the Command Line*“ pod brojem dokumenta CG-30205.

2.1.4.1. Procjena kvalitete optičkog mapiranja

Prilikom bioinformatičkog procesa *de novo* slaganja genoma (eng. *de novo genome assembly*) i detekcije SVs, molekule koje ulaze u proces se filtriraju kako bi se uklonile molekule koje se smatraju nisko kvalitetnima. Niskokvalitetne molekule su one ispod 150 kbp duljine te koje sadrže manje od 9 oznaka.

Kako bismo procijenili kvalitetu HMW DNA koje ulaze u proces optičkog mapiranja koristili smo nekoliko pokazatelja, navedenih u Tablici 3 uz preporučene referentne vrijednosti za genom čovjeka, jer preporuke, odnosno referentne vrijednosti za mišji genom trenutno ne postoje. Svi pokazatelji, što uključuje i njihove referentne vrijednosti, opisani su u BioNano dokumentu „*De Novo Assembly Informatics Report Guidelines*" pod brojem dokumenta 30255. Osim navedenih pokazatelja, pratili smo broj te prosječnu duljinu molekula prije i poslije filtriranja.

Tablica 3. Pokazatelji kvalitete OGM.

Ime pokazatelja	**Referentna vrijednost**	**Objašnjenje pokazatelja**
Prosječna duljina filtriranih molekula /kbp	> 230	-
Gustoća oznaka po 100 kbp	14 - 17	Prosječan broj oznaka detektiranih po 100 kbp duljine molekula
Pokrivenost reference prije poravnanja /X	> 100	Pokrivenost reference prije poravnavanja računa se se dijeljenjem ukupne duljine filtriranih molekula sa ukupnom dužinom reference
Udio molekula poravnatih na referencu	> 0.6 minimalan > 0.80 optimalan	Udio filtriranih molekula koje su poravnate na referencu
Pokrivenost reference poslije poravnanja /X	> 70	Pokrivenost reference poslije poravnavanja računa se dijeljenjem ukupne duljine poravnatih molekula sa ukupnom dužinom reference
N50 diploidnih genomskih mapa / Mbp	> 50	N50 diploidne mape genoma označava duljinu sklopljenih mapa genoma pri kojoj je 50 % ukupne duljine mapa genoma sadržano u mapama genoma koje su jednake ili dulje od te vrijednosti

2.1.4.2. Analiza strukturnih varijanta

Konačni rezultat detekcije SVs putem BioNano Solve alata zapisan je u SMAP datoteci. SMAP datoteka sadrži listu strukturnih varijanta detektiranih između genomskih mapa uzoraka i referentne genomske mape. Datoteka se sastoji od dvije sekcije – zaglavlja koje opisuje verziju datoteke te lokaciju referentnih odnosno genomskih mapa uzoraka, te informativni blok strukturnih varijanta u TSV obliku. Struktura datoteke opisana je u BioNano dokumentu „*SMAP File Format Specification Sheet*“ pod brojem dokumenta 30041.

U sklopu ovog istraživanja, koristili smo određene informacije sadržane unutar stupaca SMAP datoteke (Tablica 4). Pratili smo pet tipova SVs – inverzije, insercije, duplikacije (što uključuje i invertirane duplikacije) te intrakromosomske i interkromosomske translokacije. Potonje su definirane koordinatama koje se nalaze na dva referentna kontiga („RefcontigID1“ i “RefContigID2“), što odgovara dvama kromosomima.

Kao procjenu zigotnosti koristili smo stupac „VAF“ (eng. *variant allele frequency*). Vrijednosti VAF-a su u rasponu od 0 do 1 te predstavljaju udio molekula koje podržavaju određenu strukturnu varijantu u odnosu na ukupan broj molekula koje pokrivaju istu regiju. Sukladno BioNano dokumentu „*Theory of Operation: Structural Variant Calling*“ pod brojem 30110 sve SVs ispod 0.97 smatrane su heterozigotnima, a sve iznad te vrijednosti su smatrane homozigotnima. U slučaju spermija koji su haploidni, pojam zigotnosti se u ovom radu koristi uvjetno, odnosno isključivo operativno radi praćenja učestalosti varijanti u populaciji stanica unutar pojedinog uzorka.

Za funkcionalne analize koristili smo stupac „OverlapGenes“ koji sadržava imena gena (genske simbole) koji se preklapaju s detektiranim varijantama.

Unutar svake pojedine SMAP datoteke isti SV (s obzirom na tip, veličinu i položaj na referentnom genomu) može biti naveden više puta što proizlazi iz procesa detekcije SVs, u kojem se SV identificira na svakom kontigu. Dva ili više kontiga mogu biti poravnati na istu regiju referentnog genoma, ako predstavljaju različite haplotipove te regije. Međutim, između različitih haplotipova postoje dijelovi kontiga koji su identični i ako se strukturna varijanta nalazi u tom dijelu, ona će biti identificirana na svakom od kontiga koji predstavlja pojedini haplotip. Kako bismo uklonili ovakve duplikate u rezultatima, grupirali smo SVs po SMAP datoteci, odnosno biouzorku iz kojeg proizlaze, te po lokusu i tipu. Ukoliko su SVs istog tipa i sa identičnim koordinatama na referentnom genomu varirali veličinom unutar 1 kbp, spojene su u jednu strukturnu varijantu te je izračunat prosjek njihove veličine i VAF-a. Kriterij razlike u veličini odabran je arbitrarno, a uzima u obzir moć rezolucije tehnologije OGM.

Tablica 4. Pregled korištenih stupaca SMAP datoteke i njihova objašnjenja.

Stupac SMAP datoteke	Objašnjenje
SmapEntryID	Jedinstven ID dodijeljen pojedinoj SV
RefcontigID1	ID prvog referentnog kontiga (kromosoma)
RefStartPos	Koordinata početka SV na referentnom kontigu (kromosomu)
RefEndPos	Koordinata kraja SV na referentnom kontigu (kromosomu)
Type	Tip SV
SVsize	Veličina SV, nije definirana za translokacije
VAF	Frekvencija učestalosti alela SV
OverlapGenes	Geni koji se preklapaju sa SVs

Analizirali smo SVs po uzorku, s obzirom na broj, veličinu, omjer tipova, učestalosti alela, te udjela genoma kojeg obuhvaćaju.

Sličnost uzoraka smo analizirali prema broju zajedničkih SVs između dva uzoraka – dva uzorka imaju isti SV ako se oni na referentnom genomu recipročno preklapaju s više od 50 % duljine. Zbog ovakve definicije i postojanja različitih haplotipova u istom uzorku, u istoj usporedbi dva uzorka, broj zajedničkih SVs u jednom uzorku nije nužno isti broju zajedničkih SVs u drugom uzorku (primjerice jedan SV u jednom uzorku može po ovakvoj definiciji imati dva zajednička SVs u drugom uzorku). Zbog toga se broj zajedničkih SVs normalizira u raspon od 0 do 1 tako što se dijeli prosječnim brojem SVs zajedničkih SVs izmedu ta dva uzorka. Primjenom ove metrike nad skupom uzoraka dobivamo matricu sličnosti koju podvrgavamo hijerarhijsko-aglomerativnom grupiranju (eng. *hierarchical agglomerative clustering,* HAC) koristeći Wardovu metodu.

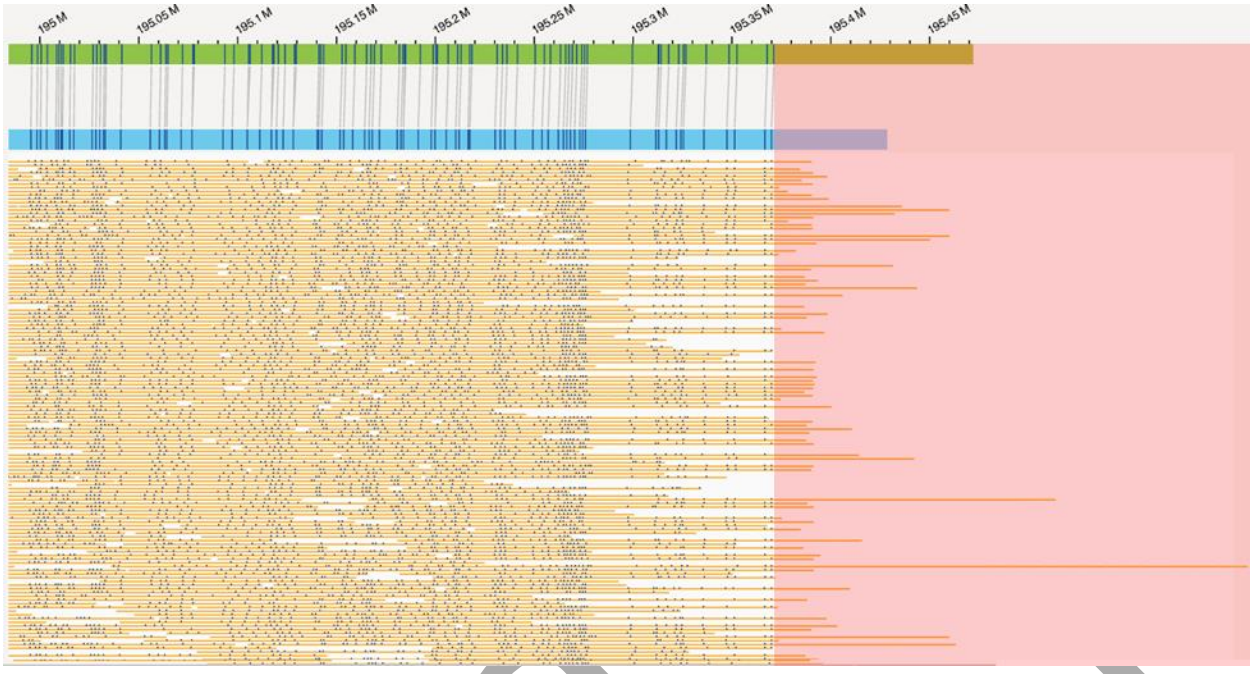
Za funkcionalne analize korištene su genske anotacije iz baze podataka GENCODE (M23 Release) za referentni genom GRCm38.p6. Za analizu ontologije gena koristili smo Python ekstenziju alata g:Profiler (Raudvere i sur., 2019.) i bazu podataka Gene Ontology za molekularne funkcije i biološke procese, te baze putanja gena „Reactome Pathway“ i Kyoto Enciklopediju Gena i Genoma (KEGG).

2.1.4.3. Analiza duljine telomera putem TelOMpy paketa

2.1.4.3.1. Definicija telomerne duljine

OGM detektira SVs pristupom slaganja genoma *de novo* – SVs se detektiraju na onim kontizima koji se samo djelomično poravnavaju s referencom. Razlike u duljini nesparenih djelova ovih molekula (ili kontiga) od duljine djelova na referetnom genomu je temelj detekcije SVs. Proširili smo ovaj koncept na računanje duljine telomera. Algoritam za odredivanje duljine telomera iz optičkih mapa genoma koji smo razvili u sklopu ove disertacije implementiran je u TelOMpy paketu za programski jezik Python. Paket, zajedno sa svim informacijama o uporabi, nalazi se na github repozitoriju - https://github.com/ivanp1994/telompy.

Optičko mapiranje je metoda temeljena na pojedinačnim molekulama, pa je za svaki uzorak i kromosom moguće izračunati duljine pojedinačnih telomera. Budući da OGM ne daje informacije o nukleotidnom slijedu sekvence već samo o položajima specifičnog DLE-1 motiva (CTTAAG) u genomu, duljina telomere se računa iz molekula koje su mapirane na krajnji motiv na referentnom kromosomskom kraku, kao duljina dijela molekule koji više nema taj motiv (Slika 5).

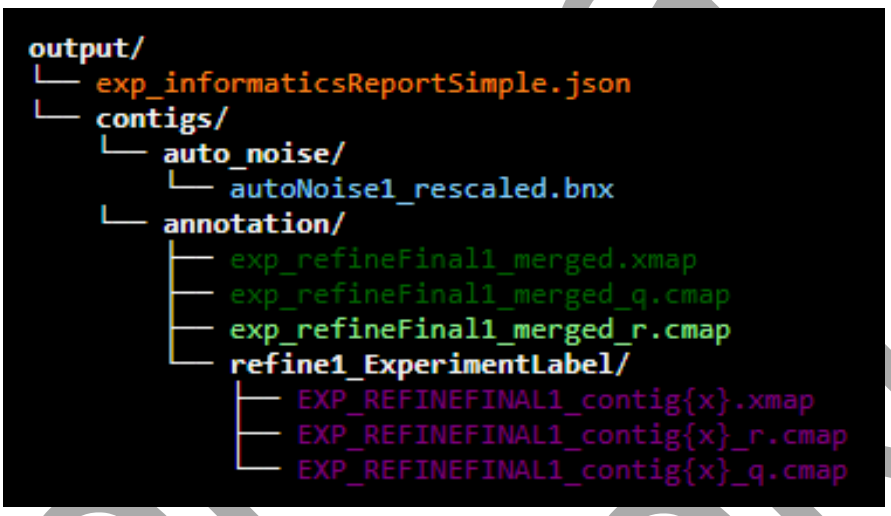


Slika 5. Prikaz subtelomerne i telomerne regije q kraka kromosoma 1 u pregledniku BioNano Access. Zelena traka predstavlja shematski prikaz referentnog genoma, plava traka predstavlja jedan složeni kontig uzorka, a krem linije predstavljaju pojedine molekule. Plave okomite linije na trakama predstavljaju sparene oznake na genomu odnosno uzorku, plave točkice na molekulama predstavljaju oznake (DLE-1 motiv) na molekulama. Brojevi i linije iznad zelene trake predstavljaju nukleotidni položaj na referentnom genomu. Duljina telomera se računa kao dio molekule nakon zadnje sparene oznake na referentnom genomu (osjenčano crveno na slici).

Ovaj izračun se temelji na nepostojanju DLE-1 motiva u telomerama, budući da se one tipično sastoje od ponavljajućeg TTAGGG motiva. Ovakav izračun osim telomerne sekvence također obuhvaća i dio subtelomerne regije u kojoj ne postoji DLE-1 motiv. Kako bismo izračunali količinu netelomerne sekvence, provjerili smo udaljenost zadnje sparene oznake na referentnom genomu od početka anotirane telomere te smo utvrdili da je količina subtelomerne sekvence u rasponu od nekoliko desetka do nekoliko tisuća baznih parova. Ovo se može smatrati zanemarivim u usporedbi s prosječnom duljinom mišjih telomera od 50 kbp (Hemann, 2000.). Osim toga, budući da je ovaj faktor prisutan u svim izračunima, ne utječe na usporedbu duljine telomera između uzoraka.

2.1.4.3.2. Teorija rada

BioNano solve poravnava molekule u nekoliko koraka: nakon uklanjanja nisko kvalitetnih molekula (molekule kraće od 150 kbp i/ili molekule s manje od 9 oznaka), molekule se međusobno poravnavaju (eng. *pairwise-alignment*) prema uzorku obilježenih DLE-1 motiva te se na temelju tih poravnanja slažu u veće jedinice nazvane *kontizi* (eng. *contigs*, odnosno *contiguous*, sekvence sačinjene od segmenata koji se preklapaju i tako formiraju jednu cjelinu). Kontizi se potom uspoređuju sa *in-silico* obilježenim referentnim genomom.



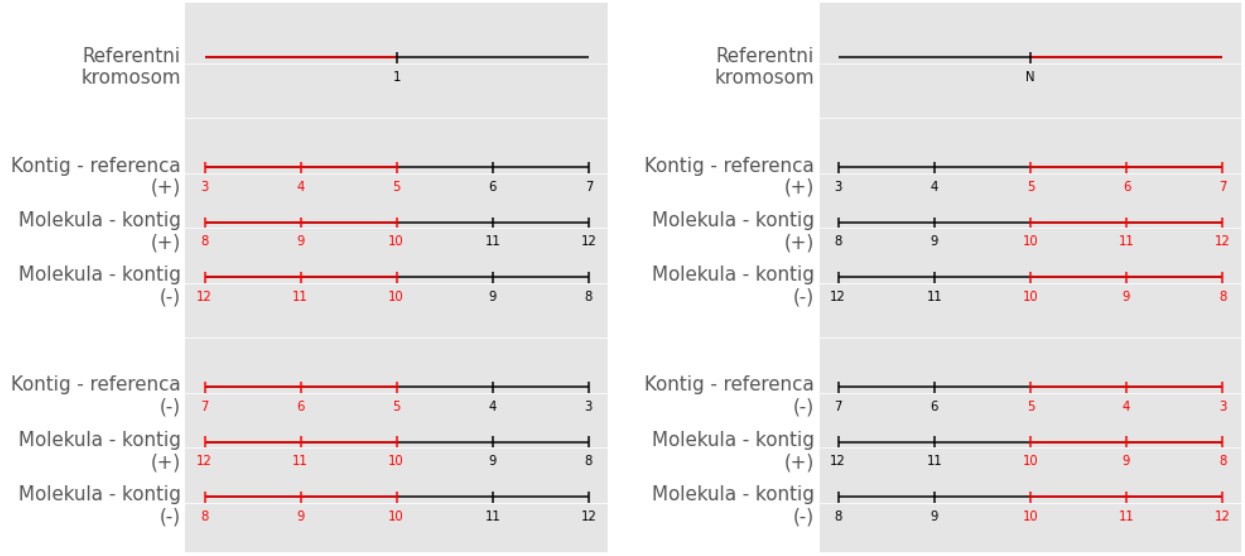
Slika 6. Struktura podataka dobivenih *de novo assembly* primjenom BioNano Solve alata. Direktoriji su napisani bijelim slovima. U direktoriju „annotation" nalazi se konačna usporedba između referentnog genoma u CMAP obliku (svijetlo zeleno) i kontiga u CMAP obliku (tamno zeleno) te je njihov rezultat u obliku XMAP datoteke (tamno zeleno). U direktoriju „refine1_ExperimentLabel" nalazi se međusobna usporedba molekula u CMAP obliku te je njihov rezultat u obliku XMAP datoteke (ljubičasto) gdje je „x" jedinstveni identifikacijski broj kontiga. Početne filtrirane i procesuirane molekule nalaze se u direktoriju „auto_noise" (svijetlo plavo). Statistike o poravnanju nalaze se u obliku JSON datoteke (narančasto).

Da bismo izdvojili molekule poravnate na krajeve kromosoma, potrebno je prvo identificirati kontige koji su poravnati na krajeve kromosoma. Te informacije se nalaze u „exp_refineFinal1_merged.xmap“ datoteci (Slika 6). Nakon identificiranja kontiga poravnatih na krajeve referentnog kromosoma, informacije o načinu poravnavanja molekula na navedene kontige nalaze se u XMAP datoteci „EXP_REFINEFINAL_contig{X}.xmap“, a informacije o duljini i nukleotidnim položajima oznaka na poravnatim molekulama nalaze se u CMAP datoteci „EXP_REFINEFINAL_contig{X}_q.cmap“ gdje je X jedinstveni identifikacijskih broj kontiga za dati skup podataka (Slika 6). Informacija o poravnavanju određene molekule ili kontiga u XMAP datotekama nalazi se u stupcu „Alignment“ u obliku parova sparenih oznaka tako da je prvi broj u paru redni broj oznake na referenci ili kontigu, a drugi broj redni broj oznake na kontigu li molekuli. Prvi član parova uvijek ide u uzlaznom redu, dok red drugog člana parova ovisi o orijentaciji sparivanja (stupac „Orientation“). Ako je orijentacija pozitivna (+), red je uzlazan, a ako je orijentacija negativna (-), red je silazan (Tablica 5).

Tablica 5. Prikaz mapiranja u XMAP datoteci. „QryContigID“ označava identifikacijski broj kontiga a „RefContigID“ označava broj kromosoma u referentnom genomu.

QryContigID	RefContigID	…	Orientation	Alignment
2921	1	…	+	(1,166)(2,167)...(4871,4623)(4872,4624)
21	2	…	-	(3525,3638)(3526,3637)...(3600,3570)(3601,3569)

Ovisno o kraku kromosoma, orijentaciji sparivanja kontiga s referentnim kromosomom, te orijentaciji sparivanja pojedinačnih molekula s kontigom, primjenjuje se jedan od 2^3 različita načina računanja duljine telomera (Slika 7).



Slika 7. Shematski prikaz svih kombinacija odnosa poravnatih molekula i kontiga s referentnim kromosomom na telomernim regijama. Orijentacija kontiga se odnosi na orijentaciju naspram reference, a orijentacija molekula na orijentaciju naspram kontiga. Crveni segmenti predstavljaju telomere, odnosno dio molekule/kontiga prije prve odnosno nakon zadnje sparene oznake na referentnom genomu. Vertikalne crtice predstavljaju oznake, a brojevi ispod njih predstavljaju redne brojeve oznaka na referentnom genomu, kontizima, ili molekulama. Na lijevom su grafu moguće kombinacije za telomeru kraćeg (p) kraka, a na desnom za telomeru duljeg (q) kraka.

Svih osam kombinacija računanja duljine telomera, na temelju kromosomskog kraka, orijentacije poravnavanja molekula na kontige te orijentacije poravnavanja kontiga na referentni genom, mogu se svesti na dvije formule:

$$duljina\ telomere = \begin{cases} LP & ako\ predznak = + \\ ML - LP & ako\ predznak = - \end{cases}$$

gdje LP predstavlja položaj oznake na molekuli koja je sparena na prvu, odnosno zadnju oznaku na referentnom genomu a ML predstavlja ukupnu duljinu molekule. Ako desnoj telomeri (na q kraku kromosoma) dodijelimo vrijednost -1 a lijevoj telomeri (p krak) vrijednost 1, predznak u gornjoj formuli je definiran kao:

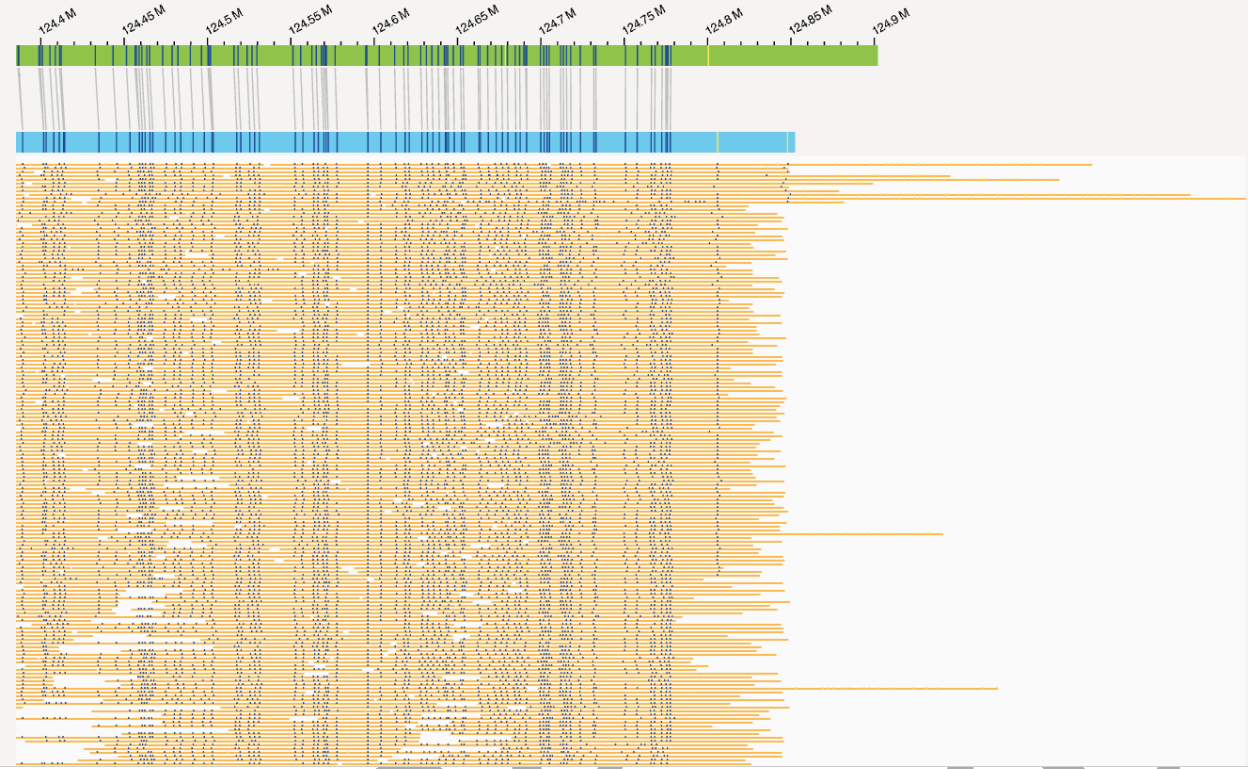
$$predznak = vrijednost\ telomere \times orijentacija\ poravnanja\ molekule\ na\ kontig \times orijentacija\ poravnanja\ kontiga\ na\ referencu$$

2.1.4.3.3. Potencijalna ograničenja

Iako je optičko mapiranje tehnologija dugih molekula, sama priroda izolacije HMW DNA znači da kraj određene molekule može biti uzrokovan degradacijom DNA, a ne prirodnim krajem molekule. Da bismo kontrolirali za ovu mogućnost, uspoređujemo duljinu telomernih molekula (molekula koje se poravnavaju na kraj kromosoma) s duljinom svih molekula koje ulaze u proces poravnavanja. Ako su molekule iz kojih računamo duljinu telomera značajno kraće od ostalih molekula, izvjesno je da je došlo do značajnije degradacije telomera. Ako nije došlo do značajnije degradacije telomera tijekom DNA izolacije, korelacija između duljine telomera i duljine telomernih molekula bi trebala biti slaba.

Postoje slučajevi gdje krajnja oznaka na referentnom genomu nije ujedno i krajnja sparena (poravnata) oznaka na kontigu (Slika 8). Takvi slučajevi se kontroliraju kroz uvođenje parametra „broj nesparenih oznaka na referenci“. U izvornim postavkama TelOMpy alata vrijednost ovog parametra iznosi 0. Međutim, ako su krajnja nesparena i krajnja sparena oznaka na referenci dovoljno blizu da molekula sparena na krajnju sparenu oznaku može u potpunosti sadržavati telomernu regiju, onda se ovaj parametar može postaviti na vrijednosti veće od nule. Korisnik se upućuje na provjeravanje ovakvih slučaja ovisno o promatranoj vrsti, o duljini telomera tih vrsta, te o prosječnoj duljini molekule koja je dobivena eksperimentalno.

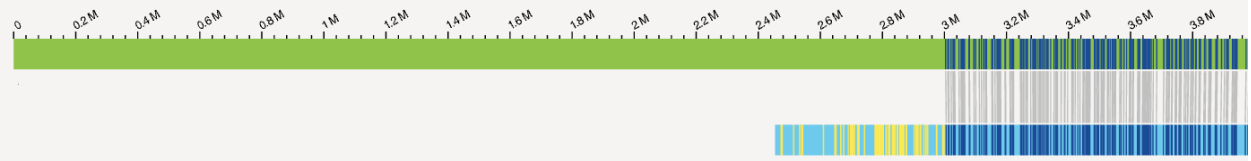
Slično, nesparene oznake na kontigu ili molekuli, distalno od krajnje sparene oznake na referentnom genomu (Slika 8), kontroliramo uvođenjem parametra „broj nesparenih oznaka na kraju molekule“ i „broj nesparenih oznaka na kontigu“. Moguće je da određena molekula ili kontig ima jednu poziciju „viška“ a uzroci mogu biti mutacija u sekvenci kojom se stvorio ciljani motiv DLE-1 ili pak nespecifično obilježavanje. U slučaju da se navedena pozicija ponavlja u više molekula, nalazit će se i u složenome kontigu. Važno je naglasiti da se ovaj broj odnosi na broj kontinuirano nesparenih oznaka lijevo (p krak), odnosno desno (q krak) od krajnje poravnate oznake na referentnom genomu, te se stoga razlikuje od ukupnog broja nesparenih oznaka na molekuli ili kontigu koji može, ali i ne mora biti veći.



Slika 8. Prikaz subtelomerne i telomerne regije q kraka kromosoma 14 u pregledniku BioNano Access. Zelena traka predstavlja shematski prikaz referentnog genoma, plava traka predstavlja jedan složeni kontig uzorka, a krem linije predstavljaju pojedine molekule. Plave okomite linije na trakama predstavljaju sparene oznake na genomu odnosno uzorku, a žute linije predstavljaju nesparene oznake. Brojevi i linije iznad zelene trake predstavljaju nukleotidni položaj na referentnom genomu. Zadnja oznaka na referentnom genomu nije ujedno i zadnja sparena oznaka na njemu. Kontig uzorka također sadrži dvije oznake koje nisu sparene na referentni genom.

Ukoliko je krajnje distalna oznaka na referentnom genomu sparena s oznakom na kontigu, ali je njena udaljenost od predvidenog kraja kromosoma u referentnom genomu znatno veća od prosječne duljine molekule, malo je vjerojatno da će molekule sparene na krajnje oznake referentnog genoma sadržavati telomernu regiju. Primjerice, prva oznaka (odnosno prvi motiv koji DLE-1 prepoznaje) na kromosomu 1 miša nalazi se 3 Mbp nakon linearnog početka kromosoma, te je prvih 3 Mbp koji sadrže i telomeru i centromeru prvog kromosoma neobilježeno, što ovaj dio čini nedostupnim za bilo kakvu analizu u kontekstu optičkog mapiranja (Slika 9). Posljedično tome, informacije o strukturnim varijantama koje se mogu javiti na kraćem kraku prvog kromosoma i o duljini telomere toga kraka su nedostupne. Ovo ograničenje kontroliramo kroz uvođenje parametra „udaljenost od kraja“, tako da udaljenost od kraja kromosoma (u referentnom genomu) ne bude veća od neke definirane vrijednosti. Ta vrijednost, osim što ovisi o prosječnoj duljini molekula, ovisi i o duljini anotirane telomere za pojedini organizam. Telomere su na referentnom genomu anotirane kao procjepi (eng. *gaps*) konzistentne veličine, koja kod miša iznosi 100

kbp a kod čovjeka 10 kbp (UCSC Table Browser) i korisnik ju treba uzeti u obzir kod definiranja vrijednosti parametra „udaljenost od kraja“. U slučaju mišjeg genoma korištena je vrijednost od 200 000 (200 kbp) što je dvostruko više od duljine anotiranih telomera, no manje od prosječne duljine molekula mapiranih na referentni genom.



Slika 9. Prikaz početka p kraka prvog mišjeg kromosoma u pregledniku BioNano Access. Zelena traka predstavlja shematski prikaz referentnog genoma, dok plava traka predstavlja jedan složeni kontig uzorka. Plave okomite linije na trakama predstavljaju sparene oznake na genomu odnosno uzorku, a žute predstavljaju nesparene oznake. Brojevi i linije iznad zelene trake predstavljaju nukleotidni položaj u referentnom genomu. Prva oznaka na referentnom genomu (što je ujedno i prva sparena oznaka) nalazi se otprilike 3 Mbp nakon početka kromosoma u referentnom genomu.

2.1.5. Korišteni puferi, reagensi, materijali, i uređaji

Tablica 6. Ime, pH, sastav, te izrada pufera korištenih u ovom istraživanju.

PUFER	pH	SADRŽAJ PUFERA	PRIPREMA
EDTA 1.0 M	8	1.0 M EDTA	372.24 g EDTA (MW=372.24) dodano je u ~800 mL reH$_2$0. Suspenzija je prebačena na magnetsku miješalicu/grijač, te se postepeno dodavalo pelete NaOH dok otopina nije postala bistra. Volumen se nadopunio do ~980 mL sa reH$_2$0. pH je namješten na 8, dopunjeno do 1 L, te autoklavirano 20 minuta.
Tris-HCL 1.0 M	8	1.0 M Tris-Hcl	24.23 g Trizma™ baze dodano je u 160 mL reH2O. pH je namješten na 8 uz koncentriranu otopinu HCL (37 %, 12 M), dopunjeno do 200 mL te autoklavirano 15 minuta.
TBE 5x	~8.3	450 mM Trizma™, 450 mM borne kiseline, 10 mM EDTA	27.82g borne kiseline (MW=61.83), 54.51g Trizma™ baze (MW=121.14) i 20mL 0.5M EDTA (pH=8) dodano je u ~800 mL reH2O te miješano na magnetnoj miješalici dok se nije otopilo. Dopunjeno je do 1 L, te autoklavirano 20 minuta.
TE 1x	8	10 mM Tris-Hcl, 1 mM EDTA	10 mL 1.0 M Tris-Hcl i 2 mL 0.5 M EDTA dodano je u ~980 mL reH2O. pH je namješten na 8, dopunjeno do konačnog volumena 1L, te autoklavirano 15 minuta.
Wash buffer	8	10 mM Tris-Hcl, 50 mM EDTA,	18.61g EDTA (M2=372.24) i 1.21 g Trizma™ baze dodano je u ~980mL reH2O. pH je namješten na 8, dopunjeno do konačnog volumena 1L, te autoklavirano 15 minuta.
TBE 0.5x / EtBr (1mg/L)	~8.3	TBE 0.5X, Etidij bromid 1 mg/L	100 mL TBE 5X razrijeđeno je do 1 L sa reH$_2$O. Za pripremu pufera sa etidij bromidom, dodano je 100 µL otopine etidij bromida koncentracije 10m g/mL.
Sarkosyl™ 25 % (w/v)	-	25 % Sarkosyl™	5 g Sarkosyl™ praha dodano je u ~12 mL reH$_2$0 te miješano uz grijanje. Volumen je dopunjen do 20 mL, te je autoklavirano 15 minuta.
Otopina proteinaze K	-	20 mg/mL proteinaze K	10 mg proteinaze K otopljeno je u 0.5 mL vode bez nukleaza.
OTOPINA DTT	-	1.0 M DTT	Otopiti 154.25 mg praha DTT po 1 mL vode bez nukleaza.
Lizni pufer	8	350 mM EDTA 2 % (w/v) Sarkosyl™, 2 mg/mL proteinaze K, 80 mM DTT	Lizni pufer priprema se netom prije digestije. Pomiješati 4.2 mL 0.5 M EDTA, 0.6mL 25 % Sarkosyl™-a, 0.6 mL 20mg/mL proteinaze K, 0.48 mL 1 M otopine DTT. Dopuniti do 6 mL vodom bez nukleaza.

Tablica 7. Korištena hrana za miševe te korištene kemikalije i reagensi.

HRANA ZA MIŠEVE	OZNAKA
Zapadnjačka hrana (eng. *Western diet*)	EF TD88137 mod. WD
Kontrolna hrana	4RF21 Mucedola

KEMIKALIJE I REAGENSI	OZNAKA
EDTA (prah)	Gram Mol P134122
Trizma™ (prah)	Sigma Aldrich T1503
Borna kiselina (prah)	Kemika EC-br. 233-139-2
NaOH (peleti)	Kemika EC-br. 215-185-5
HCl (37 %, 12 M)	Kemika EC-br. 231-595-7
DPBS pufer 1x (otopina)	Sigma Aldrich D8537
EDTA (otopina 0.5 M pH 8)	Sigma Aldrich E7889
Proteinaza K (prah)	Sigma Aldrich P2308
Sarkosyl™ (prah)	Sigma Aldrich L-5125
DTT (prah)	Thermo Scientific Catalog No. R0861
Voda bez nukleaze	Ambion AM9937
PMSF (otopina, 0.1 M u izopropilnom alkoholu)	Boston BioProducts, SKU#: BP-481
LMP Agaroza	Cambrex BioScience InCert™ Part No. 50121 Lot No. AG3952
Agaroza za gel	Bio-Rad Catalog 162-0137
Restrikcijski enzim EcoR1 (20000 U/mL)	NEB R0101S
Pufer za enzim EcoR1 (10x)	NEB R0101S
Lambda marker	NEB N0341S
Etidij bromid (otopina 10m g/mL)	Promega, H5041

Tablica 8. Korišteni uređaji i materijali

UREĐAJI I I MATERIJALI	PROIZVOĐAČ
Svjetlosni mikroskop	Leitz, Njemačka
Neubarova komora	Marienfeld, Njemačka
Hibridizacijska pećnica	Biometra (Analytik Jena), Njemačka
Magnetna miješalica	Heidolph, Njemačka
Orbitalna treskalica	Biometra (Analytik Jena), Njemačka
Vodena kupelj	Major Science, SAD
pH metar	Boeco, Njemačka
Centrifuga za mikroepruvete	Eppendorf, Njemačka
Vaga, analitička	Mettler Toledo, Švicarska/SAD
Vaga	Kern, Njemačka
CHEF III sustav za pulsnu elektroforezu	BioRad, SAD
Kalup za agarozne blokove	BioRad, SAD
Kalup za agarozni gel	BioRad, SAD
SynGene transiluminator	Fisher, UK
Hladnjak (+4 °C)	Končar, Hrvatska
Hladnjak (-20 °C)	Gorenje, Slovenija
Hladnjak (-80 °C)	Skadi, Danska
Plastične epruvete, 15 i 50 mL	Falcon, SAD
Plastične mikroeprovete, 1.5 i 5 mL	Eppendorf, Njemačka
Kriotuba	Sarsted, Njemačka
Mikropipete, različiti volumeni	Eppendorf, Njemačka
Filteri, 40 µm	Sigma Aldrich, Njemačka
Odmjerne menzure, različiti volumeni	-
Igle, 16 do 20 G	-

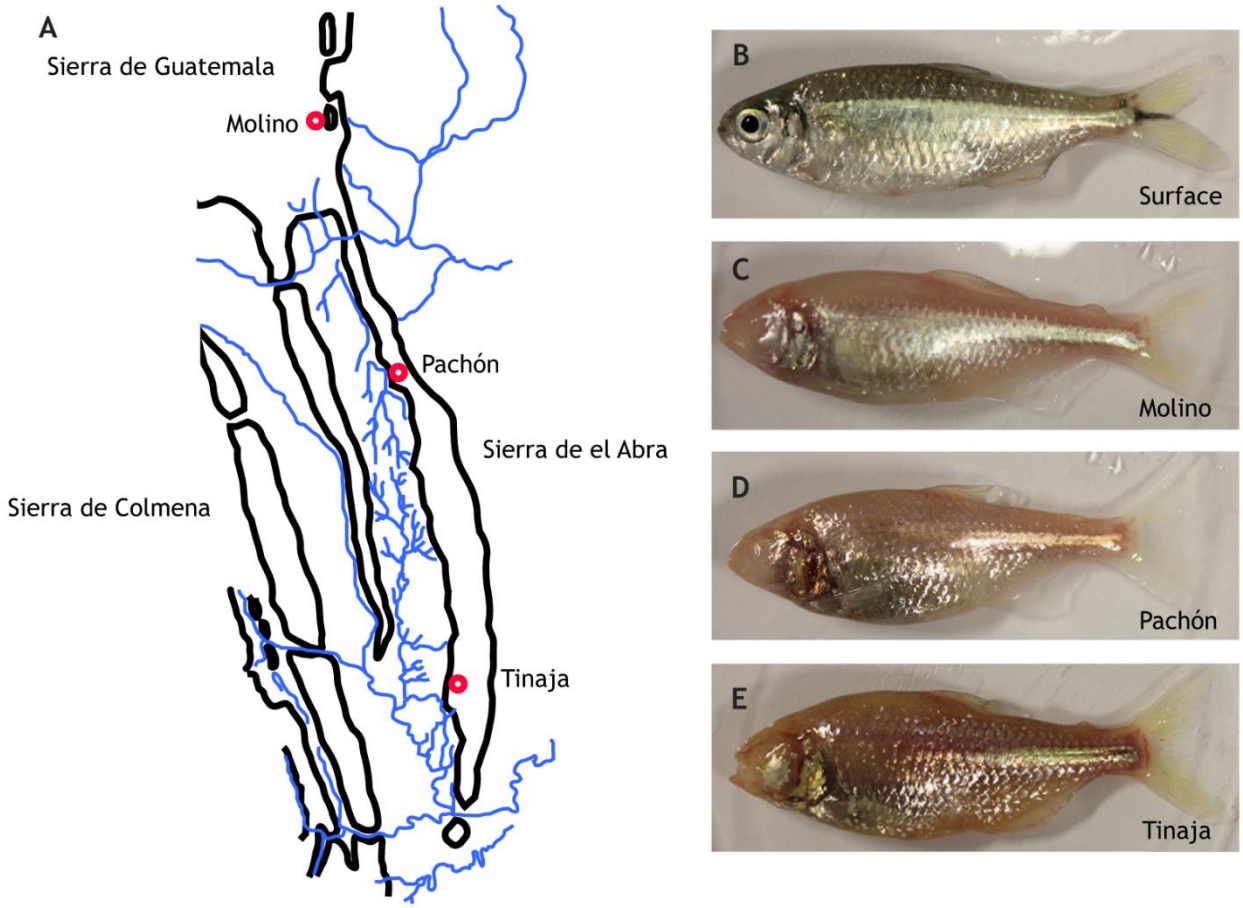
2.2. Istraživanje na razini populacije na modelu meksičke tetre (Astyanax mexicanus)

Slučajevi u kojima više nezavisnih populacija pokazuje slične karakteristike u sličnim okolišnim uvjetima prilika su za istraživanje evolucije putem prirodne selekcije - ako se isto svojstvo pojavljuje paralelno u nekoliko odvojenih populacija, vjerojatno je evoluiralo više puta kao odgovor na isti pritisak selekcije povezan sa sličnošću staništa. Unatoč potencijalu kojeg imaju, uloga SVs u kontekstu paralelne evolucije nedovoljno je istražena.

Meksička tetra (*Astyanax mexicanus*) slatkovodna je riba koja nastanjuje vode Meksika. Ova vrsta postoji u dva oblika (ekotipa): površinski ekotip, koji obitava u rijekama i jezerima Meksika i Teksasa te špiljski, slijepi ekotip, koji obitava u špiljskim vodama sjeveroistočnog Meksika. Na temelju provedenih filogenetskih analiza populacije meksičke tetre se još mogu podijeliti u stare i nove linije, ovisno o tome kada su naselile sjeverni Meksiko (Herman i sur., 2018).

Špiljski se fenotipovi smatraju derivatima ancestralnih površinskih populacija koje su kolonizirale špilje u više navrata u razdoblju od prije 10-100 tisuća godina. Poznati su po brojnim morfološkim, fiziološkim i ponašajnim značajkama, kao što su reducirani san i poremećen cirkadijalni ritam, promjene u temperaturnim preferencijama i metabolizmu, a kao najzornije promjene su redukcija ili gubitak pigmenta te regresija ili potpuni gubitak očiju (Slika 10). Određene populacije špiljskog oblika sa sličnim profilom dolaze iz neovisnih kolonizacija, te održavaju troglomorfne karakteristike unatoč priljevu genetskog materijala iz površinskih populacija što sugerira prisutnost snažnih selekcijskih pritisaka vezanih u okolišne izazove kao što su konstantna tama, niska razina kisika, te smanjena dostupnost nutrijenata.

Snažni selekcijski pritisci, raznolike karakteristike, dobro poznata povijest populacija te postojanje višestrukih populacija čine meksičku tetru prikladnim modelom za proučavanje uloge CNVs u opetovanoj prilagodbi.



Slika 10. A) Geografska distribucija meksičke tetre. B-E) Meksičke tetre iz različitih staništa. Vidljive su fenotipske razlike između špiljskih (Molino, Pachón, Tinaja) i površinskih (Surface) oblika. Preuzeto iz Kowalko, 2020.

2.2.1. Dostupnost podataka

Očitanja kratke duljine (eng. *short reads*) preuzeta su iz baze podataka European Nucleotide Archive (ENA) za genome *A.mexicanus* populacija opisanih u Herman i sur., 2018, pod šifrom projekta PRJNA260715 u obliku FASTQ formata. Ukratko, očitanja kratke duljine (100 bp) sparenih krajeva (eng. *paired-end reads*) dobivena su cjelogenomskim sekvenciranjem DNA tehnologijom sljedeće generacije (eng. *Next Generation Sequencing,* NGS). DNA je bila izolirana iz peraja ukupno 28 špiljskih riba (9 iz Molino regije, 9 iz Pachón regije, 10 iz Tinaja regije) i 15 površinskih riba (6 iz Rascon regije, 9 iz Río Choy regije) (Tablica 9).

Dodatan uzorak za Pachón populaciju dobiven je iz prethodnog referentnog genoma *A.mexicanus* koji se nalazi u NCBI bazi podataka pod oznakom GCA_004802775.1. Podaci za taj uzorak se nalaze kao 15 serija sekvenciranih na Illumina HiSeq2000 platformi izoliranih iz srca, škrga, te jetre jedne ženke u NCBI bazi podataka pod oznakom PRJNA533584. Serije su spojene u jedan uzorak, nakon čega su alatima Trimmomatic (Bolger i sur., 2014) i cutadapt (Martin, 2011) uklonjene adapterske sekvence, niskokvalitetne baze te prekratka očitanja dok ne udovoljavaju kontroli kvalitete (Herman i sur., 2018).

Referentni genom *A.mexicanus* u FASTA formatu te informacije o složenosti i anotaciji genoma preuzeti su s NCBI baze podataka pod oznakom genoma GCF_023375975.1. Informacije o anotaciji genoma sadržavale su informaciju o imenima kromosoma i njihovoj duljini, te se nalaze u tekstualnoj datoteci „assembly_report.txt“.

Anotacije gena su dobivene iz „GCF_023375975.1_AstMex3_surface_feature_table.txt.gz“ datoteke koja sadrži genomske koordinate gena, njihova imena te jedinstvene identifikacijske brojeve u NCBI Entrez bazi podataka. Anotacije eksona su dobivene iz „GCF_023375975.1_AstMex3_surface_genomic.gff.gz“ datoteke.

Tablica 9. Pregled sekvenciranih genoma korištenih u ovom radu. Sve sekvence se nalaze pod šifrom projekta PRJNA260715 na NCBI bazi podataka, a SRR ID odnosi se na šifru pod kojom se pristupa FASTQ datotekama na NCBI bazi podataka.

SRR ID	Uzorak	Regija	Ekotip	Linija
SRR1575270	Choy01	Río Choy	površina	nova
SRR1575271	Choy05	Río Choy	površina	nova
SRR1575272	Choy06	Río Choy	površina	nova
SRR1575273	Choy09	Río Choy	površina	nova
SRR1575274	Choy10	Río Choy	površina	nova
SRR1575275	Choy11	Río Choy	površina	nova
SRR1575276	Choy12	Río Choy	površina	nova
SRR1575277	Choy13	Río Choy	površina	nova
SRR1575278	Choy14	Río Choy	površina	nova
SRR1575297	Rascon02	Rascon	površina	stara
SRR1575298	Rascon04	Rascon	površina	stara
SRR1927234	Rascon13	Rascon	površina	stara
SRR1927235	Rascon15	Rascon	površina	stara
SRR1927236	Rascon8	Rascon	površina	stara
SRR1927237	Rascon6	Rascon	površina	stara
SRR1575288	Molino2a	Molino	špilja	nova
SRR1575289	Molino7a	Molino	špilja	nova
SRR1575290	Molino9b	Molino	špilja	nova
SRR1575291	Molino10b	Molino	špilja	nova
SRR1575292	Molino11a	Molino	špilja	nova
SRR1575293	Molino12a	Molino	špilja	nova
SRR1575294	Molino13b	Molino	špilja	nova
SRR1575295	Molino14a	Molino	špilja	nova
SRR1575296	Molino15b	Molino	špilja	nova
SRR1575279	Pach3	Pachón	špilja	stara
SRR1575280	Pach7	Pachón	špilja	stara
SRR1575281	Pach8	Pachón	špilja	stara
SRR1575282	Pach9	Pachón	špilja	stara
SRR1575283	Pach11	Pachón	špilja	stara
SRR1575284	Pach12	Pachón	špilja	stara
SRR1575285	Pach14	Pachón	špilja	stara
SRR1575286	Pach15	Pachón	špilja	stara
SRR1575287	Pach17	Pachón	špilja	stara
SRR1927212	Tinaja6	Tinaja	špilja	stara
SRR1927214	Tinaja12	Tinaja	špilja	stara
SRR1927215	TinajaB	Tinaja	špilja	stara
SRR1927218	Tinaja2	Tinaja	špilja	stara
SRR1927221	TinajaC	Tinaja	špilja	stara
SRR1927224	Tinaja3	Tinaja	špilja	stara
SRR1927228	TinajaD	Tinaja	špilja	stara
SRR1927232	Tinaja5	Tinaja	špilja	stara
SRR1927233	TinajaE	Tinaja	špilja	stara
SRR1927184	Tinaja11	Tinaja	špilja	stara

2.2.2. Kontrola kvalitete i mapiranje na referentni genom

Kontrola FASTQ datoteka je provjerena pomoću FastQC (Andrews, 2010) alata. FASTQ datoteke su zatim mapirane putem alata Bowtie 2 (Langmead i Salzberg, 2012) pod standardnim parametrima. SAMtools (Li i sur., 2009) alat je korišten za popravljanje nepodudarnosti sparenih očitanja, uklanjanje duplikata, te sortiranje i indeksiranje mapiranih datoteka. SAMtools alat također je korišten u računanju statistika pokrivenosti genoma. Mapirane datoteke nalaze se u obliku BAM (eng. *Binary Alignment Map*) datoteka.

2.2.3. Detekcija varijanti u broju kopija (CNVs)

Za detekciju CNVs korišten je program CNVpytor (Suvakov i sur., 2021), koji se temelji na dubini očitanja (eng. *read-depth*, RD). CNVpytor dijeli genom na uzastopne nepreklapajuće segmente, te mjeri dubinu očitanja po segmentu. Na temelju razlike u prosječnoj dubini očitanja između pojedinog segmenta i čitavog genoma identificiraju se duplikacije ili delecije. Algoritam kojeg koristi CNVpytor odlikuje se visokom točnošću pri procjeni diploidnog broja kopija (eng. *copy number*, CN), uključujući repetitivne regije, te je robustan na interindividualne razlike u pokrivenosti genoma (Garg i sur., 2021; Kosugi i sur., 2019; Pezer i sur., 2015). CN određene regije određuje se kao težinska prosječna vrijednost RD signala te regije, te je u programu definiran kao „genotip“ regije.

CNVpytor također prati ukupan broj mapiranih očitanja, te broj jedinstveno mapiranih očitanja na svakom segmentu, što omogućava izračunavanje udjela očitanja niske kvalitete (očitanja koja se jednako dobro mogu mapirati na više od jedne regije). Na taj je način moguće pratiti utjecaj očitanja niske kvalitete na rezultate. Očitanja koje se ne mogu mapirati na jedinstvenu poziciju u genomu (primjerice očitanja koja potječu iz visokorepetitivnih regija) mogu se „nagomilati“ prilikom mapiranja na određenoj genomskoj poziciji te umjetno povećati CN navedene regije.

Rezolucija detekcije CNV (broj i donja granica veličine) ovisi o veličini segmenata u kojima se mjeri dubina očitanja (eng. *bin-size*). Veća pokrivenost sekvenciranja (eng. *coverage*) omogućuje veće segmentiranje tj. mjerenje dubine očitanja u manjim segmentima genoma a time i bolju rezoluciju detekcije. Sukladno preporuci autora CNVpytora, veličinu segmenta smo algoritamski odredili za svaki pojedinačni uzorak tako da relativna standardna devijacija ukupnog RD signala iznosi između 0.2 i 0.25.

2.2.4. Reproducibilnost

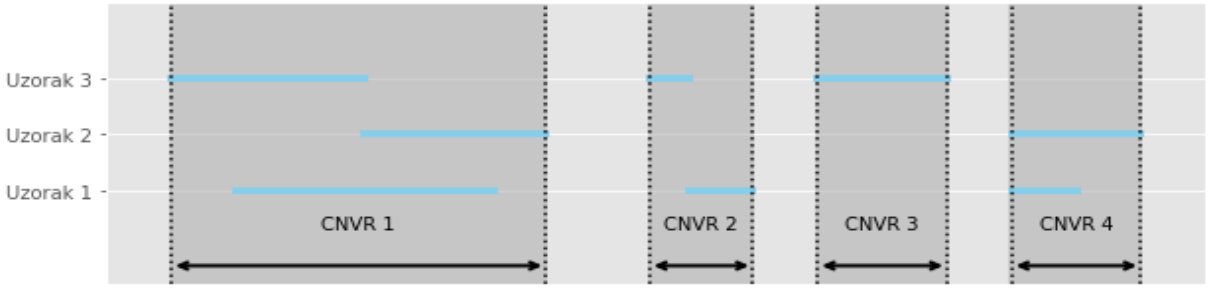
Svaki od opisanih koraka analize, zajedno sa svim računalnim okolišima te programima izvedeni su unutar *snakemake* programa za upravljanjem toka rada (eng. *workflow manager*) (Köster i Rahmann, 2018) te su dostupni na github repozitoriju na sljedećem linku – https://github.com/ivanp1994/AMEX_01_PHDTHESIS. Shematski prikaz procesa dan je u Slici 11.

Slika 11. Shematski prikaz prethodno opisanog bioinformatičkog postupka implementiranog u snakemake programu za upravljanjem toka rada. Slika je generirana sa `--rulegraph` opcijom nakon završetka postupka.

2.2.5. Definicije korištenih termina

U radu su korištene sljedeće definicije:

- **CNV gen**: protein-kodirajući gen koji se u potpunosti nalazi unutar CNVs u barem jednom uzorku
- **CNV regija (eng. *CNV region*, CNVR)**: neprekinuta genomska regija određena duljinom između najmanje i najveće genomske koordinate svih CNVs čiji se položaji u genomu preklapaju ili dodiruju između više jedinki (Slika 12)
- **duplikacija**: odnosi se na CNV detektiran CNVpytor programom kao „duplikacija“, što uključuje i višestruke amplifikacije te amplifikacije koje ne moraju biti izražene cijelim brojem
- **delecija**: odnosi se na CNV detektiran CNVpytor programom kao „delecija“, što uključuje i delecije koje nisu izražene cijelim brojem
- **zajednički CNV**: CNV koji se nalaze na istom položaju u genomu dvaju različitih jedinki, bez obzira na tip CNV (delecija ili duplikacija). Granice CNV definirane su genomskim koordinatama u referentnom genomu, a udaljenost između graničnih koordinata definira CNV duljinu. Ako se položaj CNV u jednom genomu preklapa s položajem CNV u drugom genomu po najmanje 50 % obje duljine, ti se CNVs definiraju kao zajednički između dva genoma.
- **privatni CNV**: CNV prisutan unutar CNVR u najmanje jednoj jedinci jedne grupe (populacije, ekotipa ili sl.), a nije prisutan u ostalim grupama. Prisutnost nekog CNV u CNVR određuje se na temelju bilo koje duljine preklapanja.
- **privatni CNVR**: CNVR koji je prisutan isključivo u jedoj grupi (populacija, ekotip ili sl.), a nije prisutan u ostalim grupama.
- **singleton CNV**: CNV detektiran isključivo u jednoj jedinci koji se svojim položajem u genomu ne preklapa ni s jednim CNVs detektiranim u ostalim jedinkama ni s jednim baznim parom.



Slika 12. Shematski prikaz definicije CNV regija. Prikazan je primjer za tri različita uzorka i četiri CNVR-a. Plave linije predstavljaju genomske pozicije detektiranih CNVs neovisno o tipu (duplikacija ili delecija). CNVRs se dobivaju uzimanjem najmanje početne i najveće završne koordinate preklapajućih CNVs.

2.2.6. Analiza genetičke raznolikosti

Kao pokazatelje relativne genetičke raznolikosti unutar i između populacija koristili smo:

- **udio zajedničkih CNVs:** izračunat je kao omjer broja zajedničkih CNVs između dvije jedinke i prosječnog broja detektiranih CNVs izmedju te dvije jedinke. Veći udio zajedničkih CNVs ukazuje na veću sličnost među dvije jedinke, a veći prosjek udjela zajedničkih CNVs unutar grupe na veću sličnost jedinki između te grupe.
- **udio i broj privatnih CNVRs te broj privatnih CNVs po CNVR:** veći broj, odnosno udio, privatnih CNVRs te veći broj privatnih CNVs po CNVRs unutar grupe ukazuje na veću međusobnu sličnost unutar te grupe.
- **udio singleton CNVs:** izračunat je kao omjer broja singleton CNVs u pojedinoj jedinci i ukupnog broja detektiranih CNVs u toj jedinci. Veći udio singletona unutar jedne grupe ukazuje na veću raznolikost CNVs unutar te grupe.

2.2.7. Analiza diferencijacije populacija

Diferencijaciju populacija određivali smo na temelju broja kopija CNV gena i udjela zajedničkih CNVs. Broj kopija CNV gena transformirali smo analizom principalnih komponenata (eng. *principal component analysis*, PCA) te podvrgnuli grupiranju po principu K grupa (eng. *K-means clustering*), a konačan broj grupa odredili po tome koji broj grupa ima najveći rezultat silueta (eng. *silhouette score*). Matricu CN CNV gena podvrgnuli smo hijerarhijsko aglomerativnom grupiranju (eng. *hierarchical agglomerative clustering*, HAC) koristeći Wardovu metodu nad Euklidovom metrikom normalizirane matrice CN CNV gena.

Udio zajedničkih CNVs tretirali smo kao metriku sličnosti uzoraka te smo matricu udjela zajedničkih podvrgnuli grupiranju po principu K grupa i HAC koristeći Wardovu metodu.

2.2.8. Statističke usporedbe broja kopija

Kako bismo pronašli gene i CNV regije sa značajnim razlikama u broju kopija između površinskih i špiljskih riba, kombinirali smo rezultate zasnovane na (1) usporedbi populacijskih parova i (2) rezultate kumulativne usporedbe. U prvom slučaju, koristi se Welch t-test na broju kopija gena ili CNVRs u svim kombinacijama špiljske naspram površinske populacije (primjerice Molino-Rascon, Molino-Choy, Pachón-Rascon, Pachón-Choy, itd.). U drugom slučaju, Mann-Whitneyev test se koristi u usporedbi broja kopija gena i CNVRs svih špiljskih jedinki s brojem kopija iz svih površinskih jedinki. Neparametrijski Mann-Whitneyjev test je odabran jer uzima u obzir multimodalnost podataka, odnosno potencijalnu populacijsko-specifičnu distribuciju broja kopija. Nadalje smo prilagodili p-vrijednosti koristeći FDR Benjamin-Hochenberg metodu i FWER Holmovu metodu, postavljajući razinu značajnosti na alfa 0.05. Smatrali smo da je razlika značajna kada je kriterij razine značajnosti zadovoljen u oba pristupa.

2.2.9. Analiza funkcionalnih anotacija

Funkcionalna analiza gena provedena je pomoću alata DAVID (Sherman i sur., 2022), koristeći anotacije za UP_KW_BIOLOGICAL_PROCESS, GOTERM_BP_DIRECT i KEGG_PATHWAY što odgovara UniProt i Gene Ontology bazama podataka za biološke procese i Kyoto Enciklopediji Gena i Genoma. Riječi koje se najčešće pojavljuju iz rezultata DAVID analize su izdvojene, te je sastavljen popis pojmova koji obuhvaćaju riječi ili korijene riječi. Takvi pojmovi korišteni su za daljnju analizu DAVID rezultata. Primjerice, izraz „*nerv**" korišten je za brojanje svih instanci riječi *nerve*, *nervous*, i *innervation*.

2.2.10. Permutacije

Usporedbom detektiranog broja CNVs s očekivanim brojem moguće je dobiti uvid u utjecaj prirodne selekcije. Ako je udio CNVs koji pogađaju određenu kategoriju gena sličan udjelu dobivenom nasumičnim permutacijama CNV koordinata, možemo reći da ti CNVs predstavljaju neutralne genetske varijacije. Ako je očekivani udio veći, odnosno manji od detektiranog udjela, takvi CNVs mogli bi biti pod utjecajem negativne, odnosno pozitivne selekcije.

Kako bi se utvrdio utjecaj prirodne selekcije na CNVs unutar pojedinih kategorija gena, permutirali smo genomske koordinate CNVs i analizirali njihovo preklapanje sa svakom od anotiranih genskih kategorija što uključuje protein-kodirajuće gene, pseudogene, gene koji kodiraju za ribosomalnu (rRNA) i transportnu RNA (tRNA), gene koji kodiraju za dugu nekodirajuću RNA (lncRNA) te za malu nuklearnu (snRNA) i malu nukleolarnu RNA (snoRNA). Koordinate CNVs nasumično su izmiješane na istim kromosomima, pri čemu je osigurano da distribucija veličine CNVs i omjer duplikacija i delecija odgovaraju onima u pravim podacima, i da se izbjegnu anotirane praznine (eng. *gaps*) u referentnom genomu.

2.2.11. Korišteni paketi

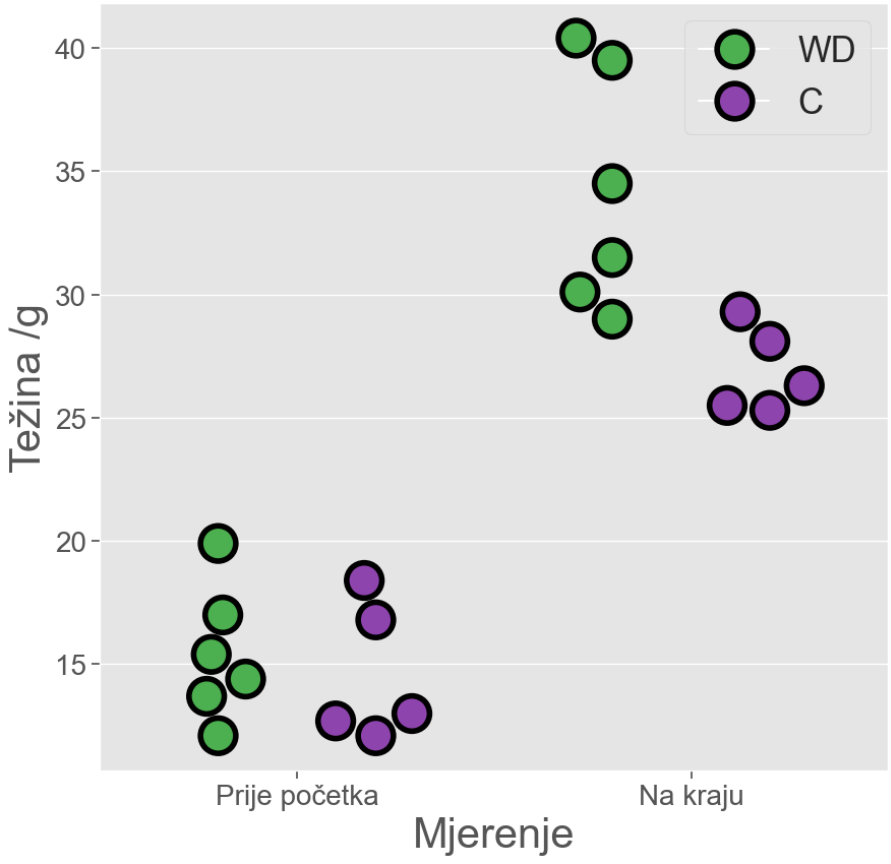
Sve analize nizvodno od detekcije CNVs, uključujući vizualizacije i statističke testove, izvršene su u programskom jeziku Python (verzija 3.8). Za vizualizacije su korišteni paketi *matplotlib* i *seaborn*, za statistički dio korišteni su paketi *sklearn* i *scipy*, za generalnu obradu tabličnih podataka korišteni su paketi *numpy* i *pandas*. Za permutacijske analize, permutacije su vršene u *numpy* paketu. Operacije nad genomskim intervalima što uključuje preklapanje genomskih intervala i pronalaženje gena najbližih kodirajućim CNVRs korištena je biblioteka *bioframe* (Abdennur i sur., 2022.).

3. REZULTATI

3.1. Utjecaj prehrane bogate mastima na strukturne varijacije

3.1.1. Utjecaj na tjelesnu težinu

Nije bilo značajne razlike u težini miševa u početku pokusa (Mann Whitney test, p vrijednost 0.58), međutim na kraju pokusa miševi u WD skupini su bili značajno teži (Mann Whitney test, p vrijednost 0.0086, Cohenov D od 1.9, CLES 0.97) od miševa u kontrolnoj skupini (Slika 13).

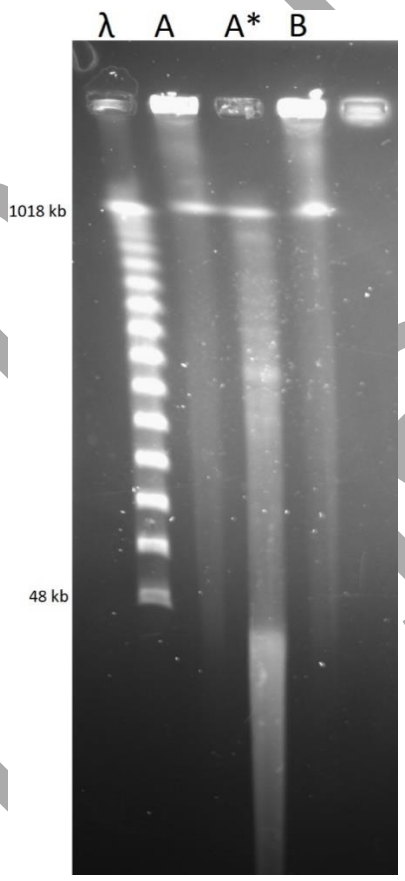


Slika 13. Težina miševa na početku i na kraju pokusa.

3.1.2. Provjera kvalitete

3.1.2.1. Kvaliteta HMW DNA nakon izolacije iz spermija

Kvalitetu svake izolacije HMW DNA pratili smo kroz procjenu duljine DNA te kroz provjeru uklanjanja proteinske komponente kromatina. Nakon svake izolacije, dio uzoraka je provjeren pomoću PFGE, prije i nakon razgradnje restrikcijskim enzimom. DNA se prije razgradnje veličinom nalazila u rasponu lambda markera, a najveći dio izolata bio je iznad 1 Mbp (Slika 14). Nakon razgradnje restrikcijskim enzimom, duljina DNA nije prelazila 1 Mbp a znatan dio nalazio se ispod 48 kbp. Zaključili smo da je HMW DNA koja se dobiva ustanovljenim protokolom za izolaciju u rasponu veličine koji zadovoljava zahtjeve OGM-a (minimalna veličina od 150 kb), te da je slobodna od proteina, čime se ostvaruje uvjet za obilježavanje DLE-1 enzimom.



Slika 14. Primjer rezultata gel-elektroforeze u pulsirajućem polju. Oznakom λ označen je lambda marker (raspon fragmenata od 48.5 do 1018.5 kb). Jažice označene A i B sadrže HMW DNA izoliranu iz otprilike 2.5 milijuna spermija. Jažica A* sadrži izolat HMW DNA pod oznakom A koji je razgrađen enzimom EcoR1.

3.1.2.2. Koncentracija HMW DNA

Kao pokazatelj kvalitete HMW DNA prije samog postupka optičkog mapiranja korištena je prosječna koncentracija DNA nakon DLS obilježavanja koja mora biti u rasponu od 4 do 16 ng/µL, a koeficijent varijacije koncentracije DNA (3 mjerenja) treba biti manji od 30 %.

DNA iz obje vrste tkiva zadovoljavala je koncentracijom, s iznimkom DNA izolirane iz bubrega miša 5458 čija je prosječna koncentracija DNA bila 3.2 ng/µL (Tablica 10).

Tablica 10. Koncentracije izoliranih HMW DNA nakon DLS obilježavanja.

Tkivo	Skupina	Miš (ID)	Koncentracija DNA (ng/µL)	Koeficijent varijacije (%)
Bubreg	C	**5458**	**3.2**	**0.2**
Bubreg	C	5460	12	9.8
Bubreg	C	5790	7.1	0.1
Bubreg	C	5792	6.6	0.2
Bubreg	WD	5455	7.9	0.1
Bubreg	WD	5456	4.8	0
Bubreg	WD	5457	8.6	0.6
Bubreg	WD	5787	5.9	0.2
Bubreg	WD	5788	4.3	0.1
Spermij	C	5460	9.37	6.2
Spermij	C	5790	6.13	20
Spermij	C	5458	6	4.2
Spermij	C	5792	9.02	26
Spermij	WD	5455	8.72	5.1
Spermij	WD	5457	10.48	9.7
Spermij	WD	5456	6.79	2.1
Spermij	WD	5787	13.45	6.8
Spermij	WD	5788	8.26	11.7

3.1.2.3. Kvaliteta podataka optičkog mapiranja

Kvalitetu podataka OGM pratili smo kroz nekoliko pokazatelja (Tablica 11). Prema preporukama Bionano Genomics protokola za OGM, prosječna duljina molekula koje ulaze u proces *de novo* slaganja genoma treba biti veća od 230 kbp što je slučaj sa svim uzorcima. Također, prosječan broj oznaka po 100 kbp je u propisanom rasponu od 14 do 17, uz iznimku tri uzorka sa nešto nižom gustoćom oznaka (Tablica 11). Međutim pokrivenost reference prije poravnavanja je iznad 100x, odnosno iznad 70x nakon poravnavanja, što je također u skladu s preporukama. Nadalje, u svim uzorcima je udio molekula poravnatih na referencu iznad 60 %, a većina (13) uzoraka nalazi se iznad optimalne vrijednosti.

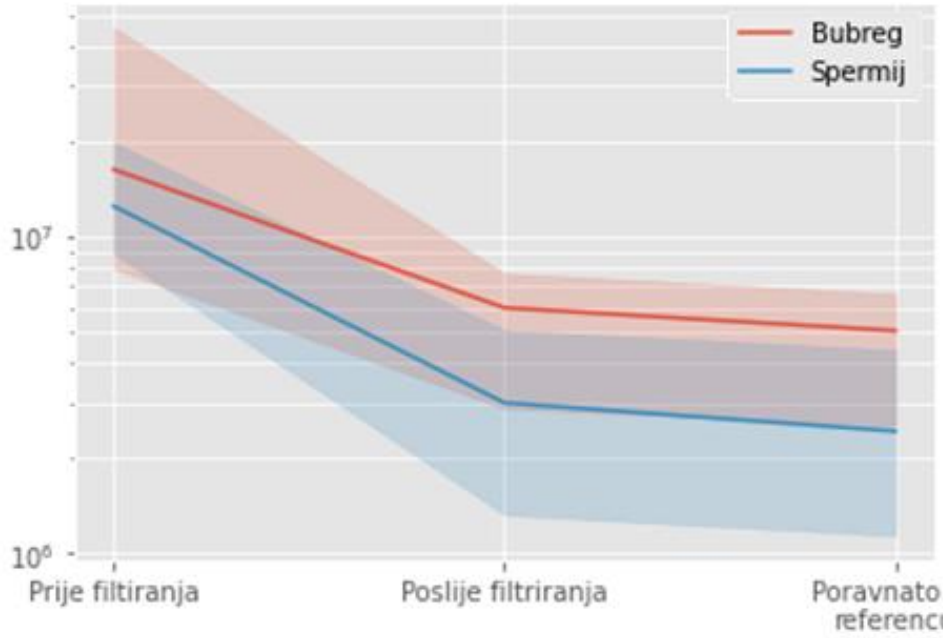
N50 je mjera kvalitete složenosti genoma gdje veći N50 označava veći kontinuitet genomskih mapa. U svim uzorcima, vrijednost N50 diploidnih genomskih mapa je bila iznad 50 s minimalnom varijacijom između uzoraka (Tablica 11).

Tablica 11. Pokazatelji kvalitete OGM.

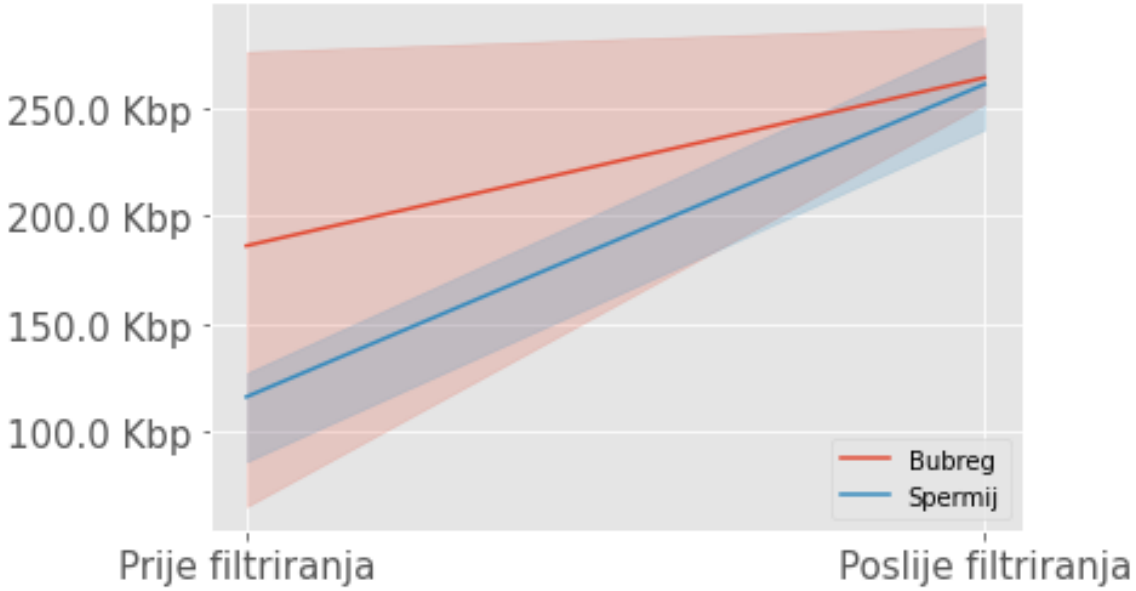
Tkivo	Tretman	Miš	Prosječna duljina filtriranih molekula /kbp	Gustoća oznaka po 100 kbp	Pokrivenost reference prije poravnanja /X	Udio molekula poravnatih na referencu	Pokrivenost reference poslije poravnanja /X	N50 diploidnih genomskih mapa / Mbp
Bubreg	C	5458	258.27	15.63	273.14	0.88	192.68	96
Bubreg	C	5460	275.44	15.32	764.61	0.88	561.91	107.96
Bubreg	C	5790	287.99	15.84	657.14	0.94	542.61	111.18
Bubreg	C	5792	255.81	15.84	598.28	0.93	472.04	117.96
Bubreg	WD	5455	252.29	**12.64**	711.93	0.63	390.58	102.05
Bubreg	WD	5456	276.48	16.44	463.34	0.91	355.96	108.84
Bubreg	WD	5457	266.08	**12.52**	616.85	0.69	391.74	102.02
Bubreg	WD	5787	253.05	16.01	574.3	0.92	459.2	102.04
Bubreg	WD	5788	252.61	15.61	554.23	0.92	427.49	102.05
Spermij	C	5458	245.98	14.1	314.64	0.79	206.27	101.28
Spermij	C	5460	270.34	16.23	219.82	0.82	144.42	101.97
Spermij	C	5790	282.9	14.13	264.17	0.83	188	106.07
Spermij	C	5792	269.94	14.31	288.19	0.78	195.26	96.39
Spermij	WD	5455	265.17	15.71	359.81	0.81	230.15	101.98
Spermij	WD	5456	258.95	14.69	232.65	0.82	171.42	95.04
Spermij	WD	5457	240.29	16.14	117.32	0.86	79.61	96.1
Spermij	WD	5787	252.17	15	470.4	0.87	320.63	95.94
Spermij	WD	5788	264.61	**13.91**	313.72	0.75	196.66	101.16

3.1.2.4. Broj i duljina molekula

S obzirom na razlike između standardnog BNG protokola za izolaciju HMW DNA iz somatskog tkiva i protokola za izolaciju HMW DNA iz spermija koji smo razvili, bilo je potrebno utvrditi na koji način te razlike eventualno utječu na postupak *de novo* sklapanja genoma. Stoga smo usporedili broj molekula na početku obrade podataka i nakon uklanjanja molekula kraćih od 150 kbp. Iako je početan broj molekula sličan u oba tkiva, filtriranjem se gubi više molekula u spermijima, te prosječan broj molekula iz spermija poravnatih na referencu iznosi trećinu prosječnog broja molekula iz bubrega (Slika 15). Prosječna duljina molekula izoliranih iz spermija je kraća od prosječne duljine molekula izoliranih iz bubrega, no raspon duljina molekula iz bubrega je puno širi. Nakon uklanjanja molekula ispod 150 kbp i molekula koje sadrže manje od 9 oznaka, prosječna duljina molekula iz oba tkiva je slična (Slika 16).



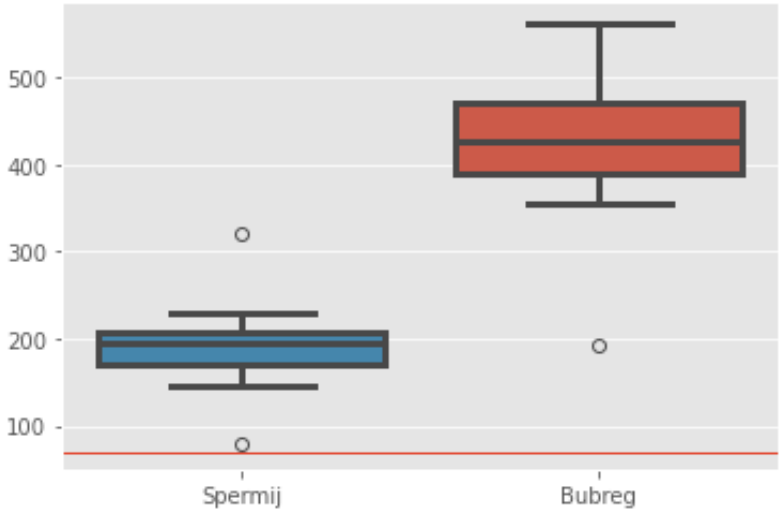
Slika 15. Broj molekula prije filtriranja, nakon filtriranja, te broj molekula poravnatih na referencu. Linije predstavljaju srednje vrijednosti, a osjenčanja pokazuju raspon vrijednosti između uzoraka.



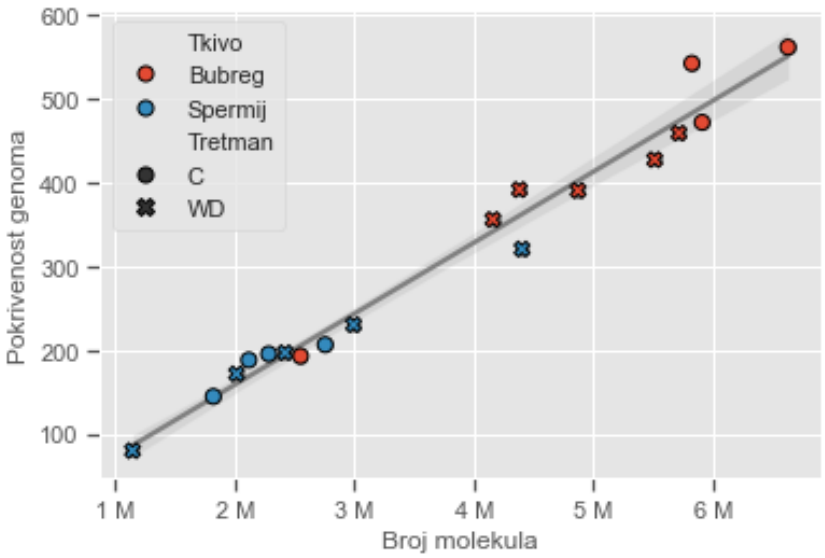
Slika 16. Prosječna duljina molekula prije i poslije filtriranja. Linije predstavljaju srednje vrijednosti, a osjenčanja pokazuju raspon vrijednosti između uzoraka.

3.1.2.5. Razlike u pokrivenosti genoma između spermija i bubrega

Pronalazimo da je pokrivenost genoma spermija drastično manja (T test za zavisne uzorke, p vrijednost 0.0007, Cohenov D 2.53) nego pokrivenost genoma bubrega (Slika 17). Ova razlika se može objasniti upola kraćom prosječnom duljinom početnih molekula iz spermija nego iz bubrega (Slika 16) te uklanjanjem molekula nezadovoljavajuće duljine (Slika 15). Sukladno tome, vidimo visoku korelaciju između broja molekula mapiranih na referencu i pokrivenosti genoma (Pearsonov r 0.987, p vrijednost 3.8×10^{-14}) (Slika 18).



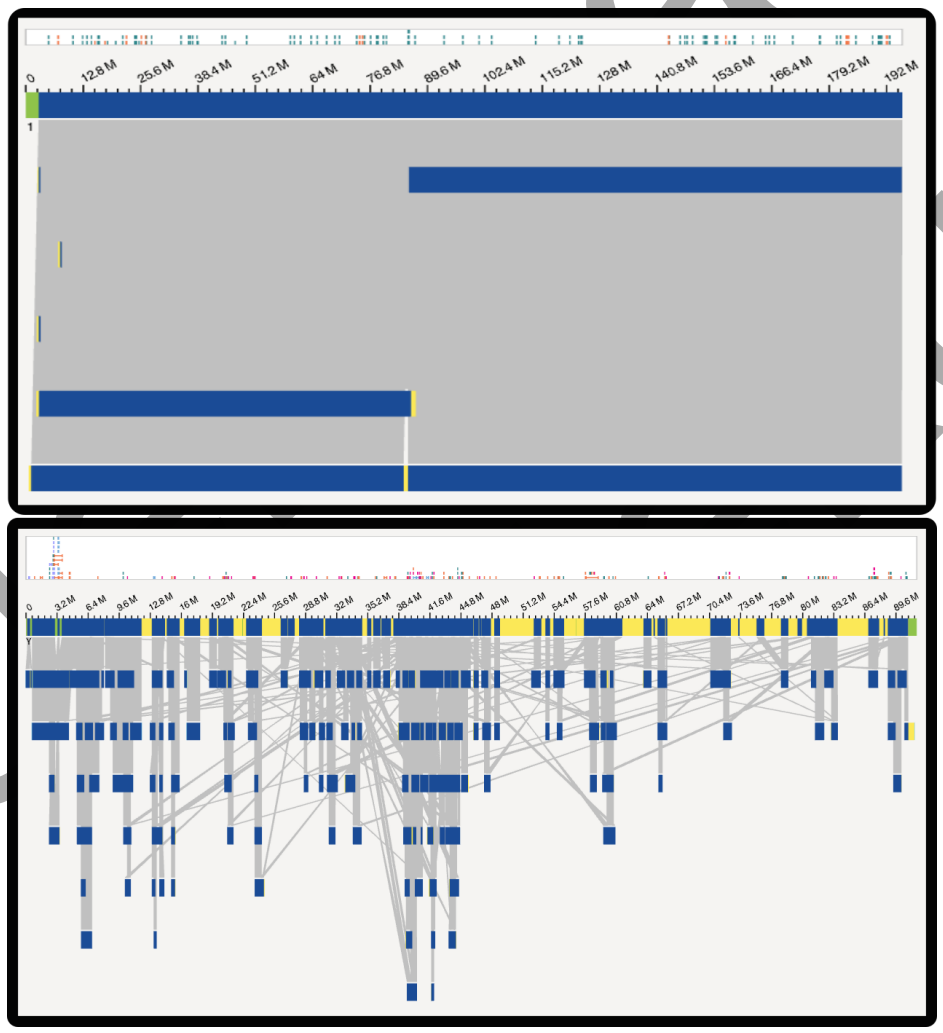
Slika 17. Pokrivenost genoma nakon poravnavanja. Crvena linija predstavlja preporučenu minimalnu pokrivenost nakon poravnavanja (70x).



Slika 18. Korelacija između broja molekula (u milijunima) i pokrivenosti genoma (M = milijun).

3.1.3. Analiza strukturnih varijacija

U svim analiziranim uzorcima, na kromosomu Y (označen kao kromosom 21) je detektirano najviše SVs: oko 95 % duljine kromosoma Y u referentnom genomu je pokriveno strukturnim varijantama (Slika 21). S obzirom na fragmentiranost genomskih mapa, ne može se s pouzdanošću utvrditi u kojoj mjeri ovaj rezultat odražava stvarnu genetičku varijabilnost a koliko je posljedica nedovoljne složenosti kromosoma Y (Slika 19).



Slika 19. Primjer prikaza složenosti genomskih mapa na kromosomu 1 (gore) i kromosomu Y (dolje) u BioNano Access pregledniku. Najgornja bijela traka prikazuje raspodjelu detektiranih SVs po cijelom kromosomu kao horizontalne crte i tocke razlicitih boja. Referentni genom je prikazan na sljedećoj traci sa naznačenim koordinatama a složeni kontizi su prikazani u pojedinim trakama ispod. Poravnate regije između kontiga i referentnog genoma prikazane su plavom a regije bez poravnanja žutom bojom. Zelenom bojom na referentnom genomu označene su regije bez DLE-1 oznaka.

3.1.3.1. Broj detektiranih strukturnih varijanti

Ukupno smo detektirali između 1332 i 1591 SVs po uzorku. Unutar tih brojeva detektiranih SVs, njih 477 u svim uzorcima imaju identične koordinate na referentnom genomu, pripadaju istom tipu varijante te imaju VAF = 1. Ovih 477 SVs stoga ne možemo smatrati stvarnim varijantama između analiziranih uzoraka, nego oni predstavljaju razliku u usporedbi s referentnim genomom. Od tih 477 SVs, glavninu (428) čine insercije, a ostatak delecije (46) i inverzije (3). Ovi SVs su izuzeti iz svih daljnjih analiza, osim u analizama razlika s referentnim genomom.

Konačno, izuzimanjem navedenih 477 SVs, pronalazimo između 855 i 1114 SVs po uzorku. Najčešći detektirani tip SVs su insercije koje čine više od 500 događaja u svim uzorcima, dok su inverzije najrjeđe (Tablica 12). Ukupno su detektirane tri interkromosomske translokacije – jedna u uzorku bubrega miša 5790, te dvije u uzorku miša 5787.

Genomi spermija sadrže statistički značajno više SVs od genoma bubrega (T test za zavisne uzorke, p vrijednost 0.027, Cohenov D od 1.515). Ova razlika je snažnija u pokusnoj skupini (p vrijednost 0.045, Cohenov D 1.560) naspram kontrolne skupine (p vrijednost 0.130, Cohenov D 1.27) međutim statistički značaj pokusne skupine ne opstaje Holmovu korekciju za višestruke usporedbe (Tablica S1 u prilozima).

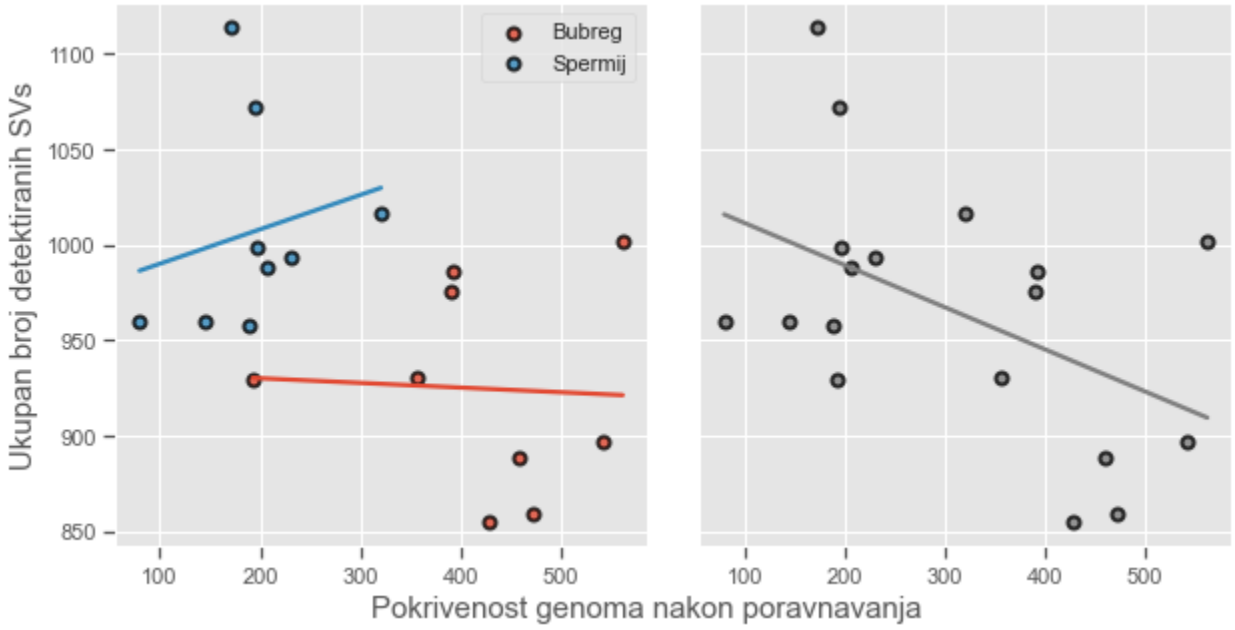
Na sličan način smo testirali broj pojedinog tipa SV između skupina i između tkiva te pronašli da su inverzije, intrakromosomske translokacije, delecije, i duplikacije statistički značajno brojnije u spermijima nego u bubrezima pokusne skupine, a da su samo inverzije statistički značajno brojnije u spermijima nego u bubrezima kontrolne skupine (Tablica S1). Međutim, samo inverzije u pokusnoj skupini opstaju kao statistički značajno brojnije u spermijima nego u bubrezima. Vrijedi istaknuti da su, unutar skupine i po tipu SVs, sve SVs brojnije u spermijima nego u bubrezima, da je ovaj efekt izraženiji u pokusnoj skupini, te da je snaga testa (naročito unutar pokusne skupine) izrazito visoka (>0.89).

Visoke p vrijednosti se u ovom kontekstu mogu objasniti malim brojem uzoraka – da bi se postigla visoka (>0.80) snagu testa, velik efekt (Cohen D > 1.5), te α od 0.05 trebalo bi po skupini imati između 6 i 10 uzoraka (Tablica S1 u prilozima).

Tablica 12. Broj SVs po tipu i po uzorku između različitih uzoraka. Ova tablica ne uključuje 409 insercija i 42 delecije koje su zajedničke svim promatranim uzorcima.

Tretman	Tkivo	Miš	delecija	duplikacija	insercija	intrakromosomska translokacija	inverzija	Ukupno
C	Bubreg	5458	211	81	589	29	19	929
C	Bubreg	5460	230	80	626	48	17	1002
C	Bubreg	5790	208	57	589	26	17	897
C	Bubreg	5792	183	67	570	23	16	859
C	Spermij	5458	221	86	633	26	22	988
C	Spermij	5460	228	78	609	21	24	960
C	Spermij	5790	208	77	632	15	26	958
C	Spermij	5792	268	87	662	31	24	1072
WD	Bubreg	5455	198	79	663	16	20	976
WD	Bubreg	5456	218	71	589	31	21	930
WD	Bubreg	5457	209	75	672	13	17	986
WD	Bubreg	5787	219	58	574	25	10	888
WD	Bubreg	5788	197	76	555	16	11	855
WD	Spermij	5455	243	91	601	29	29	993
WD	Spermij	5456	260	98	694	35	27	1114
WD	Spermij	5457	206	71	635	26	22	960
WD	Spermij	5787	248	96	593	56	23	1016
WD	Spermij	5788	220	92	634	28	25	999

S obzirom na utvrđenu značajnu razliku između tkiva u pokrivenosti genoma i ukupnom broju detektiranih SVs, bilo je potrebno utvrditi postoji li korelacija između njih, odnosno u kojoj mjeri pokrivenost genoma utječe na moć detekcije SVs u našem skupu podataka (Slika 20). Ne pronalazimo snažnu niti statistički značajnu korelaciju (Slika 20), gledano ukupno za sve uzorke (Spearmanov koeficijent -0.33, p vrijednost 0.09), ili po tkivima (Spearmanov koeficijent 0.343, p vrijednost 0.18 za spermije; Spearmanov koeficijent -0.09, p vrijednost 0.45 za bubrege). Ovi rezultati ukazuju na dobru usporedivost detektiranih SVs između tkiva unatoč značajnim razlikama u pokrivenosti genoma.

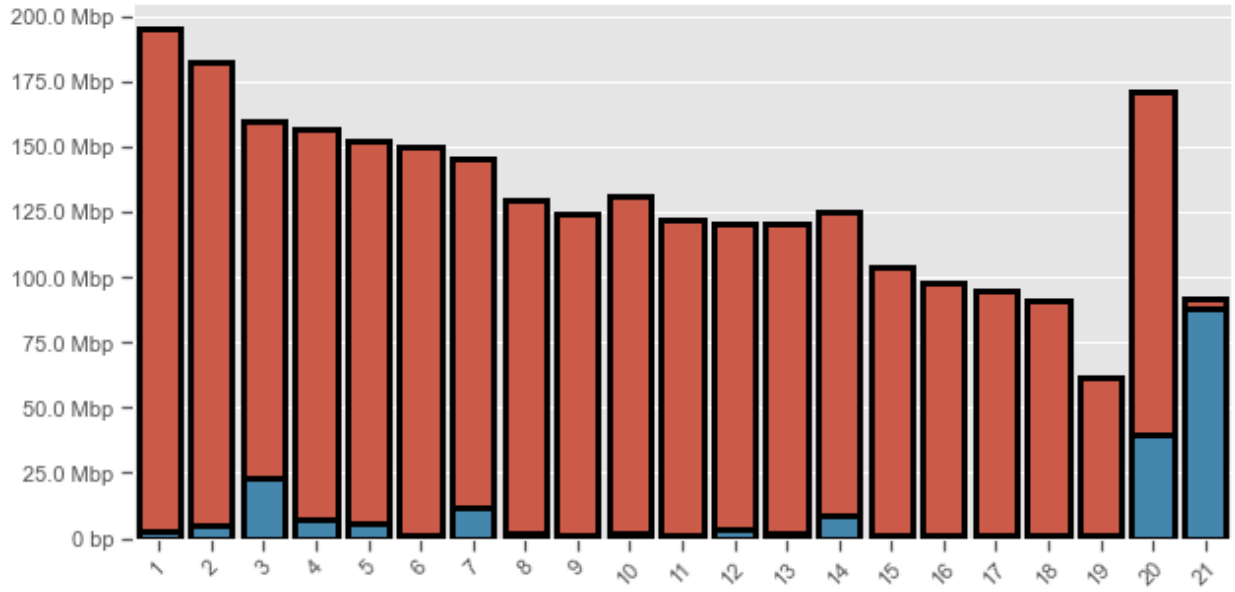


Slika 20. Korelacija između ukupnog broja detektiranih SVs i pokrivenosti genoma nakon poravnavanja. Na lijevom su grafu prikazane zasebne korelacije po tkivu, a na desnom je grafu prikazana ukupna korelacija. Iako oba grafa sadržavaju iste podatke, trend u podacima se mijenja ovisno o tome analiziramo li ih u skupinama ili kao cjelinu (tzv. Simpsonov paradoks).

3.1.3.2. Varijabilnost genoma

Kako bismo procjenili udio referentnog genoma koji je podložan strukturnim varijacijama, preklopili smo sve SVs iz svih jedinki, uključujući i 477 dijeljenih SVs, na temelju položaja SVs na referentnom genomu. Duljine regija na referentnom genomu koje nastaju takvim preklapanjem smo zbrojili i podijelili s duljinom referentnog genoma.

Ukupna duljina obuhvaćenog genoma iznosti 122.3 Mbp te dijeljenjem sa ukupnom veličinom referentnog genoma od 2.7 Gbp pronalazimo da je 4.5 % genoma C57BL/6 soja podložno strukturnim varijacijama. Ovaj izračun ne uključuje Y kromosom (kromosom 21), s obzirom na nepouzdanost rezultata koja proizlazi iz nedovoljnog kontinuiteta *de novo* sklapanja.

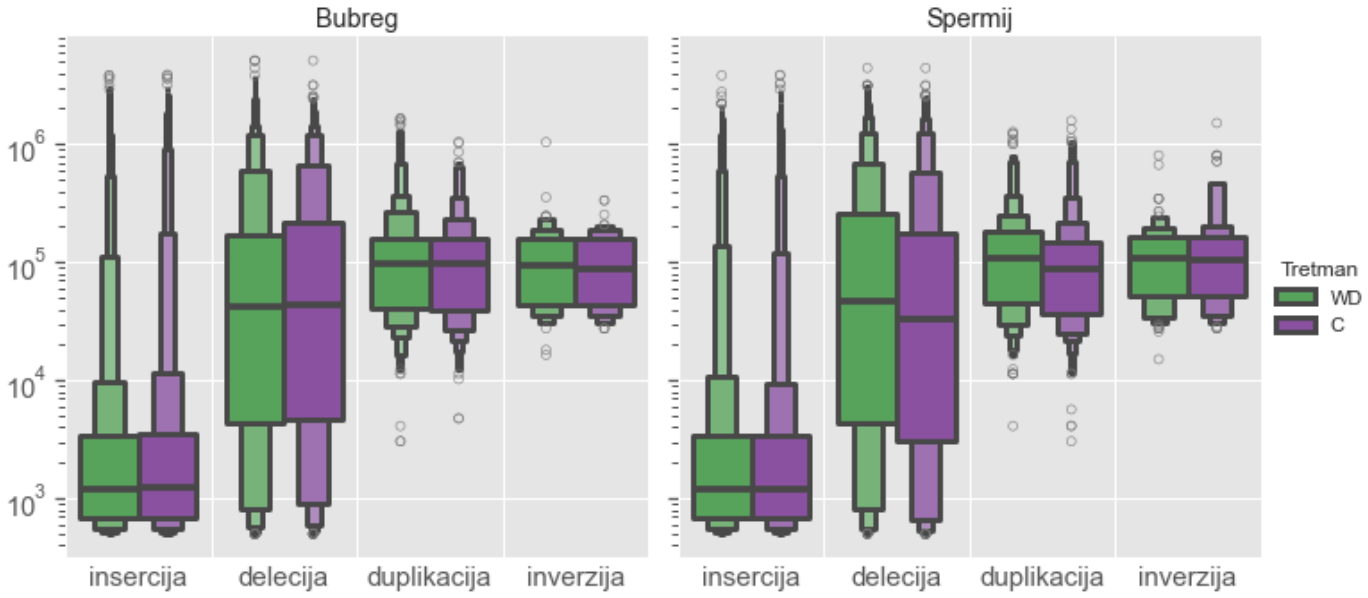


Slika 21. Količina referentnog genoma podložnog varijacijama po kromosomu. Crvenim stupcem označena je duljina kromosoma, a plavim stupcem kumulativna duljina SVs na referentnom genomu. Kromosomi X i Y su označeni kao 20 i 21.

3.1.3.3. Veličina detektiranih varijanti

Analizirali smo veličinu SVs za insercije, delecije, duplikacije i inverzije. Za translokacije oba tipa, veličina SVs je nedefinirana u izlaznoj SMAP datoteci rezultata optičkog mapiranja, pa su stoga informacije o njoj nedostupne. Delecije su najdulji tip strukturnih varijanti (u prosjeku 247.6 kbp; medijan 43.6 kbp), dok su najmanje insercije, s prosječnom duljinom od 61.9 kbp (medijan 1.2 kbp). Duljine duplikacija i delecija imaju najveće standardne devijacije, što ukazuje na veću varijabilnost (Slika 22).

Pronalazimo da su, unutar spermija, duplilkacije i delecije statistički značajno dulje u pokusnim miševima (Tablica 13). Nismo pronašli nikakve statistički značajne razlike duljina SVs između bubrega i spermija niti unutar kontrolne, niti unutar pokusne skupine.



Slika 22. Boxen dijagram distribucija duljine SVs po tipu, prikazane po tkivu i skupini.

Tablica 13. Rezultati testiranja razlike u distribuciji duljina SVs između pokusne i kontrolne skupine unutar pojedinog tkiva po tipu SVs putem Mann Whitney testa (MWU). Prikazana P vrijednost korigirana je Holmovom metodom višestruke usporedbe. Negativna CLES vrijednost označava dulje SVs u kontrolnoj skupini a pozitivna vrijednost označava dulje SVs u pokusnoj skupini.

Test	Unutar	SV	P vrijednost	CLES
MWU	Bubreg	insercija	1.0	-0.5
MWU	Bubreg	delecija	1.0	-0.51
MWU	Bubreg	duplikacija	1.0	0.5
MWU	Bubreg	inverzija	1.0	0.49
MWU	Spermij	insercija	1.0	-0.51
MWU	**Spermij**	**delecija**	**0.0356**	**0.54**
MWU	**Spermij**	**duplikacija**	**0.0466**	**0.55**
MWU	Spermij	inverzija	1.0	-0.51

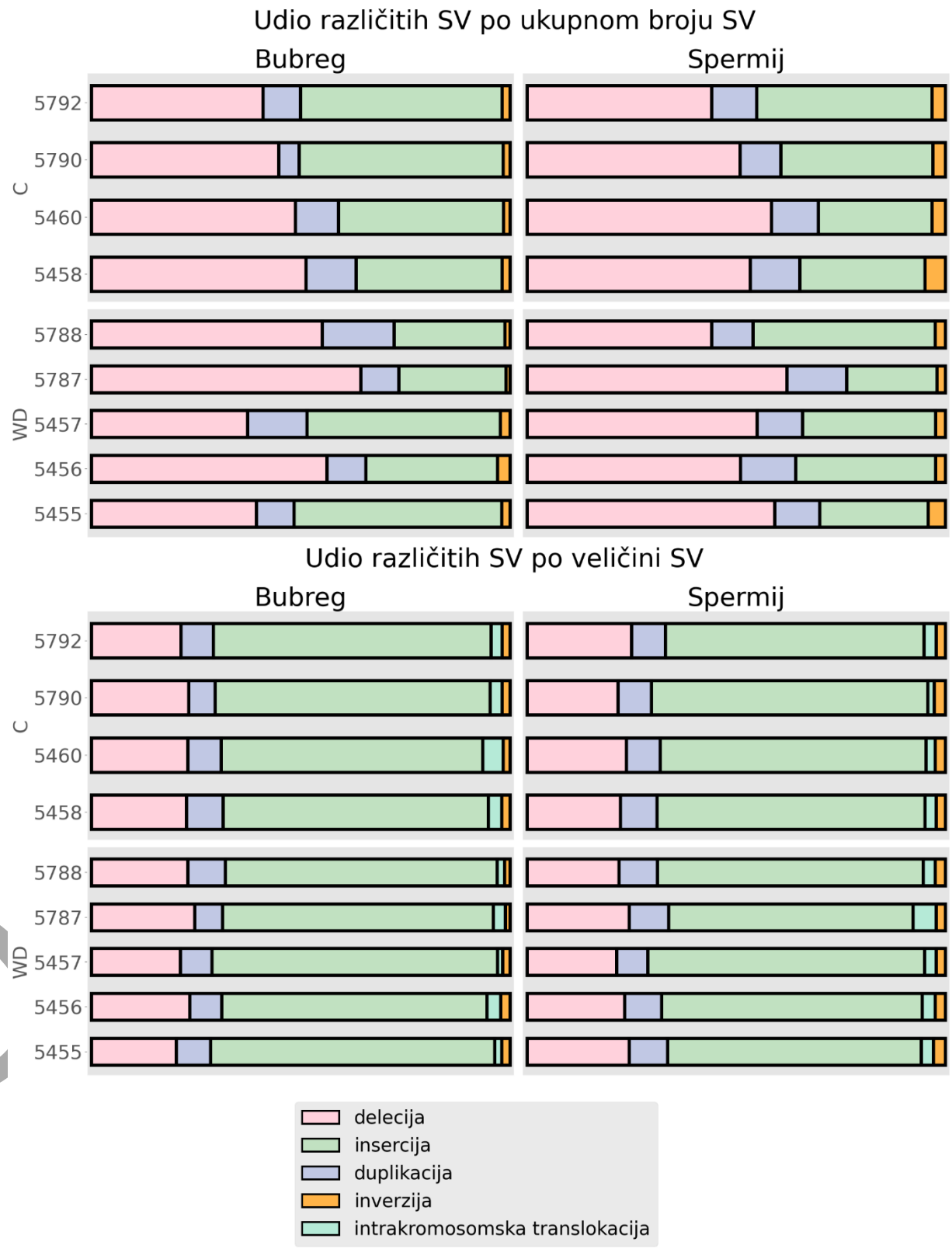
3.1.3.4. Udio strukturnih varijanta po tipu

Najčešće SVs su insercije koje čine 63.9 % svih detektiranih varijanti, a zatim slijede delecije (22.9 %), duplikacije (8.2 %), intrakromosomske translokacije (2.8 %) i inverzije (2.1 %). Međutim, s obzirom na udio u ukupnoj veličini, delecije dominiraju sa 984.1 Mbp (51.00 %) kumulativne duljine, a zatim slijede insercije s 688.6 Mbp (35.7 %) i duplikacije s 209.1 Mbp (10.8 %). Inverzije čine svega 47.9 Mbp (2.5 %) (Slika 23).

Pronalazimo statistički značajne razlike u brojčanom udjelu inverzija, duplikacija, i insercija između bubrega i spermija, ali ne pronalazimo statistički značajne razlike u brojčanom udjelu ostalih SVs niti između bubrega i spermija, ali ni niti između pokusne i kontrolne skupine (Tablica 14). Pronalazimo statistički značajne razlike u veličinskom udjelu duplikacija između bubrega i spermija, ali ne pronalazimo statistički značajne razlike u veličinskom udjelu ostalih SVs niti između bubrega i spermija, ali ni niti između skupina (Tablica 14). Međutim, ne pronalazimo nikakve statistički značajne razlike između udjela pojedinih SVs između skupina, unutar tkiva.

Tablica 14. Rezultati statističkog testiranja brojčanog i veličinskog udjela pojedinog tipa SVs između tkiva odnosno skupina. MWU se odnosi na Mann Whitney test, a Wilcoxon na Wilcoxonov test. P vrijednost prikazana korigirana je Holmovom metodom.

			Brojčani udjeli		Veličinski udjeli	
Statistički test	Između	SVs	P vrijednost	CLES	P vrijednost	CLES
Wilcoxon	Tkiva	inverzija	0.0002	0.9753	0.0011	0.9136
Wilcoxon	Tkiva	duplikacija	0.0067	0.7901	0.9064	0.6790
Wilcoxon	Tkiva	insercija	0.0239	0.7407	0.1163	0.6914
Wilcoxon	Tkiva	delecija	1.0000	0.6420	0.5428	0.6420
Wilcoxon	Tkiva	intrakromosomska translokacija	1.0000	0.5185		
MWU	Skupina	delecija	1.0000	0.5125	1.0000	0.6375
MWU	Skupina	insercija	1.0000	0.5000	0.9479	0.6750
MWU	Skupina	duplikacija	1.0000	0.6000	1.0000	0.6375
MWU	Skupina	inverzija	1.0000	0.5500	1.0000	0.5625
MWU	Skupina	intrakromosomska translokacija	1.0000	0.5250		



Slika 23. Distribucije brojčanog (gornji graf) i veličinskog (donji graf) udjela pojedinih tipova SV po tkivu i mišu.

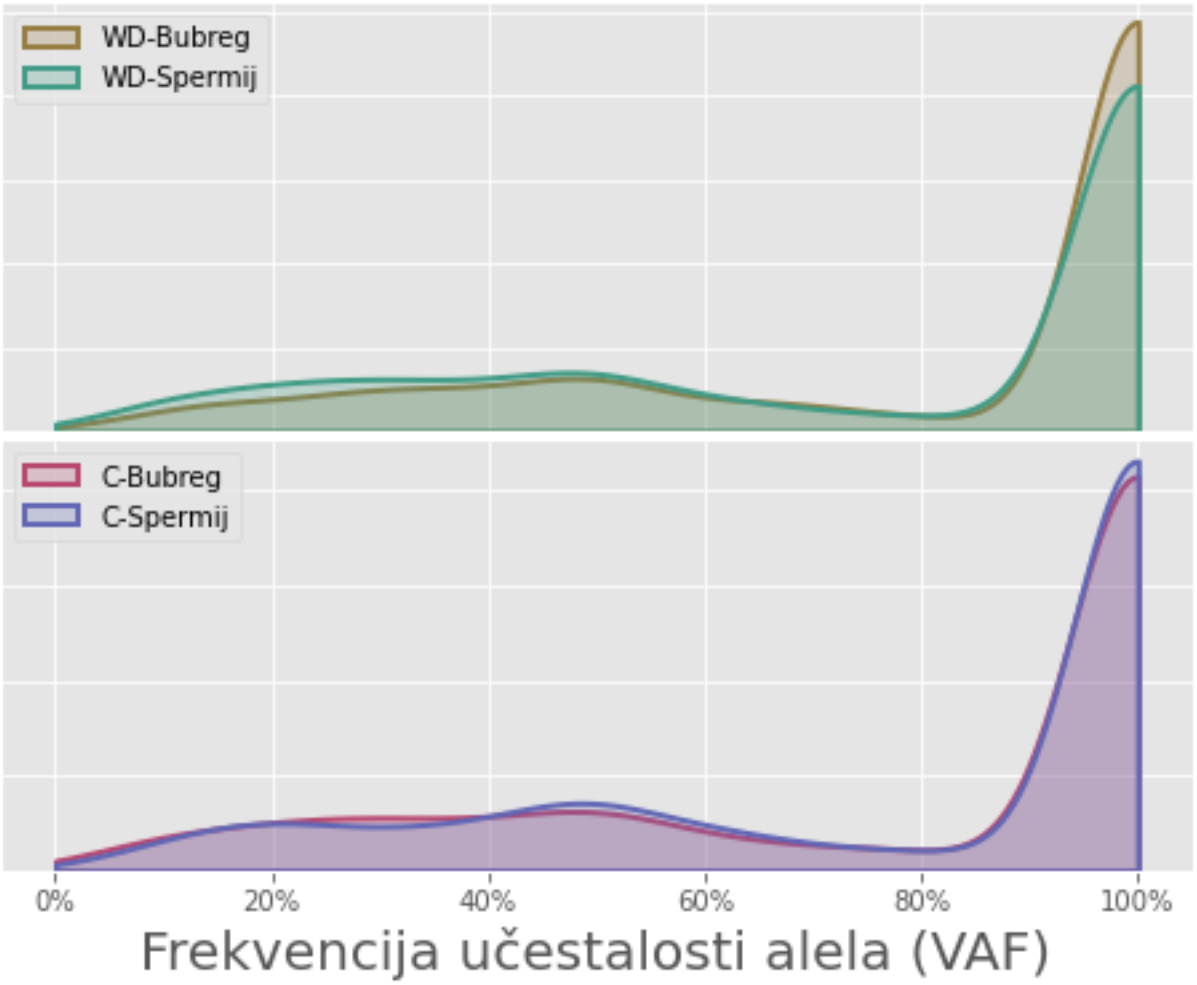
3.1.3.5. Učestalost strukturnih varijanta

Od 17 382 detektiranih SVs u cijelom skupu podataka, 6 529 imaju VAF < 0.97 te takve SVs smatramo heterozigotnima, a 10 845 imaju VAF veći ili jednak 0.97 zbog čega ih smatramo homozigotnima (Slika 24). Pet SVs imaju dodijeljenu VAF vrijednost -1, što označava nemogućnost algoritma da odredi zigotnost. Iako su homozigotne varijante skoro dvostruko češće od heterozigotnih, ukupna veličina homozigotnih varijanti je 238.2 Mbp, a heterozigotnih varijanti 1.7 Gbp. Ova razlika odraz je razlike u prosječnoj duljini homozigotnih odnosno heterozigotnih SVs: prosječna duljina homozigotnih varijanta je 22.0 kbp (medijan 1.23 kbp) a heterozigotnih 279.7.2 kbp (medijan 60.7 kbp).

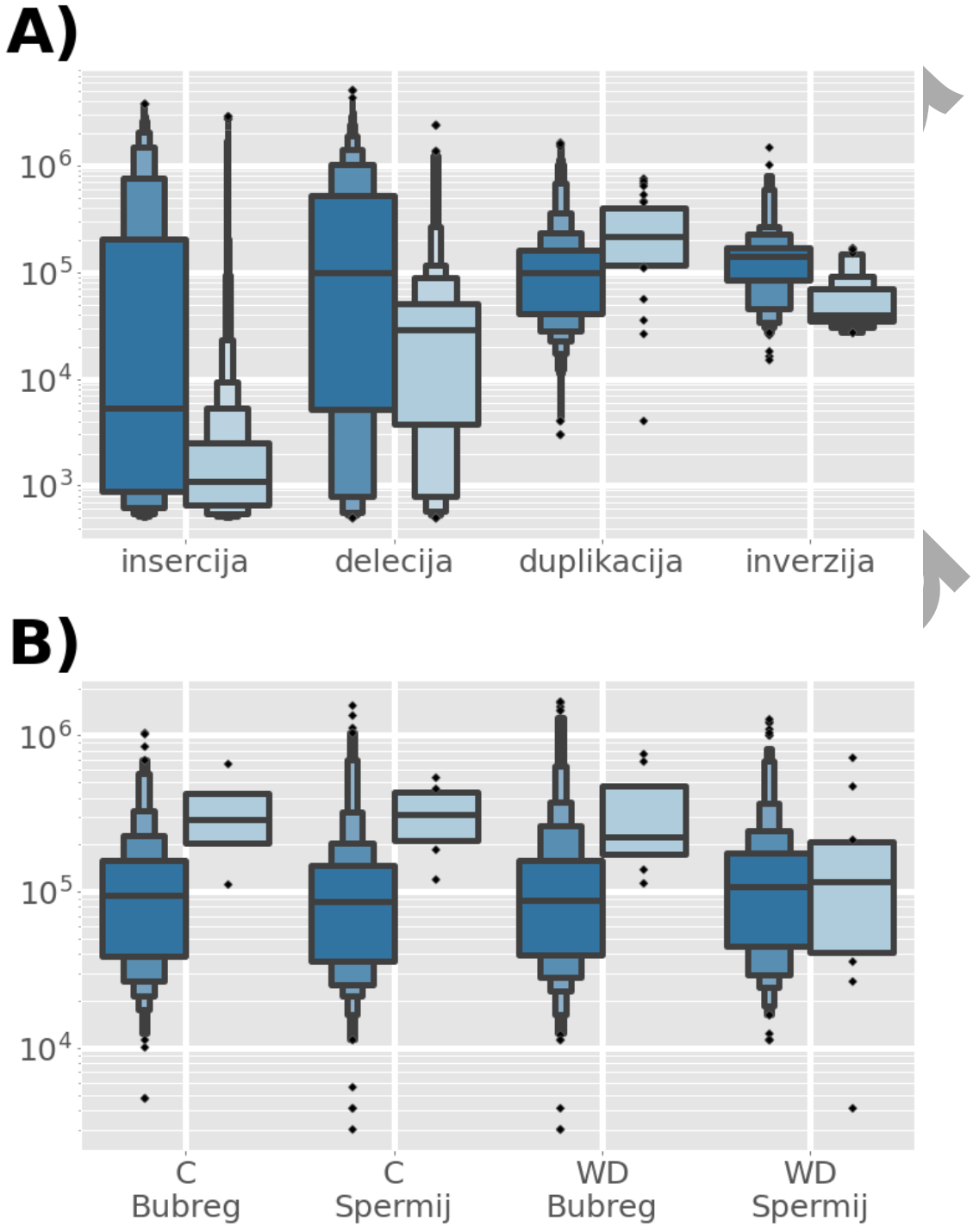
Od ukupno 1 420 detektiranih duplikacija, svega 27 (2 %) su homozigotne. Slično duplikacijama, inverzije su većinom heterozigotne (268 odnosno 72 % naspram 102 odnosno 28 % homozigotnih). Insercije su uglavnom homozigotne (8 984 odnosno 81 % naspram 2 135 odnosno 19 % heterozigotnih). Delecije su otprilike podjednako raspodijeljene – 2 252 odnosno 57 % su heterozigotni, a 1 719 odnosno 43% su homozigotni.

Kod svih tipova SVs, homozigotni SVs su značajno kraći od heterozigotnih, osim homozigotnih duplikacija koje su oko tri puta duže od heterozigotnih (Slika 25A), u bubregu životinja obje skupine i u spermijima kontrolne skupine, ali ne i spermijima pokusne skupine (Slika 25 B), u kojima su homozigotne i heterozigotne duplikacije slične veličine.

Kako bismo procijenili na koji udio referentnog genoma utječu homozigotne odnosno heterozigotne SVs, preklopili smo sve SVs iz svih jediniki na temelju položaja SVs na referentnom genomu. Količinu referentnog genoma podložnog varijacijama računali smo kao zbroj duljina regija na referentnom genomu koje nastaju takvim preklapanjem. Heterozigotni SVs zahvaćaju 103.3 Mbp, odnosno 3.8 % referentnog genoma, a homozigotni SVs (bez 477 dijeljenih SVs) pogađaju 60.3 Mbp, odnosno 2.2 % referentnog genoma. Uključivanjem dijeljenih SVs, ukupna količina referentnog genoma koje homozigotne SVs zauzimaju je 142.6 Mbp, odnosno 5.2 % referentnog genoma.



Slika 24. Distribucija učestalosti alela po grupama.



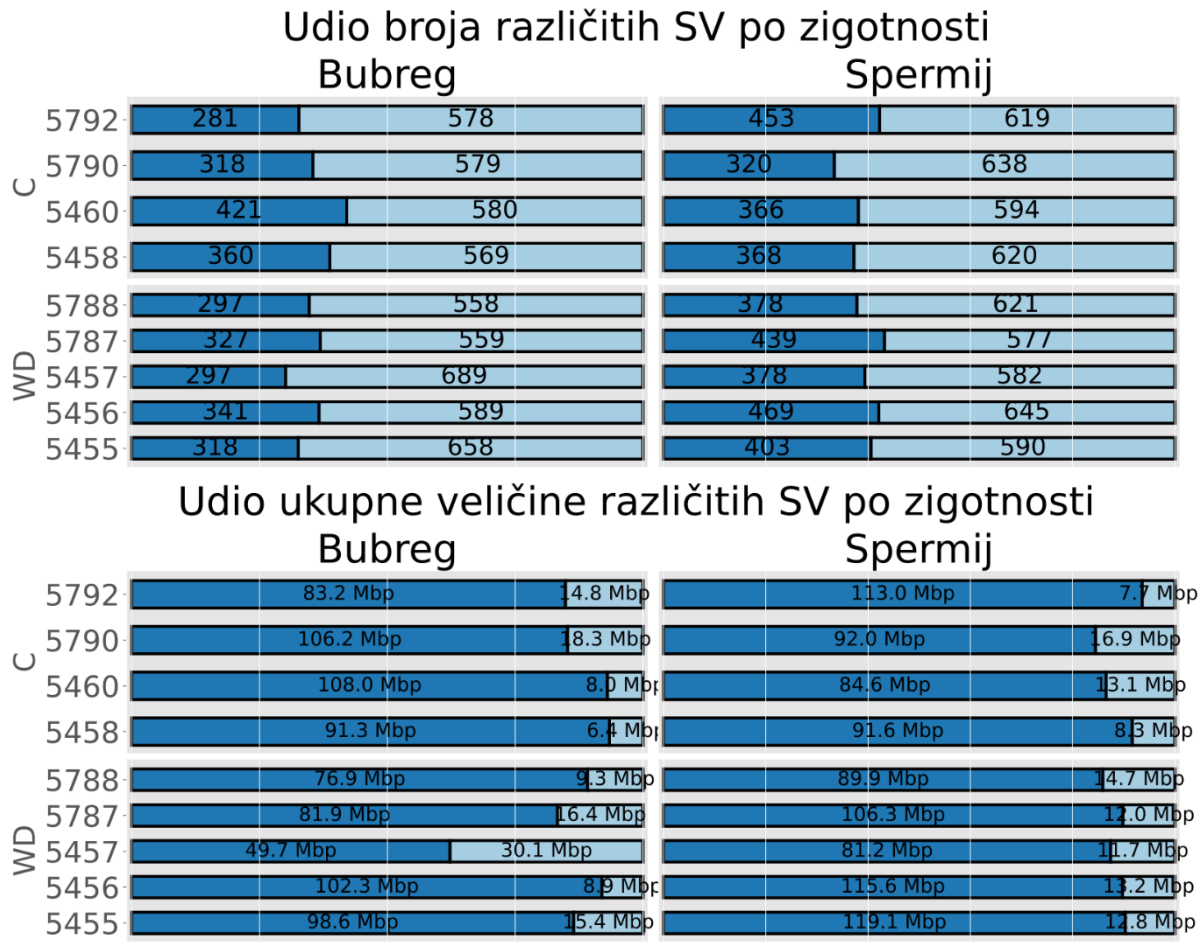
Slika 25. Boxen dijagram veličine SVs po tipu i zigotnosti. Tamno plavo predstavlja heterozigotne SVs a svijetlo plavo homozigotne. Na slici su prikazane veličine SVs po tipu u svim uzorcima (graf A), te veličine duplikacija po grupama uzoraka (graf B).

Pronalazimo da je broj heterozigotnih SVs generalno veći u spermijima u usporedbi s tkivom bubrega (Wilcoxonov test, p vrijednost 0.037, CLES 0.83, Cohen D 1.13). Ova razlika je vođena isključivo pokusnom skupinom (Wilcoxonov test, p vrijednost 0.031, CLES 1, Cohenov D 2.53), u kojoj miševi imaju za 6.4 % više heterozigotnih SVs u spermijima nego u tkivu bubrega (Tablica 15) (Slika 26). U kontrolnoj skupini (C) ne nalazimo razliku u prosječnom udjelu heterozigotnih SVs između tkiva (Wilcoxnov test, p vrijednost 0.688 Cohenov D 0.13, CLES 0.56).

Ne pronalazimo nikakve statistički značajne razlike u veličinskom udjelu homozigotnih SVs između tkiva generalno, niti između tkiva unutar pokusne i kontrolne skupine.

Tablica 15. Prosječan brojčani i veličinski udio heterozigotnih SVs i njihove standardne devijacije.

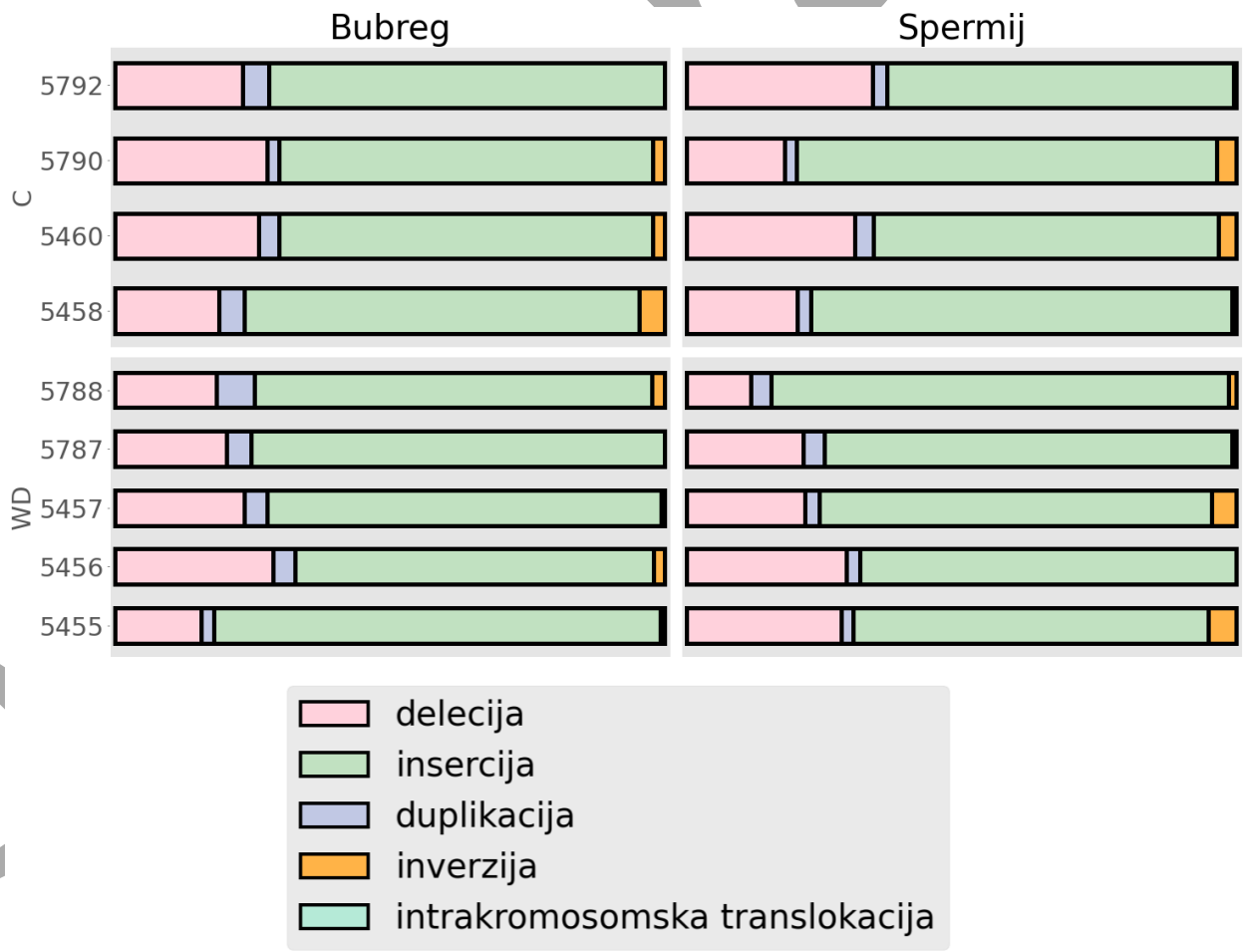
		Brojčani udio		**Veličinski udio**	
		Prosječan udio **heterozigota** /%	Standardna devijacija /%	Prosječan udio **heterozigota** /%	Standardna devijacija /%
Tretman	**Tkivo**				
C	**Bubreg**	37.23	4.04	89.16	4.73
C	**Spermij**	37.74	3.63	89.09	4.28
WD	**Bubreg**	34.19	2.88	82.64	11.83
WD	**Spermij**	40.60	2.12	88.66	1.89
-	**Bubreg**	35.54	3.58	85.54	9.5
-	**Spermij**	39.33	3.08	88.85	2.95
C	-	37.48	3.57	89.13	4.17
WD	-	37.39	4.14	85.65	8.59



Slika 26. Distribucija brojčanog (gornji graf) i veličinskog (donji graf) udjela zigotnosti varijanti po uzorcima. Tamno plavo predstavlja heterozigotne a svijetlo plavo homozigotne SVs.

3.1.3.6. *De novo* strukturne varijante

De novo SV definiramo kao SV koji je detektiran u jednom tkivu, a ne preklapa se ni sa jednim SVs koji je detektiran u drugom tkivu iste životinje. Detektirali smo između 228 i 324 *de novo* SVs (Tablica 16). Nismo pronašli nikakvu statistički značajnu razliku broja *de novo* SVs između pokusne i kontrolne skupine u bubrežnom tkivu (Mann Whitney test, p vrijednost 0.55) niti u spermijima (Mann Whitneytest, p vrijednost 0.55). Najčešći tip *de novo* SVs su insercije, a najrjeđi su inverzije (Tablica 16). Dvije interkromosomske translokacije (jedna u bubregu miša 5787 iz WD skupine, druga u bubregu miša 5460 iz kontrolne skupine) od sveukupno tri su *de novo* SVs. Nismo pronašli nikakve statistički značajne razlike između pokusne i kontrolne skupine u udjelu de novo SVs ni u jednom tipu tkiva ni po jednom tipu SVs (Slika 27).



Slika 27. Udio različitih tipova de novo SVs po uzorku.

Tablica 16. Broj detektiranih *de novo* SVs. Tablica ne uključuje dvije interkromosomske translokacije u bubrezima miševa 5460 i 5787.

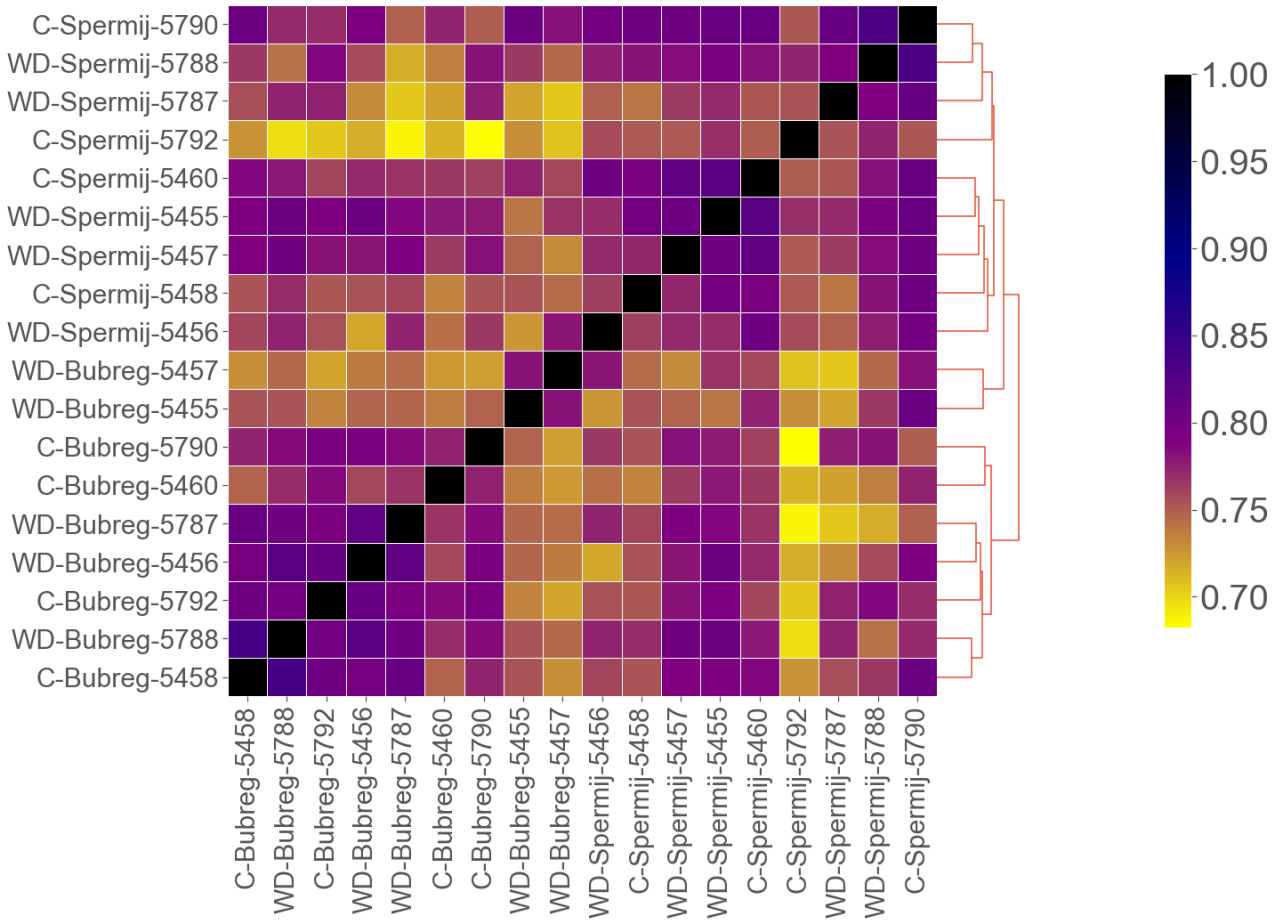
Tretman	Tkivo	Miš	delecija	duplikacija	insercija	intrakromosomska translokacija	inverzija	Ukupno
C	**Bubreg**	**5458**	16	4	61	11	4	96
C	**Spermij**	**5458**	26	3	99	3	1	132
C	**Bubreg**	**5460**	35	5	91	10	3	145
C	**Spermij**	**5460**	37	4	76	1	4	122
C	**Bubreg**	**5790**	26	2	64	5	2	99
C	**Spermij**	**5790**	25	3	107	3	5	143
C	**Bubreg**	**5792**	20	4	62	7	0	93
C	**Spermij**	**5792**	66	5	121	3	1	196
WD	**Bubreg**	**5455**	21	3	109	2	1	136
WD	**Spermij**	**5455**	27	2	62	4	5	100
WD	**Bubreg**	**5456**	29	4	66	7	2	108
WD	**Spermij**	**5456**	60	5	142	9	0	216
WD	**Bubreg**	**5457**	40	7	122	5	1	175
WD	**Spermij**	**5457**	24	3	80	5	5	117
WD	**Bubreg**	**5787**	23	5	85	2	0	116
WD	**Spermij**	**5787**	27	5	95	3	1	131
WD	**Bubreg**	**5788**	16	6	63	1	2	88
WD	**Spermij**	**5788**	16	5	114	4	2	141

3.1.3.7. Sličnosti na temelju prisutnosti strukturnih varijanti

Sličnost između dva uzorka je procijenjena na temelju SVs koji se nalaze u oba uzorka – ukoliko se položaji SVs na referentnom genomu dvaju jedinki recipročno preklapaju s barem 50 % svoje duljine onda taj SVs smatramo istim u obje jedinke.

Prema takvoj definiciji, izračunali smo da dvije jedinke dijele u prosjeku 1 478 SVs, odnosno 76.59 % ukupno detektiranih SVs pojedinog miša u prosjeku. Relativna standardna devijacija broja dijeljenih SVs je 1.9 % što ukazuje na visoku sličnost između uzoraka. Nismo našli značajne razlike između kontrolne i pokusne skupine, niti između bubrega i spermija (Slika 28).

Na sličan smo način računali prosječan broj SVs koji nisu zajednički između dva uzoraka – SV nije zajednički između dvije jedinke ako se SV iz prve jedinke nimalo ne preklapa ni sa jednim od SVs iz druge jedinke. U prosjeku, bilo koja dva uzorka razlikuju se u postojanju 251 (26 %) SVs, odnosno čine 4.5 Mbp.



Slika 28. Matrica sličnosti temeljena na prosječnom broju zajedničkih SVs između dva uzorka. Matrica udaljenosti (1 – matrica sličnosti) podvrgnuta je hijerarhijskom aglomerativnom grupiranju koristeći Wardovu metodu. Pojedinim kvadratom predstavljena je sličnost između dvije jedinke. Crna boja kvadrata predstavlja identičnost a povećanje sličnosti je prikazano pojačanjem nijansi od žute do ljubičaste (predstavljeno shemom boja s desne strane). Jedinke su razvrstane istim redoslijedom s lijeva na desno kao i odozdo prema gore.

3.1.3.8. Preklapanje s genima

Od 17 382 ukupno detektiranih SVs, njih 10 814 (62 %) preklapa barem jedan gen, djelomično ili cijelom duljinom. Gledano po genomu, u prosjeku 63 % SVs preklapaju barem jedan gen (Tablica 17) te nema razlike u sadržaju gena između pokusne i kontrolne skupine (Mann Whitney test, p vrijednost 0.23) ili tkiva (Mann Whitney test, p vrijednost 0.53). Ukupan broj gena koje preklapaju SVs je 3 276.

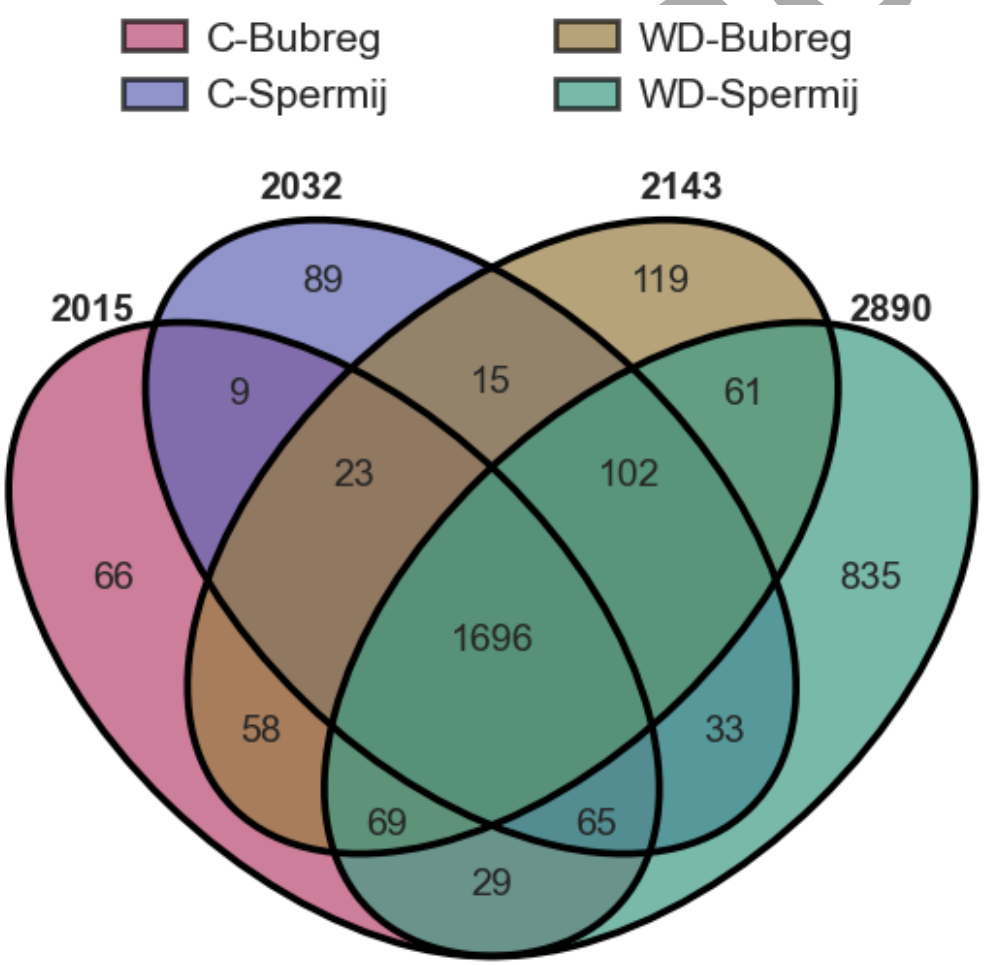
Od tih 3 276 gena, 1 454 (44 %) su geni koji kodiraju za proteine, 1 063 (33 %) su pseudogeni, a 626 (19%) su geni koji kodiraju za nekodirajuću RNA. Ostatak (manje od 4 %) čine predviđeni geni nepoznate funkcije, geni za imunoglobuline („IG gene“) te geni za receptore T stanica („TR gene“).

Ne pronalazimo razliku udjela SVs koje sadrže gene između uzoraka kontrolne i pokusne skupine (Mann Whitney test, p vrijednost 0.460) niti između tkiva (Wilcoxon test, p vrijednost 0.426). Međutim, SVs genoma unutar pokusne skupine sadrže statistički značajno više gena (Mann Whitney test, p vrijednost 0.043, CLES 0.788) nego SVs genoma unutar kontrolne skupine. Također, SVs iz bubrežnog tkiva sadrže statistički značajno manje gena nego SVs iz spermija (Wilcoxon test, p vrijednost 0.04, CLES 0.654).

Tablica 17. Pregled preklapanja SVs s genima po uzorku. U zagradi su navedeni udjeli naspram ukupnog broja SVs u uzorku odnosno naspram ukupnog broja gena (3 276) detektiranih u cijelom skupu podataka.

Tretman	Tkivo	Miš	SVs koje sadrže gen(e)	Broj gena
C	**Bubreg**	**5458**	567 (61.03 %)	1442 (44.11 %)
C	**Spermij**	**5458**	609 (61.64 %)	1479 (45.24 %)
C	**Bubreg**	**5460**	630 (62.87 %)	1702 (52.06 %)
C	**Spermij**	**5460**	583 (60.73 %)	1475 (45.12 %)
C	**Bubreg**	**5790**	568 (63.32 %)	1402 (42.89 %)
C	**Spermij**	**5790**	599 (62.53 %)	1606 (49.13 %)
C	**Bubreg**	**5792**	543 (63.21 %)	1380 (42.21 %)
C	**Spermij**	**5792**	641 (59.79 %)	1460 (44.66 %)
WD	**Bubreg**	**5455**	609 (62.4 %)	1517 (46.41 %)
WD	**Spermij**	**5455**	635 (63.95 %)	1719 (52.58 %)
WD	**Bubreg**	**5456**	571 (61.4 %)	1624 (49.68 %)
WD	**Spermij**	**5456**	686 (61.58 %)	1622 (49.62 %)
WD	**Bubreg**	**5457**	614 (62.27 %)	1439 (44.02 %)
WD	**Spermij**	**5457**	579 (60.31 %)	1591 (48.67 %)
WD	**Bubreg**	**5787**	554 (62.39 %)	1671 (51.12 %)
WD	**Spermij**	**5787**	647 (63.68 %)	2374 (72.62 %)
WD	**Bubreg**	**5788**	538 (62.92 %)	1580 (48.33 %)
WD	**Spermij**	**5788**	641 (64.16 %)	1553 (47.51 %)

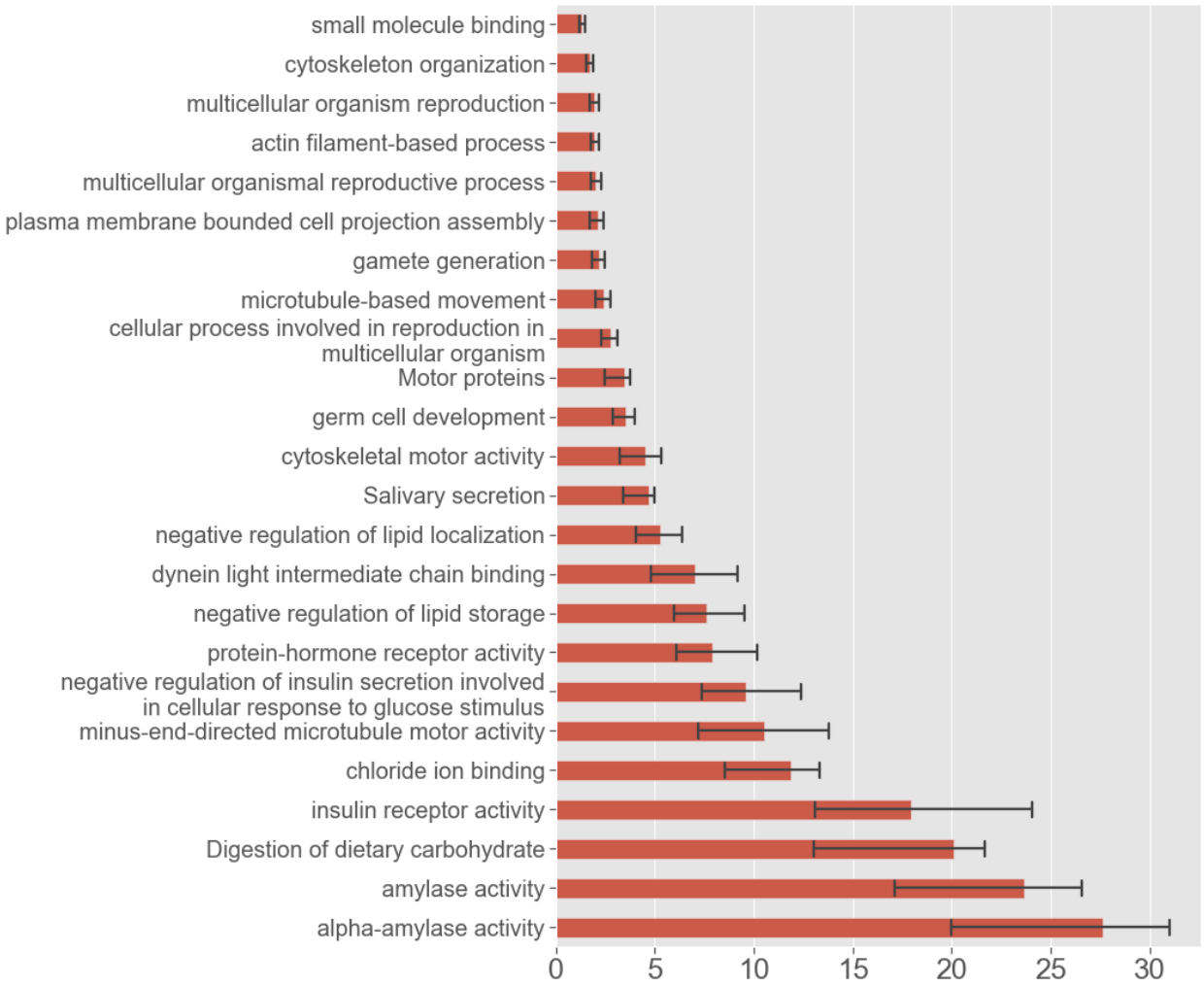
Od ukupno 3 276 gena, oko polovicu (1 696) gena preklapaju SVs iz svih promatranih kombinacija tkiva i skupina (Slika 29). U slučaju 835 gena preklapanje je jedinstveno za spermije miševa iz pokusne skupine, što je 7-13 puta više od ostalih grupa (66 gena koji se preklapaju isključivo u bubrezima kontrolne skupine; 89 u spermijima kontrolne skupine; 119 u bubrezima pokusne skupine). Međutim, od ovih 835 gena, 725 gena (87 %) proizlaze is miša 5787, te je od tih 725 gena čak 702 (96 %) isključivo mišu 5787.



Slika 29. Venov dijagram broja gena koji SVs preklapaju.

3.1.3.9. Analize ontologije gena

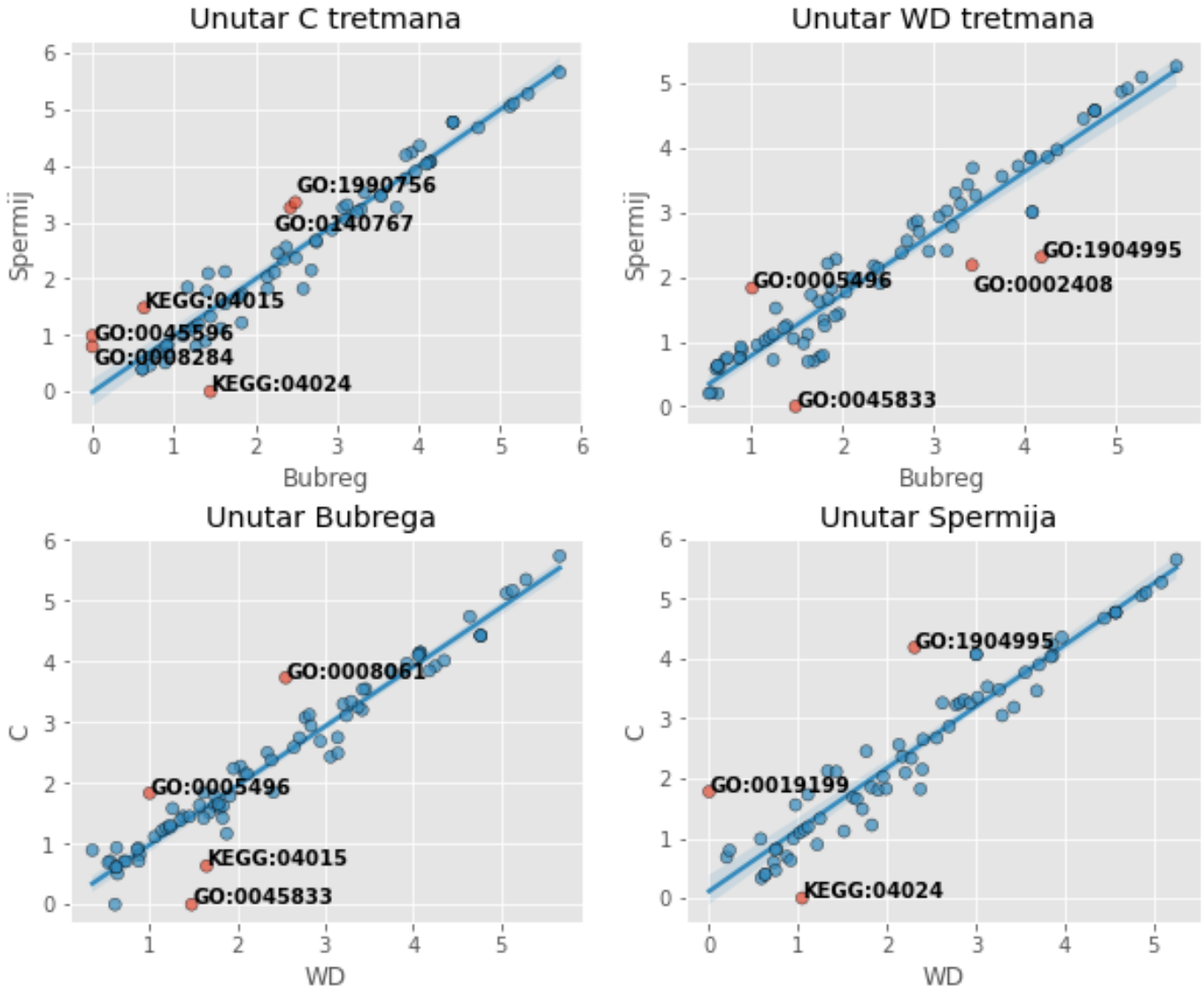
Kako bismo opisali funkcionalni sadržaj SVs u našem skupu podataka, analizirali smo gene koji su pogođeni sa SVs, s obzirom na njihovu ontologiju. U ovaj smo skup podataka uključili i 477 SVs koje su zajedničke svim uzorcima. Pronašli smo ukupno 21 termina koji su statistički značajno obogaćeni (p vrijednost nakon FDR korekcije <0.05) u svim genomima. Ovi termini se nalaze pri visokom obogaćenju u svim promatranim genomima (Slika 30) te su uglavnom vezani uz procese kao što su metabolizam šećera, pohranu lipida, aktivnost citoskeleta te reprodukciju. Procesi vezani uz aktivnost amilaze i aktivnosti inzulinskih receptora nalaze se pri najvećem prosječnom obogaćenju.



Slika 30. Termini ontologije gena koji su statistički značajno obogaćeni u SVs svih uzoraka. Crveni stupci predstavljaju prosječna obogaćenja, a crne linije raspone od minimalne do maksimalne vrijednosti obogaćenja.

Kako bismo odredili dolazi li do značajne promjene u funkcionalnom sadržaju SVs kod spermija uslijed prehrane bogate mastima, podijelili smo podatke iz uzoraka na četiri grupe: spermiji iz pokusne skupine, spermiji iz kontrolne skupine, tkivo bubrega iz pokusne skupine, i tkivo bubrega iz kontrolne skupine. Usporedili smo obogaćenost termina ontologije gena a) u spermijima između pokusne i kontrolne skupine, te b) između spermija i bubrega unutar pokusne skupine. Također smo usporedili obogaćenost termina ontologije gena a) u bubrezima između pokusne i kontrolne skupine, te b) između spermija i bubrega unutar kontrolne skupine. Pri svakoj usporedbi između dvije grupe uzeli smo u obzir samo one termine koji su značajno ($\alpha = 0.05$) obogaćeni (u SVs) u najmanje tri uzorka iz barem jedne od dvije grupe. Većina termina pokazuje visoku korelaciju između grupa, odnosno slično obogaćenje termina. Međutim, nekoliko termina značajno odstupa od pravca linearne regresije (Slika 31, Tablica 18).

Termini čije obogaćenje značajnije varira između grupa povezani su s generalnim biološkim procesima kao što su stanična komunikacija, unutarstanični prijenos signala, adhezija stanica, stanični ciklus, diferencijacija stanica, i upalni proces. Jedini termin koji je u prosjeku obogaćeniji (2x) u spermijima pokusne nego u spermijima kontrolne skupine je KEGG:04024, a odnosi se na signalni put (*cAMP signaling pathway*) koji sudjeluje u mnogim fiziološkim procesima, kao što su rast, razmnožavanje, diferencijacija i apoptoza (Yan i sur., 2016). Međutim, ovaj termin također pokazuje slično povećanje obogaćenja u bubrezima životinja iz kontrolne skupine u odnosu na spermije (Slika 31), što sugerira da povećanje obogaćenja u spermijima pokusne skupine nije uzrokovano pokusom. Unutar pokusne skupine, jedino značajnije povećanje obogaćenja u spermijima u odnosu na bubreg vezano je za termin GO:0005496 (*steroid binding*). Međutim, taj termin također pokazuje slično veće obogaćenje (2x) u bubrezima životinja iz kontrolne skupine u odnosu na pokusnu skupinu (Slika 31), što sugerira da povećanje obogaćenja u spermijima pokusne skupine nije povezano s utjecajem prehrane bogate mastima. Što se tiče promjena obogaćenja u tkivu bubrega koje bi mogle biti uzrokovane pokusom, pronalazimo dva termina koja su značajnije obogaćenija u pokusnoj skupini u odnosu na kontrolnu skupinu: KEGG:04015 (*Rap1 signaling pathway*) i GO:0045833 (*negative regulation of lipid metabolic process*). Slično obogaćenje termina KEGG:04015 nalazimo i u spermijima kontrolne skupine, što navodi na zaključak da porast obogaćenja u bubregu nije nužno efekt pokusa. Termin GO:0045833 se odnosi na metabolizam lipida i nalazi se statistički značajno obogaćen isključivo u tri uzorka, a koji su izolirani iz bubrega miševa pokusne skupine (ukupno pet životinja) (Tablica 18). Obogaćenje je u prosjeku podržano s 9 gena i 8 strukturnih varijanti koje ih pogađaju. Specifično povećanje njegova obogaćenja u bubregu pokusnih životinja bi moglo biti povezano s promjenama u regulaciji metaboličkih procesa kod životinja na masnoj hrani, a koje su uzrokovane strukturnim varijacijama.



Slika 31. Korelacija obogaćenja termina između pojedinih grupa (vrijednosti obogaćenja su logaritmirane po bazi 2). Narančastom su obojani termini koji se nalaze 1.96 standardnih devijacija od regresijskog pravca, što odgovara p vrijednosti od 0.05.

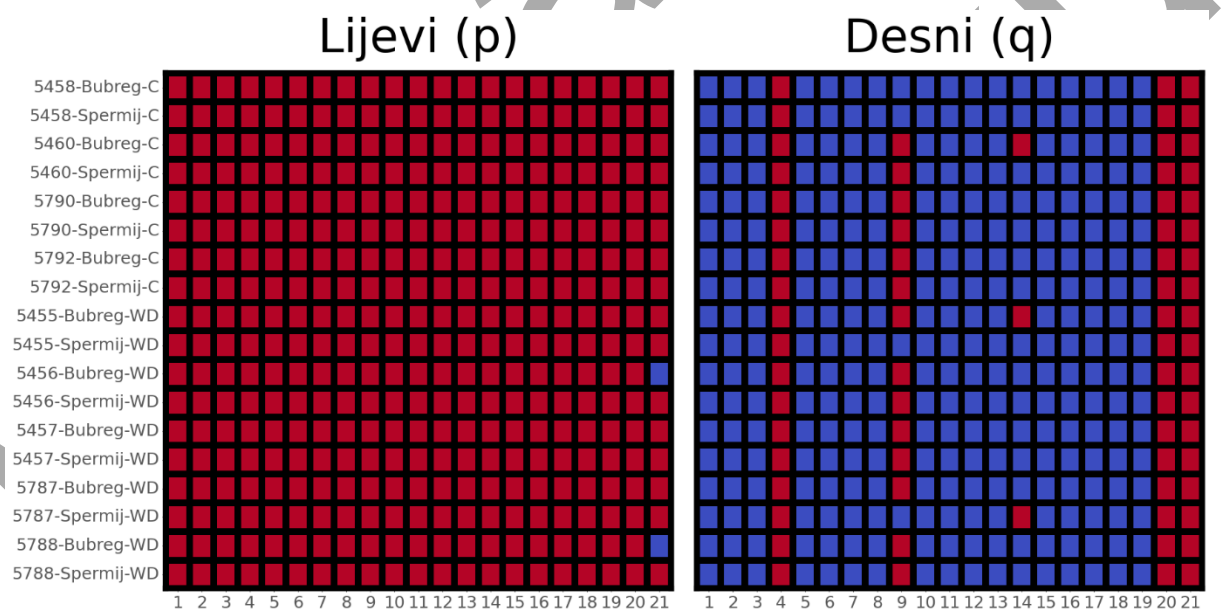
Tablica 18. Termini ontologije gena sa značajnim razlikama u obogaćenju između grupa uzoraka. Termini u stupcu „ID termina“ odnose se na točke koje značajno odstupaju od pravca linearne regresije na Slici 31. Stupci „Bubreg“ i „Spermij“ te podstupci „C“ i „WD“ odnose se na ukupan broj uzoraka u kojem je termin statistički značajno obogaćen. Stupci „Broj gena“ i „Broj SVs“ predstavlja prosječan broj gena odnosno SVs po uzorku (± jedna standardna devijacija) koji se nalaze podložni varijacijama, odnosno utječu na gene vezane uz specifičan termin.

ID termina	Termin	Broj gena	Broj SVs	Bubreg		Spermij	
				C	WD	C	WD
GO:0008284	positive regulation of cell population proliferation	43.80 ± 4.51	34.90 ± 3.05	0	3	4	3
GO:0045596	negative regulation of cell differentiation	31.75 ± 1.09	20.75 ± 1.92	0	0	3	1
GO:0140767	enzyme-substrate adaptor activity	18.53 ± 3.26	8.18 ± 2.41	3	5	4	5
GO:1990756	ubiquitin ligase-substrate adaptor activity	17.53 ± 3.26	7.18 ± 2.41	3	5	4	5
KEGG:04015	Rap1 signaling pathway	10.86 ± 2.10	11.64 ± 1.34	1	5	3	5
KEGG:04024	cAMP signaling pathway	10.90 ± 1.22	13.20 ± 1.40	3	4	0	3
GO:0002408	myeloid dendritic cell chemotaxis	3.50 ± 0.50	2.12 ± 1.17	2	3	2	1
GO:0005496	steroid binding	10.75 ± 3.00	6.42 ± 1.04	4	2	2	4
GO:0045833	negative regulation of lipid metabolic process	9.33 ± 0.47	8.33 ± 1.89	0	3	0	0
GO:1904995	negative regulation of leukocyte adhesion to vascular endothelial cell	3.31 ± 0.46	1.69 ± 1.07	3	5	4	1
GO:0008061	chitin binding	3.27 ± 0.86	1.45 ± 0.66	4	2	3	2
GO:0019199	transmembrane receptor protein kinase activity	7.33 ± 0.47	6.17 ± 2.67	2	1	3	0

3.1.4. Duljina telomera

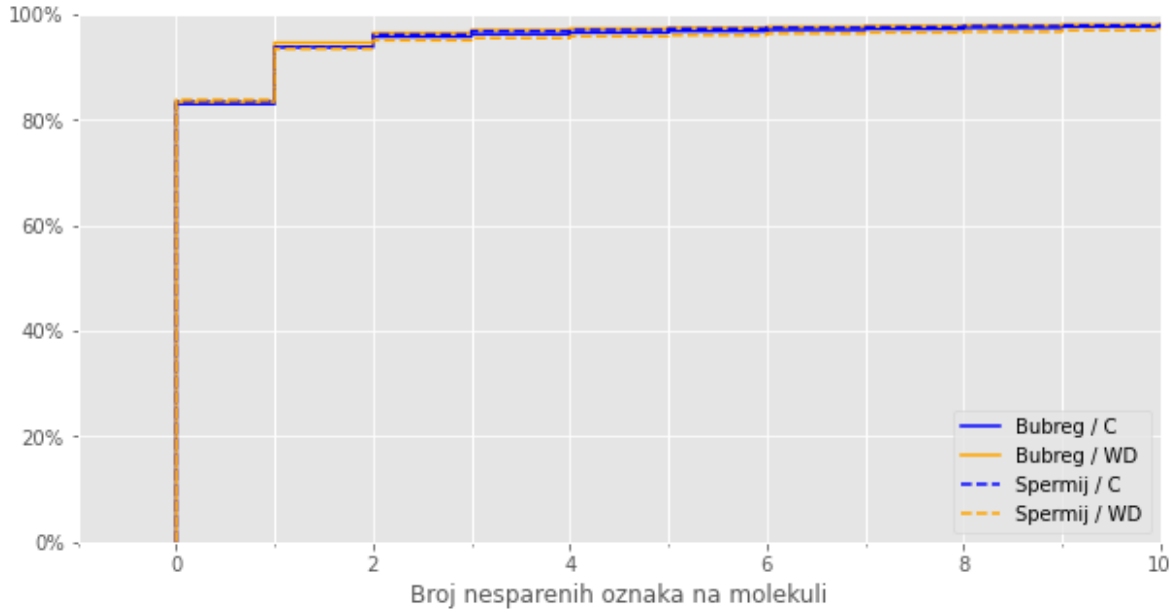
3.1.4.1. Prikladnost podataka

Informacije o duljini telomera dostupne su jedino iz molekula i kontiga koji su poravnati s referentnim genomom tako da zadovoljavaju kriterije definirane u poglavlju 2.1.4.3. S obzirom na te kriterije, nije bilo moguće odrediti duljine telomera na lijevom (p) kraku gotovo svakog od kromosoma u svim uzorcima (Slika 32). Mišji kromosomi građeni su akrocentrično tako da se između telomere na p kraku i centromere nalaze regije nepoznate sekvence i duljine koje su u referentnom genomu anotirane kao praznine arbitrarno postavljene duljine 10 kbp. U tim regijama ne nalazimo oznake koje bi odgovarale obilježenim motivima za DLE-1 enzim. Molekule dobivene optičkim mapiranjem nisu dovoljne duge da bi premostile regije telomera i centromera p kraka te pružile informacije o duljini telomera. Slično je i s desnim (q) krakom spolnih kromosoma i kromosoma 4. Kromosomi 9 i 14 nekih uzoraka imaju jednu ili više nesparenih oznaka na referenci poslije zadnje sparene oznake u određenim uzorcima (Slika 32). Ti kromosomski krakovi stoga nisu uzeti u obzir prilikom svih daljnjih analiza telomera.



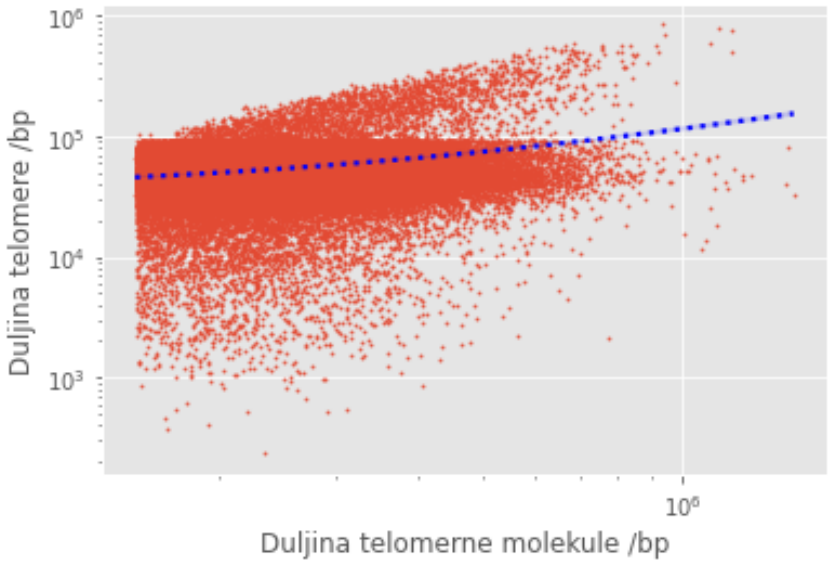
Slika 32. Pregled kromosomskih krakova koji su prikladni za određivanje duljine telomera, po uzorku. Crvenom bojom su označeni krakovi kromosoma na kojima niti jedna molekula ne udovoljava kriterijima potrebnim za određivanje duljine telomera, a plavom bojom su označeni krakovi kromosoma na kojima molekule zadovoljavaju potrebne kriterije. Prikazani podaci odnose se na autosome 1-19 te spolne kromosome (20 i 21) na lijevom (p) i desnom (q) kraku.

Među podacima koji su uzeti u obzir za analizu duljine telomera, izračunali smo udio molekula s obzirom na broj nesparenih oznaka kako bismo procijenili stupanj poravnanja na subtelomernim regijama. Gledano po broju oznaka na krajevima kontiga, oko 90 % molekula podržava kontige čija je posljednja oznaka poravnata s posljednjom oznakom na referentnom genomu, odnosno oko 10 % molekula je poravnato s kontigom koji ima samo jednu nesparenu oznaku na samom kraju. Iznimke su podaci iz tkiva bubrega na uzorcima 5460 i 5792 (kontrolni miševi) gdje oko trećinu svih telomernih molekula čine one telomerne molekule koje su poravnate s kontizima koji imaju po jednu nesparenu oznaku na svojem kraju. Gledano po broju oznaka na krajevima molekula, više od 80 % svih molekula imaju posljednju oznaku poravnatu s posljednjom oznakom na referenci, odnosno oko 98 % svih molekula ima manje od 4 oznake koje nisu poravnate s posljednjom oznakom na referentnom genomu (Slika 33). Ovi podaci sugeriraju da je stupanj poravnanja molekula na kraju q kraka većine kromosoma prikladan za analize duljine telomera.

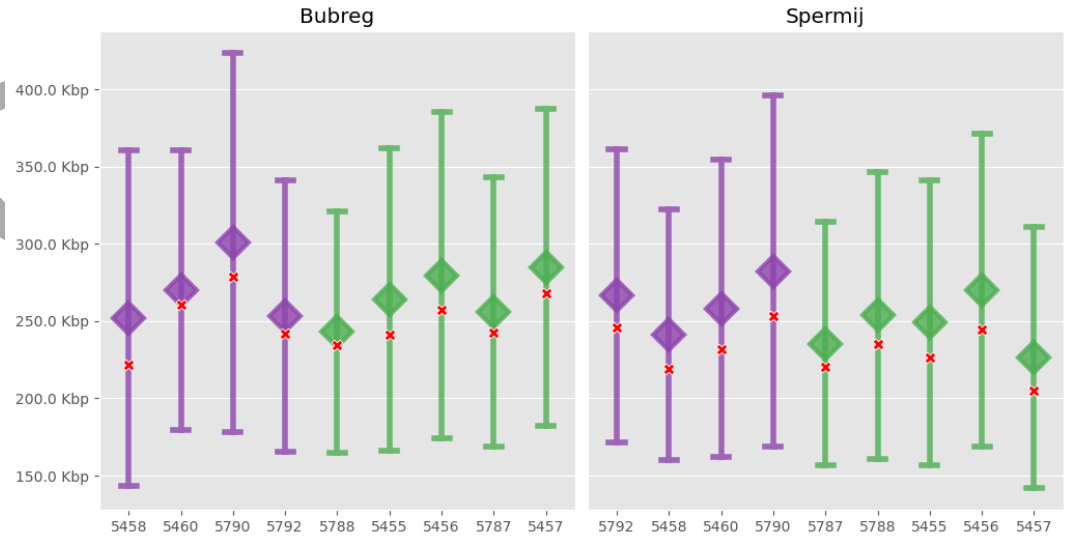


Slika 33. Dijagram kumulativne empirijske distribucije (ordinata) za broj nesparenih oznaka na molekuli.

Ne postoji značajna korelacija između duljine telomera i duljine molekula iz kojih se računa duljina telomera (Pearsonov R=0.24; Slika 34). Također, nismo našli značajnu razliku između duljine molekula koje nose telomere i prosječne duljine svih molekula koje su poravnate na referentni genom (Cohenov D < 0.3; Slika 35). Ovi rezultati ukazuju na to da duljina molekula u našem skupu podataka nema značajan utjecaj na određivanje duljine telomera.

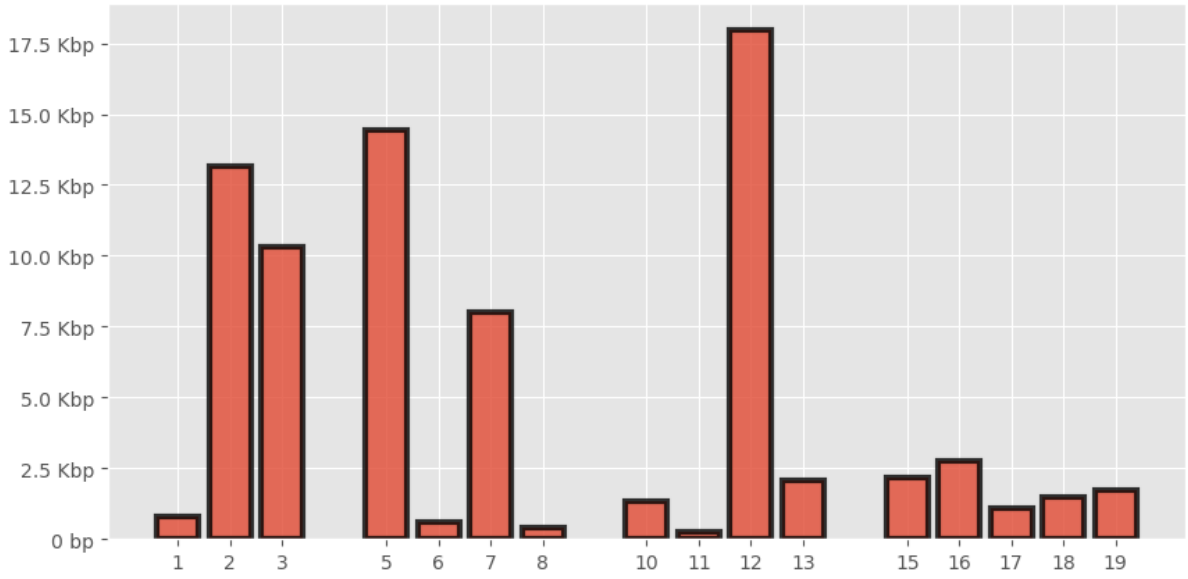


Slika 34. Korelacija duljine molekula koje nose telomere (apscisa) i duljine telomera (ordinata).



Slika 35. Distribucija duljine telomernih molekula po mišu u bubregu (lijevo) i spermijima (desno) za kontrolnu (ljubičasto) i pokusnu skupinu (zeleno). Dijamantom je prikazana prosječna duljina telomerne molekule, a kapice označavaju jednu standardnu devijaciju prosječne duljine. Crvenim znakom x označena je prosječna duljina svih molekula koje su poravnate na referencu.

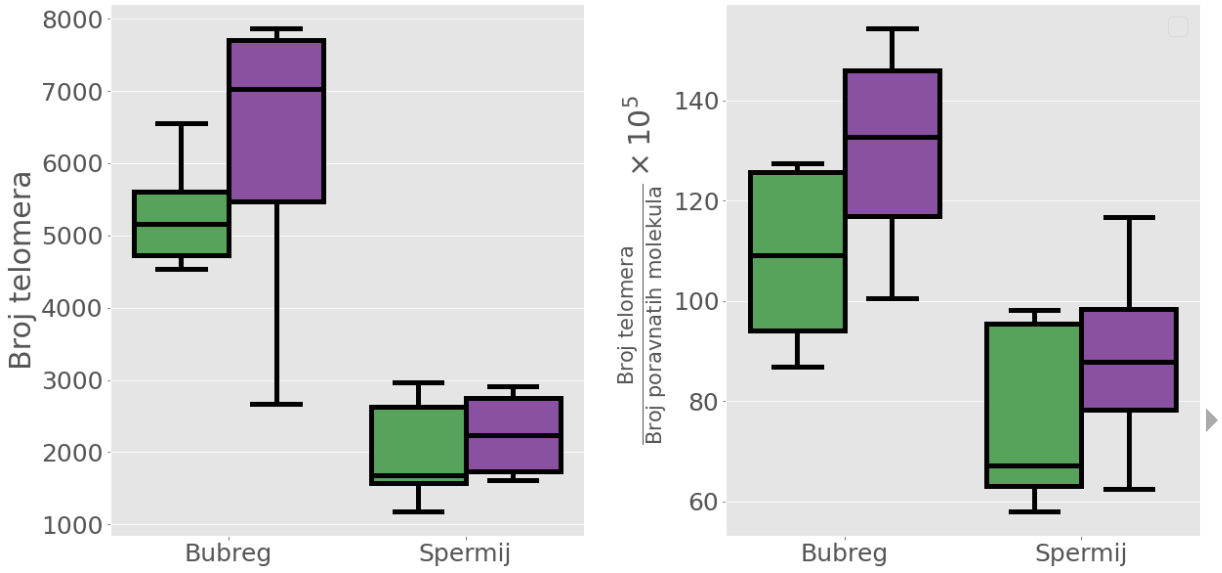
S obzirom na to da se duljina telomera iz podataka optičkog mapiranja računa od krajnjeg DLE-1 motiva do kraja molekule, svaka izmjerena duljina sadrži i dio subtelomerne regije između posljednjeg motiva i početka anotirane telomere na referentnom genomu. Taj dio varira između kromosoma na referentnom genomu i iznosi od nekoliko desetaka do nekoliko tisuća baznih parova (Slika 36).



Slika 36. Udaljenost na referentnom genomu između posljednjeg DLE-1 motiva i anotiranog početka telomere na q kraku. Prikazani su podaci za kromosome čije su telomere analizirane u ovoj disertaciji.

3.1.4.2. Razlike između kontrolne i pokusne skupine

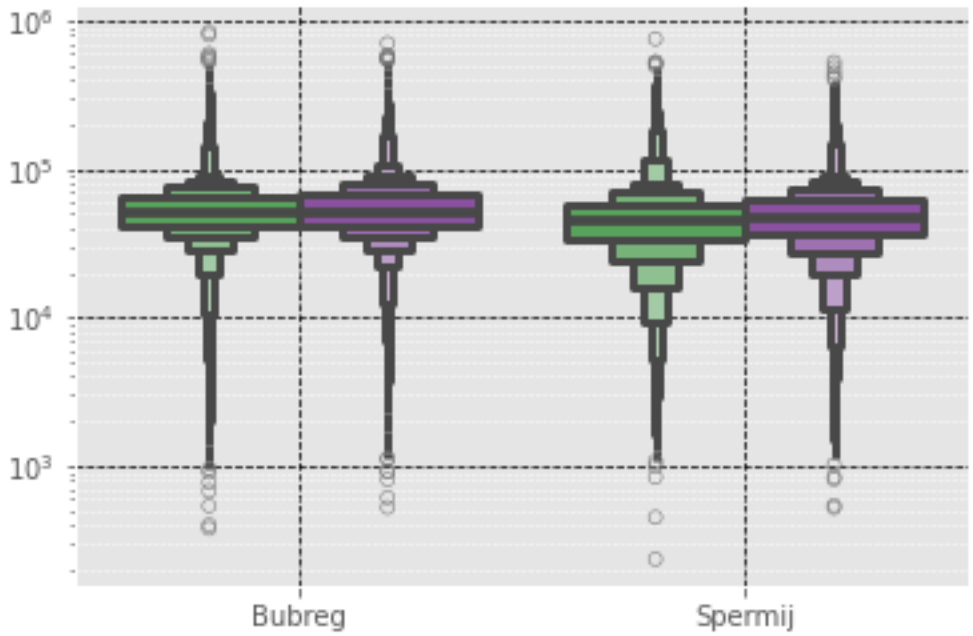
Nalazimo znatno manji broj telomera u spermijima nego u bubrezima. Osim toga, unutar istog tkiva nalazimo manji broj telomera kod miševa iz pokusne skupine. Ova razlika je još izraženija kada podijelimo broj telomera s ukupnim brojem molekula, što sugerira da ona nije uzrokovana razlikama u broju poravnatih molekula odnosno pokrivenosti genoma između uzoraka (Slika 37)



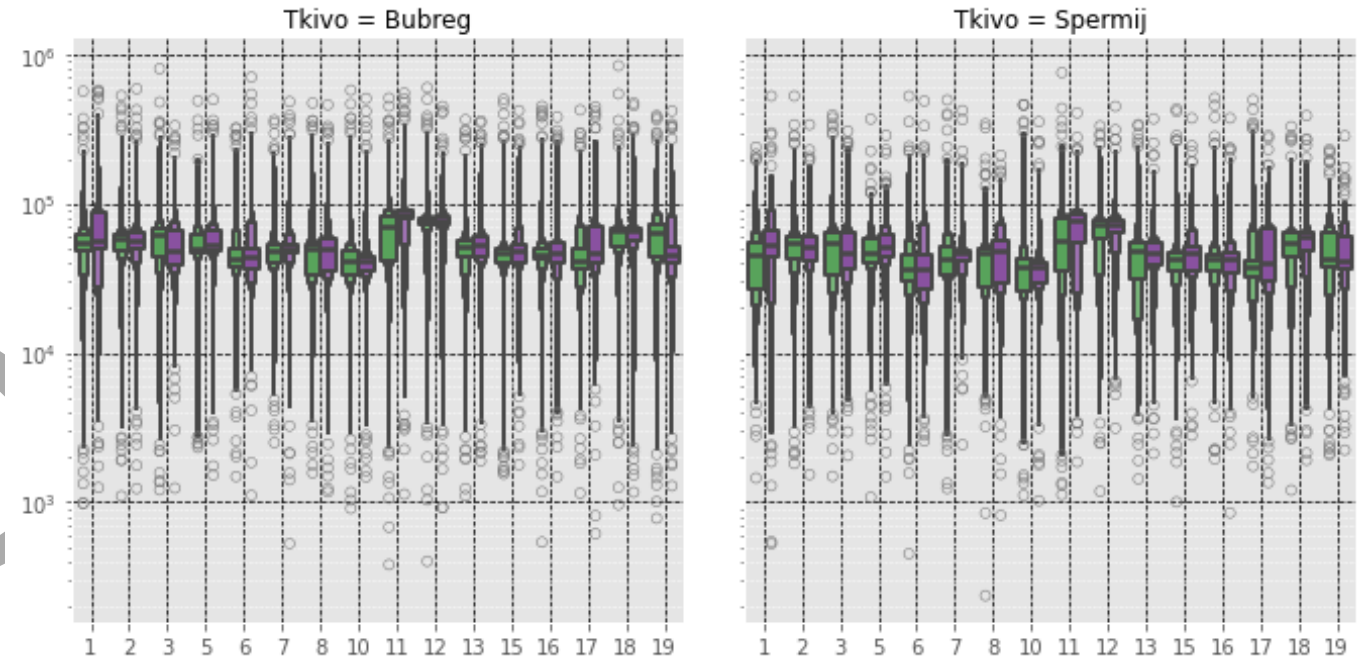
Slika 37. Broj telomera (lijevo) i broj telomera normaliziran prema ukupnom broju poravnatih molekula (desno), u pokusnoj (zeleno) i kontrolnoj skupini (ljubičasto). Ovakav normalizirani broj odgovara broju molekula koje su poravnate na (desnu) telomeru na 100 tisuća molekula ukupno poravnatih u genomu.

Izmjerena duljina pojedinačnih telomera varira između 236 bp i 840 kbp, a u prosjeku iznosi 55 kbp što je u skladu s prosječnom duljinom telomera izmjerenom prethodno kod miševa pomoću drugih metoda (Hemann, 2000). Telomere su u prosjeku kraće kod miševa iz pokusne skupine (Slika 38), za 2.6 kbp (razlika u medijanu 1.3 kbp) u tkivu bubrega (T test, p vrijednost 4×10^{-25}) te za 1.2 kbp (razlika u medijanu 2.5 kbp) u spermijima (T test, p vrijednost 4×10^{-19}). Međutim, s obzirom na veliku varijaciju, ovaj efekt je mali i za tkivo bubrega (Cohenov D = 0.08) i za spermije (Cohenov D = 0.04).

Unatoč širokom rasponu duljina telomera unutar istog kromosoma, mogu se uočiti značajne razlike u prosječnoj duljini telomera između kromosoma (Slika 39). Kromosom 10 ima u prosjeku najkraće telomere (43 kbp) a kromosom 11 najduže (74 kbp).

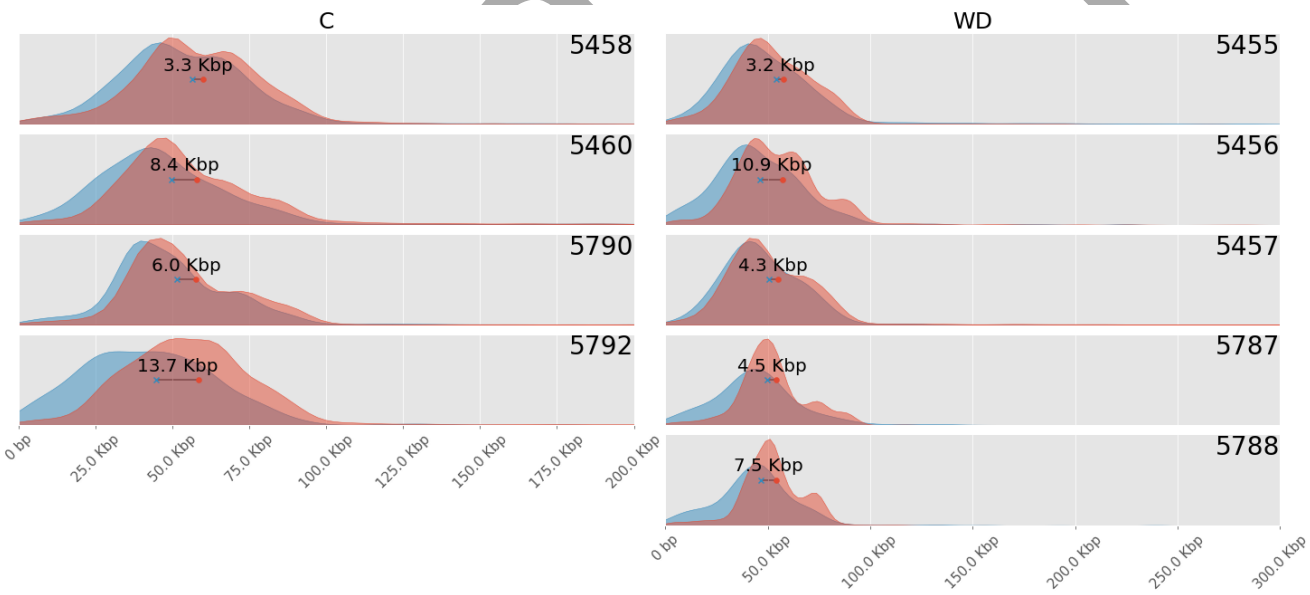


Slika 38. Distribucija duljina telomera svih miševa iz pokusne (zeleno) i kontrolne (ljubičasto) skupine prikazane „boxen“ dijagramom po tkivu. Vodoravna linija u najširem dijelu svake od prikazanih distribucija predstavlja medijan duljine telomera.



Slika 39. Distribucija duljina telomera miševa iz pokusne (zeleno) i kontrolne (ljubičasto) skupine po kromosomima prikazana „boxen“ dijagramom za svaki kromosom posebno. Vodoravna linija u najširem dijelu svake od prikazanih distribucija predstavlja medijan duljine.

Kako bismo utvrdili dolazi li do promjene u duljini telomera u spermijima miševa iz pokusne skupine u odnosu na kontrolnu skupinu, usporedili smo distribuciju duljina telomera iz spermija s distribucijom duljina telomera iz bubrega iste jedinke. Distribucija je pomaknuta prema nižim vrijednostima u spermijima u odnosu na bubrege u svim jedinkama, što ukazuje na generalno skraćivanje telomera u spermijima i/ili njihovo produljivanje u bubregu. Ovakav rezultat je iznenađujući, s obzirom na dokaze o produljivanju telomera u spermijima u odnosu na somatsko tkivo zbog aktivnosti telomeraze tijekom spermatogeneze u različitim vrstama (Fice i Robaire, 2019.; Chieffi Baccari i sur., 2023.). Nije bilo značajne razlike u ovom efektu između pokusne i kontrolne skupine - u kontrolnoj skupini (N=4) razlika prosječne duljine iznosila je 7.8 ± 4.4 kbp (razlika medijana 6.8 ± 4.2 kbp), a u pokusnoj skupini (N=5) 6.1 ± 3.1 kbp (razlika medijana 6.0 ± 2.8 kbp) (Slika 40).



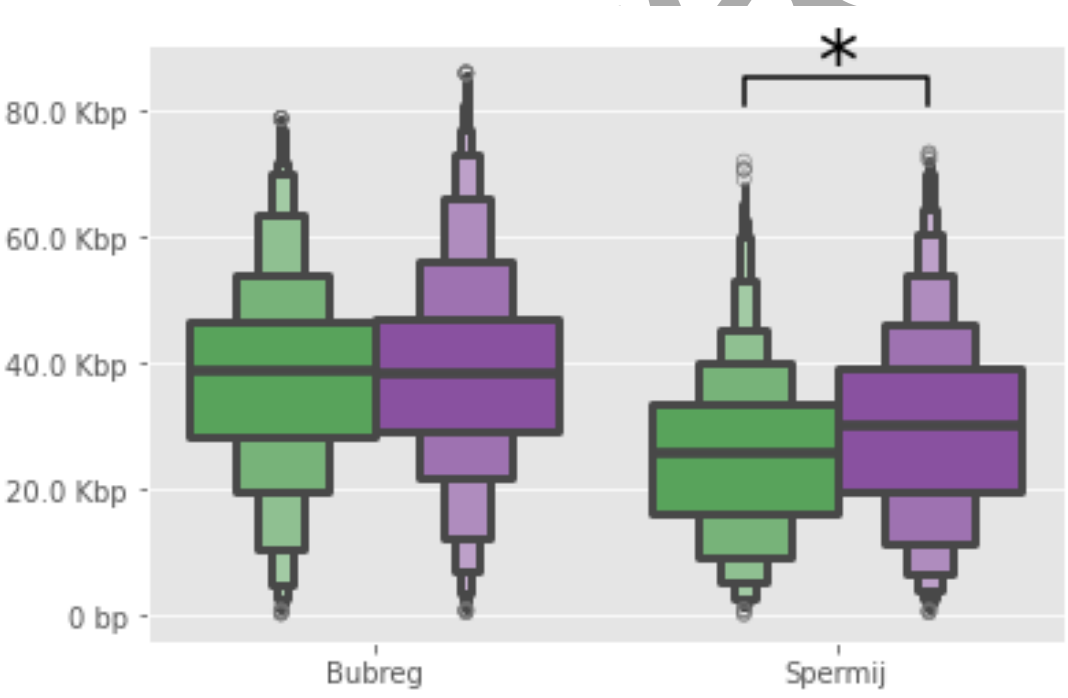
Slika 40. Usporedba distribucija duljina telomera između spermija (plavo) i bubrega (narančasto). Usporedbe su prikazane po pojedinom mišu iz kontrolne (lijevo) i pokusne (desno) skupine. Razlika u prosječnoj duljini telomera je naznačena na svakom grafu.

Poznato je da kritično kratke telomere mogu utjecati na replikativni potencijal stanice, uzrokovati genomsku nestabilnost, apoptozu, te prerano starenje i kraći životni vijek (Hemann, 2000). Kako bismo pobliže analizirali razlike u duljini kritično kratkih telomera, izdvojili smo prvih 25 % najkraćih telomera (telomere u prvom kvartilu duljine, Q1) unutar svakog kromosoma svake jedinke posebno, kako bi se uzele u obzir značajne varijacije u duljini telomera između jedinki te između kromosoma.

Unutar prvog kvartila duljine, telomere su kraće u spermijima pokusne skupine u odnosu na kontrolnu skupinu (T test, p vrijednost 9×10^{-26}; Cohenov D = 0.31) za 4.3 kbp u prosjeku (medijan 4.4 kbp). Ta razlika u tkivu bubrega iznosi 1 kb i nije statistički značajna (Tablica 19, Slika 41).

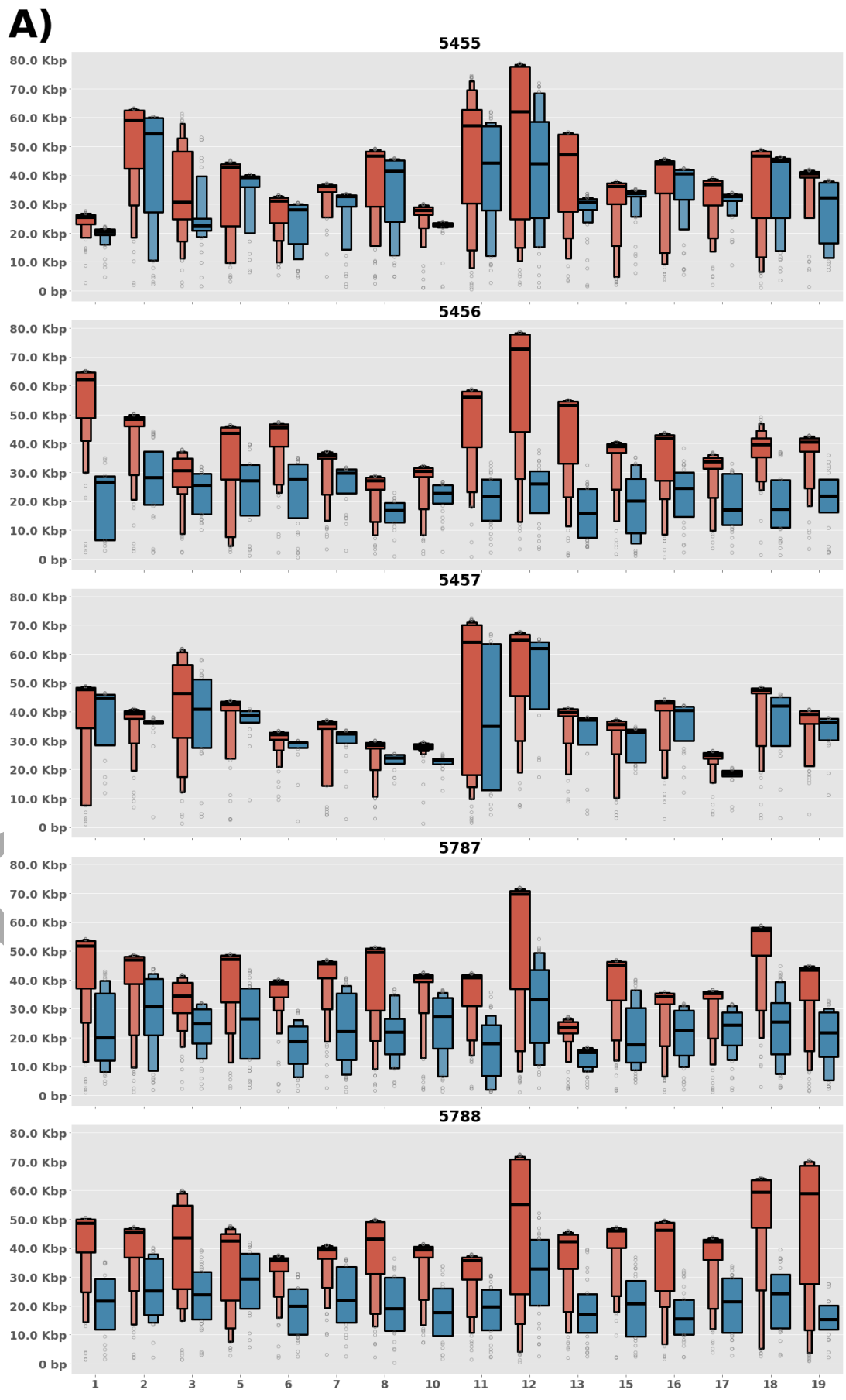
Tablica 19. Deskriptivne statistike koje opisuju duljinu telomera po skupini i uzorku. * označava da je razlika između pokusne i kontrolne skupine statistički značajna (Mann-Whitney test, P vrijednost < 0.05; Cohen D > 0.3). IQR se odnosi na interkvartalni raspon, a STD na standardnu devijaciju.

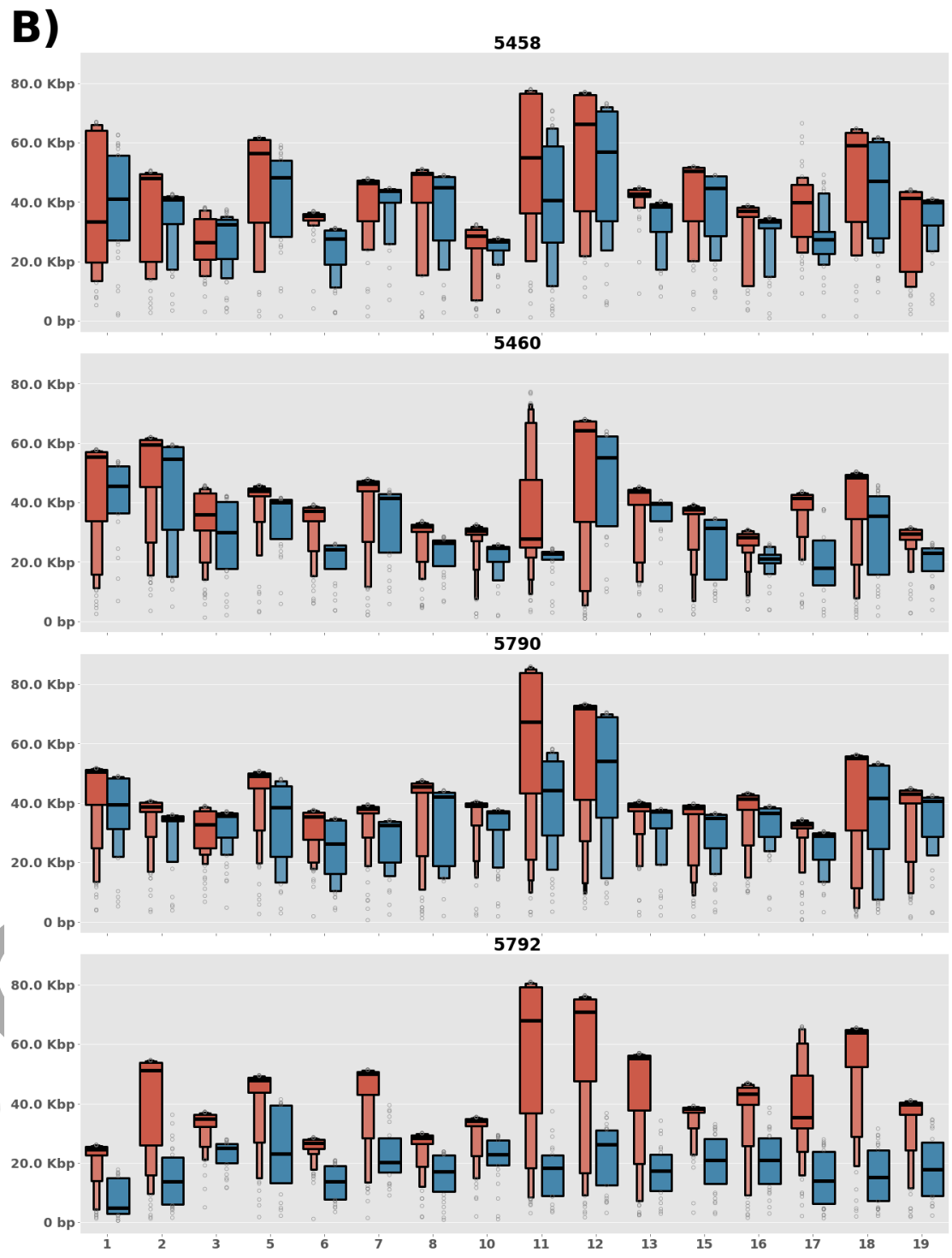
Skupina	Tkivo	prosječna duljina	STD prosječne duljine	medijan	IQR prosječne duljine
C	Bubreg	38.7 kbp	16.0 kbp	38.4 kbp	17.8 kbp
WD	Bubreg	37.7 kbp	15.5 kbp	38.8 kbp	18.1 kbp
C	Spermij*	30.0 kbp	14.9 kbp	30.0 kbp	19.4 kbp
WD	Spermij*	25.7 kbp	13.1 kbp	25.6 kbp	17.4 kbp



Slika 41. Distribucija duljina telomera u prvom kvartilu duljine (Q1) svih miševa iz pokusne (zeleno) i kontrolne (ljubičasto) skupine prikazane „boxen" dijagramom po tkivu. Vodoravna linija u najširem dijelu svake od prikazanih distribucija predstavlja medijan duljine telomera.

Razlika prosječne duljine telomera u prvom kvartilu duljine između tkiva bubrega i spermija u kontrolnoj skupini (N=4) iznosila je 10.1 ± 7.5 kbp (razlika medijana 10.0 ± 7.2 kbp), a u pokusnoj skupini (N=5) za 12.1 ± 6.6 kbp (razlika medijana 12.5 ± 7.4 kbp). Gledano po kromosomima, razlika prosječne duljine telomera iznosila je 12 kbp po kromosomu za pokusnu skupinu te 10 kbp za kontrolnu skupinu (Slika 42).


A)
5455
5456
5457
5787
5788
80.0 Kbp
70.0 Kbp
60.0 Kbp
50.0 Kbp
40.0 Kbp
30.0 Kbp
20.0 Kbp
10.0 Kbp
0 bp
1
2
3
5
6
7
8
10
11
12
13
15
16
17
18
19



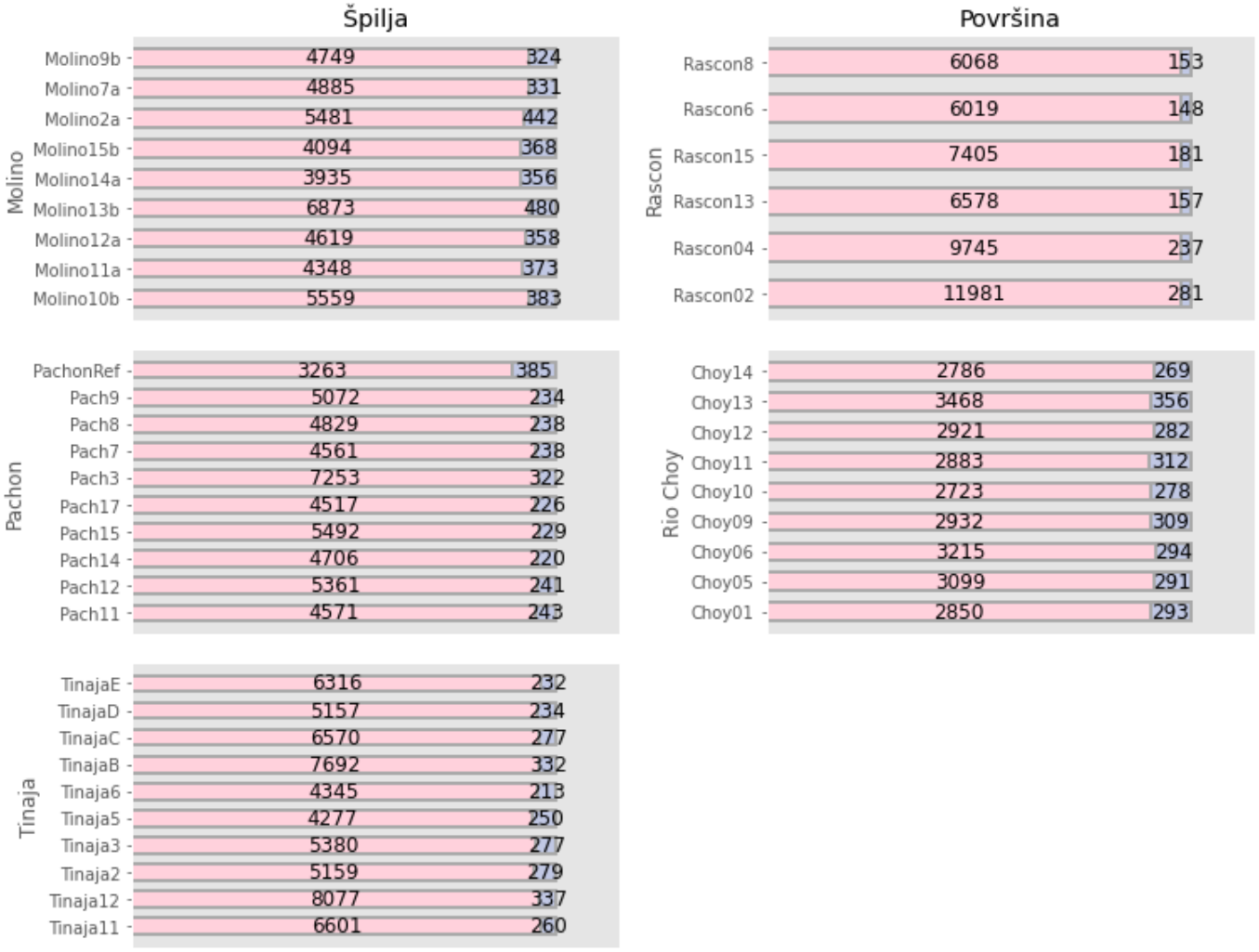
Slika 42. Distribucija duljina telomera u prvom kvartilu duljine po kromosomu. Prikazani su podaci za miševe iz pokusne (A) i kontrolne (B) skupine za spermije (plavo) i bubrege (narančasto). Identifikacijski broj miša naznačen je iznad svakog prikaza.

3.2. Varijacije u broju kopija u prilagodbi na špiljske uvjete života

3.2.1. Preliminarne analize

Nakon mapiranja kratkih NGS očitanja na referentni genom *A. mexicanus*, prosječna pokrivenost sekvencirane baze po genomu je iznosila 9.4 ± 1.8 puta (Tablica S2 u Prilogu). Optimalna veličina segmenata korištenih za detekciju CNVs sa CNVpytor-om je iznosila između 500 i 800 bp (Tablica S2 u Prilogu).

Detektirali smo između 3 001 i 12 262 CNVs po životinji (Slika 43). Nije bilo značajne razlike između linija ili ekotipova u ukupnom broju detektiranih CNVs. Veći udio duplikacija je detektiran kod životinja nove linije nego u životinja stare linije (T test, p vrijednost = 9×10^{-12}, Cohenov D = 2.9): duplikacije čine oko 8 % svih CNVs kod riba Choy i Molino populacije a u prosjeku 4 % u populacijama Rascon, Pachon i Tinaja. Ovaj rezultat može se objasniti činjenicom da referentni genom AstMex3 potječe iz Choy populacije, koja pripada novoj liniji. Stoga je za očekivati je da će manji dio referentnog genoma nedostajati u genomima riba nove linije nego u genomima riba stare linije. Nije bilo značajne razlike u udjelu duplikacija između površinskih i špiljskih ekotipova (T test, p vrijednost = 0.36).

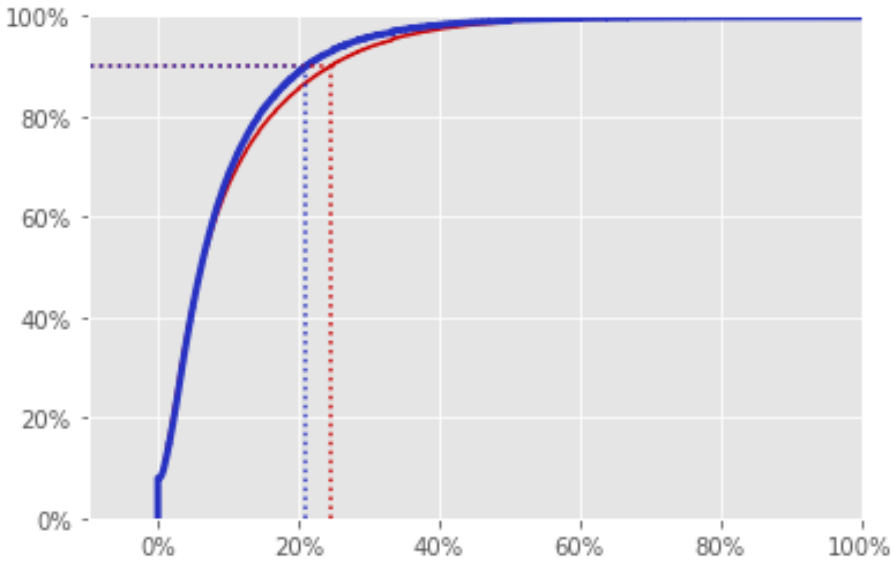


Slika 43. Udio i broj detektiranih duplikacija (ljubičasto) i delecija (ružičasto) po jednikama. Jedinke su grupirane po populacijama.

Kako bismo procijenili koji dio referentnog genoma AstMex3 je podložan varijacijama broja kopija, preklopili smo CNVs iz svih jedinki u CNVRs. Pronašli smo 15 088 CNVRs na kromosomima referentnog genoma, što ukupno iznosi 260.2 Mbp. U usporedbi s ukupnom duljinom sastavljenih kromosoma referentnog genoma (1 321 Mbp), procjenjujemo da 19.7 % referentnog genoma varira brojem kopija u prirodnim *A. mexicanus* populacijama. Čak 10 035 CNVRs se preklapa s genima (uključujući protein-kodirajuće, lncRNA, rRNA, tRNA i ostale anotirane klase gena) i zauzima ukupno 213 Mbp, dok ostala 5 053 ukupno zauzimaju 48 Mbp i mogu se smatrati nekodirajućima. Ovakav omjer kodirajućih i nekodirajućih CNVRs može se objasniti visokim udjelom kodirajućih sekvenci u AstMex3 referentnom genomu: čak 63 % sekvence na sastavljenim kromosomima je anotirano kao geni.

Pronašli smo 2 819 CNV gena u cijelom skupu podataka. Prosječna duljina CNV gena je 9 263 bp (medijan 5 335 bp), što je znatno kraće od prosječne duljine svih anotiranih protein-kodirajućih gena (34 441 bp, medijan 13 069 bp). Više od trećine CNV gena (978) je nepoznate funkcije (eng. *uncharacterized*), što je četverostruko veći udio u usporedbi s udjelom takvih gena u cijelom referentnom genomu (8.5 %). Ovo obogaćenje bi moglo biti posljedica kratke duljine ovakvih gena (u prosjeku 7.3 kbp u AstMex3 genomu) na način da je vjerojatnije da će kraći geni češće biti obuhvaćeni CNVs. Kako bismo testirali ovu pretpostavku, proveli smo permutacijske analize tako da su pozicije detektiranih CNVs nasumično raspodijeljene po genomu, pri čemu je zadržana originalna distribucija CNV duljina. U svakoj permutaciji izračunata je prosječna duljina CNV gena i udio CNV gena nepoznate funkcije. Na temelju 100 takvih permutacija, očekivali bismo pronaći oko 10 % nekarakteriziranih gena unutar skupa CNV gena, što je vrlo blizu njihovom stvarnom udjelu u cijelom AstMex3. Osim toga, očekivali bismo da će njihova prosječna duljina biti oko 15 kbp, što je niže od prosjeka genoma (34.4 kbp), ali znatno više od identificiranih CNV gena (9.3 kbp). Možemo zaključiti da postoji 3.5 x veća vjerojatnost da će CNV geni imati nepoznatu funkciju i biti 1.7 x kraći od očekivanog.

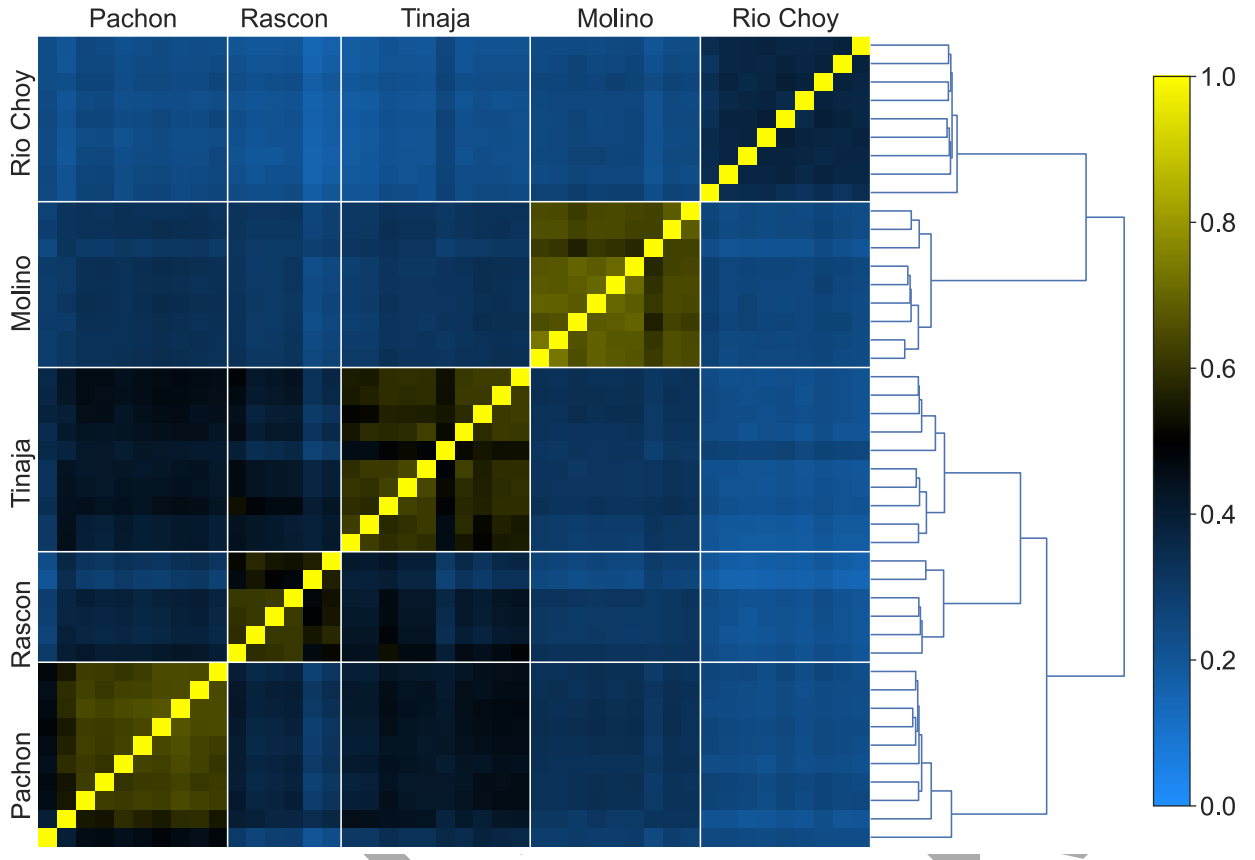
Očitanja s kvalitetom mapiranja u vrijednosti nula (MAPQ=0 u BAM datoteci) odnose se na očitanja koja se s jednakom točnošću mogu mapirati na dvije ili više pozicije u referentnom genomu. Ovakva očitanja niske kvalitete stoga mogu umjetno povećati broja kopija, na način da se nagomilaju na određenoj genomskoj poziciji iako ne potječu od nje. Kako bismo procijenili utjecaj ovih artefakata na rezultate, izračunali smo udio očitanja koji imaju MAPQ vrijednost nula, za sve detektirane CNVs i one CNVs koji preklapaju protein kodirajuće gene. U 90 % svih detektiranih CNVs, nalazimo manje od 20 % očitanja niske kvalitete, odnosno manje od 25% u CNVs koji pogađaju protein-kodirajuće gene (Slika 44).



Slika 44. Dijagram kumulativne empirijske distribucije (ordinata) za udio očitanja koja imaju MAPQ=0 (apscisa) za sve detektirane CNVs (crveno) te sve CNVs koji preklapaju protein-kodirajuće gene (plavo). Isprekidane linije označavaju odgovarajuće maksimalne udjele za 90 % CNVs.

3.2.2. Analiza genetičke raznolikosti

Genetičku sličnost između dvije jedinke možemo procijeniti na temelju broja zajedničkih CNVs, odnosno CNVs koji se u dvije jedinke nalaze na istom položaju u genomu. Takve usporedbe mogu nam dati uvid u stupanj genetičke raznolikost unutar neke populacije. Kako bismo procijenili relativnu genetičku raznolikost unutar i između populacija, analizirali smo broj zajedničkih CNVs između svih kombinacija dvaju jedinki u našem skupu podataka. Nalazimo da bilo koje dvije jedinke u našim podacima dijele u prosjeku 1 946 CNVs, odnosno 34 % svih CNVs. Jedinke su najsličnije jedna drugoj u špiljskoj populaciji Molino koja pripada novoj liniji, gdje u prosjeku dijele 3 448 CNV (65 %). Ovaj rezultat sugerira da je genetička raznolikost najniža u Molino populaciji, što je u skladu s prethodnim opažanjima temeljenim na analizama SNP podataka (Bradic i sur., 2013; Herman i sur., 2018). Površinska populacija nove linije Río Choy je genetički najraznolikija, u kojoj pojedinačni parovi u prosjeku dijele 1 198 CNVs (36 %). Na temelju udjela zajedničkih CNVs populacije se grupiraju prema svojim linijama (Slika 45), no unutar stare linije površinska Rascon populacija se grupira sa špiljskim ribama iz Tinaja populacije. Ovo posljednje se ne slaže s prethodno ustanovljenom filogenijom temeljenom na SNP podacima prema kojoj špiljske populacije Tinaja i Pachón čine monofiletsku sestrinsku skupinu površinske populacije Rascon (Herman i sur., 2018).



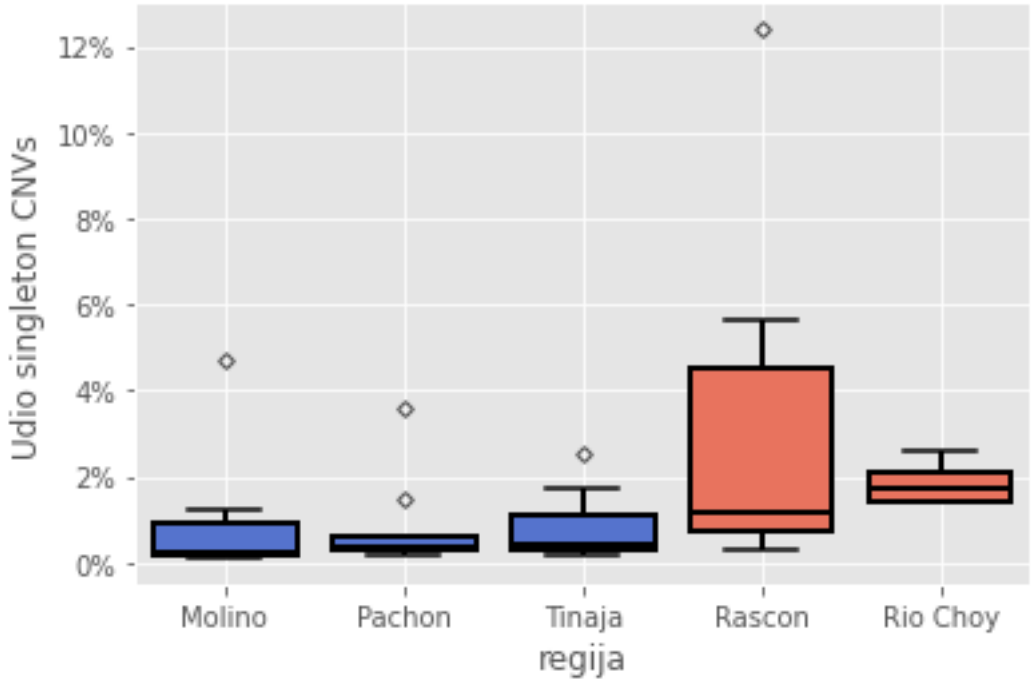
Slika 45. Matrica udaljenosti temeljena na prosječnom broju zajedničkih CNV između dva genoma. Uzorci su prikazani istim redoslijedom s desna na lijevo kao i od vrha prema dolje. Matrica udaljenosti podvrgnuta je Wardovoj metodi hijerarhijskog klasteriranja.

Kako bismo identificirali CNV lokuse koji su specifični za populacije, definirali smo pojmove privatni CNV i privatni CNVR, kao lokuse u kojima nalazimo varijacije u broju kopija u najmanje jednoj jedinki iste populacije, a niti u jednoj u ostalim populacijama. Nalazimo 4 257 CNVRs koji su specifični za špiljske i 4 728 koji su specifični za površinske ribe (Tablica 20). Udio privatnih CNVRs koji se nalaze u pojedinačnim životinjama veći je u površinskih riba (65 %) nego u špiljskih riba (41 %). To je također vidljivo u usporedbama pojedinačnih površinskih populacija (72 % u Río Choy i 68 % u Rascon populaciji) s pojedinačnim špiljskim populacijama (45 % – 62 %). Jedan privatni CNVR sadrži u prosjeku pet CNVs u špiljskim ribama i samo tri u površinskim životinjama. Ove analize sugeriraju da špiljske ribe češće dijele isti CNV, u skladu s analizom sličnosti koja se temelji na broju zajedničkih CNVs (Slika 45).

Tablica 20. Broj CNVs i CNV regija koje su privatne za populaciju, ekotip i liniju.

Skupina	Privatni CNVs	Privatni CNVRs	CNVRs s jednim CNVs	Prosječan broj CNVs po CNVRs
Molino	3,577	1,215	551 (45%)	5
Pachón	1,770	807	502 (62%)	4
Tinaja	3,142	1,266	623 (49%)	4
Río Choy	1,264	830	596 (72%)	3
Rascon	5,649	3,527	2,383 (68%)	3
špilja	15,016	4,257	1,750 (41%)	5
površina	8,243	4,728	3,056 (65%)	3
nova linija	6,346	2,326	1,197 (52%)	5
stara linija	35,583	8,752	3,633 (42%)	6

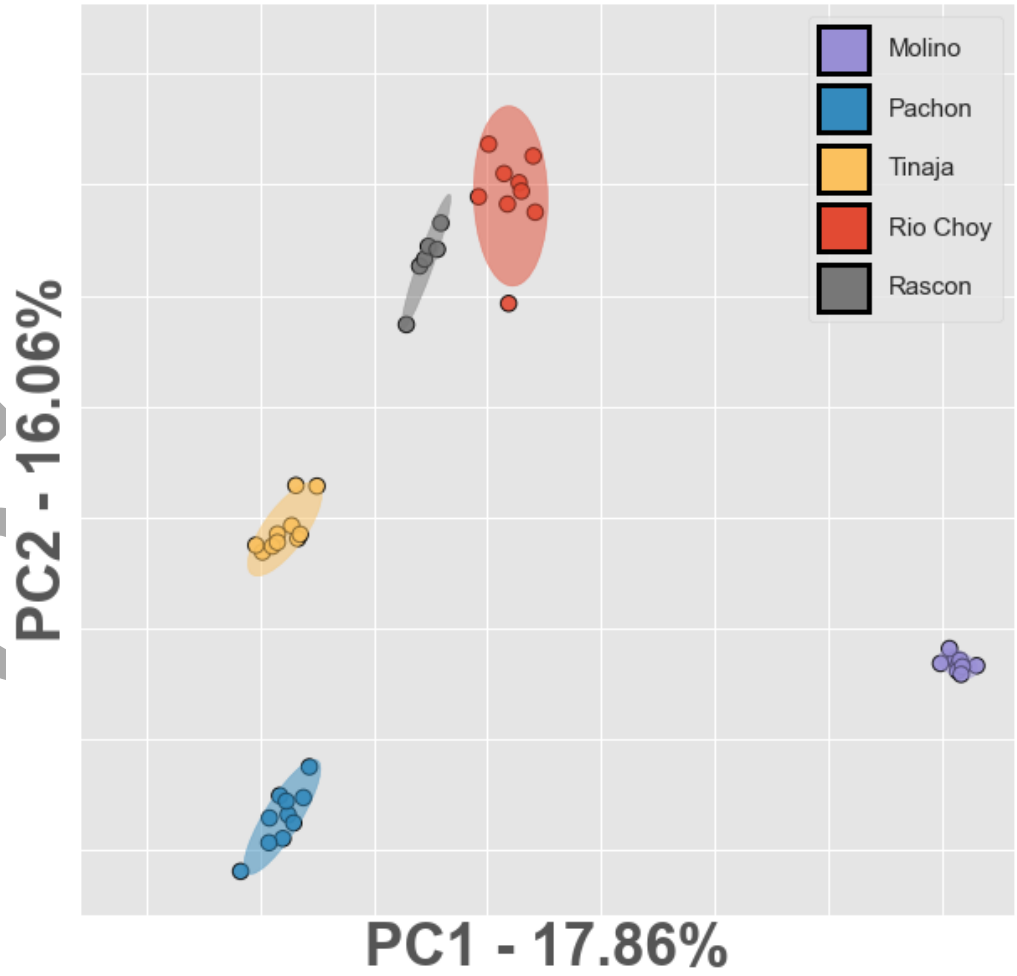
Zatim smo analizirali udio singleton CNVs, definiranih kao CNVs koji su otkriveni u samo jednoj životinji. Nalazimo između 5 i 1 537 takvih singletona po jedinci, što odgovara udjelu od 0.12 % do 2.53 % svih CNVs. Broj singletona, u odnosu na broj svih identificiranih CNVs unutar jedinke, veći je u površinskim populacijama nego u špiljskim (Slika 46).



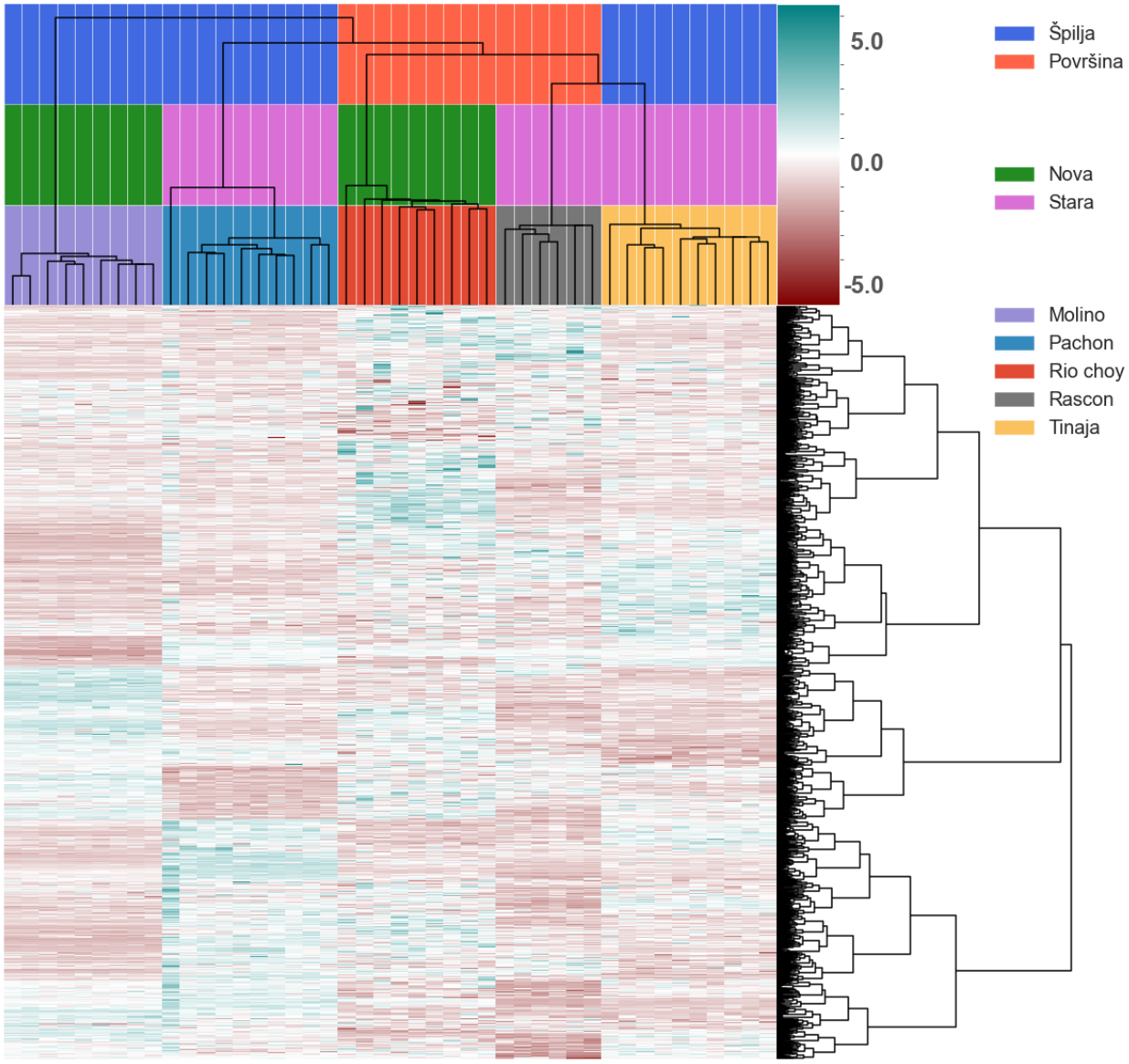
Slika 46. Distribucija udjela singleton CNVs u špiljskim (plavo) i površinskim (narančasto) populacijama.

3.2.3. Diferencijacija populacija

S pomoću opcije *-genotype* u CNVpytoru odredili smo broj kopija svakog CNV gena u svakoj jedinci. Na temelju broja kopija gena, jedinke se grupiraju prema zemljopisnom položaju (Slika 47, Slika 48). Po broju kopija gena, dvije populacije površinskih riba - nova linija Río Choy i stara linija Rascon, bliže su nego što bi se to očekivalo, s obzirom na to da se razdvajanje linija dogodilo prije najmanje 200 000 godina (Herman i sur., 2018). Štoviše, hijerarhijsko grupiranje sugerira da su površinska Rascon i špiljska Tinaja populacija najsličnije na temelju obrasca broja kopija gena i da te dvije populacije tvore sestrinsku grupu s površinskom populacijom Río Choy (Slika 48).



Slika 47. Prikaz prve dvije PCA komponente na temelju broja kopija 2,819 CNV gena. Oko klastera uzoraka prikazane su elipse pouzdanosti (95 %).

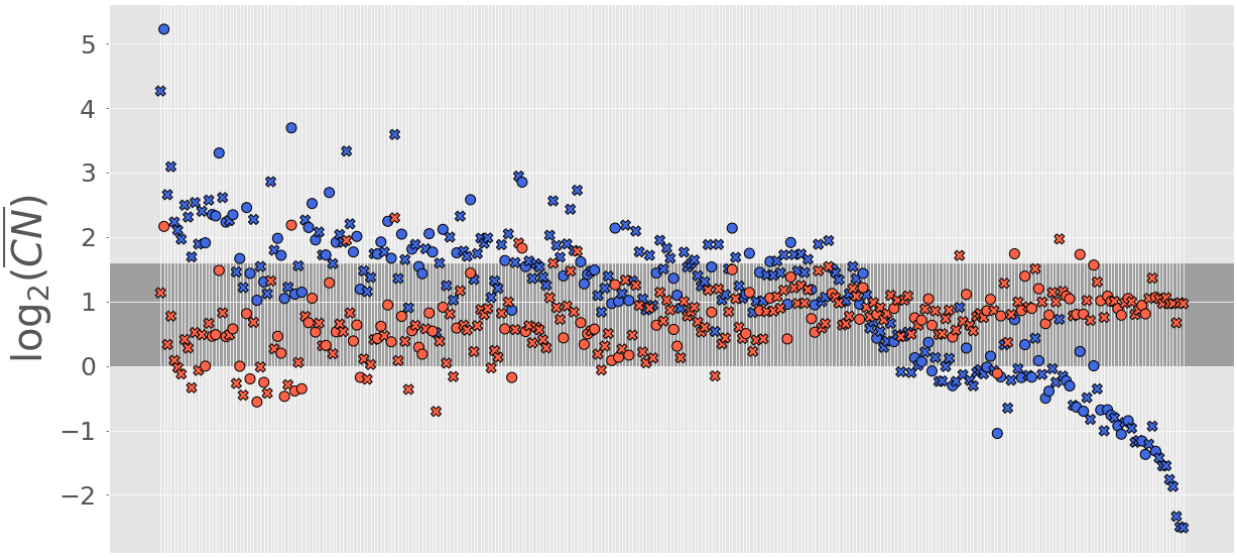


Slika 48. Prikaz normaliziranog broja kopija 2 819 CNV gena s hijerarhijskim grupiranjem na temelju gena (redovi) i uzoraka (stupci) korištenjem Wardove metode s Euklidovom metrikom. Vrijednosti su normalizirane prema retku (genu) tako da srednja vrijednost svakog retka iznosi 0, a njegova standardna devijacija iznosi 1. Predstavljene su relativne razlike u broju kopija određenog gena između uzoraka, u rasponu od najniže (crveno) do najviše (zeleno) vrijednosti. Uzorci su obojeni prema populaciji, ekotipu i liniji kako je naznačeno u legendi s desne strane.

3.2.4. Divergencija u broju kopija između ekotipova i funkcionalni sadržaj

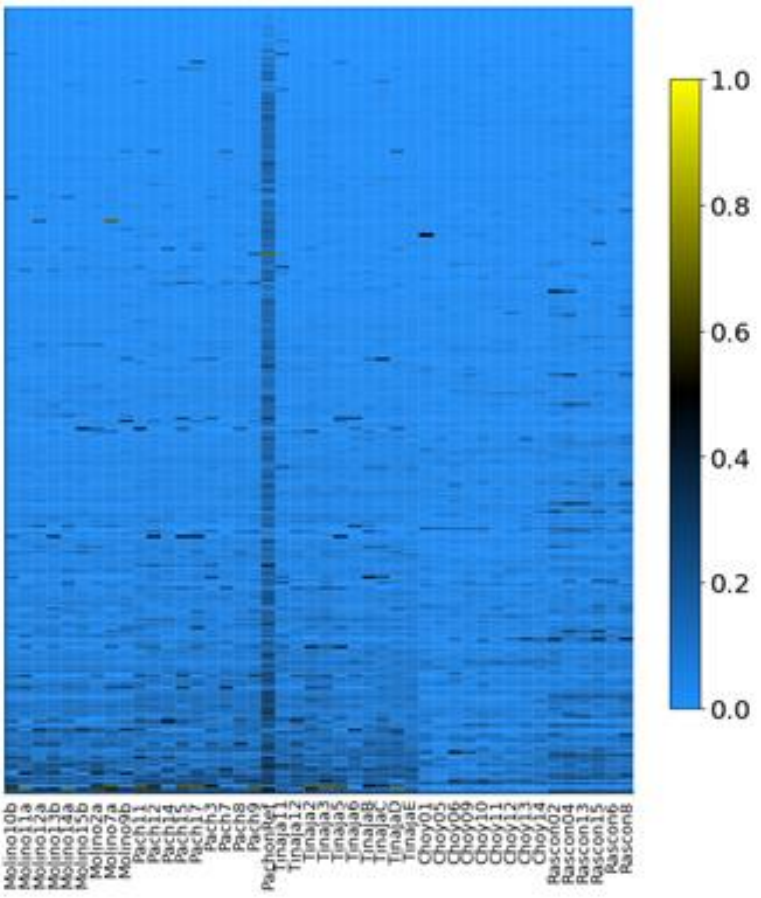
Razlike između populacija u broju kopija gena i/ili nekodirajućih genomskih regija mogu se odraziti na fenotipske razlike između tih populacija. Primjerice, ako su nukleotidne sekvence koje zahvaćaju takvi CNVs povezane s određenim biološkim procesima, to može značiti da se i ti procesi razlikuju među populacijama. Kako bismo identificirali takve procese koji bi mogli biti posebno usmjereni na divergenciju u broju kopija između površinskih i špiljskih riba, istražili smo funkcionalni sadržaj gena koji 1) su djelomično ili cijelom duljinom obuhvaćeni CNVs, 2) se nalaze u blizini CNVs, koji se brojem kopija značajno razlikuju između špiljskih i površinskih riba.

Od ukupno 15 088 CNVRs u cijelom skupu podataka, njih 292 se značajno razlikuju u broju kopija između špiljskih i površinskih životinja (Wilcoxonov test, prilagođeni pval < 0.01) (Tablica S6 u Pokrovac i sur., 2024). Oni zajedno obuhvaćaju 10.47 Mbp, odnosno 0.8 % referentnog genoma. Gotovo trećina (87) ne sadrži protein-kodirajuće sekvence, dok se većina (205) preklapa s i/ili sadrži jedan ili više gena pa ih možemo smatrati protein-kodirajućim CNVRs. Raspon prosječnih brojeva kopija ovih 292 CNVRs je širi kod špiljskih nego kod površinskih riba: dok u genomima površinskih riba ove CNVRs nalazimo uglavnom u jednoj do tri kopije po diploidu u prosjeku, iste genomske regije u genomima špiljskih riba značajno variraju brojem kopija i dosežu i do 34 kopije u prosjeku (Slika 49). Dvije trećine ovih regija (195/292) imaju veći prosječan broj kopija u špiljskom ekotipu nego u površinskom, a samo 97 CNVRs ima veći prosječan broj kopija u površinskim ribama.



Slika 49. CNVRs koji se značajno razlikuju u broju kopija između dva ekotipa. Svih 292 CNVRs prikazani su na apscisi, gdje svaka pozicija predstavlja jedan CNVR. Svaki CNVR genotipiziran je CNVpytorom, a prosječni broj kopija prikazan je plavim točkama za špiljske i narančastim za površinske ribe. Log2-transformirane vrijednosti broja kopija prikazane su na ordinati. Simbolom x označeni su CNVRs koji ne sadrže protein-kodirajuće gene a kružićima CNVRs koji sadrže protein-kodirajuće gene. Tamno sivom je osjenčan raspon od jedne do tri kopije.

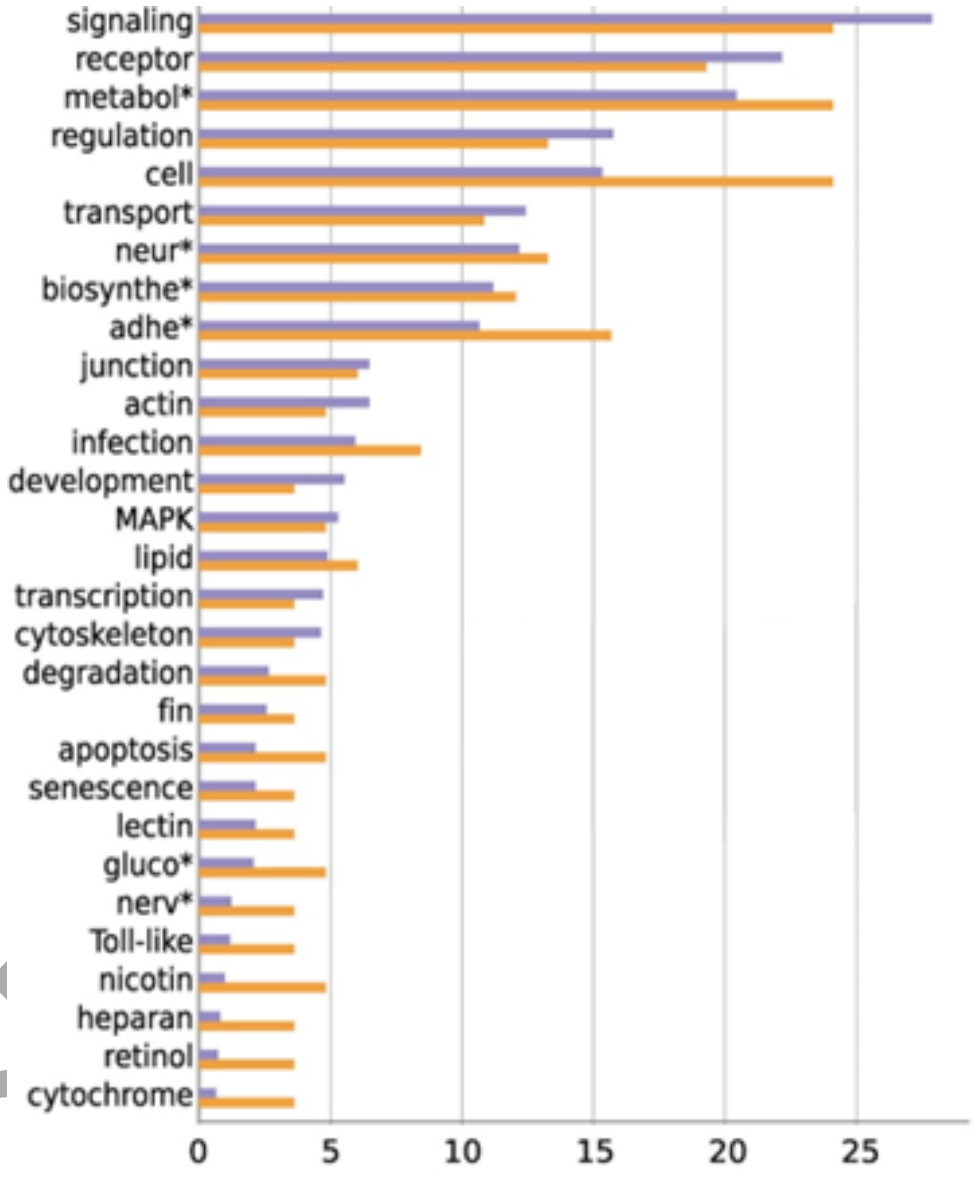
Analizirali smo udio očitanja s niskom kvalitetom mapiranja (MAPQ = 0) na svakom od 292 divergentna CNVRs (Slika 50). U većini slučajeva nije bilo mapiranih očitanja niske kvalitete, a njihov udio je manji od 5 % kod približno 90 % analiziranih CNVRs, što sugerira da očitanja niske kvalitete nemaju značajan utjecaj na ove rezultate.



Slika 50. Udio očitanja niske kvalitete za svaki od 292 divergentna CNVRs (po redovima) po pojedinom uzorku (stupcu). Udjeli su prikazani nijansama boja prema shemi na slici desno.

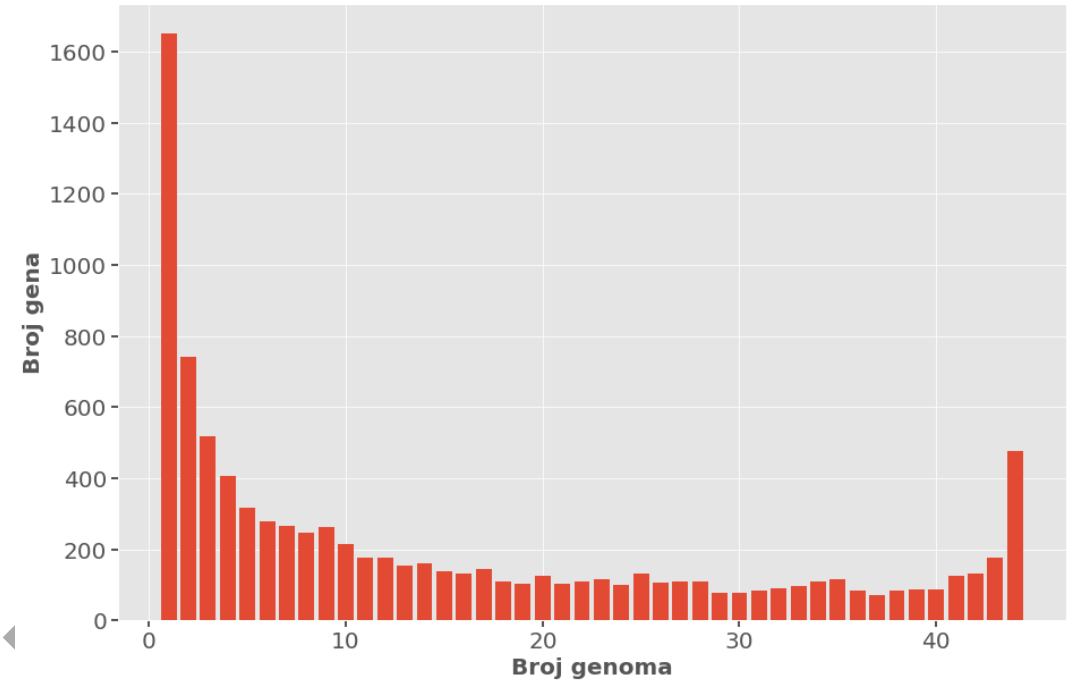
3.2.4.1. Divergencija na razini gena

Analizirali smo funkcionalne anotacije gena sadržanih u protein-kodirajućim CNVRs sa značajnim razlikama u broju kopija između ekotipova. Od 460 gena koji se preklapaju s 205 CNVR-ova, samo 83 su imala dodijeljene anotacije za biološke procese ili putove u DAVID bazi podataka. Otprilike polovica ovih gena ima ulogu u brojnim signalnim i metaboličkim putovima, uključujući MAPK, Toll-like i lektin tipa C (Slika 50; Tablica S8 u Pokrovac i sur., 2024). Ostale visoko zastupljene kategorije uključivale su transport tvari, regulaciju različitih procesa i procese povezane s funkcioniranjem živčanog sustava, kao što su interakcija neuroaktivnog liganda i receptora, te transport neurotransmitera. Međutim, sve nabrojane kategorije također su prisutne u slično visokim udjelima i u genomskim regijama koje se brojem kopija nisu značajno razlikovale između ekotipova (Slika 51). Procesi koji su više zastupljeni u genomskim regijama s različitim brojem kopija između špiljskih i površinskih životinja uključuju kategorije povezane sa staničnom adhezijom, apoptozom, metabolizmom glukoze, kao i metabolizmom citokroma, retinola i nikotinata/nikotinamida, te razgradnjom aminokiselina razgranatog lanca.



Slika 51. Pojmovi koji se najčešće pojavljuju a povezani su s genima koji se preklapaju s CNVR-ovima. Učestalost je prikazana odvojeno, za regije koje se značajno razlikuju (narančasto) i ne razlikuju (plavo) brojem kopija između špiljskih i površinskih riba. Pojmovi su izvedeni iz bioloških procesa i puteva unutar DAVID programa. Potpuni popis pojmova povezanih s riječima označenim zvjezdicama nalazi se u Tablici S8 u Pokrovac i sur., 2024. Učestalosti su prikazane na apscisi kao postotci broja gena povezanih s pojmom u odnosu na ukupni broj gena koji ima dodijeljenu anotaciju u DAVID bazi podataka.

Bilo djelomično ili cijelom duljinom, ukupno se 9 187 protein-kodirajućih gena u AstMex3 referentnom genomu preklapa s CNVs identificiranim u našim podacima. Kod 1 653 (18 %) gena CNV je detektiran u samo jednom uzorku, a kod 476 (5 %) gena CNVs su pronađeni u sva 44 analizirana genoma (Slika 52).



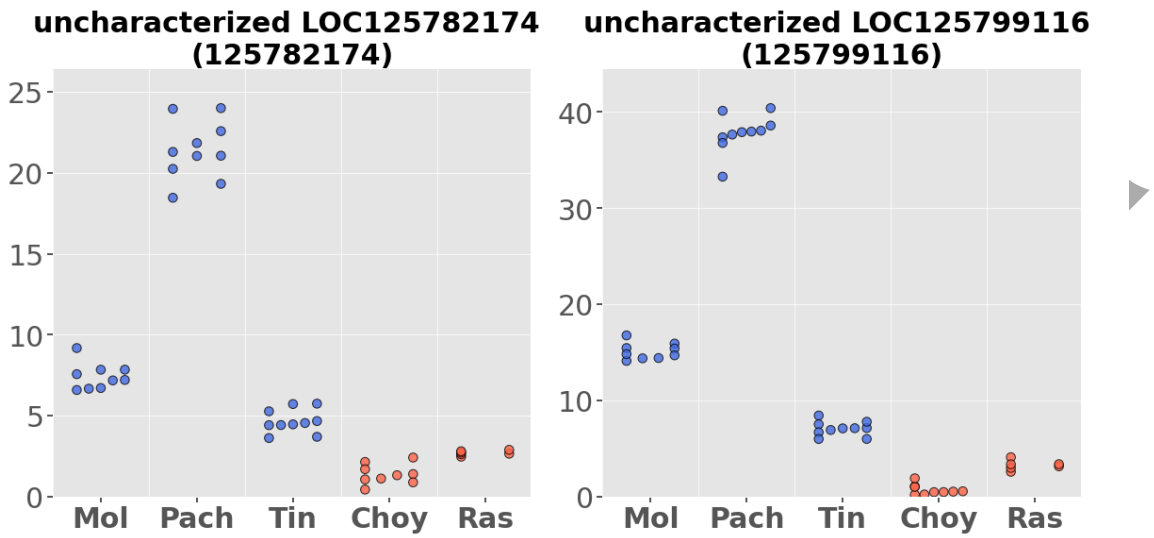
Slika 52. Broj protein-kodirajućih gena (ordinata) naspram broja jedinki (genoma) u kojem se ti geni preklapaju s jednim ili više CNVs.

Oko 15 % (1 407) tih gena se preklapa sa CNVs isključivo u špiljskim populacijama (Molino, Pachón, Tinaja) a 19 % (1 726) isključivo u površinskim populacijama (Rascon, Río Choy). Najveći udio ovih gena bio je zahvaćen samo u jednoj životinji: 42 % (589 gena) u špiljskim i 62 % (1 064 gena) u površinskim ribama. Znatno manji udio ovih gena se preklapao sa CNVs kod više jedinki i kod najmanje jedne životinje po populaciji: 4 % (58 gena) u špiljskim i 9 % (163 gena) kod površinskih riba.

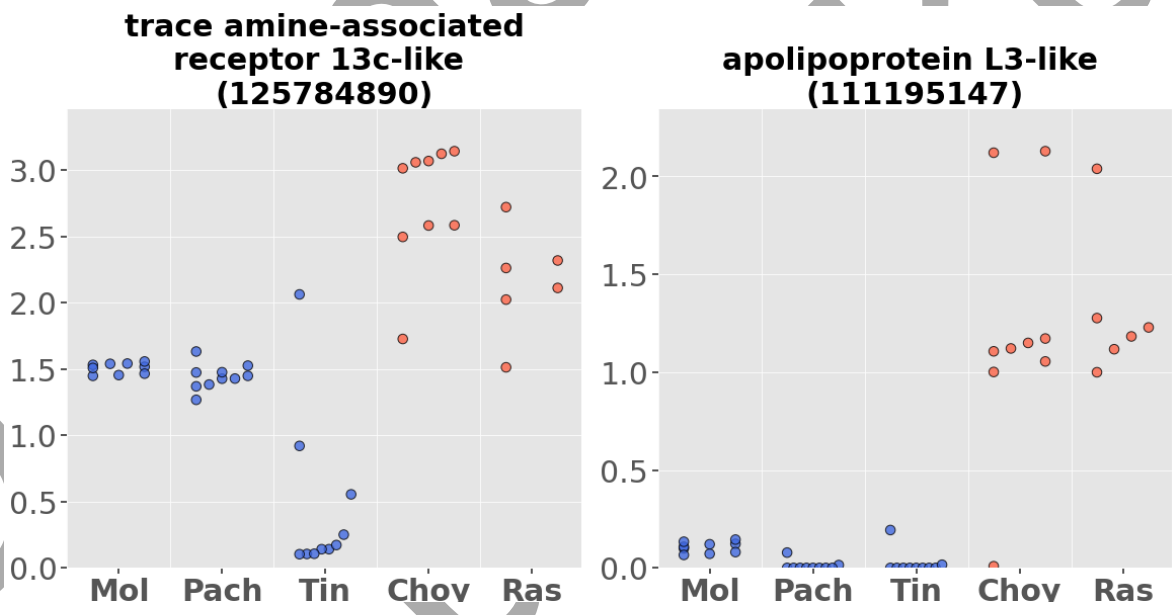
Kako bismo ustanovili koji biološki procesi su najčešće zastupljeni u ovim CNVs specifičnima za ekotip, analizirali smo učestalost preklapanja gena sa CNVs, odnosno broj jedinki u kojima je pojedini gen zahvaćen CNVs. Geni koji su najčešće zahvaćeni CNVs u oba ekotipa su oni koji sudjeluju u imunološkom odgovoru. Primjerice, geni *LOC125785663* (*NLR family CARD domain-containing protein 3-like*), *LOC111196508* (*scavenger receptor cysteine-rich type 1 protein M130*), *LOC103031898* (deleted in malignant brain tumors 1 protein) i *pikfyve* (*phosphoinositide kinase, FYVE finger containing*) povezani su s urođenim imunitetom, upalnim odgovorom i antivirusnom obranom. CNVs u ovim genima otkriveni su u 13 do 28 špiljskih jedinki od ukupno 29 (Tablica S3 u Prilogu). Slično je i sa genima koji su specifično pogođeni CNVs u površinskim ribama: *LOC103031476* (*NACHT, LRR and PYD domains-containing protein 3*), *rnf41* (*ring finger protein 411*), *LOC125801137* (*E3 SUMO-protein ligase ZBED1-like*) i *LOC125782699* (*scavenger receptor cysteine-rich type 1 protein M130-like*) zahvaćeni su CNVs u 9 do 13 od ukupno 15 površinskih jedinki (Tablica S4 u Prilogu). Među ostalim genima koji su s najvećom učestalošću pogođeni specifično u špiljskim ribama (u više od 30 % jedinki pojedine populacije), su geni koji sudjeluju u procesima kao što je vizualna percepcija (*rgrb - retinal G protein-coupled receptor b*; *LOC111194948 - TOG array regulator of axonemal microtubules protein 1*; *map2 - microtubule-associated protein 2*; *opn8a - opsin 8 group member a*; *LOC107197208 - complement C1q-like protein 3*), geni koji kodiraju za podjedinice hemoglobina (*LOC111191630 - hemoglobin subunit beta-2-like*; *LOC111191631 - hemoglobin embryonic subunit alpha*; *LOC111191628 - hemoglobin embryonic subunit alpha*) i geni povezani s neurološkim funkcijama (*plppr3a - phospholipid phosphatase related 3a*; *LOC103030484 - ras-related protein Rab-26*; *rab3c - RAS oncogene family member*).

Od 2 819 detektiranih CNV gena, njih 102 se značajno razlikuje po broju kopija između špiljskih i površinskih ekotipova (Tablica S5 u Prilogu). Više od polovice (65) su predviđeni geni nepoznate funkcije među kojima čak 89 % (58 gena) ima prosječan broj kopija veći u špiljskom nego u površinskom obliku ribe. Primjerice, nalazimo između 4 i 24 kopije gena *LOC125782174* u špiljskih jedinki, dok isti gen postoji u 1-3 kopije u površinskim ribama (Slika 53). Slično, nalazimo do 4 kopije gena *LOC125799116* u genomima površinskih riba te 6-40 kopija u genomima špiljskih riba (Slika 53).

Ostalih 57 od 102 CNV gena imaju poznatu funkciju. Ovi geni sudjeluju u različitim biološkim procesima, među kojima su najzastupljeniji imunološki odgovor (*LOC111188594 - fucolectin-1-like; LOC111194616 - polymeric immunoglobulin receptor-like; LOC111195410 - E3 ubiquitin-protein ligase DTX3L; LOC125785782, LOC125785783, LOC111188451 i LOC125785616 - B-cell receptor CD22-like; LOC111197671 - C-type lectin domain family 4 member E-like; LOC125784871, LOC125785663 i LOC125784856 - NLR family CARD domain-containing protein 3-like*), prijenos kisika (*LOC111191628, LOC111196759, LOC111191630 i LOC103027764 - encoding hemoglobin subunits; LOC125787068 - scavenger receptor cysteine-rich type 1 protein M130-like*) i metabolizam lipida (*LOC103026892 - 60 kDa lysophospholipase; LOC125799429 - phospholipase B-like 1; LOC111195147 - apolipoprotein L3-like*). Neki od njih imaju reduciran broj kopija u špiljskim ribama u odnosu na površinske, kao npr. gen *LOC125784890* (*trace amine-associated receptor 13c*) koji kodira za olfaktorni receptor (Slika 54). Slično, gen *LOC111195147* za apolipoprotein L3 je potpuno izbrisan u genomima špiljskih riba a prisutan je u jednoj ili dvije kopije u površinskim populacijama (Slika 54). Apolipoprotein L3 je kod ljudi uključen u kretanje lipida (uključujući kolesterol) unutar citoplazme i vezanje lipida na organele („Gene“ baza podataka - National Library of Medicine; Gene ID: 80833).

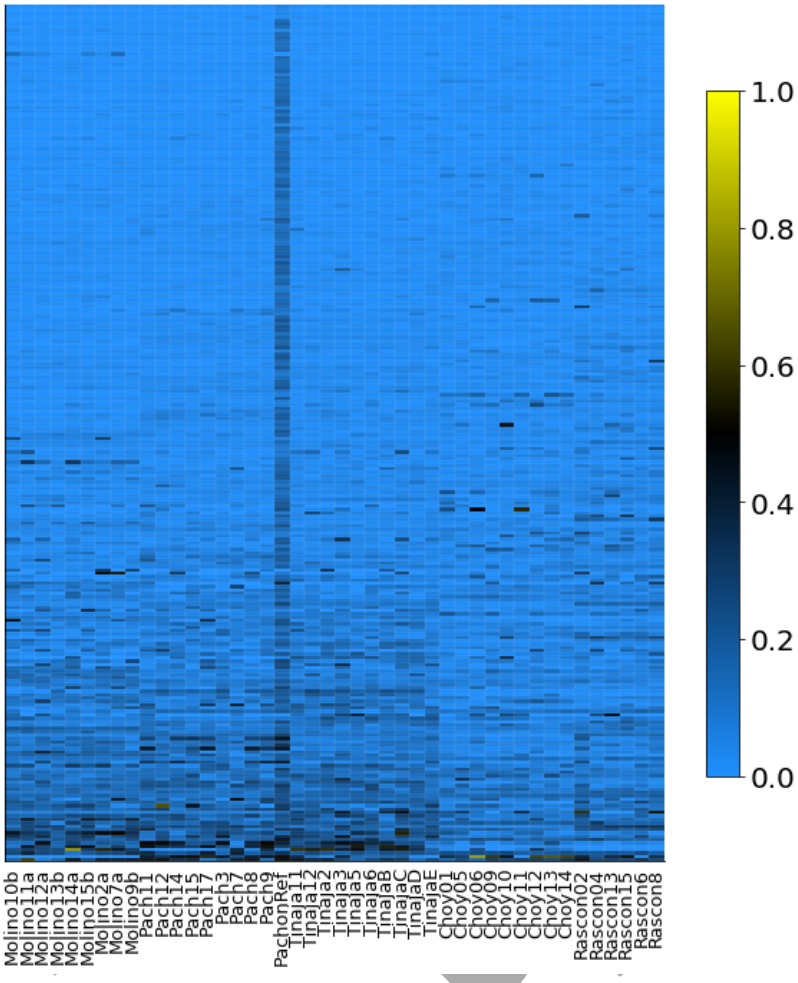


Slika 53. Neki od gena nepoznate funkcije sa značajnim razlikama u broju kopija između špiljskih (plavo) i površinskih (crveno) jedinki. Simbol gena je naznačen na vrhu svakog prikaza. Mol – Molino; Pach - Pachón; Tin – Tinaja; Choy - Río Choy; Ras – Rascon.



Slika 54. Neki od gena sa značajno većim brojem kopija u površinskih (crveno) riba u odnosu na špiljske (plavo). Simbol gena je naznačen na vrhu svakog prikaza. Mol – Molino; Pach - Pachón; Tin – Tinaja; Choy - Río Choy; Ras – Rascon.

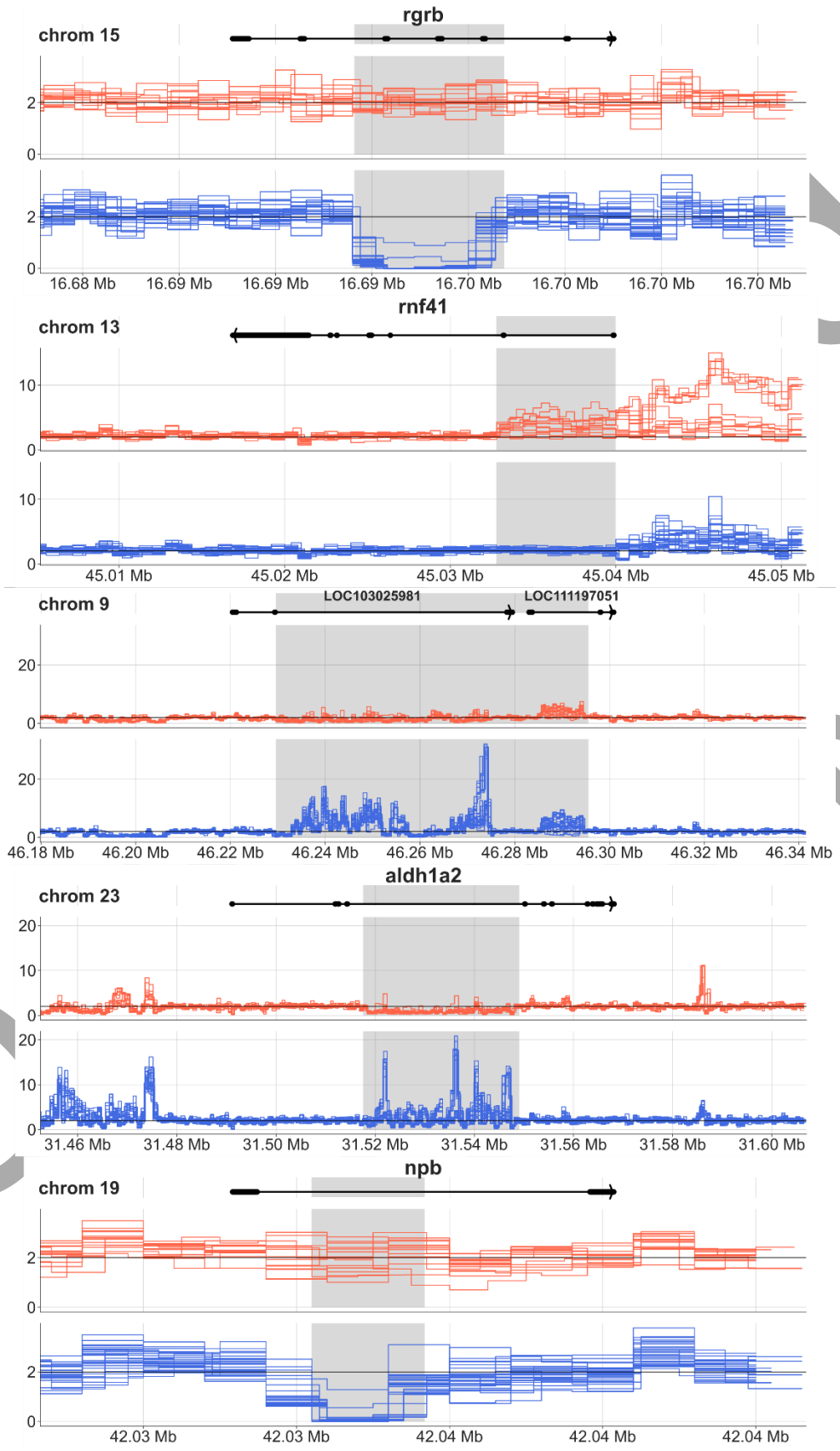
Analizirali smo udio očitanja s niskom kvalitetom mapiranja (MAPQ = 0) na svakom od 102 divergentna CNV gena (Slika 55). U većini slučajeva nije bilo mapiranih očitanja niske kvalitete, a njihov udio je manji od 5 % kod približno 90 % analiziranih CNVRs, što sugerira da očitanja niske kvalitete nemaju značajan utjecaj na ove rezultate.



Slika 55. Udio očitanja niske kvalitete za svaki od 102 divergentna CNV gena (po redovima) po pojedinom uzorku (stupcu). Udjeli su prikazani nijansama boja prema shemi na slici desno.

Među CNVs s visokom učestalošću u pojedinom ekotipu nalazimo i one koji zahvaćaju gene dijelom njihove duljine. Jedan od takvih primjera je gore spomenuti gen *rgrb* koji sudjeluje u procesu fototransdukcije. U svim analiziranim genomima špilja, dio ovog gena koji obuhvaća eksone 4 i 5 je deletiran (Slika 56). Sličan primjer predstavlja i gen *rnf41* u kojem su prvi intron i prva dva eksona umnoženi u svim analiziranim površinskim ribama (Slika 56).

LOC111197051 i *LOC103025981* su u AstMex3 anotirani kao dvije kopije istog gena koji kodira gvanilil ciklaza-aktivirajući protein 2, a koji je uključen u fototransdukciju. Ovi geni se preklapaju sa ~66 kbp dugom genomskom regijom koja je duplicirana kod špiljskih riba (Slika 56). Slično, broj kopija genomske regije duge 31 kbp koja obuhvaća dio gena *aldh1a2* je specifično povećan kod špiljskih jedinki (Slika 56). Ovaj gen kodira retinaldehid dehidrogenazu 2, enzim neophodan za pravilnu morfogenezu tijekom embrionalnog razvoja (Niederreither i sur. 2002). Dio gena *npb* duljine ~1 kbp djelomično je deletiran u špiljskim ribama, odnosno prisutan u jednoj kopiji po diploidu (Slika 56). Ovaj gen kodira za neuropeptid B koji se eksprimira u središnjem živčanom sustavu i uključen je u regulaciju ponašanja pri hranjenju (Singh i Davenport 2006).

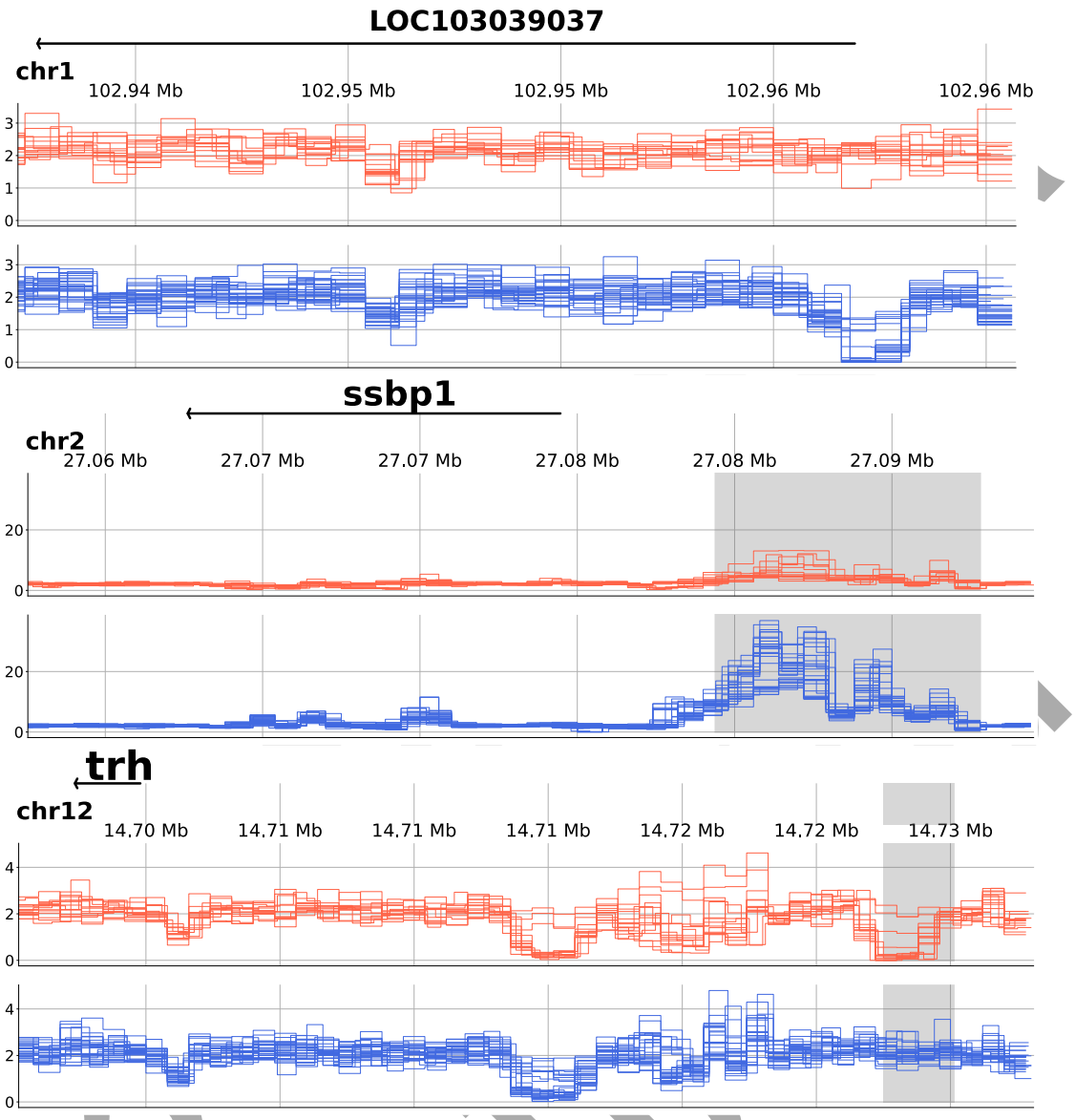


Slika 56. Grafički prikaz dubine očitanja na regijama odabranih gena. Normalizirana dubina očitanja prikazana je za špiljske (plavo) i površinske (narančasto) jedinke, kao diploidni broj kopija (ordinata) po segmentu referentnog genoma (apscisa). Struktura i orijentacija gena prikazani su strelicama iznad kojih su naznačeni simboli gena. Regije koje su divergentne u broju kopija osjenčane su sivom bojom na grafičkim prikazima.

3.2.4.2. Divergencija u nekodirajućim regijama

Nekodirajuće regije genoma koje se značajno razlikuju brojem kopija između ekotipova mogu također biti važne za adaptaciju. Primjerice, ako zahvaćaju regulatorne elemente, mogu utjecati na razlike u regulaciji obližnjih gena koji su uključeni u specifične biološke procese.

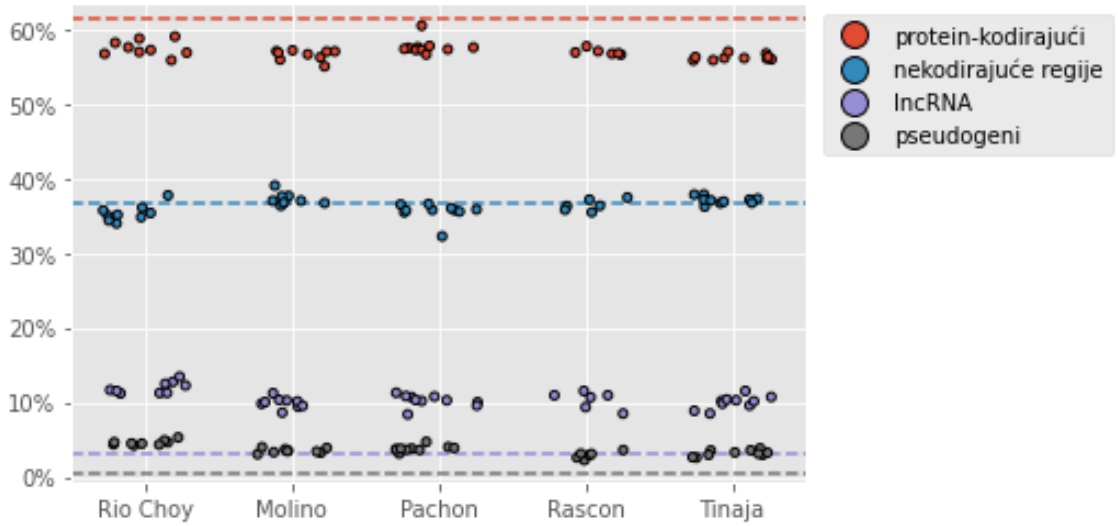
Analizirali smo funkcionalni sastav gena koji su najbliži svakoj od 87 nekodirajućih CNVRs koje smo identificirali kao divergentne po broju kopija. Iako je samo 27 gena imalo pridružene anotacije za biološke procese i puteve, njihov funkcionalni sadržaj je sličan sadržaju protein-kodirajućih CNVRs (Tablica S9 u Pokrovac i sur., 2024). Mnogi od ovih gena su povezani s procesima koji se mijenjaju uslijed prilagodbe na špiljske uvjete života. Primjerice, gen koji kodira za podjedinicu 4 citokrom c oksidaze (*LOC103039037*) nalazi se oko 350 bp nizvodno od regije duge oko 1.4 kb koja je deletirana isključivo kod špiljskih riba (Slika 57). Ovaj gen sudjeluje u oksidativnoj fosforilaciji i kontrakciji srčanog mišića. Delecija genomske regije u blizini ovog gena kod špiljskih riba mogla bi biti povezana s promjenama u njegovoj regulaciji, kao odgovor na hipoksične uvjete u špiljama. Još jedna razlika koja bi mogla biti povezana s prilagodbom na smanjenu razinu kisika u podzemnom okolišu je amplifikacija genomske regije u blizini gena *ssbp1*. Oko 5 kbp uzvodno od ovog gena, i otprilike 8 kbp duljine, ovu regiju nalazimo u 9 do 17 kopija u špiljskim jedinkama (Slika 57). Gen *ssbp1* igra važnu ulogu u biogenezi mitohondrija, i moguće je da amplifikacija ove regije utječe na regulaciju ovog gena a time i na proizvodnju mitohondrija, kao mehanizam kompenzacije na hipoksične uvjete (Gutsaeva i sur. 2008; Gamboa i Andrad 2009). Kao još jedan primjer potencijalne regulacije gena putem CNVs ističe se regija duljine oko 2 kbp koja se nalazi otprilike 27 kbp uzvodno od *trh* gena. Ova genomska regija je deletirana u većini površinskih riba (Slika 57). *Trh* se eksprimira u hipotalamusu kao hormon koji ima važne uloge u mnogim biološkim procesima, uključujući metaboličku i lokomotornu aktivnost, termoregulaciju, percepciju boli i regulaciju spavanja (Wozniak i Quinnell 2015).



Slika 57. Grafički prikaz dubine očitanja na odabranim nekodirajućim CNVRs. Normalizirana dubina očitanja prikazana je za špiljske (plavo) i površinske (narančasto) jedinke, kao diploidni broj kopija (ordinata) po segmentu referentnog genoma (apscisa). Položaj i orijentacija gena prikazani su strelicama iznad kojih su naznačeni simboli gena. Regije divergentne u broju kopija osjenčane su sivom bojom na grafičkim prikazima.

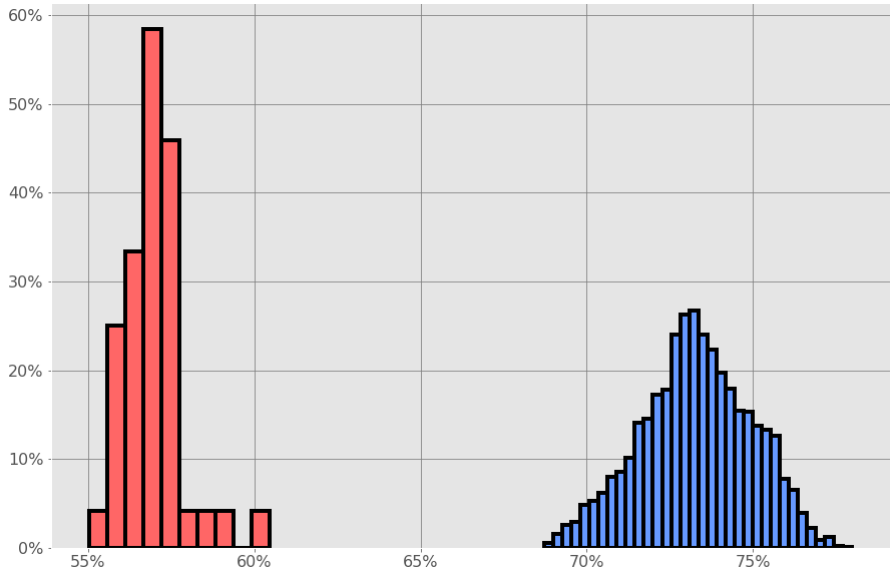
3.2.5. Rezultati permutacijskih analiza

U prosjeku, 57 % svih otkrivenih CNVs pogađa protein-kodirajuće gene (Slika 58), a u prosjeku 12 % zahvaća gene cijelom duljinom. Druga po redu najzastupljenija kategorija su geni koji kodiraju za duge nekodirajuće RNA (eng. *long non-coding RNA*, lncRNA), koje se preklapaju s približno 10 % detektiranih CNVs. Otprilike jedna trećina svih CNVs ne pogađa ni jednu kategoriju gena (Slika 58).

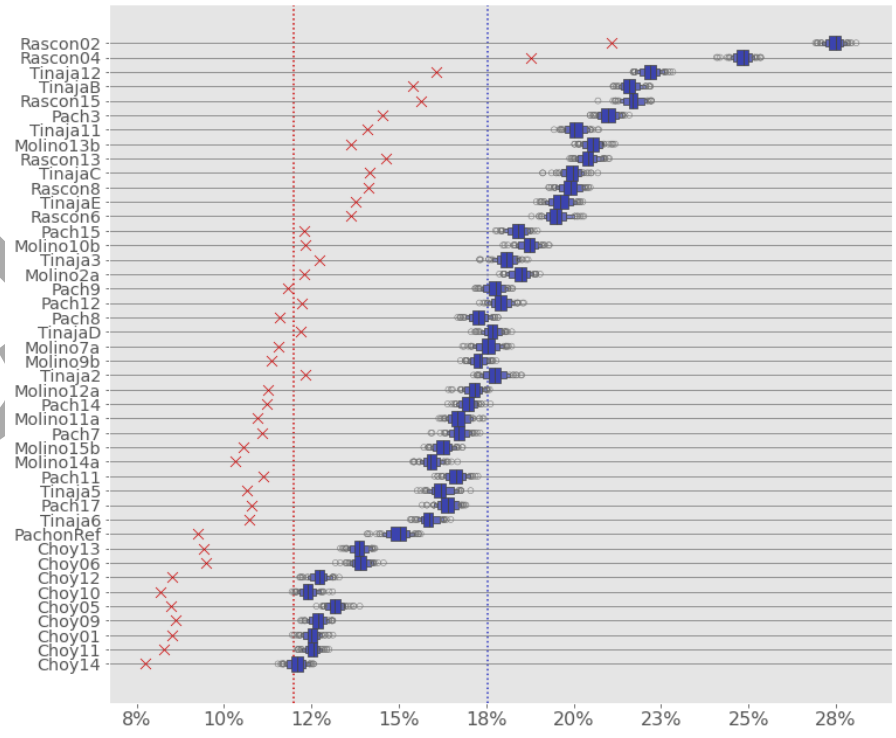


Slika 58. Zastupljenost pojedinih kategorija gena u detektiranim CNVs. Različitim bojama prikazani su udjeli CNVs koji pogađaju tri najzastupljenije kategorije gena i nekodirajuće regije. Isprekidane linije označavaju udio pojedine kategorije u AstMex3 genomu.

Permutacijske analize pokazuju da očekivani udio CNVs koji pogađaju protein-kodirajuće gene iznosi od 69 % do 78 % (Slika 59). Međutim, udjeli detektirani u analiziranim genomima se kreću od 55 % do 61 %. Ova razlika sugerira da su, generalno, CNVs koji pogađaju protein-kodirajuće gene pod utjecajem negativne prirodne selekcije. To potvrđuju i udijeli zahvaćenih gena: na temelju permutacija, za očekivati je da će 18 % svih gena koji kodiraju proteine biti zahvaćeno CNVs, u usporedbi s 12 % koliko ih nalazimo u prosjeku u stvarnim podacima (Slika 60).

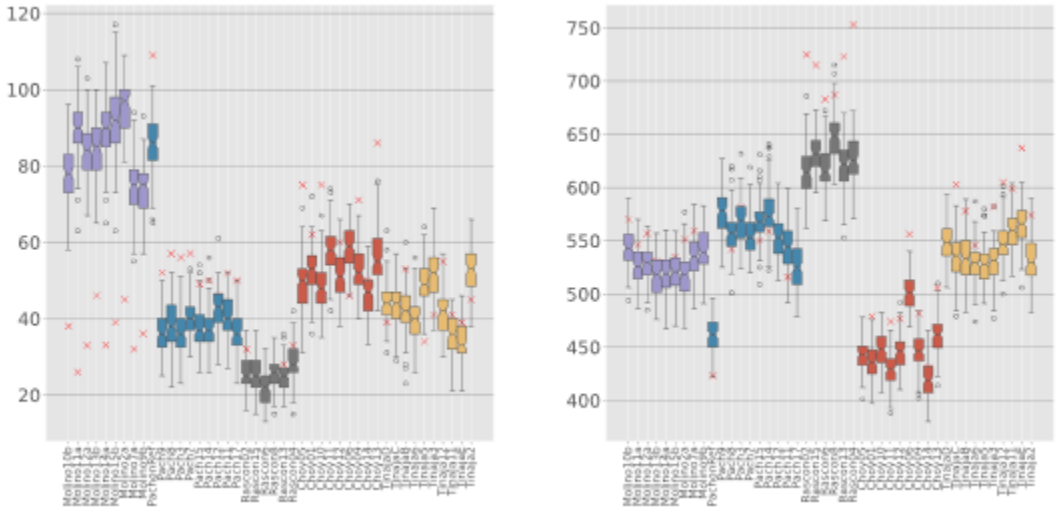


Slika 59. Distribucija postotka CNVS koji pogađaju protein-kodirajuće gene po pojedinom genomu. Distribucija je prikazana za sve analizirane genome (crveno) i za 100 permutacija (plavo).

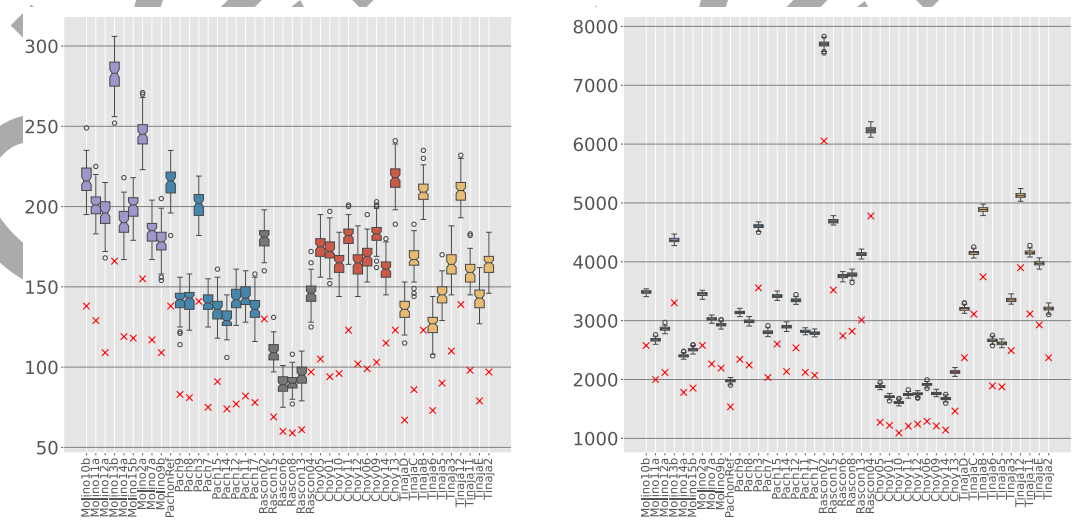


Slika 60. Udio protein-kodirajućih gena koji su pogođeni sa CNVs u stvarnim podacima (crveni križići) i distribucija očekivanih udjela na temelju permutiranih podataka (plavi *boxplot* prikazi). Podaci su prikazani po pojedinačnim genomima, označenim imenima uzoraka s lijeve strane. Isprekidane linije predstavljaju prosječnu vrijednosti za stvarni (crveno) odnosno permutirani (plavo) skup podataka.

Uz iznimku Molino populacije, broj CNVs koji sadrže kompletne gene je u skladu s očekivanim brojem, prema rezultatima permutacijskih analiza (Slika 61). Međutim, broj duplikacija i delecija koje se djelomično preklapaju s genom je znatno manji od očekivanog (Slika 62).

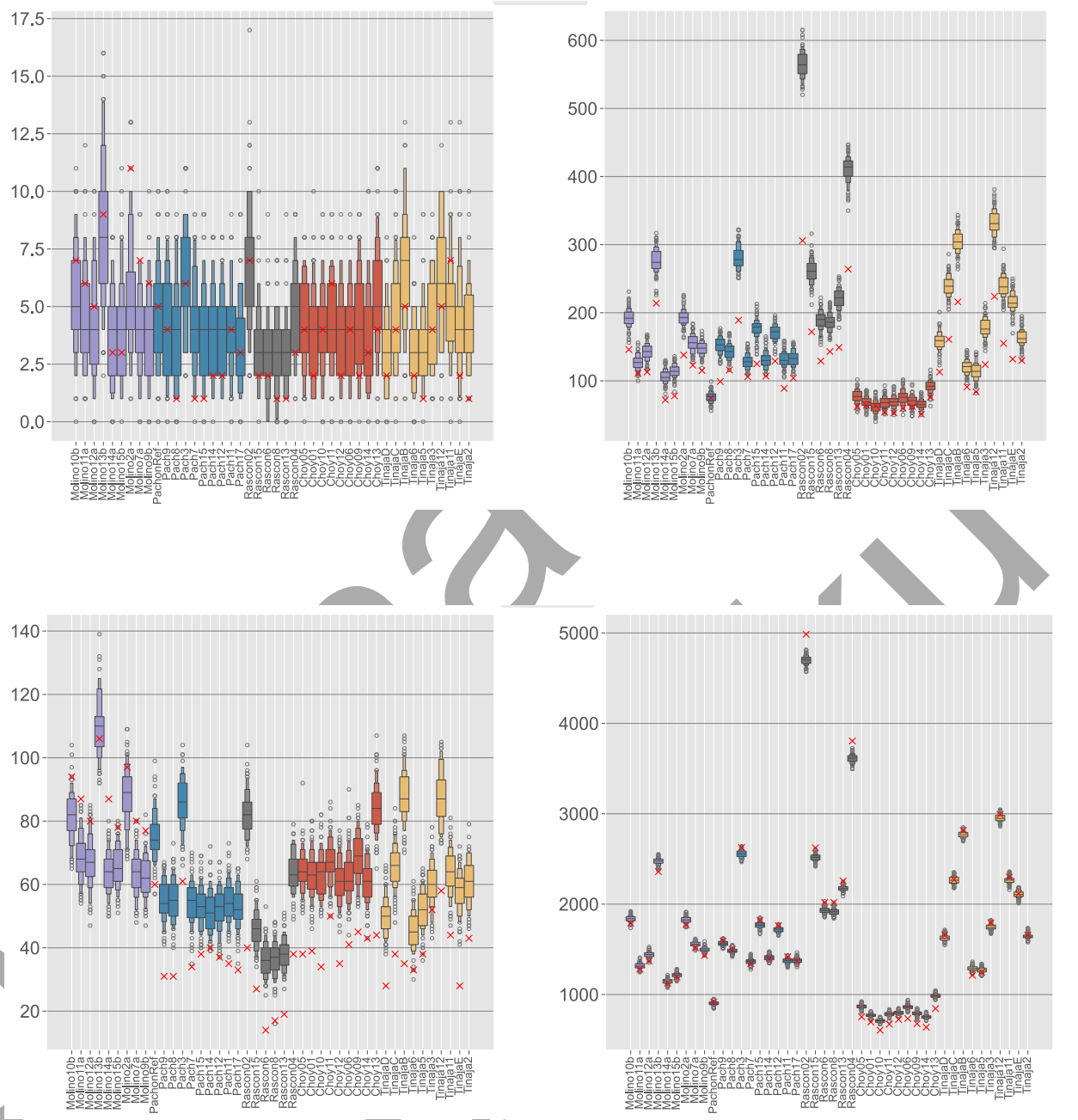


Slika 61. Broj duplikacija (lijevi dijagram) i delecija (desni dijagram) koje sadrže čitave gene. Vrijednosti su naznačene crvenim križićima za stvarne podatke. *Boxplot* prikazima (obojeni prema populacijama) predstavljene su distribucije vrijednosti dobivenih na temelju 100 permutacija.

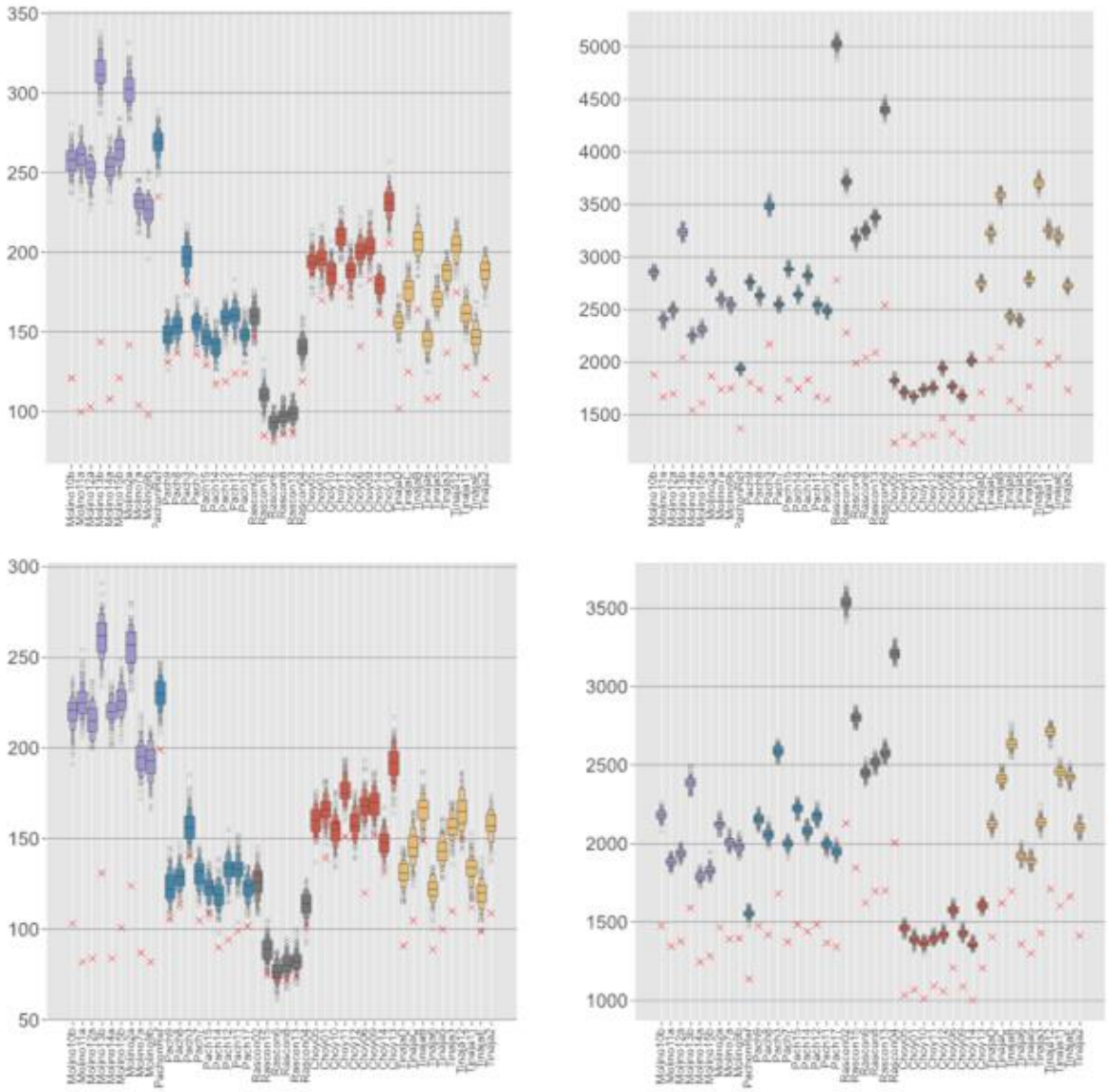


Slika 62. Broj duplikacija (lijevi dijagram) i delecija (desni dijagram) koje pogađaju gene dijelom njihove duljine. Vrijednosti su naznačene crvenim križićima za stvarne podatke. *Boxplot* prikazima (obojeni prema populacijama) predstavljene su distribucije vrijednosti dobivenih na temelju 100 permutacija.

Kako bismo ustanovili jesu li ovi rezultati uvjetovani neobično velikom duljinom introna u meksičkoj tetri (Jakt i sur. 2022), analizirali smo preklapanje CNVs sa eksonima i intronima. Usporedba permutiranih i stvarnih podataka sugerira da je broj CNVs koji pogađaju dio introna ili eksona u stvarnim podacima očekivan (Slika 63). Slično vrijedi i za duplikacije koje zahvaćaju cijele introne i eksone, uz iznimku Molino populacije u kojoj nalazimo manji broj takvih CNVs od očekivanog. Međutim, delecije koje zahvaćaju cijele introne i eksone nalazimo u manjem broju nego što bismo očekivali prema njihovoj nasumičnoj raspodjeli u genomu (Slika 64).



Slika 63. Broj duplikacija (lijevi dijagrami) i delecija (desni dijagrami) koje pogađaju eksone (gore) i introne (dolje) dijelom njihove duljine. Vrijednosti su naznačene crvenim križićima za stvarne podatke. *Boxplot* prikazima (obojeni prema populacijama) predstavljene su distribucije vrijednosti dobivenih na temelju 100 permutacija.



Slika 64 Broj duplikacija (lijevi dijagrami) i delecija (desni dijagrami) koje pogađaju čitave eksone (gore) i introne (dolje). Vrijednosti su naznačene crvenim križićima za stvarne podatke. *Boxplot* prikazima (obojeni prema populacijama) predstavljene su distribucije vrijednosti dobivenih na temelju 100 permutacija.

Na sličan način smo usporedili permutirane i stvarne podatke za preklapanje CNVs s ostalim anotiranim kategorijama gena u AstMex3 genomu: lncRNA, pseudogeni, tRNA, snoRNA (eng. *small nucleolar RNA*), snRNA (eng. *small nuclear RNA*) i rRNA (Slika S1 u Prilogu). Usporedba rezultata permutacijskih analiza sa stvarnim podacima za sve genske kategorije sažeta je u Tablici 21.

Tablica 21. Usporedba očekivanog i stvarnog broja CNVs koji zahvaćaju anotirane kategorije gena. U usporedbi s permutiranim podacima, broj u stvarnim podacima je očekivan (0), manji (-), ili veći (+). Zbog male duljine gena, u nekim kategorijama nije bilo moguće detektirati djelomična preklapanja sa CNVs te su označeni sa Ø. Kategorije koje sadrže iznimke od naznačenog rezultata označene su sa * (za detalje vidjeti like 60 – 63 i slike u Prilogu).

Anotirana genska kategorija	Broj anotiranih elemenata	Delecije cijelog elementa	Duplikacije cijelog elementa	Duplikacije dijela elementa	Delecije dijela elementa
Protein-kodirajući	26,735	0*	0*	–	–
Eksoni	749,966	-	0*	0	0
Introni	280,474	-	0*	0*	0
tRNA	9,987	+	0	Ø	Ø
rRNA	9,252	+	0	Ø	Ø
lncRNA	3,603	+	0*	0	0
Pseudogeni	1,376	+	0*	0	+
snRNA	1,314	0	0	Ø	Ø
snoRNA	297	0	0	Ø	Ø

4. RASPRAVA

4.1. Utjecaj okolišnih čimbenika na strukturne varijacije u genomu unutar jedne generacije

U ovom istraživanju koristili smo srođeni soj kućnog miša u pokusu prehrane bogate mastima kako bismo istražili na koji način okolišni čimbenici mogu izravno utjecati na strukturne varijacije u genomu. Glavni fokus istraživanja su bile promjene koje nastaju u genomima spermija, kao izravnih nositelja nasljednog materijala. Optičko mapiranje je odabrano kao prikladna metoda koja se može primijeniti za detekciju strukturnih varijanti u genomima tako genetički heterogene populacije stanica kao što su spermiji, budući da omogućava praćenje varijanti na razini pojedinačnih DNA molekula.

4.1.1. Protokol za izolaciju visokomolekularne DNA iz spermija za potrebe optičkog mapiranja genoma

U sklopu ove disertacije razvijen je protokol za izolaciju HMW DNA iz spermija za potrebe optičkog mapiranja genoma. Spermiji se odlikuju strukturom kromatina koja je znatno čvršća od kromatina somatskih stanica, zahvaljujući protaminskoj komponenti koja zamjenjuje histone tijekom spermatogeneze. Postojeći komercijalni protokol za izolaciju visokomolekularne DNA iz somatskih stanica (*Bionano Genomics*) nije stoga prikladan za primjenu na spermijima.

S obzirom na to da je ovo prvi put da je tehnologija optičkog mapiranja primijenjena na genome spermija, bilo je potrebno procijeniti kvalitetu izlaznih podataka i usporedivost s podacima dobivenima na somatskom tkivu primjenom standardnog protokola kompanije *Bionano Genomics*. Prosječna duljina molekula izoliranih iz spermija je znatno manja od one iz tkiva bubrega, što nakon uklanjanja niskokvalitetnih molekula rezultira manjom pokrivenošću genoma spermija nego genoma bubrega. No, unatoč tome, nakon provedenog de novo slaganja genoma, kvaliteta dobivenih podataka je zadovoljavala preporučene vrijednosti po svim potrebnim parametrima, uključujući konačnu prosječnu duljinu preostalih molekula, gustoću obilježenosti, udio molekula poravnatih molekula, i kontinuitet složenosti genomskih mapa. Osim toga, po svim pokazateljima kvalitete, osim po pokrivenosti genoma (koja je također bila iznad preporučene vrijednosti), genomske mape spermija se nisu razlikovali od genomskih mapa bubrega. Ovo ukazuje na usporedivost rezultata između spermija i bubrega, odnosno visoku pouzdanost nizvodnih analiza. U širem smislu, pokazuje da je razvijeni protokol za izolaciju visokomolekularne DNA iz spermija prikladan za potrebe optičkog mapiranja, i efikasan za dobivanje genomskih mapa visokog kontinuiteta koje su usporedive s mapama iz somatskog tkiva.

Razvijeni protokol omogućava istraživanje strukturnih varijacija u stanicama koje direktno prenose genetički materijal u sljedeću generaciju, kroz izravnu detekciju *de novo* varijanti i pomaka u frekvenciji postojećih varijanti, koje bi mogle biti povezane s različitim bolestima i poremećajima i/ili uzrokovane okolišnim čimbenicima. U tom smislu, optičko mapiranje genoma spermija po prvi put otvara mogućnost istraživanja interakcije genoma s okolišem na razini strukturnih varijacija, koja može utjecati na zdravlje potomstva ali i evoluciju.

4.1.2. Strukturna varijabilnost genoma C57BL/6 soja

Od svih mišjih kromosoma, na kromosomu Y je detektiran daleko najveći broj strukturnih varijanti u svim analiziranim uzorcima. Ovaj rezultat je u skladu s prethodnim analizama kod miša (Soh i sur., 2014., Morgan i Pardo-Manuel de Villena, 2017.) i čovjeka (Hallast i sur., 2023.) koje su pokazale iznimnu i varijaciju u strukturi i veličini Y kromosoma. Međutim, ove studije su također sugerirale da su takve varijacije posljedica čestih rearanžmana zbog velikog udjela visoko repetitivnih sekvenci koje izgrađuju većinu sekvence Y kromosoma, a u slučaju miša i ekstremne ekspanzije genskih obitelji. Visoki stupanj repetitivnosti sekvence također uzrokuje lošu složenost Y kromosoma u referentnom genomu, a u našim podacima se onda odražava kao nerazmjerno veliki broj kratkih kontiga, odnosno nedovoljan kontinuitet genomskih mapa da bi se detektirani SVs smatrali pouzdanima. Stoga su strukturne varijante na kromosomu Y izuzete iz svih opisanih analiza.

U međusobnoj usporedbi bilo koja dva uzorka, oko dvije trećine detektiranih SVs se nalazi u oba uzorka. Pronalazimo da se bilo koja dva uzorka u prosjeku razlikuju u postojanju 251 SVs koje ukupno pogađaju 0.17 % genoma u prosjeku. Ovakva međusobna sličnost se čini visokom no treba uzeti u obzir da je riječ o srođenom organizmu. Primjerice, SVs pogađaju 0.33 % ljudskog genoma u prosjeku (Audano i sur., 2019), što je samo dva puta više od onoga što smo našli za C57BL/6 miševe. Gledano u cijelom skupu podataka, pronalazimo da je otprilike 4.5 % C57BL/6 genoma podložno strukturnim varijacijama te da je više od 3.8 % genoma heterozigotno s obzirom na SVs. Izvjesno je da bi ovi procijenjeni udijeli bili i znatno veći, ako bi se uzele u obzir i varijante detektirane na Y kromosomu. Stoga, procijenjena razina strukturne varijacije je iznenađujuća naspram očekivanja da su jedinke srođenog soja izogenične u više od 98 % svog genoma (Casellas, 2011). Razlike u strukturnim varijacijama u našim uzorcima mogle bi biti temelj razlikama u fenotipu, primjerice značajnoj varijaciji u težini miševa prije stavljanja u pokus. S obzirom na to da referentni genom potječe od istog soja kao i miševi iz našeg pokusa, ovakvo odudaranje od referentnog genoma je neočekivano i sugerira da se kolonija miševa koju održava Institut Ruđer Bošković genetički značajno odvojila od originalnog soja. Budući da je ovo istraživanje obuhvaćalo isključivo strukturne varijante, ostaje nepoznato doprinose li drugi oblici genetičkih varijacija tom odvajanju, primjerice varijante jednog nukleotida (SNVs).

4.1.3. Prehrana bogata mastima pojačava strukturne varijacije u genomu

Nalazimo značajno više SVs u spermijima nego u bubrezima životinja pokusne skupine, dok tu razliku ne detektiramo u kontrolnoj skupini. Ovo sugerira da prehrana bogata mastima povećava stopu strukturnih varijacija. Nismo pronašli značajne razlike u broju *de novo* varijanti između pokusne i kontrolne skupine niti između tkiva. Stoga, naši rezultati sugeriraju da prehrana masnom hranom utječe na povećanje učestalosti varijacije u genomskim regijama postojećih strukturnih varijacija.

U genomima spermija, delecije i duplikacije su u prosjeku dulje u pokusnoj skupini u odnosu na kontrolnu skupinu. Iako ove razlike nisu velike, značajne su i ne nalazimo ih u genomima bubrega niti za jedan tip varijanti. Stoga je moguće da prehrana bogata mastima utječe na povećanje broja i duljine delecija i duplikacija.

U pokusnoj skupini, broj heterozigotnih SVs je statistički značajno veći u spermijima u odnosu na bubrege. Ovu razliku ne nalazimo u kontrolnoj skupini, što sugerira da prehrana bogata mastima utječe na povećanje heterozigotnosti u spermijima.

Gledano generalno u cijelom skupu uzoraka, homozigotne insercije, delecije i inverzije su značajno kraće od heterozigotnih. Suprotno tome, homozigotne duplikacije su oko tri puta dulje od heterozigotnih duplikacija. Ovakve razlike u duljini mogu se objasniti selekcijskim pritiskom: zbog potencijalno većeg štetnog efekta, očekivali bismo jači pritisak na homozigote nego heterozigote, pa stoga i kraće homozigotne varijante u odnosu na heterozigotne. Međutim, u slučaju duplikacija, zbog njihove repetitivne strukture, heterozigotnost na velikim duplikacijama bi mogla uzrokovati genomsku nestabilnost i tako izazvati veću štetu od homozigotnosti. Zanimljivo je da samo u genomima spermija iz uzoraka pokusne skupine ne vidimo ovu razliku u duljini između heterozigotnih i homozigotnih duplikacija: prosječna duljina homozigotnih duplikacija je u ovim uzorcima slična prosječnoj duljini heterozigotnih duplikacija svih uzoraka. Ovo specifično smanjenje duljine homozigotnih duplikacija u spermijima pokusnih miševa, u kombinaciji s povećanim udjelom heterozigotnih SVs u odnosu na homozigotne, ukazuje na to da se tijekom spermatogeneze kod pokusnih miševa povećala učestalost nealelne homologne rekombinacije (NAHR) na velikim homozigotnim duplikacijama, koja stvara raspon duljina heterozigotnih produkata.

Geni su značajno češće pogođeni strukturnim varijantama u pokusnoj skupini. Pronalazimo da je u prosjeku 85 više gena pogođeno u genomu bubrega miševa iz pokusne skupine nego u genomu bubrega miševa iz kontrolne skupine. U genomima spermija ta razlika je još i veća: čak 267 gena više je u prosjeku

zahvaćeno strukturnim varijacijama u pokusnoj skupini nego u kontrolnoj. Gledano unutar pokusne skupine, u prosjeku je 206 gena više pogođeno strukturnim varijantama u genomima spermija nego u genomima bubrega, dok ta razlika u kontrolnoj skupini iznosi samo 24 gena u prosjeku. Ovi rezultati sugeriraju da prehrana bogata mastima uzrokuje povećanje varijacije u strukturi gena, i to posebice u spermijima. Unatoč ovim razlikama u brojčanom sadržaju gena, analiza njihove ontologije nije pokazala specifičnosti pojedinih skupina, vezane uz funkcionalni sadržaj. Drugim riječima, nema razlike u skupu gena između pokusne i kontrolne skupine s obzirom na biološke procese u koje su uključeni. Nalazimo da su strukturne varijante generalno obogaćene genima koji sudjeluju u metabolizmu šećera, pohrani lipida, aktivnosti citoskeleta te reprodukciji. Jedina značajna razlika u obogaćenju gena koja bi moglo biti uzrokovana pokusom je vezana uz negativnu regulaciju lipidnog metaboličkog procesa (GO:0045833). Specifično povećanje obogaćenja ovog termina u bubregu pokusnih životinja bi moglo biti povezano s promjenama u regulaciji metaboličkih procesa kod životinja na masnoj hrani, a koje su uzrokovane strukturnim varijacijama.

4.1.4. Varijacije u duljini telomera

U sklopu ove disertacije razvijen je bioinformatički alat TelOMpy, za određivanje duljine telomera iz podataka optičkog mapiranja genoma. TelOMpy koristi izlazne podatke BioNano Solve paketa (Bionano Genomics), a koji su rezultat standardiziranog postupka za *de novo* sastavljanja genoma iz produciranih optičkih mapa genoma (Bionano Genomics). Iz tih podataka TelOMpy izračunava apsolute duljine pojedinačnih telomera na pojedinačnim kromosomima, što je velika prednost koja izdvaja ovaj alat naspram drugih često korištenih metoda za određivanje i uspoređivanje duljine telomera. Naime, jedna od najčešće korištenih metoda za mjerenje duljine ponavljanja telomernog motiva je qPCR (Cawthon, 2002; Cawthon, 2009). Iako zahtjeva znatno manje količine DNA i jednostavniji te jeftiniji postupak za pripremu uzorka, qPCR metodom je moguće odrediti samo relativnu duljinu telomera, odnosno utvrditi razliku u duljini u odnosu na referentni uzorak. Osim toga, ova metoda daje procjenu prosječne duljine telomera između molekula DNA koje su prisutne u qPCR reakciji. Terminalni restrikcijski fragment (TRF) je također metoda koja daje procjenu prosječne duljine telomera u populaciji stanica i neosjetljiva je na vrlo kratke telomere (Montpetit i sur., 2014). STELA (*Single-telomere length analysis*) i Q-FISH (*quantitative fluorescence in situ hybridization*) su metode za mjerenje duljine pojedinačnih telomera s rezolucijom od 0,1 kb (Aubert i sur., 2012; Montpetit i sur., 2014). Međutim, prva je obično ograničena na telomere nekolicine kromosoma a druga na stanice u metafazi (Aubert i sur., 2012; Montpetit i sur., 2014). Dosad su opisane dvije metode za određivanje duljina pojedinačnih telomera iz podataka optičkog mapiranja. Jedna od njih zahtjeva CRISPR/Cas9 sustav za obilježavanje telomernog motiva, što stvara dodatne komplikacije i trošak u provedbi (McCaffrey i sur., 2017). Druga pak nije dovoljno detaljno opisana da bi se mogla reproducirati, niti je implementirana u softver koji bi bio javno dostupan (Young i sur., 2017). Za razliku od navedenih metoda, TelOMpy je javno dostupan alat za određivanje apsolutne duljine telomera širokog raspona iz optičkih mapa genoma, koji ne zahtijeva korištenje dodatne tehnologije. TelOMpy je moguće implementirati bez dodatnog troška u bilo koje istraživanje koje uključuje standardne postupke *de novo* sastavljanja genoma i detekcije strukturnih varijanti iz podataka optičkog mapiranja. Primjerice, optičko mapiranje se često koristi za istraživanje genomskih rearanžmana u tumorskim tkivima, u kojima promjene u duljini telomera čine dio kancerogenih procesa (Fan i sur., 2021).

Iz istih podataka optičkog mapiranja genoma na kojima smo proveli analize strukturnih varijanti u pokusu prehrane bogate mastima, usporedili smo životinje u pokusnoj i kontrolnoj skupini s obzirom na duljinu telomera i pokazali izvanrednu moć TelOMpy alata za detekciju suptilnih razlika. Gledano ukupno, telomere spermija u pokusnoj skupini su u prosjeku značajno kraće od telomera spermija u kontrolnoj skupini, što

je moć detekcije na razini metode qPCR. Ovaj rezultat bi sugerirao da je prehrana bogata mastima povezana sa skraćivanjem telomera u spermijima. Međutim, detaljnije usporedbe - na razini pojedinačnih kromosoma između tkiva bubrega i spermija pojedinog miša, ne podupiru ovaj zaključak jer ne pokazuju konzistentno značajno skraćivanje telomera isključivo u spermijima pokusnih miševa.

Sposobnost TelOMpy alata da izračuna duljine pojedinačnih telomera nam je omogućila uvid u njihovu izvanrednu varijabilnost. Pronalazimo da telomere variraju duljinom u rasponu od tri reda veličine - od nekoliko stotina do nekoliko stotina tisuća baznih parova. Ovu varijaciju nalazimo između i unutar jedinki i tkiva, te između i unutar kromosoma što je u skladu s polimorfizmom duljine telomera unutar somatskih stanica koji je prethodno pokazan kod ljudi (Lansdorp i sur., 1996; Schmidt i sur., 2024).

Telomere spermija su u pravilu dulje od telomera somatskih stanica, zbog aktivnosti telomeraze u spermatogenim stanicama koja produljuje telomere (Fice i Robaire, 2019). Međutim, kod svih analiziranih miševa u našem istraživanju nalazimo da su telomere spermija konzistentno kraće od telomera bubrega. Kontrolne analize podataka na temelju duljine molekula sugeriraju da ovi rezultati nisu tehnički artefakti, nego predstavljaju stvarno biološko stanje. Iako rijetko, skraćivanje telomera u spermijima zabilježeno je prethodno, kao rezultat starenja (de Frutos i sur., 2016) ili cirkulacijskog poremećaja u testisima (Tahamtan i sur., 2019). Osim toga, pokazano je i da mutacije u različitim genima mogu dovesti do smanjenja aktivnosti telomeraze i skraćivanja telomera (Margalef i sur., 2018; Mirabello i sur., 2010; Soerensen i sur., 2012; Bojesen i sur., 2013; Grill i Nandakumar, 2021). Stoga je moguće da je relativno kraća duljina telomera u spermijima u odnosu na bubrege u našim uzorcima rezultat mutacije u jednom ili više gena među mnogima koji sudjeluju u održavanju duljine telomera. Međutim, za potvrdu ovoga trebalo bi analizirati varijante jednog nukleotida (SNVs) a podaci optičkih mapa genoma koje smo imali na raspolaganju ne pružaju tu mogućnost. Važno je istaknuti da eventualne mutacije koje bi umanjile aktivnost telomeraze ne bi nužno bile detrimentalne, budući da su miševi deficijentni za telomerazu vijabilni, fertilni i bez vidljivih morfoloških abnormalnosti (Blasco i sur., 1997). Alternativno, moguće je da je relativno veća duljina telomera u tkivu bubrega nego u spermijima odraz aktivnosti telomeraze u stanicama bubrega. Poznato je da miševi eksprimiraju telomerazu u nekim somatskim tkivima (Prowse i Greider, 1995) a aktivnost telomeraze u različitim somatskim tkivima - uključujući bubreg, zabilježena je i kod kokoši (Venkatesan i Price, 1998).

Sasvim iznenađujuće, i u spermijima i u tkivu bubrega pronalazimo značajno manji broj telomera u pokusnoj nego u kontrolnoj skupini, što bi moglo biti odraz gubitka čitavih telomera uslijed prehrane masnom hranom. Potpuni gubitak telomera na kromosomu je fenomen koji je prethodno zabilježen u

tumorskim stanicama (Fouladi i sur., 2000; Lo i sur., 2002; Bai i Murnane, 2003; Ferreira i sur., 2004; Lin i sur., 2010; Raseley i sur., 2023) i mutantima (Vannier i sur., 2012). Naše istraživanje po prvi put sugerira da prehrana može inducirati gubitak čitavih telomera, u somatskim i reproduktivnim stanicama.

4.2. Utjecaj okolišnih čimbenika na varijacije u broju kopija na mikroevolucijskoj razini

Iako su strukturne varijante česti oblik genetičkih varijacija, proučavanje njihove uloge u ekološkoj prilagodbi kod kompleksnih višestaničnih organizama je otežano, prvenstveno zbog nemogućnosti primjene metoda eksperimentalne evolucije. Slučajevi paralelne prilagodbe na slične okolišne uvjete koji su obilježeni jakim selektivnim pritiscima predstavljaju priliku za proučavanje uloge strukturnih varijanti u prilagodbi. U ovoj disertaciji koristili smo model paralelne prilagodbe ribe meksičke tetre (*Astyanax mexicanus*) na špiljske uvjete života, kako bismo identificirali strukturne varijante koje opetovano doprinose prilagodbi u višestrukim populacijama koje su neovisno prešle iz površinskih u špiljske vode. Ovo istraživanje temeljilo se na postojećim podacima sekvenciranih genoma a ograničeno je na analize varijacija u broju kopija (CNVs), odnosno duplikacija (amplifikacija) i delecija, što je uvjetovano vrstom podataka (NGS) i metodom za detekciju CNVs koja se prethodno pokazala pouzdanom i robusnom za ovakav tip analiza u drugim vrstama (Abyzov i sur., 2011; Pezer i sur., 2015; Kosugi i sur., 2019; Garg i sur., 2021).

4.2.1. Varijabilnost *A. mexicanus* genoma

Procjenjujemo da je jedna petina referentnog genoma podložna varijacijama u broju kopija u prirodnim populacijama *A. mexicanus*, od kojih većina pogađa gene. Međutim, ovo je direktna posljedica složenosti referentnog genoma: samo trećinu AstMex3 genoma izgrađuju sekvence izvan gena. Prethodne studije na ljudima pokazale su da su upravo nekodirajuće repetitivne regije one koje su izrazito bogate strukturnim varijantama, uključujući CNVs (Huddleston i sur., 2016.; Audano i sur., 2019.; Ebert i sur., 2021.). Kako se dijelovi referentnog genoma meksičke tetre budu upotpunjavali sekvencama koje trenutačno nedostaju a anotirani su kao praznine (eng. *gaps*), možemo očekivati i povećanje udjela CNVs u nekodirajućim regijama.

Među genima koji variraju brojem kopija, nalazimo neočekivano veliki broj onih sa značajkama evolucijski mladih gena. Prethodno istraživanje na koljuški pokazalo je da su CNVs obogaćeni evolucijski novim genima koji su općenito kraći i za koje nisu pronađeni dokazi homologije u drugim vrstama (Chain i sur., 2014.). Njihova je funkcija često nepoznata pa se u anotacijama ti geni opisuju kao "nekarakterizirani" (eng. *uncharacterized*). Budući da je AstMex3 genom relativno nedavno sastavljen pa se anotacija gena može smatrati „nezrelom“, moguće je da neki od ovih gena nisu doista evolucijski mladi. Međutim, nepoznata funkcija ovih gena, kratka duljina i snažno obogaćenje u usporedbi s očekivanjima temeljenim na permutacijama u našem istraživanju dodatno podupiru hipotezu da su evolucijski mladi geni skloniji varijaciji u broju kopija (Chain i sur., 2014.).

4.2.2. Genetička raznolikost *A. mexicanus* populacija

Rezultati naših CNV analiza dosljedno ukazuju na manju genetičku raznolikost špiljskih populacija meksičke tetre u odnosu na površinske. Ovo je u skladu s rezultatima prethodnih istraživanja koja su se temeljila na SNP-ovima i mikrosatelitskim lokusima, a može se objasniti kombinacijom male efektivne veličine populacije, ograničene dostupnosti hranjivih tvari i prostora u špiljama, kao i mogućih događaja uskog grla (Bradic i sur., 2012; Bradic i sur., 2013; Herman i sur., 2018). Pretpostavlja se da većina genetičke varijacije u špiljama predstavlja jedan dio postojeće genetičke varijacije iz fonda površinskih predaka (Bradic i sur., 2012). Međutim, otkrili smo velik i usporediv broj genomskih regija koje pokazuju varijaciju broja kopija koja je specifična za ekotip: otkrili smo na tisuće CNVs što u špiljskom, što u površinskom genomu, koji se ne nalaze istodobno u oba. Takvo opažanje otvara mogućnost da se veliki broj CNVs javlja neovisno u oba ekotipa. Međutim, ovi rezultati uvelike ovise o veličini uzorka i potrebno je analizirati mnogo više genoma po ekotipu kako bi se dobila pouzdanija procjena i čvrsto tumačenje ovog rezultata.

4.2.3. Uloga varijanti broja kopija u prilagodbi na špiljske uvjete života

Identificirali smo 292 genomske regije koje sadrže CNVs čiji broj kopija je divergirao između špiljskih i površinskih riba, što sugerira da bi one mogle biti pod selekcijskim pritiskom. Ove regije kumulativno obuhvaćaju gotovo 1 % genoma. Nedavno istraživanje na populacijama koljuške u kontekstu kolonizacije slatkih voda iz morskih staništa je procijenilo da se sličan udio genoma koljuške nalazi pod utjecajem prirodne selekcije u obliku CNVs (Lowe i sur., 2018.). Nalazimo oko stotinu gena čiji broj kopija se razlikuje između ekotipova meksičke tetre, od kojih većina pokazuje svojstva evolucijski mladih gena i veći broj kopija u špiljskim populacijama. Dvije trećine od 292 genomskih regija koje su divergirale brojem kopija imaju veći prosječan broj kopija u špiljskim genomima. Zanimljivo je da je povećanje broja kopija gena, a ne smanjenje, također dominantno kod koljuški koje su se prilagodile slatkovodnim staništima (Hirase i sur., 2014; Lowe i sur., 2018; Ishikawa i sur., 2022). Ove studije zajedno s našim rezultatima sugeriraju da je povećanje broja kopija gena jedan od mehanizama prilagodbe populacije nakon kolonizacije novih i ekstremnih okruženja. Prethodno istraživanje ljudskih CNVs sugeriralo je da su duplikacije pod opuštenijim evolucijskim pritiskom od delecija, pa stoga predstavljaju veću metu za adaptivnu selekciju (Sudmant i sur., 2015.). Štoviše, mehanistički gledano, veća je vjerojatnost da će duplikacije pokazati veće stope mutacije zbog osjetljivosti na nealelnu homolognu rekombinaciju između izravno orijentiranih dupliciranih sekvenci. To im omogućuje da često mijenjaju svoje stanje broja kopija tijekom kratkog vremena (Sudmant i sur.,

2015) i tako opstaju kao multialelni CNVs unutar populacije. Većina divergentnih genomskih regija u našem istraživanju prisutna je u višestrukim stanjima broja kopija u cijelom skupu podataka i mnoge obuhvaćaju čitave gene. Ovi multialelni CNV geni mogu biti osobito relevantni u brzoj prilagodbi na nove okolišne izazove, na način da povećaju fitnes od trenutka njihova nastanka, primjerice kroz povećanje proteinske doze (eng. *protein dosage*) (Kondrashov i sur., 2002; Handsaker i sur., 2015). Iako nalazimo znatan udio CNVs specifičnih za špilje u cijelom skupu podataka, oni koje smo identificirali u divergentnim regijama također donekle variraju brojem kopija i u površinskim populacijama. Naši rezultati su u skladu s prethodnim istraživanjima koja su predložila da se brza prilagodba populacije koja je kolonizirala novo stanište može postići odabirom iz postojeće genetičke varijacije u ancestralnoj populaciji (Jones i sur., 2012.; Lai i sur., 2019.; Zong i sur., 2020.).

Mnoge genomske regije koje su divergentne po ekotipu povezane su s genima uključenima u biološke procese koji su prethodno identificirani kao važni za prilagodbu meksičke tetre životu u špiljama. Pronalazimo CNVs koji pogađaju različite gene (ili se nalaze u njihovoj blizini) čija je zajednička značajka uključenost u obradu informacija o okolišu (Kondrashov i sur., 2002.). Primjerice, produkti ovih gena sudjeluju u funkcijama središnjeg živčanog sustava, obradi vizualnih signala, metabolizmu, potrošnji kisika i imunološkom odgovoru. Paralelna divergencija CNVs povezanih s ovim procesima u špiljskim populacijama sugerira njihovu ulogu u odgovoru na okolišne izazove kao što su stalna tama, niska dostupnost hranjivih tvari, niska razina kisika i razlike u sastavu parazita.

4.2.4. Utjecaj evolucijskih sila na varijacije u broju kopija

Primjećujemo snažnu stratifikaciju CNVs između *A. mexicanus* populacija, u skladu s brojnim istraživanjima provedenima na različitim vrstama (Sudmant i sur., 2015; Pezer i sur., 2015; Xu i sur., 2016; Dorant i sur., 2020; Zhu i sur., 2020; Jang i sur., 2021.; Yang i sur., 2023). Klasteriranje na temelju prisutnosti CNVs podudara se s prethodno utvrđenim filogenetskim odnosima između nove i stare linije *A. mexicanus* populacija na temelju SNP-ova (Herman i sur., 2018). Međutim, unutar stare linije, špiljska populacija Pachón manje je slična špiljskoj populaciji Tinaja, budući da potonja dijeli veći udio CNVs s površinskom populacijom Rascon. Ovo odstupanje je još očitije kada se uzmu u obzir CNV geni: hijerarhijsko grupiranje temeljeno na broju kopija gena smješta populaciju stare linije Pachón bliže populaciji nove linije Molino. Te se razlike mogu objasniti u svjetlu povećane stope mutacije CNVs u usporedbi sa SNPs (Zhang i sur., 2009). Osim toga, drugi čimbenici mogu pridonijeti uočenim obrascima, poput veličine populacije i stupnja migracije između populacija. Primjerice, nedavno istraživanje upućuje na to da se populacija Pachón sastoji od samo nekoliko stotina jedinki koje nastanjuju relativno izolirano područje, što rezultira niskim genetskim polimorfizmom i ograničenim protokom gena s površine ili iz drugih špiljskih populacija (Legendre i sur., 2023.). Sukladno tome, u našim analizama uočavamo nisku raznolikost Pachón populacije na razini CNVs. Stoga, čimbenici poput male veličine populacije i izolacije u kombinaciji s visokom stopom mutacije CNVs su vjerojatno povećali učinak genetskog pomaka (eng. *drift*) na CNVs u populaciji Pachón, što je rezultiralo neočekivanim položajem Pachóna na filogenetskom stablu.

Kako bismo utvrdili postoje li razlike u evolucijskim pritiscima koji djeluju na različite genske kategorije unutar genoma, uspoređivali smo stvarne podatke s rezultatima nasumičnih permutacija CNV koordinata, kao zamjenom za neutralni model evolucije. Pronalazimo da CNVs koji obuhvaćaju čitave protein-kodirajuće gene evoluiraju neutralno, dok su CNVs koji pogađaju dijelove gena pod negativnom selekcijom. Ovo je jednim dijelom u skladu s prethodnim istraživanjem provedenim na ljudskim populacijama koje je sugeriralo da negativna selekcija djeluje na sve strukturne varijante koje pogađaju dijelove protein-kodirajućih gena ali ne i duplikacije koje zahvaćaju čitave gene (Collins i sur., 2020). Također, istraživanja na kolješki su pokazala da duplikacije češće zahvaćaju čitave gene nego delecije (Lowe i sur., 2018). Neutralni model varijacije u broju kopija gena mogao bi objasniti stratifikaciju populacija meksičke tetre koju vidimo na temelju CNV gena, a koja odražava demografsku sliku.

Naši rezultati permutacijskih analiza sugeriraju da se duplikacije čitavih introna ili eksona bolje toleriraju u populaciji nego delecije. Stoga se negativni pritisak na CNVs koji pogađaju dijelove gena može konkretnije objasniti negativnim pritiskom na delecije čitavih eksona ili introna.

Generalno, naše istraživanje sugerira da duplikacije čitavih gena - bez obzira na analiziranu gensku kategoriju, evoluiraju neutralno, dok su delecije čitavih gena pod neutralnim ili pozitivnim pritiskom, kao npr. kod gena koji kodiraju za lncRNA, rRNAs, tRNAs i pseudogene. Jedini negativni pritisak nalazimo na CNVs koji pogađaju dijelove protein-kodirajućih gena.

5. ZAKLJUČCI

Ova disertacija je obuhvatila istraživanje utjecaja okolišnih čimbenika na dvije razine: unutar jedne generacije i na razini populacije. Ovakav pristup zahtijevao je korištenje dva različita modela: 1) srođeni mišji soj, za praćenje genetičkih promjena koje su izravno izazvane u pokusu prehrane bogate mastima kao okolišnom čimbeniku, te 2) prirodne populacije meksičke tetre, kao model paralelne prilagodbe na špiljske uvjete života, odnosno na okolišne čimbenike koji predstavljaju snažni selekcijski pritisak. Zahvaljujući ovakvom dvojakom pristupu, ustanovili smo da okolišni čimbenici ne uzrokuju *de novo* događaje, već pojačavaju učestalost strukturnih varijacija na mjestima u genomu gdje su varijacije već prisutne (eng. *standing genetic variation*). Činjenica da se te promjene ne događaju konzistentno na istim genomskim regijama u pokusu s masnom hranom, govori u prilog tome da je izravan utjecaj okoliša na strukturne varijacije u genomu nespecifičan, a ne usmjeren na određene regije, gene ili biološke funkcije i procese. Naše istraživanje na razini populacija sugerira da je tek evolucijski pritisak taj koji odabire varijante povezane sa specifičnim funkcijama ovisno o tome koliko su korisne ili štetne u danim okolišnim uvjetima. Rezultati na obje razine istraživanja ukazuju na utjecaj okolišnih čimbenika na varijacije u broju kopija, a rezultati istraživanja na mikroevolucijskoj razini sugeriraju da duplikacije imaju posebnu adaptivnu vrijednost.

Iako telomere predstavljaju specijalizirane strukture u kromosomu, tehnički gledano, možemo ih smatrati strukturnim varijacijama u genetičkom smislu, zbog varijabilne duljine i mehanizama koji mogu dovesti do te varijacije a koji su povezani s repetitivnošću telomerne sekvence (Choo i sur., 2023; Brewer i sur., 2024; Waitkus i sur., 2024). Bioinformatički alat TelOMpy koji smo razvili i predstavili ovdje omogućio nam je određivanje apsolutne duljine pojedinačnih telomera iz optičkih genomskih mapa miševa u pokusu. Pomoću TelOMpy-a ustanovili smo da prehrana bogata mastima utječe na smanjenje broja telomera u stanici, što sugerira gubitak čitavih telomera – fenomen koji je dosad zabilježen u tumorskim stanicama.

U ovom istraživanju je po prvi put tehnologija optičkog mapiranja primijenjena na genome spermija. U tu svrhu razvijen je protokol za izolaciju visokomolekularne DNA iz spermija koji je kompatibilan s tehnologijom optičkog mapiranja genoma (*Bionano Genomics*), čime se omogućava uvid u izravnu interakciju genoma s okolišem na razini strukturnih varijacija u reproduktivnim stanicama. Ovakav pristup, usmjeren na genetičke promjene u nasljednom materijalu koje su nastale pod utjecajem okolišnih čimbenika je stoga od neizmjerne važnosti za napredak istraživanja u području zdravlja i evolucije.

6. LITERATURA

Advani, A. S., & Pendergast, A. M. (2002). Bcr–Abl variants: biological and clinical aspects. Leukemia Research, 26(8), 713–720. https://doi.org/10.1016/s0145-2126(01)00197-7

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology, 12(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. Nature Reviews Genetics, 12(5), 363–376. https://doi.org/10.1038/nrg2958

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research, 21, 974–984

Aubert, G., Hills, M., & Lansdorp, P. M. (2012). Telomere length measurement—Caveats and a critical assessment of the available technologies and tools. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 730(1–2), 59–67. https://doi.org/10.1016/j.mrfmmm.2011.04.003

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Adewoye, A. B., Lindsay, S. J., Dubrova, Y. E., & Hurles, M. E. (2015). The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. Nature Communications, 6(1). https://doi.org/10.1038/ncomms7684

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell, 176(3), 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019

Abdennur, N., Fudenberg, G., Flyamer, I., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. (2022). Bioframe: Operations on Genomic Intervals in Pandas Dataframes. https://doi.org/10.1101/2022.02.16.480748

Blasco, M. A., Lee, H.-W., Hande, M. P., Samper, E., Lansdorp, P. M., DePinho, R. A., & Greider, C. W. (1997). Telomere Shortening and Tumor Formation by Mouse Cells Lacking Telomerase RNA. Cell, 91(1), 25–34. https://doi.org/10.1016/s0092-8674(01)80006-4

Bai, Y., & Murnane, J. P. (2003). Telomere instability in a human tumor cell line expressing a dominant-negative WRN protein. Human Genetics, 113(4), 337–347. https://doi.org/10.1007/s00439-003-0972-y

Balhorn, R. (2007). The protamine family of sperm nuclear proteins. Genome Biology, 8(9), 227. https://doi.org/10.1186/gb-2007-8-9-227

Bradic, M., Beerli, P., García-de León, F. J., Esquivel-Bobadilla, S., & Borowsky, R. L. (2012). Gene flow and population structure in the Mexican blind cavefish complex (Astyanax mexicanus). BMC Evolutionary Biology, 12(1). https://doi.org/10.1186/1471-2148-12-9

Bradic, M., Teotónio, H., & Borowsky, R. L. (2013). The Population Genomics of Repeated Evolution in the Blind Cavefish Astyanax mexicanus. Molecular Biology and Evolution, 30(11), 2383–2400. https://doi.org/10.1093/molbev/mst136

Bojesen, S. E., Pooley, K. A., Johnatty, S. E., Beesley, J., Michailidou, K., Tyrer, J. P., Edwards, S. L., Pickett, H. A., Shen, H. C., Smart, C. E., Hillman, K. M., … Dunning, A. M. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature Genetics, 45(4), 371–384. https://doi.org/10.1038/ng.2566

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brody, S., Dubuc, A. M., & Kim, A. S. (2023). Optical genome mapping: A tool with significant potential from discovery to diagnostics. College of American Pathologists. https://www.cap.org/member-resources/articles/optical-genome-mapping-a-tool-with-significant-potential-from-discovery-to-diagnostics

Brewer, B. J., Dunham, M. J., & Raghuraman, M. K. (2024). A unifying model that explains the origins of human inverted copy number variants. PLOS Genetics, 20(1), e1011091. https://doi.org/10.1371/journal.pgen.1011091

Cawthon, R. M. (2002). Telomere measurement by quantitative PCR. Nucleic Acids Research, 30(10), 47e–447. https://doi.org/10.1093/nar/30.10.e47

Cheung, J., Wilson, M. D., Zhang, J., Khaja, R., MacDonald, J. R., Heng, H. H., Koop, B. F., & Scherer, S. W. (2003). Recent segmental and gene duplications in the mouse genome. Genome Biology, 4(8). https://doi.org/10.1186/gb-2003-4-8-r47

Cawthon, R. M. (2009). Telomere length measurement by a novel monochrome multiplex quantitative PCR method. Nucleic Acids Research, 37(3), e21–e21. https://doi.org/10.1093/nar/gkn1027

Casellas, J. (2011). Inbred mouse strains and genetic stability: a review. Animal, 5(1), 1–7. https://doi.org/10.1017/s1751731110001667

Chain, F. J. J., Feulner, P. G. D., Panchal, M., Eizaguirre, C., Samonte, I. E., Kalbe, M., Lenz, T. L., Stoll, M., Bornberg-Bauer, E., Milinski, M., & Reusch, T. B. H. (2014). Extensive Copy-Number Variation of Young Genes across Stickleback Populations. PLoS Genetics, 10(12), e1004830. https://doi.org/10.1371/journal.pgen.1004830

de Castro Barbosa, T., Ingerslev, L. R., Alm, P. S., Versteyhe, S., Massart, J., Rasmussen, M., Donkin, I., Sjögren, R., Mudry, J. M., Vetterli, L., Gupta, S., Krook, A., Zierath, J. R., & Barrès, R. (2016). High-fat diet reprograms the epigenome of rat spermatozoa and transgenerationally affects metabolism of the offspring. Molecular Metabolism, 5(3), 184–197. https://doi.org/10.1016/j.molmet.2015.12.002

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics, 17(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. Nature, 581(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Chieffi Baccari, G., Iurato, G., Santillo, A., & Dale, B. (2023). Male Germ Cell Telomeres and Chemical Pollutants. Biomolecules, 13(5), 745. https://doi.org/10.3390/biom13050745

Choo, Z.-N., Behr, J. M., Deshpande, A., Hadi, K., Yao, X., Tian, H., Takai, K., Zakusilo, G., Rosiene, J., Da Cruz Paula, A., Weigelt, B., Setton, J., Riaz, N., Powell, S. N., Busam, K., Shoushtari, A. N., Ariyan, C., Reis-Filho, J., de Lange, T., & Imieliński, M. (2023). Most large structural variants in cancer genomes can be detected without long reads. Nature Genetics, 55(12), 2139–2148. https://doi.org/10.1038/s41588-023-01540-6

Dias, B. G., & Ressler, K. J. (2013). Parental olfactory experience influences behavior and neural structure in subsequent generations. Nature Neuroscience, 17(1), 89–96. https://doi.org/10.1038/nn.3594

De Coster, W., & Van Broeckhoven, C. (2019). Newest Methods for Detecting Structural Variations. Trends in Biotechnology, 37(9), 973–982. https://doi.org/10.1016/j.tibtech.2019.02.003

Dorant, Y., Cayuela, H., Wellband, K., Laporte, M., Rougemont, Q., Mérot, C., Normandeau, E., Rochette, R., & Bernatchez, L. (2020). Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. Molecular Ecology, 29(24), 4765–4782. Portico. https://doi.org/10.1111/mec.15565

Dremsek, P., Schwarz, T., Weil, B., Malashka, A., Laccone, F., & Neesen, J. (2021). Optical Genome Mapping in Routine Human Genetic Diagnostics—Its Advantages and Limitations. Genes, 12(12), 1958. https://doi.org/10.3390/genes12121958

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends in Genetics, 17(11), 661–669. https://doi.org/10.1016/s0168-9525(01)02492-1

Eichler, E. E., Clark, R. A., & She, X. (2004). An assessment of the sequence gaps: Unfinished business in a finished human genome. Nature Reviews Genetics, 5(5), 345–354. https://doi.org/10.1038/nrg1322

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. Briefings in Functional Genomics, 14(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren,

J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science, 372(6537). https://doi.org/10.1126/science.abf7117

Fouladi, B., Sabatier, L., Miller, D., Pottier, G., & Murnane, J. P. (2000). The Relationship Between Spontaneous Telomere Loss and Chromosome Instability in a Human Tumor Cell Line. Neoplasia, 2(6), 540–554. https://doi.org/10.1038/sj.neo.7900107

Ferreira, M. G., Miller, K. M., & Cooper, J. P. (2004). Indecent Exposure. Molecular Cell, 13(1), 7–18. https://doi.org/10.1016/s1097-2765(03)00531-8

Finegersh, A., & Homanics, G. E. (2014). Paternal Alcohol Exposure Reduces Alcohol Drinking and Increases Behavioral Sensitivity to Alcohol Selectively in Male Offspring. PLoS ONE, 9(6), e99078. https://doi.org/10.1371/journal.pone.0099078

de Frutos, C., López-Cardona, A. P., Fonseca Balvís, N., Laguna-Barraza, R., Rizos, D., Gutierrez-Adán, A., & Bermejo-Álvarez, P. (2016). Spermatozoa telomeres determine telomere length in early embryos and offspring. REPRODUCTION, 151(1), 1–7. https://doi.org/10.1530/rep-15-0375

Fice, H., & Robaire, B. (2019). Telomere Dynamics Throughout Spermatogenesis. Genes, 10(7), 525. https://doi.org/10.3390/genes10070525

Fan, H.-C., Chang, F.-W., Tsai, J.-D., Lin, K.-M., Chen, C.-M., Lin, S.-Z., Liu, C.-A., & Harn, H.-J. (2021). Telomeres and Cancer. Life, 11(12), 1405. https://doi.org/10.3390/life11121405

Grill, S., & Nandakumar, J. (2021). Molecular mechanisms of telomere biology disorders. Journal of Biological Chemistry, 296, 100064. https://doi.org/10.1074/jbc.rev120.014017

Garg, P., Martin-Trujillo, A., Rodriguez, O. L., Gies, S. J., Hadelia, E., Jadhav, B., Jain, M., Paten, B., & Sharp, A. J. (2021). Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. The American Journal of Human Genetics, 108(5), 809–824. https://doi.org/10.1016/j.ajhg.2021.03.016

Hemann, M. T. (2000). Wild-derived inbred mouse strains have short telomeres. Nucleic Acids Research, 28(22), 4474–4478. https://doi.org/10.1093/nar/28.22.4474

Hemann, M. T. (2000). Wild-derived inbred mouse strains have short telomeres. Nucleic Acids Research, 28(22), 4474–4478.

Haraksingh, R. R., & Snyder, M. P. (2013). Impacts of Variation in the Human Genome on Gene Regulation. Journal of Molecular Biology, 425(21), 3970–3977. https://doi.org/10.1016/j.jmb.2013.07.015

Hirase, S., Ozaki, H., & Iwasaki, W. (2014). Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes. BMC Genomics, 15(1), 735. https://doi.org/10.1186/1471-2164-15-735

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. Nature Genetics, 47(3), 296–303. https://doi.org/10.1038/ng.3200

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C.-S., Korlach, J., Wilson, R. K., & Eichler, E. E. (2016). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Research, 27(5), 677–685. https://doi.org/10.1101/gr.214007.116

Hull, R. M., Cruz, C., Jack, C. V., & Houseley, J. (2017). Environmental change drives accelerated adaptation through stimulated copy number variation. PLOS Biology, 15(6), e2001333. https://doi.org/10.1371/journal.pbio.2001333

Herman, A., Brandvain, Y., Weagley, J., Jeffery, W. R., Keene, A. C., Kono, T. J. Y., Bilandžija, H., Borowsky, R., Espinasa, L., O'Quin, K., Ornelas-García, C. P., Yoshizawa, M., Carlson, B., Maldonado, E., Gross, J. B., Cartwright, R. A., Rohner, N., Warren, W. C., & McGaugh, S. E. (2018). The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, Astyanax mexicanus. Molecular Ecology, 27(22), 4397–4416. Portico. https://doi.org/10.1111/mec.14877

Huang, Y., Feulner, P. G. D., Eizaguirre, C., Lenz, T. L., Bornberg-Bauer, E., Milinski, M., Reusch, T. B. H., & Chain, F. J. J. (2019). Genome-Wide Genotype-Expression Relationships Reveal Both Copy Number and Single Nucleotide Differentiation Contribute to Differential Gene Expression between Stickleback Ecotypes. Genome Biology and Evolution, 11(8), 2344–2359. https://doi.org/10.1093/gbe/evz148

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., dePamphilis, C. W., & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of

Theobroma cacao, the chocolate tree. Proceedings of the National Academy of Sciences, 118(35). https://doi.org/10.1073/pnas.2102914118

Hallast, P., Ebert, P., Loftus, M., Yilmaz, F., Audano, P. A., Logsdon, G. A., Bonder, M. J., Zhou, W., Höps, W., Kim, K., Li, C., Hoyt, S. J., Dishuck, P. C., Porubsky, D., Tsetsos, F., Kwon, J. Y., Zhu, Q., Munson, K. M., Hasenfeld, P., … Lee, C. (2023). Assembly of 43 human Y chromosomes reveals extensive complexity and variation. Nature, 621(7978), 355–364. https://doi.org/10.1038/s41586-023-06425-6

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., & Eichler, E. E. (2010). De novo rates and selection of large copy number variation. Genome Research, 20(11), 1469–1481. https://doi.org/10.1101/gr.107680.110

Iskow, R. C., Gokcumen, O., & Lee, C. (2012). Exploring the role of copy number variants in human adaptation. Trends in Genetics, 28(6), 245–257. https://doi.org/10.1016/j.tig.2012.03.002

Ishikawa, A., Kabeya, N., Ikeya, K., Kakioka, R., Cech, J. N., Osada, N., Leal, M. C., Inoue, J., Kume, M., Toyoda, A., Tezuka, A., Nagano, A. J., Yamasaki, Y. Y., Suzuki, Y., Kokita, T., Takahashi, H., Lucek, K., Marques, D., Takehana, Y., … Kitano, J. (2019). A key metabolic gene for recurrent freshwater colonization and radiation in fishes. Science, 364(6443), 886–889. https://doi.org/10.1126/science.aau5656

Ishikawa, A., Yamanouchi, S., Iwasaki, W., & Kitano, J. (2022). Convergent copy number increase of genes associated with freshwater colonization in fishes. Philosophical Transactions of the Royal Society B: Biological Sciences, 377(1855). https://doi.org/10.1098/rstb.2020.0509

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. Nature, 484(7392), 55–61. https://doi.org/10.1038/nature10944

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology, 36(4), 338–345. https://doi.org/10.1038/nbt.4060

Jang, J., Terefe, E., Kim, K., Lee, Y. H., Belay, G., Tijjani, A., Han, J., Hanotte, O., & Kim, H. (2021). Population differentiated copy number variation of Bos taurus, Bos indicus and their African hybrids. BMC Genomics, 22(1). https://doi.org/10.1186/s12864-021-07808-7

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. Genome Biology, 3(2). https://doi.org/10.1186/gb-2002-3-2-research0008

Köster, J., & Rahmann, S. (2018). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics, 34(20), 3600–3600. https://doi.org/10.1093/bioinformatics/bty350

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1720-5

Kowalko, J. (2020). Utilizing the blind cavefish Astyanax mexicanus to understand the genetic basis of behavioral evolution. Journal of Experimental Biology, 223(Suppl_1). https://doi.org/10.1242/jeb.208835

Lansdorp, P. M., Verwoerd, N. P., van de Rijke, F. M., Dragowska, V., Little, M. T., Dirks, R. W., Raap, A. K., & Tanke, H. J. (1996). Heterogeneity in telomere length of human chromosomes. Human Molecular Genetics, 5(5), 685–691. https://doi.org/10.1093/hmg/5.5.685

Lo, A. W. l., Sabatier, L., Fouladi, B., Pottier, G., Ricoul, M., & Mumane, J. P. (2002). DNA Amplification by Breakage/Fusion/Bridge Cycles Initiated by Spontaneous Telomere Loss in a Human Cancer Cell Line. Neoplasia, 4(6), 531–538. https://doi.org/10.1038/sj.neo.7900267

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lin, T. T., Letsolo, B. T., Jones, R. E., Rowson, J., Pratt, G., Hewamana, S., Fegan, C., Pepper, C., & Baird, D. M. (2010). Telomere dysfunction and fusion during the progression of chronic lymphocytic leukemia: evidence for a telomere crisis. Blood, 116(11), 1899–1907. https://doi.org/10.1182/blood-2010-02-272104

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., & Kwok, P.-Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology, 30(8), 771–776. https://doi.org/10.1038/nbt.2303

Lowe, C. B., Sanchez-Luege, N., Howes, T. R., Brady, S. D., Daugherty, R. R., Jones, F. C., Bell, M. A., & Kingsley, D. M. (2018). Detecting differential copy number variation between groups of samples. Genome research, 28(2), 256–265. https://doi.org/10.1101/gr.206938.116

Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., Cao, H.-F., Zhang, B.-W., Yeh, C.-F., Hung, C.-M., Hung, H.-Y., Yang, M.-Y., Liang, W., Hsu, Y.-C., Yao, C.-T., Dong, L., Lin, K., & Li, S.-H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. Proceedings of the National Academy of Sciences, 116(6), 2152–2157. https://doi.org/10.1073/pnas.1813597116

Legendre, L., Rode, J., Germon, I., Pavie, M., Quiviger, C., Policarpo, M., Leclercq, J., Père, S., Fumey, J., Hyacinthe, C., Ornelas-García, P., Espinasa, L., Rétaux, S., … Casane, D. (2023). Genetic identification and reiterated captures suggest that the <i>Astyanax mexicanus</i> El Pachón cavefish population is closed and declining. Zoological Research, 44(4), 701–711. https://doi.org/10.24272/j.issn.2095-8137.2022.481

Mirabello, L., Yu, K., Kraft, P., De Vivo, I., Hunter, D. J., Prescott, J., Wong, J. Y. Y., Chatterjee, N., Hayes, R. B., & Savage, S. A. (2010). The association of telomere length and genetic variation in telomere biology genesa. Human Mutation, 31(9), 1050–1058. https://doi.org/10.1002/humu.21314

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10. https://doi.org/10.14806/ej.17.1.200

Marchetti, F., Rowan-Carroll, A., Williams, A., Polyzos, A., Berndt-Weis, M. L., & Yauk, C. L. (2011). Sidestream tobacco smoke is a male germ cell mutagen. Proceedings of the National Academy of Sciences, 108(31), 12811–12814. https://doi.org/10.1073/pnas.1106896108

Montpetit, A. J., Alhareeri, A. A., Montpetit, M., Starkweather, A. R., Elmore, L. W., Filler, K., Mohanraj, L., Burton, C. W., Menzies, V. S., Lyon, D. E., & Jackson-Cook, C. K. (2014). Telomere Length. Nursing Research, 63(4), 289–299. https://doi.org/10.1097/nnr.0000000000000037

Morgan, A. P., & Pardo-Manuel de Villena, F. (2017). Sequence and Structural Diversity of Mouse Y Chromosomes. Molecular Biology and Evolution, 34(12), 3186–3204. https://doi.org/10.1093/molbev/msx250

McCaffrey, J., Young, E., Lassahn, K., Sibert, J., Pastor, S., Riethman, H., & Xiao, M. (2017). High-throughput single-molecule telomere characterization. Genome Research, 27(11), 1904–1915. https://doi.org/10.1101/gr.222422.117

Margalef, P., Kotsantis, P., Borel, V., Bellelli, R., Panier, S., & Boulton, S. J. (2018). Stabilization of Reversed Replication Forks by Telomerase Drives Telomere Catastrophe. Cell, 172(3), 439-453.e14. https://doi.org/10.1016/j.cell.2017.11.047

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1828-7

Oakberg, E. F. (1957). Duration of Spermatogenesis in the Mouse. Nature, 180(4595), 1137–1138. https://doi.org/10.1038/1801137a0

Prowse KR, Greider CW (1995) Developmental and tissue-specific regulation of mouse telomerase and telomere length. Proc. Natl Acad. Sci. USA 92, 4818–4822

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. Nature Genetics, 39(10), 1256–1260. https://doi.org/10.1038/ng2123

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2013). A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics, 15(2), 256–278. https://doi.org/10.1093/bib/bbs086

Pezer, Ž., Harr, B., Teschke, M., Babiker, H., & Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. Genome Research, 25(8), 1114–1124. https://doi.org/10.1101/gr.187187.114

Pokrovac, I., & Pezer, Ž. (2022). Recent advances and current challenges in population genomics of structural variation in animals and plants. Frontiers in Genetics, 13. https://doi.org/10.3389/fgene.2022.1060898

Pokrovac, I., Rohner, N., & Pezer, Ž. (2024). The prevalence of copy number increase at multiallelic copy number variants associated with cave colonization. Molecular Ecology, 33(9). Portico. https://doi.org/10.1111/mec.17339

Rompala, G. R., Mounier, A., Wolfe, C. M., Lin, Q., Lefterov, I., & Homanics, G. E. (2018). Heavy Chronic Intermittent Ethanol Exposure Alters Small Noncoding RNAs in Mouse Sperm and Epididymosomes. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00032

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Research, 47(W1), W191–W198. https://doi.org/10.1093/nar/gkz369

Raseley, K., Jinwala, Z., Zhang, D., & Xiao, M. (2023). Single-Molecule Telomere Assay via Optical Mapping (SMTA-OM) Can Potentially Define the ALT Positivity of Cancer. Genes, 14(6), 1278. https://doi.org/10.3390/genes14061278

Somers, C. M., Yauk, C. L., White, P. A., Parfett, C. L. J., & Quinn, J. S. (2002). Air pollution induces heritable DNA mutations. Proceedings of the National Academy of Sciences, 99(25), 15904–15907. https://doi.org/10.1073/pnas.252499499

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. Trends in Genetics, 18(2), 74–82. https://doi.org/10.1016/s0168-9525(02)02592-1

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., & Eichler, E. E. (2005). Segmental Duplications and Copy-Number Variation in the Human Genome. The American Journal of Human Genetics, 77(1), 78–88. https://doi.org/10.1086/431652

Soerensen, M., Thinggaard, M., Nygaard, M., Dato, S., Tan, Q., Hjelmborg, J., Andersen-Ranberg, K., Stevnsner, T., Bohr, V. A., Kimura, M., Aviv, A., Christensen, K., & Christiansen, L. (2012). Genetic variation in TERT and TERC and human leukocyte telomere length and longevity: a cross-sectional and longitudinal analysis. Aging Cell, 11(2), 223–227. Portico. https://doi.org/10.1111/j.1474-9726.2011.00775.x

Soh, Y. Q. S., Alföldi, J., Pyntikova, T., Brown, L. G., Graves, T., Minx, P. J., Fulton, R. S., Kremitzki, C., Koutseva, N., Mueller, J. L., Rozen, S., Hughes, J. F., Owens, E., Womack, J. E., Murphy, W. J., Cao, Q., de Jong, P., Warren, W. C., Wilson, R. K., … Page, D. C. (2014). Sequencing the Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex Chromosomes. Cell, 159(4), 800–813. https://doi.org/10.1016/j.cell.2014.09.052

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J.,

Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), 75–81. https://doi.org/10.1038/nature15394

Suvakov, M., Panda, A., Diesh, C., Holmes, I., & Abyzov, A. (2021). CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. GigaScience, 10(11). https://doi.org/10.1093/gigascience/giab074

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research, 50(W1), W216–W221. https://doi.org/10.1093/nar/gkac194

Schmidt, T. T., Tyer, C., Rughani, P., Haggblom, C., Jones, J. R., Dai, X., Frazer, K. A., Gage, F. H., Juul, S., Hickey, S., & Karlseder, J. (2024). High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. Nature Communications, 15(1). https://doi.org/10.1038/s41467-024-48917-7

Tahamtan, S., Tavalaee, M., Izadi, T., Barikrow, N., Zakeri, Z., Lockshin, R. A., Abbasi, H., & Nasr- Esfahani, M. H. (2019). Reduced sperm telomere length in individuals with varicocele is associated with reduced genomic integrity. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-40707-2

Venkatesan, R. N., & Price, C. (1998). Telomerase expression in chickens: Constitutive activity in somatic tissues and down-regulation in culture. Proceedings of the National Academy of Sciences, 95(25), 14763–14768. https://doi.org/10.1073/pnas.95.25.14763

Vannier, J.-B., Pavicic-Kaltenbrunner, V., Petalcorin, M. I. R., Ding, H., & Boulton, S. J. (2012). RTEL1 Dismantles T Loops and Counteracts Telomeric G4-DNA to Maintain Telomere Integrity. Cell, 149(4), 795–806. https://doi.org/10.1016/j.cell.2012.03.030

Wevrick, R., & Willard, H. F. (1989). Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. Proceedings of the National Academy of Sciences, 86(23), 9394–9398. https://doi.org/10.1073/pnas.86.23.9394

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics, 14(2), 125–138. https://doi.org/10.1038/nrg3373

Waitkus, M. S., Erman, E. N., Reitman, Z. J., & Ashley, D. M. (2024). Mechanisms of telomere maintenance and associated therapeutic vulnerabilities in malignant gliomas. Neuro-Oncology, 26(6), 1012–1024. https://doi.org/10.1093/neuonc/noae016

Xu, L., Hou, Y., Bickhart, D. M., Zhou, Y., Hay, E. H. abdel, Song, J., Sonstegard, T. S., Van Tassell, C. P., & Liu, G. E. (2016). Population-genetic properties of differentiated copy number variations in cattle. Scientific Reports, 6(1). https://doi.org/10.1038/srep23161

Yan, K., Gao, L., Cui, Y., Zhang, Y., & Zhou, X. (2016). The cyclic AMP signaling pathway: Exploring targets for successful drug discovery (Review). Molecular Medicine Reports, 13, 3715-3723. https://doi.org/10.3892/mmr.2016.5005

Young, E., Pastor, S., Rajagopalan, R., McCaffrey, J., Sibert, J., Mak, A. C. Y., Kwok, P.-Y., Riethman, H., & Xiao, M. (2017). High-throughput single-molecule mapping links subtelomeric variants and long-range haplotypes with specific telomeres. Nucleic Acids Research, 45(9), e73–e73. https://doi.org/10.1093/nar/gkx017

Yuste-Lisbona, F. J., Fernández-Lozano, A., Pineda, B., Bretones, S., Ortíz-Atienza, A., García-Sogo, B., Müller, N. A., Angosto, T., Capel, J., Moreno, V., Jiménez-Gómez, J. M., & Lozano, R. (2020). ENOregulates tomato fruit size through the floral meristem development network. Proceedings of the National Academy of Sciences, 117(14), 8187–8195. https://doi.org/10.1073/pnas.1913688117

Yuan, Y., Chung, C. Y.-L., & Chan, T.-F. (2020). Advances in optical mapping for genomic research. Computational and Structural Biotechnology Journal, 18, 2051–2062. https://doi.org/10.1016/j.csbj.2020.07.018

Yang, L., Han, J., Deng, T., Li, F., Han, X., Xia, H., Quan, F., Hua, G., Yang, L., & Zhou, Y. (2023). Comparative analyses of copy number variations between swamp buffaloes and river buffaloes. Animal Genetics, 54(2), 199–206. Portico. https://doi.org/10.1111/age.13288

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics, 10(1), 451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

Zong, S.-B., Li, Y.-L., & Liu, J.-X. (2020). Genomic Architecture of Rapid Parallel Adaptation to Fresh Water in a Wild Fish. Molecular Biology and Evolution, 38(4), 1317–1329. https://doi.org/10.1093/molbev/msaa290

Zhu, C., Li, M., Qin, S., Zhao, F., & Fang, S. (2020). Detection of copy number variation and selection signatures on the X chromosome in Chinese indigenous sheep with different types of tail. Asian-Australasian Journal of Animal Sciences, 33(9), 1378–1386. https://doi.org/10.5713/ajas.18.0661

7. SAŽETAK

Strukturne varijante su promjene u linearnoj strukturi genoma. Mogu značajno utjecati na gensku funkciju i ekspresiju, tako pridonoseći genetskoj raznolikosti, ali i razvoju bolesti. Unatoč značajnom utjecaju na fenotip i adaptaciju, proučavanje strukturnih varijanti otežano je zbog njihove heterogenosti i ograničene primjenjivosti standardnih evolucijskih metoda temeljenih na varijantama jednog nukleotida, a izravan utjecaj okoliša na varijante uglavnom se istražuje posredno putem komparativne genomike i obiteljskih studija, koje su ograničene na fiksirane varijante ili maladaptivne SVs, uz nedostatak informacija o pojedinačnim alelima.

U sklopu ove disertacije istražujemo utjecaj okolišnih čimbenika na strukturne varijante genoma kroz dvije perspektive: unutar jedne generacije laboratorijskog miša i na mikroevolucijskoj razini u prirodnim populacijama špiljskog i površinskog ekotipa ribe meksičke tetre (Astyanax mexicanus).

Prvi dio disertacije proveden je na laboratorijskom soju miševa C57BL/6, izloženih prehrani bogatoj mastima, kao primjeru stresora koji uzrokuje značajnu promjenu fenotipa. U ovom smo djelu prvi puta primijenili optičko mapiranje genoma na spermijima, zbog čega smo razvili protokol za izolaciju visokomolekularne DNA iz spermija za potrebe optičkog mapiranja. Rezultati optičkog mapiranja spermija udovoljavali su svim kriterijima kvalitete, te su po ovim kriterijima bili usporedivi s rezultatima dobivenim iz somatskog tkiva. Među promatranim miševima otkriveno je da oko 5 % genoma podliježe nekom obliku strukturne varijacije, što je iznenađujuće s obzirom na očekivanu izogeničnost ovog soja i također predstavlja genetsko udaljavanje kolonije na IRB-u od referentnog soja. Pronašli smo da prehrana bogata mastima povećava broj i duljinu delecija i duplikacija u spermijima pokusne skupine. Iako razlika u *de novo* varijantama između tkiva nije uočena, pronađeno je da je brojčani udio heterozigotnih varijanti značajno veći u spermijima pokusne skupine. Primjenom alata TelOMpy, razvijenog u sklopu disertacije, pronašli smo da prehrana bogata mastima uzrokuje smanjenje ukupnog broja telomera u oba analizirana tkiva Naše istraživanje po prvi put sugerira potencijalnu povezanost prehrane i gubitka čitavih telomera.

U drugom smo dijelu disertacije koristili meksičku tetru (Astyanax mexicanus) kao model prirodnog evolucijskog eksperimenta. Meksička tetra, s površinskim i špiljskim ekotipovima, idealan je model za istraživanje uloge strukturnih varijanti u paralelnoj evoluciji zbog dobro poznate povijesti populacija i prilagodbi na ekstremne okolišne uvjete poput tame i nedostatka nutrijenata. Identificirali smo 292 genomske regije s divergentnim brojem kopija između fenotipova, što sugerira selekcijske pritiske u procesu prilagodbe na špiljske uvjete. Rezultati su ukazivali na nižu genetičku raznolikost špiljskih populacija zbog ograničene veličine populacije i potencijalnih genetskih uskih grla, ali i na značajnu divergenciju varijanta broja kopija između špiljskih i površinskih populacija koje su povezane sa prilagodbom na špiljske uvjete. Povećanje, a ne smanjenje, broja kopija gena predstavlja dominantan mehanizam prilagodbe, s mogućim utjecajem na povećanje proteinske doze i brzu ekološku specijalizaciju.

Primjenom ove dvojne perspektive ustanovili smo da okolišni čimbenici ne uzrokuju de novo događaje, već pojačavaju učestalost varijacije na mjestima gdje su one već prisutne. Izravan utjecaj okoliša na strukturne varijacije nije usmjeren na određene regije i biološke funkcije, a istraživanje na razini populacije pokazuje da je evolucijski pritisak onaj koji odabire varijante povezane sa specifičnim funkcijama ovisno o njihovoj koristi ili šteti u danim okolišnim uvjetima.

8. SUMMARY

Structural variants (SVs) are changes in the linear structure of the genome. They can significantly impact gene function and expression, contributing to genetic diversity and the development of diseases. Despite their profound influence on phenotype and adaptation, the study of SVs is hindered by their heterogeneity and the limited applicability of standard evolutionary methods based on single nucleotide variants. The direct environmental impact on SVs is mostly explored indirectly through comparative genomics and family studies, which are restricted to fixed variants or maladaptive SVs and often lack information on individual alleles.

This dissertation explores the influence of environmental factors on structural genome variants from two perspectives: within a single generation of laboratory mice and at the microevolutionary level in natural populations.

The first part of the dissertation was conducted on the laboratory mouse strain C57BL/6, exposed to a high-fat diet as an example of a stressor causing significant phenotypic changes. In this research, we were the first to apply optical genome mapping to spermatozoa and in order to do so we developed a protocol for isolating high-molecular-weight DNA from sperm cells for optical mapping purposes. The resulting sperm genome optical maps met all quality criteria and were comparable to those obtained from somatic tissue. Among the examined mice, about 5 % of the genome exhibited some form of variation, which is surprising given the expected isogeneity of this strain and may represent genetic divergence of the used mouse colony from the reference strain. We found that a high-fat diet increased the number and length of deletions and duplications in the sperm of the experimental group. Although no difference in de novo variants between tissues was observed, the numerical proportion of heterozygous variants was significantly higher in the sperm of the experimental group. Using the TelOMpy tool developed as part of the dissertation, we found that a high-fat diet causes a reduction in the total number of telomeres in both analyzed tissues. Our research suggests a potential link between diet and the loss of entire telomeres.

In the second part of the dissertation, we used the Mexican tetra (Astyanax mexicanus) as a model of a natural evolutionary experiment. The Mexican tetra, with surface and cave ecotypes, is an ideal model for studying the role of SVs in parallel evolution due to the well-documented history of its populations and adaptations to extreme environmental conditions such as darkness and nutrient scarcity. We identified 292 genomic regions with divergent copy numbers between ecotypes, suggesting selective pressures in the adaptation process to cave environments. The results indicated lower genetic diversity in cave populations due to limited population size and potential genetic bottlenecks, but also a significant divergence of copy number variants between cave and surface populations associated with adaptation to cave conditions. An increase, rather than a decrease, in gene copy numbers emerged as the dominant adaptation mechanism, potentially impacting protein dosage and rapid ecological specialization.Through this dual perspective, we established that environmental factors do not cause de novo events but rather increase the frequency of variation at loci with pre-existing genetic variation. The direct environmental impact on structural variations is not targeted to specific regions or biological functions, while population-level research shows that evolutionary pressure selects from standing genetic variation at loci associated with specific functions depending on their benefits or detriments under given environmental conditions.

9. PRILOZI

9.1. Tablice

9.1.1. Statističko testiranje broja SVs, ukupno i po tipu SVs

Tablica S1. Rezultati sparenog T testa broja SVs između spermija i bubrega unutar skupine. Prikazane su p vrijednosti prije korekcije i poslije korekcije Holmovom metodom. U tablici je prikazan minimalan broj uzoraka potreban za α 0.05, te snagu testa i veličinu efekta prikazanu u istom retku.

	Tretman	P vrijednost	P vrijednost (Holm)	Cohenov D	Snaga testa	Broj uzoraka	Minimalni broj uzoraka
	WD	0,045	0,089	1,560	0,885	5	8
	C	0,130	0,130	1,270	0,618	4	6
SV							
inverzija	**WD**	**0,003**	**0,032**	**2,280**	**0,993**	**5**	**8**
inverzija	C	0,007	0,064	4,630	1,000	4	5
intrakromosomska translokacija	WD	0,015	0,121	1,430	0,833	5	8
delecija	WD	0,017	0,121	1,580	0,891	5	8
duplikacija	WD	0,033	0,198	1,856	0,956	5	8
duplikacija	C	0,074	0,368	1,209	0,582	4	6
insercija	C	0,084	0,368	1,792	0,851	4	6
delecija	C	0,172	0,516	1,019	0,472	4	6
insercija	WD	0,276	0,553	0,442	0,207	5	8
intrakromosomska translokacija	C	0,828	0,828	0,885	0,001	4	10

9.1.2. Statistike pokrivenosti genoma i veličine segmenta za detekciju CNVs

Tablica S2. Statistike pokrivenosti (medijan, interkvartilni raspon, prosjek, standardna devijacija po bazi) korištenih genoma te veličina binova korištenih za detekciju.

uzorak	Median pokrivenosti	IQR pokrivenosti	Prosjek pokrivenosti	Standardna devijacija pokrivenosti	Veličina bina za detekciju CNV /bp
Choy01	9	5	8,83	5,61	800
Choy05	10	5	9,68	5,58	800
Choy06	8	6	8,17	6,63	800
Choy09	9	5	8,92	5,95	800
Choy10	8	5	8,65	4,99	800
Choy11	9	6	9,10	7,50	800
Choy12	9	6	9,15	5,68	800
Choy13	10	6	9,82	6,01	700
Choy14	8	5	8,65	5,29	800
Molino10b	10	7	10,06	11,40	700
Molino11a	8	5	8,06	9,39	800
Molino12a	8	6	8,99	10,85	800
Molino13b	11	6	11,78	14,49	600
Molino14a	7	5	7,46	9,68	800
Molino15b	7	6	7,58	8,68	800
Molino2a	9	6	9,82	12,94	700
Molino7a	9	6	9,41	10,67	800
Molino9b	9	5	9,02	10,27	800
Pach11	7	5	7,71	8,68	800
Pach12	9	6	9,35	10,48	800
Pach14	7	5	7,71	8,80	800
Pach15	9	6	9,39	10,77	800
Pach17	7	5	7,69	8,54	800
Pach3	11	7	11,01	12,34	600
Pach7	8	6	8,31	9,13	800
Pach8	9	5	8,95	10,40	800
Pach9	8	6	8,42	9,96	800
PachonRef	17	15	17,69	16,23	800
Rascon02	12	8	12,20	10,79	500
Rascon04	11	7	11,53	10,66	600
Rascon13	9	6	9,32	8,27	800
Rascon15	9	6	9,61	9,77	700

Rascon6	9	6	9,08	8,19	800
Rascon8	8	5	7,98	7,84	800
Tinaja11	10	7	10,08	12,12	700
Tinaja12	11	7	11,16	14,00	600
Tinaja2	8	5	7,97	9,33	800
Tinaja3	8	6	8,90	10,25	800
Tinaja5	6	5	6,64	7,83	800
Tinaja6	8	5	8,89	9,74	800
TinajaB	10	6	10,65	12,61	600
TinajaC	10	7	10,00	12,21	700
TinajaD	7	5	7,95	9,49	800
TinajaE	9	6	9,60	10,65	700

9.1.3. Geni koji preklapaju CNVs isključivo špiljskih genoma

Tablica S2. Svi protein-kodirajući geni koji preklapanja CNVs pronađene isključivo u špiljskim genomima. Od 1407 ukupnih gena, prikazano je samo 157 gena koji preklapaju CNV iz barem 9 (minimalno 30 %) različitih špiljskih genoma.

Genski simbol (NCBI)	Genski ID (NCBI)	Ime gena (NCBI)	špilja	Molino	Pachon	Tinaja
LOC125785663	125785663	NLR family CARD domain-containing protein 3-like	28/29	9/9	10/10	9/10
rgrb	103043895	retinal G protein coupled receptor b	27/29	9/9	9/10	9/10
cd34	103036525	CD34 molecule	26/29	8/9	9/10	9/10
LOC125806175	125806175	zona pellucida sperm-binding protein 3-like	20/29	5/9	9/10	6/10
cds2	103035110	CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2	20/29	5/9	6/10	9/10
plppr3a	103047787	phospholipid phosphatase related 3a	20/29	3/9	9/10	8/10
atp2b2	103025027	ATPase plasma membrane Ca2+ transporting 2	20/29	0/9	10/10	10/10
pikfyve	103023917	phosphoinositide kinase, FYVE finger containing	19/29	1/9	8/10	10/10
apaf1	103032437	apoptotic peptidase activating factor 1	19/29	0/9	9/10	10/10
LOC125799514	125799514	perforin-1-like	19/29	0/9	9/10	10/10
LOC111194948	111194948	TOG array regulator of axonemal microtubules protein 1	19/29	1/9	8/10	10/10
gpr83	111196039	G protein-coupled receptor 83	19/29	0/9	10/10	9/10
LOC111195514	111195514	perforin-1-like	18/29	0/9	9/10	9/10
tgfbr3	103043240	transforming growth factor, beta receptor III	18/29	0/9	9/10	9/10
LOC103033205	103033205	SH3 and multiple ankyrin repeat domains protein 2	18/29	0/9	9/10	9/10
pde10a	103038843	phosphodiesterase 10A	18/29	0/9	8/10	10/10
opn8b	111197225	opsin 8, group member b	17/29	0/9	7/10	10/10
map7d1a	103035851	MAP7 domain containing 1a	17/29	0/9	9/10	8/10
galnt7	103041790	UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase 7	17/29	9/9	0/10	8/10

kcnc3a	103030038	potassium voltage-gated channel, Shaw-related subfamily, member 3a	17/29	0/9	8/10	9/10
LOC103031898	103031898	deleted in malignant brain tumors 1 protein	16/29	4/9	8/10	4/10
map2	103026463	microtubule-associated protein 2	16/29	9/9	4/10	3/10
dchs1b	103027380	dachsous cadherin-related 1b	16/29	3/9	10/10	3/10
kics2	103023944	KICSTOR subunit 2	16/29	0/9	8/10	8/10
LOC111191630	111191630	hemoglobin subunit beta-2-like	16/29	7/9	7/10	2/10
LOC103046298	103046298	nuclear factor 7, brain-like	16/29	0/9	10/10	6/10
opn8a	111191907	opsin 8, group member a	16/29	1/9	6/10	9/10
LOC103030484	103030484	ras-related protein Rab-26	16/29	9/9	4/10	3/10
trim108	103038292	tripartite motif containing 108	16/29	8/9	4/10	4/10
LOC111190875	111190875	uncharacterized LOC111190875	15/29	7/9	8/10	0/10
rab3c	103035406	RAB3C, member RAS oncogene family	15/29	7/9	4/10	4/10
whrnb	103021854	whirlin b	15/29	8/9	0/10	7/10
sulf1	103047505	sulfatase 1	15/29	0/9	8/10	7/10
LOC103031348	103031348	protein NLRC3-like	15/29	9/9	0/10	6/10
tmem132e	103046758	transmembrane protein 132E	15/29	0/9	10/10	5/10
fbxw7	103046885	F-box and WD repeat domain containing 7	14/29	0/9	7/10	7/10
LOC103030318	103030318	putative gustatory receptor clone PTE03	14/29	6/9	0/10	8/10
abcg2b	103038929	ATP-binding cassette, sub-family G (WHITE), member 2b	14/29	0/9	7/10	7/10
LOC111191063	111191063	cytochrome P450 2K1	14/29	0/9	9/10	5/10
LOC111191317	111191317	gamma-crystallin M2-like	13/29	9/9	0/10	4/10
LOC103033534	103033534	gamma-crystallin M2	13/29	9/9	0/10	4/10
LOC103037059	103037059	probable G-protein coupled receptor 153	13/29	0/9	9/10	4/10
LOC111196508	111196508	scavenger receptor cysteine-rich type 1 protein M130	13/29	2/9	8/10	3/10
LOC103045991	103045991	uncharacterized LOC103045991	13/29	0/9	8/10	5/10
LOC111189811	111189811	gamma-crystallin M2-like	13/29	9/9	0/10	4/10
LOC111191631	111191631	hemoglobin embryonic subunit alpha	13/29	7/9	5/10	1/10
LOC103041875	103041875	gamma-crystallin M2-like	13/29	9/9	0/10	4/10

wt1a	103045853	WT1 transcription factor a	13/29	0/9	7/10	6/10
LOC125803842	125803842	uncharacterized LOC125803842	13/29	0/9	9/10	4/10
clasp1a	103026413	cytoplasmic linker associated protein 1a	13/29	0/9	7/10	6/10
LOC111191312	111191312	gamma-crystallin M2-like	13/29	9/9	0/10	4/10
LOC107197208	107197208	complement C1q-like protein 3	12/29	5/9	3/10	4/10
LOC125801092	125801092	myelin basic protein-like	12/29	7/9	0/10	5/10
sult6b1	103031948	sulfotransferase family, cytosolic, 6b, member 1	12/29	9/9	1/10	2/10
cadm1a	103030031	cell adhesion molecule 1a	12/29	0/9	5/10	7/10
trabd2b	103041038	TraB domain containing 2B	12/29	0/9	8/10	4/10
ryr1a	103047751	ryanodine receptor 1a (skeletal)	12/29	0/9	6/10	6/10
cuedc1a	103041127	CUE domain containing 1a	12/29	0/9	6/10	6/10
LOC103033609	103033609	uncharacterized LOC103033609	11/29	0/9	3/10	8/10
chrm2a	103039136	cholinergic receptor, muscarinic 2a	11/29	0/9	4/10	7/10
wfs1b	103038069	Wolfram syndrome 1b (wolframin)	11/29	0/9	3/10	8/10
carhsp1	111193795	calcium regulated heat stable protein 1	11/29	0/9	6/10	5/10
LOC111191628	111191628	hemoglobin embryonic subunit alpha	11/29	1/9	6/10	4/10
kpnb3	103035615	karyopherin (importin) beta 3	11/29	0/9	1/10	10/10
myo1d	103046626	myosin 1D	11/29	0/9	4/10	7/10
si:dkey-174m14.3	103027838	uncharacterized protein LOC563117 homolog	11/29	1/9	4/10	6/10
foxe1	103021819	forkhead box E1	11/29	0/9	2/10	9/10
LOC125780624	125780624	histone H2A-like	10/29	0/9	8/10	2/10
LOC111193082	111193082	protocadherin alpha-3-like	10/29	0/9	0/10	10/10
pth2ra	103034973	parathyroid hormone 2 receptor a	10/29	9/9	0/10	1/10
LOC111188524	111188524	G2/M phase-specific E3 ubiquitin-protein ligase-like	10/29	0/9	3/10	7/10
LOC103027764	103027764	hemoglobin subunit beta-2	10/29	0/9	6/10	4/10
dnah6	111195267	dynein, axonemal, heavy chain 6	10/29	0/9	3/10	7/10
nfatc2a	103038616	nuclear factor of activated T cells 2a	10/29	6/9	4/10	0/10
LOC103030412	103030412	GTPase IMAP family member 4-like	10/29	1/9	1/10	8/10

lrriq1	103032138	leucine-rich repeats and IQ motif containing 1	10/29	0/9	10/10	0/10
grm3	103035018	glutamate receptor, metabotropic 3	10/29	0/9	6/10	4/10
LOC103027414	103027414	gastricsin	10/29	7/9	3/10	0/10
LOC107197522	107197522	proline-rich protein 5	10/29	0/9	3/10	7/10
znf710b	103040070	zinc finger protein 710b	10/29	8/9	0/10	2/10
LOC125784919	125784919	trace amine-associated receptor 13c-like	10/29	0/9	9/10	1/10
LOC125799049	125799049	protocadherin alpha-8-like	10/29	0/9	0/10	10/10
slc6a17	103043338	solute carrier family 6 member 17	10/29	0/9	0/10	10/10
LOC103043615	103043615	cytosolic sulfotransferase 3	10/29	0/9	0/10	10/10
LOC103031223	103031223	uncharacterized LOC103031223	10/29	1/9	9/10	0/10
LOC103044416	103044416	cytosolic sulfotransferase 3	10/29	0/9	0/10	10/10
papss1	103043441	3'-phosphoadenosine 5'-phosphosulfate synthase 1	10/29	0/9	6/10	4/10
si:ch73-362m14.4	103027289	wiskott-Aldrich syndrome protein family member 3	10/29	0/9	6/10	4/10
ednraa	103047095	endothelin receptor type Aa	10/29	0/9	1/10	9/10
LOC111196910	111196910	NLR family CARD domain-containing protein 3	10/29	0/9	10/10	0/10
LOC107197546	107197546	C-type mannose receptor 2-like	10/29	0/9	10/10	0/10
rsad2	103041061	radical S-adenosyl methionine domain containing 2	10/29	0/9	6/10	4/10
LOC103039501	103039501	putative C-type lectin domain family 20 member A	10/29	0/9	0/10	10/10
cmpk2	103046469	cytidine monophosphate (UMP-CMP) kinase 2, mitochondrial	10/29	0/9	6/10	4/10
LOC103026293	103026293	NLR family CARD domain-containing protein 3-like	10/29	5/9	3/10	2/10
st6galnac2	103031766	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 2	10/29	4/9	1/10	5/10
pitpnc1a	111193245	phosphatidylinositol transfer protein cytoplasmic 1a	10/29	0/9	9/10	1/10
hsph1	103040797	heat shock 105/110 protein 1	10/29	0/9	5/10	5/10
LOC103030249	103030249	solute carrier family 45 member 3	9/29	9/9	0/10	0/10

LOC111188268	111188268	tripartite motif-containing protein 16	9/29	2/9	3/10	4/10
gse1b	103032434	Gse1 coiled-coil protein b	9/29	9/9	0/10	0/10
LOC103040220	103040220	histidine N-acetyltransferase	9/29	9/9	0/10	0/10
LOC111189810	111189810	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
si:ch211-66i15.4	103036214	uncharacterized protein LOC560315 homolog	9/29	1/9	8/10	0/10
gusb	103037105	glucuronidase, beta	9/29	7/9	0/10	2/10
rhocb	103031338	ras homolog family member Cb	9/29	0/9	8/10	1/10
si:dkey-63d15.12	125806161	uncharacterized si:dkey-63d15.12	9/29	0/9	3/10	6/10
klf17	103035114	Kruppel-like factor 17	9/29	9/9	0/10	0/10
uhrf1bp1	103036129	UHRF1 binding protein 1	9/29	0/9	0/10	9/10
LOC125804890	125804890	transmembrane 4 L6 family member 4-like	9/29	9/9	0/10	0/10
LOC103026930	103026930	RAC-alpha serine/threonine-protein kinase	9/29	9/9	0/10	0/10
LOC103034380	103034380	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
lpgat1	103023845	lysophosphatidylglycerol acyltransferase 1	9/29	9/9	0/10	0/10
pex16	103037472	peroxisomal biogenesis factor 16	9/29	8/9	0/10	1/10
myog	103037490	myogenin	9/29	0/9	0/10	9/10
si:ch211-86h15.1	103039085	uncharacterized si:ch211-86h15.1	9/29	0/9	0/10	9/10
LOC111189809	111189809	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC125780580	125780580	histone H2A-like	9/29	0/9	8/10	1/10
LOC103035974	103035974	cytochrome P450 2B4-like	9/29	0/9	0/10	9/10
LOC103034696	103034696	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC103032533	103032533	NACHT, LRR and PYD domains-containing protein 12	9/29	9/9	0/10	0/10
LOC125782367	125782367	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC103035341	103035341	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC111191325	111191325	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC125782313	125782313	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC103035960	103035960	gamma-crystallin M2	9/29	9/9	0/10	0/10
ywhag1	103041195	3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide 1	9/29	0/9	5/10	4/10

tnfaip8l3	103022034	tumor necrosis factor, alpha-induced protein 8-like 3	9/29	1/9	7/10	1/10
LOC111191307	111191307	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC111194828	111194828	skin secretory protein xP2-like	9/29	9/9	0/10	0/10
madd	103030336	MAP-kinase activating death domain	9/29	9/9	0/10	0/10
syngap1a	103024487	synaptic Ras GTPase activating protein 1a	9/29	0/9	4/10	5/10
LOC111188912	111188912	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
npy2rl	103047056	neuropeptide Y receptor Y2, like	9/29	0/9	9/10	0/10
avpr2b.1	103036026	arginine vasopressin receptor 2b, tandem duplicate, 1	9/29	9/9	0/10	0/10
atp13a1	103035743	ATPase 13A1	9/29	0/9	2/10	7/10
LOC111194995	111194995	trace amine-associated receptor 6-like	9/29	0/9	0/10	9/10
arhgef1b	103031326	Rho guanine nucleotide exchange factor (GEF) 1b	9/29	0/9	0/10	9/10
LOC125782685	125782685	uncharacterized LOC125782685	9/29	2/9	7/10	0/10
rgs9a	103026135	regulator of G protein signaling 9a	9/29	9/9	0/10	0/10
LOC125799046	125799046	protocadherin beta-16-like	9/29	0/9	9/10	0/10
LOC111195000	111195000	ras-related protein Rab-3C	9/29	9/9	0/10	0/10
baiap3	103037679	BAI1 associated protein 3	9/29	8/9	0/10	1/10
tmem65	103046307	transmembrane protein 65	9/29	0/9	4/10	5/10
nmd3	103047645	NMD3 ribosome export adaptor	9/29	0/9	0/10	9/10
LOC103045325	103045325	CBY1-interacting BAR domain-containing protein 1-like	9/29	9/9	0/10	0/10
LOC125785603	125785603	uncharacterized LOC125785603	9/29	6/9	0/10	3/10
LOC103036591	103036591	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC111191315	111191315	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
LOC111191323	111191323	gamma-crystallin M2-like	9/29	9/9	0/10	0/10
dennd4c	103042516	DENN/MADD domain containing 4C	9/29	0/9	0/10	9/10
LOC103038137	103038137	transcription factor AP-2-alpha	9/29	0/9	0/10	9/10
LOC103035084	103035084	PDZ domain-containing protein 4	9/29	9/9	0/10	0/10
LOC103026283	103026283	E3 ubiquitin-protein ligase TRIM69-like	9/29	8/9	1/10	0/10

fibcd1a	103041154	fibrinogen C domain containing 1a	9/29	9/9	0/10	0/10
LOC111192659	111192659	uncharacterized LOC111192659	9/29	0/9	8/10	1/10
LOC103021927	103021927	coiled-coil domain-containing protein 60	9/29	2/9	0/10	7/10

9.1.4. Geni koji preklapaju CNVs isključivo površinskih genoma

Tablica S4. Svi protein-kodirajući geni koji preklapanja CNVs pronađene isključivo u površinskim genomima. Od 1726 ukupnih gena, prikazano je samo 128 gena koji preklapaju CNV iz barem 5 (minimalno 30 %) različitih špiljskih genoma.

Genski simbol (NCBI)	Genski ID (NCBI)	Ime gena (NCBI)	površina	Rascon	Rio Choy
aldh1a2	103042804	aldehyde dehydrogenase 1 family, member A2	14/15	5/6	9/9
tmem147	103031793	transmembrane protein 147	13/15	6/6	7/9
LOC103031476	103031476	NACHT, LRR and PYD domains-containing protein 3	13/15	6/6	7/9
rnf41	103035049	ring finger protein 41	12/15	6/6	6/9
LOC125801242	125801242	zinc finger protein 585A-like	12/15	5/6	7/9
LOC125782686	125782686	uncharacterized LOC125782686	12/15	5/6	7/9
LOC107196865	107196865	E3 ubiquitin-protein ligase TRIM16-like	11/15	4/6	7/9
LOC125806043	125806043	mucin-22-like	10/15	6/6	4/9
LOC125801137	125801137	E3 SUMO-protein ligase ZBED1-like	9/15	3/6	6/9
LOC125782699	125782699	scavenger receptor cysteine-rich type 1 protein M130-like	9/15	3/6	6/9
LOC103023352	103023352	GDP-L-fucose synthase	9/15	2/6	7/9
LOC125782464	125782464	uncharacterized LOC125782464	8/15	4/6	4/9
LOC125782465	125782465	uncharacterized LOC125782465	8/15	4/6	4/9
si:dkeyp-97a10.3	103027437	uncharacterized si:dkeyp-97a10.3	8/15	3/6	5/9
LOC111193848	111193848	zinc finger protein 850	8/15	5/6	3/9
LOC107197812	107197812	zinc finger protein 501	8/15	5/6	3/9
ccnf	103029415	cyclin F	8/15	2/6	6/9
LOC111189980	111189980	uncharacterized LOC111189980	8/15	2/6	6/9
LOC125801531	125801531	uncharacterized LOC125801531	8/15	6/6	2/9
b3gnt2a	103037893	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2a	8/15	5/6	3/9
lyrm1	111188442	LYR motif containing 1	8/15	5/6	3/9
LOC111195209	111195209	uncharacterized LOC111195209	8/15	1/6	7/9
cgnl1	103041976	cingulin-like 1	8/15	6/6	2/9
LOC125784863	125784863	uncharacterized LOC125784863	8/15	5/6	3/9
LOC125801535	125801535	uncharacterized LOC125801535	8/15	3/6	5/9
LOC125780599	125780599	epidermal differentiation-specific protein-like	7/15	3/6	4/9
LOC125801119	125801119	uncharacterized LOC125801119	7/15	1/6	6/9
itsn2b	103045212	intersectin 2b	7/15	6/6	1/9

LOC111189493	111189493	uncharacterized LOC111189493	6/15	3/6	3/9
LOC103027630	103027630	odorant receptor 131-2-like	6/15	5/6	1/9
LOC125801670	125801670	zinc finger protein 850-like	6/15	6/6	0/9
slc19a2	103043267	solute carrier family 19 member 2	6/15	6/6	0/9
LOC111190803	111190803	protein NYNRIN-like	6/15	6/6	0/9
nrxn3a	103035103	neurexin 3a	6/15	5/6	1/9
stxbp6	103022399	syntaxin binding protein 6 (amisyn)	6/15	6/6	0/9
LOC125802188	125802188	uncharacterized protein K02A2.6-like	6/15	2/6	4/9
selenow2a	103033471	selenoprotein W, 2a	6/15	3/6	3/9
spag17	103025850	sperm associated antigen 17	6/15	6/6	0/9
LOC125799173	125799173	uncharacterized LOC125799173	6/15	3/6	3/9
LOC111195922	111195922	NAD(P)(+)--arginine ADP-ribosyltransferase 2-like	6/15	3/6	3/9
septin9b	103042219	septin 9b	6/15	6/6	0/9
basp1	103023200	brain abundant, membrane attached signal protein 1	6/15	6/6	0/9
zgc:92749	103026880	elongation of very long chain fatty acids protein	6/15	5/6	1/9
im:7138535	103041104	uncharacterized protein LOC797998 homolog	6/15	4/6	2/9
zdhhc14	103029743	zinc finger DHHC-type palmitoyltransferase 14	6/15	6/6	0/9
kcnh1b	103034506	potassium voltage-gated channel, subfamily H (eag-related), member 1b	6/15	1/6	5/9
LOC111194360	111194360	uncharacterized LOC111194360	6/15	6/6	0/9
LOC111193354	111193354	uncharacterized LOC111193354	6/15	6/6	0/9
LOC125804596	125804596	uncharacterized LOC125804596	6/15	2/6	4/9
LOC111197170	111197170	scavenger receptor cysteine-rich type 1 protein M130-like	6/15	5/6	1/9
LOC125798967	125798967	uncharacterized LOC125798967	6/15	0/6	6/9
LOC103046458	103046458	matrix metalloproteinase-16-like	6/15	6/6	0/9
LOC103042106	103042106	cytochrome c oxidase assembly factor 5	6/15	6/6	0/9
LOC111188177	111188177	uncharacterized LOC111188177	6/15	2/6	4/9
casp23	103021832	caspase 23, apoptosis-related cysteine peptidase	6/15	6/6	0/9
zgc:194655	103022147	uncharacterized protein LOC794380 homolog	6/15	6/6	0/9
LOC125801099	125801099	deleted in malignant brain tumors 1 protein-like	6/15	5/6	1/9
LOC125782292	125782292	uncharacterized LOC125782292	6/15	5/6	1/9
LOC111196805	111196805	uncharacterized LOC111196805	6/15	4/6	2/9
si:ch211-212k18.15	103036756	uncharacterized si:ch211-212k18.15	6/15	6/6	0/9
pcsk6	103021799	proprotein convertase subtilisin/kexin type 6	6/15	6/6	0/9

mapk14a	103046996	mitogen-activated protein kinase 14a	6/15	6/6	0/9
kcnma1a	103039470	potassium large conductance calcium-activated channel, subfamily M, alpha member 1a	6/15	6/6	0/9
ltk	103040582	leukocyte receptor tyrosine kinase	6/15	6/6	0/9
LOC103044661	103044661	ADAMTS-like protein 1	6/15	6/6	0/9
LOC103025888	103025888	cAMP-specific 3',5'-cyclic phosphodiesterase 4C	6/15	6/6	0/9
pgap1	103031160	post-GPI attachment to proteins inositol deacylase 1	6/15	6/6	0/9
pam	103031431	peptidylglycine alpha-amidating monooxygenase	6/15	6/6	0/9
efl1	103028390	elongation factor like GTPase 1	6/15	2/6	4/9
LOC107197471	107197471	putative protocadherin beta-18	5/15	3/6	2/9
LOC125799167	125799167	uncharacterized LOC125799167	5/15	0/6	5/9
LOC111196429	111196429	uncharacterized LOC111196429	5/15	0/6	5/9
LOC103029182	103029182	striatin	5/15	5/6	0/9
LOC103033885	103033885	EMILIN-3	5/15	5/6	0/9
LOC103033531	103033531	latent-transforming growth factor beta-binding protein 4	5/15	5/6	0/9
LOC125780640	125780640	uncharacterized LOC125780640	5/15	4/6	1/9
LOC111196250	111196250	trace amine-associated receptor 1-like	5/15	0/6	5/9
abcc5	103033123	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	5/15	5/6	0/9
si:dkey-34e4.1	103035680	carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase protein	5/15	5/6	0/9
LOC111191480	111191480	granzyme K-like	5/15	4/6	1/9
etfbkmt	103045035	electron transfer flavoprotein subunit beta lysine methyltransferase	5/15	5/6	0/9
LOC111193382	111193382	uncharacterized LOC111193382	5/15	1/6	4/9
nck2b	103023033	NCK adaptor protein 2b	5/15	5/6	0/9
gabrg2	103027255	gamma-aminobutyric acid type A receptor subunit gamma2	5/15	4/6	1/9
lsp1a	103031193	lymphocyte specific protein 1 a	5/15	4/6	1/9
vps33a	103022932	VPS33A core subunit of CORVET and HOPS complexes	5/15	0/6	5/9
LOC103021841	103021841	insulin gene enhancer protein ISL-1	5/15	0/6	5/9
LOC125786877	125786877	galectin-8-like	5/15	1/6	4/9
LOC103033154	103033154	uncharacterized LOC103033154	5/15	3/6	2/9
rbm11	103039150	RNA binding motif protein 11	5/15	5/6	0/9
LOC103030317	103030317	amyloid beta (A4) precursor protein-binding, family B, member 1 interacting protein	5/15	5/6	0/9
LOC111192936	111192936	uncharacterized LOC111192936	5/15	2/6	3/9
clec16a	103042374	C-type lectin domain containing 16A	5/15	5/6	0/9

LOC125804738	125804738	uncharacterized LOC125804738	5/15	5/6	0/9
gng12a	107196998	guanine nucleotide binding protein (G protein), gamma 12a	5/15	5/6	0/9
LOC103045308	103045308	E3 ubiquitin-protein ligase HECW1	5/15	5/6	0/9
rims2b	103023006	regulating synaptic membrane exocytosis 2b	5/15	5/6	0/9
adora2b	103021368	adenosine A2b receptor	5/15	0/6	5/9
gpbar1	103021352	G protein-coupled bile acid receptor 1	5/15	5/6	0/9
LOC103028418	103028418	microsomal triglyceride transfer protein	5/15	5/6	0/9
kcna4	111188195	potassium voltage-gated channel, shaker-related subfamily, member 4	5/15	5/6	0/9
LOC103026444	103026444	uncharacterized LOC103026444	5/15	4/6	1/9
LOC103026764	103026764	mucin-12	5/15	5/6	0/9
erg	103046049	ETS transcription factor ERG	5/15	1/6	4/9
LOC125782580	125782580	rho-related GTP-binding protein RhoG-like	5/15	2/6	3/9
LOC103030984	103030984	L-amino-acid oxidase	5/15	1/6	4/9
LOC125799106	125799106	uncharacterized LOC125799106	5/15	1/6	4/9
pgap3	103047346	post-GPI attachment to proteins phospholipase 3	5/15	0/6	5/9
LOC111193568	111193568	coiled-coil domain-containing protein 106	5/15	5/6	0/9
nalf1a	103032664	NALCN channel auxiliary factor 1a	5/15	5/6	0/9
LOC111194136	111194136	uncharacterized LOC111194136	5/15	5/6	0/9
atad2b	103025295	ATPase family AAA domain containing 2B	5/15	5/6	0/9
LOC125801712	125801712	uncharacterized LOC125801712	5/15	5/6	0/9
timmdc1	103033360	translocase of inner mitochondrial membrane domain containing 1	5/15	5/6	0/9
cpxm2	103041423	carboxypeptidase X (M14 family), member 2	5/15	5/6	0/9
acox3	103033929	acyl-CoA oxidase 3, pristanoyl	5/15	5/6	0/9
thrap3a	103043409	thyroid hormone receptor associated protein 3a	5/15	5/6	0/9
LOC125801334	125801334	zinc finger protein 239-like	5/15	0/6	5/9
wdfy4	103039810	WDFY family member 4	5/15	5/6	0/9
acsf3	103034152	acyl-CoA synthetase family member 3	5/15	3/6	2/9
slc5a6a	103042151	solute carrier family 5 member 6a	5/15	4/6	1/9
LOC103030678	103030678	zinc finger protein 585A	5/15	0/6	5/9
tmed3	111192548	transmembrane p24 trafficking protein 3	5/15	4/6	1/9
LOC111191577	111191577	uncharacterized LOC111191577	5/15	2/6	3/9
ldlra	103022705	low density lipoprotein receptor a	5/15	5/6	0/9
golph3b	103021521	golgi phosphoprotein 3b	5/15	0/6	5/9
LOC125799291	125799291	uncharacterized LOC125799291	5/15	3/6	2/9

pip5k1bb	103036273	phosphatidylinositol-4-phosphate 5-kinase, type I, beta b	5/15	5/6	0/9

9.1.5. Geni divergentni između ekotipova

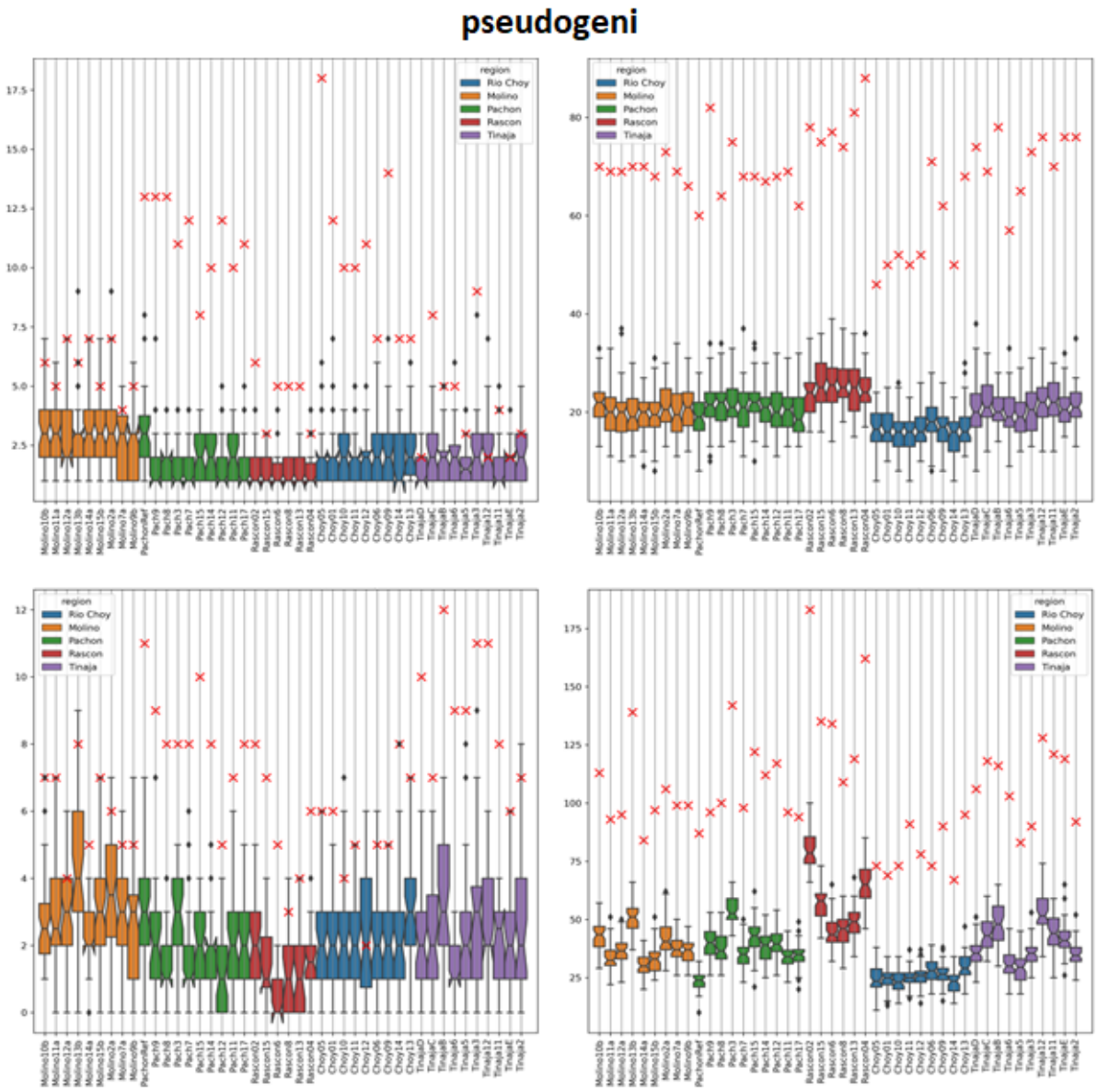
Tablica S4. Rezultati statističkog testiranja CNV gena. Od 2819 CNV gena, 165 je bilo statistički značajno (α<0.05) između svih kombinacija parova populacija špilja-površina. Od tih 165 gena, 102 gena su značajno različita brojem kopija i u grupiranim usporedbama špilja-površina te su navedeni u ovoj tablici.

Genske koordinate	Genski ID (NCBI)	Ime gena (NCBI)
chrom 1:113 316 352-113 317 935	111195209	uncharacterized LOC111195209
chrom 1:127 770 015-127 778 305	125785580	stonustoxin subunit beta-like
chrom 1:129 917 718-129 920 059	125785664	neoverrucotoxin subunit alpha-like
chrom 1:130 059 992-130 068 264	111188594	fucolectin-1-like
chrom 1:131 531 427-131 551 041	125784856	NLR family CARD domain-containing protein 3-like
chrom 1:132 912 592-132 972 820	125784871	protein NLRC3-like
chrom 1:13 836 759-13 844 029	103025050	uncharacterized si:dkey-3n22.9
chrom 1:38 611 364-38 614 391	125801153	mucin-2-like
chrom 1:39 575 235-39 582 158	125801397	uncharacterized LOC125801397
chrom 1:40 302 506-40 304 059	125802291	chromobox protein homolog 5-like
chrom 1:51 299 073-51 314 467	125802261	uncharacterized LOC125802261
chrom 2:29 243 080-29 248 623	125799116	uncharacterized LOC125799116
chrom 2:38 685 381-38 722 757	125799138	scavenger receptor cysteine-rich type 1 protein M130-like
chrom 2:39 108 115-39 145 193	125799140	scavenger receptor cysteine-rich type 1 protein M130-like
chrom 2:73 676 522-73 682 045	111189493	uncharacterized LOC111189493
chrom 2:73 681 407-73 685 494	125799173	uncharacterized LOC125799173
chrom 3:13 494 296-13 510 170	125799429	phospholipase B-like 1
chrom 3:21 289 929-21 291 857	111190345	uncharacterized LOC111190345
chrom 3:28 905 107-28 920 079	103026012	uncharacterized LOC103026012
chrom 3:31 018 336-31 038 649	111197155	uncharacterized LOC111197155
chrom 3:31 041 070-31 044 321	125799490	uncharacterized LOC125799490
chrom 3:31 061 218-31 073 301	125799489	uncharacterized LOC125799489
chrom 3:31 081 104-31 085 999	111197153	uncharacterized LOC111197153
chrom 3:35 945 525-35 956 823	103026892	60 kDa lysophospholipase
chrom 3:64 360 227-64 361 046	111191630	hemoglobin subunit beta-2-like
chrom 3:64 367 441-64 368 168	111191628	hemoglobin embryonic subunit alpha
chrom 3:64 369 203-64 370 021	103027764	hemoglobin subunit beta-2
chrom 3:64 391 590-64 392 417	111196759	hemoglobin subunit beta-2-like
chrom 4:11 811 318-11 813 593	125801369	zinc finger protein 501-like
chrom 4:13 906 386-13 910 289	125801672	uncharacterized LOC125801672
chrom 4:14 269 225-14 278 619	125801160	uncharacterized LOC125801160
chrom 4:14 594 434-14 597 830	125801673	uncharacterized LOC125801673
chrom 4:15 129 439-15 131 144	125801525	uncharacterized LOC125801525

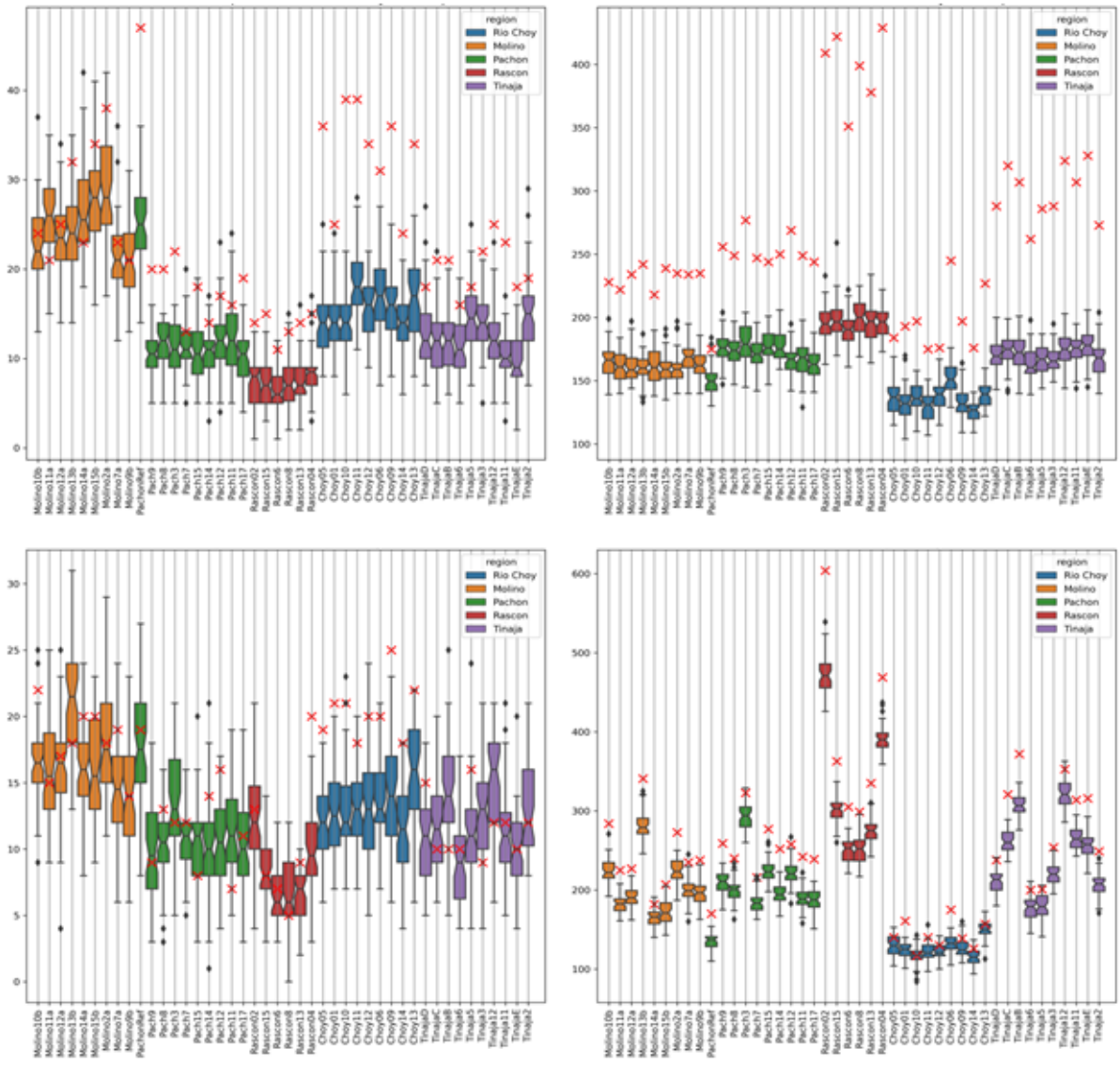
chrom 4:16 954 084-16 955 581	111194136	uncharacterized LOC111194136
chrom 4:18 364 477-18 366 863	103044012	uncharacterized LOC103044012
chrom 4:18 606 155-18 612 254	125801392	zinc finger protein 239-like
chrom 4:21 294 725-21 303 162	125801098	uncharacterized LOC125801098
chrom 4:26 432 007-26 433 675	125801744	uncharacterized LOC125801744
chrom 4:32 961 185-32 964 138	111193700	THAP domain-containing protein 2
chrom 4:50 505 037-50 512 387	111194616	polymeric immunoglobulin receptor-like
chrom 4:7 458 307-7 463 069	125801127	uncharacterized LOC125801127
chrom 4:7 462 867-7 464 696	125801657	uncharacterized LOC125801657
chrom 4:8 548 502-8 572 333	125801157	zinc finger protein 585A-like
chrom 4:9 643 839-9 649 988	125801665	uncharacterized LOC125801665
chrom 5:28 757 954-28 774 326	111193711	membrane-spanning 4-domains subfamily A member 4A
chrom 5:43 624 764-43 629 381	111189980	uncharacterized LOC111189980
chrom 6:13 508 151-13 515 294	111196955	uncharacterized protein C14orf93-like
chrom 6:21 864 374-21 866 847	125802482	uncharacterized LOC125802482
chrom 6:3 175 716-3 191 471	111193392	meprin A subunit beta-like
chrom 6:57 137 416-57 140 951	125802504	uncharacterized LOC125802504
chrom 7:23 641 892-23 646 686	111190362	uncharacterized LOC111190362
chrom 8:32 368 219-32 376 206	103026072	uncharacterized LOC103026072
chrom 8:34 010 777-34 018 984	111197052	uncharacterized LOC111197052
chrom 8:49 849 392-49 857 315	125803858	uncharacterized LOC125803858
chrom 9:41 652 575-41 654 054	125804571	protein FAM214A-like
chrom 10:52 748 997-52 750 258	125804738	uncharacterized LOC125804738
chrom 12:15 752 078-15 757 800	125806212	uncharacterized LOC125806212
chrom 12:17 178 887-17 182 553	111188878	uncharacterized LOC111188878
chrom 12:17 430 602-17 434 201	125806165	uncharacterized LOC125806165
chrom 13:9 187 469-9 201 867	125780609	uncharacterized LOC125780609
chrom 13:9 753 164-9 758 183	125780583	uncharacterized LOC125780583
chrom 13:9 780 541-9 786 450	111195202	uncharacterized LOC111195202
chrom 14:6 860 994-6 866 571	125781121	uncharacterized LOC125781121
chrom 14:6 866 112-6 869 709	125781133	uncharacterized LOC125781133
chrom 15:1 507 236-1 514 155	111194352	uncharacterized LOC111194352
chrom 15:15 967 007-15 982 662	111196589	uncharacterized LOC111196589
chrom 15:16 399 481-16 436 353	111197390	uncharacterized LOC111197390
chrom 15:8 110 427-8 115 370	125781258	uncharacterized LOC125781258
chrom 15:8 246 685-8 256 455	125781257	uncharacterized LOC125781257
chrom 16:1 680 590-1 683 959	111194172	PWWP domain-containing DNA repair factor 3B-like
chrom 16:19 136 508-19 142 042	125782174	uncharacterized LOC125782174
chrom 16:30 383 731-30 396 272	111192539	golgin subfamily A member 6-like protein 22

chrom 16:36 297 783-36 301 828	125782233	endogenous retrovirus group FC1 Env polyprotein-like
chrom 17:25 196 282-25 198 651	125782385	uncharacterized protein LOC550590 homolog
chrom 18:16 490 143-16 494 851	111192142	uncharacterized LOC111192142
chrom 18:32 569 708-32 575 563	125782686	uncharacterized LOC125782686
chrom 19:10 526 435-10 534 170	125784445	uncharacterized LOC125784445
chrom 19:9 181 567-9 183 327	125784514	uncharacterized LOC125784514
chrom 20:19 298 527-19 302 638	125784832	uncharacterized LOC125784832
chrom 20:35 544 626-35 554 959	111196706	uncharacterized LOC111196706
chrom 20:7 737 711-7 755 053	125784890	trace amine-associated receptor 13c-like
chrom 21:34 238 847-34 248 037	125785565	uncharacterized LOC125785565
chrom 21:34 248 127-34 265 088	111194392	uncharacterized LOC111194392
chrom 21:34 256 652-34 265 088	125785564	uncharacterized LOC125785564
chrom 21:38 813 965-38 815 897	111188451	B-cell receptor CD22-like
chrom 21:38 859 465-38 865 870	125785616	B-cell receptor CD22-like
chrom 21:38 929 168-38 931 726	125785783	B-cell receptor CD22-like
chrom 21:39 315 199-39 317 614	125785782	B-cell receptor CD22-like
chrom 21:40 236 788-40 238 726	125785665	uncharacterized LOC125785665
chrom 21:40 709 493-40 718 690	125785663	NLR family CARD domain-containing protein 3-like
chrom 21:6 136 430-6 151 764	125785568	uncharacterized LOC125785568
chrom 22:11 287 414-11 300 385	125787068	scavenger receptor cysteine-rich type 1 protein M130-like
chrom 22:1 700 138-1 704 398	111197139	uncharacterized LOC111197139
chrom 22:1 948 624-1 950 844	125787031	uncharacterized LOC125787031
chrom 22:235 600-237 498	125787050	uncharacterized LOC125787050
chrom 22:4 132 260-4 136 851	125787058	uncharacterized LOC125787058
chrom 23:23 022 172-23 029 575	103043358	myelin-oligodendrocyte glycoprotein-like
chrom 23:31 084 380-31 087 978	111195059	uncharacterized LOC111195059
chrom 24:16 937 696-16 943 270	111195410	E3 ubiquitin-protein ligase DTX3L
chrom 25:26 304 122-26 306 905	111195147	apolipoprotein L3-like
chrom 25:8 612 394-8 621 300	111197671	C-type lectin domain family 4 member E-like
chrom 25:9 429 831-9 435 415	111193335	uncharacterized LOC111193335

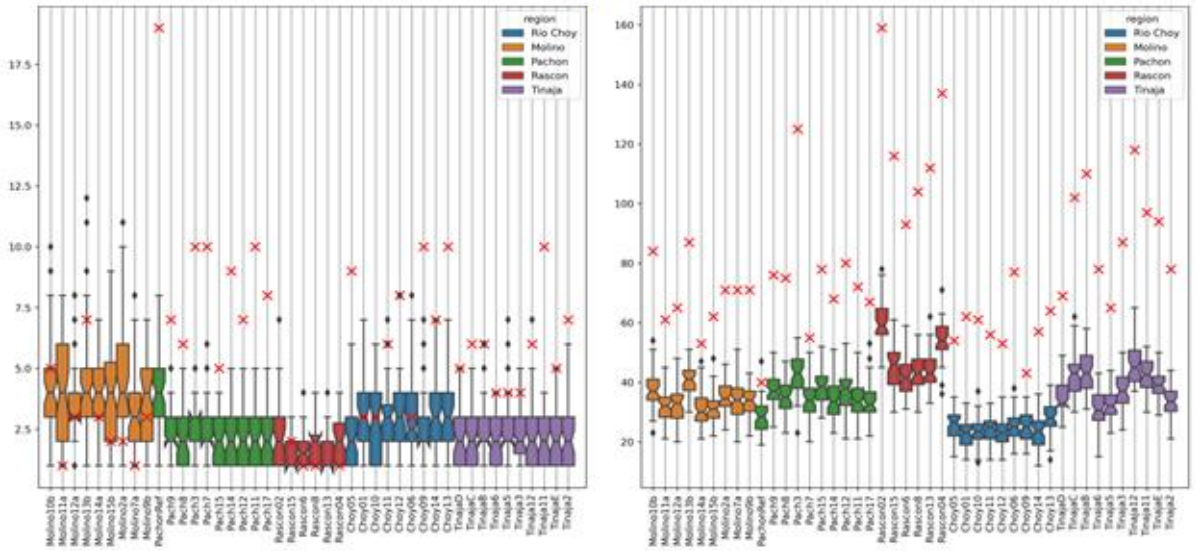
9.2. Slike



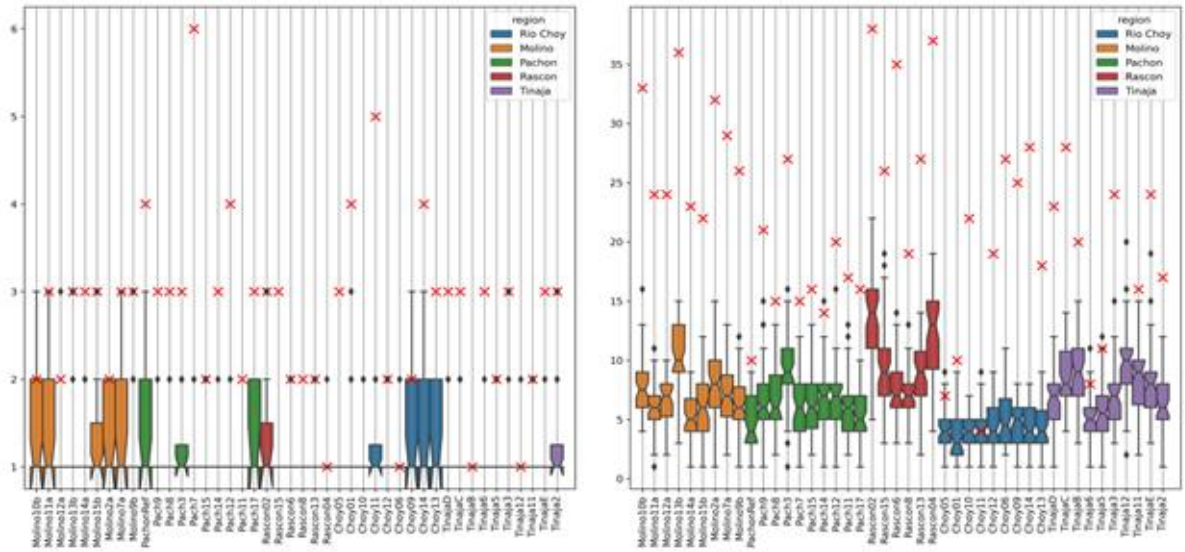
lncRNA

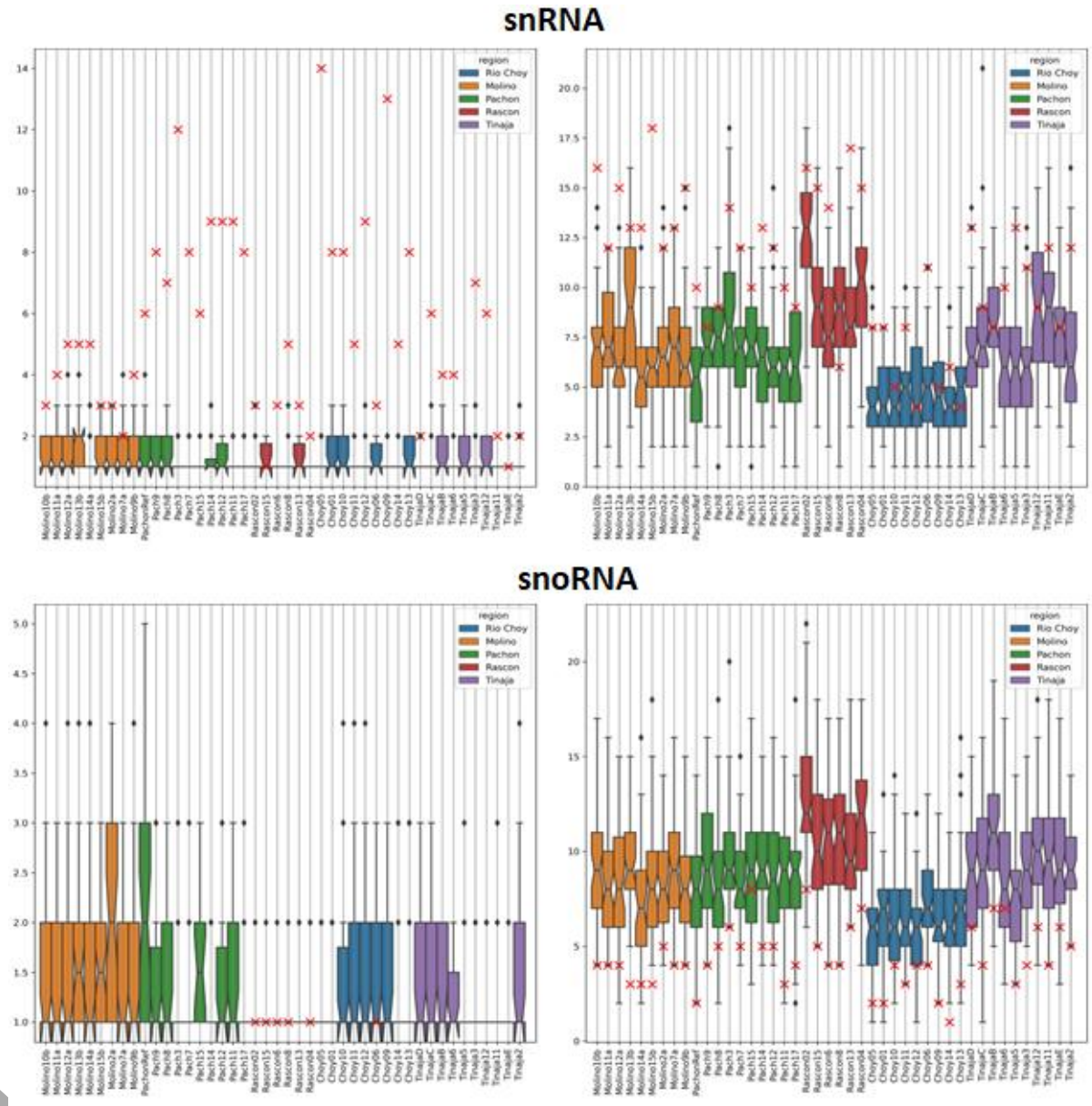


tRNA



rRNA





Slika S1. Rezultati permutacijskih analiza za sve promatrane kategorije gena (sažeto u Tablici 21). Duplikacije su prikazane lijevo, a delecije desno. Gornji grafovi predstavljaju duplikacije, odnosno delecije, koje u potpunosti zahvaćaju anotirani element. Donji grafovi predstavljaju duplikacije, odnosno delecije, koje samo djelomično zahvaćaju anotirani element, te su odsutni u svim kategorijama osim lncRNA i pseudogena zbog male duljine elementa. Vrijednosti su naznačene crvenim križićima za stvarne podatke. *Boxplot* prikazima (obojeni prema populacijama) predstavljene su distribucije vrijednosti dobivenih na temelju 100 permutacija.

10. ŽIVOTOPIS I POPIS PUBLIKACIJA

10.1. Obrazovanje

Sveučilište Josipa Jurja Strossmayera u Osijeku

Doktorski studij, Molekularne bioznanosti (MOBI)

2020./2021. -

Osijek, Hrvatska

Sveučilište u Zagrebu, Farmaceutsko-biokemijski fakultet
Magistar farmacije (Mag.Pharm)

2013./2014. – 2018./2019.
Zagreb, Hrvatska

- **Diplomski rad:** „Antimikrobni učinak koloidnih otopina nanočestica elementarnih metala i metalnih oksida na MSSA i MRSA“, pod mentorstvom prof. dr. sc. Ivana Kosalca i izv. prof. dr. sc. Ive Rezić.
- **Rektorova nagrad**a Sveučilišta u Zagrebu za rad „Utjecaj stereoizomera mentola na staničnu stijenku vrste Candida albicans te ljudskih eritrocita *in vitro*“, 2015./2016.

10.2. Publikacije

Pokrovac I, Pezer Ž. „Recent advances and current challenges in population genomics of structural variation in animals and plants“. Front Genet. 2022;13.

Pokrovac I, Rohner N, Pezer Ž. „The prevalence of copy number increase at multiallelic CNVs associated with cave colonization“. bioRxiv. 2023.

Rezić I, Majdak M, Ljoljić Bilić V, **Pokrovac I**, Martinaga L, Somogyi Škoc M, Kosalec I. „Development of Antibacterial Protective Coatings Active against MSSA and MRSA on Biodegradable Polymers“. Polymers. 2021;13(4):659.

Martinaga Pintarić L, Somogyi Škoc M, Ljoljić Bilić V, **Pokrovac I**, Kosalec I, Rezić I. „Synthesis, Modification and Characterization of Antimicrobial Textile Surface Containing ZnO Nanoparticles“. Polymers. 2020;12(6):1210.

10.3. Sudjelovanja na konferencijama

Pokrovac I, Pezer Ž. *Environmental impact on gene function in Mexican tetra*, 25th EMBL PhD Symposium, Heidelberg, 2023. (Oral Poster)

Pokrovac I, Pezer Sakač Ž. *Gene copy number variation contributes to environmental adaptation*, 14th Croatian Biological Congress, 2022. (Poster)

Postersko priopćenje na 7. Simpoziju studenata farmacije i medicinske biokemije (FARMEBS) 2018, Zagreb

Usmeno izlaganje na 5. Simpoziju studenata farmacije i medicinske biokemije (FARMEBS) 2016, Zagreb

10.4. Certifikati i radionice

Nvidia Fundamentals of Deep Learning

MedILS Bioinformatics School in Transcriptomics

EMBO Research Integrity Workshop

LabAnim tečaj za osposobljavanje osoba koje rade s pokusnim životinjama (FELASA ekvivalent), B kategorija

10.5. *Open-source* projekti

Svi projekti navedeni u tablici nalaze se na github stranici - https://github.com/ivanp1994.

Projekt	Opis	Tech stack
GuavaMuseFCS	Kreirao GUI alat za čitanje i analizu vlasničkih podataka o protočnoj citometriji (.FCS datoteke). Alat omogućuje odabir pragova, klaster analizu i vizualizaciju podataka.	Python (Tkinter, Numpy, Seaborn)
PyPerMANOVA	Implementirao PERMANOVA statističku metodu za neparametrijsku multivarijatnu analizu varijance u Pythonu.	Python (Numpy, Scipy)
ReadDepthVisualizer	Razvio alat za vizualizaciju podataka sekvenciranja nove generacije koji omogućuje interaktivni pregled genomskih regija s duplikacijama i delecijama.	Python (Tkinter, Seaborn, NCBI API)
TelOMPy	Razvio Python alat za mjerenje duljine telomera iz genomskih podataka korištenjem optičkog mapiranja genoma (OGM). Trenutno ga portiram u Rust za bolju izvedbu.	Python (Pandas), Rust
Bnxtools	Izradio Rust alatni set za obradu izlaznih datoteka optičkog mapiranja genoma (OGM), optimiziran za rad s velikim skupovima podataka.	Rust, BASH
QSAR Pipeline za kumarinske derivate	Razvio *open-source* analog QSAR softvera za predviđanje inhibitornog potencijala kumarinskih derivata na Dipeptidil Peptidazu III (hDP3). Implementiran je genetski algoritam za odabir deskriptora i integrirani su modeli strojnog učenja za predviđanje aktivnosti spojeva.	Python (Mordred, Scikit-learn), Snakemake, Conda
CitationTool	Skripta koja vraća APA format citata literature iz linka na *Digital Object Identifier* (DOI) znanstvenog rada. Nalazi se i u obliku grafičkog sučelja.	Python (REST API, PyQt5)
PyMart	Python sučelje za BioMart server ENSEMBL genomske baze podataka.	Python (REST API, Pandas)