Identifikacija periodičnosti višega reda u insektu Tribolium castaneum pomoću računalne grm metode

Vlahović, Ines

Sveučilište u Zagrebu

PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

INES VLAHOVIĆ

IDENTIFIKACIJA PERIODIČNOSTI VIŠEGA REDA U INSEKTU *TRIBOLIUM CASTANEUM* POMOĆU RAČUNALNE GRM METODE

DOKTORSKI RAD

Zagreb, 2014.

University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF PHYSICS

INES VLAHOVIĆ

IDENTIFICATION OF HIGHER-ORDER PERIODICITY IN INSECT *TRIBOLIUM CASTANEUM* USING COMPUTATIONAL GRM METHOD

DOCTORAL THESIS

Zagreb, 2014.

Sveučilište u Zagrebu

PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

INES VLAHOVIĆ

IDENTIFIKACIJA PERIODIČNOSTI VIŠEGA REDA U INSEKTU *TRIBOLIUM CASTANEUM* POMOĆU RAČUNALNE GRM METODE

DOKTORSKI RAD

Mentor:
Akademik, Prof. dr. sc. Vladimir Paar

Zagreb, 2014.

University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF PHYSICS

INES VLAHOVIĆ

IDENTIFICATION OF HIGHER-ORDER PERIODICITY IN INSECT *TRIBOLIUM CASTANEUM* USING COMPUTATIONAL GRM METHOD

DOCTORAL THESIS

Supervisor:
Academician, Prof. dr.sc. Vladimir Paar

Zagreb, 2014.

Zahvale

Od početka mojeg obrazovanja, osnovne i srednje škole pa do fakulteta i dalje, najveća potpora bili su mi moja obitelj (tata Velimir, mama Nediljka i brat Davorin) kao i od fakultetskih dana Tomislav Kadežabek moj budući suprug, koji su se stalno odricali svojeg slobodnog vremena kako bi mi omogućili kvalitetno obrazovanje i profesionalni rast te rješavanje svih problema tokom tog teškog puta. Također želim zahvaliti svojem mentoru, akademiku, prof. dr. sc. Vladimiru Paaru, koji me tokom cijelog poslijediplomskog studija hrabrio i usmjeravao, te Matku Glunčiću dobrom suradniku i prijatelju. Doktorici Mariji Rosandić-Pilaš želim zahvaliti na dugotrajnim razgovorima o problemima svih vrsta i poticanja za ustrajnost kada je to bilo najpotrebnije. Želim zahvaliti svim zaposlenicima PMF-a Fizičkog odsjeka, koji su mi pomagali tokom cijelog poslijediplomskog studija i ubrzavali sve potrebne papirologije za pravovremeni dovršetak studija, i mojim novo stečenim prijateljima Sanji, Janji, Kseniji, Marini, Vlasti, Marini, Slavici, Kreši, Ivanu, Leu, Željku, Gorjani, Hrvoju.

Puno vam hvala!

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Fizički odsjek

Doktorska disertacija

SAŽETAK

IDENTIFIKACIJA PERIODIČNOSTI VIŠEGA REDA U INSEKTU *TRIBOLIUM CASTANEUM* POMOĆU RAČUNALNE GRM METODE

INES VLAHOVIĆ

Prirodoslovno-matematički fakultet, Zagreb

Koristeći računalnu metodu GRM, Global Repeat Map, u disertaciji sam analizirala genom insekta *Tribolium castaneum,* malog brašnara, kako bih identificirala periodičnosti višega reda (HOR-ove). Uz algoritme za detekciju tandemne DNK, koji rade na principu dinamičkih matrica poravnanja, kompresije podataka, mapiranja u numeričku sekvencu i FFT (Fast Fourier Transform), naša metoda čini učinkovit alat za analizu DNK sekvence. Prednost korištene metode, GRM, je u direktnom mapiranju DNK simboličke sekvence u frekventnu domenu stvarajući globalnu mapu te preko kompletnog ansambla „ključnih riječi", s obzirom da nema ulaznih parametara, može identificirati repeticije svih duljina bez obzira da li postoje devijacije kopija od savršenog uzorka. Pomoću ove metode identificirala sam veliki broj periodičnosti višega reda u insektu *T. castaneum* koje su karakteristične za sisavce. Te periodičnosti višega reda zasnivaju se na monomerima duljina ~360 bp što su zapravo često istraživani sateliti insekta *T.castaneum*. Identifikacija HOR-ova je bitna za daljnja istraživanja regulatornih uloga tj. ekspresiji gena i u mehanizmima evolucijskog razvoja.

(119 stranica, 78 slika, 27 tablica, 156 literaturnih navoda, jezik izvornika: hrvatski jezik)

Ključne riječi: tandemne repeticije, periodičnosti višega reda – HOR, GRM – Global Repeat Map, *Tribolium castaneum*

Mentor: Akademik Vladimir Paar, Prirodoslovno-matematički fakultet, Hrvatska akademija znanosti i umjetnosti

Povjerenstvo za obranu:

1. akademik, Vladimir Paar, Prirodoslovno-matematički fakultet, HAZU
2. prof.dr.sc. Đurđica Ugarković, Institut Ruđer Bošković
3. doc. dr. sc. Matko Glunčić, Prirodoslovno-matematički fakultet

Rad prihvaćen: 2014.

University of Zagreb
Faculty of Science
Department of Physics

Doctoral Thesis

ABSTRACT

IDENTIFICATION OF HIGHER-ORDER PERIODICITY IN INSECT *TRIBOLIUM CASTANEUM* USING COMPUTATIONAL GRM METHOD

INES VLAHOVIĆ

Faculty of science, Zagreb

With GRM computational method, in thesis, I have analyzed insect *Tribolium castaneum* genome, in order to identify higher order repeats (HORs). Along the other algorithms for detection of tandem repeats, which works on principles of dynamical matrix alignments, data compression, mapping sequence in numerical one and Fast Fourier Transform, our method is effective tool for analysis a DNA sequence. Advantage of GRM method is in direct mapping of symbolic DNA sequence in frequency domain making global map and with the complete assemble of "key words", with no input parameters, it can detect repeats of all lengths in spite of copy deviations of perfect sample. With this method I have identified a great number of higher order repeats in insect *T.castaneum*, which are characteristic for mammals. Those HORs are assembled from monomers of length ~360 bp, which are often investigated satellites of insect *T.castaneum*. Identification of HOR's is relevant for further investigations of regulatory roles ie. gene expression and mechanisms of evolution development.

(119 pages, 78 figures, 27 tables, 156 references, original in croatian)

Keywords: tandem repeats, higher order repeats – HOR, GRM - Global Repeat Map, *Tribolium castaneum*

Supervisor: Academician Vladimir Paar, HAZU, Faculty of Science

Thesis Committee:

1. academician, Vladimir Paar, Faculty of Science, Croatian Academy of Science and Art
2. prof.dr.sc. Đurđica Ugarković, Ruđer Bošković Institute
3. doc. dr. sc. Matko Glunčić, Faculty of Science

Thesis accepted: 2014.

Sadržaj

1. UVOD 1

2. LITERATURNI PREGLED 4

3. MATERIJALI I METODE 26

3.1 TANDEMNE REPETICIJE I PERIODIČNOSTI VIŠEGA REDA 26
3.2 FLEKSIBILNI STATISTIČKI ALGORITMI USPOREDBE NIZOVA 27
3.2.1.Tandem Repeat Finder 28
3.3 ALGORITMI ZA PROCESIRANJE SIGNALA 29
3.3.1 Spectral Repeat Finder 29
3.4 GLOBAL REPEAT MAP RAČUNALNA METODA - GRM 31
3.5 NEEDLEMAN – WUNSCH ALGORITAM 36
3.6 BLAST - BASIC LOCAL ALIGNMENT SEARCH TOOL 38

4. REZULTATI I RASPRAVA 39

4.1 TCAST SATELITI ~ 360 BP 45
4.1.1 GG695826.1 48
4.1.2 DS497953.1 51
4.1.3 Ostale HOR strukture zasnovane na TCAST satelitima 59
4.2 PIK 51 BP 63
4.2.1 GG694243.1 63
4.2.2 GG695755.1 65
4.3. PIK 63 67
4.3.1 CM000284.2 67
4.3.2 GG694162.1 69
4.4 PIK 73 70
4.4.1 GG695679.1 70
4.4.2 GG695418.1 71
4.4.3 GG695223.1 72
4.4.4 GG695058.1 74
4.4.5 GG695890.1 75
4.4.6 GG694624.1 77
4.5 PIK 170 BP 78
4.5.1 GG695869.1 78
4.5.2 CM000279.1. 79
4.5.3 CM000277.2 82
4.5.4 DS497720.1 82
4.5.5 DS497688.1 83
4.6. PIK 180 BP 85
4.6.1 GG695637.1 85
4.7 PIK 218 BP 86
47.1 CM000279.1 86
4.8. PIK 311 BP 87
4.8.1 GG694292.1 87
4.9 PIK 328 88
4.9.1 DS497673.1 88
4.10 PIK 332 BP 89
4.10.1 GG695436.1 89
4.11 PIK 721 90

4.11.1 GG694249.1 90
4.12. PIK 1110 BP 91
4.1.12 GG695437.1 91
5. ZAKLJUČAK 93
6. DODATAK 95
7. POPIS LITERATURE 105
8. ŽIVOTOPIS 117

1. Uvod

Sekvenca deoksiribonukleinske kiseline tj. DNK, se sastoji od kodirajućeg i nekodirajućeg dijela (nekada zvanog i „junk“ DNK [1] koji čini veći dio eukariotskog genoma). Repeticije čine većinu DNK sekvence i one se mogu nalaziti i u kodirajućem i u nekodirajućem dijelu, što je naročito vidljivo na primjeru genoma čovjeka gdje je više od polovice ukupnog genoma sastavljeno od repeticija [2], ali i u genomima ostalih sisavaca. Postoje različite vrste repeticija u DNK. One mogu biti klasificirane kao tandemne repeticije, raspršene repeticije, segmentne duplikacije te kompleksna ponavljanja. Tandemne repeticije se dodatno, prema svojim duljinama, mogu klasificirati kao mikrosateliti (1 do ~6 bp), minisateliti (~6 do ~100 bp), sateliti (~100 do ~2 kb) i makrosateliti (veći od ~2 kb). Sve te repeticije mogu sadržavati supstitucije nukleotida, umetanja ili brisanja nukleotida u odnosu na savršeni uzorak tj. konsenzus. Pokazalo se da repeticije imaju značajnu regulatornu ulogu kao jedan od glavnih uzročnika evolucijskog razvoja [3 - 9] i zbog toga potrebna je njihova identifikacija i analiza. Tandemne repeticije smještene unutar regulatornog područja ili onog kodirajućeg su bitne za upravljanje funkcijom ili ekspresijom gena [10], ali su važne i za fenotipske varijabilnosti i bolesti kod viših eukariota [11, 12].

S napretkom tehnologije za sekvencioniranje genoma, razvijali su se različiti bioinformatički alati i algoritmi za bolju identifikaciju repeticija DNK sekvenci, a zajedno s njima su se mogle odrediti i periodičnosti višega reda. Algoritmi za identifikaciju repeticija mogu se podijeliti u dvije kategorije, fleksibilne statističke algoritme usporedbe nizova (od kojih je najpoznatiji Tandem Repeat Finder, TRF [13]) i algoritme za procesiranje signala (od kojih je najpoznatiji Spectral Repeat Finder, SRF [14]). Svi ti algoritmi za identifikaciju tandemnih repeticija imaju svoje nedostatke, koji su vezani uz njihove metode rada. Problem pronalaska takvih repeticija je kompleksan zbog nesavršenih kopija monomera (osnovne ponavljajuće jedinice u tandemu). Primjenom različitih algoritama moguće je dobiti i različite informacije vezane uz broj kopija, naročito ako se istražuju repeticije velikih duljina. Fleksibilni statistički algoritmi usporedbe nizova imaju probleme s identifikacijom repeticija u slučaju velikih supstitucija, umetanja i brisanja nukleotida kod dugačkih monomera (iznad 2 kb) te mogu davati nekoliko mogućih struktura iste sekvence. U slučaju algoritama za procesiranje signala mogu nastati numerički artefakti i slaba rezolucija spektralne analize.

Od tandemnih repeticija najviše su istraživani sateliti. Jedan takav primjer istraživanja su sateliti u ljudskom genomu, tzv. alfa sateliti duljine ~171 bp [15, 16, 17, 18], koji se nalaze u području centromere u svakom kromosomu. Također su se proučavali sateliti i u višim

primatima te su se među njima pokušavale naći sličnosti i različitosti u njihovoj strukturi u odnosu na ljudske alfa satelite. Ljudski alfa sateliti mogu se pojaviti u dva oblika, kao monomeri i kao periodičnosti višega reda tzv. HOR-ovi (koji su hijerarhijski organizirani u sekundarnu ponavljajuću jedinicu). Razlike između pojedinih monomera koji čine HOR su u rasponu od ~20% - do ~35%, dok su one između HOR kopija manje od 5% [19]. Za HOR-ove ljudskih alfa satelita smatra se da su nedavno nastali u ljudskom genomu pomoću mehanizma nejednolikog „crossing-overa“ [19, 20]. Pomoću tog mehanizma mogu se objasniti stvaranje i lokalna homogenizacija jedinica HOR kopija te razlike između različitih kromosoma u njihovoj veličini. Jedno od značajnijih rezultata dobivenih s računalnom metodom Global Repeat Map (GRM), vezano uz istraživanje HOR-ova, je identifikacija tzv. ljudskog akceleriranog područja (HAHOR) u neuroblastoma „breakpoint“ obitelji gena koje daje naznaku za razliku između razvoja ljudskog i mozga viših primata [21] te također razliku između Y kromosoma čovjeka i čimpanze [22] kao i ostalih kromosoma [23].

Algoritmi za identifikaciju tandemnih repeticija, osim na ljudima i primatima, su se primjenjivali i na drugim eukariotima [24, 25, 26]. Jedne od također dosta istraživanih sekvenci DNK su sateliti insekta *Tribolium castaneum*, tzv. malog brašnara, koji napada spremišta hrane, naročito ona s pohranjenim žitaricama [27, 28, 29, 30]. Genomi iz obitelji *Tenebrionid* insekata općenito imaju velike blokove pericentromernog heterokromatina, na čijem području se i nalaze sateliti koji se proučavaju radi boljeg razumijevanja njihove uloge i evolucije. Problem kod proučavanja satelita kod insekata, ali i drugih eukariota, je njihovo teško sastavljanje prilikom sekvencioniranja što je upravo i vidljivo na primjeru *T. castaneum* u čijem sekvencioniranom genomu je sastavljeno samo 0,3% [29] od ~35% genoma koji im pripada [31]. Prijašnja istraživanja, koja su koristila restrikcijske enzime [31], pokazala su da od struktura višega reda postoje dimeri, trimeri i pentameri koji su ukazivali na periodičnosti višega reda, kao što postoje kod sisavaca odnosno viših primata te čovjeka.

Svrha disertacije je dati prikaz svih identificiranih periodičnosti višega reda (HOR) pomoću računalne metode Global Repeat Map (GRM) u insektu *Tribolium castaneum* (slika 1). Za analizu ovog insekta potrebno je bilo preuzeti sekvencionirani genom TCAS 3.0. s javno web dostupne baze podataka NCBI (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AAJJ01). TCAS 3.0. sastavljen je od kontiga AAJJ01000001-AAJJ02009708 u kromosome CM000276-CM000285 („linkage grupe“ – označene s L1-L10 i napravljenih od 140 spojenih kontiga (scaffolda) koji predstavljaju 70% sekvencioniranog genoma te su sastavljeni s visoko rezolucijskim rekombinacijskim tehnikama

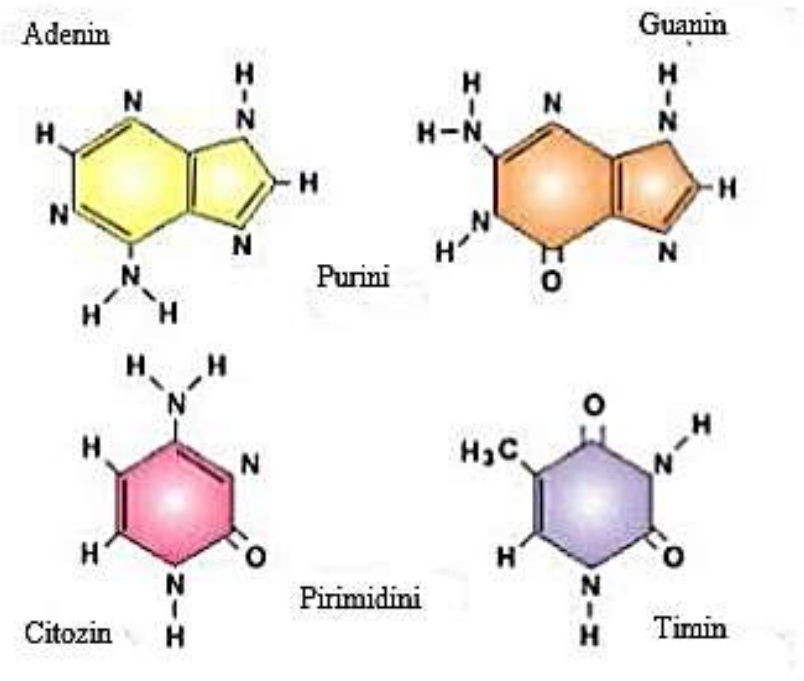
mapiranja pomoću bakterijskih umjetnih kromosoma i pomoću ekspresije „sekvence tag“ markera), u nepovezane višestruko – komponentne povezane kontige („unlinked multi-component scaffolds“ DS47665-DS497969) i nepoznate jednostruke spojene kontige („unknown singleton scaffolds“ GG694051- GG694051). Prednost primjene GRM metode za identifikaciju HOR-ova leži u simboličkom mapiranju sekvence u frekventnu domenu te stvaranju GRM dijagrama iz kojih se mogu pomoću dominantnih ključnih riječi identificirati tandemne, raspršene, pravilne i kompleksne periodičnosti višega reda u rasponu svih duljina fragmenata. Za razliku od drugih algoritama, metoda je robusna na supstitucije, brisanja i umetanja nukleotida te daje veliku mogućnost pri identifikaciji jednostavnih i kompleksnih struktura višega reda (HOR-ova). U kombinaciji s programom BLAST (Basic Local Alignment Search Tool) [32], možemo za identificirane repeticije pomoću GRM metode, vidjeti da li se one pojavljuju i kod nekih drugih eukariota te također za obrađeni genom nekog eukariota, odrediti da li se one nalaze unutar ili izvan genskog dijela DNK. Identifikacija HOR-ova u *T. castaneum* je zanimljiva iz tog razloga što se smatralo da su HOR-ovi općenito karakteristični za sisavce, pogotovo kod viših primata i čovjeka te da su oni nedavno nastali na evolucijskoj ljestvici.

Slika 1. Prikaz *T. castaneum*. Preuzeto sa: http://en.wikipedia.org/wiki/File:Tribolium_castaneum87-300.jpg#file.

Tandemne repeticije su zanimljive za istraživanje zbog svojih regulatornih uloga kao jedan od glavnih faktora evolucijskog razvoja (na primjer razvoj ljudskog mozga u odnosu na više primate) te primjeni u medici u otkrivanju uzroka različitih bolesti kao posljedicu nepravilne ekspresije gena. Periodičnosti višega reda, nastalih s nejednolikim „ crossing-over“ mehanizmom (koji objašnjava stvaranje i lokalnu homogenizaciju HOR-ova i ubraja velike varijacije između HOR-ova na homolognim kromosomima), omogućuju brzi evolucijski razvoj eukariota. Identifikacija HOR-ova, s obzirom da su upravo oni karakteristični za primate i čovjeka i smatra se da su nastali nedavno na evolucijskoj skali, u genomu insekta *T. castaneum* otvara nove mogućnosti istraživanja mogućih mehanizma evolucije kod vrsta nastalih prije sisavaca.

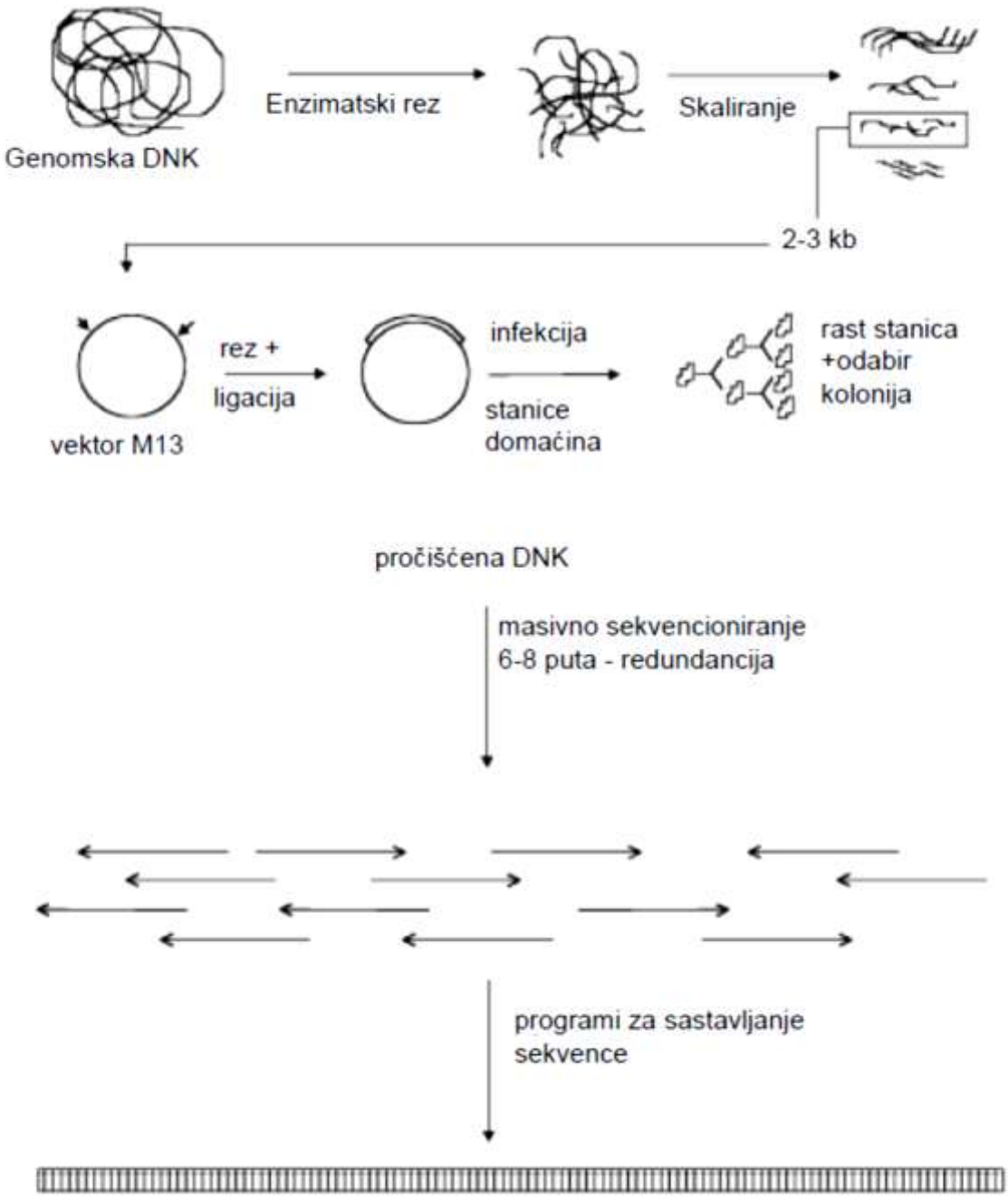
2. Literaturni pregled

Kako bi dešifrirali genome svih eukariota, a naročito genom čovjeka, razvijale su se metode sekvencioniranja DNK odnosno načini određivanja redoslijeda nukleotida, sastavljenih od purinskih i pirimidinskih baza (adenina - A, guanina – G, citozina - C i timina – T - slika 2), šećera i fosfata.

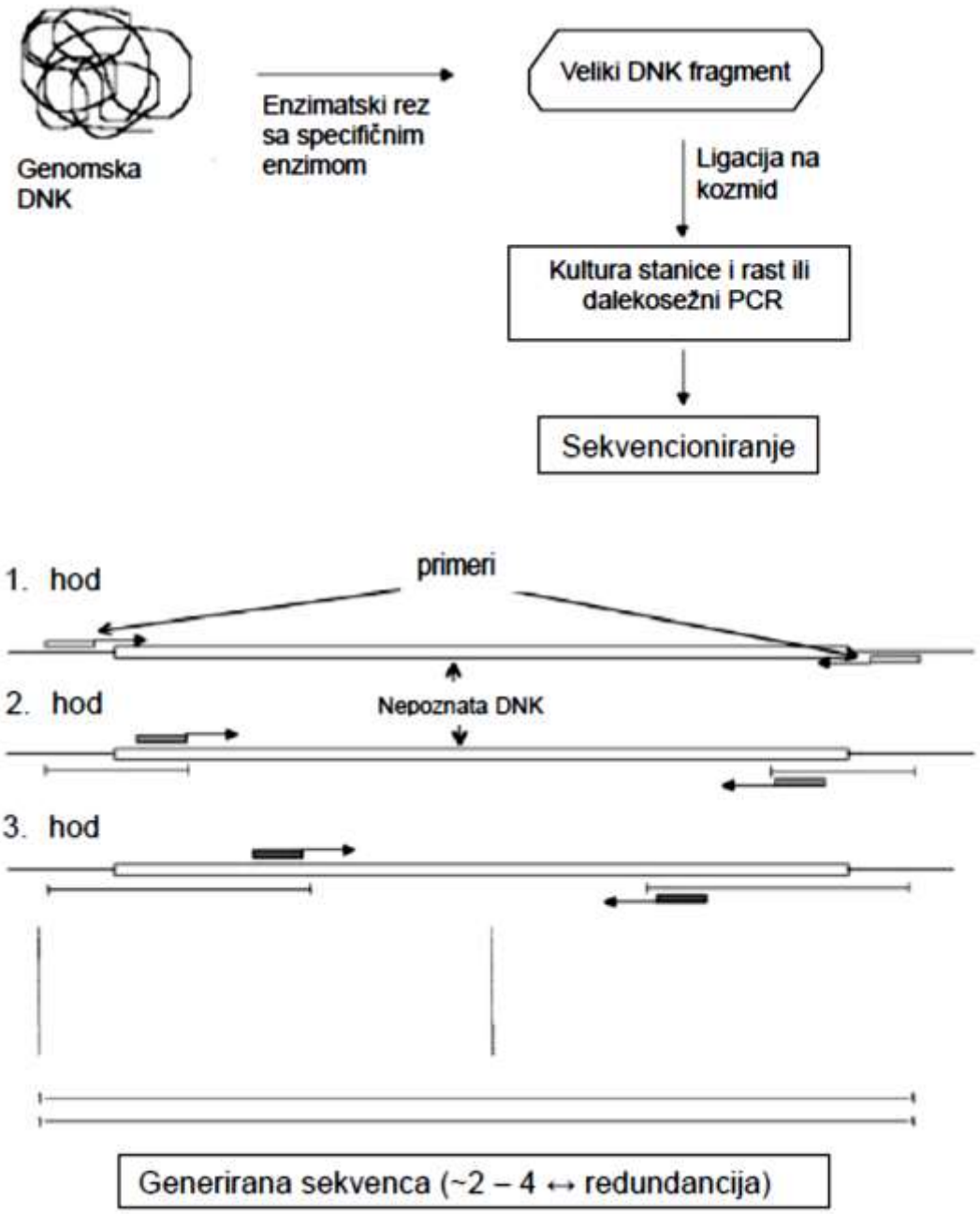


Slika 2. Građa baza. Preuzeto i prerađeno s http://javiciencias.blogspot.com/2012/10/biomolecules-our-earth-is-as-far-as-we.html.

Razvojem tehnika sekvencioniranja omogućio se napredak u različitim poljima znanosti, od arheologije, antropologije, genetike, biotehnologije, molekularne biologije i forenzike. Nobelovac Frederick Sanger sa suradnikom Coulsonom 1975. g. [33] je napravio prvu metodu sekvencioniranja zvanu „plus i minus“ koristeći *E. coli*. DNK polimerazu I i DNK polimerazu iz bakteriofage T4. Zbog neučinkovitosti metode, Sanger je sa suradnicima napravio novu metodu sekvencioniranja dvije godine kasnije [34], temeljenu na enzimskoj polimerizaciji, tzv. „chain termination method“. U ovoj metodi postoje dva glavna pristupa, nasumični (shotgun sekvencioniranje) i direktni (hod primera) pristupi [35] koji su prikazani na slici 3 i slici 4.



Slika 3. Pristup nasumičnog sekvencioniranja ili shotgun metoda. Različiti procesi prvo uključuju fragmentaciju DNK u fragmente duljina 2 – 3 kbp koji su zatim klonirani u vektore i umetnute u stanice domaćine za amplifikaciju. Nakon pročišćenja, DNK iz pojedinačne kolonije je sekvencionirana, a krajnji rezultat je dobiven s programima za sastavljane (poravnanje) sekvenci. Preuzeto i prerađeno s [36].



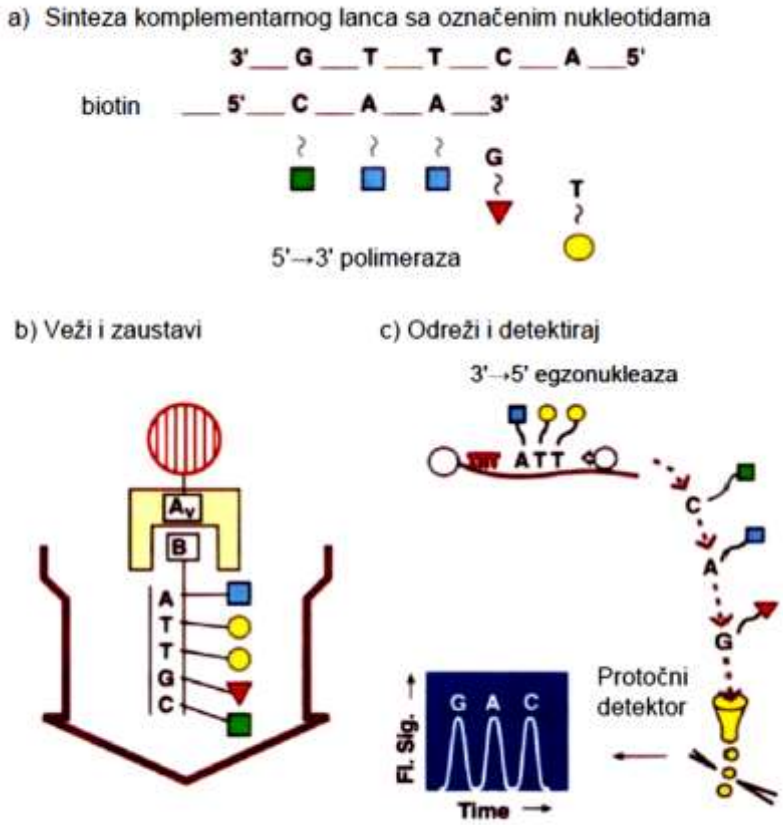
Slika 4. Pristup direktnog sekvencioniranja ili metoda hodanja primera. U ovoj metodi, DNK je izrezana na velike komade od ~ 40 kbp i umetnuta u kozmid za rast. Sekvencioniranje se radi u koracima počevši od poznate pozicije kozmida. Nakon što se urede podaci iz prvog kruga, određuje se novo mjesto primera u novoj stvorenoj sekvenci. Postupak se ponavlja dok hod ne dođe do suprotne početne točke. Preuzeto i prerađeno iz [36].

S vremenom se Sangerova metoda dorađivala kako bi se omogućilo čitanje dužih sekvenci (broj čitanja baza po hodu), kratko vrijeme analize, malu cijenu analize i bolju preciznost i to u enzimskoj tehnologiji, pripremanju samih uzoraka, DNK označavanju i separaciji fragmenata i analizi [36].

Osim Sangerove metode sekvencioniranja, razvijale su se i druge metode, kao što je Maxam i Gilbert metoda zasnovana na kemijskoj degradaciji [37], u kojima nije bilo potrebno subkloniranje i pročišćivanje uzoraka [38]. U kemijskim metodama neke od mana, pri sekvencioniranju, su nedovršene reakcije koje smanjuju duljinu čitanja fragmenata te sporost zbog pažnje koja se mora obratiti na rad s kemikalijama.

Treća metoda sekvencioniranja se zasniva na pirofosfatima, odnosno na DNK sekvencioniranju u realnom vremenu s detekcijom otpuštenih PPi (anion $P_2O_7^{4-}$, pirofosfata) tokom reakcije DNK polimerizacije [39, 40, 41].

Sekvencioniraje jedne molekule s egzonukleazama je još jedna metoda, koja omogućuje brzo sekvencioniranje velikih fragmenata DNK (40 kb) s brzinama od 100 do 1000 baza u sekundi [42] čiji princip rada je prikazan na slici 5.



Slika 5. Prikaz koraka u postupku sekvencioniranja jedne molekule zasnovan na laserski induciranoj fluorescenciji. Preuzeto i prerađeno iz [36].

Danas se upotrebljavaju i tehnike sekvencioniranja slijedeće generacije, čije se osobine mogu vidjeti u tablici 1.

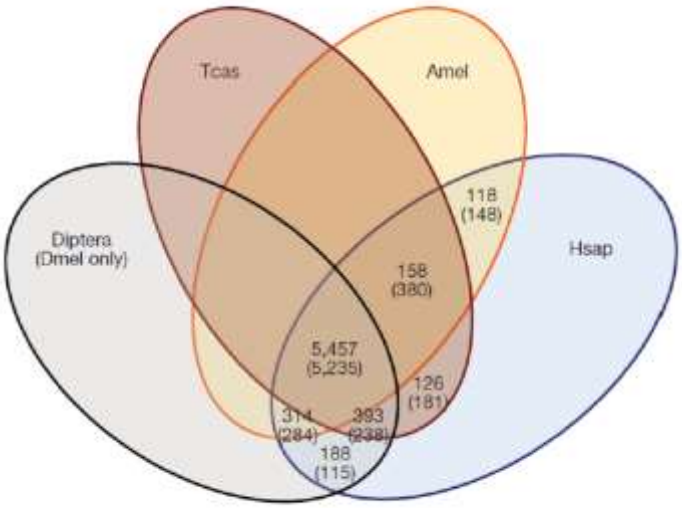
Tablica 1. Međusobna usporedba sekvencioniranja slijedeće generacije. Preuzeto i prerađeno s http://en.wikipedia.org/wiki/DNA_sequencing.

Metoda	Duljina očitanja	Točnost	Čitanje po hodu	Vrijeme po hodu	Cijena po 1 milijunu baza (u US$)	Prednosti	Nedostaci
Single-molecule real-time sekvencioniranje (Pacific Bio)	5,500 do 8,500 bp	99.999% konsenzusna točnost, 87% pojedinačna točnost	50,000 po SMRT stanici, ili ~400 megabaza	30 minuta do 2 sata	$0.33–$1.00	Najdulja dužina čitanja. Brza. Detektira 4mC, 5mC, 6mA.	Umjerena propusnost. Oprema je skupa.
Ionsko semikonduktorsko (Ion Torrent sequencing)	do 400 bp	98%	do 80 milijuna	2 sata	$1	Manje skupa oprema. Brza.	Greške homopolimera.
Pirosekvencioniranje (454)	700 bp	99.9%	1 milijun	24 sata	$10	Velika duljina čitanja. Brza.	Skupi hodovi. Greške homopolimera.
Sekvencioniranje sa sintezom (Illumina)	50 do 300 bp	98%	do 3 milijardi	1 do 10 dana	$0.05 do $0.15	Potencijal za visoki doprinos sekvencioniranja, ovisno o modelu sekvencera i primjeni.	Skupa oprema. Potrebne visoke DNK koncentracije.
Sekvencioniranje s ligacijom (SOLiD sequencing)	50+35 ili 50+50 bp	99.9%	1.2 do 1.4 milijardi	1 do 2 tjedna	$0.13	Niska cijena po bazi.	Sporija od drugih metoda. Problemi s palindromskim sekvencama.
Zaustavljanje lanca (Sanger sequencing)	400 do 900 bp	99.9%	N/A	20 minuta do 3 sata	$2400	Dugačka individualna očitanja. Velike primjene.	Skuplja i nepraktična za veće projekte sekvencioniranja.

Sa razvojem metoda sekvencioniranja, nastala je potreba za pohranom svih tih DNK sekvenci tj. stvaranja baza podataka genoma. Tri najbitnije baze podataka u koje se genomi pohranjuju su: DNA Data Bank of Japan (National Institute of Genetics [43]), EMBL (European Bioinformatics Institute [44]), GenBank (National Center for Biotechnology Information [45]), na kojima se mogu pronaći genomi svih sekvencioniranih organizama. Zajedno sa stalnim dodavanjem genoma u ove baze podataka, započinjani su i mnogi drugi projekti sekvencioniranja genoma različitih organizama. Jedan od značajnih primjera projekta je HGP (Human Genome Project [46]) koji su pokrenuli 1990.g U.S. Department of Energy (DOE) i National Institutes of Health (NIH) u trajanju od 15 godina. Prva draft sekvenca je objavljena

2001.g. [47], a sam projekt je završio 2003. godine. HGP je potaknuo pokretanje i drugih sličnih projekata. Kao rezultat HGP – a, otkriveno je oko 1800 gena odgovornih za neke bolesti te je stvoreno novih 350 biotehnoloških produkata koji su dospjeli u klinička istraživanja [48]. Zajedno s projektima sekvencioniranja drugih organizama omogućene su komparativne analize koje daju bolje razumijevanje strukture i funkcije gena i ujedno evolucijskih promjena između organizama [49].

Projekt koji je bitan za ovu disertaciju je projekt genoma insekta *Tribolium castaneum*, koji je jedan od organizama modela. On se koristi u svrhu općenitog istraživanja razvoja insekata. *T. castaneum* je jedan od najraširenijih štetnika koji napada pohranjenu hranu, naročito žitarice s čime se stvaraju milijardske štete u gospodarstvu, s obzirom da može preživjeti zahtjevne uvjete okoliša, zahvaljujući razvijenom organu sličnom bubregu. Zbog svojeg kratkog životnog vijeka, velike plodnosti, pogodan je za genetička istraživanja i to pretežno u razvojnim studijama. Da je dobar odabir za proučavanje u komparativnim studijama, vidljivo je iz slike 6, na kojoj se vidi da 126 ortolognih grupa gena, prisutnih u malom brašnaru i čovjeku, su izostavljeni iz drugih sekvencioniranih genoma insekata od kojih je 44 prisutno u svim kralježnjacima [50] od ukupnog konsenzusnog seta od 16404 gena.



Slika 6. Zajednički ortologni geni u genomima insekta i čovjeka. Preuzeto iz [50].

Od 16404 gena, njih 47% su ansestralni koji se mogu pratiti u insektima i kralježnjacima te 9% ortlognih gena se nalaze samo kod insekata [50]. Proučavanjem kemo osjetilnih gena i gena detoksikacije, traže se bolji načini izrade pesticida u svrhu zaustavljanja širenja ovog insekta.

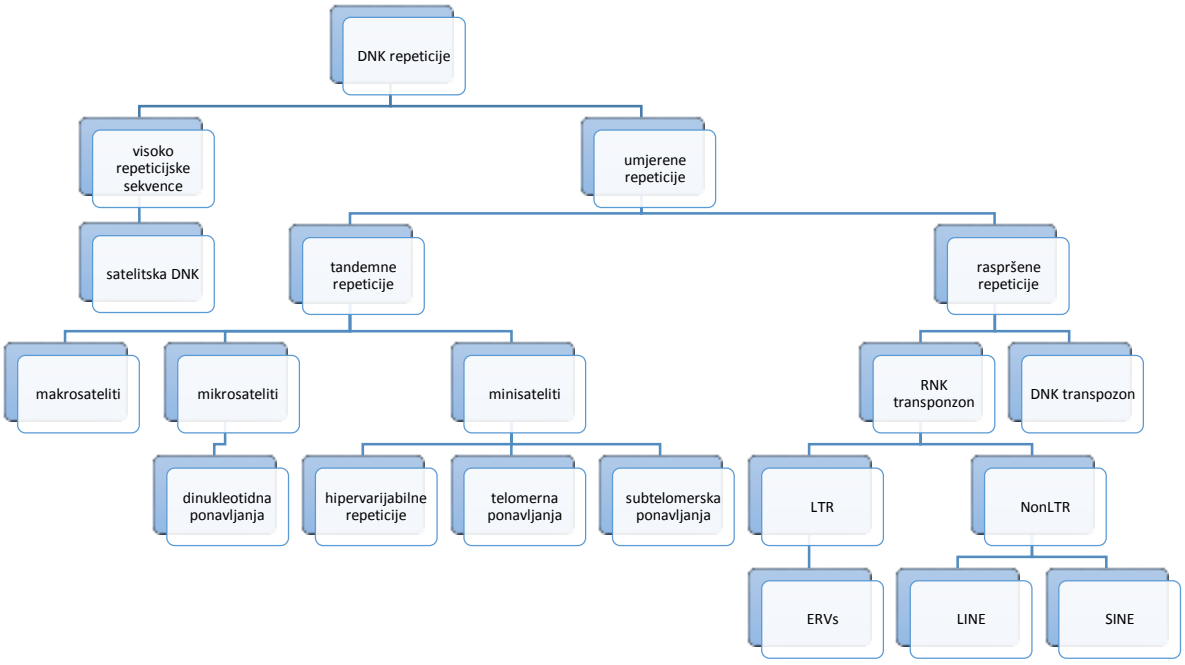
Zajedno uz razvoj metoda sekvencioniranja, počeli su se razvijati i bioinformatički alati, koji bi mogli analizirati i interpretirati podatke dobivene s projektima sekvencioniranja genoma, i razvoj bioinformatike. Sama bioinformatika je interdisciplinarna, i ona uključuje matematiku, fiziku, biologiju te računalne znanosti za razne primjene u medicini (npr. otkrivanje i razvoj lijekova kao što je Gleevec, koji interagira s abnormalnim proteinom stvorenim u kroničnoj mieloidnoj leukemiji [51]). Osnovni bioinformatički alati služe za analizu i DNK sekvence, ali i proteina uz javno dostupne baze podataka. Interpretacija podataka pomoću bioinformatičkih alata, od analize DNK sekvence ide sve do analize ukupnog broja proteina i određivanja njihove strukture (proteotomika) do analize mRNA (transkriptomika), analiza genetske varijacije i ekspresije, predviđanje i detekcije regulatornih mreža gena i njihove dinamike, analize molekularnih puteva kako bi se razumjele bolesti vezane uz gene, dizajniranje primera za predviđanje funkcije produkata gena te za simulacije u modeliranju dinamike stanica. Neki od bioinformatičkih alata prikazani su u tablici 2.

Tablica 2. Popis nekih bioinformatičkih alata.

Alati	Opis
Basic Local Alignment Search Tool (BLAST) [52]	Pronalazi slična područja između bioloških sekvenci s računanjem statističke značajnosti pogodaka.
Electronic PCR (e-PCR) [53]	Računalna metoda za identifikaciju sekvencijski označenih mjesta STS (sequence tagged sites) u DNK sekvenci.
Open Reading Frame Finder (ORF Finder) [54]	Grafička analiza koja traži ORF u sekvenci.
Conserved Domain Search Service (CD Search) [55]	Identificira očuvane domene prisutne u proteinskoj sekvenci.
ENSEMBL [56]	Identificira gene i druga svojstva sekvence.
COMBOSA3D – (Coloring Of Molecules Based On Sequence Alignment) [57]	Bojanje molekula prema poravnanju sekvence.
SeWeR: Sequence analysis [58]	Integrirani portal za česte web servise u bioinformatici.

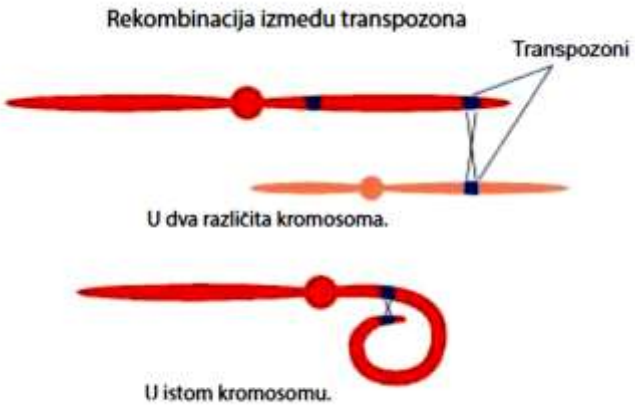
Jedna grupa bioinformatičkih alata se posebno počela razvijati za identifikaciju repeticija u DNK sekvenci. Postoje razlike u kopijama repeticija u obliku točkastih mutacija, umetanja, brisanja, zamjena nukleotida koje su nastale s različitim mehanizmima. Pomoću dinamike repeticija, može se doći do evolucijskih razlika koje se mogu iskoristiti za

proučavanje mutacija i prirodne selekcije [59]. Repeticije se mogu podijeliti u dvije glavne skupine [59] koje se dalje mogu dijeliti kako je prikazano na slici 7.



Slika 7. Shema različitih vrsta repeticija. Preuzeto i prerađeno iz [59].

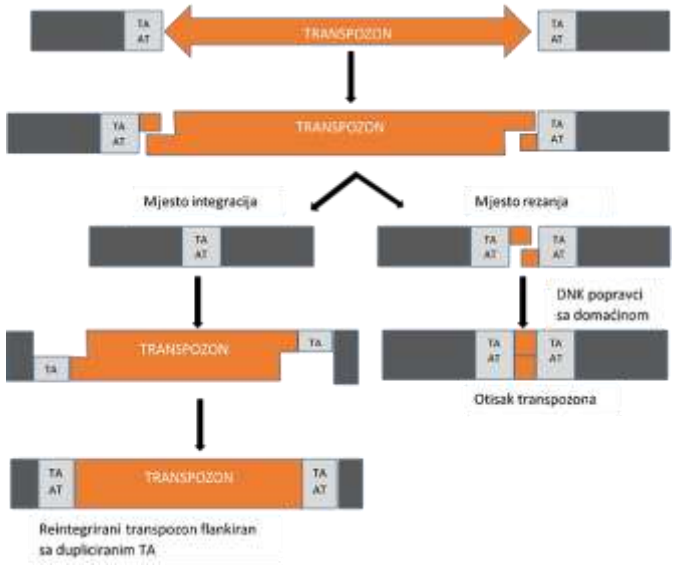
Raspršene repeticije su nastale s mehanizmom premještanja s jednog mjesta na drugo unutar genoma – transpozicijom na dva načina, s RNK transpozonom i s DNK transpozonom s čime se može započeti proces rekombinacije [60].



Slika 8. Prikaz započinjanja procesa rekombinacije na istom ili različitim kromosomima. Preuzeto i prerađeno iz [60].

RNK transpozoni se mogu dalje podijeliti na LTR (jedna podklasa su ERVs – endogeni retrovirusi) koje imaju potencijalnu regulatornu ulogu i kodiraju mRNK molekule [60]), i NonLTR elementa (LINE – Long Interspersed Nuclear Elements, SINE – Short Interspersed Nuclear Elements). Jedne od najpoznatijih SINE sekvenca u čovjeka su tzv. alu sekvence, duljine 350 bp, s jednim milijunom kopija, a od LINE je L1 s duljinom od 7 – kbp i više od 500 000 kopija (što čini 17% ljudskog genoma) [61].

DNK transpozoni su bitniji za prokariote i oni se mogu pomicati po genomu s mehanizmom „izreži i zalijepi“ (cut-and-paste) koji se sastoji od nekoliko koraka (slika 9).



Slika 9. Izreži i zalijepi mehanizam (cut-and-paste). Na slici se vidi da je transpozon izrezan s jednog mjesta i zalijepljen na drugo na TA ciljnoj dinukleotidi koji je dupliciran i popravljen s DNK mehanizmom popravka. Preuzeto i prerađeno iz [62].

Općenito transpozoni mogu uzrokovati genetske izmjene zbog transpozicijskih procesa, kao što je vidljivo na DNK transpozonima koji mogu deaktivirati ili izmijeniti ekspresiju gena umetanjem u introne, egzone ili regulatorna područja [62].

Tandemne repeticije se sastoje od uzoraka nizova baza tzv. monomera, poslaganih linearno jedan iza drugoga u genomu i mogu zauzimati do nekoliko tisuća ponavljajućih jedinica (monomera). Prema duljini monomera mogu se podijeliti na mikrosatelite (1 do ~6 bp), minisatelite (~6 do ~100 bp), satelite (~100 do ~2 kb, na slici 7 su predstavljeni kao visoko repeticijske sekvence) i makrosatelite (veći od ~2 kb).

Istraživanja su pokazala da makrosateliti imaju strukturalnu i regulatornu ulogu i da varijacije u njihovoj duljini mogu uzrokovati bolesti kao što je facioskapulohumeralna mišićna distrofija, kao i nekih drugih [63].

Od otkrića prvih minisatelita, Jeffrey 1985. godine, u području introna gena mioglobina, slične strukture su se počele otkrivati i u drugim organizmima. Većina minisatelita su GC bogata i većina mutacija se sastoji od brisanja ili umetanja monomera (odnosno osnovne jedinice ponavljanja). Za sada se zna da su neki minisateliti bitni zbog njihovih asocijacija sa strukturalnim svojstvima kromosoma [60] kao što su telomerska ponavljanja bitna za DNK replikaciju i održavanje integriteta kromosomskih krajeva, motiva 5′-TTAGGG-3′ ponovljenog u stotinama kopija. U genomu čovjeka, minisateliti su smješteni u krajevima subtelomernog područja, dok su kod nekih drugih organizama, kao što su miš i šupljorošci, rasprostranjeni u cijelom genomu [64]. U slučaju *T. castaneum*, motiv je (TCAGG)n koji se pojavljuje u svim kromosomima [65]. Tokom mejoze minisateliti mijenjaju svoju ukupnu duljinu i sastav repeticija [66] te su korisna za mapiranje genoma [67].

Subtelomerska ponavljanja nalaze se u području između telomera i kromatina. Područje subtelomere može sadržavati više tipova repeticija, a neke od njih su segmentne duplikacije, sateliti i sekvence oblika (TTAGGG)n. Segmentne duplikacije su ponavljanja duljine ~1kb sa sličnosti kopija većim od 90% i tvore 5% eukromatina u ljudskom genomu [68]. Za njih se zna da su one zapravo spojeni duplikoni kod čovjeka te da su to područja genomske nestabilnosti kao i brza izmjenjujuća područja genomske sekvence smještene blizu telomere. Neke od ovih sekvenci mogu biti specifične za pojedine kromosome, dok su druge prisutne na svim ljudskim kromosomima [68]. Za nastanak ovih segmentnih duplikacija prvo se moraju dogoditi primarne višestruke duplikacije kopija koje se zatim agregiraju i nakon agregacije prolaze kroz sekundarne duplikacije. Segmentne duplikacije su najviše istraživani u genomu čovjeka i *D. Melanogaster*. Prvi rezultati vezani uz segmentne duplikacije u insektima, vezanim uz istraživanje asocijacija mikrosatelita sa segmentnim duplikacijama, pokazala su da u genomu *T. castaneum* postoji 7 mikrosatelitskih parova u asocijaciji sa segmentnim duplikacijama [69] (prikazano u tablici 3).

Tablica 3. Broj repeticijskih mikrosatelitskih parova (rPM) i mikrosatelitskih parova u asocijaciji sa segmentnim duplikacijama (mSD) u insektima. Preuzeto i prerađeno iz [69].

Vrsta	rMP	mSD	Postotak (%)
A.aeg	4644	1023	22.03
A.gam	6674	891	13.35
A.mel	2861	193	6.75
A.pis	1961	40	2.04

B.mor	4920	545	11.08
C.qui	4828	573	11.87
D.ana	1731	108	6.24
D.ere	550	49	8.91
D.gri	21863	790	3.61
D.mel	2618	290	11.08
D.moj	24975	686	2.75
D.per	4579	184	4.02
D.pse	2125	97	4.56
D.sec	2062	57	2.76
D.sim	902	10	1.11
D.vir	16699	628	3.76
D.wil	11936	532	4.46
D.yak	1000	30	3
N.vit	23314	2759	11.83
T.cas	789	7	0.89

Mikrosateliti tzv. SSR (Short Sequence Repeats) se mogu pronaći isto kao i ostali tipovi repeticija, i u kodirajućem i češće u nekodirajućem dijelu DNK sekvence. U nekim slučajevima se mikrosateliti češće nalaze u egzonima (pozitivna selekcija) što može upućivati na njihovu funkciju. Oni imaju različite uloge, npr. asocijacija sa segmentnim duplikacijama te regulatornim ulogama genske ekspresije (zbog velike brzine mutacija) te daju mogućnosti nastajanja mehanizma odmotavanja DNK [70], zbog varijacije u broju kopija mogu utjecati na transkripciju [71, 72], ali također mogu poslužiti za izradu genetskih profila za svaku osobu (jedinku), pošto ne postoje dvije osobe s jednakim brojem kombinacija varijanti njihovih duljina, odnosno imaju značajnu ulogu u forenzici [60]. Mikrosateliti su bitni za proučavanje bolesti [73] koje mogu uzrokovati, i općenito su se sve tandemne repeticije počele intezivnije istraživati s otkrićem bolesti uzrokovanih s polimorfizmom broja kopija kao što je to slučaj kod Huntingtonove bolesti, bolesti koja utječe na rad mišića i uzrokuje kognitivno propadanje i

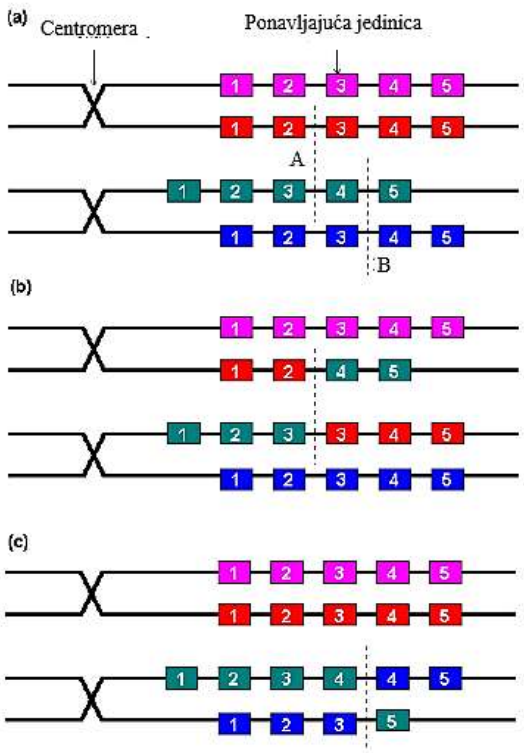
promjene u ponašanju. Promjene broja kopija trinukleotida CAG (kodira amino kiselinu glutamin) u Huntingtinovom genu (HTT), koji kodira protein Huntingtin (Htt), uzrokuje bolest (tablica 4, [74]).

Tablica 4. Klasifikacija trinukleotidnih ponavljanja i status bolesti ovisi o broju kopija CAG. Preuzeto i prerađeno[74].

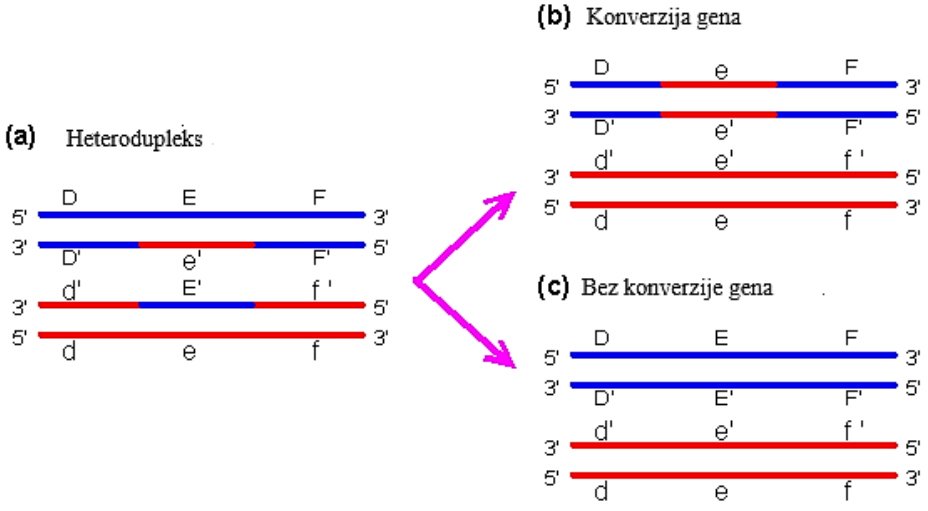
Broj kopija	Klasifikacija	Status bolesti	Rizik za potomke
<26	Normalan	Nema	Nikakav
27–35	Umjeren	Nema	<<50%
36–39	Smanjeno prodiranja	Može i ne mora imati	50%
40+	Potpuno prodiranje	Ima	50%

Istraživanja su pokazala da se u genomu *T. castaneum* nalazi više od 12000 mikrosatelita, od kojih su trinukleotide najraširenije, među kojima je najduži motiv $(AAT)_{195}$ [75]. Većina mikrosatelita u *T. castaneum*, nalaze se u intergenskom dijelu (63%) ili u intronima (20%). Pomoću mikrosatelita (509 polimorfnih) iz ovog insekta, dobili su se markeri koji se koriste u populacijskim istraživanjima [50].

Od tandemnih repeticija, još postoje i sateliti koji su na slici 7 predstavljene kao visoko repetitivne sekvence. Ovaj tip tandemnih repeticija najčešće je smješten u pericentromernom ili telomerskom heterkromatinskom području [76], interagira sa specifičnim proteinima te sudjeluju u organizaciji kromosoma. Veliki broj istraživanja, vezana uz analizu satelita, su se provodila na usporedbi između blisko povezanih vrsta kao što su čovjek i čimpanza [21, 22, 23] te da su promjene u broju kopija i mutacijama unutar njih bitne za njihovu evoluciju, unutar pojedine vrste i između različitih vrsta [77]. Sateliti, veličine ~ megabaza, nastaju s procesom koordinirane evolucije („concerted evolution“) s kojom dolazi do homogenizacije mutacija unutar genoma i njihove fiksacije u populaciji. S procesom „molecular drive“ [78] dolazi do izmjene genetskog sadržaja pomoću procesa okretanja (DNK „turnover mechanism“) koji ima utjecaj na strukturu DNK molekule (točkaste mutacije, kromosomske anomalije (inverzija, translokacija, rekombinacija), genomske anomalije, i genska ekspresija) i fenotipske osobine [79]. Pomoću procesa nejednolikog „crossing-over-a“ (slika 10) i „gene conversion“ (slika 11) mogu nastati razlike između broja kopija satelita [80] i drugih mutacija unutar samih satelita.



Slika 10. Proces nejednolikog „crossing-over-a“ i izmjena sestara kromatida. (a) Poravnanjem sestrinskih kromatida tokom mejoze, ne moraju se sve repeticije poravnati savršeno. (b) Slamanje lanca na nesestrinskoj kromatidi (duž linije A) uzrokuje nejednoliki „crossing-over“, stvarajući različiti broj kopija repeticija. (c) Slamanje lanca na sestrinskoj kromatidi (duž linije B) također uzrokuje nejednoliki broj kopija. Preuzeto i prerađeno s [81].



Slika 11. Nastanak konverzija gena - „gene conversion”. a) Heterodupleks. b) DNK koristi segment *e'* za mehanizam popravka odnosno konverziju gena (označeno plavom bojom). c) Obje DNK molekule koriste svoju originalnu sekvencu za popravak bez genske ekspresije. Preuzeto i prerađeno s [81].

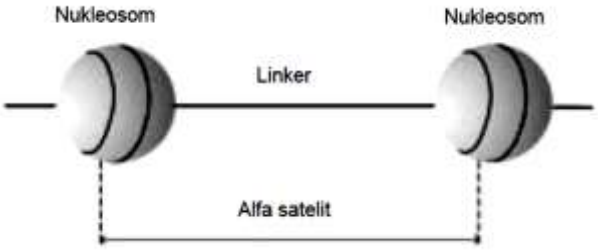
Preko evolucije satelitske DNK i praćenjem njihove dinamike došlo se do zaključka da postoji inverzna korelacija između brzine nejednolikog „crossing-overa“ i očuvanosti satelita među vrstama [82] te na smanjenje ili povećanje njihovog broja kopija s različitim mehanizmima (pomaknuto sparivanje lanaca DNK - replication slippage [83], kružne replikacije - rolling circle replication [84]) u kratkom evolucijskom razdoblju.

U jednoj vrsti mogu postojati više različitih satelita koji se tokom evolucije mogu u bliskim vrstama nalaziti u sličnom broju kopija ili se mogu značajno razlikovati [85, 86, 21, 22, 23] kao i postojanje biblioteka satelita [87]. Svaki od satelita je nezavisna jedinica koje s „turnover“ mehanizmom mogu postati specifični za pojedine kromosome (gdje mogu postojati podobitelji satelita) ili mogu nastati novi s velikom preraspodjelom sekvence ili s npr. umetanjem dodatne sekvence (kod *Tribolium Madens* [88]). Razlike između sekvenci nastaju s mutacijama postupno, pri čemu u cijeloj toj preraspodijeli neki sateliti mogu ostati očuvani kroz široki raspon vrsti kao što je „dodeca“ satelit koji je očuvan od *D.melanogaster* do čovjeka [89]. Njihova proučavanja daju nam mogućnost razumijevanja njihove dinamike i evoluciju vrsta.

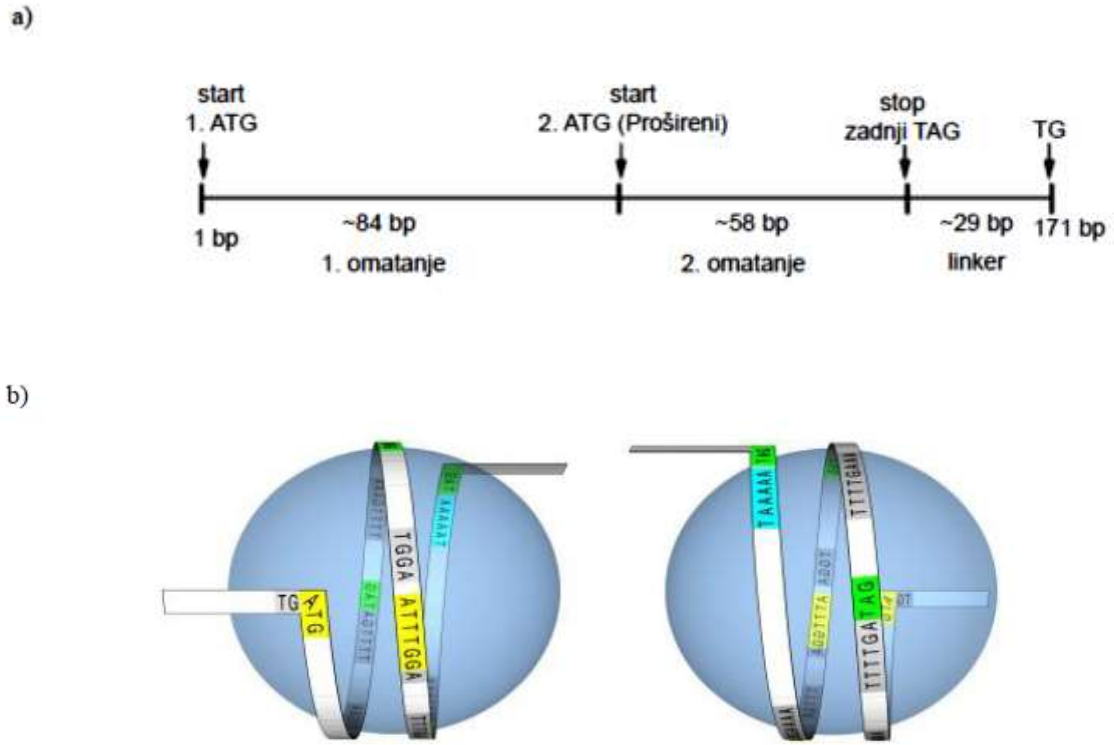
Komparativne studije su pokazale da u rodu *Tribolium* postoje zajednička svojstva u satelitima kao što je 20 - 42 bp s ~95% A - T nizom nukleotida i tercijarna struktura koja omogućuje gusto pakiranje DNK i proteina u heterokromatinu [77].

Sateliti koji su se očuvali među vrstama milijunima godina (60 Mya), ukazuju na postojanje *cis* regulatornih elemenata kod eukariota kao i sudjelovanje u formiranju centromere (odgovorne za pravilnu segregaciju kromosoma u mitozi i mejozi), kao što je to slučaj sa satelitima kod insekta *Palorus ratzeburgii* i *Palorus subdepressus* [90] i ljudskih alfa satelita očuvanih i kod kokoši i ribe zebrice [91]. Jedan primjer očuvanog motiva među satelitima je CENP-B box motiv, duljine 17 bp -TTTCGTTGGAAGCGGGA, kod ljudskih alfa satelita, a u drugim vrstama se nalaze njemu slični motivi (*D. Melanogaster*, Cid motiv [92]). Alfa sateliti duljine ~171 bp imaju bitnu ulogu u formiranju heterokromatina i služe kao gradivni elementi centromere. Sateliti u području centromere su brzo evoluirajući i iz tog razloga motivi, kao što je CENP-B box, također se moraju brzo prilagoditi kako bi održali svoju funkciju pri vezanju sa svojim veznim proteinom. Kod alfa satelita, njih 23% ima funkcionalni CENP-B box i istraživanja su pokazala da takozvane periodičnosti višega reda (HOR) su bitna za njihovo pravilno omatanje oko histona [93] kao što se vidi na slici 12. Dodatnim analizama pokazalo se da kod omatanja alfa satelita oko proteina postoji unutrašnja struktura, okarakterizirana s proširenim nakupinama trinukleotidnih ponavljanja (CLT – codon like trinucleotides) gdje predominira TGA s multipliciranim T i A nukleotidama [94]. U usporedbi s ostalim

sekvenciranim genomima primata, zaključilo se da se oni međusobno razlikuju [94] upravo u raširenosti start/stop ekstenzija. Na slici 13 se može vidjeti utjecaj CLT-ova na omatanje alfa satelita na proteine.



Slika 12. Shematski prikaz omatanja alfa satelita oko centromernih proteina (CENP). Preuzeto i prerađeno iz [93].



Slika 13. Omatanje alfa satelita. a) shematski prikaz start/stop CLT-ova kod alfa satelita. b) shematski prikaz omatanja alfa satelita oko nukleosoma. Preuzeto i prerađeno iz [95].

Kod insekata, ali i ostalih eukariota, sekvencioniranje genoma je uspješno napravljeno na eukromatinskom dijelu, dok je područje heterokromatina manje pokriveno radi repeticija koje se u njemu nalaze te ih je i teže sekvencionirati pomoću tehnika kao što je PCR (Polymerase Chain Reaction) zbog teškoća u kloniranju i sastavljanju sekvence. Samo heterokromatinsko područje je važno za analizu zbog sudjelovanja u pravilnoj kromosomskoj segregaciji tokom mitoze i mejoze. Za satelite, kao glavne elemente heterokromatina,

pretpostavilo se da imaju regulatorne uloge te da kod *Schizosaccharomyces pombe* njihovi transkripti mogu sudjelovati u RNAi – vođenom sastavljanju heterokromatina [96].

Sateliti kod insekata se mogu klasificirati kao jednostavni i kao kompleksni, prema njihovim duljinama repeticijske jedinice – monomera. Kod insekata te dvije veličine opadaju u prvi raspon od 140 bp do 190 bp i u drugi raspon od 300 bp – 400 bp [97], ali se među njima mogu naći i iznimke kao što su sateliti duljine od 24 bp kod *Musca domestica* (domaća muha) [98] pa sve do 2.5 kbp satelita u *Monomorium Subopacum* [99]. Općenito kao i kod eukariota, sateliti mogu biti specifični za pojedini insekt (*Gryllus bimaculatus* – cvrčak [100]), a neki mogu biti rašireni među slabije povezanim vrstama u manjem broju kopija (*Drosophila ambigua, D. tristis, i D. obscura* [101]) koje se mogu detektirati sa PCR metodom (Polymerase Chain Reaction). U svim pregledanim vrstama insekata, sateliti postoje u svim kromosomima [102], odnosno oni mogu biti specifični za pojedinu vrstu ili mogu biti specifični za pojedini kromosom (*D. Melanogaster* [103]) te također mogu biti specifični za kromosome za određivanje spola.

Periodičnosti višega reda (HOR – ovi) se također mogu pojavljivati i kod insekata, gdje monomeri mogu biti značajno različiti. Kod insekata prijavljene su neke HOR strukture i to kod špiljske bube *Pholeuon proserpinae* 532 bp HOR (2 tipa monomera duljine 266 bp [104]), *Chrysolina carnifex (*čije repeticije se mogu pojaviti kao monomeri - 211 bp, dimeri i trimeri napravljeni od 6 različitih tipova monomera 211 bp i čiji svaki tip ponavljanja je izmiješan u području heterokromatina [105]), kao i 1061 bp HOR iz *Tribolium brevicornis* (napravljen od dva 470 bp s dodatna dva izmjenjujuća segmenta od 56 bp i 65 bp [106]).

Transkripti satelita insekta RNAi, isto kao i kod kralježnjaka, mogu sudjelovati u formaciji kromatina te ih najviše ima kod ranih embrija što upućuje na dramatične promjene heterokromatina u razvoju jedinke, npr. kod insekta *Drosophila melanogaster* [107] te da neki RNAi transkripti satelitske DNK u jajnicima mogu uzrokovati prigušeno stanje centromernih i pericentromernih ponavljanja [108]. Za neke transkripte satelita se pokazalo da postoji veza s intronima mega gena (nekoliko megabaza) u Y kromosomu kod *D. melanogaster, D. hydei i D. Eohydei* [103] koji imaju ulogu u plodnosti. U špiljskom cvrčku *Dolichopoda schiavazzii,* transkripti 500 bp satelita mogu funkcionirati kao ribosomi bez razjašnjene fiziološke uloge [109]. Transkripcija satelitske DNK može započeti u satelitima u kojima se mogu nalaziti transkripcijski regulatorni elementi za RNK polimerazu II i III kao što je to slučaj kod satelita *Diadromus*-a [110].

Istraživanja su pokazala da transponirajući elementi (TE) mogu imati bitnu ulogu u nastanku, umnožavanju i homogenizaciji satelita kod insekata [111] kao i moguću pretvorbu TE-ova u funkcionalno područje telomere i centromere (kod *melanogaster* grupi vrsta očuvani transpozon BEL [112] (za koji nije jasno da li je očuvan zbog nedavne insercije u satelit ili je očuvan kao posljedica funkcionalnog ograničenja zbog aktivnosti centromere) i *C. Capitata* - sličan transpozon BEL [113]).

Istraživanja satelita kod insekata su pokazala da varijacije unutar njih mogu biti 1-13% [97], ali mogu biti jako male (*Eyprepocnemis plorans 100%* [114]) ili jako velike (68% sličnosti kod *Reticulitermes taxa* - termita *[115]).* Održavanje duljine monomera kod satelita je bitno zbog pozicioniranja nukleosoma, kondenzacije heterokromatina, i funkciji centromere [92] kao i zbog održavanja struktura višega reda [116], iako mogu postojati razlike među njima u blisko povezanim vrstama [96].

Kod *Tribolium* vrsti sateliti imaju promjenjiva i očuvana područja kao i nekoliko zajedničkih karakteristika kao što su kratka obrnuta ponavljanja u blizini A - T niza baza, nenasumična ponavljanja A-T nizova baza i motiv sličan CENP-B box-u [106], koji se također pojavljuje u drugim insektima [117] i mogu biti rezultat selekcijskog pritiska [18]. Drugi satelitski DNK dijelovi genoma mogu biti i u interakciji s proteinima koji podilaze adaptivnu selekciju kao što je to CENH-3 [118, 119].

Kod insekata sateliti su mogli nastati od kratkih ponavljajućih uzoraka pomoću koordinirane evolucije, kao što je slučaj kod *D. virilis, D. simulans* i *D. Melanogaster* uzorak 7-9 bp dugačak [120], ali mogu nastati i od kompleksnijih ponavljanja s drugim mehanizmima [76] kao što su duplikacija, inverzija, insercija kod repeticija u *T. brassicae* [121].

Proučavanja satelita insekata pokazala su da postoje očuvani evolucijski segmenti bez obzira na heterogenost sekvence, kao što su duljina monomera i različiti očuvani motivi koji mogu sudjelovati u vezanju proteina potrebnih za epigenetske promjene i funkciju centromere [122]. Bitna uloga kod interakcije proteina i satelitske DNK kod insekata je ona za oblikovanje heterokromatina i moguće kontrole ekspresije gena [18, 123, 124], odnosno upućuje na postojanje selekcijskih ograničenja. Kod insekta potrebna su još brojna istraživanja njihovih satelita. Zbog kratkog životnog vijeka, istraživanja u insektima su pogodna za komparativne analize radi boljeg razumijevanja njihove funkcije i evolucije.

Sekvencionirani genom *Tribolium castaneum*, u potrazi za repeticijama je istraživao mali broj znanstvenika [27, 29, 31, 50, 75]. Većina istraživanja je napravljena na satelitima

TCAST i otkrivanju njihovih uloga kod razvoja *T.castaneum* kao i u komparativnim analizama sa nekim drugim insektima. Wang i suradnici su 2008 g. istraživali sekvencionirani genom *T. castaneum kako* bi identificirali repeticije, pošto se znalo preko reasocijacijske kinetike da je 42% genoma u nekom obliku repeticija [125, 126] (što je slično razini ljudskog genoma), koristeći alate TRF, TEpipe and RepeatScout. Njihovo istraživanje je pokazalo da je 30% sekvencioniranog dijela genoma (Tcas 2.0) sastavljeno od repeticija, od kojih je 17% u obliku tandemnih repeticija, a ostali dio pripada raspršenim repeticijama (transpozoni – 5%). Od 31 visoko repeticijskih područja dužih od 100 bp, identificiranih s Repeat Scout alatom, od njih 65% visoko repeticijskih područja i 75% transponirajućih elemenata se nalaze u kromosomima koji čine 70% sastavljenog genoma. Kao i kod drugih eukariota, područje pericentromernog heterokromatina kod *T. castaneum* sadrži veliki broj repeticija koji može zauzimati od 25% - 58% genoma [127]. Broj tandemnih repeticija koje su dobili Wang i suradnici se nalaze u tablici 5.

Tablica 5. Broj identificiranih tandemnih repeticija u Tcas 2.0. s TRF-om. Preuzeto i obrađeno iz [29].

Tribolium castaneum	**Broj baznih parova (bp)**	**Postotak genoma**	**Broj mjesta**	**Prosječna gustoća (mjesta/Mb)**
Mikrosateliti	591 105	0,4	17 328	114
Minisateliti	3 112 304	2,1	120 474	796
Sateliti	3 775 523	2,5	4 272	28
Ukupni broj tandemnih repeticija	7 478 923	4,9	142 074	939
Genom	151 333 735			

Njihova istraživanja s Repeat Scout-om, koji je maskirao 25% sekvencioniranog dijela genoma su dala naslutiti da postoje nove repeticije u genomu *T. castaneum* koje još treba identificirati. Biblioteka dobivena s Repeat Scoutom sadrži 4 475 obitelji repeticija ukupne duljine 1.41 Mb za koje su pretpostavili da pripadaju nekoj vrsti satelita koje su najvjerojatnije bogate AT bazama u odnosu na transpozone, koji su bogati sa GC bazama. Repeticije koje nisu

transponirani element, su podijelili prema postotku genoma koji zauzimaju tri klase – visoku (više od 0,1%), srednju (između 0,01% i 0,1%) i nisku (manje od 0,01%) kao što je prikazano u tablici 6. U sekvencioniranom genomu za TCAST satelite su vidjeli da zauzimaju samo 0.3%, što je manje od eksperimentalno određenog postotka (prvo procijenjeno da je to 17%, dok su novija istraživanja pokazala da se radi o ~35% genoma [31]).

Tablica 6. Analiza *Tribolium* genoma s Repeat Scout-om. Preuzeto i prerađeno iz [29].

Klasa ponavljanja	Ukupna duljina obitelji repeticija	Broj obitelji repeticija	Postotak biblioteke Repeat Scouta	Postotak genoma	Duljina ponavljajuće obitelji (bp)	Prosjek ponavljajuće obitelji (bp)	Raspon broja kopija u obitelji	Prosječan broj kopija u obitelji
HighA	26,1	31	1,9	7,1	160-1771	841	323 - 4337	1,368
Mid	220,3	304	15,6	7,4	67-4881	725	11-1746	204
Low	738,2	3237	52,3	4,7	51-4520	228	3-215	14
HighB	4,6	5	0,3	1,6	982-1277	921	432-3531	1306
360 bp satelit	0,4	1	0,2	0,3	-	-	1122	-
TS	406,2	896	28,9	4,4	51-11289	453,3	3-2471	27

52% od nesastavljenog genoma se također sastoji od repeticija koje zbog tehničkih poteškoća nije bilo moguće sastaviti u veće komponente kao kromosome. Preko analize repeticija na svakom kromosomu, Wang i suradnici su tražili područja koja pripadaju heterokromatinu (što se svelo na pregledavanje 137,7 Mb) čija distribucija je prikazana u tablici 7.

Tablica 7. Distribucija repeticijske DNK u putativnom heterokromatinu i eukromatinu u sastavljenom genomu *T.castaneum*. Preuzeto i prerađeno iz [29].

Element repeticije	Ukupna duljina (kb)	Količina u heterokromtinu (kb)	Količina u eukromatinu (kb)	Postotak u heterokromatinu	Postotak u eukromatinu
Ukupni DNK	137758	54754	83004	39,70	60,30
HighA	8729	5633	3096	64,53	35,47
Mid	8769	5633	3096	59,00	41,00
Low	4915	2893	2022	58,86	41,14
HighB	2045	567	1778	13,06	86,94
Non-LTR	1370	962	408	70,22	29,78
LTR	1042	896	312	74,17	25,83
DNK transpozon	2579	1963	616	76,11	23,89
Mikro sateliti	439	188	251	42,82	57,18
Mini sateliti	2593	1152	1441	44,43	55,57
Tandemni sateliti	2621	646	1975	24,65	75,35

2010. godine objavljena je nadopunjena verzija genoma označena sa Tcas 3.0 [128]. Sekvence iz 9686 kontiga sastavljene su u 10 kromosoma (CM000276–CM000285, označenih još s LG1-LG10), 305 nepovezanih višestruko – komponentnih „scaffold-a" (unmapped scaffolds DS497665–DS497969) i 1848 nepoznatih jednostrukih „scaffolde" (unmapped single contigs GG694051–GG695897) s brojevima pristupa u GenBank-u AAJJ01000001–AAJJ01009708. Cijeli genom je procijenjen na ~160 Mb sa statistikom prikazanom u tablici 8.

Tablica 8. Statistika Tcas 3.0. Preuzeto i prerađeno iz [128].

Kromosom	Duljina (baze)	Broj kontiga (baze)	Broj „uhvaćenih“ razmaka (baze)	Broj „neuhvaćenih“ razmaka baza
ChLGX	10 877 635	338 (7 017 036)	299 (265951)	12 (3 600 000)
ChLG2	20 218 415	393 (14 025 453)	338 (505072)	19 (5 700 000)
ChLG3	38 731 480	1560 (27 070 658)	1355 (1 568 829)	34 (10 200 000)
ChLG4	13 894 384	331 (11 543 342)	299 (554 338)	6 (1 800 000)
ChLG5	19 135 781	338 (13 841 583)	335 (502 879)	16 (4 800 000)
ChLG6	13 176 827	667 (8 259 034)	549 (747 290)	14 (4 200 000)
ChLG7	20 532 854	445 (14 850 616)	401 (591 423)	17 (5 100 000)
ChLG8	18 020 898	676 (12 793 837)	570 (761 081)	15 (4 500 000)
ChKG9	21 459 655	695 (14 607 456)	598 (892 186)	20 (6 000 000)
ChLG10	11 386 040	585 (7 061 652)	495 (442 098)	13 (3 9000 000)
Nepoznati	41 251 169	3616 (20 543 639)	1254 (2 031 250)	1848 (18 480 000)

Od svih tandemnih repeticija najviše su proučavani TCAST sateliti u genomu *Tribolium castaneum* [27, 31, 129]. TCAST sateliti zauzimaju 35% genoma u pericentromernom području i području eukromatina. Zanimljivi su za proučavanje pošto njihovi transkripcijski regulatorni elementi mogu biti modulatori ekspresije gena u eukromatinu [129]. TCAST sateliti su sastavljeni od dva podtipa Tcast1a (377 bp) i Tcast1b (362 bp) sa sličnosti od 79% [31] koje su autori tražili u sastavljenom genomu u kromosomima, te su pronašli 68 TCAST raspršenih elemenata koji se nalaze u blizini gena. Tih 68 sekvenci su podijelili u dvije grupe na temelju strukture, odnosno da li se pojavljuju kao monomeri ili u tandemu (dimer, trimer, tetramer (jedan duljine 1440 bp)).

Iz prikazanih rezultata, može se vidjeti da tandemne repeticije imaju značajne uloge u genomima eukariota. Također, nedavna istraživanja su pokazala da tandemne repeticije, koje mogu tvoriti periodičnosti višega reda, kod ljudskih alfa satelita, su potrebne za pravilno funkcioniranje centromere preko motiva CENB-P box te također da upravo oni mogu tvoriti evolucijsku razliku između blisko povezanih vrsta kao što su čimpanza i čovjek, tvoreći takozvane ljudske ubrzane periodičnosti višeg reda – HAHOR [21].

Cilj istraživanja

Metoda koju koristim u disertaciji je računalna GRM metoda (Global Repeat Map), razvijena u našoj grupi (PMF – Fizika, V.Paar, M. Glunčić, M. Rosandić, I. Vlahović) za identifikaciju periodičnosti višega reda u cijelom sekvencioniranom genomu insekta *Tribolium castaneum* (Tcas 3.0). Pomoću ove metode mogu se identificirati tandemne kao i raspršene repeticije u sekvencioniranim genomima. Prednost metode je mogućnost identifikacije repeticija na velikim udaljenostima u genomu i to neovisno o duljini repeticijske jedinice (monomera). Temelj učinkovitosti metode je direktno mapiranje simboličke sekvence u frekventnu domenu pomoću koje možemo odrediti repeticije svih vrsta preko GRM dijagrama. Za razliku od drugih bioinformatičkih alata, metoda je robusna na supstitucije, brisanja i umetanja nukleotida te daje veliku mogućnost pri identifikaciji jednostavnih i kompleksnih periodičnosti višega reda (HOR-ova).

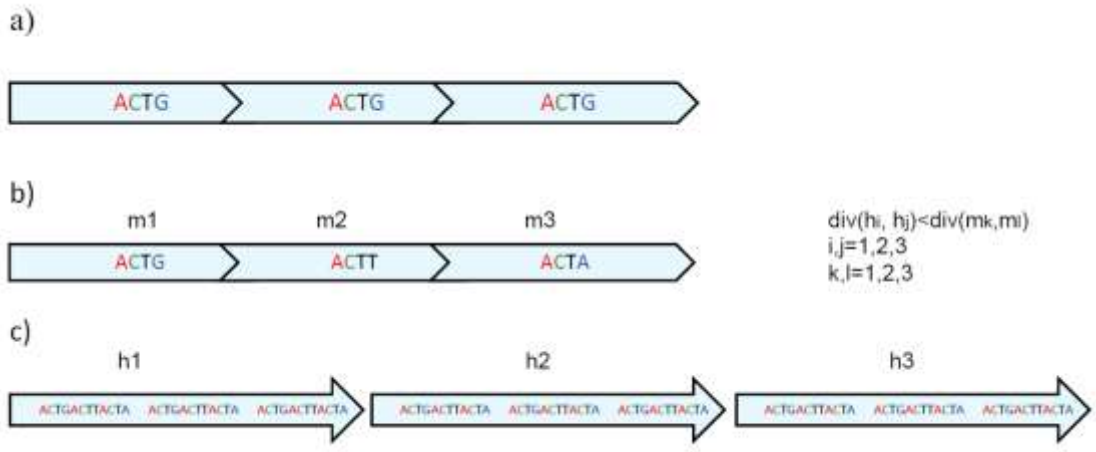
Identifikacija HOR-ova u *T. castaneum* je zanimljiva iz tog razloga što se smatralo da su HOR-ovi općenito karakteristični za sisavce, više primate i čovjeka. Također, identifikacija HOR-ova u genomu *T. castaneum* bi upućivala na postojanje HOR-ova i u ostalim insektima i vrstama nastalim prije sisavaca i koja bi poslužila kao pomoć pri određivanju njihove uloge i evolucije.

3. Materijali i metode

Sa napretkom u tehnologiji sekvencioniranja genoma, razvijali su se i algoritmi za njihovu analizu. Posebno su se počeli razvijati algoritmi za analizu različitih vrsta repeticija koji su se specijalizirali ili samo za traženje raspršenih repeticija ili samo za traženje tandemnih repeticija ili obje. Algoritme za traženje tandemnih repeticija možemo podijeliti u fleksibilne statističke algoritme usporedbe nizova i algoritme za procesiranje signala. U našoj grupi razvijena je nova računalna metoda Global Repeat Map – GRM, korištena za analizu genoma *T.castaneum* u ovoj disertaciji, koja koristi elemente iz obje skupine algoritama.

3.1 Tandemne repeticije i periodičnosti višega reda

Tandemne repeticije su repeticije u kojima se pojavljuju nizovi uzorka nukleotida, u više kopija jedna za drugom, kao što je prikazano na slici 14. Ti uzorci mogu biti kratki ili dugački, odnosno možemo ih podijeliti u mikrosatelite (1 - ~6 bp), minisatelite (~6 – 100 bp), satelite (100 bp – ~2 kbp), makrosatelite (veće od ~2 kbp). Tandemne repeticije, s različitim mehanizmima kao što su mutacije, insercije i brisanja nukleotida, se mogu tokom svoje evolucije izmijeniti. Također u nekim slučajevima mogu tvoriti i periodičnosti višega reda - HOR kopije, pri čemu se monomeri međusobno razlikuju oko 20% - 35% [19], ali u strukturi višega reda razlike između HOR kopija su manje od 5%. HOR – ovi, kao što je prikazano u poglavlju „Literaturni pregled“ mogu imati bitne uloge.



Slika 14. Prikazi a) tandem, b) tandem koji se može s različitim mehanizmima promijeniti unutar kopija, c) tandemne repeticije kao periodičnost višega reda – HOR-a.

3.2 Fleksibilni statistički algoritmi usporedbe nizova

Detektiranje tandemnih repeticija i periodičnosti višega reda je bitan zadatak u analizi genomske sekvence, a s time su nastali zahtjevi za razvijanjem novih algoritama koji omogućuju prepoznavanje i tandemnih repeticija koje međusobno variraju. Algoritmi za prepoznavanje takvih repeticija (aproksimativnih) se zasnivaju na prepoznavanju ponavljanja unutar riječi (w) koja se može sastojati od podriječi (uv) koji su udaljeni za k. Prema tipu udaljenosti mogu se napisati definicije za tandemne repeticije (konkretno njih tri, ovisno da li se zasnivaju na tzv. Hammingovoj udaljenosti ili prema udaljenosti uređivanja). Za prve dvije, Hammingova udaljenost [130] je označena s $h(\cdot,\cdot)$ i označava udaljenost između riječi jednakih duljina, dok u trećoj definiciji se mora definirati udaljenost između dva niza [131, 132] (definicije su preuzete iz [133]).

- Riječ $r[1 \dots n]$ je *K-repeticija* perioda p, gdje je $p \leq n/2$, ako je $h(r[1 \dots n-p], r[p+1 \dots n]) \leq K$. Odnosno riječ $r[1 \dots n]$ je *K-repeticija* perioda p, ako je broj razlika (i) takav da vrijedi $r[i] \neq r[i+p]$ je najviše K.
- Riječ $r[1 \dots n]$ je „*K-run*“ (pokretanja), perioda p, gdje je $p \leq n/2$ ako za svaki $i \in [1 \dots n-2p+1]$, imamo da je $h(r[i \dots i+p-1], r[i+p \dots i+2p-1]) \leq K$. Za „*K-run*“ vrijedi da je $|u| = |v| = p$ gdje se u i v mogu najviše razlikovati za K razlika (koje nisu ograničene u „*K-run*“-u).
- Riječ je „*K-edit*“ ponavljanje (gdje je edit – udaljenost definirana kao udaljenost između dva niza), ako se ona može podijeliti na podnizove tako da vrijedi $r = v' w_1 w_2 \dots w_l v''$, gdje je $l \geq 2$, tako da $ed(v', w_1') + \sum_{i=1}^{l-1} ed(w_i, w_{i+1}) + ed(w_l'', v'') \leq K$, gdje w_1' neki sufiks od w_1 i w_l'' prefiks od w_l. U „*K-edit*“ ponavljanju mogu se pojavljivati najviše K insercija, brisanja i razlika u svim kopijama.

Općenito je svaka K - repeticija dio „K-run-a” istog perioda i svaki „K-run” je unija svih K - repeticija koje sadrži. Za danu duljinu n riječi w mogu se izračunati vremena i prostorne granice „*K-run*“-a, *K-repeticija i „K-edit*“ repeticija. Neki od algoritama iz ove skupine su Tandem Repeat Finder [13], Mreps [134], STAR [135] itd.

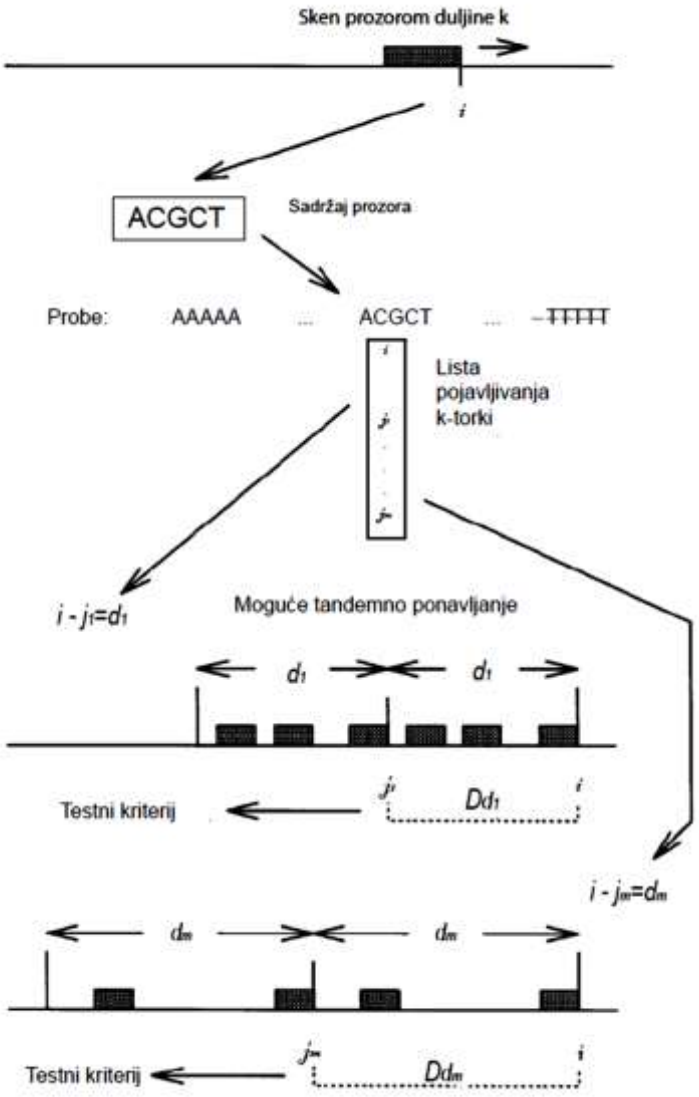
Tablica 9. Prikaz vremena i prostornih granica [133]. *S* označava izlaznu veličinu sekvence.

	Vrijeme	**Prostor**
„K-run“	O(nKlogK+S)	O(n)
„K“ repeticija	O(nKlogK+S)	O(n)
„K-edit“ repeticije	O(nKlogKlog(n/k)+S)	$O(n+K^2)$

3.2.1.Tandem Repeat Finder

Jedan od najpopularnijih programa za identifikaciju tandemnih repeticija koji pripada u statističke algoritme je Tandem Repeat Finder – TRF. On radi u dva dijela, odnosno ima komponentu detekcije i komponentu analize.

Kod detekcije, za traženje tandemnih repeticija, ovaj algoritam ima pretpostavku da susjedne kopije imaju pogotke na određenim pozicijama. Za određivanje tandema, mora se fiksirati vrijednost pogodaka p_M i udaljenosti između njih p_I, koje se određuju preko statističkih kriterija. Algoritam radi na principu pogodaka *k-torka* koji se sastoji od *k* uzastopnih nukleotida čije poklapanje se traži za udaljenost *d* kao što je prikazano na slici 15.



Slika 15. Pomicanjem prozorom veličine *k* po sekvenci, mogu se odrediti udaljenosti između točnih pogodaka te korištenjem statističkih kriterija mogu se odrediti tandemne repeticije. Preuzeto i prerađeno iz [13].

Statistički kriteriji koje koristi TRF se zasnivaju na Bernoulijevoj sekvenci ovisno o detektiranim *k - torkama* pohranjenim u listu i zasnovani su na četiri distribucije koje ovise o duljini uzorka *d*, vjerojatnosti poklapanja p_M, vjerojatnosti indela (umetanje ili brisanja nukleotida) p_I i veličini *k-torki*. Te distribucije su suma distribucije glava (temeljena na normalnoj distribuciji, sa zadanim početnim parametrima duljine *d*, vjerojatnosti pogodaka i veličini *k*), distribucija nasumičnog hoda (koja opisuje kako mogu varirati udaljenosti između pogodaka *k-torki*), distribucija vidljive veličine (koja se koristi za razlikovanje tandemnih i raspršenih repeticija) i distribucija vremena čekanja (koja služi za odabir veličine prozora odnosno *k-torki* radi bolje iskoristivosti algoritma).

U komponenti analize, pronađene repeticije se pokušavaju poravnati s omotanim dinamičkim programiranjem (WDP – wraparound dynamic programming [136]) ako se zadovolje statistički kriteriji (postotak identičnosti i indela, sastav). Pomoću WDP-a, moguće je pojavljivanje nekoliko veličina uzorka tandemnih repeticija. Samo poravnanje radi se u odnosu na konsenzusnu sekvencu (većina uzorka sa istom veličinom). U odabiru statističkih kriterija koristi se „cut-off" vrijednost u odnosu na nasumične sekvence te se mogu pojaviti problemi kod varijacija u monomerima.

3.3 Algoritmi za procesiranje signala

Druga skupina programa za identifikaciju tandemnih repeticija pripada algoritmima za procesiranje signala. Ova grupa programa mapira genomsku sekvencu u numeričku te ju analizira s alatima poput brze Fourierove metode (FFT). Jedan takav algoritam je Spectral Repeat Finder (SRF) [14], ali postoje i mnogi drugi [137-142].

3.3.1 Spectral Repeat Finder

Ova metoda za identifikaciju repeticija se može svesti na nekoliko koraka. U prvom koraku kao ulazna informacija se uzima DNK sekvenca, koja se može napisati kao $\alpha_1\alpha_2...\alpha_n$, ($\alpha_i$ može biti A, T, G ili C). U drugom koraku mora se izračunati „power" spektar definiran kao:

$$S(f) = \sum_{\alpha} \frac{1}{n} \left| \sum_{j=1}^{n} U_\alpha(j) e^{2\pi i f j} \right|^2 ,$$

gdje je frekvencija f=1/N što nam daje naznaku u spektru da korelacijska funkcija ima periodičnost od N baza [143]. Korelacijska funkcija je oblika :

$$C_{\alpha\beta}(r) = \langle U_\alpha(i) U_\beta(i+r) \rangle, \qquad \alpha, \beta \in \{A, T, G, C\},$$

gdje je $U_\alpha(i) = 1$ za α_i da je simbol na poziciji i, ili je $U_\alpha(i) = 0$ za α_i različit od simbola na poziciji i te označava prosjek preko niza. Za repeticije, korelacijska funkcija [144] ima oblik

$$C(r) \approx C(N + r).$$

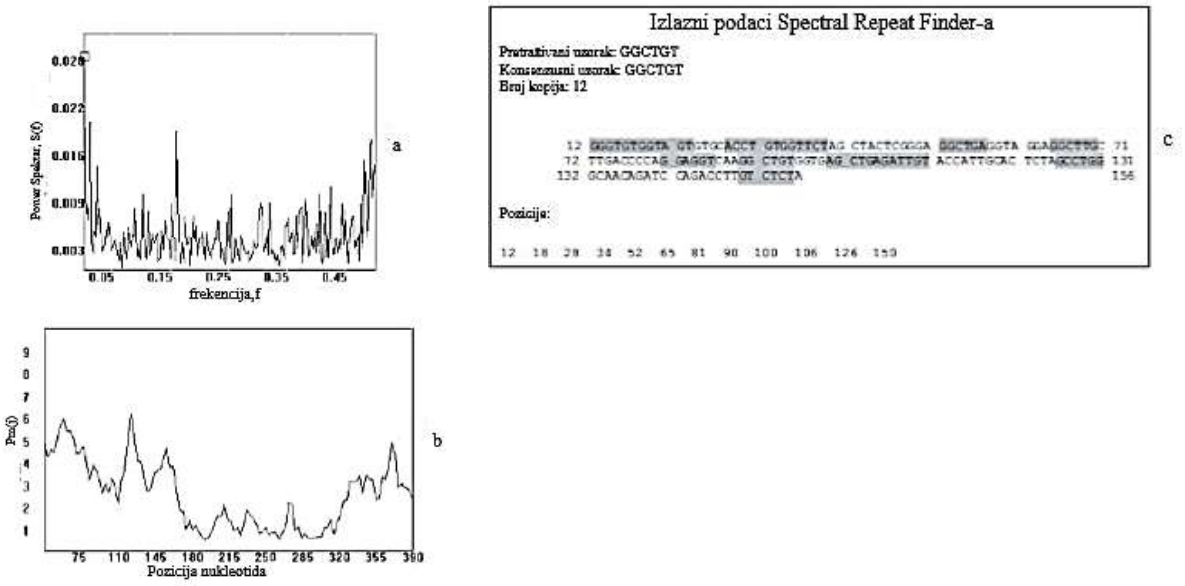
Također se računa i spektralni prosjek $\bar{S}$ cijele sekvence preko izraza:

$$\bar{S} = \frac{1}{n}\left(1 + \frac{1}{n} - \sum_{\alpha} \rho_\alpha^2\right),$$

gdje je ρ_α frekvencija nukleotide α u sekvenci.

Treći korak služi za identifikaciju svih pikova sa $S(f_i)/\bar{S} > T$, (T je omjer S/N koji nam određuje bitnost pojedinog pika i općenito se uzima iznos 4 [145], premda su već i pikovi za taj omjer veći od dva, bitni) s time da se može izračunati za svaki f_i repeticije duljine $N_i = 1/f_i$ kao i periodičnosti višega reda.

U slijedećem koraku računa se omjer „power" spektra i spektralnog prosjeka $P_m(j) = S(f_i)/\bar{S}$ u prozoru veličine m centriranom na poziciji j u sekvenci. Računanjem $P_m(j)$ omogućuje se identifikacija područja u kojemu se nalazi repeticija. Daljnji korak koristi egzaktnu metodu određivanja jedinice repeticije (monomera) koja dopušta pojavljivanje brisanja nukleotida ili umetanje nukleotida. Primjer primjene SFR programa je prikazan na slici 16.



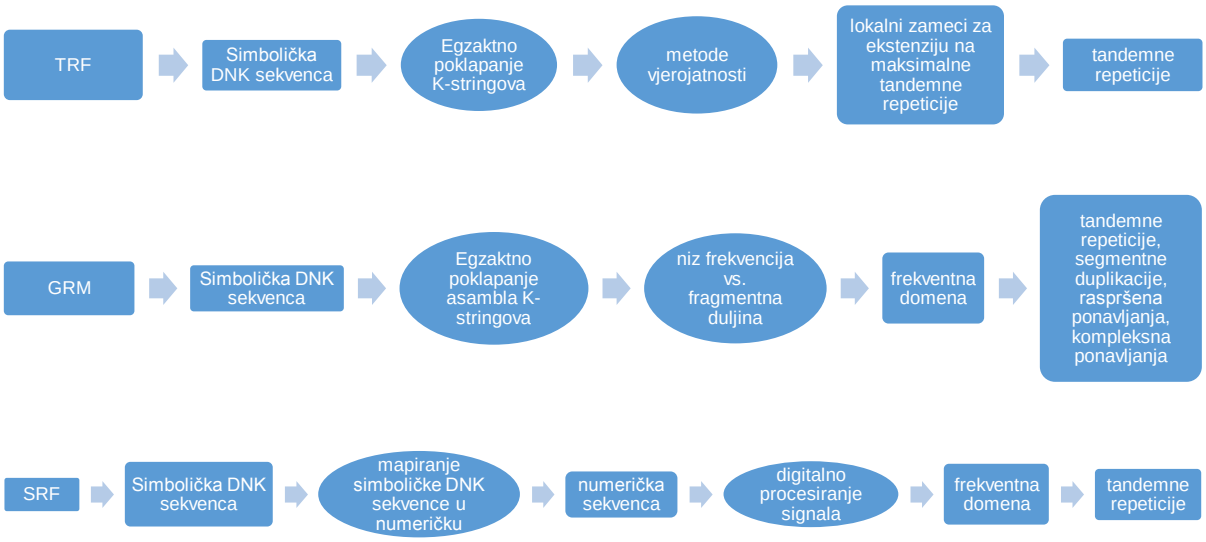
Slika 16. Prikaz a) analize mikrosatelita M96445 sa SRF-om s glavnim dobivenim pikovima na $f = 1/2$ i $f = 1/6$, b) analiza s pomicanjem prozora duljine 96 baza, c) točnih lokacija heksamera koji daju peak za 1/6 iz a). Preuzeto i prerađeno iz [14].

3.4 Global Repeat Map računalna metoda - GRM

Računalna metoda, Global Repeat Map – GRM, još je jedan od novijih alata koji služi za identifikaciju repeticija svih vrsta te se posebno pokazao učinkovitim u identifikaciji periodičnosti višega reda (HOR-ova), pogotovo na primjeru čovjeka i ostalih sekvencioniranih primata [21, 22, 23, 93].

Ova računalna metoda razvijena je na Prirodoslovno - matematičkom fakultetu, Fizički odsjek, Sveučilište u Zagrebu (www.hazu.hr/grm/software/win/grm2012.exe). U ovoj disertaciji je i glavna metoda s kojom se analizirao genom *Tribolium castaneum*-a.

Kao što je već bilo spomenuto, ova metoda koristi elemente i iz grupe fleksibilnih statističkih algoritama usporedbe nizova i algoritama za procesiranje signala, dajući rezultate preko kojih se mogu identificirati HOR-ovi, kao što je prikazano na slici 17.



Slika 17. Prikaz usporedbe GRM metode i alata iz grupe fleksibilnih statističkih algoritama i algoritama za procesiranje signala. Preuzeto i prerađeno iz [146].

Algoritam radi u nekoliko koraka. Prvo, simboličku sekvencu možemo napisati, isto kao kod ostalih algoritama, kao niz nukleotida (A, T, G, C) i K-stingova (uzorak od K elementa u simboličkoj sekvenci):

$$S_K(j) = \alpha_1(K,j)\alpha_2(K,j)\alpha_3(K,j) \ldots,$$

gdje $\alpha_i(K,j)$ označava jednu od četiri nukleotide na i-toj poziciji u j-tom K-stringu kojih u simboličkoj sekvenci može biti 4^K. Svi mogući K-stringovi tvore ansambl (E_K). Zatim se svi

pogodci, za ukupnu sekvencu DNK duljine L, zasebno određuju za svaki K-string $S_K(j)$, pomičući se za 1 nukleotidu u sekvenci, pri čemu se zapisuju njihove početne pozicije:

$$\{X_K(j)\} = [X_K(j)]_1, [X_K(j)]_2, \ldots., [X_K(j)]_n, [X_K(j)]_{n+1},$$

kao i razmak između susjednih pogodaka pojedinog K-stringa:

$$[d_K(j)]_n = [X_K(j)]_{n+1} - [X_K(j)]_n \ldots, n = 1,2,3 \ldots$$

Upravo ove udaljenosti između pogodaka za svaki K-string su bitne, pošto pomoću njih simboličku DNK sekvencu možemo izraziti preko frekventne domene i za cijeli ansambl K-stringova možemo napisati:

$$\{d_K(j)\} = [d_K(j)]_1, [d_K(j)]_2, [d_K(j)]_3 \ldots \; j = 1, 2, \ldots 4^K.$$

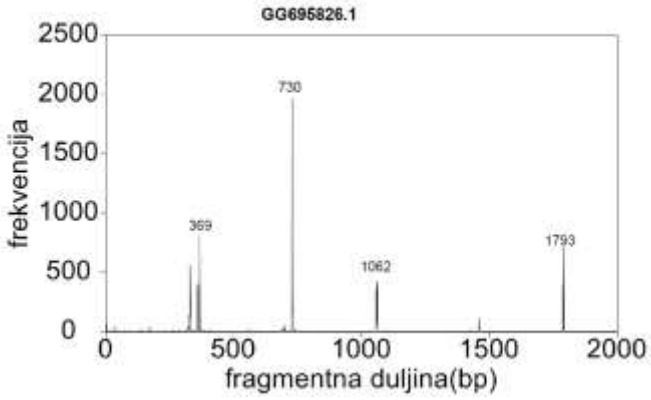
Sada s zabilježenim duljinama, možemo za svaku duljinu razmaka dobiti odgovarajuću frekvenciju jednostavnim prebrojavanjem broja puta pojavljivanja određene duljine u sekvenci, prvo za svaki K-string zasebno:

$$\{f_K(j)\} = [f_K(j)]^1 [f_K(j)]^2, \ldots., [f_K(j)]^\nu,$$

a zatim i za cijeli ansambl K-stringova s superpozicijom zasebnih:

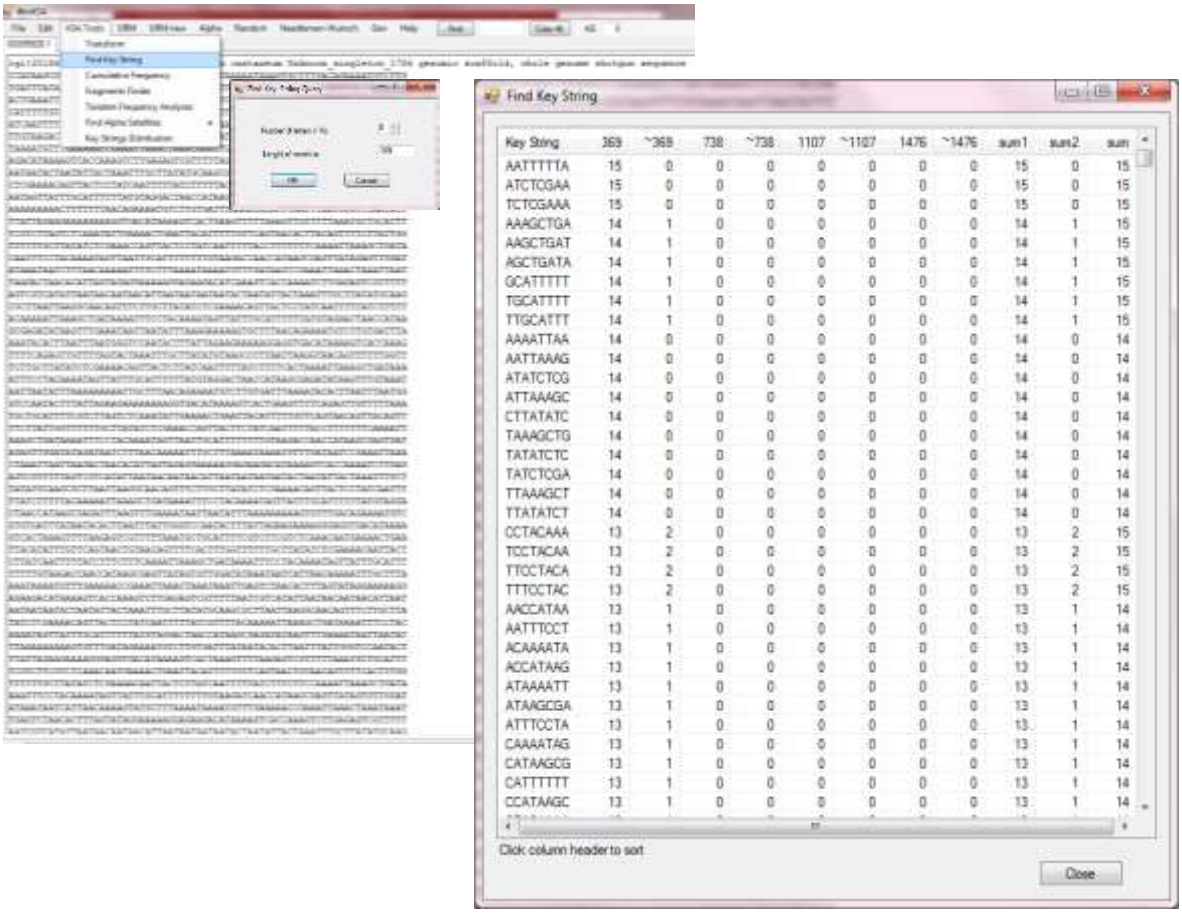
$$\{f_{K(E)}\} = \sum_{j=1}^{N} [f_K(j)]^1, \sum_{j=1}^{N} [f_K(j)]^2, \ldots\ldots, \sum_{j=1}^{N} [f_K(j)]^\nu, \; N = 4^K.$$

Ovim postupkom Global Repeat Map mapira simboličku DNK sekvencu u frekventnu domenu, koja se može prikazati na GRM dijagramu kao globalna mapa. Iz GRM dijagrama nakon identifikacije pikova, rade se daljnje analize sekvence. Sami pikovi mogu predstavljati raspršena ponavljanja, tandemne repeticije i kompleksna ponavljanja (slika 18).



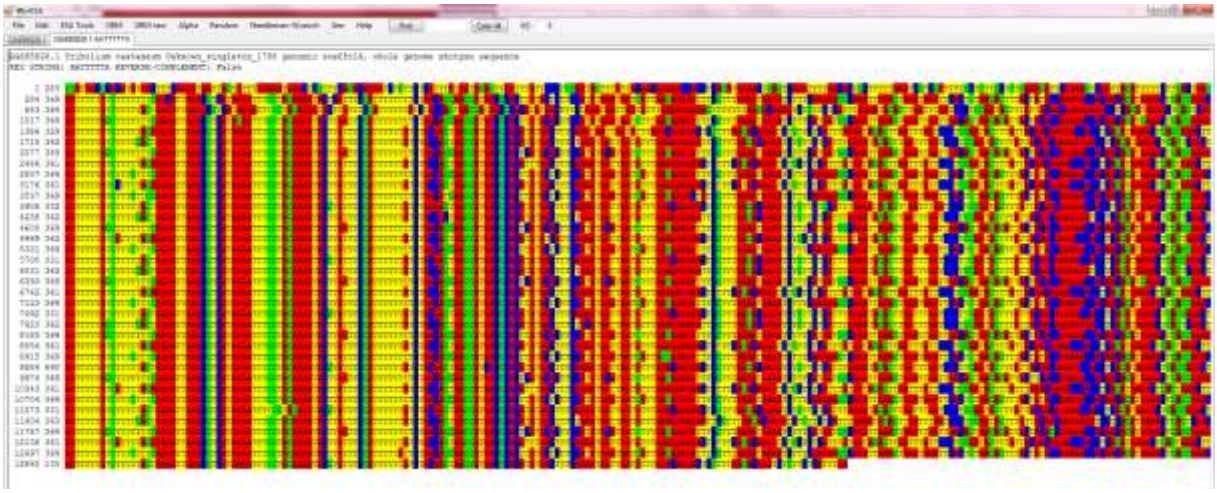
Slika 18. Prikaz GRM dijagrama s identificiranim pikovima u sekvenci GG695826.1 iz *T.castaneum* genoma.

Preko GRM dijagrama mogu se identificirati pikovi za koje se treba odrediti vrsta repeticija kojoj oni pripadaju. U slijedećem koraku analize trebamo odrediti dominantni K-string, pomoću algoritma iz grm2012, odnosno onaj K-string koji ima najvišu frekvenciju ponavljanja za traženu fragmentnu duljinu odnosno za $max[f_K(j)^m]$. U programu grm2012 taj K - string se odabere iz već pohranjenog niza kao što je prikazano na slici 19.



Slika 19. Prikaz određivanja dominantnog K-stringa pomoću programa grm2012.

Nakon pronalaska K-stringa, da bi vidjeli da li je određeni pik s GRM dijagrama tandemna repeticija ili raspršena, cijelu sekvencu „režemo“ s dominantnim K-stringom te na taj način dobivamo točne pozicije unutar DNK sekvence i duljine fragmenata monomera (osnovne jedinice ponavljanja) bez obzira da li se među njima nalaze umetanja, brisanja ili supstitucije nukleotida (slika 20).



Slika 20. Prikaz fragmentirane sekvence s dominantnim K-stringom AATTTTTA.

Ovako fragmentirane sekvence pogodne su za daljnju analizu, odnosno određivanja konsenzusne sekvence, računanja divergencije između svih kopija u odnosu na konsenzus te međusobne divergencije (pomoću Needleman-Wunsch algoritma) između monomera kao i za slaganje periodičnosti višega reda, ako one postoje, te provjeru s BLAST-om da li se one pojavljuju u nekom drugom području sekvencioniranog genoma istog ili različitih organizama.

GRM metoda može analizirati velike DNK sekvence, za koje nisu potrebni nikakvi ulazni parametri, i to na običnom PC računalu preko intuitivnog sučelja programa grm2012.

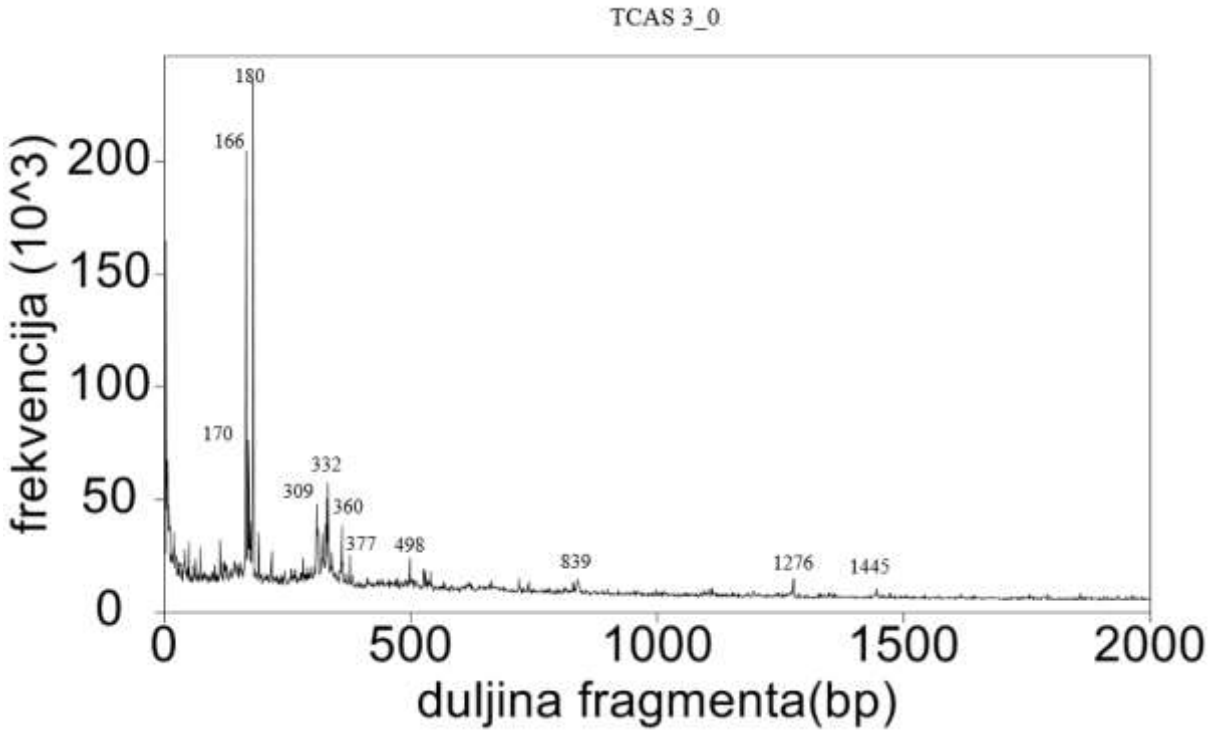
Za razliku od fleksibilnih statističkih algoritama usporedbe nizova, koji imaju problema s velikim umetanjima i brisanjima nukleotida na dugačkim sekvencama, i algoritama za procesiranje signala, u kojima se mogu javiti numerički artefakti, GRM metoda direktno mapira simboličku DNK sekvencu u frekventnu domenu, stvarajući globalnu mapu pomoću ansambla K-stringova bez upotrebe statističkih metoda (čija je mana unos nekih parametara, a s time i točnost analize pri određivanju repeticija).

Primjenom ove metode identificiran je veliki broj periodičnosti višega reda (HOR) u čovjeku i primatima [23]. GRM metoda se primijenila za identifikaciju alfa satelita u kromosomu 7 kao tandemnih repeticija i HOR-ova, alu raspršenih kopija i 2 alu tandema, kompleksnih HOR-ova u ljudskom Y kromosomu zasnovanom na ~2.4 kbp monomeru i primjeni na velike segmentne duljine (identifikacija 0.6 Mbp segmentne duplikacije s 1Mbp razmakom u čovjekovom kromosomu Y).

Također smo kod čovjeka pronašli 3mer HOR zasnovan na ~1.6 kb dugačkom monomeru (u cijelosti umetnutom u neuroblastoma „breakpoint“ obitelji gena koji je povezan s funkcijom mozga) odsutnih iz genoma primata. Generalni zaključak je da se čovjek i

čimpanza razlikuju značajno za tandemne repeticije (HOR-ovima) nego za genski dio sekvence. Važnost ljudskih akceleriranih HOR-ova (HAHOR) kao komponente u ekspresiji gena, koje nisu primijećene kod primata upućuju na evolucijski skok između njih [21].

U ovoj disertaciji primijenila sam GRM metodu na cijeli sekvencionirani genom insekta *T.castaneum* radi identifikacije periodičnosti višega reda. Nakon svih preuzetih komponenti iz Tcas 3.0, izradila sam GRM analizu za svaku, čije rezultate sam svela na jedinstveni GRM dijagram iz kojeg sam u ostalim komponentama tražila one duljine fragmenata kojima najistaknutiji pikovi pripadaju (slika 21).

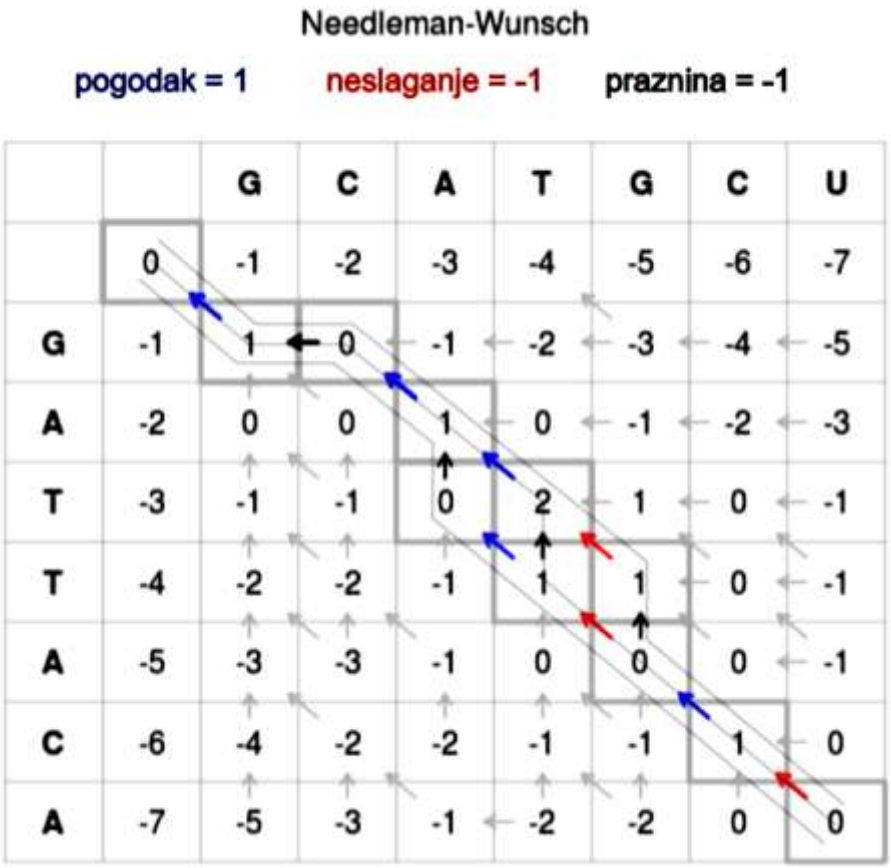


Slika 21. Sumarni dijagram za sve komponente *T.castaneum*. Preuzeto i obrađeno iz [147].

3.5 Needleman – Wunsch algoritam

Program za poravnavanje amino kiselina ili nukleotida su napravili 1970.g. Needleman i Wunsch temeljen na dinamičkom programiranju koji nalazi najbolja moguća poravnanja između dvije sekvence. Ovaj algoritam je zasnovan na računaju pogodaka (match), neslaganja (msimatch) i praznina (gap) koji se računaju preko dvodimenzionalne matrice (slika 22). Preko izraza u kojem se traži maksimalna vrijednost elementa matrice F_{ij} u odnosu na tri pomaka (d je iznos penala za prazninu, S daje naznaku da li se radi o pogotku ili neslaganju):

$$F_{ij} = \max\left(F_{i-1,j-1} + S(A_i, B_j), F_{i,j-1} + d, F_{i-1,j} + d\right).$$



Slika 22. Prikaz dvodimenzionalne matrice F za računanje poravnanja. Podatak u redu *i* i stupcu *j* je F_{ij}. Jedan stupac označava sekvencu A, a red sekvencu B u matrici. Sekvenci A u stupcu se dodaju vrijednosti i=0,…n, a sekvenci B vrijednosti u redu j=0,…n. Preuzeto i prerađeno s [148].

Pomoću Needleman-Wunsch programa (nakon identifikacije repeticija s GRM računalnom metodom i nakon određivanja dominantnog K-stringa s kojim se „reže“ dana DNK sekvenca na fragmente određenih duljina), radimo tablice usporedbe konsenzusne sekvence

repeticija sa svim monomerima kao i međusobnu usporedbu monomera u tandemu kako bi lakše identificirali periodičnosti višega reda. Usporedbe međusobnih kopija prikazujemo pomoću tzv. „heat“ mape koja nam daje informaciju o postojanju periodičnosti višega reda prema iznosu postotka divergencije (prema definiciji monomeri tvore repeticiju višega reda ako se oni međusobno razlikuju za ~20-~35%, a kada su složeni u strukturu višega reda taj iznos je manji od ~5%, kao što je prikazano na slici 23) što se također vidi i iz GRM dijagrama.

		m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13
		338	667	999	1361	1734	2063	2395	2758	3135	3507	3837	4169	4531
m1	338	0	22,8	14,36	30	1	23	15	29	28	1,2	23	15	30
m2	667		0	25,48	18	23	1	25	18	16	23	0,6	26	18
m3	999			0	27	14	25	2	27	25	14	25	2,5	28
m4	1361				0	29	18	27	5	3	29	18	27	5,6
m5	1734					0	23	15	29	27	0,6	23	15	30
m6	2063						0	25	17	16	23	0	26	18
m7	2395							0	26	25	15	25	1,1	28
m8	2758								0	8	29	17	27	2,1
m9	3135									0	27	16	25	8,5
m10	3507										0	23	15	30
m11	3837											0	26	18
m12	4169												0	28
m13	4531													0

Slika 23. Primjer „heat“ mape dobivene s Needleman – Wunsch algoritmom za sekvencu pika 332, GG695436.1 iz genoma *T.castaneum* (m1-m13 predstavljaju redoslijed monomera).

Needleman – Wunsch algoritam je dodan u grm2012 program, s time da za potrebe naših analiza osim računanja matrica s direktnim kopijama monomera, možemo isto raditi s obrnutim komplementom sekvenci.

3.6 BLAST - Basic Local Alignment Search Tool

Prvi programi za poravnavanje sekvenci temeljili su se na algoritmima dinamičkog programiranja (Needleman – Wunsch [149], Smith-Waterman [150, 151]) za čiji rad su bila potrebna super računala zbog svoje sporosti. BLAST je povećao brzinu poravnanja sa smanjenjem područja rada tako da koristi riječi kao zametke pretrage te na temelju stvorene liste riječi u sekvenci upita omogućuje manji broj pretraga u sekvenci subjektu [152]. Nakon postavljanja zametka, pomoću podataka za najvišu dopuštenu vrijednost koje postavlja korisnik, zametak se širi kako bi se dobili bolji rezultati poravnanja. Program zatim koristi sekvence koje imaju pogotke iznad granične vrijednosti, što upućuje na homologne sekvence.

Prednost programa je što računa statističku značajnost za svaki pojedini pogodak poravnanja (očekivana vrijednost – E i vrijednost vjerojatnosti – P). Od početka razvoja, stvorile su se varijante programa kao što su BLASTN (uspoređivanje nukleotida sa cjelokupnom bazom podataka nukleotida), BLASTP (uspoređivanje sekvence proteina s bazom podataka proteinskih sekvenci), BLASTX (sekvencu nukleotida pretvara u proteinsku i uspoređuje ju s bazom podataka). Sve verzije programa mogu se pronaći na web stranici: http://blast.ncbi.nlm.nih.gov/Blast.cgi.

4. Rezultati i rasprava

GRM metodu primijenili smo na Tcas 3.0 verziju genoma *T.castaneum* koja je dostupna na NCBI web stranici http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AAJJ01. Tcas 3.0. sastavljen je od kontiga AAJJ01000001-AAJJ02009708 u kromosome CM000276-CM000285 („linkage grupe" – označene s L1-L10 i napravljenih od 140 spojenih kontiga (scaffolda) koji predstavljaju 70% sekvencioniranog genoma, sastavljenih s visoko rezolucijskim rekombinacijskim tehnikama mapiranja pomoću bakterijskih umjetnih kromosoma (BAC) i pomoću ekspresije sekvence „tag" markera), u nepovezane višestruko – komponentne povezane kontige („unlinked multi-component scaffolds" DS47665-DS497969) i nepoznate jednostruke spojene kontige („unknown singleton scaffolds" GG694051–GG695897). Statistika za Tcas 3.0 genom je prikazana u tablici 10.

Tablica 10. Nazivi komponenti u genomu Tcas 3.0 i njihova statistika. Preuzeto i prerađeno s [153].

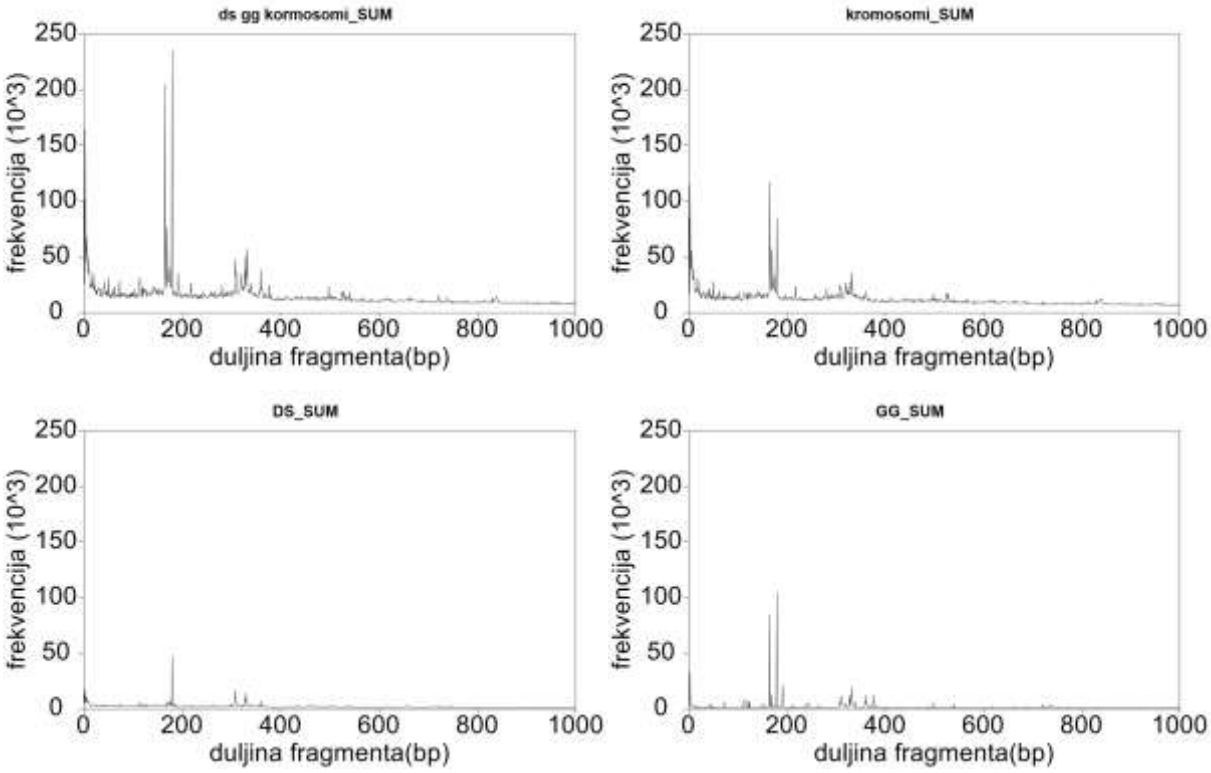
Tip	Naziv	INSDC	Veličina (Mb)	GC%	Proteini	rRNA	tRNA	Drugi RNK	Geni	Pseudogeni
	master WGS	AAJJ00000000.1	210.25	33.9	16,530	-	-	-	16,540	12
Linkage Group	LG1=X	CM000276.2	10.88	41.1	1,095	-	3	134	698	1
Linkage Group	LG2	CM000277.2	20.22	39.6	1,854	-	29	90	1,337	4
Linkage Group	LG3	CM000278.2	38.79	37.5	2,108	-	44	164	1,614	2
Linkage Group	LG4	CM000279.1	13.89	38.1	1,782	-	8	63	1,245	2
Linkage Group	LG5	CM000280.2	19.14	39.4	2,126	-	25	106	1,550	2
Linkage Group	LG6	CM000281.2	13.18	39.3	1,098	-	9	62	898	2
Linkage Group	LG7	CM000282.2	20.53	39.3	2,240	-	39	105	1,670	2
Linkage Group	LG8	CM000283.2	18.02	38.5	1,870	-	28	71	1,361	2
Linkage Group	LG9	CM000284.2	21.46	39.2	1,796	-	33	92	1,243	3
Linkage Group	LG10	CM000285.2	11.39	38.9	714	-	4	42	502	1
Chr	MT	AJ312413.2	0.015881	28.3	13	2	22	-	13	-
Un	-	-	22.75	34.2	1,380	-	3	195	1,113	7

Duljine baza komponenti DS47665-DS497969 i GG694051–GG695897 genoma Tcas 3.0 nalaze se u tablici 1 u poglavlju „Dodatak".

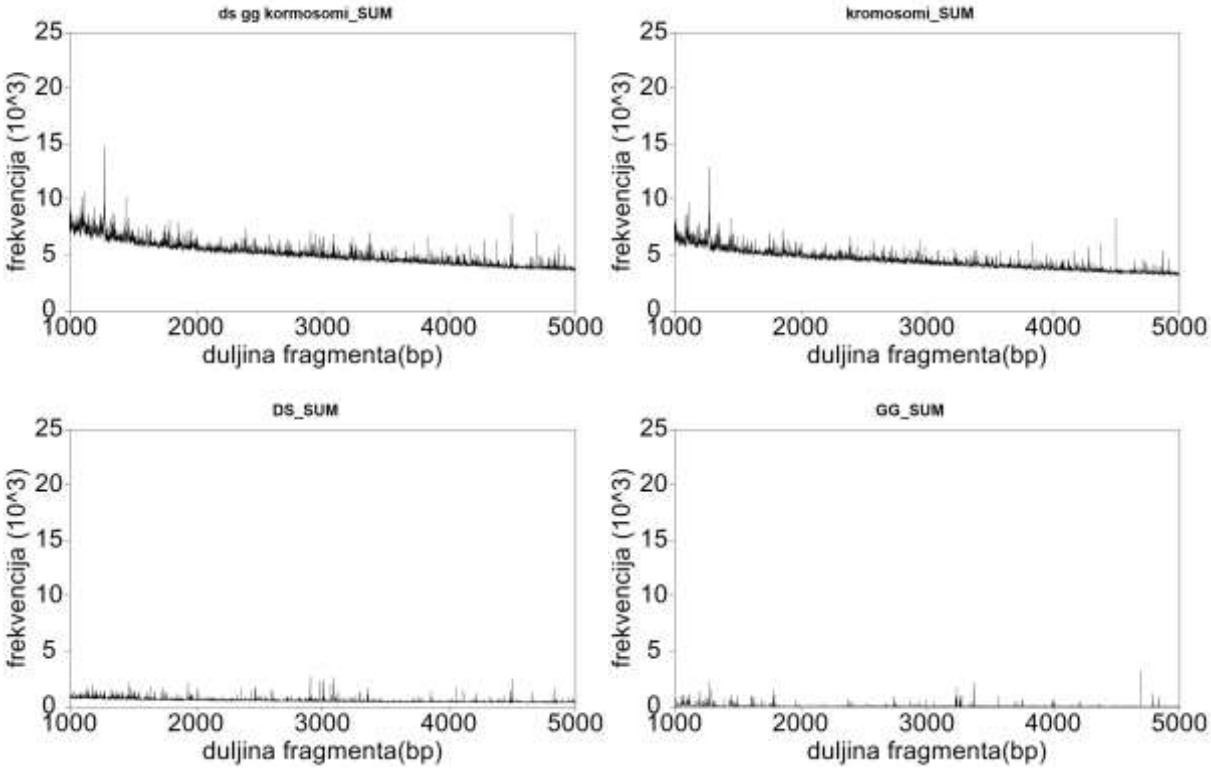
Analizom s GRM metodom svih komponenti stvorili smo superpozicijom jedan GRM dijagram na kojemu smo vidjeli koje su najčešće repeticije koje se pojavljuju unutar

sekvencioniranog genoma. Zatim smo stvorili GRM dijagrame za svaku skupinu komponenti zasebno kako bi vidjeli, u odnosu na ukupni GRM dijagram, u kojoj skupini komponenti se ističe pojedina repeticija. Radi bolje usporedbe GRM dijagrami su podijeljeni na četiri raspona duljina fragmenata – od 0 - 1000 bp, 1000 bp - 5000 bp, 5000 bp – 10000 bp i od 10 kbp - 100 kbp. Pomoću GRM metode najčešće se koristi upravo raspon duljina fragmenata od 0 - 100 000 bp, iako program može gledati i duljine fragmenata koje su puno veće (~milijun bp).

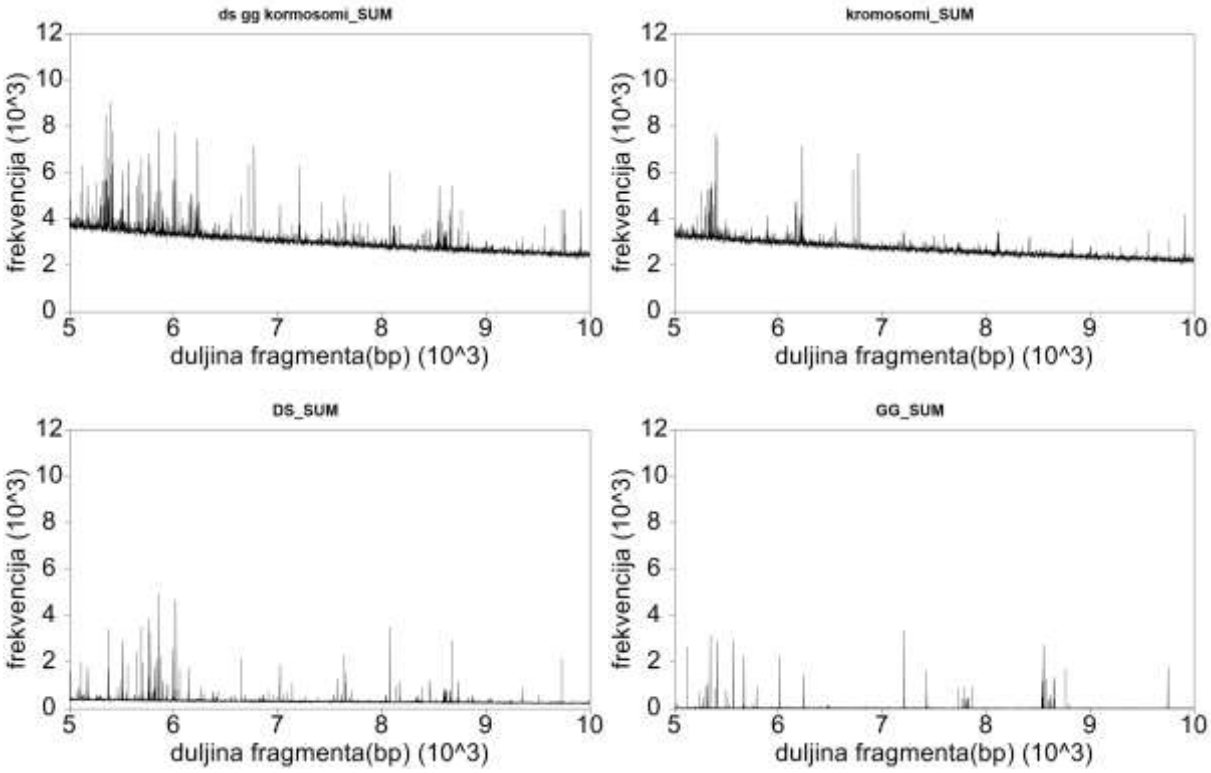
a)



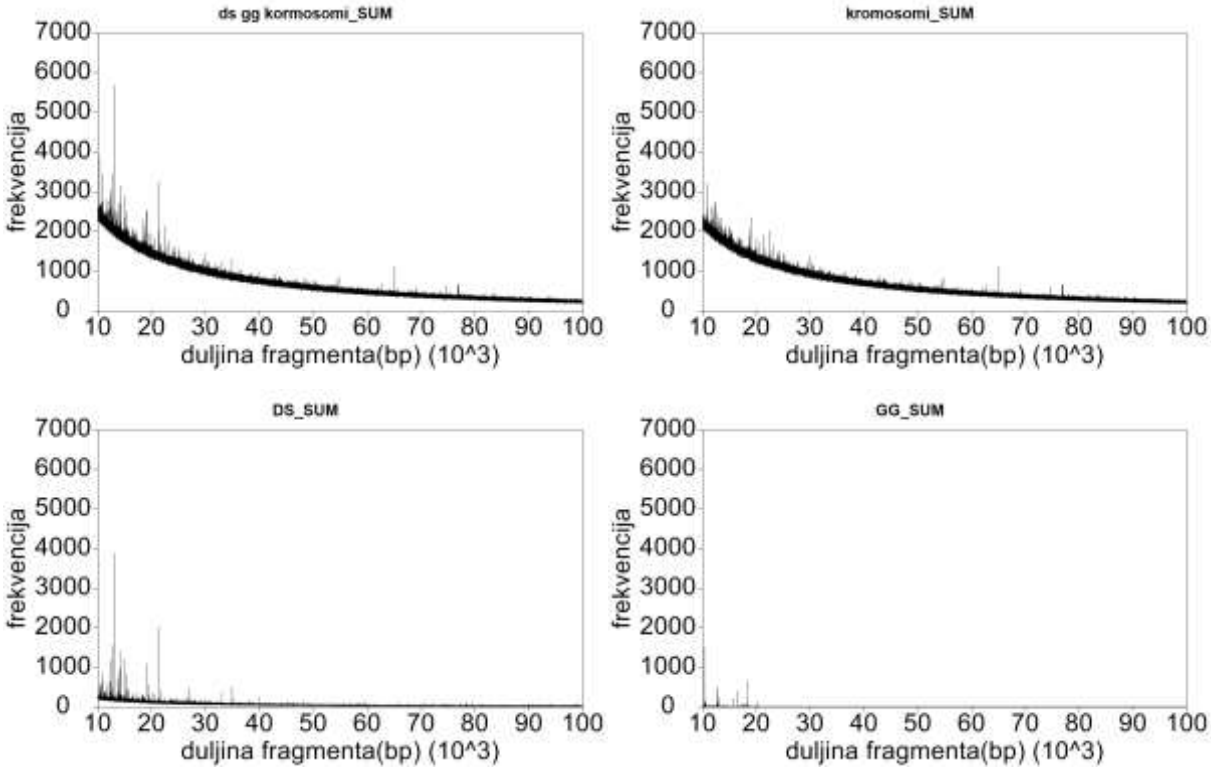
b)


ds gg kormosomi_SUM
kromosomi_SUM
DS_SUM
GG_SUM
frekvencija (10^3)
duljina fragmenta(bp)

c)


ds gg kormosomi_SUM
kromosomi_SUM
DS_SUM
GG_SUM
frekvencija (10^3)
duljina fragmenta(bp) (10^3)

d)

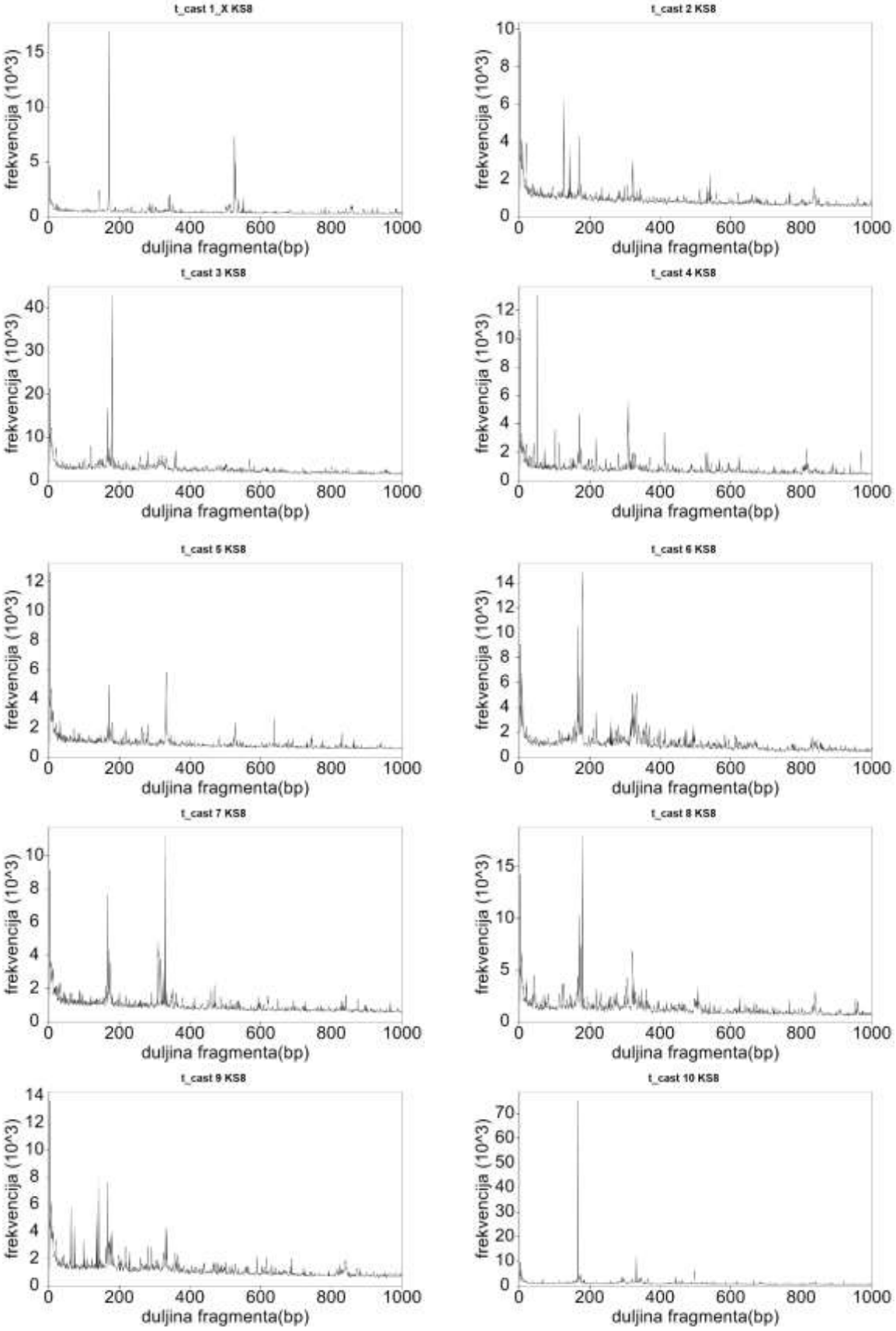


Slika 24. Prikaz GRM dijagrama za cijeli sekvencionirani genom Tcas 3.0 za duljine fragmenta od a) 0 -1000 bp, b) 1000 bp - 5000 bp, c) 5000 bp – 10000 bp, d) 10000 bp – 100000 bp.

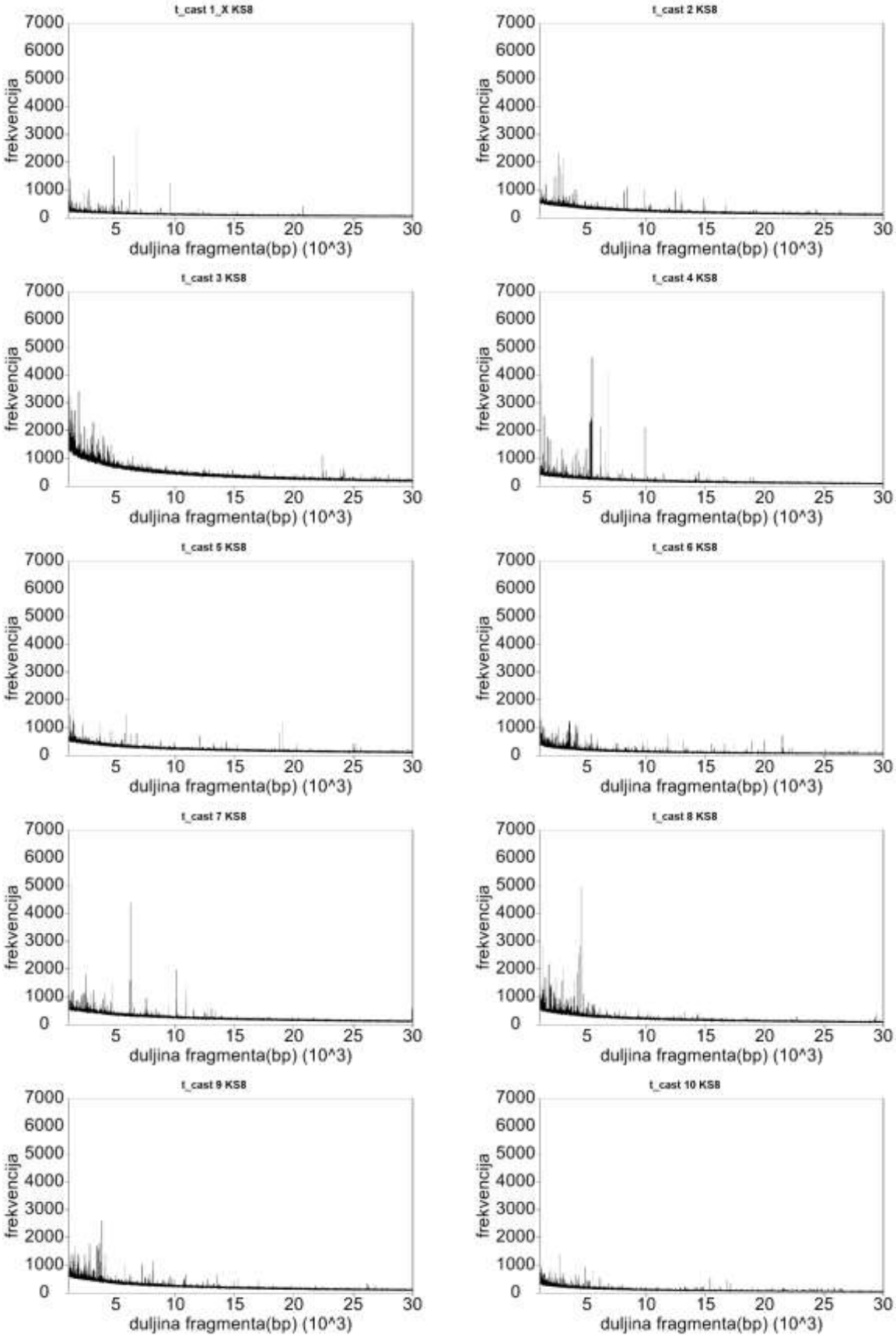
Također se sa slike 24 može vidjeti da su pikovi u nepovezanim višestruko – komponentnim povezanim kontizima (scaffolds) i nepoznatim jednostrukim spojenim kontizima izraženiji za duljine fragmenta do 20 kbp, što je logično pošto su njihove ukupne duljine manje od duljina kromosoma (tablica 1 u poglavlju „Dodatak“).

Nakon gledanja frekvencija ukupnog GRM dijagrama, isti postupak se radi i za pregledavanje komponenti koje pripadaju u pojedinu grupu, npr. za sve kromosome (10) kao što je prikazano na slici 25, kao i sve nepovezane višestruko – komponentne povezane kontige (DS47665 - DS497969, njih 305) i nepoznate jednostruke spojene kontige (GG694051–GG695897, njih 1848).

a)


t_cast 1_X KS8
t_cast 2 KS8
t_cast 3 KS8
t_cast 4 KS8
t_cast 5 KS8
t_cast 6 KS8
t_cast 7 KS8
t_cast 8 KS8
t_cast 9 KS8
t_cast 10 KS8
frekvencija (10^3)
duljina fragmenta(bp)

b)



Slika 25. Prikaz GRM dijagrama za raspon od a) 0 - 1000 bp, b) 1000 bp - 30000 bp.

4.1 TCAST sateliti ~ 360 bp

Sateliti TCAST iz insekta *T.castaneum* nalaze se velikim dijelom u pericentromernom području kao i u centromernom području na kromosomima u dvije obitelji Tcast1a (377 bp) i Tcast1b (362 bp), čije konsenzusne sekvence imaju međusobnu prosječnu sličnost od 79% te zauzimaju ~35% genoma, kao što su pokazala prethodna istraživanja [31]. Ovi sateliti se mogu pojaviti kao raspršene kopije ili kao tandemi. Prijašnja istraživanja repeticija pomoću različitih bioinformatičkih algoritama (TRF, TePipe, Repeat Scout) u sekvencioniranom genomu ovog insekta, pokazala su da TCAST sateliti čine 0.3% sastavljenog genoma, s naglaskom da su se gledale samo komponente kromosoma.

Konsenzusne sekvence TCAST satelita prikazane su u tablici 11.

Tablica 11. Prikaz konsenzusne sekvence za Tcast1a i Tcast1b satelite. Slovo Y može mijenjati nukleotide C ili T, slovo W nukleotide A ili T, slovo R nukleotide A ili G i slovo K nukleotide T ili G. Preuzeto iz [31].

Tcast 1a	**AACCATAAGCGAGTTATARAGTTGGATATAAATAATATTTAACAAAAACTTGCTTAAAAATAAAATGTTTGAAAT ACTGAAATTAAACTAAATTAATTGTGTCTAACACTTTAGTAGAGRAAAAAGRAARAGACATAAAAGTCACYAAAG KYTTCAGARTCGTYTTTAGTCGTCATATTAATAACAATAACATTAATAATAATAATACTAATAGTACTAAATYTGC TTATATGCAARCGCTTWATTAAGGTAACAGTKTTTTGGTTTCTTGCTTATATCTCGAAAACAGTTACTCCTATSAA TTTTTRTCTTTYTTTCAAAATTAAAGCTGATAAAATWTYCTACAAAATAGTTATTTGCATTTTTYATGTAGGACT**
Tcast 1b	AACCATAAGCGAGATATAAGTTTGAAAATAATTAATATTTAAMAAAAAAGTGCTTTAACAGAAAATGTCTTGTG- ATTTAAAATACACTTAATTTATTGGGTCCAATACTTTATTARAARAAAAAGGRGGKGACATAAAAGTCMCTGAAGTCYT CARAGTCGTTTTTAAATGCTGCATTTTCGTCTTCGTCTCATACATTGAAAAC-TGAATTACATT--- TTTGTTCWGTAACWGTCACAGTTTT--------- YWYWTTGGTTTTTTGCTTATATCTCRAAAACAGTTACTCCTATCAATTTTTATCTTTYTKTYAAAATTAAAGCTGATARM ATTTCCTACAAAATAGTTATTKGCATTTTTYTTGTAGGACC

Primjenom GRM metode identificirali smo satelite u Tcas 3.0, koje se nalaze i u komponentama koje nisu dodane na kromosome. Prema podacima o duljinama satelita Tcast1a i Tcast1b, tražili smo te pikove u svim komponentama Tcas 3.0. Pregledavanjem GRM dijagrama kromosoma i traženjem dominantnog K-stringa u njima za duljine 362 bp i 377 bp, nismo pronašli tandemne repeticije koje pripadaju TCAST satelitima.

Iz ostalih analiziranih komponenta genoma Tcast 3.0 - DS47665 - DS497969, GG694051–GG695897, preko podataka za frekvencije za pik 377 bp i 362 bp, koje smo u MS Excelu filtrirali s uvjetom da je frekvencija >100 dobili smo da takvih pikova ima samo u tri komponente (tablica 12.).

Tablica 12. Prikaz frekvencija za peak fragmentne duljine 362 bp i 377 bp većih od 100 u svim komponentama Tcas 3.0. U tablici se vidi i broj pogodaka s BLAST-om za frekvencije veće od 100 za ove fragmentne duljine.

Komponenta	Frekvencija pika 362	Frekvencija pika 377	Broj BLAST pogodaka
ds gg kormosomi_SUM.grm	22338	25017	
DS497867.1 KS8.grm	133	131	5
GG694360.1 KS8.grm	190	764	
GG695145.1 KS8.grm	163	524	
t_cast 10 KS8.grm	1097	796	
t_cast 1_X KS8.grm	375	358	
t_cast 2 KS8.grm	961	834	
t_cast 3 KS8.grm	3159	2630	
t_cast 4 KS8.grm	654	698	
t_cast 5 KS8.grm	841	909	
t_cast 6 KS8.grm	1034	776	
t_cast 7 KS8.grm	1605	1432	
t_cast 8 KS8.grm	1911	926	
t_cast 9 KS8.grm	1454	1263	

Ako uzmemo u obzir da se u nekoj komponenti genomske sekvence ne moraju nalaziti i Tcast1a i Tcast1b sateliti, već samo jedan od njih, primjenom filtera za svaku fragmentnu duljinu, posebno za Tcast1a i Tcast1b, dobili smo slijedeći rezultat prikazan u tablici 13.

Tablica 13. Prikaz filtriranih frekvencija >100 samo za a) pik 362, b) 377.

a)

Komponenta	Frekvencija pika 362	Frekvencija pika 377	Broj BLAST pogodaka
ds gg kormosomi_SUM.grm	22338	25017	
DS497672.1 KS8.grm	803	16	2
DS497678.1 KS8.grm	182	12	
DS497699.1 KS8.grm	127	44	5
DS497867.1 KS8.grm	133	131	5
GG694191.1 KS8.grm	112	0	
GG694275.1 KS8.grm	224	4	4
GG694323.1 KS8.grm	106	40	
GG694360.1 KS8.grm	190	764	
GG694430.1 KS8.grm	629	0	5
GG694708.1 KS8.grm	393	0	
GG694816.1 KS8.grm	152	0	1
GG694900.1 KS8.grm	119	0	2
GG694903.1 KS8.grm	116	0	5

GG695047.1 KS8.grm	157	0	
GG695097.1 KS8.grm	148	78	
GG695145.1 KS8.grm	163	524	
GG695161.1 KS8.grm	101	29	
GG695393.1 KS8.grm	108	0	4
GG695424.1 KS8.grm	167	8	
GG695436.1 KS8.grm	106	77	
GG695639.1 KS8.grm	220	0	
GG695747.1 KS8.grm	242	0	3
GG695753.1 KS8.grm	167	0	
GG695826.1 KS8.grm	368	1	30
t_cast 10 KS8.grm	1097	796	
t_cast 1_X KS8.grm	375	358	
t_cast 2 KS8.grm	961	834	
t_cast 3 KS8.grm	3159	2630	
t_cast 4 KS8.grm	654	698	
t_cast 5 KS8.grm	841	909	
t_cast 6 KS8.grm	1034	776	
t_cast 7 KS8.grm	1605	1432	
t_cast 8 KS8.grm	1911	926	
t_cast 9 KS8.grm	1454	1263	

b)

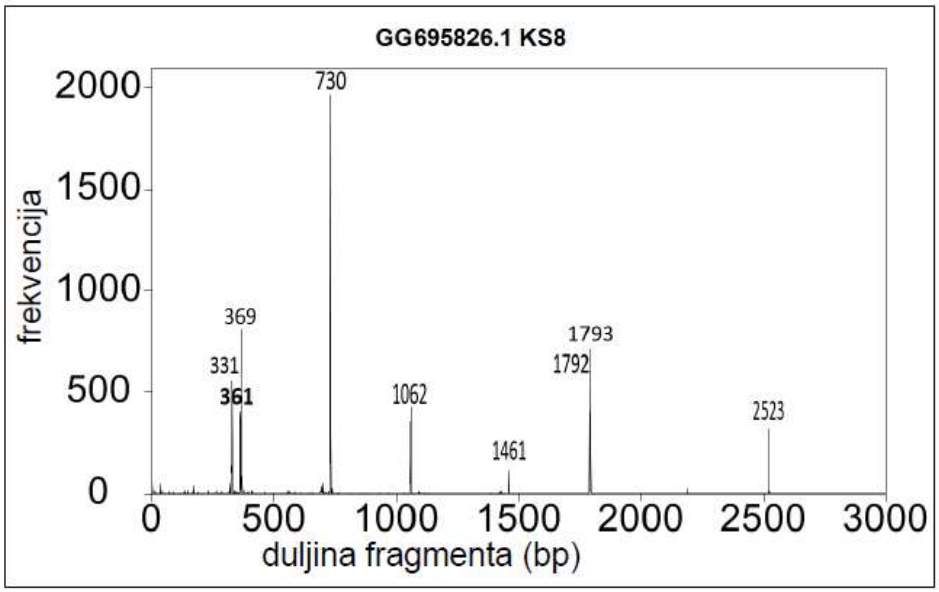
Komponenta	**Frekvencija pika 362**	**Frekvencija pika 377**	**Broj BLAST pogodaka**
ds gg kormosomi_SUM.grm	22338	25017	
DS497697.1 KS8.grm	10	232	3
DS497845.1 KS8.grm	6	112	12
DS497867.1 KS8.grm	133	131	5
DS497896.1 KS8.grm	7	263	
GG694069.1 KS8.grm	3	494	8
GG694147.1 KS8.grm	75	132	
GG694160.1 KS8.grm	3	1348	
GG694238.1 KS8.grm	84	176	
GG694256.1 KS8.grm	1	246	
GG694329.1 KS8.grm	2	176	
GG694360.1 KS8.grm	190	764	
GG694388.1 KS8.grm	15	147	6
GG694393.1 KS8.grm	17	278	
GG694403.1 KS8.grm	1	112	
GG694453.1 KS8.grm	2	114	7
GG694602.1 KS8.grm	0	125	
GG694766.1 KS8.grm	2	275	
GG694774.1 KS8.grm	1	132	
GG694835.1 KS8.grm	2	179	4
GG694837.1 KS8.grm	1	128	
GG694841.1 KS8.grm	2	256	

GG694849.1 KS8.grm	0	145
GG694896.1 KS8.grm	1	245
GG694898.1 KS8.grm	1	222
GG694901.1 KS8.grm	1	171
GG694950.1 KS8.grm	0	197
GG695096.1 KS8.grm	0	151
GG695145.1 KS8.grm	163	524
GG695212.1 KS8.grm	0	445
GG695357.1 KS8.grm	1	196
GG695513.1 KS8.grm	0	234
GG695659.1 KS8.grm	7	113
GG695706.1 KS8.grm	1	375
GG695732.1 KS8.grm	14	476
GG695836.1 KS8.grm	0	245
t_cast 10 KS8.grm	1097	796
t_cast 1_X KS8.grm	375	358
t_cast 2 KS8.grm	961	834
t_cast 3 KS8.grm	3159	2630
t_cast 4 KS8.grm	654	698
t_cast 5 KS8.grm	841	909
t_cast 6 KS8.grm	1034	776
t_cast 7 KS8.grm	1605	1432
t_cast 8 KS8.grm	1911	926
t_cast 9 KS8.grm	1454	1263

Pregledavanjem ovih komponenti pronašli smo slijedeće HOR strukture.

4.1.1 GG695826.1

U ovoj komponenti identificirali smo pravilnu periodičnost višega reda sastavljenu od pet tipa monomera duljine 331 bp, 362 bp, 369 bp, 361 bp i 369 bp fragmentacijom sa K-stringom AACCATAA. HOR struktura se sastoji od devet kopija dajući konsenzusnu duljinu od 1792 bp kao što se vidi i na GRM dijagramu za ovu komponentu (slika 26). Zanimljivo je primijetiti da se duljine monomera međusobno razlikuju za ~11%, za razliku od duljina monomera kod ljudskih alfa satelita koja iznosi ~2% i upravo te razlike upućuje na teže otkrivanje ovih kompleksnih struktura s postojećim drugim alatima za identifikaciju repeticija.



Slika 26. Prikaz frekventne domene komponente GG695826.1 s pikovima koji karakteriziraju pet tipa monomera (TCAST sateliti ~360 bp) kao i udaljenosti početnih pozicija homolognih monomera i duljinu HOR strukture.

Usporedbom kopija s Needleman – Wunsch programom dobili smo divergencije među kopijama koje su prikazane na „heat“ mapi (slika 27) te smo izračunali prosječne divergencije između monomera istih tipova (tablica 14).

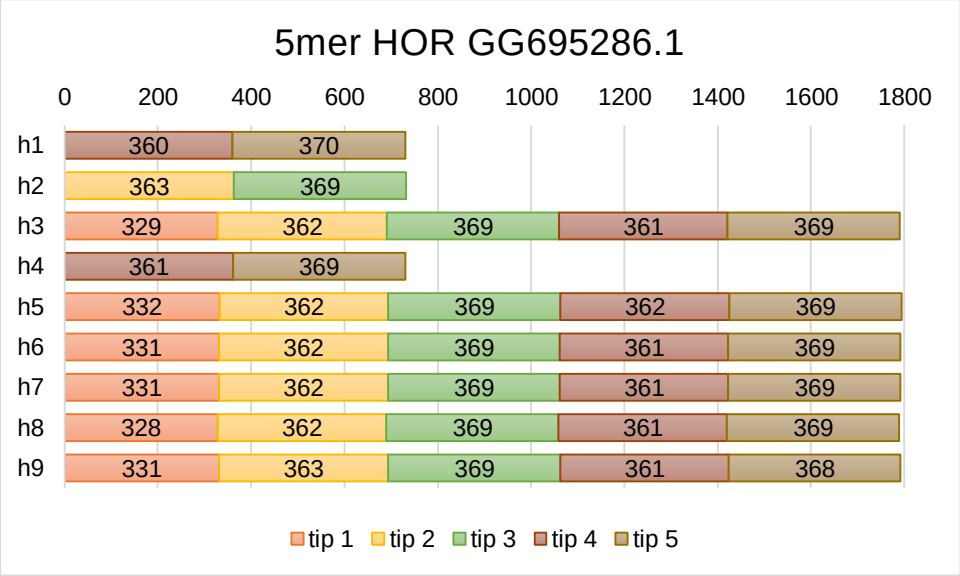
		m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24	m25	m26	m27	m28	m29	m30	m31	m32	m33	m34	m35	m36
		1	361	731	1094	1463	1792	2154	2523	2884	3253	3614	3983	4315	4677	5046	5408	5777	6108	6470	6839	7200	7569	7900	8262	8631	8992	9361	9689	10051	10420	10781	11150	11481	11844	12213	12574
m1	1	0	27	9	26	19	8	26	4	27	4,1	27	19	8,6	27	4,1	27	19	8,3	27	3,6	27	19	7,7	26	3,6	27	19	8,3	27	4,4	28	19	8,5	27	4,1	27
m2	361		0	26	5	25	26	5	26	2	26	2,2	25	26	5,1	25	1,6	24	26	4,9	26	1,4	25	25	4,3	26	2,2	25	26	5,1	26	2,2	25	26	5,1	26	1,6
m3	731			0	26	18	2	26	9	26	9,4	26	18	2,8	26	9,1	26	18	2,5	27	8,8	26	18	2,5	26	8,8	26	19	2,5	27	9,6	27	18	2,5	27	9,4	26
m4	1094				0	25	26	1	25	4	25	5,2	25	26	2,2	25	4,6	25	26	0,8	25	4,3	25	25	0,8	25	5,2	25	26	1,6	26	5,1	25	26	1,6	26	4,6
m5	1463					0	17	25	17	25	17	26	1,5	18	25	17	26	2,7	17	25	17	25	2,1	17	25	17	26	3,3	17	25	18	26	0,9	18	25	18	26
m6	1792						0	26	8	25	8,3	26	17	0,6	26	7,7	26	17	0,3	26	7,7	25	17	0,8	25	7,7	26	18	0,3	26	8,6	26	17	0,6	26	8,3	26
m7	2154							0	25	4	25	5,7	25	26	1,1	25	5,2	25	26	0,3	25	4,9	25	25	0,8	25	5,7	25	26	0,5	26	5,7	25	26	0,5	26	5,2
m8	2523								0	25	1,7	25	18	8,8	25	1,7	26	17	8,6	26	0,6	25	18	8	25	1,1	25	18	8,6	26	0,8	26	17	8,8	26	0,6	26
m9	2884									0	25	2,4	25	26	3,5	25	1,9	25	26	4,1	25	1,6	25	25	3,5	25	2,4	25	26	4,1	26	2,4	25	26	4,1	26	1,9
m10	3253										0	25	17	8,8	25	2,2	25	17	8,6	26	1,1	25	18	8	25	1,1	25	18	8,6	26	1,9	26	17	8,8	26	1,7	25
m11	3614											0	26	26	5,7	25	1,1	25	26	6	25	0,8	26	25	5,4	25	1,6	26	26	6,2	26	1,6	26	26	6,2	26	1,1
m12	3983												0	18	25	17	25	2,4	17	25	17	25	1,8	17	25	17	25	3	17	26	18	26	0,9	18	25	18	25
m13	4315													0	26	8,3	26	18	0,3	26	8,3	26	18	1,4	26	8,3	26	19	0,3	27	9,1	27	17	0,6	26	8,8	26
m14	4677														0	25	5,2	25	26	1,4	25	4,9	25	25	1,9	25	5,7	25	26	1,1	26	5,7	25	26	1,1	26	5,2
m15	5046															0	25	17	8	25	1,1	25	18	7,5	25	1,7	25	18	8	25	1,9	26	17	8,3	25	1,7	25
m16	5408																0	25	26	5,4	26	0,3	25	25	4,9	26	1,1	26	26	5,7	26	0,5	25	26	5,7	26	0
m17	5777																	0	17	25	17	25	1,2	17	24	17	25	2,4	17	25	18	26	2,1	18	25	18	25
m18	6108																		0	26	8	26	17	1,1	26	8	26	18	0	26	8,8	26	17	0,3	26	8,6	26
m19	6470																			0	26	5,2	25	25	1,1	26	6	26	26	0,8	26	6	25	26	0,8	26	5,4
m20	6839																				0	25	18	7,5	25	0,6	25	18	8	26	0,8	26	17	8,3	26	0,6	26
m21	7200																					0	25	25	4,6	25	0,8	26	26	5,4	26	0,8	25	26	5,4	26	0,3
m22	7569																						0	17	25	18	25	1,8	17	25	18	26	1,5	18	25	18	25
m23	7900																							0	25	7,5	25	18	1,1	26	8,3	26	17	1,4	25	8	25
m24	8262																								0	25	5,4	25	26	1,4	26	5,4	25	26	1,4	26	4,9
m25	8631																									0	25	18	8	26	1,4	26	17	8,3	26	1,1	26
m26	8992																										0	26	26	6,2	26	1,6	25	26	6,2	26	1,1
m27	9361																											0	18	26	19	26	2,7	18	25	18	26
m28	9689																												0	26	8,8	26	17	0,3	26	8,6	26
m29	10051																													0	26	6,2	25	27	0,5	26	5,7
m30	10420																														0	27	18	9,1	26	0,8	26
m31	10781																															0	26	27	6,2	27	0,5
m32	11150																																0	17	25	18	25
m33	11481																																	0	26	8,8	26
m34	11844																																		0	26	5,7
m35	12213																																			0	26
m36	12574																																				0

Slika 27. Prikaz „heat“ mape za GG695826.1. Prosječna divergencija među monomerima je 16,82±9,99%.

Tablica 14. Prosječna divergencija (%) između kopija monomera istog tipa i između pet različitih tipova monomera. Preuzeto iz [147].

	m1	**m2**	**m3**	**m4**	**m5**
m1	2,03	17,64	25,03	17,73	25,38
m2		1,14	25,96	8,44	25,82
m3			1,08	25,65	5,17
m4				1,94	25,78
m5					1,39

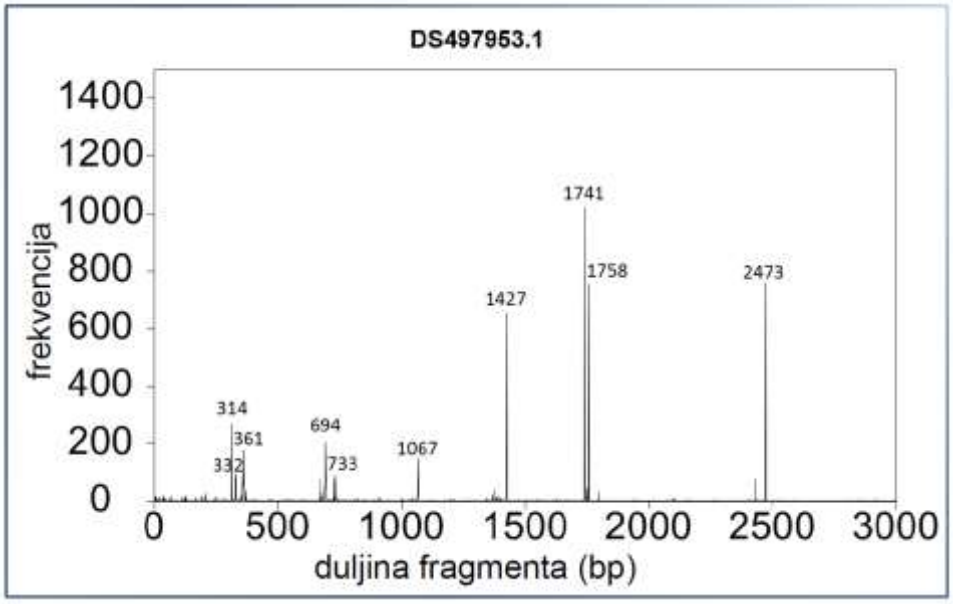
Nakon računanja divergencija između kopija za "heat" map, vidjeli smo da ti monomeri tvore periodičnost višega reda odnosno 5mer HOR duljine ~1792 bp kao što je prikazano na slici 28.



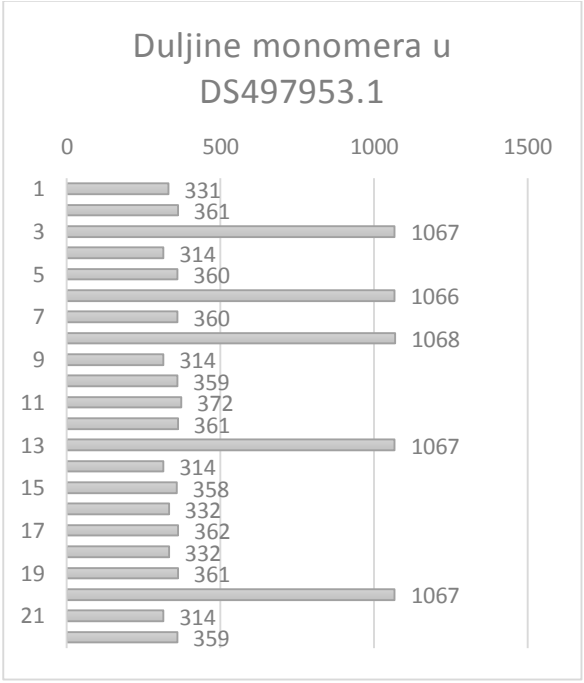
Slika 28. Shematski prikaz poravnatih TCAST monomera u pravilnu 5mer HOR strukturu. Preuzeto i prerađeno iz [147]. Divergencija između cijelih HOR kopija je 1,16±0,26%.

4.1.2 DS497953.1

Pregledavanjem GRM dijagrama komponente DS497953.1 (slika 29) za pikove 372 i 362 te odabirom dominantnog K-stringa ACTCCTAT dobili smo monomere duljina 361 bp, 314 bp, 332 bp i 1067 bp poredane kao na slici 30. Prva tri tipa monomera odgovaraju TCAST satelitima.



Slika 29. Prikaz frekventne domene komponente DS497953.1 s vrhovima koji karakteriziraju četiri tipa monomera (TCAST sateliti ~360 bp i jedan ~1067 bp). Preuzeto i prerađeno iz [147].

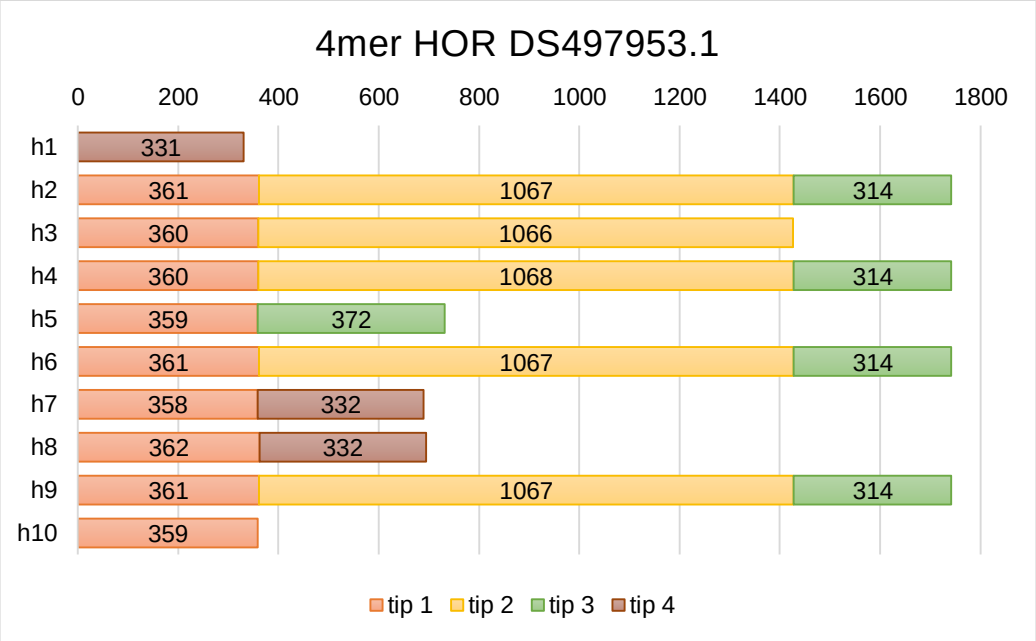


Slika 30. Duljine monomera dobivene sa K-stringom ACTCCTAT u DS497953.1. Preuzeto i obrađeno iz [147].

Iz rezultata „heat“ mape (slika 31) odredili smo kompleksnu strukturu periodičnosti višega reda kao što je prikazano na slici 32.

		m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22
		3714	4045	4406	5473	5787	6147	7213	7573	8641	8955	9314	9686	10047	11114	11428	11786	12118	12480	12812	13173	14240	14554
m1	3714	0	25	80	15	25	82	25	80	15	25	20	24	82	16	24	6,6	25	3,6	25	82	15	25
m2	4045		0	77	31	5	80	1	80	31	6,6	27	3,3	77	31	6,9	27	2,8	26	1,1	77	30	6,6
m3	4406			0	83	80	1	80	1	83	80	78	80	1,4	83	80	81	80	81	77	0,6	83	80
m4	5473				0	32	83	31	83	0	30	20	30	83	1,3	30	17	30	16	30	83	0,6	30
m5	5787					0	77	6	81	32	2,8	25	5,5	77	32	4,4	25	6,1	25	5,8	77	32	2,8
m6	6147						0	80	1	83	81	77	80	1,4	82	80	83	81	83	81	0,9	82	80
m7	7213							0	81	31	6,6	27	3,1	81	31	6,4	27	2,5	26	1,4	77	30	6,6
m8	7573								0	83	81	77	80	1,7	82	80	83	80	82	80	1,4	83	80
m9	8641									0	30	20	30	83	1,3	30	17	30	16	30	83	0,6	30
m10	8955										0	25	6,1	81	30	3,1	25	6,4	25	6,6	80	30	0,6
m11	9314											0	25	77	22	25	16	26	20	26	77	21	25
m12	9686												0	80	30	6,9	25	3,9	25	3,3	80	30	6,1
m13	10047													0	82	80	83	80	82	80	1,2	83	80
m14	11114														0	29	18	30	17	31	82	1,9	30
m15	11428															0	25	6,6	25	6,4	80	30	3,1
m16	11786																0	26	6,6	27	83	17	25
m17	12118																	0	25	2,2	80	30	6,4
m18	12480																		0	25	83	16	25
m19	12812																			0	80	30	6,6
m20	13173																				0	83	80
m21	14240																					0	30
m22	14554																						0

Slika 31. Prikaz rezultata usporedbe monomera iz DS497953.1. Prosječna divergencija monomera je 40,66±31,90%, zbog najduljeg monomera. Divergencije između homolognih monomera su za tip 1, 2, 3 i 4 redom 4,66±2,02%, 1,24±0,33%, 8,91±10,28% (s obrisanih 58 baza u 11-tom monomeru divergencija je 3,92±3,87%) i 5,61±1,74%.



Slika 32. Prikaz kompleksne strukture 4mer HOR-a u DS497953.1. Nijedna HOR kopija nema sva četiri tipa monomera. Divergencija između h2, h3, h4 i h9 je 1,35±0,48%.

Na ovu komponentu primijenili smo K-string AACCATAA kako bi usporedili ove monomere sa satelitima Tcast1a i Tcast1b pomoću BLASTN programa sa standardnim postavkama. Monomer za koji smo dobili najveći identitet sa Tcast1b satelitom nalazi se na poziciji 5873 u komponenti DS497953.1. Ovaj monomer nazvali smo DS360/5873 (tablica 15) i s njime smo pomoću programa BLAST identificirali raspršene i tandemne kopije u ostalim dijelovima Tcas 3.0. genoma, s time da smo ručno odabirali preuzete sekvence svih onih komponenti koje su dale s GRM metodom frekvenciju veću od 1 kao sekvencu „subjekt“ i rezultate smo prikazali u tablici 16.

Tablica 15. DS360/5873 sekvenca, dobivena sa K-stringom AACCATAA. Preuzeto iz [147].

AACCATAAGCGAGATATAAGTTTGAAAATAATTAATATTAAAAAAAAGTACTT
TGACAGAAAATGTCTTGTGATTTATAATACACTTAATTTATTGGGTCCAATATTT
TATTAGAAGAAAAAGGAGGTGACATAAAAGTCACTGAAGTCTTCAGAGTCGTT
TTTAAATGCTGCATTTTTGTCTTAGTCTCAAATATTGAAAACTGAATTACATTTT
TGTTCAGTAACAGTTACAGTTTTCTTATTGGTTTTTTGCTTATATCTCGAAACCA
GTTATTCCAAATAATTTTTATCTTTATTTCAAAATTAATGCTGATAAAATTTCCT
ACAAAATCGTTATTTGCATTTTTCATGTAGGACC

Tablica 16. Prikaz rezultata dobivenog s BLASTN programom za identifikaciju TCAST satelita raspršenih i tandemnih u a) LG (linkage grupe) kromosomima, b) DS47665 - DS497969, c) GG694051–GG695897. Broj monomera je prikazan u pojedinoj komponenti za monomere koji su veći od 355 bp. Preuzeto i prerađeno iz [147].

LG1 –CM000276.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orijentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092770.1	1137974	1137879	-	91	95	-
2		1138101	1137987	-	84	114	
3		1138622	1138739	+	81	117	
4		1138741	1138938	+	79	197	
5		1270059	1269972	-	79	87	
6	001093327.1	877178	877276	+	82	98	-
7		877278	877514	+	88	236	
8		878078	877885	-	79	193	
9		1049744	1049453	-	92	291	
10		1049885	1049762	-	84	123	

LG2 - CM000277.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orijentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092771.1	54424	54263	-	75	161	1
2		54784	54425	-	91	359	
3		537267	537392	+	84	125	
4		560539	560649	+	78	110	
5		561039	560951	-	89	88	
6		561261	561141	-	74	120	
7		690622	690739	+	79	117	
8		691240	691030	-	80	210	
9	001092773.1	146830	146718	-	74	112	-
10		146942	146838	-	87	104	
11		570585	570678	+	72	93	
12		570680	570918	+	75	238	
13		571244	571171	-	81	73	
14		645442	645510	+	81	68	
15		646187	646310	+	81	123	
16		646312	646543	+	76	231	
17		648427	648556	+	83	129	
18	001092774.1	232019	232142	+	81	123	-

19		232149	232319	+	74	170	
20		234811	234698	-	74	113	
21	001092775.1	2679711	2679832	+	82	121	-
22		2679834	2680069	+	79	235	
23		2680385	2680289	-	78	96	
24	001092778.1	1554476	1554593	+	74	117	-
25		1555017	1554875	-	79	142	
26	001093459.1	46628	46471	-	74	157	-
27		46756	46629	-	89	127	
28	001093470.1	25700	25847	+	82	147	-
29		26273	26174	-	79	99	
30	001093833.1	5	231	+	78	226	-
31		474	358	-	76	116	
32		601	476	-	80	125	
33		1134	908	-	75	226	
34		1264	1137	-	85	127	
35		1631	1445	-	82	186	
36		1756	1633	-	84	123	

LG3 – CM000278.2

Rb.	**Kontig (NW)**	**Početna pozicija**	**Krajnja pozicija**	orjentacija	**Identitet (%)**	**Duljina sekvence**	Broj Tcast>355
1	001092794.1	1789869	1789990	+	80	121	-
2		1790473	1790284	-	81	189	
3		2305436	2305201	-	74	235	
4		2305549	2305438	-	75	111	
5	001092796.1	208809	208573	-	78	236	-
6		208924	208811	-	77	113	
7		322389	322510	+	83	121	
8		322516	322751	+	77	235	
9		323164	323057	-	81	107	
10		437562	437672	+	74	110	
11		438148	437976	-	78	172	
12		456368	456244	-	77	124	
13		475066	475180	+	79	114	
14		475187	475418	+	77	231	
15		475845	475724	-	76	121	
16		481364	481468	+	78	104	
17		482027	481798	-	75	229	
18		838792	838917	+	83	125	
19		838919	839129	+	71	210	
20		1058751	1058870	+	84	119	
21		1058975	1059122	+	82	147	
22		1623782	1623655	-	81	127	
23	001092797.1	474169	474296	+	73	127	-
24		596499	596382	-	77	117	
25		845986	846110	+	82	124	
26		846113	846353	+	76	240	
27		846775	846658	-	80	117	
28		1134020	1134140	+	77	120	
29		1136308	1136151	-	74	157	
30		1136417	1136312	-	75	105	
31		1390563	1390670	+	81	107	
32		1390675	1390903	+	83	228	
33	001092798.1	76394	76163	-	80	231	
34	001092800.1	24779	24668	-	74	111	-
35		339772	339546	-	79	226	
36		339902	339774	-	84	128	
37		340551	340671	+	84	120	
38		763462	763589	+	75	127	
39		764144	763905	-	76	239	
40		764256	764146	-	81	110	
41		2172183	2172300	+	80	117	
42		2172828	2172593	-	78	235	
43		2172953	2172830	-	82	123	
44	001092801.1	678177	678286	+	74	109	2
45		678361	678497	+	78	136	
46		835467	835579	+	87	112	
47		835580	835745	+	74	165	
48		835752	836109	+	89	357	
49		836111	836469	+	92	358	
50		836470	836637	+	76	167	
51		836856	836994	+	86	138	
52	001092802.1	122810	122935	+	83	125	-
53		122937	123165	+	77	228	
54		123576	123459	-	77	117	
55	001092803.1	455887	455997	+	79	110	-
56		456539	456305	-	79	234	
57		456659	456541	-	82	118	
58	001092806.1	421683	421571	-	84	112	-
59		913040	912814	-	82	226	
60		913161	913048	-	84	113	
61	00109805.1	159091	158973	-	88	118	-
62	00109808.1	343918	343681	-	78	237	-
63		344030	343920	-	81	110	
64	001092811.1	1254220	1254003	-	91	217	-
65		1254347	1254222	-	84	125	
66	001092816.1	333240	333602	+	91	362	1
67		333603	333770	+	77	167	
68		1175375	1175429	+	80	54	
69		1175431	1175667	+	78	236	
70	001093371.1	9768	9892	+	80	124	-
71		9895	10131	-	76	236	
72		10554	10458	-	84	96	
73		30674	30879	+	74	205	
74		88572	88689	+	76	117	
75		88691	88916	+	71	225	
76		111754	111852	+	81	98	
77		112414	112179	-	81	235	
78		112541	112418	-	82	123	
79		542128	542321	+	80	193	
80		545324	545215	-	79	109	
81		1093804	1094037	+	73	233	
82		1094459	1094342	-	80	117	
83		1201003	1200783	-	76	220	
84	001093514.1	110676	110909	+	77	233	-
85		121151	121260	+	72	109	
86	001093965.1	538086	537777	-	70	309	-
87		538379	538087	-	90	292	
88	001092810.1	182318	182432	+	75	114	-
89	001093448.1	213141	213251	+	79	110	-
90		214666	214444	-	75	222	
91		214794	214668	-	83	126	
92		249500	249308	-	81	192	

LG4 –CM000279.1

Rb.	**Kontig (NW)**	**Početna pozicija**	**Krajnja pozicija**	orjentacija	**Identitet (%)**	**Duljina sekvence**	Broj Tcast>355
1	001092817.1	539017	539386	+	89	369	1
2	001092818.1	370560	370684	+	82	124	-
3		370687	370913	+	90	226	
4		412412	412231	-	95	181	
5		412767	412414	-	73	353	
6		1267934	1268064	+	95	130	
7	001092820.1	1780771	1780888	+	79	117	-
8		1781393	1781192	-	72	201	
9	001092825.1	166404	166295	-	77	109	-
10		166530	166406	-	83	124	
11		166975	167086	+	88	111	
12	001092821.1	1735418	1735223	-	73	195	-
13		1735539	1735435	-	76	104	
14		1858770	1858888	+	77	118	
15		1858911	1859111	+	74	200	
16		1984625	1984507	-	82	118	
17	001092819.1	680256	680039	-	76	217	-
18		680382	680258	-	85	124	
19		680906	681035	+	85	129	
20		681037	681228	+	76	191	
21		1045818	1045700	-	82	118	
22	001092822.1	234779	234904	+	81	125	-
23		234908	235141	+	76	233	
24		278383	278240	-	72	143	
25		278503	278385	-	72	118	
26		438545	438660	+	82	115	
27		438662	438899	+	81	237	

28		523650	523774	+	83	124	
29		523777	524012	+	78	235	

LG5 – CM000280.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092826.1	162126	162294	+	72	168	-
2		163884	164009	+	84	125	
3		164011	164245	+	78	234	
4		164654	164555	-	80	99	
5	001092827.1	996888	996645	-	81	243	-
6		997014	996890	-	83	124	
7	001092831.1	175147	175039	-	74	108	-
8	001092836.1	671120	670880	-	93	240	-
9	001093404.1	138654	138836	+	71	182	-
10		342916	342787	-	85	129	
11	001092834.1	182588	182357	-	78	231	-

LG6 – CM000281.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092848.1	779584	779448	-	73	136	1
2		779942	779586	-	90	356	
3	001092850.1	372714	372834	+	85	120	-
4		372838	373072	+	80	234	
5		374407	374299	-	74	108	
6		409202	408966	-	79	236	
7		409329	409204	-	83	125	
8		421656	421781	+	77	125	
9		422321	422085	-	78	236	
10		422445	422323	-	84	122	
11	001092852.1	397138	397252	+	78	114	-
12		397791	397557	-	77	234	
13		397914	397790	-	81	124	
14		805870	805767	-	86	103	
15		806404	806528	+	86	124	
16		806530	806769	+	85	239	
17		832816	832940	+	80	124	
18		832942	833178	+	78	236	
19		834132	834012	-	77	120	
20		1236082	1236189	+	70	107	
21	001092853.1	168531	168679	+	76	148	-
22		239474	239584	+	76	110	
23		239643	239767	+	73	124	
24	001092851.1	587038	586928	-	83	110	-

LG7 – CM000282.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092855.1	3437247	3437049	-	88	198	-
2	001092861.1	82354	82639	+	291	285	1
3		82641	82999	+	90	358	
4	001092864.1	369228	369062	-	79	166	3
5		369606	369230	-	74	376	
6		369983	369607	-	73	376	
7		370359	369984	-	73	375	
8		370485	370360	-	84	125	
9	001092868.1	885764	885882	+	77	118	-
10		886421	886187	-	78	234	
11		886536	886423	-	80	113	
12		892758	892872	+	71	114	
13		893438	893202	-	78	236	
14		893558	893440	-	85	118	
15	001093888.1	277891	277718	-	78	173	-
16		407464	407579	+	76	115	
17		408098	407865	-	77	233	
18		408208	408100	-	80	108	
19	001092856.1	222081	221960	-	77	121	-
20		274535	274418	-	80	117	
21		420620	420476	-	70	144	
22	001092857.1	421516	421611	+	75	95	-
23		422605	422492	-	74	113	
24	001092867.1	789682	789561	-	80	121	-

LG8 – CM000283.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092869.1	71093	71219	+	86	126	1
2		71221	71583	+	92	362	
3		71584	71711	+	72	127	
4		517978	518096	+	87	118	
5		518142	518317	+	81	175	
6		518735	518622	-	79	113	
7		1665995	1665773	-	71	222	
8		1666122	1665997	-	83	125	
9		2064091	2064287	+	80	196	
10		2064449	2064351	-	84	98	
11		2338442	2338548	+	77	106	
12		2339188	2338953	-	80	235	
13		2339315	2339195	-	84	120	
14		2425520	2425637	+	77	117	
15	001092875.1	631104	631301	+	79	197	-
16	001092876.1	337575	337690	+	77	115	-
17		337691	337920	+	71	229	
18		338300	338192	-	74	108	
19	001092871.1	9740	9858	+	79	118	-
20		9860	10095	+	78	235	
21	001093360.1	774899	775015	+	80	116	-
22		775022	775248	+	79	226	
23		896125	896318	+	90	193	
24	001093679.1	253099	253337	+	81	238	-

LG9 – CM000284.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092882.1	80778	80543	-	77	235	-
2		81135	81019	-	81	116	
3	001092884.1	255889	255764	-	85	125	-
4	001092885.1	45535	45410	-	84	125	-
5	001092888.1	1678954	1679066	+	88	112	-
6		1679067	1679403	+	70	336	
7		1679405	1679667	+	95	262	
8	001092892	924168	924363	+	87	195	-
9	001092889.1	25792	25919	+	77	127	-
10		26444	26224	-	78	220	
11		26562	26446	-	80	116	
12		156604	156475	-	83	129	
13		157138	157241	+	81	103	
14		157255	157348	+	84	93	
15		811501	811693	+	79	192	
16		812119	811990	-	78	129	
17		813003	812877	-	88	126	
18		815982	815787	-	92	195	
19		816090	815984	-	89	106	
20		962761	962417	-	92	344	
21		962883	962760	-	85	123	
22	001092893.1	927292	927667	+	73	375	3
23		927669	928030	+	91	361	
24		928032	928199	+	77	167	
25		1090345	1090463	+	91	118	
26		1090465	1090826	+	93	361	
27		1090827	1090994	+	76	167	
28	001092894.1	391748	391950	+	93	202	-
29	001093393.1	37098	36865	-	77	233	-
30		37211	37104	-	76	107	
31	001093415.1	78304	78175	-	87	129	
32		78814	78922	+	85	108	
33		369157	368925	-	76	232	
34		369284	369159	-	79	125	
35		371809	371575	-	77	234	
36		371936	371811	-	80	125	
37	001093481.1	25586	25685	+	86	99	1
38		25904	26027	+	83	123	
39		26029	26376	+	87	347	
40		44783	44418	-	85	365	
41		74778	74999	+	70	221	
42		553720	553549	-	76	171	
43		553823	553722	-	73	101	
44	001093602.1	92386	92496	+	77	110	-
45		93037	92801	-	77	236	
46		93164	93039	-	82	125	
47		352145	352027	-	81	118	
48	001093789.1	27744	27534	-	79	210	-
49		27900	27774	-	84	126	
50		514496	514283	-	78	213	
51		514637	514528	-	83	109	

LG10 – CM000285.2

Rb.	Kontig (NW)	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	001092756.1	29131	28945	-	72	186	2
2		29274	29169	-	81	105	
3		442487	442701	+	75	214	
4		606983	607207	+	87	224	
5		1009264	1009014	-	92	250	
6		1009641	1009266	-	72	375	
7		1166004	1165877	-	73	127	
8		1166361	1166005	-	91	356	
9	001092765.1	181510	181617	+	76	107	-
10		182127	181911	-	76	216	
11		182258	182129	-	78	129	
12	001092764.1	77877	77984	+	78	107	-
13	001093338.1	40336	40554	+	93	218	1
14		40555	40930	+	73	375	
15		40932	41039	+	88	107	
16	001094327.1	70105	69870	-	77	235	-
17		70232	70108	-	78	124	

b) DS47665-DS497969

Rb.	DS	Početna pozicija	Krajnja pozicija	orjentacija	Identitet (%)	Duljina sekvence	Broj Tcast>355
1	DS497666.1	114091	114210	+	79	119	-
2		114221	114410	+	72	189	
3	DS497669.1	158530	158645	+	75	115	-
4		158651	158884	+	76	233	
5		1559430	159313	-	84	117	
6	DS497672.1	181212	181099	-	89	113	2
7		181367	181216	-	89	151	
8		181729	181369	-	88	360	
9		182091	181731	-	72	360	
10		182359	182092	-	75	267	
11	DS497674.1	328911	329241	+	89	330	-
12		329368	329505	+	76	137	
13	DS497678.1	2934	3165	+	91	231	-
14	DS497684.1	69596	69445	-	69	151	-
15		69720	69603	-	74	117	
16	DS497690.1	95689	95806	+	79	117	-
17		96327	96110	-	76	217	
18		96451	96329	-	82	122	
19		97424	97265	-	80	159	
20		97555	97426	-	81	129	
21	DS497695.1	53332	53096	-	84	236	-
22		53457	53334	-	80	123	
23		56267	56134	-	81	133	
24		56489	56375	-	76	114	
25		56913	57042	+	81	129	
26		57811	57632	-	81	179	
27	DS497697.1	251476	251309	-	74	167	3
28		251853	251480	-	70	373	
29		252213	251855	-	86	358	
30		252590	252215	-	72	375	
31		252710	252591	-	87	119	
32	DS497699.1	111	1	-	85	110	5
33		13170	13005	-	78	165	
34		13533	13172	-	89	361	
35		13910	13535	-	71	375	
36		14272	13911	-	90	361	
37		14648	14273	-	73	375	
38		15024	14649	-	73	375	
39		15141	15025	-	90	116	
40		190371	190478	+	75	107	
41	DS497703.1	1934	1816	-	82	118	-
42		26114	26322	+	75	208	
43		26740	26624	-	77	116	
44	DS497722.1	4793	4915	+	87	122	-
45		4921	5061	+	93	140	
46		5060	5350	+	90	290	
47		5352	5486	+	71	134	
48		12486	123776	-	94	110	
49		12832	12488	-	69	344	
50	DS497727.1	8962	9088	+	88	126	-
51		61588	61394	-	76	194	
52		61734	61616	-	78	118	
53	DS497760.1	365	1	-	71	364	5
54		728	366	-	84	362	
55		880	765	-	88	115	
56		5785	5452	-	83	333	
57		6162	5878	-	71	284	
58		6526	6163	-	84	363	
59		6887	6527	-	90	360	
60		7262	6889	-	72	373	
61		38176	38296	+	86	120	
62		38302	38437	+	90	135	

63	DS497763.1	3185	2988	-	77	**197**	-
64		3315	3188	-	82	**127**	
65		3861	4045	+	91	**184**	
66		4789	4457	-	90	**332**	
67		4926	4791	-	86	**135**	
68		13443	13141	-	71	**302**	
69	DS497765.1	42944	43293	+	91	**349**	-
70		42816	42942	+	85	**126**	
71	DS497786.1	14182	14328	+	84	**146**	-
72		15389	15190	-	75	**199**	
73		15545	15391	-	82	**154**	
74	DS497799.1	49801	49924	+	81	**123**	**1**
75		50548	50322	-	79	**226**	
76		50672	50555	-	83	**117**	
77		51198	51327	+	76	**129**	
78		51329	51689	+	87	**360**	
79		51689	51852	+	75	**163**	
80		52233	52363	+	80	**130**	
81		52367	52574	+	76	**207**	
82		53050	52941	-	79	**109**	
83	DS497832.1	3115	2763	-	71	**352**	**1**
84		9794	9627	-	73	**167**	
85		10155	9795	-	88	**360**	
86		10280	10157	-	81	**123**	
87	DS497845.1	141	1	-	90	**140**	**12**
88		269	143	-	87	**126**	
89		406	764	+	91	**358**	
90		765	1129	+	73	**364**	
91		1133	1492	+	92	**359**	
92		1493	1860	+	72	**367**	
93		1862	2222	+	91	**360**	
94		2223	2590	+	72	**367**	
95		2592	2729	+	93	**137**	
96		3135	2807	-	73	**328**	
97		3512	3137	-	70	**375**	
98		3869	3513	-	89	**356**	
99		4163	3871	-	72	**292**	
100		4721	4554	-	73	**167**	
101		5082	4722	-	92	**360**	
102		5457	5084	-	74	**373**	
103		5835	5458	-	70	**377**	
104		6195	5836	-	92	**359**	
105		6488	6197	-	72	**291**	
106		7046	6879	-	72	**167**	
107	DS497847.1	5653	6022	+	72	**369**	**1**
108	DS497867.1	6150	6025	-	73	**125**	**5**
109		6497	6330	-	78	**167**	
110		6859	6498	-	89	**361**	
111		7236	6861	-	73	**375**	
112		7598	7237	-	89	**361**	
113		7975	7600	-	74	**375**	
114		8335	7976	-	88	**359**	
115		8452	8337	-	89	**115**	
116	DS497897.1	22180	22013	-	74	**167**	
117		22308	22181	-	86	**127**	
118	DS497953.1	2603	2733	+	95	**130**	**10**
119		3800	4129	+	73	**329**	
120		4131	4491	+	95	**360**	
121		5759	5871	+	88	**112**	
122		5873	6232	+	100	**359**	
123		7209	7296	+	88	**87**	
124		7298	7658	+	95	**360**	
125		8637	8729	+	88	**92**	
126		8927	9039	+	88	**112**	
127		9041	9399	+	94	**358**	
128		9400	9769	+	72	**369**	
129		9771	10132	+	93	**361**	
130		11110	11199	+	88	**89**	
131		11400	11512	+	88	**112**	
132		11514	11871	+	93	**357**	
133		11872	12202	+	73	**330**	
134		12204	12566	+	93	**362**	
135		12567	12896	+	73	**329**	
136		12898	13258	+	95	**360**	
137		14236	14325	+	88	**89**	
138		14526	14638	+	88	**112**	
139		14640	14998	+	94	**358**	
140	DS497963.1	7128	7476	+	92	**348**	-
141		7477	7612	+	71	**135**	

c) GG694051–GG695897

Rb.	**GG**	**Početna pozicija**	**Krajnja pozicija**	orijentacija	**Identitet (%)**	**Duljina sekvence**	Broj Tcast>355
1	GG694069.1	409	34	-	72	375	8
2		785	410	-	80	375	
3		1273	915	-	92	358	
4		1650	1275	-	73	375	
5		2025	1651	-	80	374	
6		2402	2027	-	73	375	
7		2778	2403	-	79	375	
8	GG694086.1	293	1	-	93	292	2
9		625	295	-	73	330	
10		978	626	-	92	352	
11		1338	982	-	91	356	
12		1715	1340	-	72	375	
13	GG694112.1	140	1	-	68	139	2
14		517	141	-	72	376	
15		897	518	-	94	379	
16		952	881	-	90	71	
17	GG694130.1	1	106	+	85	105	2
18		109	484	+	73	375	
19		486	846	+	88	360	
20	GG694457.1	217	1	-	89	216	4
21		594	221	-	72	373	
22		1205	1566	+	88	361	
23		1567	1942	+	72	375	
24		1945	2307	+	89	362	
25		2308	2475	+	75	167	
26	GG695075.1	341	442	+	85	101	-
27		444	637	+	93	193	
28	GG694064.1	30	405	+	71	375	1
29		407	679	+	91	272	
30	GG694072.1	375	14	-	92	361	2
31		1056	786	-	93	270	
32		751	377	-	71	374	
33		1170	1058	-	88	112	
34	GG694074.1	172	5	-	75	167	2
35		548	173	-	72	375	
36		908	549	-	91	359	
37	GG694139.1	314	1	-	89	313	1
38		691	316	-	71	375	
39		888	692	-	90	196	
40	GG694156.1	422	67	-	89	355	2
41		799	424	-	70	375	
42	GG694158.1	273	2	-	92	271	1
43		650	275	-	70	375	
44		945	651	-	91	294	
45	GG694258.1	1	97	+	87	96	1
46		98	428	+	73	330	
47		430	789	+	89	359	
48		1246	1143	-	91	103	
49		2015	1729	-	89	286	
50	GG694275.1	1	107	+	87	106	4
51		321	154	-	75	167	
52		652	322	-	83	330	
53		1030	656	-	72	374	
54		1362	1031	-	82	331	
55		1724	1363	-	93	361	
56		2084	1726	-	92	358	
57		2420	2086	-	72	334	
58		2752	2421	-	82	331	
59		3130	2756	-	71	374	
60		3275	3131	-	88	144	
61	GG694431.1	22	389	+	72	367	2
62		391	711	+	82	320	
63		776	905	+	84	129	
64		907	1005	+	94	98	
65		1022	1148	+	85	126	
66		1150	1509	+	91	359	
67		1510	1676	+	74	166	
68	GG694460.1	57	165	+	85	108	2
69		166	521	+	73	355	
70		525	886	+	90	361	
71		887	1254	+	70	367	
72	GG694477.1	416	53	-	74	363	3

73		793	418	-	72	375	
74		1155	794	-	93	361	
75	GG694489.1	343	1	-	92	342	2
76		469	346	-	84	123	
77		910	552	-	94	358	
78		1242	912	-	74	330	
79		1602	1243	-	90	359	
80	GG694816.1	1	168	+	89	167	1
81		169	337	+	75	168	
82		464	711	+	74	247	
83		713	1071	+	92	358	
83		1073	1423	+	93	350	
84	GG694903.1	43	413	+	71	370	5
85		415	744	+	81	329	
86		746	1106	+	94	360	
87		1107	1437	+	73	330	
88		1438	1808	+	75	370	
89		1810	2139	+	80	329	
90		2141	2502	+	94	361	
91		2503	2829	+	72	326	
92		2831	3192	+	92	361	
93		3193	3339	+	72	146	
94	GG695009.1	424	49	-	80	375	1
95		892	554	-	92	338	
96	GG695026.1	1	236	+	89	235	3
97		237	402	+	74	165	
98		409	771	+	88	362	
99		772	1143	+	72	371	
100		1145	1475	+	82	330	
101		1476	1641	+	73	165	
102		1648	2009	+	88	361	
103		2010	2173	+	74	163	
104	GG695046.1	109	1	-	93	108	2
105		470	111	-	90	359	
106		847	472	-	70	375	
107		1190	848	-	84	342	
108	GG695146.1	345	1	-	93	344	2
109		714	347	-	71	367	
110		992	715	-	94	277	
111		1171	993	-	96	178	
112		1547	1173	-	70	374	
113		1672	1548	-	83	124	
114	GG695117.1	1299	940	-	90	359	3
115		1674	1301	-	82	373	
116		2051	1676	-	70	375	
117		2380	2052	-	87	328	
118	GG695498.1	345	1	-	89	344	
119		677	359	-	73	318	
120		1009	678	-	73	331	
121		1340	1010	-	74	330	
122		1689	1341	-	73	348	
123	GG695747.1	4	375	+	71	371	3
124		376	706	+	73	330	
125		708	1066	+	88	358	
126		1068	1429	+	88	361	
127	GG695796.1	1	85	+	82	84	1
128		91	450	+	90	359	
129		451	677	+	79	226	
130		919	804	-	72	115	
131		1054	921	-	77	133	
132		1560	1339	-	79	221	
133		1914	1561	-	87	353	
134		2037	1916	-	82	121	
135	GG694363.1	62	264	+	80	202	2
136		451	270	-	77	181	
137		827	452	-	72	375	
138		1187	828	-	92	359	
139		1315	1189	-	85	126	
140	GG694388.1	1	193	+	92	192	6
141		194	569	+	71	375	
142		571	930	+	93	359	
143		931	1306	+	70	375	
144		1308	1667	+	92	359	
145		1669	2027	+	92	358	
146		2030	2405	+	70	375	
147	GG694430.1	430	69	-	92	361	5
148		792	432	-	92	360	
149		1166	794	-	71	372	
150		1528	1167	-	91	361	
151		1890	1530	-	91	360	
152	GG694432.1	319	1	-	92	318	6

153		693	321	-	71	372	
154		1069	694	-	76	375	
155		1444	1071	-	71	373	
156		1802	1445	-	91	357	
157		2176	1804	-	72	372	
158		2533	2177	-	91	356	
159		2650	2535	-	87	115	
160	GG694453.1	17	309	+	73	292	7
161		311	671	+	92	360	
162		672	1047	+	72	375	
163		1049	1409	+	91	360	
164		1410	1785	+	72	375	
165		1787	2145	+	91	358	
166		2146	2521	+	71	375	
167		2523	2883	+	91	360	
168	GG694526.1	1	302	+	90	301	-
169		303	581	+	70	278	
170	GG694559.1	350	2	-	92	348	3
171		682	352	-	73	330	
172		1058	683	-	72	375	
173		1418	1059	-	89	359	
174		1750	1420	-	72	330	
175		2109	1751	-	89	358	
176		2442	2112	-	73	330	
177	GG694569.1	15	374	+	93	359	3
178		375	746	+	72	371	
179		748	1097	+	88	349	
180		1098	1473	+	70	375	
181		1475	1594	+	93	119	
182	GG694617.1	223	71	-	77	152	2
183		334	705	+	70	371	
183		707	1066	+	90	359	
184		1067	1273	+	76	206	
185	GG94835.1	260	1	-	80	259	4
186		629	262	-	72	367	
187		1006	630	-	71	376	
188		1365	1007	-	91	358	
189		1496	1367	-	84	129	
190		1882	1561	-	80	321	
191		2251	1884	-	71	367	
192		2379	2252	-	85	127	
193	GG694900.1	1	335	+	93	334	2
194		336	662	+	73	326	
195		664	993	+	80	329	
196		995	1356	+	94	361	
197		1357	1683	+	72	326	
198		1685	2045	+	93	360	
199	GG694920.1	889	546	-	90	343	5
200		1266	891	-	72	375	
201		1629	1267	-	83	362	
202		1989	1630	-	90	359	
203		2350	1991	-	91	359	
204		2727	2354	-	72	373	
205		2820	2728	-	85	92	
206	GG695045.1	1	250	+	94	249	-
207		292	576	+	73	284	
208		793	928	+	77	135	
209	GG695178.1	135	1	-	86	134	-
210		560	393	-	73	167	
211		895	561	-	92	334	
212	GG695206.1	228	558	+	85	330	-
213		603	883	+	71	280	
214		885	1231	+	91	346	
215	GG695229.1	30	156	+	84	126	2
216		158	524	+	80	366	
217		526	883	+	91	357	
218	GG695393.1	109	221	+	87	112	4
219		223	552	+	80	329	
220		554	915	+	92	361	
221		916	1283	+	71	367	
222		1285	1614	+	81	329	
223		1616	1977	+	92	361	
224		1978	2345	+	71	367	
225		2347	2663	+	90	316	
226	GG695542.1	336	5	-	92	331	1
227		464	338	-	85	126	
228		956	614	-	92	342	
229		1288	958	-	73	330	
230		1650	1289	-	93	361	
231		1982	1652	-	72	330	
232		2176	1983	-	93	193	

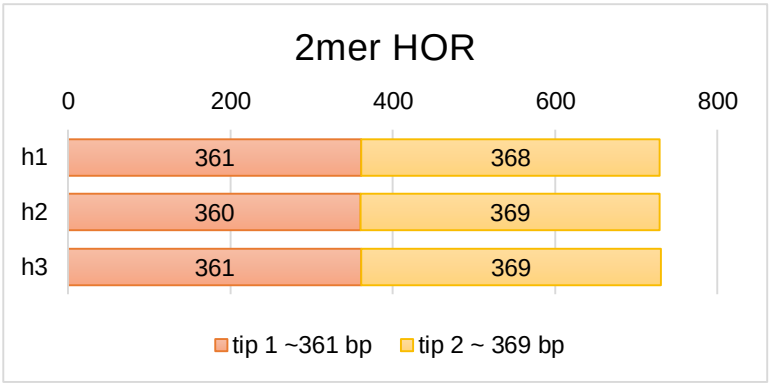
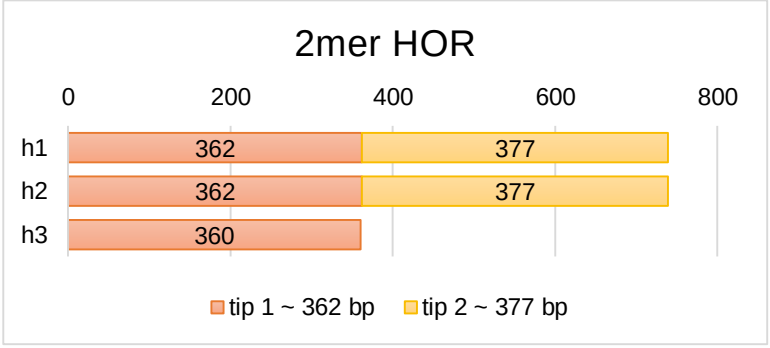
233	GG695616.1	172	299	+	84	127	6
234		300	630	+	73	330	
235		632	993	+	91	361	
236		994	1367	+	72	373	
237		1368	1698	+	74	330	
238		1700	2059	+	89	359	
239		2586	2398	-	93	188	
240		2918	2588	-	74	330	
241		3292	2919	-	71	373	
242		3654	3293	-	92	361	
243		3986	3656	-	73	330	
244		4360	3987	-	71	373	
245		4565	4361	-	90	204	
246	GG695732.1	464	90	-	74	374	5
247		841	465	-	73	376	
248		1218	852	-	72	366	
249		1578	1219	-	92	359	
250		1954	1580	-	81	374	
251		2082	1956	-	86	126	
252	GG695826.1	1	360	+	91	359	30
253		361	729	+	73	368	
254		731	1093	+	91	362	
255		1094	1461	+	71	367	
256		1463	1790	+	80	327	
257		1792	2153	+	93	361	
258		2154	2521	+	71	367	
259		2523	2883	+	91	360	
260		2884	3251	+	73	367	

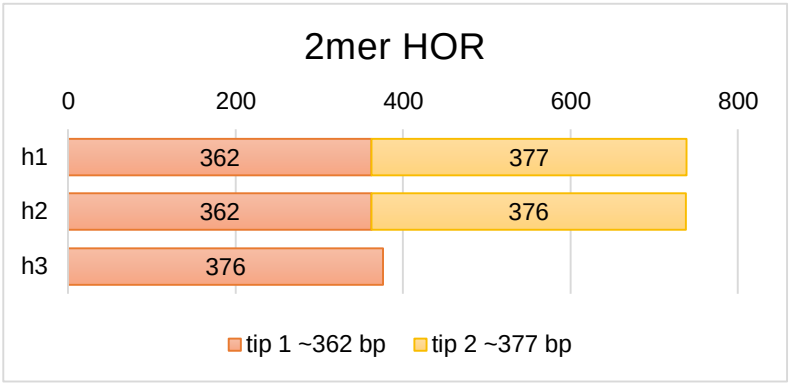
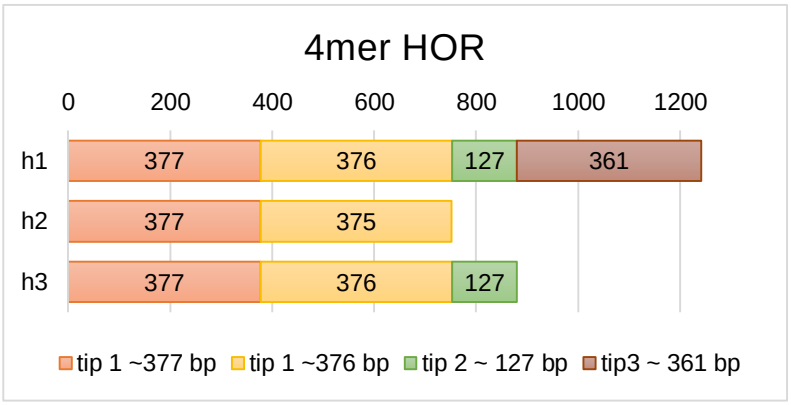
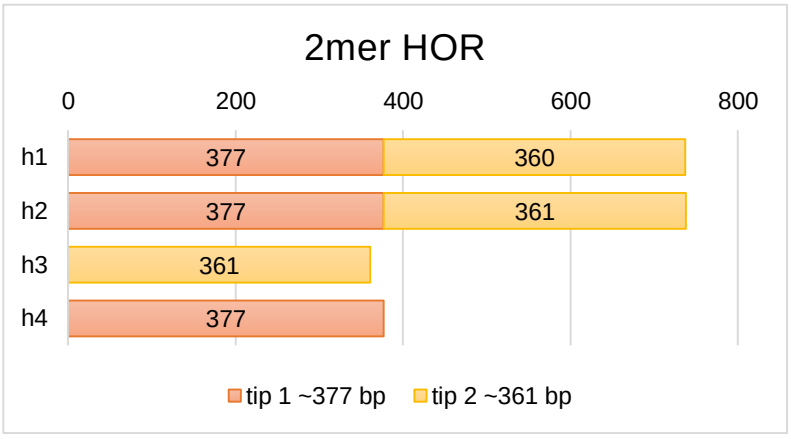
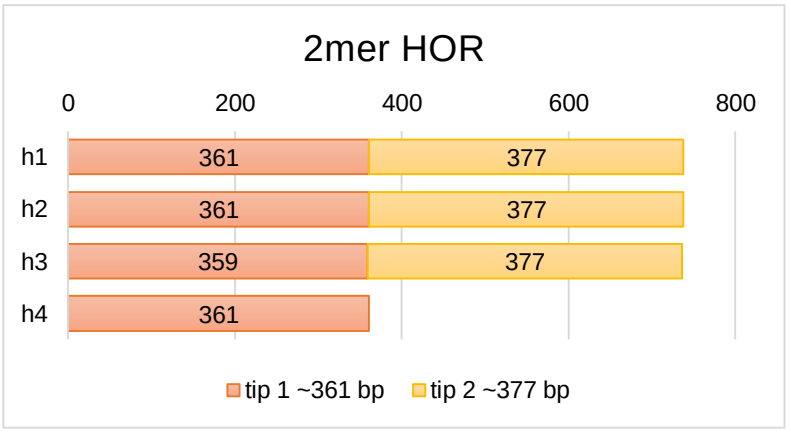
261		3253	3611	+	91	358	
262		3614	3977	+	73	363	
263		3983	4312	+	80	329	
264		4315	4676	+	92	361	
265		4677	5044	+	72	367	
266		5046	5407	+	91	361	
267		5408	5775	+	73	367	
268		5777	6106	+	81	329	
269		6108	6469	+	92	361	
270		6470	6837	+	71	367	
271		6839	7199	+	91	360	
272		7200	7567	+	73	367	
273		7569	7898	+	80	329	
274		7900	8261	+	93	361	
275		8262	8629	+	71	367	
276		8631	8991	+	91	360	
277		8992	9359	+	72	367	
278		9361	9687	+	80	326	
279		9689	10050	+	92	361	
280		10051	10418	+	71	367	
281		10420	10780	+	90	360	
282		10781	11148	+	73	367	
283		11150	11479	+	80	329	
284		11481	11843	+	92	362	
285		11844	12211	+	71	367	
286		12213	12573	+	91	360	
287		12574	12941	+	73	367	

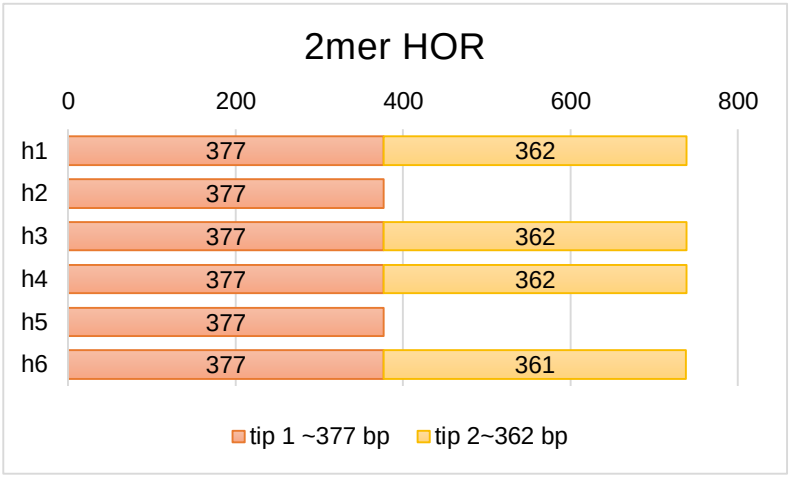
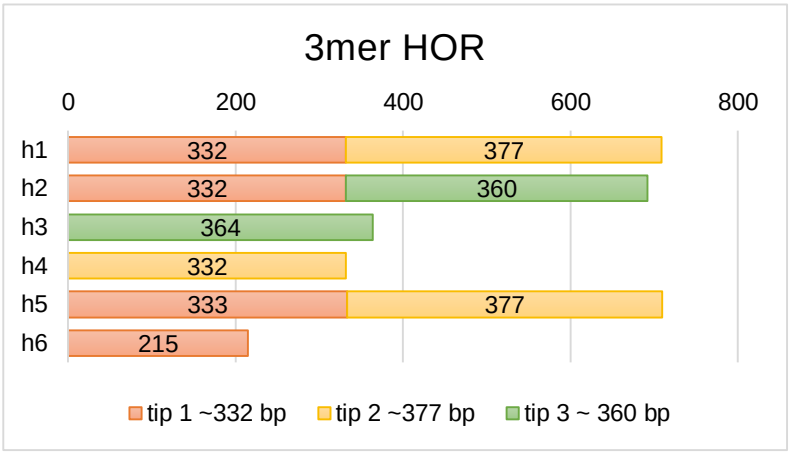
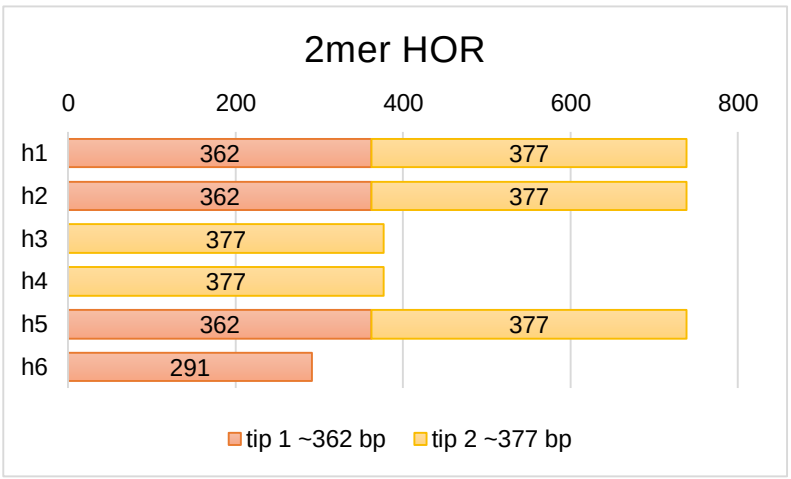
4.1.3 Ostale HOR strukture zasnovane na TCAST satelitima

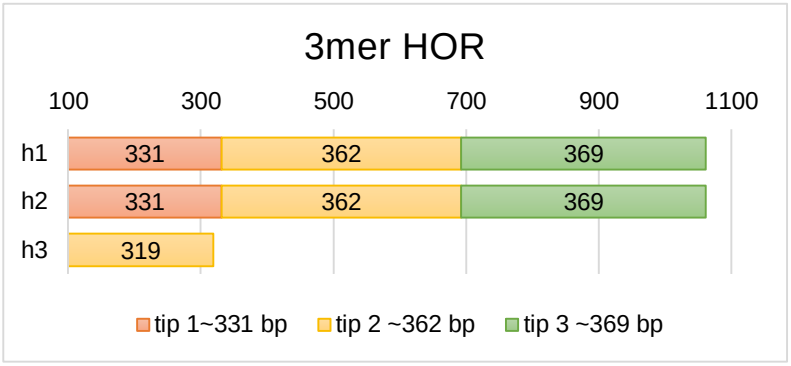
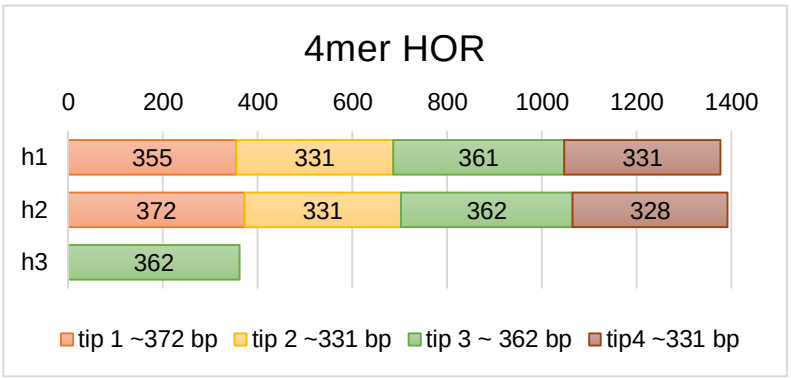
Temeljeno na rezultatima BLAST-a, provjerili smo ostale komponente za koje smo dobili pogotke, ali i za one pogotke koje smo dobili s GRM metodom (tablica 17).

Tablica 17. Prikaz identificiranih HOR struktura temeljenih na TCAST satelitima.

DS497845.1	Prosječna divergencija između monomera: 15,74±10,63%. Monomeri tipa 1 i 2 su obrnuti komplement od DS360/5873 sekvence (91% i 73% identiteta). Prosječna divergencija između ~ 730 bp HOR kopija: 3,20±0,29%.	
DS497867.1	Prosječna divergencija između monomera : 16,40±10,97%. Monomeri tipa 1 i 2 su obrnuti komplement od DS360/5873 sekvence (89% i 73% identiteta). Prosječna divergencija između ~ 739 bp HOR kopija: 2,97±0,00%.	

DS497699.1	Prosječna divergencija između monomera : 19,20±9,04%. Monomeri tipa 1 i 2 su obrnuti komplement od DS360/5873 sekvence (89% i 71% identiteta). Prosječna divergencija između ~ 739 bp HOR kopija:9,19±0,00%.	
GG694069.1	Prosječna divergencija između monomera : 33,93±30,53%. Monomeri tipa 1, 2 i 4 su obrnuti komplement od DS360/5873 sekvence (70%, 82% i 92 % identiteta). Prosječna divergencija između ~1241 bp HOR kopija 0,91±0,00%.	
GG694388.1	Prosječna divergencija između monomera: 17,56±10,46. Monomeri tipa 1, 2 sa DS360/5873 sekvencom imaju ~ 90% identiteta. Prosječna divergencija između ~738 bp HOR kopija 4,88±0,00%.	
GG694453.1	Prosječna divergencija između monomera: 15,14±11,53%. Monomeri tipa 1, 2 sa DS360/5873 sekvencom imaju 92% i 70% identiteta. Prosječna divergencija između ~738 bp HOR kopija 1,90±0,23%.	

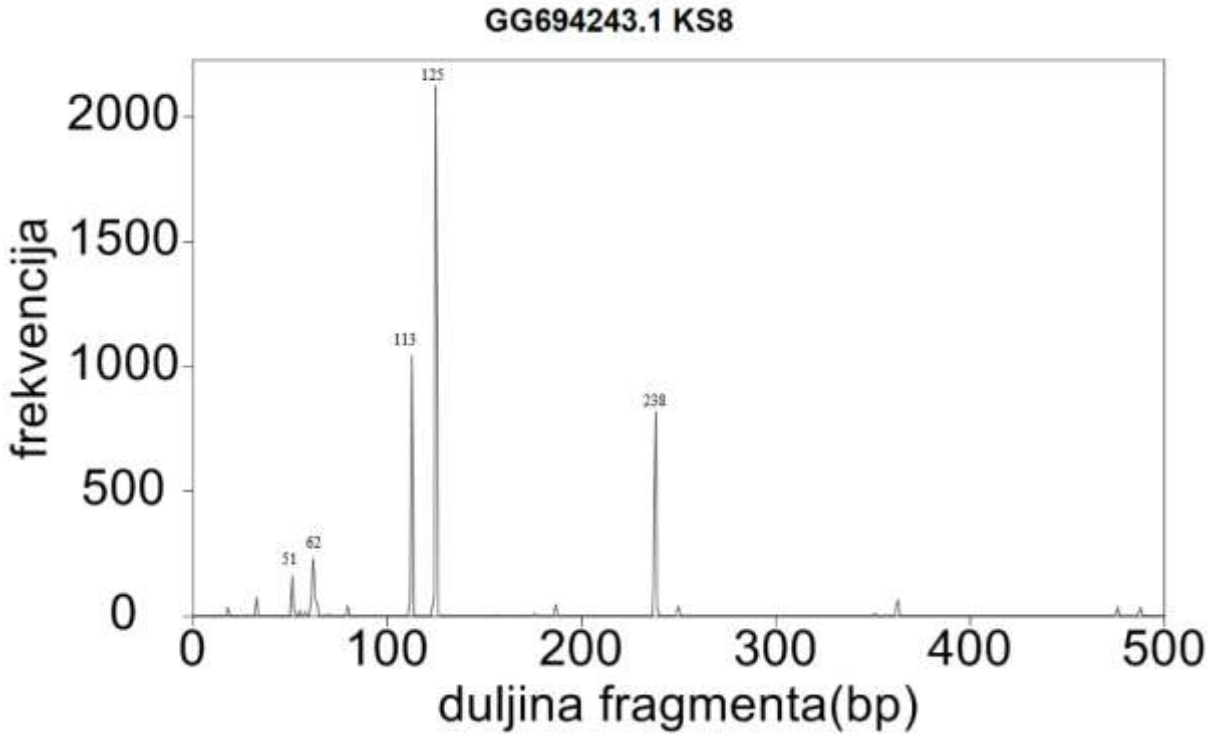
GG695145.1	Prosječna divergencija između monomera: 15,41±10,06%. Monomeri tipa 1, 2 sa DS360/5873 sekvencom imaju 92% i 72% identiteta.	
	Prosječna divergencija između ~739 bp HOR kopija 3,27±0,82%.	
GG694275.1	Prosječna divergencija između monomera : 27,39±16,91%. Monomeri tipa 1, 2 i 3 sa DS360/5873 sekvencom imaju 72%, 82% i 93% identiteta i ujedno su obrnuti komplement.	
	Prosječna divergencija između ~1069 bp HOR kopija je 11,36±8,86% (uzete su h1, h2 i h3 za račun divergencije).	
GG694360.1	Prosječna divergencija između monomera: 18,57±12,10%. Monomeri tipa 1, 2 sa DS360/5873 sekvencom imaju 90%, 73% identiteta i ujedno su obrnuti komplement.	
	Prosječna divergencija između ~739 bp HOR kopija je 2,39±0,51%.	

GG695393.1	Prosječna divergencija između monomera: 21,03±9,31%. Monomeri tipa 1, 2 i 3 sa DS360/5873 sekvencom imaju 80%, 92 i 71% identiteta.	
	Prosječna divergencija između ~1062 bp HOR kopija je 2,45±0,00%.	
GG694903.1	Prosječna divergencija između monomera: 19,08±8,49%. Monomeri tipa 1, 2, 3 i 4 sa DS360/5873 sekvencom imaju 75%, 80%, 94% i 69% identiteta.	
	Prosječna divergencija između ~1396 bp HOR kopija je 4,79±0,00%.	

Periodičnosti višega reda, kao što se vidi i iz tablica dobivenih s BLAST-om, se ne pojavljuju u kromosomima, već se nalaze u ostalim komponentama genoma Tcas 3.0. Ovo se podudara s prethodnim istraživanjima, ali također daje naslutiti da se područje heterokromatina sastoji od velikog broja repeticija što čini taj dio genoma i tehnički težim za sastavljanje i određivanje pozicije na kromosomima. Postojanje HOR-ova u tom području, sastavljenih od TCAST satelita za koje se pretpostavlja da imaju ulogu u organizaciji kromosoma, te njihovo najčešće pojavljivanje kao dimer mogu uputiti na postojanje uloge kakve imaju i alfa sateliti kod čovjeka.

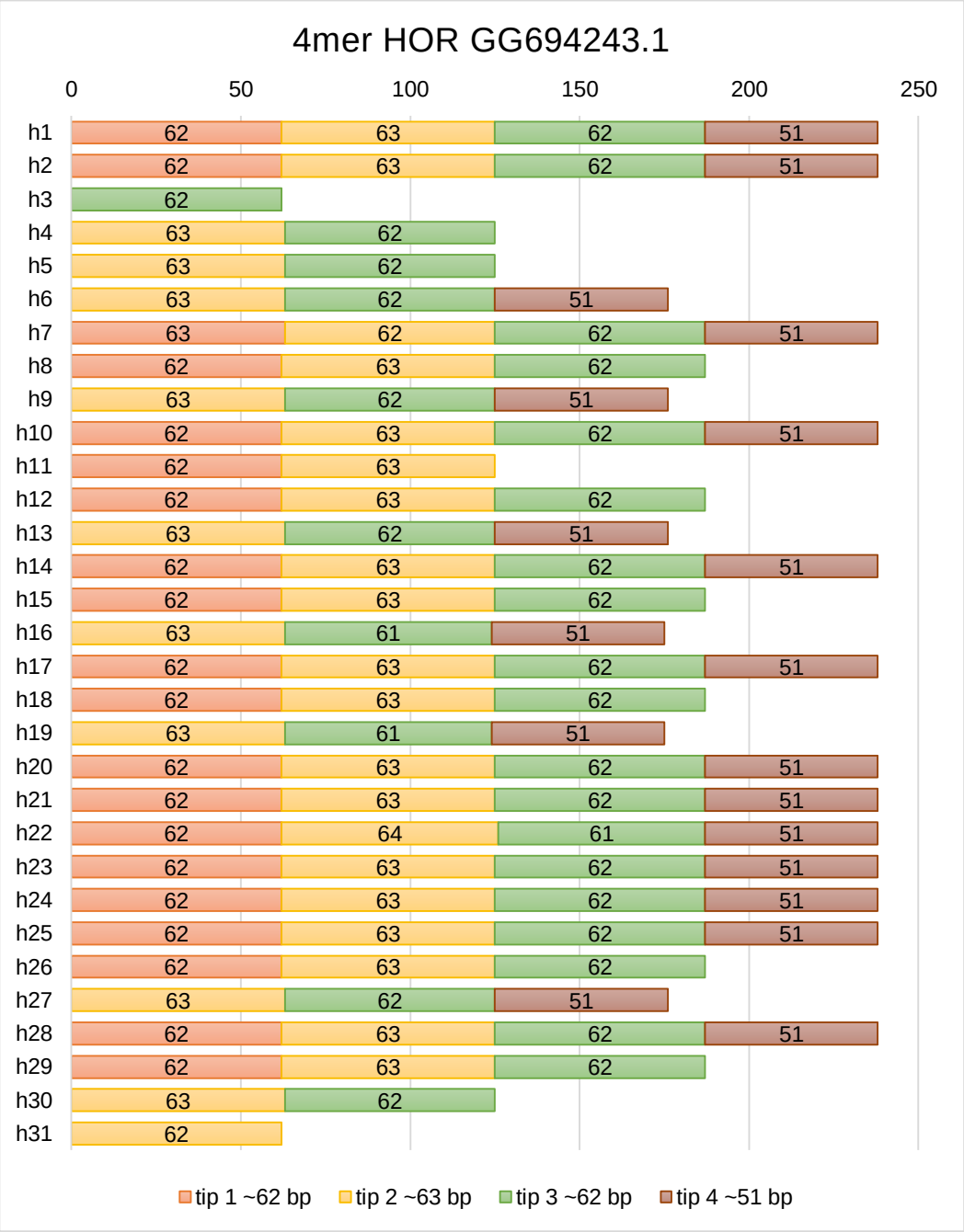
4.2 Pik 51 bp

4.2.1 GG694243.1



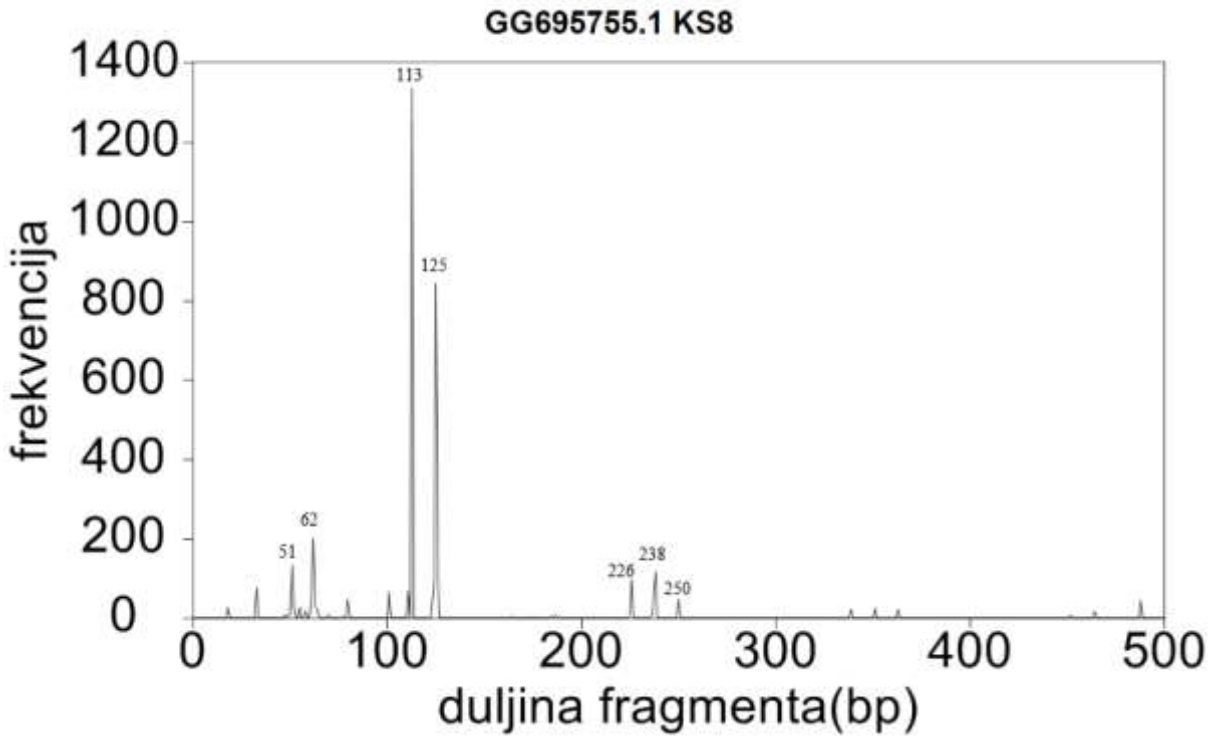
Slika 33. Prikaz GRM dijagrama za GG694243.1

Na GRM dijagramu prikazani su pikovi za GG694243.1. Pomoću dobivenog dominantnog K-stringa TAGTGCGA za pik 51 bp, dobili smo četiri tipa monomera duljina 62 bp, 63 bp, 62 bp i 51 bp. Divergencije između ovih monomera su 21,46±15,29% te smo iz „heat“ mape dobili HOR strukturu ~238 bp 4mer HOR-a kao što je prikazano na slici 34. Prosječna divergencija HOR kopija je 1,44±1,09%. Iz slike 34 vidi se da se mogu objasniti pikovi 125 bp, kao zbroj monomera 62 + 63, 113 bp kao 62+51 i 176 bp kao 62+63+51 kao rezultat principa rada GRM metode.



Slika 34. Struktura 4mer HOR u GG694243.1. Ova struktura ima 31 kopiju, od kojih 13 ima sve tipove monomera. Zbroj nepotpunih kopija daje ostale vrhove na GRM dijagramu na slici 33.

4.2.2 GG695755.1



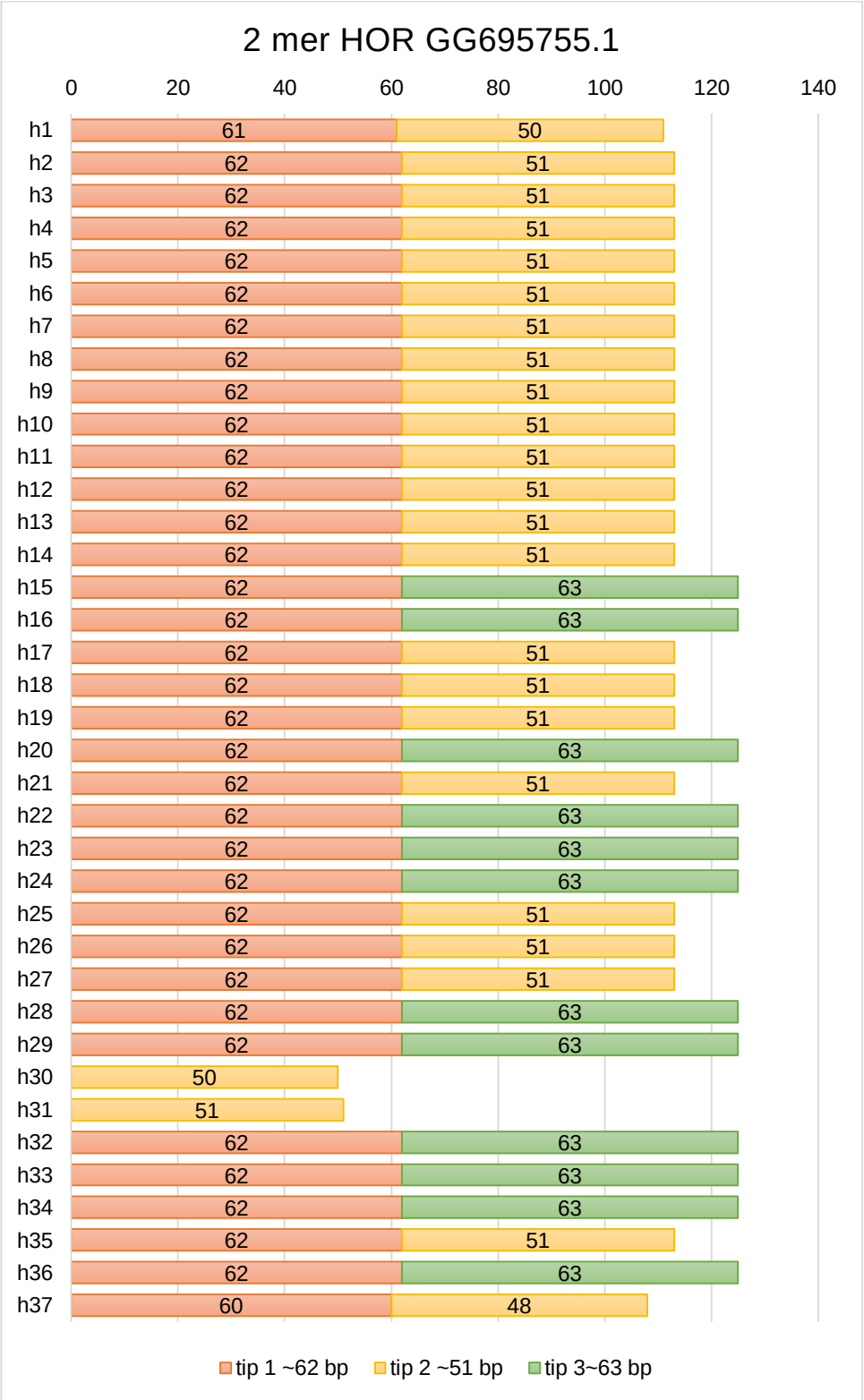
Slika 35. Prikaz GRM dijagrama za GG695755.1.

Sa K-stringom CGCACTAA dobili smo tri tipa monomera duljina ~62 bp, ~51 bp i ~63 bp s međusobnom divergencijom od 22,76±15,67%. Preko „heat" mape dobili smo 2mer HOR strukturu konsenzusne duljine ~ 113 bp, s time da se tip 2 od ~51 bp izmjenjuje s trećim tipom monomera duljine 63 bp što objašnjava peak 125 bp kao što je prikazano na slici 36.

Poravnanjem konsenzusnih sekvenci monomera tipa 2 i 3, dobili smo s BLAST-om identitet 92% kao što je prikazano u tablici 18. Divergencija monomera tipa 1 s tipovima 2 i 3 je redom 33,33% i 26,98% i kod poravnanja monomera tipa 1 s tipom 3 BLAST nije davao za standardne postavke poravnanje.

Tablica 18. Usporedba konsenzusne sekvence monomera tipa 2 i tipa 3 u GG695755.1 pomoću BLAST-a.

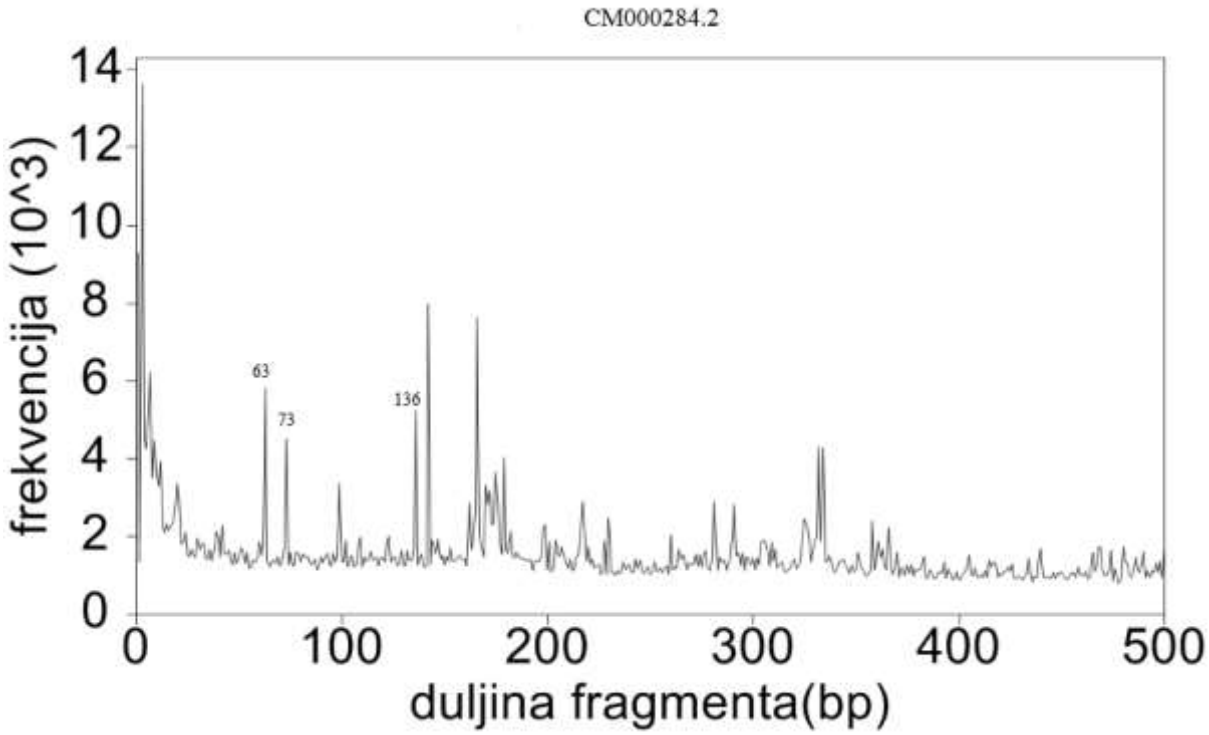
Tip monomera	Početna pozicija	Poravnanje	Konačna pozicija
Tip 2	4	ACTAAACCTAAAAATGCCACGTGTTTCTCGTACTGCCAACCTAAAAAT	51
		|||||| | |||||||||||||||||| ||||||||||||||	
Tip 3	16	ACTAAATTTTAAAATGCCACGTGTTTCTCGTAATGCCAACCTAAAAAT	63



Slika 36. Prikaz strukture 2mer HOR-a u GG695755.1. Divergencija između kopija HOR-a je 7,24±5,00.

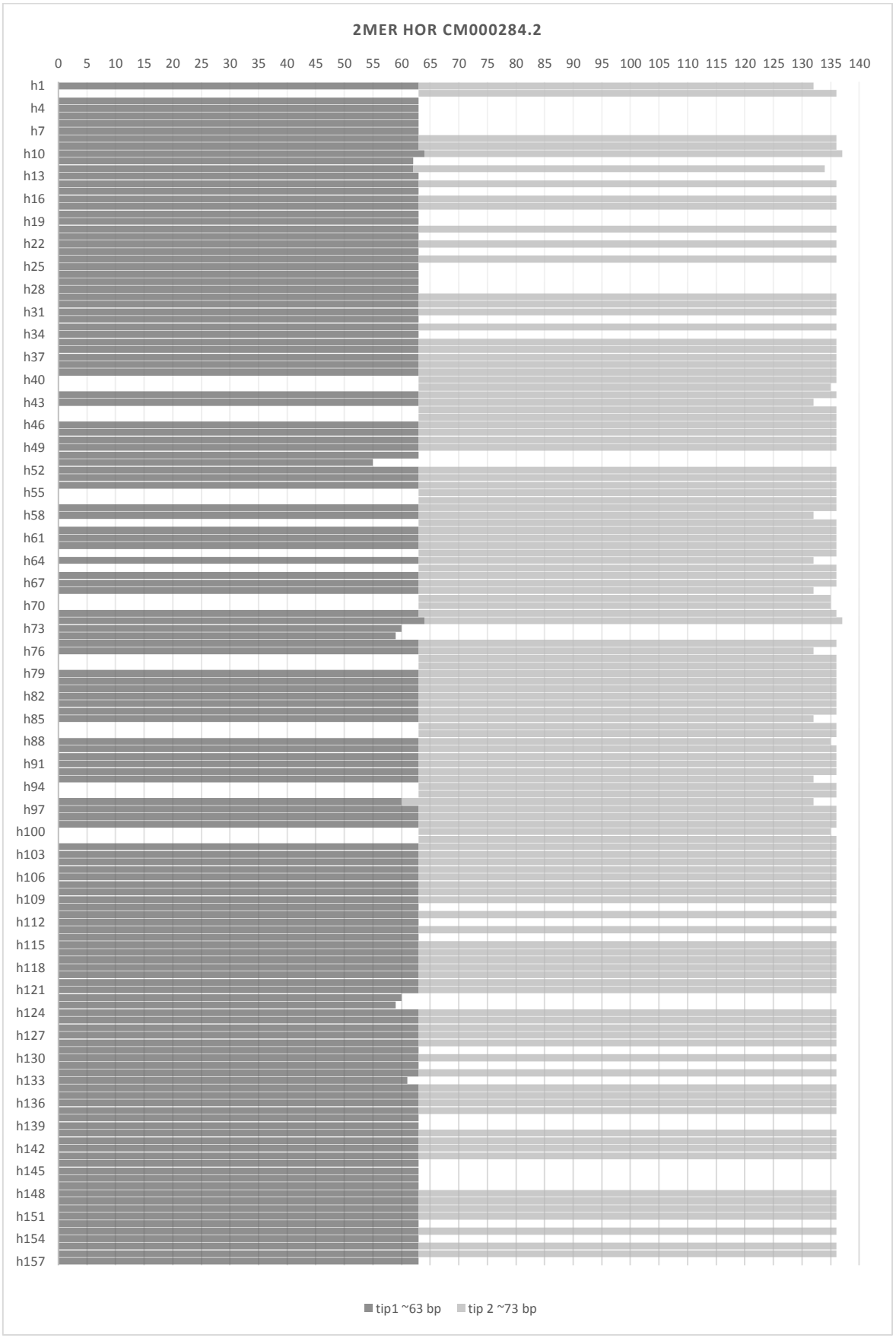
4.3. Pik 63

4.3.1 CM000284.2



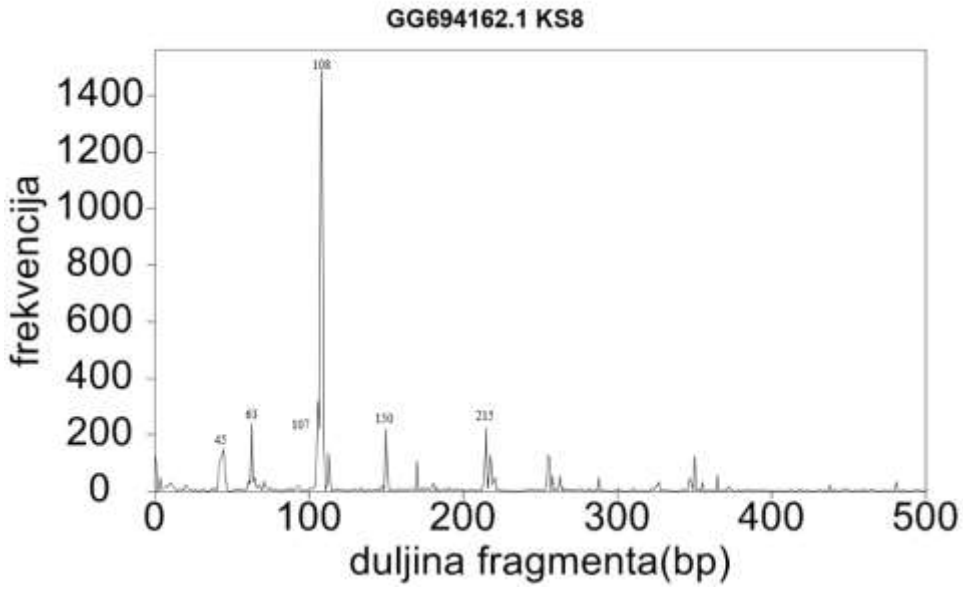
Slika 37. GRM dijagram komponente CM000284.2.

Pomoću K-stringa ATTAAATT za pik 63 bp u kromosomu CM000284.2 (LG9), dobili smo 255 monomera duljina 63 bp i 73 bp od pozicije 20166028 do pozicije 20183143, koje daju 157 kopija 2mer HOR-a duljine ~136 bp (slika 37, 38). Pomoću Needleman-Wunsch algoritma dobili smo za divergenciju između kopija monomera 13,78±8,63%, a za HOR kopije 5,17±2,37%.



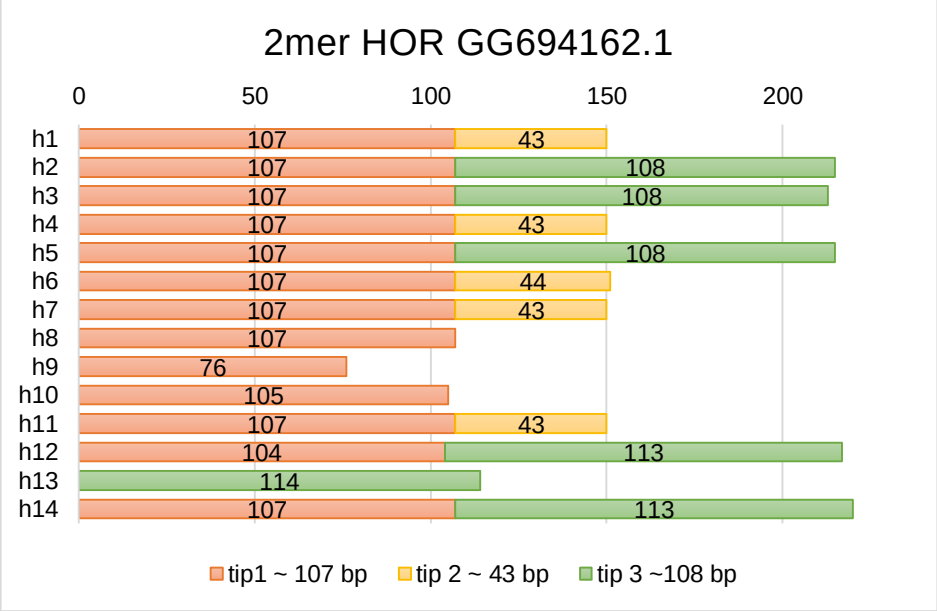
Slika 38. 2mer HOR struktura iz CM000284.2.

4.3.2 GG694162.1



Slika 39. GRM dijagram komponente GG694162.1.

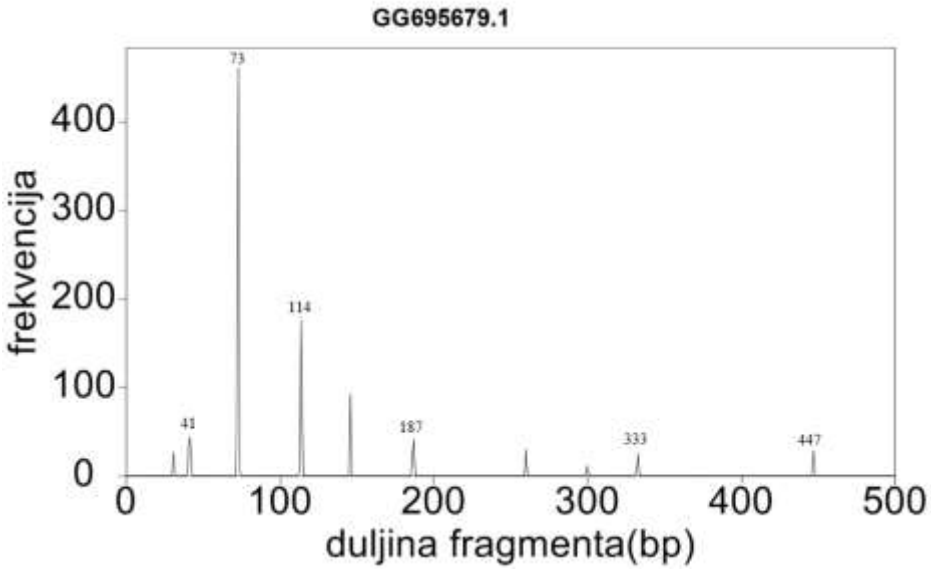
Sa dominantnim K-stringom ATTTATTT od pozicije 1174 do pozicije 3296 dobili smo tri tipa monomera duljina ~107 bp, ~43 bp i ~108 bp, koji tvore 2mer HOR. U ovom HOR-u izmjenjuju se monomeri tipa 2 i 3 što objašnjava pikove na GRM dijagramu (slika 39) duljina ~150 bp i ~215 bp. HOR struktura je prikazana na slici 40. Prosječna divergencija monomera je 33,99±29,09%.



Slika 40. Struktura 2 mer HOR-a. Prvih 38 baza u tipu monomera 2 ima 92% identiteta s tipom monomera 3. Divergencija HOR kopija ~150 bp je 2,39±1,26%, a HOR kopija ~215 bp je 4,44±3,52%.

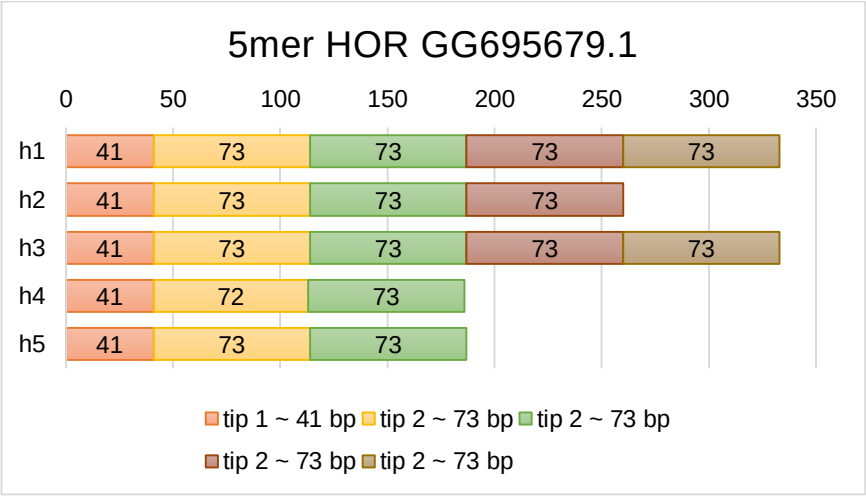
4.4 Pik 73

4.4.1 GG695679.1



Slika 41. GRM dijagram komponente GG695679.1

Od pozicije 5 do 1231 pomoću K-stringa TTTAAAAC dobili smo 20 monomera duljina ~ 73 bp i ~41 bp međusobne divergencije 25,18±24,13%. Pomoću „heat“ mape konstruirali smo periodičnost višega reda konsenzusne duljine ~333 bp, zasnovanu na jednom tipu monomera ~41 bp i 4 monomera tipa 2 duljine ~73 bp kao što je prikazano na slikama 41 i 42. Divergencija između monomera tipa 1 je 2,93±2,52%, a tipa 2 je 6,16±3,02%. Zanimljivo je poravnanje između monomera tipa 1 i 2, naime konsenzusne sekvence tipa 1 i tipa 2 pokazuje zapravo da je tip 1 isto što i tip 2 s obrisanim bazama kao što je prikazano u tablici 19.

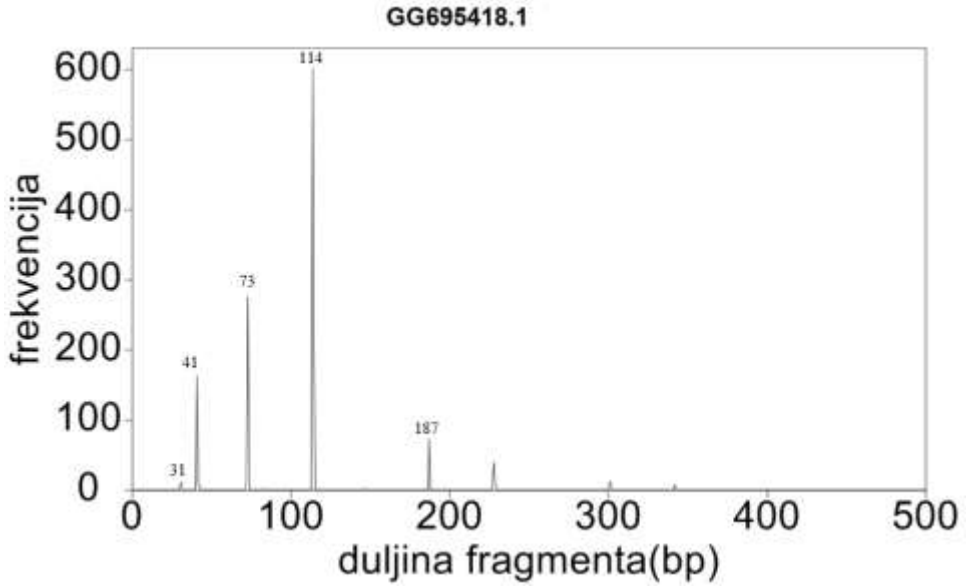


Slika 42. Prikaz 5mer HOR strukture u GG695679.1. Divergencija cijelih HOR kopija je 3,60±0,00%.

Tablica 19. Poravnanje konsenzusne sekvence tipa 1 i tipa 2 s BLAST-om. Identitet prvih 8 baza je 100%, a ostalih 96 % od kojih je u tipu 1 od 14-te pozicije do zadnje poravnato s 28 zadnjih u tipu monomera 2.

Tip 1 1 TTTAAAAC 8 |||||||| Tip 2 1 TTTAAAAC 8	Tip 1 14 TAAATTTAAAAAATCGCGCGGGCACGTG 41 |||||||||||||| |||||||||||||| Tip 2 46 TAAATTTAAAAAATGGCGCGGGCACGTG 73

4.4.2 GG695418.1

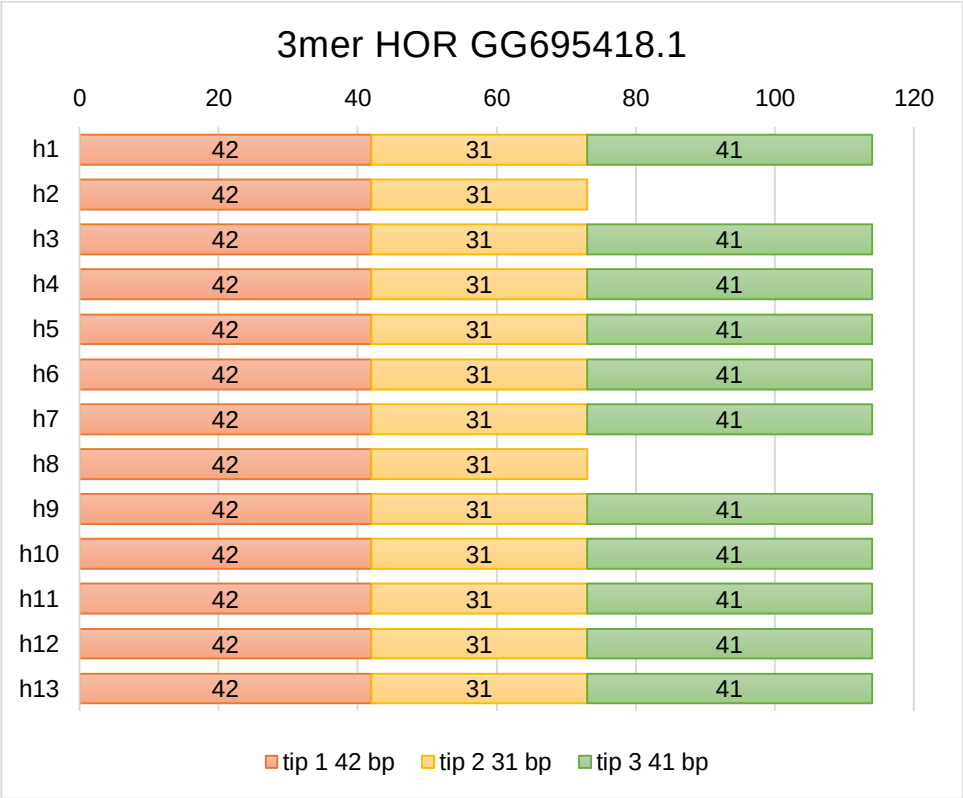


Slika 43. GRM dijagram za GG695418.1.

Prema grm dijagramu sa slike 43 i pomoću dominantnog K-stringa TAAAAAAT, dobili smo 37 monomera podijeljenih u tri tipa s duljinama redom 41 bp, 31 bp i 42 bp od pozicije 12 do pozicije 1371, s međusobnom divergencijom od 26,50±17,13%. Na temelju „heat“ mape dobili smo ~114 bp kao što je prikazano na slici 44. Kopije HOR-a imaju divergenciju od 3,48±3,98%. Usporedbe konsenzusnih sekvenci su prikazane u tablici 20.

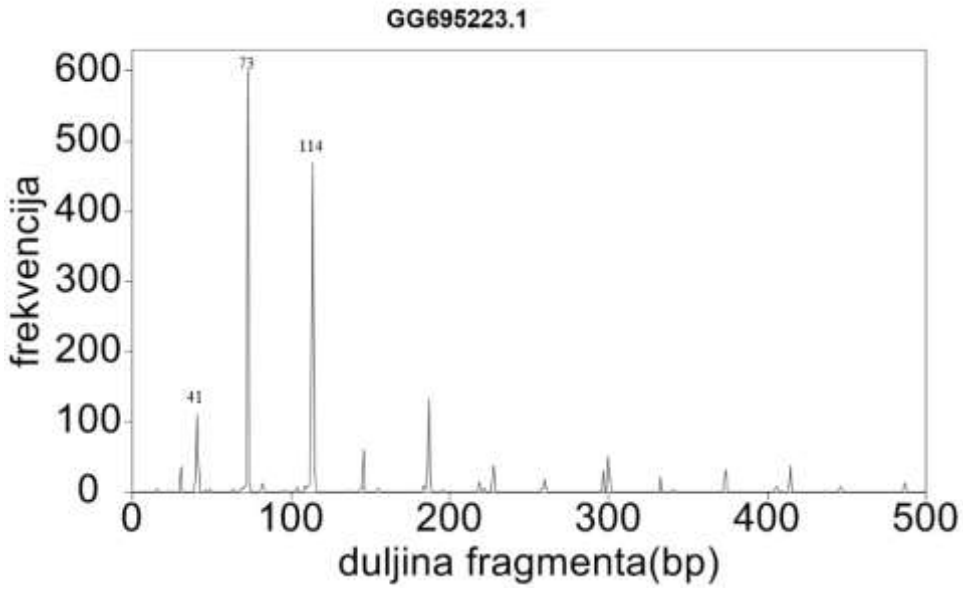
Tablica 20. Prikaz poravnanja konsenzusnih sekvenci tri tipa monomera s Needleman – Wunsch programom.

tip 1 vs. tip 2 – divergencija je 45,23%.	TAAAAAATGGCGCGGGCACGTGTTTAAAACACGTTGCCAACC ||||||| ||||| || | | | | || || TAAAAAATCGCGCGAAATCG_G__T__A_A__TT__AAATT
tip 1 vs. tip 3 – divergencija je 23,26%.	TAAAAAATGGCGCGGGCACGTGTTTAAAACACGTTGC_CAACC |||||||||||||| |||||||| ||| ||| | || TAAAAAATGGCGCGGACACGTGTTT_AAA_GCGTGACTAAATT
tip 2 vs. tip 3 – divergencija je 34,15%.	TAAAAAATCGCGCGAAATCG_G__T_AA___T___TAAATT ||||||| ||||| | || | || | |||||| TAAAAAATGGCGCGGACACGTGTTTAAAGCGTGACTAAATT



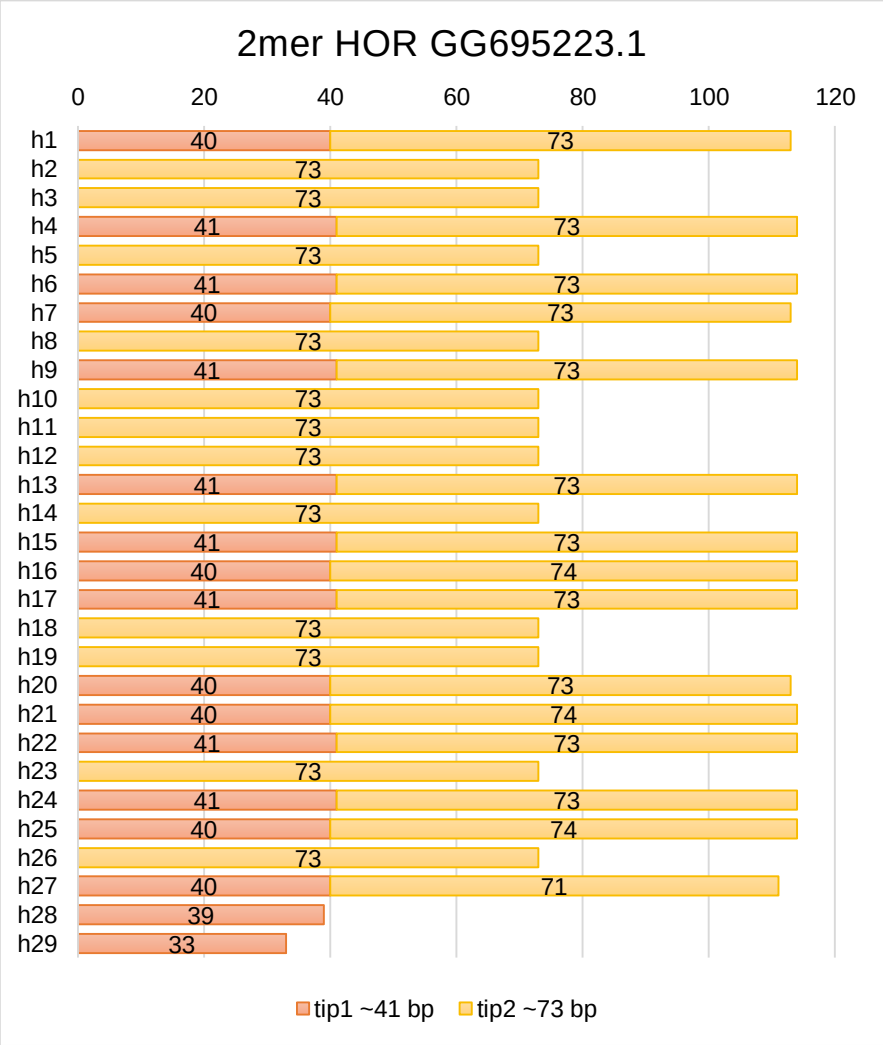
Slika 44. Prikaz 13 kopija 3mer HOR-a duljine ~114 bp. Monomer tipa 1 i 2 daju prvo pik 73 a on s trećim tipom monomera daje pik od 114 bp, što je ujedno i konsenzusna duljina HOR-a.

4.4.3 GG695223.1



Slika 45. GRM prikaz za GG695223.1.

Od pozicije 119 do pozicije 2588 nalazi se 42 monomera duljine 73 i 41 baze dobiveni s K-stringom TTTTTAAA (prema GRM dijagramu na slici 45). Poravnanjem monomera tipa 1 i tipa 2 pomoću BLAST-a vidi se da je u tipu 1 obrisano oko 40 baza (tablica 21). Divergencija monomera je 33,34±25,64%. Pomoću „heat" mape dobili smo strukturu periodičnosti višega reda duljine ~114 bp, prikazanu na slici 46.

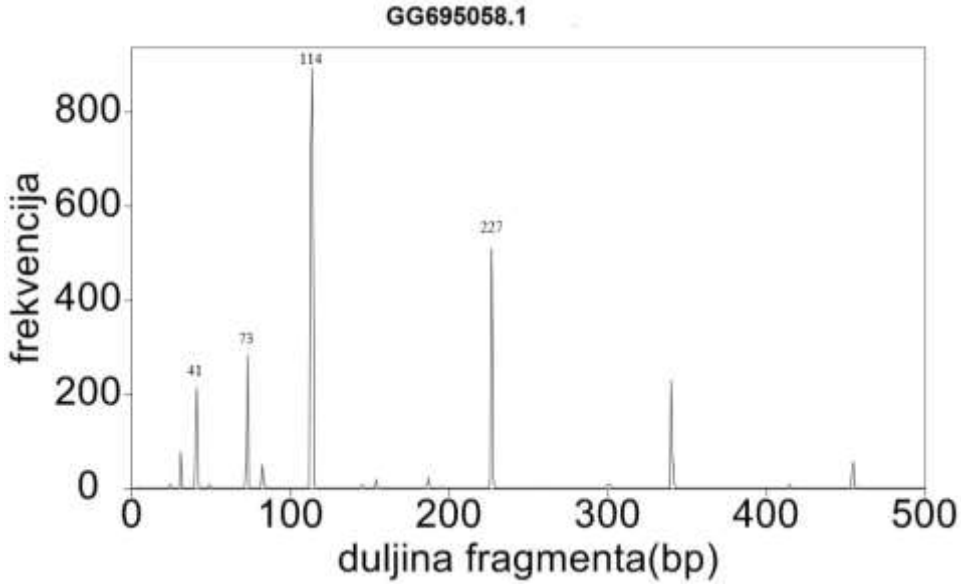


Slika 46. Prikaz 29 kopija 2mer HOR-a u GG695223.1. Divergencija među cijelim HOR kopijama je 8,23±3,20%.

Tablica 21. Poravnanje monomera tipa 1 i tipa 2 s BLAST-om, daje za prvih 12 baza identitet od 100%, a za zadnjih 18 baza od 95%.

Tip 1 1 TTTTTAAATTTA 12 |||||||||||| Tip 2 1 TTTTTAAATTTA 12	Tip 1 20 TTTAAACACGTGTCCGCGC 38 ||||||||||| |||||| Tip 2 52 TTTAAACACGTGCCCGCGC 70

4.4.4 GG695058.1

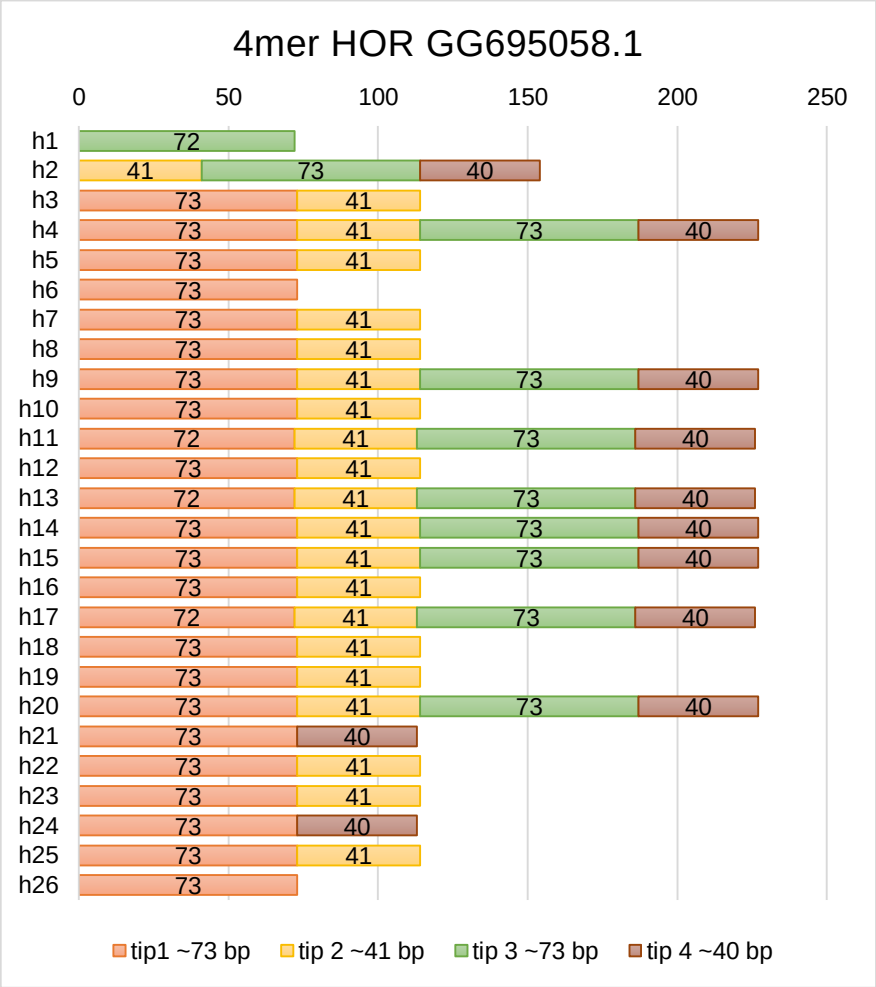


Slika 47. GRM prikaz GG695058.1

Od pozicije 22 do pozicije 3728 pomoću K-stringa TTTTTAAA dobili smo 66 monomera duljina 41 bp i 73 bp, međusobne divergencije 32,49±26,08%. Pomoću „heat“ mape vidjeli smo da ove monomere možemo pripisati različitim tipovima monomera i to tipu 1 ~ 73 bp, tipu 2 ~ 41 bp , tipu 3 ~73 bp i tipu 4 ~ 40 bp što objašnjava pikove na GRM dijagramu (slika 47), 73+41=114 i 73+41+73+40= 227, što je dodatno prikazano na slici 48. Divergencija između kopija tipa 1 je 2,95±2,02%, tipa 2 je 2,16±1,90%, tipa 3 je 3,29±6,06% i tipa 4 je 0,91±1,95%. U tablici 22 su prikazana poravnanja konsenzusnih sekvenci 4 tipa monomera.

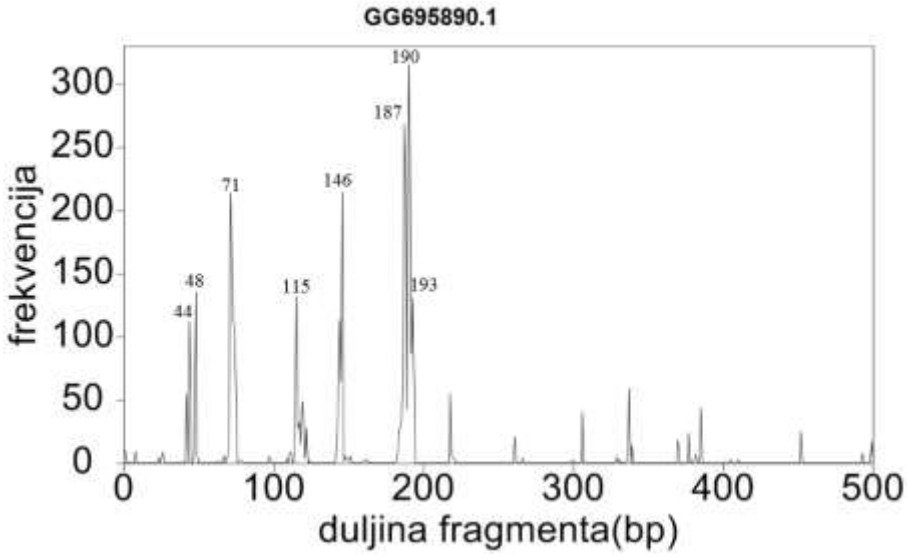
Tablica 22. Poravnanje konsenzusnih sekvenci četiri tipa monomera s BLAST-om.

Tip 1 1 TTTTTAAATTTA 12 |||||||||||| Tip 2 1 TTTTTAAATTTA 12	Tip 1 52 TTTAAACACGTGCCCGCGC 70 |||||||||||| |||||| Tip 2 20 TTTAAACACGTGTCCGCGC 38	Identitet prvih 12 baza je 100% a zadnjih 18 je 95%.
Tip 1 1 TTTTTAAATTTAATTACCGATTTCATGCGATTTTTTAGGTTGGCAACGTGTTTTAAACAC 60 |||||||||||||||||||||||| |||||||||||| |||||||||||||||||| Tip 3 1 TTTTTAAATTTAATTACCGATTTCGTGCGATTTTTTAAGTTGGCAACGTGTTTTAAATGT 60 Tip 1 61 GTGCCCGCGCGAG 73 ||| |||||||| Tip 3 61 GTGTTCGCGCGAG 73		Identitet je 90%.
Tip 1 50 GTTTTAAACACGTGCCCGCGCGA 72 |||||||||||||| |||||||| Tip 4 17 GTTTTAAACACGTATCCGCGCGA 39		Identitet je 91%.
Tip 2 1 TTTTTAAATTTA 12 |||||||||||| Tip 3 1 TTTTTAAATTTA 12	Tip 2 23 AAACACGT 30 |||||||| Tip 3 53 AAACACGT 46	Oba dijela imaju identitet od 100%, s time da je drugi dio obrnute orijentacije.
Tip 2 1 TTTTTAAATTTAGTCACGCTTTAAACACGTGTCCGCGC 38 |||||||||| || |||| |||||||||| ||||||| Tip 4 1 TTTTTAAATT-AGACACGTTTTAAACACGTATCCGCGC 37		Identitet je 89%.



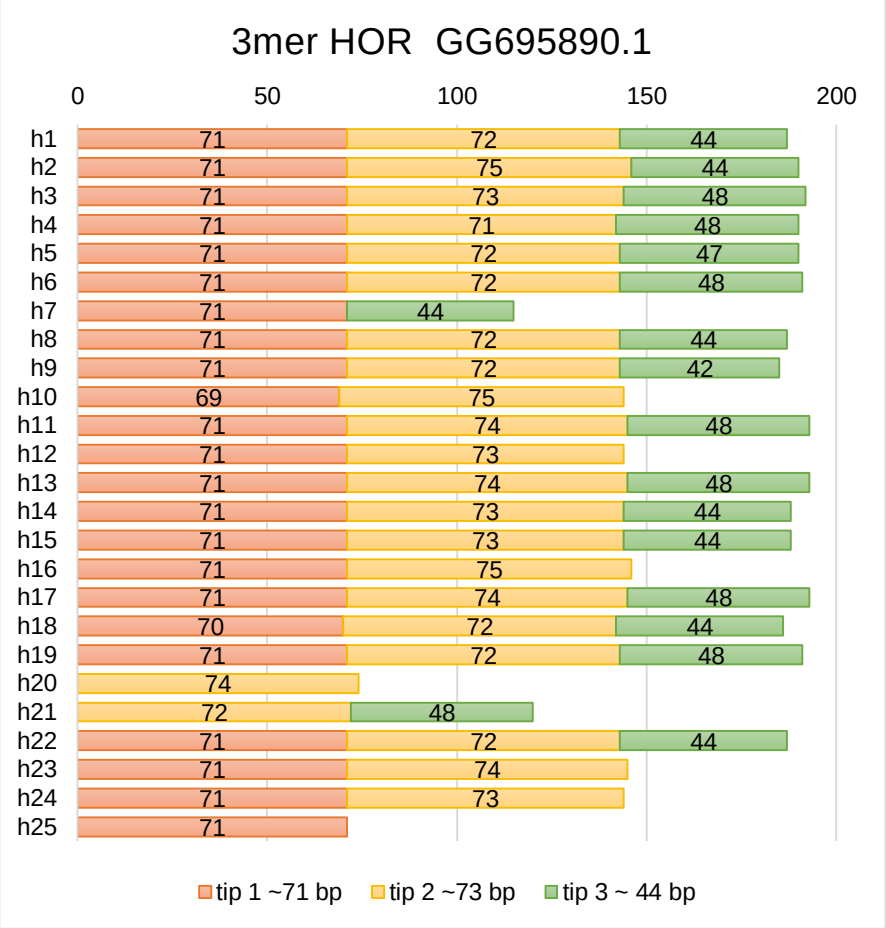
Slika 48. Prikaz 26 kopija periodičnosti višega reda sastavljene od 4 tipa monomera. Divergencija među cijelim HOR kopijama je 1,31±0,66%.

4.4.5 GG695890.1



Slika 49. GRM dijagram za GG695890.1.

Od pozicije 34 do pozicije 4096 nalazi se 65 kopija monomera ~73 bp i ~48 bp dobivenih pomoću dominantnog K-stringa CCGCGCCA. Međusobna divergencija neporavnatih kopija je 33,79±26,01%. Pomoću "heat" mape utvrdili smo da se zapravo ovi monomeri mogu podijeliti u tri tipa monomera duljina redom ~71 bp, ~72 bp i ~44 bp u skladu s GRM dijagramom (slika 49). Struktura periodičnosti višega reda ima konsenzusnu duljinu ~187 bp kao što je prikazano na slici 50.

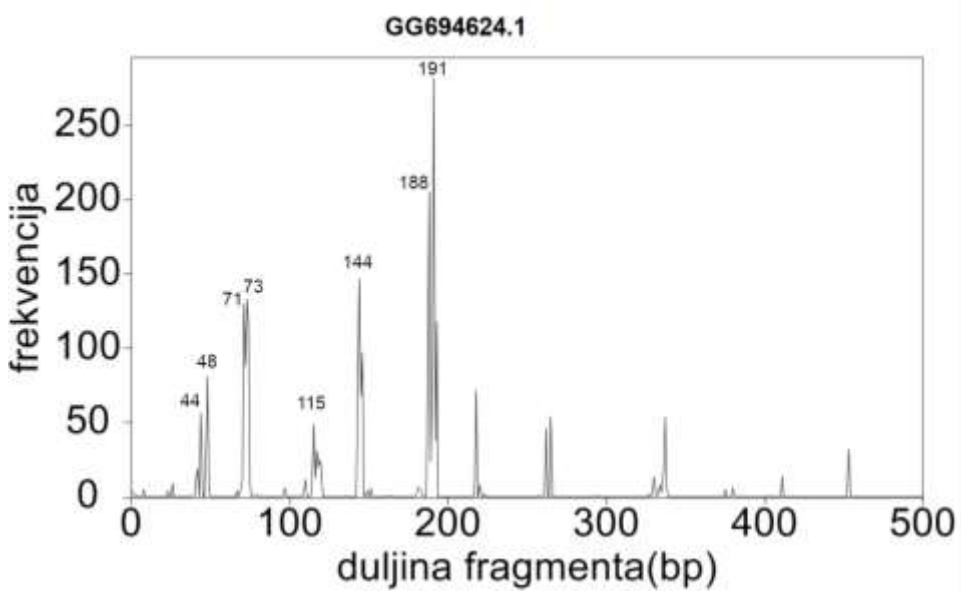


Slika 50. Prikaz 25 kopija 3mer HOR-a u GG695890.1. Divergencija cijelih HOR kopija je 3,60±1,79.

Tablica 23. Prikaz poravnanja konsenzusa monomera tipa 1 i 2.

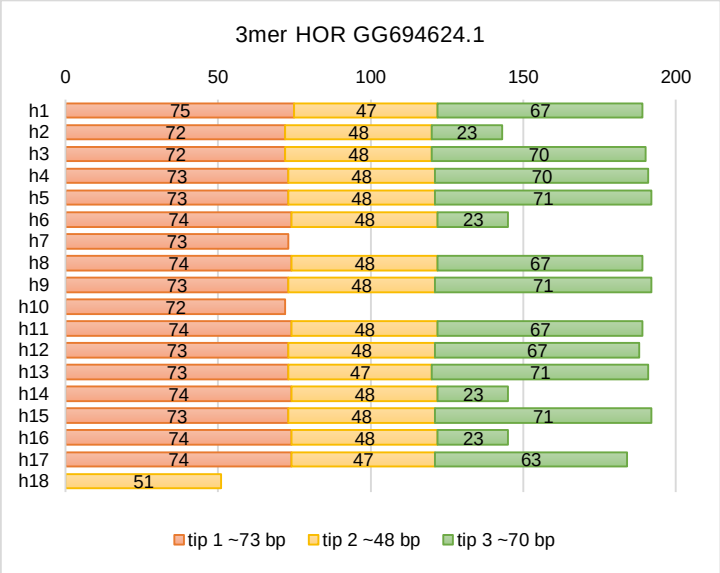
Tip 1 1 CCGCGCCATTTTTTAAATTTTGTTACCGATTTAGAGCAATTTTTAGGTT-GGCATTACTTT 59 Tip 2 1 CCGCGCCATTTTTTTAATTTAGTTACAGGATTAGCGTGATTTTTAGGTTAGAAATTGTTTT 60 Tip 1 60 TTAAACACGT 69 Tip 2 61 TTAAACACGT 70	Identitet je 81%.
Tip 1 1 CCGCGCCA--TTTTTAAATTTTGTTACCGATTTAG 33 Tip 3 1 CCGCGCCATTTTTTTTTTACTTATTTACCGATTAAG 35	Identitet je 77%.

4.4.6 GG694624.1



Slika 51. GRM dijagram za GG694624.1.

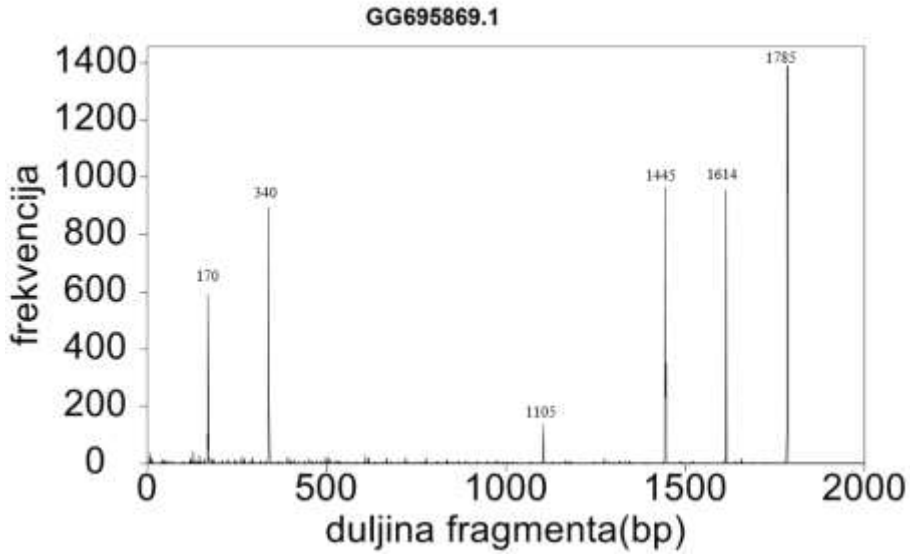
Od pozicije 5 do pozicije 2815 nalazi se 48 kopija duljina ~73 bp i ~48 bp. Međusobna divergencija neporavnatih kopija je 41,84±26,66% zbog razlika u duljinama monomera. Prema „heat“ mapi mogu se podijeliti u tri tipa monomera duljina redom ~73 bp, ~48 bp i ~70 bp (u trećem tipu pojavljuju se 4 kopije duljine 23 bp gdje je obrisano ~52 baze od pozicije 7) od kojih je sastavljena struktura koja objašnjava GRM dijagram (slika 51). Usporedbom monomera tipa 1 i 3 s Needleman-Wunsch dobili smo divergenciju od 36,05±3,12%. Cijele HOR kopije, prikazane na slici 52, imaju divergenciju od 4,25±2,36%.



Slika 52. Prikaz 3mer HOR-a konsenzusne duljine ~191 bp.

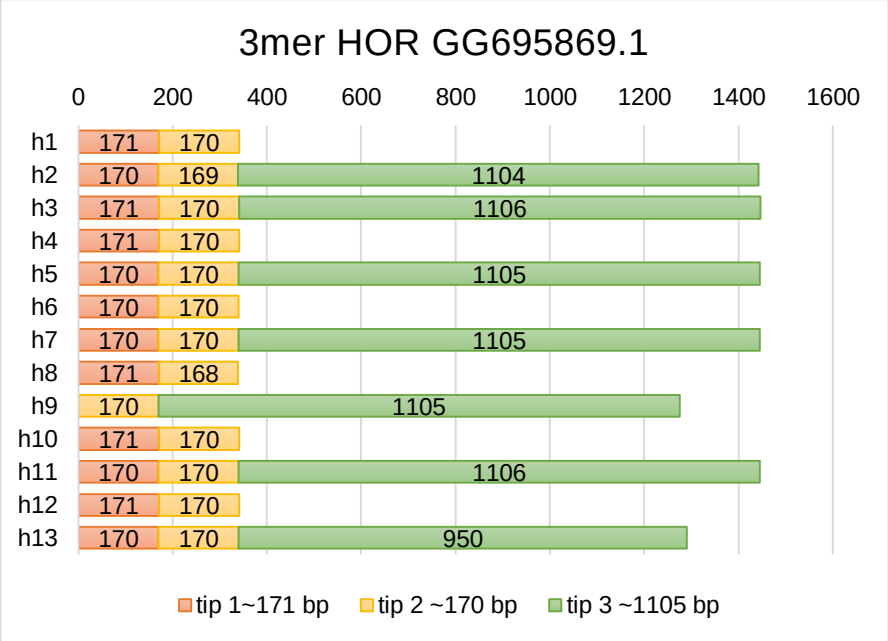
4.5 Pik 170 bp

4.5.1 GG695869.1



Slika 53. Prikaz GRM dijagrama GG695869.1.

Pomoću dominantnog K-stringa AAACCGCA od pozicije 787 do pozicije 11670, dobili smo 32 monomera duljina 171bp, 170 bp i 1105 bp, za koje se preko divergencija u „heat“ mapi vidi da pripadaju trima različitim tipovima monomera na temelju kojih je identificirana periodičnost višega reda konsenzusne duljine ~1445 bp kao što je prikazano na slici 54 i pri čemu se mogu objasniti pikovi na GRM dijagramu na slici 53.



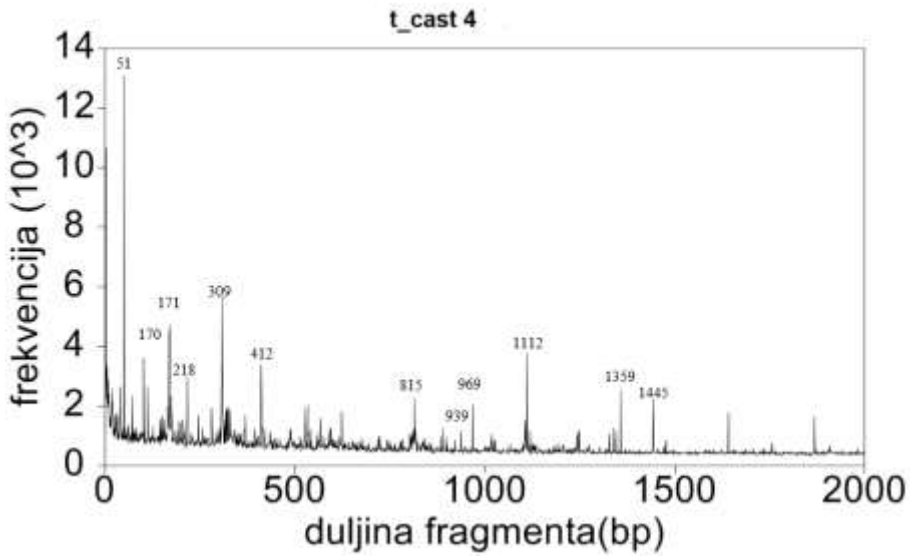
Slika 54. Prikaz 3mer HOR-a konsenzusne duljine ~1445 bp. Divergencije među cijelim HOR kopijama je 5,04±5,74%. Divergencija monomera tipa 1 i 2 je 11,57±9,32%.

Usporedbom monomera tipa 1 i tipa 3, tipa 2 i 3 pomoću BLAST-a, dobili smo kao rezultat poravnanje prikazano u tablici 24.

Tablica 24. Prikaz BLAST poravnanja monomera tipa 1 s tipom 3 i tipa 2 s tipom monomera 3, pri čemu se vidi da se monomeri tipa 1 i 2 slažu za prvih ~63 bp i zadnjih ~84 bp s monomerom tipa 3.

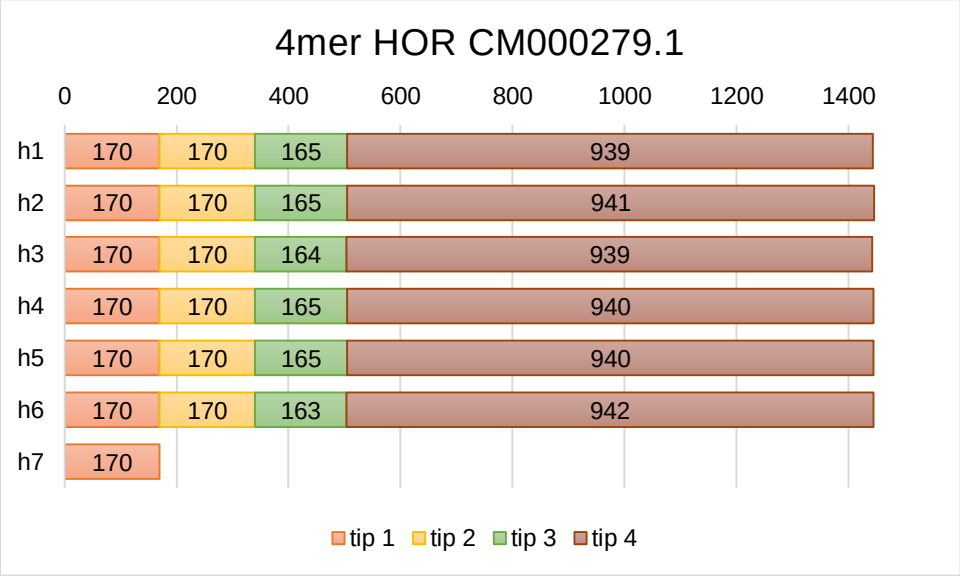
Tip 1 1 AAACCGCATTGCTCTACGACTTTTAGTTTTCAAGTTATGAttttttttttGTTGAATAAAA 60 Tip 3 1 AAACCGCATTGCTCTACGACTTTTAGTTATTTTTTTATGA-ATTTTTTTATTGAA-AAAC 58 Tip 1 61 TTT 63 Tip 3 59 TTT 61	Identitet je 84%.
Tip 1 88 AATTTTTTGCAAAAAATGTTATAACTTTTGATCTCATTAAATGTCGGTATAATACAGAGA 147 Tip 3 1022 AAATTTTAGTAAAAAAA-TCAAATTTTTAAATCTCATGAAAAGTTGTTATAATACAGAAA 1080 Tip 1 148 AAAAGCCGCTGAATCCAATGGAAC 171 Tip 3 1081 ATGAGCCGCTGAATTCAATGGAAC 1104	Identitet je 77%.
Tip 2 1 AAACCGCATTGCTCTACGACTGTTAGTTTTTGAGTTATAAATTTTTTGTTGGATAAAATT 60 Tip 3 1 AAACCGCATTGCTCTACGACTTTTAGTTATTTTTTTATGAATTTTTTTATTGAAAAACTT 60 Tip 2 61 T 61 Tip 3 61 T 61	Identitet je 82%.
Tip 2 86 aaatttaggcaaaaaattttaaaattttaattcttgtgaaaaattgatattatatgtaaa 145 Tip 3 1023 AATTTTAGTAAAAAAATC--AAATTTTTAAATCTCATGAAAAGTTGTTATAATACAGAAA 1080 Tip 2 146 atGAGCCGCTGAATTCAATGGAAC 169 Tip 3 1081 ATGAGCCGCTGAATTCAATGGAAC 1104	Identitet je 81%.

4.5.2 *CM000279.1.*



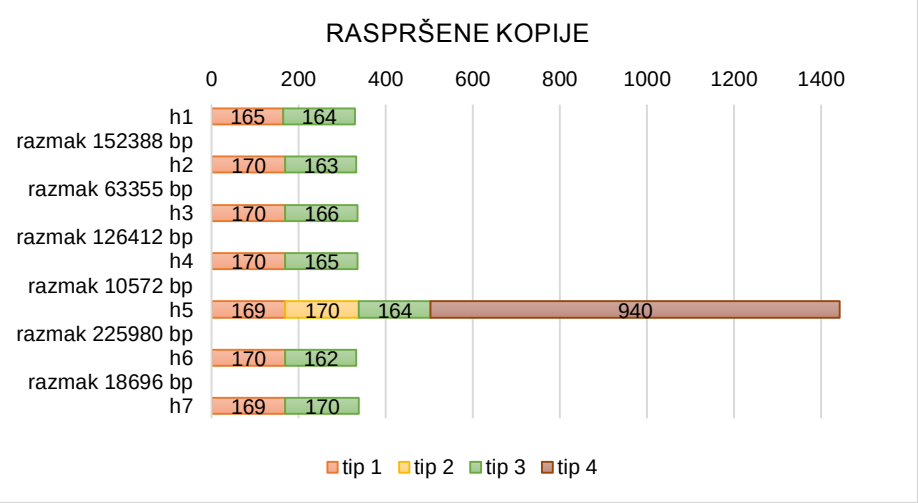
Slika 55. GRM dijagram za CM000279.1 – LG4.

U CM000279.1 (LG4), slika 55, identificirali smo 4mer HOR strukturu, od pozicije 2692170 bp do 2700838 bp pomoću key stringa AACTAAA, zasnovanu na tri tipa monomera od ~170 bp i jednom ~940 bp konsenzusne duljine ~1445 bp (objavljeno u [147]). HOR struktura sastoji se od 6 potpunih kopija kao što je prikazano na slici 56.



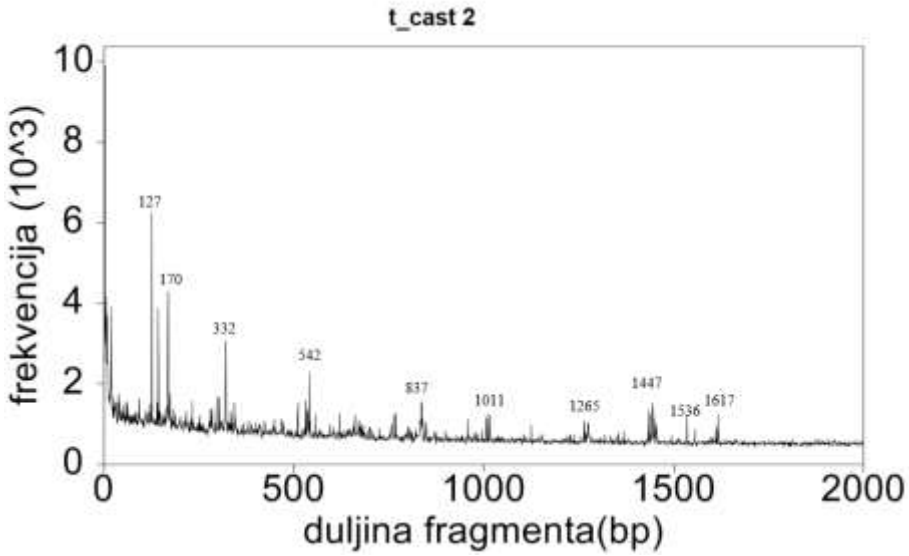
Slika 56. Prikaz HOR-strukture u CM000279.1. Divergencija među HOR kopijama je 3,22±3,06%. Preuzeto i prerađeno iz [147].

Divergencije između kopija su od 2% do 20%, ovisno o tome da li se radi o monomerima istog ili različitih tipova. Također smo pronašli i sedam raspršenih kopija ispred HOR –a na udaljenosti od 14 kb u sekvenci. Unutar raspršenih kopija monomeri tipa 2 i 4 su često obrisani kao što se može vidjeti na slici 57.

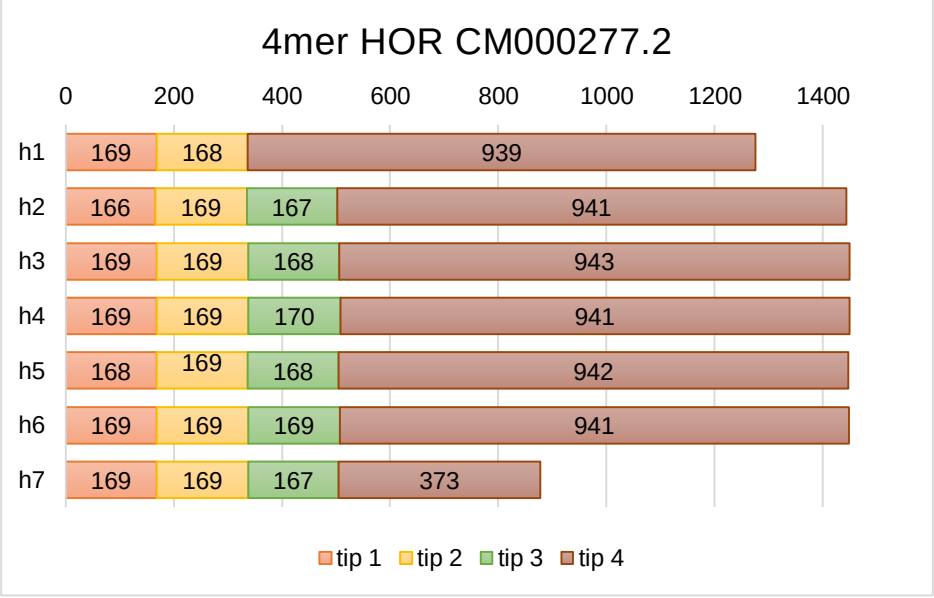


Slika 57. Prikaz raspršenih kopija od pozicije 1908980 bp do 2677761 bp, s naznačenim razmacima između njih. Preuzeto iz [147].

U CM000277.2 – LG2 (slika 58) nalazi se obrnuti komplement ovog HOR-a koji se sastoji od sedam kopija s konsenzusnom duljinom ~1447 bp (slika 59). Monomeri istog tipa imaju divergenciju manju od 2%, a različiti imaju ~20%, s time da se u ovom HOR-u monomeri tipa 2 i tipa 3 razlikuju za 3% što može biti rezultat nedavne duplikacije ovih monomera.



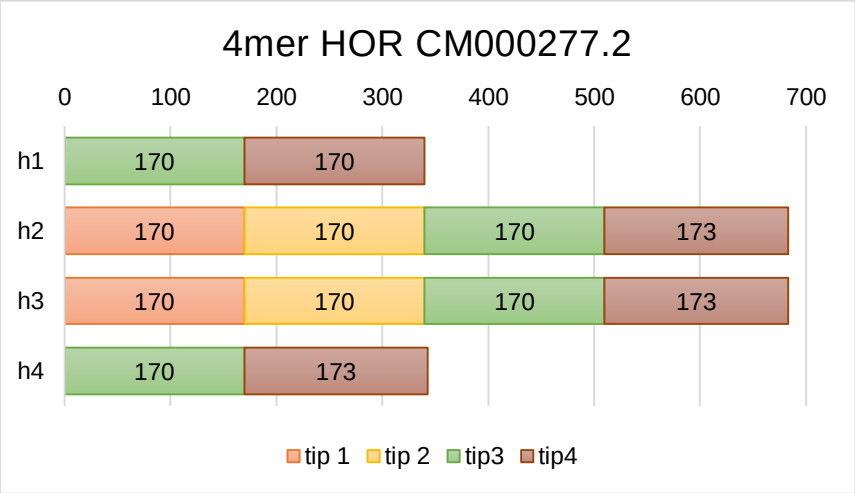
Slika 58. GRM dijagram za CM000277.2 (LG2). Preuzeto i prerađeno iz [147].



Slika 59. Obrnuti komplement HOR-a iz CM000279.1 u CM000277.2 konsenzusne duljine ~1447 bp. Divergencija između HOR kopija je 1,17±0,29%. Preuzeto i prerađeno iz [147].

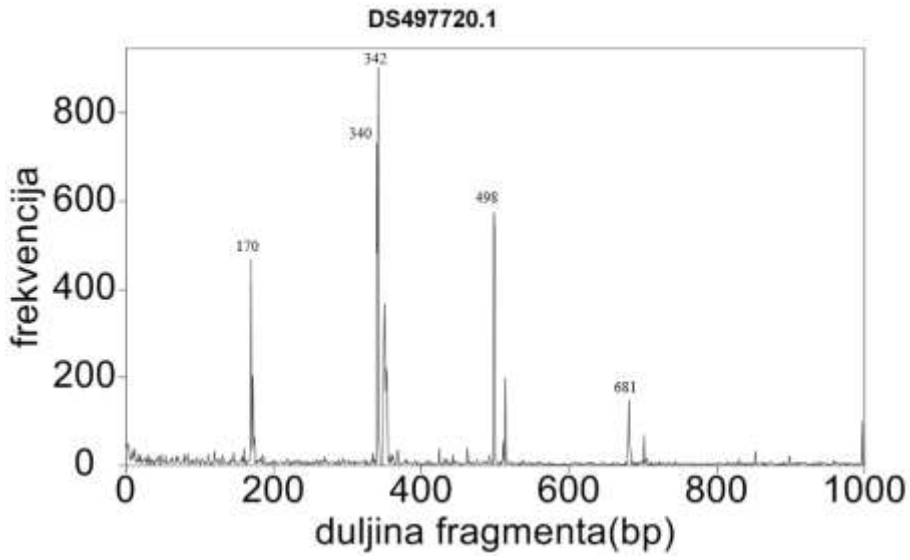
4.5.3 CM000277.2

U CM000277.2 identificirali smo periodičnost višega reda zasnovanu na četiri različita tipa monomera duljine ~170 bp u četiri kopije konsenzusne duljine ~680 bp (dvije cijele kopije) prikazane na slici 60. Divergencija između različitih tipova monomera je ~16%.



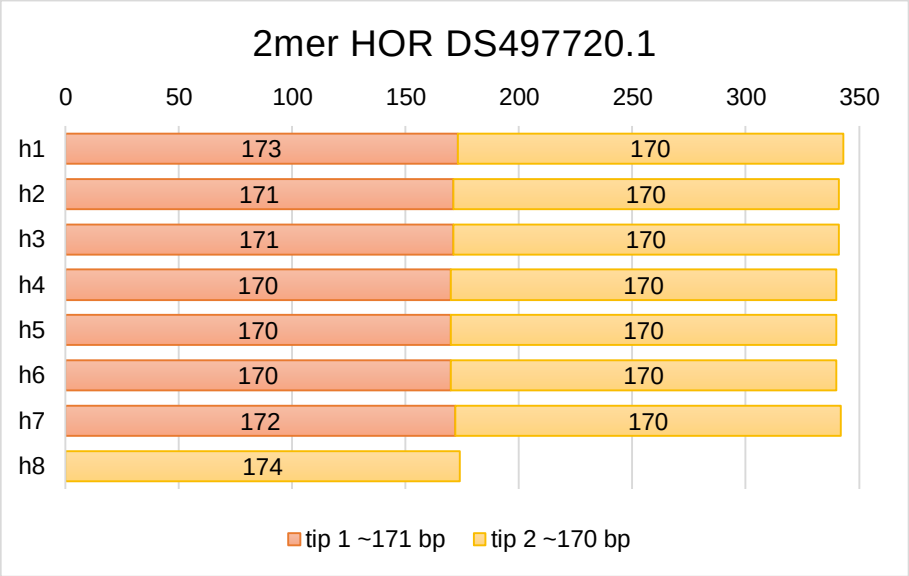
Slika 60. HOR struktura zasnovana na četiri tipa monomera duljina ~170 bp s divergencijom HOR kopija od 0,44±0,00% za cijele h2 i h3.

4.5.4 DS497720.1



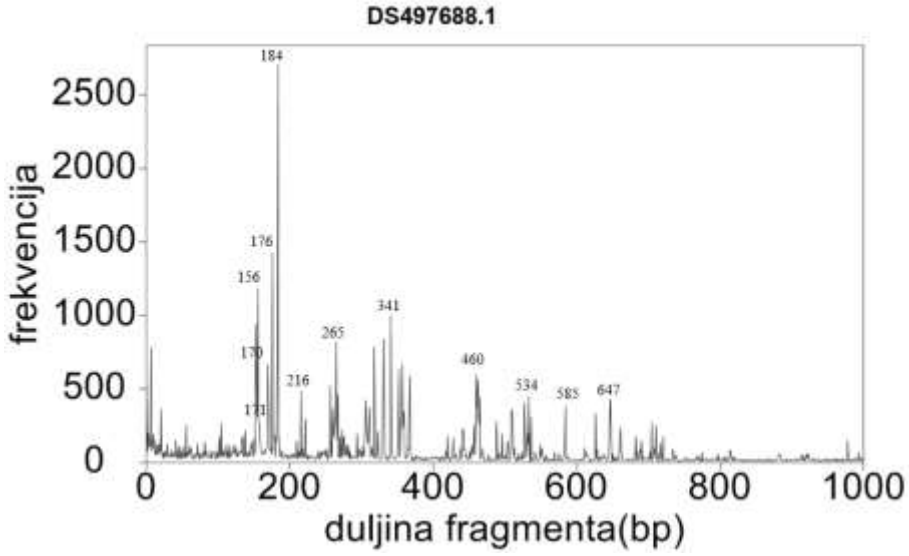
Slika 61. GRM dijagram DS497720.1.

Od pozicije 725 do 3112 dobili smo pomoću K-stringa TTGTTCCA dimer koji se sastoji od dva tipa monomera duljina ~171 bp i ~170 bp s međusobnom divergencijom od 22,17±13,54%. Ova dva tipa tvore dimer konsenzusne duljine ~340 bp kao što je prikazano na slici 62.



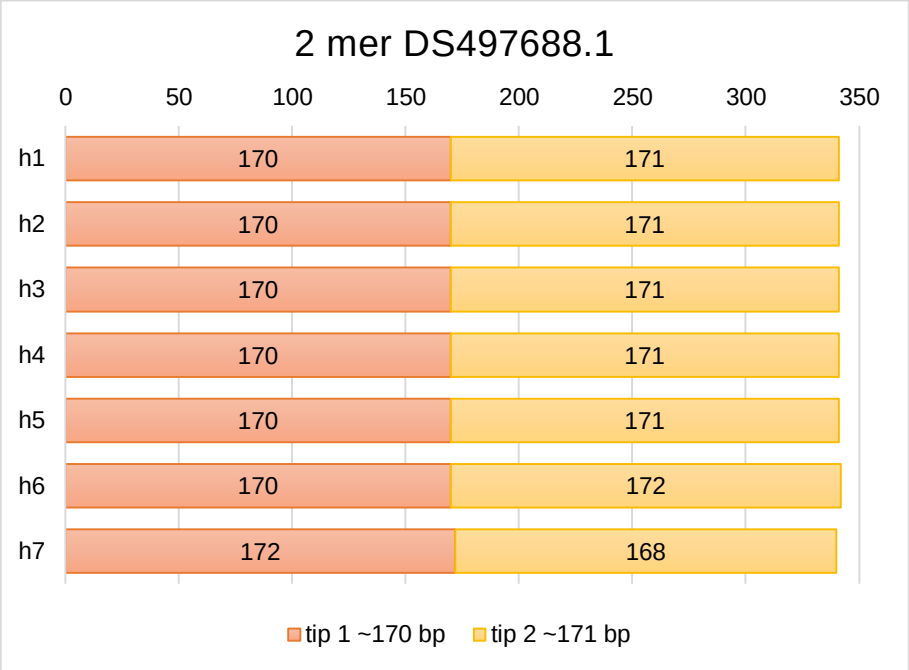
Slika 62. HOR struktura dimera u DS497720.1, složenih u osam kopija s međusobnom divergencijom od 6,53±2,57%.

4.5.5 DS497688.1



Slika 63. GRM dijagram za DS497668.1.

Od pozicije 74406 do pozicije 76625 dobili smo s K-stringom AAAAATAA 14 kopija, složenih u 2 tipa monomera duljina ~170 bp i ~171 bp s međusobnom divergencijom od 12,29±7,98% i složenih u 7 HOR kopija duljine ~341 bp što objašnjava sliku 63. Struktura je prikazana na slici 64. Poravnanje konsenzusa ova dva tipa monomera prikazano je u tablici 25.



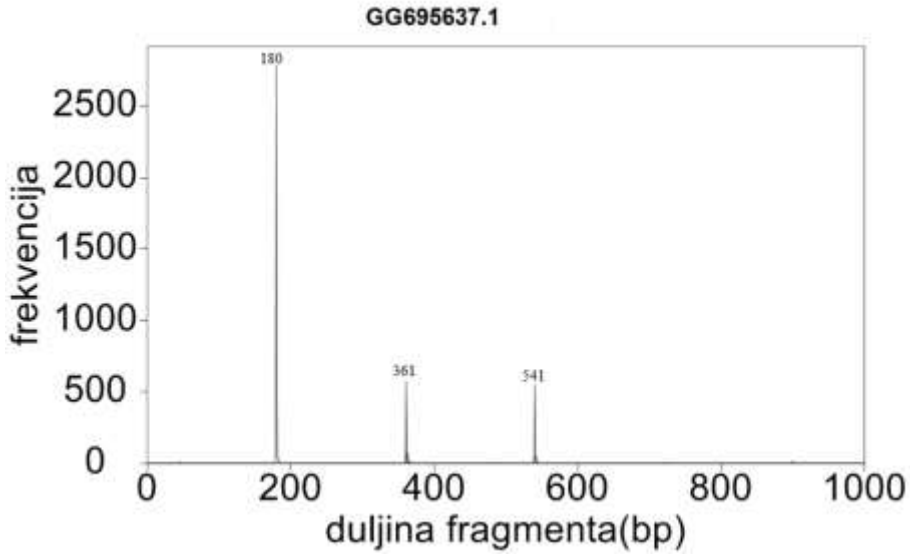
Slika 64. Prikaz strukture dimera u DS497688.1. Divergencija među kopijama je 7,70±7,78%.

Tablica 25. Poravnanje konsenzusnih sekvenci monomera tipa 1 i 2 u DS497688.1.

Tip 1 1 AAAAATAAAAACACAACCGAATTCCCCTTTTAATTATCTATCGGACACCGAAAACCGAAA 60 Tip 2 1 AAAAATAAAAACACCATCGCATTCCCCTGTTAATTATCTATCGGACAGTGAAAACCGGAA 60 Tip 1 61 GCCAATCGAACTTATAGTTTCCATACGGAAGCGTTTTAAAATATCTCCGTAGATT--TTT 118 Tip 2 61 GCTAATCGGACTGATAGTTTCCGTAGGGAAGCGTTTTAAAGTATACGCTTAGAATCGCTA 120 Tip 1 119 CGGCTATCATATTTTCTAACGaaaaaaaCTCCAGttttttttCGTC 163 Tip 2 121 GAAGTACCATATTTTCTAATG-AAAAAACTCCGGTTTTTTTCGTC 164	Identitet je 82%.

4.6. Pik 180 bp

4.6.1 GG695637.1.



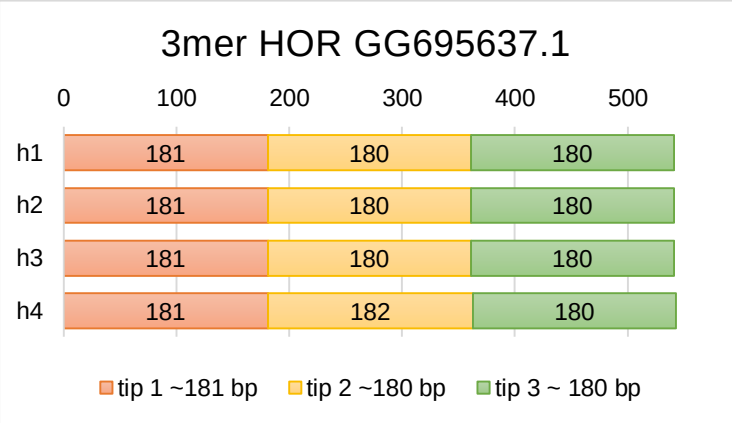
Slika 65. GRM dijagram za GG695637.1.

Od pozicije 40 do pozicije 5454 nalazi se 31 monomer duljina ~180 bp koje smo dobili pomoću K-stringa AAAAATTA. Iako je međusobna divergencija među kopijama monomera 6,63±3,64%, koja je dobivena s Needleman-Wunsch programom i upućuje na tandemne repeticije, može se vidjeti zapravo slabi HOR signal (slika 65) koji je objašnjen s „heat“ mapom (slika 66) u kojoj se vidi da od monomera m20 do monomera m31 postoji naznaka periodičnosti višega reda.

		m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24	m25	m26	m27	m28	m29	m30	m31
		40	226	406	586	766	946	1126	1307	1487	1668	1848	2028	2208	2388	2568	2748	2928	3108	3288	3468	3649	3829	4009	4190	4370	4550	4731	4911	5091	5272	5454
m1	40	0	11	12	11	12	10	10	12	10	12	8,6	11	10	10	11	10	10	11	11	9,7	13	11	9,7	12	11	9,7	12	11	9,7	13	11
m2	226		0	7	7	8	7	12	3	12	2,8	6,1	6,7	6,1	6,1	6,7	6,1	6,1	6,7	7,2	12	3,9	1,7	12	3,3	1,7	12	2,8	1,7	12	4,4	7,2
m3	406			0	3	2	3	11	7	11	7,8	4,4	3,3	3,9	2,8	3,3	3,9	2,8	2,2	2,8	11	8,9	5,6	11	8,3	5,6	9,9	7,8	5,6	11	9,3	2,8
m4	586				0	3	1	9	6	9	7,2	3,3	1,1	1,7	0,6	1,1	1,7	0,6	2,2	2,8	9,4	8,3	5	9,4	7,8	5	9,9	7,2	5	9,4	8,8	1,7
m5	766					0	3	11	7	11	8,3	3,9	3,3	3,9	2,8	3,3	3,9	2,8	1,1	1,7	11	8,8	6,1	11	8,9	6,1	11	8,3	6,1	11	9,9	2,8
m6	946						0	8	6	8	7,2	3,3	1,1	1,7	0,6	1,1	1,7	0,6	2,2	2,8	8,3	8,3	5	8,3	7,8	5	8,8	7,2	5	8,3	8,8	1,7
m7	1126							0	12	0	11	8,3	8,3	8,8	8,8	8,3	8,8	8,8	9,9	11	0	12	11	0	12	11	1,1	11	11	0,6	13	9,4
m8	1307								0	12	5,6	5,6	6,1	5,6	5,6	6,1	5,6	5,6	5,5	6,1	12	4,4	1,1	12	6,1	1,1	12	5,6	1,1	12	7,1	6,7
m9	1487									0	11	8,3	8,3	8,8	8,8	8,3	8,8	8,8	9,9	11	0	12	11	0	12	11	1,1	11	11	0,6	13	9,4
m10	1668										0	6,7	7,2	6,7	6,7	7,2	6,7	6,7	7,2	7,8	11	1,1	4,4	11	0,6	4,4	11	0	4,4	11	1,7	7,8
m11	1848											0	3,3	2,8	2,8	3,3	2,8	2,8	2,8	3,3	8,3	7,8	4,4	8,3	7,2	4,4	7,7	6,7	4,4	8,3	8,2	3,9
m12	2028												0	1,7	0,6	0	1,7	0,6	2,2	2,8	8,3	8,3	5	8,3	7,8	5	8,8	7,2	5	8,3	8,8	1,7
m13	2208													0	1,1	1,7	0	1,1	2,8	3,3	8,8	7,8	4,4	8,8	7,2	4,4	9,4	6,7	4,4	8,8	8,2	2,2
m14	2388														0	0,6	1,1	0	1,7	2,2	8,8	7,8	4,4	8,8	7,2	4,4	9,4	6,7	4,4	8,8	8,2	1,1
m15	2568															0	1,7	0,6	2,2	2,8	8,3	8,3	5	8,3	7,8	5	8,8	7,2	5	8,3	8,8	1,7
m16	2748																0	1,1	2,8	3,3	8,8	7,8	4,4	8,8	7,2	4,4	9,4	6,7	4,4	8,8	8,2	2,2
m17	2928																	0	1,7	2,2	8,8	7,8	4,4	8,8	7,2	4,4	9,4	6,7	4,4	8,8	8,2	1,1
m18	3108																		0	0,6	9,9	7,7	5	9,9	7,8	5	9,4	7,2	5	9,9	8,8	1,7
m19	3288																			0	11	8,3	5,6	11	8,3	5,6	9,9	7,8	5,6	11	9,3	2,2
m20	3468																				0	12	11	0	12	11	1,1	11	11	0,6	13	9,4
m21	3649																					0	5,6	12	1,7	5,6	12	1,1	5,6	12	2,8	8,9
m22	3829																						0	11	5	0	11	4,4	0	11	6	5,6
m23	4009																							0	12	11	1,1	11	11	0,6	13	9,4
m24	4190																								0	5	12	0,6	5	12	2,2	8,3
m25	4370																									0	11	4,4	0	11	6	5,6
m26	4550																										0	11	11	0,6	13	9,9
m27	4731																											0	4,4	11	1,7	7,8
m28	4911																												0	11	6	5,6
m29	5091																													0	13	9,4
m30	5272																														0	9,3
m31	5454																															0

Slika 66. Prikaz "heat" mape za GG695637.1. Monomeri od pozicije 3468 do 5454 daju 3mer HOR.

Identificirani HOR sastoji se od tri tipa monomera duljina redom 181 bp, ~180 bp i 180 bp koji daju konsenzusnu duljinu ~541 bp, kao što je prikazano na slici 67.

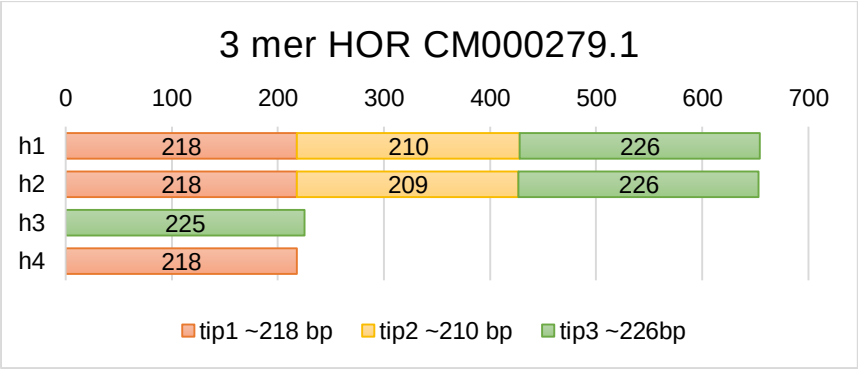


Slika 67. Prikaz 3mer HOR-a konsenzusne duljine ~541 bp, s divergencijom između HOR kopija od 1,69±1,18%.

4.7 Pik 218 bp

47.1 CM000279.1

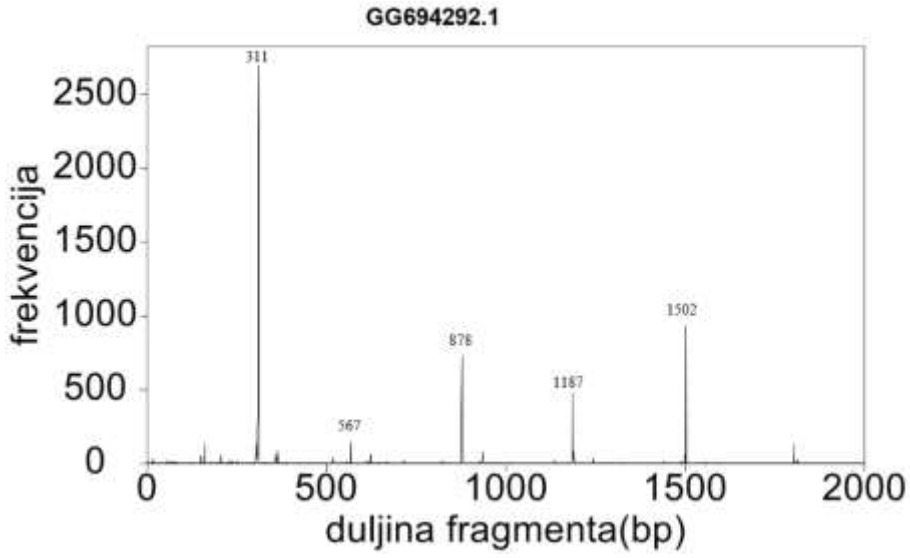
Na slici 55, kao jedan od pikova je i pik 218 bp. Od pozicije 2187547 do pozicije 2189082 nalazi se osam kopija monomera duljina ~218 bp, ~210 bp i 226 bp dobivenih sa K-stringom AATAAAAA. Ovi monomeri imaju međusobnu divergenciju od 18,63±4,95%. Monomere smo složili u 3mer HOR prema „heat“ mapi kao što je prikazano na slici 68.



Slika 68. 3mer HOR struktura konsenzusne duljine ~654 bp u CM000279.1. Divergencija između dvije potpune HOR kopije je 5,34±0.00%.

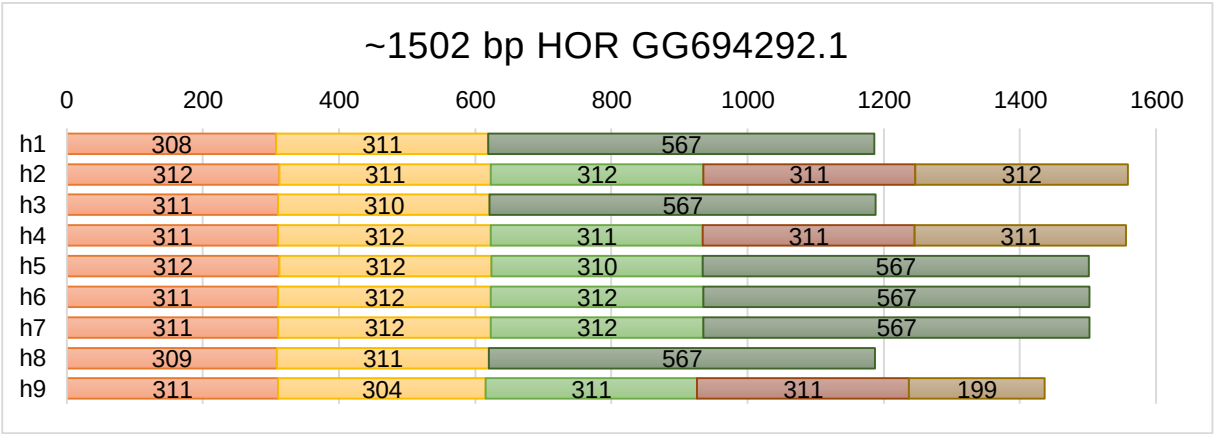
4.8. Pik 311 bp

4.8.1 GG694292.1



Slika 69. GRM dijagram za GG694292.1.

Od pozicije 28 do pozicije 12444 dobili smo 36 monomera duljina ~311 bp i ~567 bp sa K-stringom AAGCGATA. Međusobna divergencija ovih monomera je 21,87±26,99%. Prema „heat" mapi konstruirali smo periodičnost višega reda koja se sastoji od jednog tipa monomera ~311 bp i drugog tipa ~567 bp, s time da se neke kopije sastoje samo od monomera tipa 1 ponovljenog pet puta ili tri puta ponovljenog monomera tipa 1 i jednog tipa 2, a neke od dva monomera tipa 1 i jednog tipa 2 kao što je prikazano na slici 70.



Slika 70. HOR struktura u GG694292.1. Divergencije HOR kopija s 2 monomera tipa 1 i jednim monomerom tipa 2 je 1,85±0,96%, divergencija 3 monomera tipa 1 i jednog tipa 2 je 9,03±4,99% i divergencija kopija sastavljenog od pet monomera tipa 1 je 1,44±0,51%.

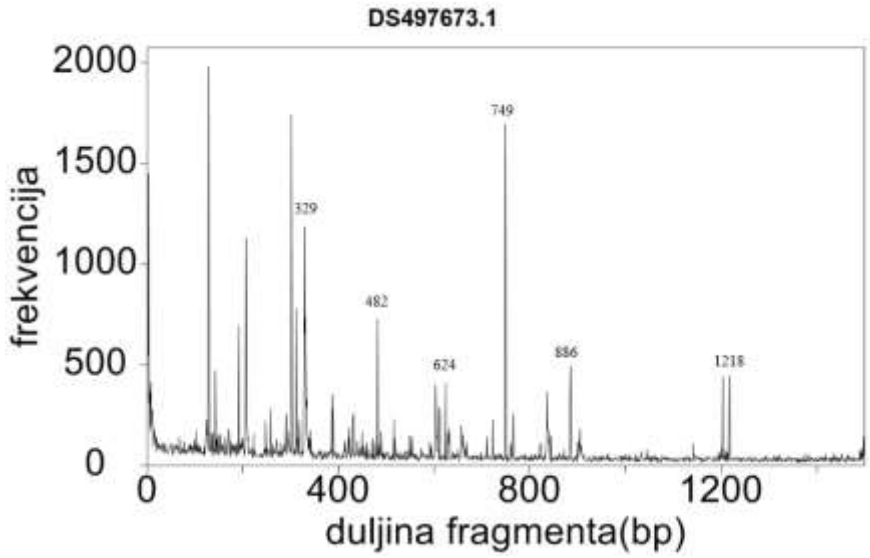
Ova kompleksna struktura objašnjava GRM dijagram na slici 69, pri čemu se vidi na koji način se mogu dobiti pikovi kako je prikazano u tablici 26.

Tablica 26. Prikaz objašnjenja pikova na dijagramu na slici 68 odnosno HOR strukture u GG694292.1.

	h1	h2	h3	h4	h5	h6	h7	h8	h9
tip1	308	312	311	311	312	311	311	309	311
	311	311	310	312	312	312	312	311	304
		312		311	310	312	312		311
		311		311					311
		312		311					199
tip2	567		576		567	567	567	567	
suma	**1186**	**1558**	**1197**	**1556**	**1501**	**1502**	**1502**	**1187**	**1436**

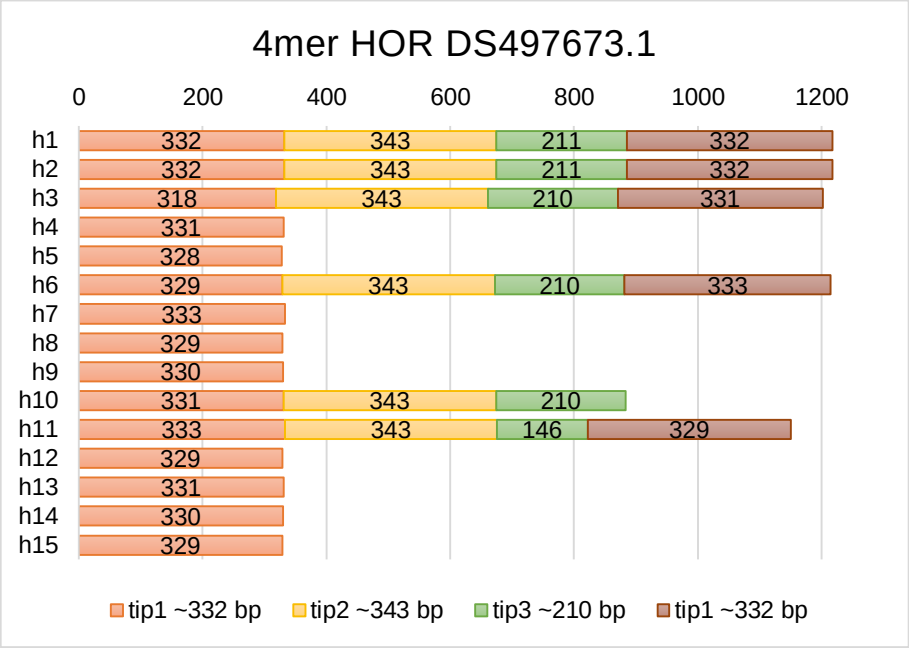
4.9 Pik 328

4.9.1 DS497673.1



Slika 71. GRM dijagram DS497673.1.

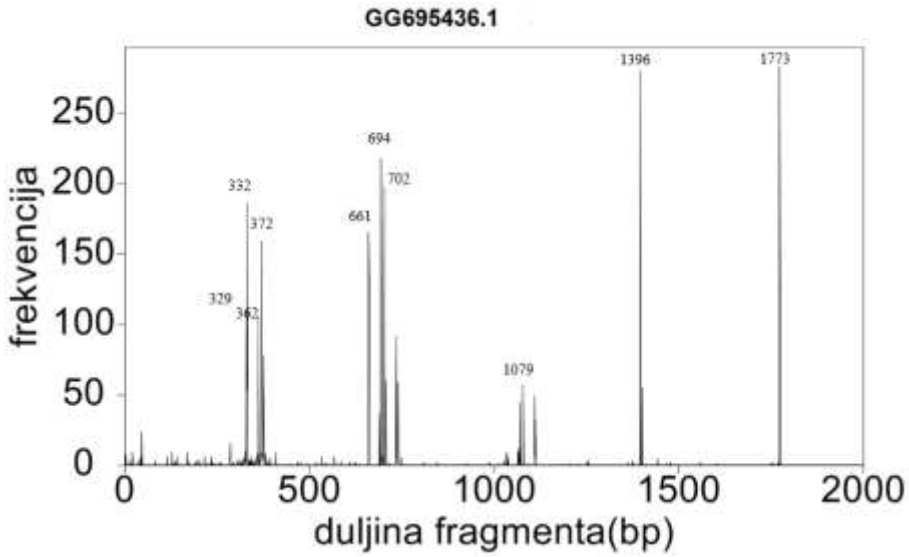
Od pozicije 13 do pozicije 9542 nalaze se 32 monomera duljina ~332 bp, ~211 bp i 343 bp dobivenih sa K-stringom TAAAAAAA. Međusobna divergencija između monomera je 29,84±24,49%. Ovi monomeri se mogu podijeliti u tri tipa i prema „heat“ mapi može se vidjeti da se i u ovom slučaju radi o 4mer HOR-u konsenzusne duljine ~1218 bp čija struktura je prikazana na slici 72.



Slika 72. 4mer HOR struktura konsenzusne duljine ~1218 bp. Divergencija između cijelih kopija je 2,19±0,48%.

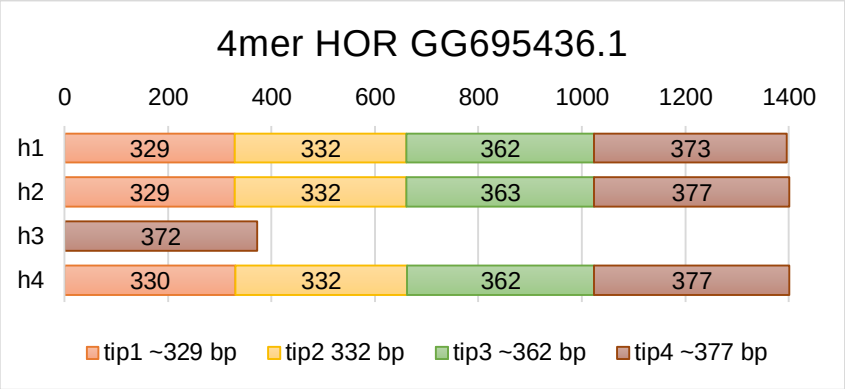
4.10 Pik 332 bp

4.10.1 GG695436.1



Slika 73. GRM dijagram za GG695436.1.

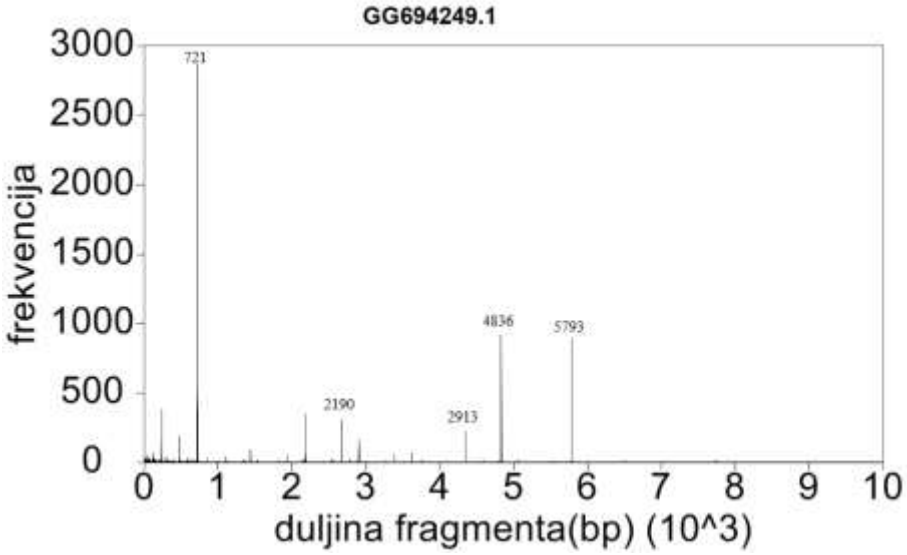
Od pozicije 338 do pozicije 4531 dobili smo 13 monomera s međusobnim divergencijama 19,04±9,36%. Pomoću "heat" mape vidi se da se tih 13 monomera može svrstati u 4 tipa kojima pripadaju duljine redom ~329 bp, 332 bp, ~362 bp i ~377 bp. Ta četiri tipa monomera tvore četiri kopije, 4mer HOR-a konsenzusne duljine ~1396 bp, kao što je prikazano na slici 74 i prema čemu se mogu objasniti pikovi sa slike 73.



Slika 74. 4mer HOR struktura konsenzusne duljine ~1396 bp s divergencijom cijelih HOR kopija od 1,88±0,80%.

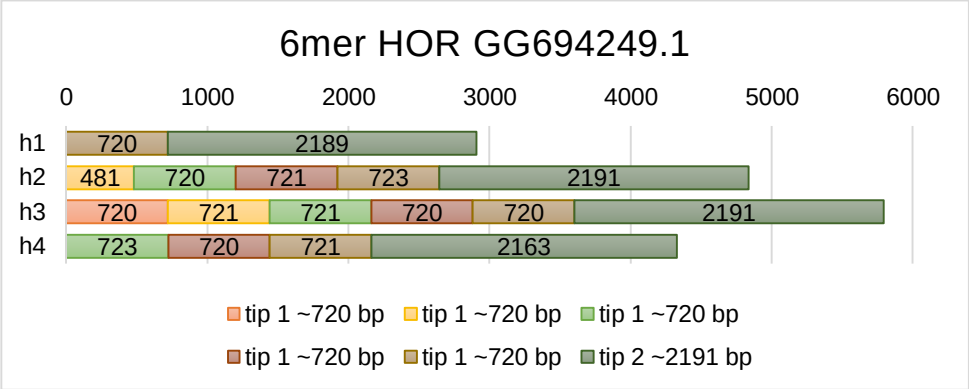
4.11 Pik 721

4.11.1 GG694249.1



Slika 75. GRM dijagram za GG694249.1.

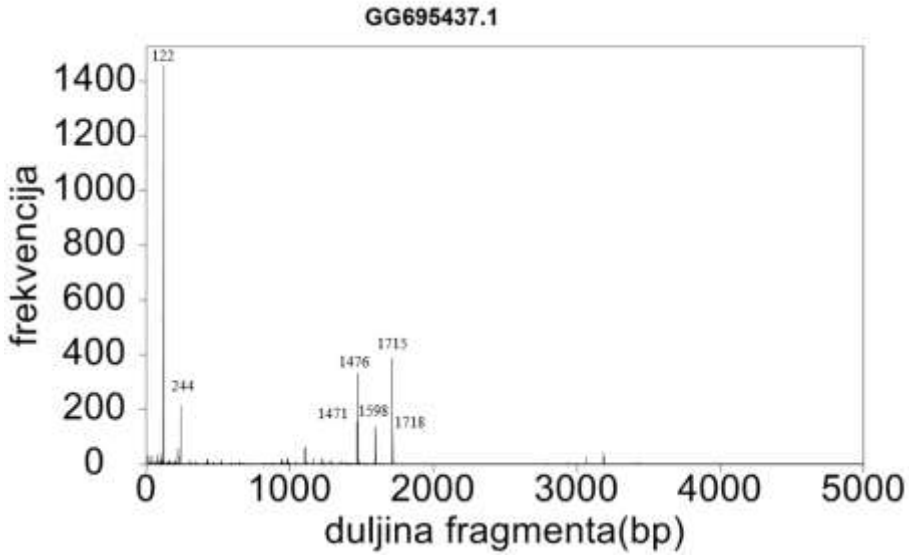
Od pozicije 76 do pozicije 15777 nalazi se 17 monomera duljina ~720 bp i ~2190 bp dobiveni sa K-stringom GGCCTTAT (s međusobnom divergencijom 39,25±35,03%). Pomoću „heat“ mape vidi se da oni tvore periodičnost višega reda konsenzusne duljine ~5793 bp. Struktura HOR-a je prikazana na slici 76 i ona ujedno objašnjava pikove prikazane na GRM dijagramu na slici 75. Ova kompleksna struktura sastoji se od 1, 3 ili 5 monomera tipa 1 i jednog monomera tipa 2. Pomoću BLAST-a smo usporedili monomer tipa 1 i tipa 2 i dobili smo da se prvih 671 bp poklapa s identitetom od 94%.



Slika 76. Prikaz kompleksne strukture u GG694249.1.

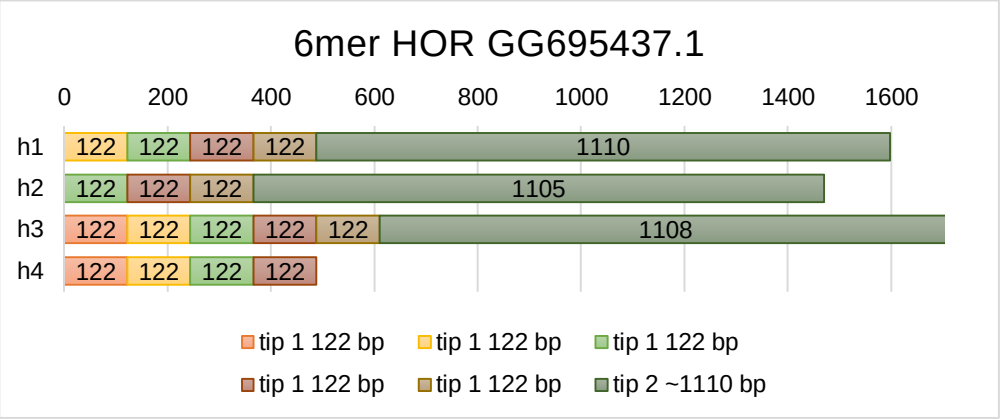
4.12. Pik 1110 bp

4.1.12 GG695437.1



Slika 77. GRM dijagram za GG695437.1.

Od pozicije 425 do pozicije 5578 pomoću K-stringa AAAATCAC dobili smo 19 monomera duljina 122 bp i ~1110 bp s međusobnom divergencijom između istog tipa monomera 4,85±2,62%, za tip 122 bp, i 2,22±0,45% za tip 2 ~1110 bp. Pomoću „heat“ mape zaključili smo da se radi o ~6mer HOR-u duljine ~1718 bp, kao što se vidi na slici 78 i što objašnjava GRM dijagram na slici 77.



Slika 78. 6mer HOR struktura u GG695437.1. U četiri HOR kopije, samo jedna ima svih šest monomera ukupne duljine 1718 bp, dok ostale imaju ukupne duljine za HOR prvu kopiju 1598 bp, drugu kopiju 1471 bp i za četvrtu kopiju 488 bp. Divergencija između HOR kopija h1 i h3 je 2,19%.

5. Zaključak

Pomoću računalne metode Global Repeat Map pregledali smo sekvencionirani genom insekta *T. castaneum*, Tcas 3.0, kako bi identificirali periodičnosti višega reda koje do sada nisu bile istraživane. Dobiveni rezultati upućuju da postoje periodičnosti višega reda, posebno za one komponente genoma koje sadrže TCAST satelite što ukazuje da upravo ti sateliti imaju bitnu ulogu u strukturi i funkciji centromere, kao što je to slučaj kod ljudskih alfa satelita. Identifikacija tih HOR-ova bila je otežana zbog tehničkih poteškoća pri sastavljanju sekvencioniranog dijela genoma. Oni se nalaze u komponentama koje nisu uspješno smještene na kromosome pomoću današnjih tehnika. Poboljšanjem tih tehnika mogla bi se dobiti u budućnosti potpuna slika svih periodičnosti višega reda. Pomoću GRM metode identificirali smo pravilne i kompleksne HOR strukture sastavljene od TCAST satelita. Jedan od razloga zbog kojih oni nisu ranije detektirani su varijacije u duljini monomera koje postojeći algoritmi nisu mogli uočiti. Za slučaj TCAST satelita prosječna divergencija između monomera je oko 20%, a periodičnosti višega reda koje su konstruirane na temelju njih imaju prosječnu divergenciju od oko 4 %. Ljudski HOR-ovi zasnovani na alfa satelitima duljine 171 bp, imaju sličnu divergenciju monomera. Divergencije između alfa satelita su oko 20%, a HOR-ova oko 2% kao što se može vidjeti na primjeru ljudskih kromosoma 1 i Y [21, 23]. Jedna bitna razlika između HOR-ova kod čovjeka i *T.castaneum* su varijacije između duljina monomera u tim satelitima koje kod čovjeka iznose oko 2% a kod *T.castaneum* taj postotak može biti i do 15%. Ovo otkriće velikih divergencija u duljinama monomera daje nam mogućnost novog razmišljanja o periodičnostima višega reda i njihovoj identifikaciji, kako bi se razjasnila njihova evolucija i uloga. Strukture koje su zasnovane na TCAST satelitima pojavljuju se kao 2 mer, 3 mer, 4 mer, 5mer, 6mer s malim brojem cijelih kopija (~3 kopije) osim 4 mer HOR u DS497953.1 s 4 cijele kopije i 5mer HOR u GG695826.1 s 6 cijelih kopija.

Osim HOR-ova zasnovanih na TCAST satelitima, otkrili smo i druge periodičnosti višega reda u obliku 2mer, 3mer, 4mer, 5 mer i 6mer HOR struktura, s time da se ove periodičnosti višega reda nalaze u većem broju kopija (do 16 cijelih u pojedinim HOR strukturama) u odnosu na TCAST satelite.

Za pojedine identificirane HOR strukture karakteristično je da izmjenjuju tipove monomera jedne s drugim, kao što je slučaj kod HOR struktura identificiranih u GG694292.1 (3mer ili 5mer HOR), GG694162.1 (2mer HOR) i GG695755.1 (2mer HOR). Također pojedine HOR strukture mogu se sastojati od nekoliko tipova monomera sličnih duljina i jedne duže kao

što je to slučaj sa strukturama u DS497953.1 (4mer HOR), CM000279.1 (dva 4mer HOR), CM000277.2 (4 mer HOR), GG695869.1(3mer HOR) i GG694292.1(3mer ili 5mer HOR).

U cijelom pregledanom genomu Tcas 3.0 pomoću računalne GRM metode, pronađene HOR strukture čine 0,08% ukupnog genoma (u odnosu na ukupni broj baza u svim pregledanim komponentama), dok je taj postotak za ljudski prvi kromosom 0,18%. Prema prosječnim divergencijama HOR-ova kod čovjeka (monomer duljine ~171 bp) i *T. castaneum* (TCAST satelit ~360 bp), gdje je ona u *T.castaneum* dva puta veća od one u čovjeka, može se zaključiti da su se HOR strukture u *T.castaneum* pojavile ranije na evolucijskoj skali u odnosu na one u čovjeka, premda se upravo za njih smatralo da su karakteristične za čovjeka i primate te da su nedavno nastali.

Ovaj rad može poslužiti kao početna točka za određivanje da li se kod tih periodičnosti višega reda u kukcu *T.castaneum* radi o kontinuiranom razvoju ili evolucijskom skoku, odnosno da li one predstavljaju vrh evolucijskog razvoja ili postoji neki lokalni maksimum. Za odgovore na ta pitanja potrebne su daljnje sustavne analize genoma blisko povezanih vrsta u potrazi za periodičnostima višega reda.

6. Dodatak

Tablica 1. Nazivi i duljine komponenata. a) Nepovezane višestruko – komponentne povezane kontige DS47665 - DS497969, b) nepoznate jednostruke spojene kontige GG694051–GG695897.

a)

Br.	Komponenta	Duljina (bp)
1.	DS497665.1	1175385
2.	DS497666.1	324118
3.	DS497667.1	179620
4.	DS497668.1	218890
5.	DS497669.1	327011
6.	DS497670.1	150934
7.	DS497671.1	207115
8.	DS497672.1	252662
9.	DS497673.1	635691
10.	DS497674.1	329991
11.	DS497675.1	145421
12.	DS497676.1	281566
13.	DS497677.1	298863
14.	DS497678.1	233905
15.	DS497679.1	205575
16.	DS497680.1	143519
17.	DS497681.1	252324
18.	DS497682.1	179222
19.	DS497683.1	163124
20.	DS497684.1	239692
21.	DS497685.1	270706
22.	DS497686.1	107027
23.	DS497687.1	209029
24.	DS497688.1	315570
25.	DS497689.1	241962
26.	DS497690.1	113643
27.	DS497691.1	118857
28.	DS497692.1	242491
29.	DS497693.1	72172
30.	DS497694.1	260580
31.	DS497695.1	68328
32.	DS497696.1	175539
33.	DS497697.1	265344
34.	DS497698.1	184673
35.	DS497699.1	199664
36.	DS497700.1	167053
37.	DS497701.1	137173
38.	DS497702.1	159174
39.	DS497703.1	261922
40.	DS497704.1	258350
41.	DS497705.1	215799
42.	DS497706.1	92541
43.	DS497707.1	91379
44.	DS497708.1	104333
45.	DS497709.1	85306
46.	DS497710.1	83343
47.	DS497711.1	324113
48.	DS497712.1	129677
49.	DS497713.1	39916
50.	DS497714.1	156511
51.	DS497715.1	185203
52.	DS497716.1	117267
53.	DS497717.1	48660
54.	DS497718.1	95181
55.	DS497719.1	151068
56.	DS497720.1	72142
57.	DS497721.1	148858
58.	DS497722.1	71516
59.	DS497723.1	251196
60.	DS497724.1	150287
61.	DS497725.1	156722
62.	DS497726.1	57680
63.	DS497727.1	87717
64.	DS497728.1	72281
65.	DS497729.1	20563
66.	DS497730.1	28811
67.	DS497731.1	186310
68.	DS497732.1	31684
69.	DS497733.1	155158
70.	DS497734.1	68148
71.	DS497735.1	24257
72.	DS497736.1	54391
73.	DS497737.1	101176
74.	DS497738.1	64617
75.	DS497739.1	21177
76.	DS497740.1	51733
77.	DS497741.1	65225
78.	DS497742.1	36186
79.	DS497743.1	15147
80.	DS497744.1	115727
81.	DS497745.1	54736
82.	DS497746.1	112887
83.	DS497747.1	16044
84.	DS497748.1	19271
85.	DS497749.1	16749
86.	DS497750.1	25259
87.	DS497751.1	61914
88.	DS497752.1	39386
89.	DS497753.1	33741
90.	DS497754.1	28505
91.	DS497755.1	42240
92.	DS497756.1	166251
93.	DS497757.1	28282
94.	DS497758.1	10717
95.	DS497759.1	16141
96.	DS497760.1	41942
97.	DS497761.1	18441
98.	DS497762.1	18940
99.	DS497763.1	22076
100.	DS497764.1	54238
101.	DS497765.1	45171
102.	DS497766.1	18619
103.	DS497767.1	22684
104.	DS497768.1	16756
105.	DS497769.1	35178
106.	DS497770.1	34632
107.	DS497771.1	118057
108.	DS497772.1	15018
109.	DS497773.1	60029
110.	DS497774.1	13400
111.	DS497775.1	13946
112.	DS497776.1	10304
113.	DS497777.1	34571
114.	DS497778.1	20790
115.	DS497779.1	140909
116.	DS497780.1	17992
117.	DS497781.1	28592
118.	DS497782.1	101378
119.	DS497783.1	21807
120.	DS497784.1	25297
121.	DS497785.1	13532
122.	DS497786.1	15545
123.	DS497787.1	26581
124.	DS497788.1	25901
125.	DS497789.1	12496
126.	DS497790.1	20916
127.	DS497791.1	19111
128.	DS497792.1	16630
129.	DS497793.1	57899
130.	DS497794.1	11059
131.	DS497795.1	24150
132.	DS497796.1	15323
133.	DS497797.1	39987
134.	DS497798.1	12372
135.	DS497799.1	54804
136.	DS497800.1	14030
137.	DS497801.1	6495
138.	DS497802.1	28921
139.	DS497803.1	9827
140.	DS497804.1	29226
141.	DS497805.1	8673
142.	DS497806.1	68640
143.	DS497807.1	10296
144.	DS497808.1	9432
145.	DS497809.1	12723
146.	DS497810.1	21238
147.	DS497811.1	12191
148.	DS497812.1	9146
149.	DS497813.1	14882
150.	DS497814.1	13565
151.	DS497815.1	33323
152.	DS497816.1	13542
153.	DS497817.1	27874
154.	DS497818.1	17233
155.	DS497819.1	65089
156.	DS497820.1	130039
157.	DS497821.1	9763
158.	DS497822.1	31461
159.	DS497823.1	10460
160.	DS497824.1	29642
161.	DS497825.1	12573
162.	DS497826.1	11368
163.	DS497827.1	8494
164.	DS497828.1	16879
165.	DS497829.1	11890
166.	DS497830.1	27641
167.	DS497831.1	13056
168.	DS497832.1	13638
169.	DS497833.1	15245
170.	DS497834.1	9367
171.	DS497835.1	10559
172.	DS497836.1	18640
173.	DS497837.1	18300
174.	DS497838.1	7526
175.	DS497839.1	10089
176.	DS497840.1	17763
177.	DS497841.1	19147
178.	DS497842.1	18828
179.	DS497843.1	5497
180.	DS497844.1	6289
181.	DS497845.1	7096
182.	DS497846.1	67896
183.	DS497847.1	6092
184.	DS497848.1	12449
185.	DS497849.1	7300
186.	DS497850.1	5005
187.	DS497851.1	4966
188.	DS497852.1	6422
189.	DS497853.1	7157
190.	DS497854.1	12260
191.	DS497855.1	7326
192.	DS497856.1	7366
193.	DS497857.1	22617
194.	DS497858.1	6095
195.	DS497859.1	11625
196.	DS497860.1	5595
197.	DS497861.1	4963
198.	DS497862.1	17193
199.	DS497863.1	5706
200.	DS497864.1	10004

201.	DS497865.1	8072
202.	DS497866.1	8591
203.	DS497867.1	8454
204.	DS497868.1	11825
205.	DS497869.1	5801
206.	DS497870.1	45434
207.	DS497871.1	132576
208.	DS497872.1	5129
209.	DS497873.1	22996
210.	DS497874.1	55563
211.	DS497875.1	9930
212.	DS497876.1	16592
213.	DS497877.1	5523
214.	DS497878.1	16665
215.	DS497879.1	22226
216.	DS497880.1	5727
217.	DS497881.1	6251
218.	DS497882.1	5506
219.	DS497883.1	6803
220.	DS497884.1	7412
221.	DS497885.1	8964
222.	DS497886.1	13009
223.	DS497887.1	12763
224.	DS497888.1	19325
225.	DS497889.1	38459
226.	DS497890.1	12159
227.	DS497891.1	13093
228.	DS497892.1	5069
229.	DS497893.1	6499
230.	DS497894.1	16808
231.	DS497895.1	7289
232.	DS497896.1	12849
233.	DS497897.1	26849
234.	DS497898.1	10529
235.	DS497899.1	27791
236.	DS497900.1	7540
237.	DS497901.1	40903
238.	DS497902.1	7141
239.	DS497903.1	7948
240.	DS497904.1	13007
241.	DS497905.1	6846
242.	DS497906.1	7241
243.	DS497907.1	5515
244.	DS497908.1	14114
245.	DS497909.1	5510
246.	DS497910.1	19548
247.	DS497911.1	6277
248.	DS497912.1	7822
249.	DS497913.1	8476
250.	DS497914.1	15533
251.	DS497915.1	11057
252.	DS497916.1	6252
253.	DS497917.1	7125
254.	DS497918.1	5211
255.	DS497919.1	5732
256.	DS497920.1	130058
257.	DS497921.1	5737
258.	DS497922.1	4918
259.	DS497923.1	15518
260.	DS497924.1	5597
261.	DS497925.1	4796
262.	DS497926.1	21235
263.	DS497927.1	15907
264.	DS497928.1	9475
265.	DS497929.1	5572
266.	DS497930.1	9125
267.	DS497931.1	5846
268.	DS497932.1	6133
269.	DS497933.1	8509
270.	DS497934.1	12125
271.	DS497935.1	12197
272.	DS497936.1	5818
273.	DS497937.1	5909
274.	DS497938.1	25138
275.	DS497939.1	50964
276.	DS497940.1	7206
277.	DS497941.1	5740
278.	DS497942.1	5816
279.	DS497943.1	16023
280.	DS497944.1	10854
281.	DS497945.1	51138
282.	DS497946.1	17100
283.	DS497947.1	6483
284.	DS497948.1	11273
285.	DS497949.1	5340
286.	DS497950.1	5611
287.	DS497951.1	36830
288.	DS497952.1	14761
289.	DS497953.1	15078
290.	DS497954.1	9299
291.	DS497955.1	29442
292.	DS497956.1	6388
293.	DS497957.1	5295
294.	DS497958.1	20439
295.	DS497959.1	9224
296.	DS497960.1	12635
297.	DS497961.1	22538
298.	DS497962.1	8595
299.	DS497963.1	18488
300.	DS497964.1	8141
301.	DS497965.1	11177
302.	DS497966.1	12055
303.	DS497967.1	12350
304.	DS497968.1	21413
305.	DS497969.1	11368
306.	DS497665.1	18573726

b)

Br.	Komponenta	Duljina (bp)
1.	GG694051.1	859
2.	GG694052.1	1293
3.	GG694053.1	1738
4.	GG694055.1	1509
5.	GG694057.1	2059
6.	GG694059.1	1794
7.	GG694061.1	1146
8.	GG694063.1	961
9.	GG694065.1	986
10.	GG694067.1	844
11.	GG694069.1	2997
12.	GG694071.1	1079
13.	GG694073.1	641
14.	GG694075.1	1191
15.	GG694076.1	1525
16.	GG694078.1	1230
17.	GG694080.1	931
18.	GG694082.1	1954
19.	GG694084.1	2045
20.	GG694086.1	1734
21.	GG694087.1	2633
22.	GG694089.1	884
23.	GG694090.1	1833
24.	GG694092.1	1675
25.	GG694094.1	1675
26.	GG694096.1	1043
27.	GG694097.1	1339
28.	GG694098.1	1631
29.	GG694100.1	1228
30.	GG694102.1	1496
31.	GG694103.1	2630
32.	GG694105.1	905
33.	GG694107.1	1468
34.	GG694108.1	1086
35.	GG694110.1	1150
36.	GG694112.1	970
37.	GG694114.1	1242
38.	GG694116.1	1409
39.	GG694117.1	1013
40.	GG694119.1	1423
41.	GG694121.1	975
42.	GG694123.1	1576
43.	GG694124.1	1599
44.	GG694126.1	1143
45.	GG694127.1	1225
46.	GG694128.1	1978
47.	GG694130.1	967
48.	GG694132.1	1897
49.	GG694134.1	2310
50.	GG694136.1	1310
51.	GG694138.1	1224
52.	GG694140.1	5620
53.	GG694141.1	1441
54.	GG694143.1	1169
55.	GG694145.1	1621
56.	GG694146.1	1875
57.	GG694148.1	932
58.	GG694149.1	10003
59.	GG694151.1	903
60.	GG694153.1	1652
61.	GG694155.1	2343
62.	GG694157.1	1600
63.	GG694159.1	1136
64.	GG694161.1	4007
65.	GG694162.1	26212
66.	GG694164.1	4440
67.	GG694166.1	898
68.	GG694168.1	1968
69.	GG694169.1	1375
70.	GG694171.1	3853
71.	GG694172.1	1350
72.	GG694174.1	1766
73.	GG694176.1	2377
74.	GG694178.1	9002
75.	GG694180.1	722
76.	GG694182.1	1872
77.	GG694184.1	924
78.	GG694186.1	1019
79.	GG694188.1	888
80.	GG694190.1	933
81.	GG694192.1	918
82.	GG694194.1	1402
83.	GG694196.1	974
84.	GG694198.1	2341
85.	GG694199.1	1172
86.	GG694201.1	4379
87.	GG694202.1	12705
88.	GG694203.1	16642
89.	GG694204.1	1792
90.	GG694205.1	870
91.	GG694206.1	4095
92.	GG694208.1	832
93.	GG694209.1	1985
94.	GG694210.1	2659
95.	GG694212.1	490
96.	GG694214.1	713
97.	GG694216.1	2271
98.	GG694218.1	917
99.	GG694219.1	1217
100.	GG694221.1	1267
101.	GG694223.1	3547
102.	GG694225.1	1452
103.	GG694237.1	13175
104.	GG694238.1	29772
105.	GG694242.1	1248
106.	GG694249.1	17940
107.	GG694251.1	7504
108.	GG694257.1	19595
109.	GG694262.1	7446
110.	GG694263.1	1047
111.	GG694268.1	10044
112.	GG694293.1	8834
113.	GG694301.1	2340

114.	GG694303.1	887
115.	GG694304.1	831
116.	GG694309.1	1861
117.	GG694310.1	1333
118.	GG694311.1	3971
119.	GG694312.1	6727
120.	GG694318.1	9416
121.	GG694324.1	16229
122.	GG694325.1	2579
123.	GG694335.1	14419
124.	GG694340.1	1105
125.	GG694343.1	878
126.	GG694347.1	1363
127.	GG694348.1	17920
128.	GG694351.1	2634
129.	GG694353.1	6783
130.	GG694356.1	789
131.	GG694368.1	2844
132.	GG694384.1	1389
133.	GG694385.1	1540
134.	GG694394.1	1336
135.	GG694399.1	1306
136.	GG694404.1	885
137.	GG694415.1	1851
138.	GG694416.1	4916
139.	GG694421.1	2934
140.	GG694422.1	2647
141.	GG694425.1	921
142.	GG694426.1	1855
143.	GG694427.1	1472
144.	GG694433.1	3458
145.	GG694434.1	1604
146.	GG694439.1	833
147.	GG694441.1	6746
148.	GG694449.1	5927
149.	GG694450.1	1296
150.	GG694452.1	1402
151.	GG694456.1	1119
152.	GG694457.1	2788
153.	GG694463.1	937
154.	GG694478.1	1776
155.	GG694481.1	2335
156.	GG694483.1	3408
157.	GG694486.1	980
158.	GG694508.1	1495
159.	GG694510.1	1909
160.	GG694519.1	3316
161.	GG694525.1	11910
162.	GG694529.1	884
163.	GG694590.1	1289
164.	GG694591.1	1658
165.	GG694600.1	1236
166.	GG694613.1	2145
167.	GG694618.1	1326
168.	GG694632.1	1998
169.	GG694635.1	816
170.	GG694646.1	1226
171.	GG694656.1	2400
172.	GG694661.1	782
173.	GG694662.1	2091
174.	GG694664.1	1104
175.	GG694667.1	1726
176.	GG694669.1	11045
177.	GG694672.1	1751
178.	GG694673.1	1188
179.	GG694676.1	3784
180.	GG694696.1	2137
181.	GG694697.1	1339
182.	GG694700.1	11035
183.	GG694701.1	13021
184.	GG694706.1	3015
185.	GG694709.1	1247
186.	GG694711.1	35221
187.	GG694716.1	5882
188.	GG694719.1	3099
189.	GG694722.1	779
190.	GG694732.1	1487
191.	GG694744.1	743
192.	GG694745.1	3724
193.	GG694748.1	957
194.	GG694750.1	952
195.	GG694753.1	1885
196.	GG694755.1	1996
197.	GG694757.1	1290
198.	GG694764.1	1645
199.	GG694770.1	1444
200.	GG694773.1	5983
201.	GG694774.1	6235
202.	GG694776.1	29903
203.	GG694777.1	35899
204.	GG694783.1	4837
205.	GG694786.1	5296
206.	GG694788.1	1312
207.	GG694793.1	10456
208.	GG694810.1	5982
209.	GG694818.1	1710
210.	GG694821.1	1386
211.	GG694822.1	969
212.	GG694825.1	1304
213.	GG694828.1	7712
214.	GG694832.1	9145
215.	GG694842.1	1079
216.	GG694849.1	1306
217.	GG694850.1	1173
218.	GG694851.1	4420
219.	GG694853.1	4036
220.	GG694854.1	16214
221.	GG694859.1	2255
222.	GG694860.1	1953
223.	GG694872.1	1306
224.	GG694874.1	15063
225.	GG694877.1	1100
226.	GG694884.1	2118
227.	GG694887.1	9111
228.	GG694892.1	905
229.	GG694899.1	7347
230.	GG694913.1	1315
231.	GG694915.1	2237
232.	GG694927.1	2384
233.	GG694932.1	1888
234.	GG694933.1	1515
235.	GG694939.1	1664
236.	GG694942.1	1939
237.	GG694943.1	1244
238.	GG694946.1	1132
239.	GG694959.1	8032
240.	GG694962.1	4792
241.	GG694963.1	21889
242.	GG694966.1	18156
243.	GG694971.1	3002
244.	GG694980.1	2828
245.	GG695003.1	1126
246.	GG695006.1	975
247.	GG695007.1	2112
248.	GG695024.1	1705
249.	GG695042.1	2107
250.	GG695044.1	2368
251.	GG695050.1	15659
252.	GG695053.1	2396
253.	GG695059.1	2026
254.	GG695063.1	6590
255.	GG695065.1	9096
256.	GG695072.1	1578
257.	GG695075.1	7042
258.	GG695084.1	1512
259.	GG695095.1	732
260.	GG695107.1	2048
261.	GG695114.1	10167
262.	GG695118.1	1366
263.	GG695123.1	2756
264.	GG695126.1	2102
265.	GG695140.1	3095
266.	GG695144.1	1058
267.	GG695148.1	1536
268.	GG695151.1	1763
269.	GG695159.1	903
270.	GG695162.1	832
271.	GG695171.1	844
272.	GG695174.1	15132
273.	GG695176.1	1236
274.	GG695184.1	1278
275.	GG695192.1	898
276.	GG695195.1	1302
277.	GG695199.1	1657
278.	GG695211.1	1373
279.	GG695232.1	1415
280.	GG695246.1	2482
281.	GG695248.1	13776
282.	GG695261.1	1558
283.	GG695262.1	1823
284.	GG695269.1	1800
285.	GG695272.1	32629
286.	GG695280.1	1622
287.	GG695282.1	5220
288.	GG695289.1	41267
289.	GG695290.1	909
290.	GG695305.1	1593
291.	GG695320.1	1212
292.	GG695328.1	1408
293.	GG695330.1	869
294.	GG695332.1	2112
295.	GG695334.1	4512
296.	GG695336.1	2198
297.	GG695341.1	4480
298.	GG695350.1	4984
299.	GG695353.1	1107
300.	GG695365.1	6015
301.	GG695366.1	6457
302.	GG695368.1	1026
303.	GG695371.1	1234
304.	GG695376.1	4640
305.	GG695379.1	9377
306.	GG695381.1	5650
307.	GG695382.1	21755
308.	GG695383.1	11065
309.	GG695385.1	20802
310.	GG695386.1	5018
311.	GG695396.1	3166
312.	GG695399.1	1817
313.	GG695401.1	4242
314.	GG695402.1	4124
315.	GG695410.1	2839
316.	GG695412.1	1869
317.	GG695417.1	3210
318.	GG695422.1	20436
319.	GG695427.1	16669
320.	GG695437.1	5709
321.	GG695448.1	1474
322.	GG695449.1	15562
323.	GG695458.1	1947
324.	GG695460.1	1177
325.	GG695465.1	6529
326.	GG695473.1	5531
327.	GG695482.1	4021
328.	GG695483.1	1636
329.	GG695487.1	1404
330.	GG695491.1	1797
331.	GG695495.1	1523
332.	GG695507.1	1684
333.	GG695512.1	1922
334.	GG695519.1	3395
335.	GG695522.1	996
336.	GG695530.1	1294
337.	GG695545.1	1789
338.	GG695546.1	1871
339.	GG695547.1	2517
340.	GG695553.1	1061
341.	GG695558.1	1698
342.	GG695559.1	1127
343.	GG695563.1	1444
344.	GG695568.1	3878
345.	GG695573.1	1059
346.	GG695576.1	1962
347.	GG695580.1	1406
348.	GG695587.1	1667
349.	GG695589.1	2594
350.	GG695595.1	1267
351.	GG695597.1	2633
352.	GG695599.1	1024
353.	GG695604.1	1007
354.	GG695619.1	1489
355.	GG695625.1	8744
356.	GG695629.1	1337

357.	GG695634.1	1930
358.	GG695646.1	1237
359.	GG695648.1	1588
360.	GG695657.1	1301
361.	GG695660.1	1455
362.	GG695663.1	1266
363.	GG695670.1	1537
364.	GG695674.1	2125
365.	GG695698.1	1519
366.	GG695701.1	1234
367.	GG695703.1	1055
368.	GG695709.1	3831
369.	GG695712.1	1506
370.	GG695713.1	1097
371.	GG695723.1	1566
372.	GG695724.1	2361
373.	GG695725.1	11296
374.	GG695736.1	1328
375.	GG695738.1	1402
376.	GG695758.1	1590
377.	GG695769.1	1741
378.	GG695773.1	6950
379.	GG695778.1	8039
380.	GG695781.1	1395
381.	GG695787.1	2892
382.	GG695788.1	1240
383.	GG695789.1	839
384.	GG695793.1	5744
385.	GG695816.1	1792
386.	GG695822.1	7395
387.	GG695843.1	1315
388.	GG695855.1	22356
389.	GG695857.1	29871
390.	GG695858.1	1411
391.	GG695859.1	6775
392.	GG695860.1	11735
393.	GG695866.1	5537
394.	GG695872.1	19506
395.	GG695874.1	840
396.	GG695875.1	7082
397.	GG695879.1	1160
398.	GG695883.1	1657
399.	GG695886.1	20473
400.	GG695896.1	1598
401.	GG694053.1	1622177
402.	GG694227.1	1440701
403.	GG694054.1	297
404.	GG694056.1	878
405.	GG694058.1	998
406.	GG694060.1	1485
407.	GG694062.1	1508
408.	GG694064.1	2518
409.	GG694066.1	1338
410.	GG694068.1	994
411.	GG694070.1	1994
412.	GG694072.1	1309
413.	GG694074.1	938
414.	GG694077.1	546
415.	GG694079.1	2609
416.	GG694081.1	814
417.	GG694083.1	1308
418.	GG694085.1	1367
419.	GG694088.1	5832
420.	GG694091.1	2037
421.	GG694093.1	1464
422.	GG694095.1	816
423.	GG694099.1	1165
424.	GG694101.1	1134
425.	GG694104.1	1528
426.	GG694106.1	1454
427.	GG694109.1	1256
428.	GG694111.1	970
429.	GG694113.1	2385
430.	GG694115.1	2001
431.	GG694118.1	1221
432.	GG694120.1	709
433.	GG694122.1	1429
434.	GG694125.1	1078
435.	GG694129.1	2767
436.	GG694131.1	1416
437.	GG694133.1	1560
438.	GG694135.1	959
439.	GG694137.1	2026
440.	GG694139.1	888
441.	GG694142.1	947
442.	GG694144.1	1217
443.	GG694147.1	1802
444.	GG694150.1	5734
445.	GG694152.1	916
446.	GG694154.1	976
447.	GG694156.1	890
448.	GG694158.1	945
449.	GG694160.1	3145
450.	GG694163.1	1247
451.	GG694165.1	1162
452.	GG694167.1	788
453.	GG694170.1	1606
454.	GG694173.1	1657
455.	GG694175.1	1541
456.	GG694177.1	1010
457.	GG694179.1	2239
458.	GG694181.1	1288
459.	GG694183.1	1198
460.	GG694185.1	2324
461.	GG694187.1	6042
462.	GG694189.1	857
463.	GG694191.1	1656
464.	GG694193.1	864
465.	GG694195.1	9544
466.	GG694197.1	977
467.	GG694200.1	1887
468.	GG694207.1	883
469.	GG694211.1	683
470.	GG694213.1	922
471.	GG694215.1	1175
472.	GG694217.1	822
473.	GG694220.1	1183
474.	GG694222.1	1924
475.	GG694224.1	811
476.	GG694226.1	1367
477.	GG694227.1	1362
478.	GG694228.1	1490
479.	GG694229.1	929
480.	GG694230.1	1465
481.	GG694231.1	777
482.	GG694232.1	1164
483.	GG694233.1	2792
484.	GG694234.1	3556
485.	GG694235.1	926
486.	GG694236.1	822
487.	GG694239.1	1659
488.	GG694240.1	851
489.	GG694241.1	764
490.	GG694243.1	5917
491.	GG694244.1	8361
492.	GG694245.1	12328
493.	GG694246.1	1556
494.	GG694247.1	802
495.	GG694248.1	1658
496.	GG694250.1	1052
497.	GG694252.1	939
498.	GG694253.1	750
499.	GG694255.1	1246
500.	GG694256.1	2423
501.	GG694258.1	2016
502.	GG694259.1	1712
503.	GG694260.1	1337
504.	GG694261.1	1448
505.	GG694264.1	834
506.	GG694265.1	1158
507.	GG694266.1	925
508.	GG694267.1	2064
509.	GG694269.1	820
510.	GG694270.1	1231
511.	GG694271.1	2092
512.	GG694272.1	2733
513.	GG694273.1	964
514.	GG694274.1	1263
515.	GG694275.1	3275
516.	GG694276.1	868
517.	GG694277.1	4806
518.	GG694278.1	4336
519.	GG694279.1	890
520.	GG694280.1	1147
521.	GG694281.1	2468
522.	GG694282.1	689
523.	GG694283.1	1110
524.	GG694284.1	9690
525.	GG694285.1	803
526.	GG694286.1	1665
527.	GG694287.1	1357
528.	GG694288.1	1400
529.	GG694289.1	2968
530.	GG694290.1	983
531.	GG694291.1	849
532.	GG694292.1	12643
533.	GG694294.1	930
534.	GG694295.1	1062
535.	GG694296.1	913
536.	GG694297.1	925
537.	GG694298.1	901
538.	GG694299.1	15899
539.	GG694300.1	885
540.	GG694302.1	1059
541.	GG694305.1	784
542.	GG694306.1	1250
543.	GG694307.1	1286
544.	GG694308.1	9799
545.	GG694313.1	943
546.	GG694314.1	1551
547.	GG694315.1	1906
548.	GG694316.1	1553
549.	GG694317.1	2760
550.	GG694319.1	1185
551.	GG694320.1	1113
552.	GG694321.1	2293
553.	GG694322.1	2462
554.	GG694323.1	2863
555.	GG694326.1	861
556.	GG694327.1	1404
557.	GG694328.1	1464
558.	GG694329.1	2052
559.	GG694330.1	670
560.	GG694331.1	2127
561.	GG694332.1	7622
562.	GG694333.1	862
563.	GG694334.1	3895
564.	GG694337.1	1371
565.	GG694339.1	1051
566.	GG694342.1	2560
567.	GG694345.1	969
568.	GG694349.1	5745
569.	GG694352.1	7213
570.	GG694355.1	1254
571.	GG694358.1	976
572.	GG694360.1	3355
573.	GG694362.1	769
574.	GG694364.1	1741
575.	GG694366.1	930
576.	GG694369.1	6484
577.	GG694371.1	1225
578.	GG694373.1	1930
579.	GG694375.1	819
580.	GG694377.1	1637
581.	GG694379.1	1162
582.	GG694381.1	1565
583.	GG694383.1	1214
584.	GG694387.1	1654
585.	GG694389.1	660
586.	GG694391.1	1280
587.	GG694393.1	1800
588.	GG694396.1	2281
589.	GG694398.1	841
590.	GG694401.1	821
591.	GG694403.1	1689
592.	GG694406.1	897
593.	GG694408.1	1144
594.	GG694410.1	1668
595.	GG694412.1	1164
596.	GG694414.1	878
597.	GG694418.1	1031
598.	GG694420.1	1496
599.	GG694424.1	1536

600.	GG694429.1	1418
601.	GG694431.1	1775
602.	GG694435.1	892
603.	GG694437.1	859
604.	GG694440.1	932
605.	GG694443.1	511
606.	GG694445.1	2505
607.	GG694447.1	997
608.	GG694451.1	1882
609.	GG694454.1	19173
610.	GG694458.1	1332
611.	GG694460.1	1583
612.	GG694462.1	2237
613.	GG694465.1	1303
614.	GG694467.1	961
615.	GG694469.1	848
616.	GG694471.1	1199
617.	GG694473.1	1160
618.	GG694475.1	2964
619.	GG694477.1	1218
620.	GG694480.1	1656
621.	GG694484.1	1766
622.	GG694487.1	1165
623.	GG694489.1	1632
624.	GG694491.1	692
625.	GG694493.1	3989
626.	GG694495.1	1469
627.	GG694497.1	2213
628.	GG694499.1	802
629.	GG694501.1	838
630.	GG694503.1	1143
631.	GG694505.1	1305
632.	GG694507.1	1335
633.	GG694511.1	1932
634.	GG694513.1	1591
635.	GG694515.1	972
636.	GG694517.1	1473
637.	GG694520.1	937
638.	GG694522.1	1015
639.	GG694524.1	1117
640.	GG694527.1	744
641.	GG694530.1	840
642.	GG694532.1	948
643.	GG694534.1	1593
644.	GG694536.1	1788
645.	GG694538.1	888
646.	GG694540.1	1295
647.	GG694542.1	1478
648.	GG694544.1	1744
649.	GG694546.1	1284
650.	GG694548.1	1407
651.	GG694550.1	879
652.	GG694552.1	1788
653.	GG694554.1	2012
654.	GG694556.1	745
655.	GG694558.1	1103
656.	GG694560.1	906
657.	GG694562.1	2186
658.	GG694564.1	1849
659.	GG694566.1	972
660.	GG694568.1	1718
661.	GG694570.1	1950
662.	GG694572.1	1482
663.	GG694574.1	1701
664.	GG694576.1	868
665.	GG694578.1	1805
666.	GG694580.1	980
667.	GG694582.1	844
668.	GG694584.1	1831
669.	GG694586.1	1245
670.	GG694588.1	3095
671.	GG694592.1	3218
672.	GG694594.1	1063
673.	GG694596.1	1400
674.	GG694598.1	1573
675.	GG694601.1	1559
676.	GG694603.1	1395
677.	GG694605.1	1487
678.	GG694607.1	1556
679.	GG694609.1	989
680.	GG694611.1	1630
681.	GG694614.1	1921
682.	GG694616.1	1287
683.	GG694619.1	892
684.	GG694621.1	1125
685.	GG694623.1	1501
686.	GG694625.1	2511
687.	GG694627.1	1164
688.	GG694629.1	2218
689.	GG694631.1	909
690.	GG694634.1	1047
691.	GG694637.1	1868
692.	GG694639.1	1843
693.	GG694641.1	1756
694.	GG694643.1	1778
695.	GG694645.1	985
696.	GG694648.1	601
697.	GG694650.1	881
698.	GG694652.1	1412
699.	GG694654.1	824
700.	GG694659.1	2542
701.	GG694663.1	1477
702.	GG694666.1	1429
703.	GG694670.1	1306
704.	GG694675.1	842
705.	GG694678.1	1764
706.	GG694680.1	892
707.	GG694682.1	1877
708.	GG694684.1	1477
709.	GG694687.1	1283
710.	GG694689.1	2181
711.	GG694691.1	385
712.	GG694693.1	847
713.	GG694695.1	1152
714.	GG694699.1	867
715.	GG694703.1	354
716.	GG694705.1	928
717.	GG694708.1	2055
718.	GG694712.1	1269
719.	GG694714.1	4127
720.	GG694717.1	772
721.	GG694720.1	2249
722.	GG694723.1	1422
723.	GG694725.1	932
724.	GG694727.1	1216
725.	GG694729.1	1512
726.	GG694731.1	915
727.	GG694734.1	1774
728.	GG694736.1	761
729.	GG694738.1	844
730.	GG694740.1	878
731.	GG694742.1	1728
732.	GG694746.1	1509
733.	GG694749.1	280
734.	GG694752.1	2996
735.	GG694756.1	1678
736.	GG694759.1	1143
737.	GG694761.1	849
738.	GG694763.1	1764
739.	GG694766.1	3174
740.	GG694768.1	1235
741.	GG694771.1	925
742.	GG694775.1	1099
743.	GG694779.1	1336
744.	GG694781.1	669
745.	GG694784.1	1197
746.	GG694787.1	882
747.	GG694790.1	4274
748.	GG694792.1	900
749.	GG694795.1	1320
750.	GG694797.1	1298
751.	GG694799.1	1847
752.	GG694801.1	8365
753.	GG694803.1	2035
754.	GG694805.1	1876
755.	GG694807.1	979
756.	GG694809.1	732
757.	GG694812.1	2746
758.	GG694814.1	932
759.	GG694816.1	1424
760.	GG694819.1	719
761.	GG694823.1	1189
762.	GG694826.1	1866
763.	GG694829.1	814
764.	GG694831.1	3850
765.	GG694834.1	778
766.	GG694836.1	991
767.	GG694838.1	849
768.	GG694840.1	988
769.	GG694843.1	1660
770.	GG694845.1	1562
771.	GG694847.1	1537
772.	GG694852.1	6247
773.	GG694856.1	743
774.	GG694858.1	865
775.	GG694862.1	718
776.	GG694864.1	858
777.	GG694866.1	2492
778.	GG694868.1	933
779.	GG694870.1	1465
780.	GG694873.1	1197
781.	GG694876.1	1470
782.	GG694879.1	1340
783.	GG694881.1	6096
784.	GG694883.1	1002
785.	GG694886.1	3767
786.	GG694889.1	888
787.	GG694891.1	673
788.	GG694894.1	1281
789.	GG694896.1	1068
790.	GG694898.1	982
791.	GG694901.1	1061
792.	GG694903.1	3342
793.	GG694905.1	1636
794.	GG694907.1	963
795.	GG694909.1	1596
796.	GG694911.1	1491
797.	GG694914.1	1618
798.	GG694917.1	910
799.	GG694919.1	1376
800.	GG694921.1	1217
801.	GG694923.1	876
802.	GG694925.1	1427
803.	GG694928.1	983
804.	GG694930.1	1060
805.	GG694934.1	1468
806.	GG694936.1	921
807.	GG694938.1	2931
808.	GG694941.1	3707
809.	GG694945.1	1357
810.	GG694948.1	1598
811.	GG694950.1	925
812.	GG694952.1	2716
813.	GG694954.1	9790
814.	GG694956.1	1439
815.	GG694958.1	793
816.	GG694961.1	1146
817.	GG694965.1	804
818.	GG694968.1	1057
819.	GG694970.1	1314
820.	GG694973.1	928
821.	GG694975.1	2416
822.	GG694977.1	1137
823.	GG694979.1	1398
824.	GG694982.1	1971
825.	GG694984.1	1330
826.	GG694986.1	3635
827.	GG694988.1	840
828.	GG694990.1	1164
829.	GG694992.1	1206
830.	GG694994.1	2810
831.	GG694996.1	912
832.	GG694998.1	1027
833.	GG695000.1	586
834.	GG695002.1	877
835.	GG695005.1	2112
836.	GG695009.1	892
837.	GG695011.1	1135
838.	GG695013.1	1074
839.	GG695015.1	6130
840.	GG695017.1	1428
841.	GG695019.1	953
842.	GG695021.1	1086

843.	GG695023.1	851
844.	GG695026.1	2200
845.	GG695028.1	913
846.	GG695030.1	1940
847.	GG695032.1	881
848.	GG695034.1	1533
849.	GG695036.1	1020
850.	GG695038.1	1793
851.	GG695040.1	1540
852.	GG695043.1	1369
853.	GG695046.1	1786
854.	GG695048.1	1330
855.	GG695051.1	8493
856.	GG695054.1	920
857.	GG695056.1	1605
858.	GG695058.1	3827
859.	GG695061.1	1508
860.	GG695064.1	1147
861.	GG695067.1	1387
862.	GG695069.1	889
863.	GG695071.1	727
864.	GG695074.1	1318
865.	GG695077.1	806
866.	GG695079.1	2077
867.	GG695081.1	728
868.	GG695083.1	2402
869.	GG695086.1	1929
870.	GG695088.1	964
871.	GG695090.1	1485
872.	GG695092.1	1640
873.	GG695094.1	857
874.	GG695097.1	2526
875.	GG695099.1	692
876.	GG695101.1	13357
877.	GG695103.1	1639
878.	GG695105.1	4635
879.	GG695108.1	674
880.	GG695110.1	1316
881.	GG695112.1	995
882.	GG695115.1	1020
883.	GG695117.1	2380
884.	GG695120.1	995
885.	GG695122.1	1272
886.	GG695125.1	1310
887.	GG695128.1	1150
888.	GG695130.1	1005
889.	GG695132.1	1277
890.	GG695134.1	817
891.	GG695136.1	981
892.	GG695138.1	1007
893.	GG695141.1	2393
894.	GG695143.1	2316
895.	GG695146.1	1714
896.	GG695149.1	1018
897.	GG695152.1	1196
898.	GG695154.1	855
899.	GG695156.1	721
900.	GG695158.1	972
901.	GG695161.1	1942
902.	GG695164.1	1591
903.	GG695166.1	6353
904.	GG695168.1	1112
905.	GG695170.1	762
906.	GG695173.1	1678
907.	GG695177.1	944
908.	GG695179.1	1670
909.	GG695181.1	1118
910.	GG695183.1	598
911.	GG695186.1	1588
912.	GG695188.1	1378
913.	GG695190.1	926
914.	GG695193.1	896
915.	GG695196.1	1022
916.	GG695198.1	2319
917.	GG695201.1	2899
918.	GG695203.1	1501
919.	GG695205.1	898
920.	GG695207.1	1368
921.	GG695209.1	1620
922.	GG695212.1	2207
923.	GG695214.1	875
924.	GG695216.1	2219
925.	GG695218.1	1526
926.	GG695220.1	1062
927.	GG695222.1	3049
928.	GG695224.1	12777
929.	GG695226.1	731
930.	GG695228.1	1230
931.	GG695230.1	1240
932.	GG695233.1	1947
933.	GG695235.1	2104
934.	GG695237.1	937
935.	GG695239.1	930
936.	GG695241.1	971
937.	GG695243.1	942
938.	GG695245.1	1617
939.	GG695249.1	1344
940.	GG695251.1	1457
941.	GG695253.1	866
942.	GG695255.1	2561
943.	GG695257.1	1238
944.	GG695259.1	1098
945.	GG695263.1	1008
946.	GG695265.1	5550
947.	GG695267.1	1303
948.	GG695270.1	920
949.	GG695273.1	1075
950.	GG695275.1	1508
951.	GG695277.1	1632
952.	GG695279.1	882
953.	GG695283.1	748
954.	GG695285.1	1097
955.	GG695287.1	1498
956.	GG695291.1	993
957.	GG695293.1	1010
958.	GG695295.1	943
959.	GG695297.1	347
960.	GG695299.1	896
961.	GG695301.1	1742
962.	GG695303.1	926
963.	GG695306.1	4333
964.	GG695308.1	1693
965.	GG695310.1	800
966.	GG695312.1	884
967.	GG695314.1	709
968.	GG695316.1	916
969.	GG695318.1	937
970.	GG695321.1	778
971.	GG695323.1	711
972.	GG695325.1	617
973.	GG695327.1	2185
974.	GG695331.1	1203
975.	GG695335.1	1578
976.	GG695338.1	1417
977.	GG695340.1	1475
978.	GG695343.1	1713
979.	GG695345.1	951
980.	GG695347.1	762
981.	GG695349.1	6960
982.	GG695352.1	1545
983.	GG695355.1	1735
984.	GG695357.1	1710
985.	GG695359.1	7169
986.	GG695361.1	723
987.	GG695363.1	896
988.	GG695367.1	1385
989.	GG695370.1	1022
990.	GG695373.1	1288
991.	GG695375.1	2250
992.	GG695378.1	7101
993.	GG695384.1	3759
994.	GG695388.1	761
995.	GG695390.1	985
996.	GG695392.1	946
997.	GG695394.1	2861
998.	GG695397.1	1564
999.	GG695400.1	833
1000.	GG695404.1	2548
1001.	GG695406.1	1540
1002.	GG695408.1	1102
1003.	GG695411.1	931
1004.	GG695414.1	1249
1005.	GG695416.1	822
1006.	GG695419.1	1456
1007.	GG695421.1	576
1008.	GG695424.1	2738
1009.	GG695426.1	826
1010.	GG695429.1	6292
1011.	GG695431.1	27688
1012.	GG695433.1	765
1013.	GG695435.1	901
1014.	GG695438.1	1279
1015.	GG695440.1	752
1016.	GG695442.1	1087
1017.	GG695444.1	954
1018.	GG695446.1	1373
1019.	GG695450.1	1707
1020.	GG695452.1	1710
1021.	GG695454.1	810
1022.	GG695456.1	1170
1023.	GG695459.1	2084
1024.	GG695462.1	1472
1025.	GG695464.1	1026
1026.	GG695467.1	1810
1027.	GG695469.1	7822
1028.	GG695471.1	916
1029.	GG695474.1	2945
1030.	GG695476.1	672
1031.	GG695479.1	881
1032.	GG695481.1	1156
1033.	GG695485.1	940
1034.	GG695488.1	1219
1035.	GG695490.1	1893
1036.	GG695493.1	1322
1037.	GG695496.1	926
1038.	GG695498.1	1700
1039.	GG695500.1	938
1040.	GG695502.1	790
1041.	GG695504.1	893
1042.	GG695506.1	919
1043.	GG695509.1	7529
1044.	GG695511.1	593
1045.	GG695514.1	888
1046.	GG695516.1	1692
1047.	GG695518.1	1409
1048.	GG695521.1	2695
1049.	GG695524.1	1558
1050.	GG695526.1	893
1051.	GG695528.1	861
1052.	GG695531.1	852
1053.	GG695533.1	1292
1054.	GG695535.1	2673
1055.	GG695537.1	1136
1056.	GG695539.1	1482
1057.	GG695541.1	20397
1058.	GG695543.1	1828
1059.	GG695548.1	1557
1060.	GG695550.1	927
1061.	GG695554.1	1752
1062.	GG695556.1	876
1063.	GG695560.1	1198
1064.	GG695562.1	1522
1065.	GG695565.1	891
1066.	GG695567.1	1192
1067.	GG695570.1	897
1068.	GG695572.1	814
1069.	GG695575.1	1386
1070.	GG695578.1	3018
1071.	GG695581.1	1843
1072.	GG695583.1	1422
1073.	GG695585.1	814
1074.	GG695588.1	349
1075.	GG695591.1	1483
1076.	GG695593.1	1396
1077.	GG695596.1	819
1078.	GG695600.1	1340
1079.	GG695602.1	1438
1080.	GG695605.1	1562
1081.	GG695607.1	1567
1082.	GG695609.1	920
1083.	GG695611.1	2270
1084.	GG695613.1	929
1085.	GG695615.1	6772

1086.	GG695617.1	955
1087.	GG695620.1	2494
1088.	GG695622.1	899
1089.	GG695624.1	922
1090.	GG695627.1	892
1091.	GG695630.1	1385
1092.	GG695632.1	4234
1093.	GG695635.1	1287
1094.	GG695637.1	5704
1095.	GG695639.1	2463
1096.	GG695641.1	881
1097.	GG695643.1	735
1098.	GG695645.1	3559
1099.	GG695649.1	2771
1100.	GG695651.1	796
1101.	GG695653.1	2322
1102.	GG695655.1	1081
1103.	GG695658.1	1713
1104.	GG695661.1	740
1105.	GG695664.1	7937
1106.	GG695666.1	1594
1107.	GG695668.1	2207
1108.	GG695671.1	1730
1109.	GG695673.1	925
1110.	GG695676.1	1079
1111.	GG695678.1	815
1112.	GG695680.1	1363
1113.	GG695682.1	821
1114.	GG695684.1	1146
1115.	GG695686.1	936
1116.	GG695688.1	1080
1117.	GG695690.1	1194
1118.	GG695692.1	1647
1119.	GG695694.1	1899
1120.	GG695696.1	2126
1121.	GG695699.1	1323
1122.	GG695702.1	962
1123.	GG695705.1	6604
1124.	GG695707.1	776
1125.	GG695710.1	1020
1126.	GG695714.1	2168
1127.	GG695716.1	842
1128.	GG695718.1	1343
1129.	GG695720.1	1045
1130.	GG695722.1	966
1131.	GG695727.1	1349
1132.	GG695729.1	994
1133.	GG695731.1	1160
1134.	GG695733.1	876
1135.	GG695735.1	1210
1136.	GG695739.1	911
1137.	GG695741.1	1824
1138.	GG695743.1	1484
1139.	GG695745.1	1078
1140.	GG695747.1	1659
1141.	GG695749.1	2106
1142.	GG695751.1	1498
1143.	GG695753.1	1318
1144.	GG695755.1	4120
1145.	GG695757.1	2073
1146.	GG695760.1	1566
1147.	GG695762.1	925
1148.	GG695764.1	1485
1149.	GG695766.1	1167
1150.	GG695768.1	571
1151.	GG695771.1	2742
1152.	GG695774.1	1790
1153.	GG695776.1	788
1154.	GG695779.1	1479
1155.	GG695782.1	2362
1156.	GG695784.1	1703
1157.	GG695786.1	1069
1158.	GG695791.1	893
1159.	GG695794.1	1315
1160.	GG695796.1	6438
1161.	GG695798.1	783
1162.	GG695800.1	790
1163.	GG695802.1	1830
1164.	GG695804.1	868
1165.	GG695806.1	2007
1166.	GG695808.1	992
1167.	GG695810.1	1160
1168.	GG695812.1	1753
1169.	GG695815.1	6213
1170.	GG695818.1	1145
1171.	GG695820.1	2108
1172.	GG695823.1	1446
1173.	GG695825.1	1459
1174.	GG695827.1	889
1175.	GG695829.1	1090
1176.	GG695831.1	879
1177.	GG695833.1	1289
1178.	GG695835.1	2039
1179.	GG695837.1	1143
1180.	GG695839.1	2007
1181.	GG695841.1	1028
1182.	GG695844.1	782
1183.	GG695846.1	945
1184.	GG695848.1	1840
1185.	GG695850.1	2592
1186.	GG695852.1	1236
1187.	GG695854.1	1313
1188.	GG695861.1	1036
1189.	GG695863.1	883
1190.	GG695865.1	1416
1191.	GG695868.1	1687
1192.	GG695870.1	1868
1193.	GG695873.1	825
1194.	GG695877.1	4129
1195.	GG695880.1	1953
1196.	GG695882.1	2115
1197.	GG695885.1	1467
1198.	GG695888.1	1568
1199.	GG695890.1	4132
1200.	GG695892.1	2182
1201.	GG695894.1	2106
1202.	GG695897.1	1023
1203.	GG694336.1	1117160
1204.	GG694336.1	6965
1205.	GG694338.1	1185
1206.	GG694341.1	895
1207.	GG694344.1	2374
1208.	GG694346.1	1621
1209.	GG694350.1	3104
1210.	GG694354.1	854
1211.	GG694357.1	1186
1212.	GG694359.1	2057
1213.	GG694361.1	1754
1214.	GG694363.1	1332
1215.	GG694365.1	999
1216.	GG694367.1	2847
1217.	GG694370.1	509
1218.	GG694372.1	2393
1219.	GG694374.1	2275
1220.	GG694376.1	3036
1221.	GG694378.1	794
1222.	GG694380.1	910
1223.	GG694382.1	1720
1224.	GG694386.1	4017
1225.	GG694388.1	2472
1226.	GG694390.1	7551
1227.	GG694392.1	856
1228.	GG694395.1	3922
1229.	GG694397.1	731
1230.	GG694400.1	714
1231.	GG694402.1	981
1232.	GG694405.1	941
1233.	GG694407.1	1003
1234.	GG694409.1	1197
1235.	GG694411.1	1976
1236.	GG694413.1	779
1237.	GG694417.1	975
1238.	GG694419.1	1416
1239.	GG694423.1	1281
1240.	GG694428.1	1489
1241.	GG694430.1	1970
1242.	GG694432.1	2841
1243.	GG694436.1	884
1244.	GG694438.1	897
1245.	GG694442.1	441
1246.	GG694444.1	903
1247.	GG694446.1	1870
1248.	GG694448.1	2054
1249.	GG694453.1	2898
1250.	GG694455.1	769
1251.	GG694459.1	2086
1252.	GG694461.1	7398
1253.	GG694464.1	975
1254.	GG694466.1	861
1255.	GG694468.1	1585
1256.	GG694470.1	1296
1257.	GG694472.1	1371
1258.	GG694474.1	2867
1259.	GG694476.1	914
1260.	GG694479.1	488
1261.	GG694482.1	894
1262.	GG694485.1	2099
1263.	GG694488.1	1353
1264.	GG694490.1	1642
1265.	GG694492.1	1506
1266.	GG694494.1	2040
1267.	GG694496.1	918
1268.	GG694498.1	990
1269.	GG694500.1	2320
1270.	GG694502.1	1518
1271.	GG694504.1	1233
1272.	GG694506.1	916
1273.	GG694509.1	1605
1274.	GG694512.1	469
1275.	GG694514.1	1772
1276.	GG694516.1	1062
1277.	GG694518.1	1726
1278.	GG694521.1	1329
1279.	GG694523.1	1479
1280.	GG694526.1	1206
1281.	GG694528.1	852
1282.	GG694531.1	1259
1283.	GG694533.1	2017
1284.	GG694535.1	7890
1285.	GG694537.1	1002
1286.	GG694539.1	1606
1287.	GG694541.1	1374
1288.	GG694543.1	1777
1289.	GG694545.1	1986
1290.	GG694547.1	4123
1291.	GG694549.1	1492
1292.	GG694551.1	1993
1293.	GG694553.1	959
1294.	GG694555.1	1562
1295.	GG694557.1	954
1296.	GG694559.1	2471
1297.	GG694561.1	1595
1298.	GG694563.1	3263
1299.	GG694565.1	1250
1300.	GG694567.1	1994
1301.	GG694569.1	1594
1302.	GG694571.1	1720
1303.	GG694573.1	2146
1304.	GG694575.1	3736
1305.	GG694577.1	617
1306.	GG694579.1	853
1307.	GG694581.1	1590
1308.	GG694583.1	1361
1309.	GG694585.1	1464
1310.	GG694587.1	1432
1311.	GG694589.1	1093
1312.	GG694593.1	323
1313.	GG694595.1	1097
1314.	GG694597.1	1626
1315.	GG694599.1	1846
1316.	GG694602.1	1417
1317.	GG694604.1	1405
1318.	GG694606.1	883
1319.	GG694608.1	1286
1320.	GG694610.1	1153
1321.	GG694612.1	675
1322.	GG694615.1	1446
1323.	GG694617.1	1273
1324.	GG694620.1	1419
1325.	GG694622.1	885
1326.	GG694624.1	2866
1327.	GG694626.1	1423
1328.	GG694628.1	1102

1329.	GG694630.1	1845
1330.	GG694633.1	854
1331.	GG694636.1	1804
1332.	GG694638.1	1315
1333.	GG694640.1	945
1334.	GG694642.1	1835
1335.	GG694644.1	1245
1336.	GG694647.1	2198
1337.	GG694649.1	1287
1338.	GG694651.1	1925
1339.	GG694653.1	1612
1340.	GG694655.1	1498
1341.	GG694660.1	1731
1342.	GG694665.1	1555
1343.	GG694668.1	599
1344.	GG694674.1	944
1345.	GG694677.1	679
1346.	GG694679.1	631
1347.	GG694681.1	1492
1348.	GG694683.1	1516
1349.	GG694685.1	1605
1350.	GG694688.1	3671
1351.	GG694690.1	1273
1352.	GG694692.1	734
1353.	GG694694.1	1936
1354.	GG694698.1	1010
1355.	GG694702.1	1015
1356.	GG694704.1	1434
1357.	GG694707.1	901
1358.	GG694710.1	1470
1359.	GG694713.1	7625
1360.	GG694715.1	2932
1361.	GG694718.1	891
1362.	GG694721.1	940
1363.	GG694724.1	443
1364.	GG694726.1	1782
1365.	GG694728.1	1410
1366.	GG694730.1	2107
1367.	GG694733.1	841
1368.	GG694735.1	1410
1369.	GG694737.1	872
1370.	GG694739.1	5677
1371.	GG694741.1	812
1372.	GG694743.1	1272
1373.	GG694747.1	964
1374.	GG694751.1	3077
1375.	GG694754.1	833
1376.	GG694758.1	1192
1377.	GG694760.1	906
1378.	GG694762.1	1330
1379.	GG694765.1	861
1380.	GG694767.1	1233
1381.	GG694769.1	1841
1382.	GG694772.1	914
1383.	GG694778.1	2428
1384.	GG694780.1	1483
1385.	GG694782.1	1207
1386.	GG694785.1	1160
1387.	GG694789.1	933
1388.	GG694791.1	956
1389.	GG694794.1	843
1390.	GG694796.1	1465
1391.	GG694798.1	835
1392.	GG694800.1	1108
1393.	GG694802.1	1932
1394.	GG694804.1	1453
1395.	GG694806.1	1612
1396.	GG694808.1	1775
1397.	GG694811.1	848
1398.	GG694813.1	1537
1399.	GG694815.1	839
1400.	GG694817.1	806
1401.	GG694820.1	998
1402.	GG694824.1	1616
1403.	GG694827.1	6669
1404.	GG694830.1	859
1405.	GG694833.1	796
1406.	GG694835.1	2507
1407.	GG694837.1	893
1408.	GG694839.1	1280
1409.	GG694841.1	1724
1410.	GG694844.1	3485
1411.	GG694846.1	1022
1412.	GG694848.1	2827
1413.	GG694855.1	6027
1414.	GG694857.1	1500
1415.	GG694861.1	955
1416.	GG694863.1	953
1417.	GG694865.1	1588
1418.	GG694867.1	1195
1419.	GG694869.1	1519
1420.	GG694871.1	1373
1421.	GG694875.1	2641
1422.	GG694878.1	1558
1423.	GG694880.1	2736
1424.	GG694882.1	1221
1425.	GG694885.1	1100
1426.	GG694888.1	898
1427.	GG694890.1	625
1428.	GG694893.1	2084
1429.	GG694895.1	794
1430.	GG694897.1	1056
1431.	GG694900.1	2069
1432.	GG694902.1	992
1433.	GG694904.1	1312
1434.	GG694906.1	831
1435.	GG694908.1	1669
1436.	GG694910.1	1480
1437.	GG694912.1	1645
1438.	GG694916.1	1131
1439.	GG694918.1	2257
1440.	GG694920.1	2820
1441.	GG694922.1	1217
1442.	GG694924.1	1536
1443.	GG694926.1	1004
1444.	GG694929.1	1306
1445.	GG694931.1	1422
1446.	GG694935.1	1077
1447.	GG694937.1	3016
1448.	GG694940.1	741
1449.	GG694944.1	843
1450.	GG694947.1	1031
1451.	GG694949.1	1855
1452.	GG694951.1	2663
1453.	GG694953.1	892
1454.	GG694955.1	4698
1455.	GG694957.1	1126
1456.	GG694960.1	933
1457.	GG694964.1	735
1458.	GG694967.1	2498
1459.	GG694969.1	945
1460.	GG694972.1	2489
1461.	GG694974.1	2223
1462.	GG694976.1	1940
1463.	GG694978.1	606
1464.	GG694981.1	909
1465.	GG694983.1	1503
1466.	GG694985.1	1069
1467.	GG694987.1	2876
1468.	GG694989.1	732
1469.	GG694991.1	878
1470.	GG694993.1	1594
1471.	GG694995.1	2038
1472.	GG694997.1	7776
1473.	GG694999.1	915
1474.	GG695001.1	900
1475.	GG695004.1	1238
1476.	GG695008.1	2680
1477.	GG695010.1	1173
1478.	GG695012.1	979
1479.	GG695014.1	743
1480.	GG695016.1	1336
1481.	GG695018.1	1411
1482.	GG695020.1	1356
1483.	GG695022.1	607
1484.	GG695025.1	903
1485.	GG695027.1	930
1486.	GG695029.1	754
1487.	GG695031.1	1487
1488.	GG695033.1	1621
1489.	GG695035.1	897
1490.	GG695037.1	1616
1491.	GG695039.1	1620
1492.	GG695041.1	2455
1493.	GG695045.1	1090
1494.	GG695047.1	1148
1495.	GG695049.1	1091
1496.	GG695052.1	959
1497.	GG695055.1	2266
1498.	GG695057.1	1296
1499.	GG695060.1	943
1500.	GG695062.1	1655
1501.	GG695066.1	808
1502.	GG695068.1	1455
1503.	GG695070.1	2331
1504.	GG695073.1	709
1505.	GG695076.1	869
1506.	GG695078.1	827
1507.	GG695080.1	1154
1508.	GG695082.1	1674
1509.	GG695085.1	928
1510.	GG695087.1	1083
1511.	GG695089.1	2366
1512.	GG695091.1	901
1513.	GG695093.1	846
1514.	GG695096.1	1698
1515.	GG695098.1	1015
1516.	GG695100.1	964
1517.	GG695102.1	868
1518.	GG695104.1	2425
1519.	GG695106.1	1237
1520.	GG695109.1	1135
1521.	GG695111.1	1275
1522.	GG695113.1	923
1523.	GG695116.1	1027
1524.	GG695119.1	1183
1525.	GG695121.1	750
1526.	GG695124.1	925
1527.	GG695127.1	1588
1528.	GG695129.1	567
1529.	GG695131.1	1064
1530.	GG695133.1	1018
1531.	GG695135.1	3014
1532.	GG695137.1	741
1533.	GG695139.1	248
1534.	GG695142.1	952
1535.	GG695145.1	3934
1536.	GG695147.1	1725
1537.	GG695150.1	1026
1538.	GG695153.1	480
1539.	GG695155.1	4777
1540.	GG695157.1	982
1541.	GG695160.1	1209
1542.	GG695163.1	892
1543.	GG695165.1	5858
1544.	GG695167.1	848
1545.	GG695169.1	457
1546.	GG695172.1	972
1547.	GG695175.1	829
1548.	GG695178.1	895
1549.	GG695180.1	1516
1550.	GG695182.1	937
1551.	GG695185.1	1050
1552.	GG695187.1	1630
1553.	GG695189.1	1346
1554.	GG695191.1	822
1555.	GG695194.1	1191
1556.	GG695197.1	990
1557.	GG695200.1	2338
1558.	GG695202.1	879
1559.	GG695204.1	1035
1560.	GG695206.1	1233
1561.	GG695208.1	1339
1562.	GG695210.1	692
1563.	GG695213.1	1793
1564.	GG695215.1	928
1565.	GG695217.1	779
1566.	GG695219.1	1262
1567.	GG695221.1	785
1568.	GG695223.1	2730
1569.	GG695225.1	904
1570.	GG695227.1	1433
1571.	GG695229.1	887

1572.	GG695231.1	1545
1573.	GG695234.1	1330
1574.	GG695236.1	1840
1575.	GG695238.1	1479
1576.	GG695240.1	994
1577.	GG695242.1	857
1578.	GG695244.1	1316
1579.	GG695247.1	916
1580.	GG695250.1	985
1581.	GG695252.1	1267
1582.	GG695254.1	1268
1583.	GG695256.1	2702
1584.	GG695258.1	875
1585.	GG695260.1	1015
1586.	GG695264.1	8161
1587.	GG695266.1	857
1588.	GG695268.1	985
1589.	GG695271.1	798
1590.	GG695274.1	2171
1591.	GG695276.1	6779
1592.	GG695278.1	7673
1593.	GG695281.1	899
1594.	GG695284.1	616
1595.	GG695286.1	367
1596.	GG695288.1	345
1597.	GG695292.1	1216
1598.	GG695294.1	1534
1599.	GG695296.1	575
1600.	GG695298.1	1405
1601.	GG695300.1	607
1602.	GG695302.1	851
1603.	GG695304.1	1769
1604.	GG695307.1	1293
1605.	GG695309.1	712
1606.	GG695311.1	2487
1607.	GG695313.1	812
1608.	GG695315.1	2267
1609.	GG695317.1	1649
1610.	GG695319.1	954
1611.	GG695322.1	1170
1612.	GG695324.1	1132
1613.	GG695326.1	960
1614.	GG695329.1	1529
1615.	GG695333.1	2765
1616.	GG695337.1	2176
1617.	GG695339.1	1116
1618.	GG695342.1	820
1619.	GG695344.1	910
1620.	GG695346.1	913
1621.	GG695348.1	2633
1622.	GG695351.1	1729
1623.	GG695354.1	918
1624.	GG695356.1	1726
1625.	GG695358.1	793
1626.	GG695360.1	894
1627.	GG695362.1	1240
1628.	GG695364.1	1642
1629.	GG695369.1	1571
1630.	GG695372.1	1650
1631.	GG695374.1	4185
1632.	GG695377.1	1126
1633.	GG695380.1	959
1634.	GG695387.1	985
1635.	GG695389.1	1321
1636.	GG695391.1	1051
1637.	GG695393.1	2665
1638.	GG695395.1	1191
1639.	GG695398.1	976
1640.	GG695403.1	1511
1641.	GG695405.1	1727
1642.	GG695407.1	4461
1643.	GG695409.1	2032
1644.	GG695413.1	904
1645.	GG695415.1	30375
1646.	GG695418.1	1429
1647.	GG695420.1	5465
1648.	GG695423.1	1498
1649.	GG695425.1	1888
1650.	GG695428.1	1201
1651.	GG695430.1	1297
1652.	GG695432.1	2421
1653.	GG695434.1	1353
1654.	GG695436.1	4986
1655.	GG695439.1	916
1656.	GG695441.1	1422
1657.	GG695443.1	874
1658.	GG695445.1	4614
1659.	GG695447.1	928
1660.	GG695451.1	883
1661.	GG695453.1	18388
1662.	GG695455.1	932
1663.	GG695457.1	922
1664.	GG695461.1	1183
1665.	GG695463.1	1399
1666.	GG695466.1	810
1667.	GG695468.1	2003
1668.	GG695470.1	873
1669.	GG695472.1	1039
1670.	GG695475.1	2136
1671.	GG695477.1	1471
1672.	GG695480.1	1284
1673.	GG695484.1	1702
1674.	GG695486.1	894
1675.	GG695489.1	8995
1676.	GG695492.1	1446
1677.	GG695494.1	915
1678.	GG695497.1	1317
1679.	GG695499.1	880
1680.	GG695501.1	985
1681.	GG695503.1	2294
1682.	GG695505.1	1119
1683.	GG695508.1	796
1684.	GG695510.1	964
1685.	GG695513.1	1308
1686.	GG695515.1	1773
1687.	GG695517.1	2496
1688.	GG695520.1	981
1689.	GG695523.1	1261
1690.	GG695525.1	1571
1691.	GG695527.1	512
1692.	GG695529.1	1902
1693.	GG695532.1	827
1694.	GG695534.1	863
1695.	GG695536.1	2246
1696.	GG695538.1	961
1697.	GG695540.1	1002
1698.	GG695542.1	2201
1699.	GG695544.1	36208
1700.	GG695549.1	1571
1701.	GG695552.1	1627
1702.	GG695555.1	1335
1703.	GG695557.1	2397
1704.	GG695561.1	988
1705.	GG695564.1	1280
1706.	GG695566.1	1165
1707.	GG695569.1	898
1708.	GG695571.1	3969
1709.	GG695574.1	907
1710.	GG695577.1	1520
1711.	GG695579.1	3681
1712.	GG695582.1	1091
1713.	GG695584.1	1112
1714.	GG695586.1	1493
1715.	GG695590.1	372
1716.	GG695592.1	1723
1717.	GG695594.1	987
1718.	GG695598.1	1229
1719.	GG695601.1	1792
1720.	GG695603.1	1218
1721.	GG695606.1	1423
1722.	GG695608.1	995
1723.	GG695610.1	1015
1724.	GG695612.1	3293
1725.	GG695614.1	1018
1726.	GG695616.1	4565
1727.	GG695618.1	1585
1728.	GG695621.1	1043
1729.	GG695623.1	1384
1730.	GG695626.1	561
1731.	GG695628.1	1269
1732.	GG695631.1	1217
1733.	GG695633.1	752
1734.	GG695636.1	2757
1735.	GG695638.1	1225
1736.	GG695640.1	1992
1737.	GG695642.1	1985
1738.	GG695644.1	878
1739.	GG695647.1	869
1740.	GG695650.1	1425
1741.	GG695652.1	1287
1742.	GG695654.1	814
1743.	GG695656.1	1576
1744.	GG695659.1	1613
1745.	GG695662.1	1264
1746.	GG695665.1	10185
1747.	GG695667.1	1229
1748.	GG695669.1	328
1749.	GG695672.1	1883
1750.	GG695675.1	2284
1751.	GG695677.1	1169
1752.	GG695679.1	1327
1753.	GG695681.1	1636
1754.	GG695683.1	898
1755.	GG695685.1	8273
1756.	GG695687.1	1760
1757.	GG695689.1	1661
1758.	GG695691.1	882
1759.	GG695693.1	1466
1760.	GG695695.1	817
1761.	GG695697.1	912
1762.	GG695700.1	1430
1763.	GG695704.1	1548
1764.	GG695706.1	3883
1765.	GG695708.1	1373
1766.	GG695711.1	1084
1767.	GG695715.1	1221
1768.	GG695717.1	863
1769.	GG695719.1	895
1770.	GG695721.1	834
1771.	GG695726.1	876
1772.	GG695728.1	635
1773.	GG695730.1	930
1774.	GG695732.1	3259
1775.	GG695734.1	1133
1776.	GG695737.1	981
1777.	GG695740.1	638
1778.	GG695742.1	1002
1779.	GG695744.1	1481
1780.	GG695746.1	1789
1781.	GG695748.1	946
1782.	GG695750.1	1519
1783.	GG695752.1	2127
1784.	GG695754.1	1164
1785.	GG695756.1	1381
1786.	GG695759.1	901
1787.	GG695761.1	741
1788.	GG695763.1	1346
1789.	GG695765.1	1005
1790.	GG695767.1	1319
1791.	GG695770.1	2716
1792.	GG695772.1	1067
1793.	GG695775.1	1245
1794.	GG695777.1	799
1795.	GG695780.1	728
1796.	GG695783.1	1474
1797.	GG695785.1	1559
1798.	GG695790.1	919
1799.	GG695792.1	2770
1800.	GG695795.1	812
1801.	GG695797.1	1734
1802.	GG695799.1	619
1803.	GG695801.1	1566
1804.	GG695803.1	995
1805.	GG695805.1	1582
1806.	GG695807.1	1476
1807.	GG695809.1	1691
1808.	GG695811.1	1750
1809.	GG695813.1	652
1810.	GG695817.1	1425
1811.	GG695819.1	829
1812.	GG695821.1	874
1813.	GG695824.1	735
1814.	GG695826.1	13020

1815.	GG695828.1	762
1816.	GG695830.1	1766
1817.	GG695832.1	2329
1818.	GG695834.1	1039
1819.	GG695836.1	2139
1820.	GG695838.1	2776
1821.	GG695840.1	867
1822.	GG695842.1	752
1823.	GG695845.1	1254
1824.	GG695847.1	1293
1825.	GG695849.1	848
1826.	GG695851.1	1298
1827.	GG695853.1	1681
1828.	GG695856.1	898
1829.	GG695862.1	2402
1830.	GG695864.1	918
1831.	GG695867.1	1243
1832.	GG695869.1	12620
1833.	GG695871.1	1344
1834.	GG695876.1	1088
1835.	GG695878.1	942
1836.	GG695881.1	4394
1837.	GG695884.1	2543
1838.	GG695887.1	1449
1839.	GG695889.1	1578
1840.	GG695891.1	2764
1841.	GG695893.1	1632
1842.	GG695895.1	1094
1843.	GG695898.1	729

7. Popis literature

[1] Ohno, S. (1972) So much "junk" DNA in our genome. Brookhaven. Symp. Biol., 23,366 - 370.

[2] Jelinek, W. R., Toomey, T.P., Leineand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L. et.al. (1980) Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Natl. Acad. Sci. U.S.A., 77, 1398 - 1402.

[3] Pennacchio, L. A. and Rubin, E. M. (2001) Genomic strategies to identify mammalian regulatory sequences. Nat. Rev. Genet., 2, 100 - 109.

[4] Batzer, M. A. and Deininger, P. L. (2002) Alu repeats and human genomic diversity. Nature Genet. , 3, 370 - 379.

[5] Gelfand, Y., Rodriguez, A. and Benson, G. (2007) TRDB – the tandem repeats database. Nucleic Acid Res., 35, D80 - D87.

[6] King, D. C., Taylor, J., Zhang, Y., Cheng, Y., Lawson, H. A., Martin, J., Chiaromonte, F., Miller, W. and Hardison, R.C. (2007). Finding cis – regulatory elements using comparative genomics: Some lessons from ENCODE data. Genome Res., 17, 775 - 786.

[7] Mercer, T., R., Dinger, M.E. and Mattick, J. S. (2009) Long non – coding RNAs: insights into functions. Nat. Rev. Genet., 10, 155 - 159.

[8] Garfield, D.A. and Wray, G. A. (2010) The evolution of gene regulatory interactions. BioScience, 60, 15-23.

[9] Noonan, J. P. and McCallion, A. S. (2010) Genomics of long – range regulatory elements. Annu. Rev. Genomics Hum. Genet., 11, 1 - 23.

[10] Kashi, Y and King, D. G. (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet., 22, 253 - 259.

[11] Gemayel, R., Vinces, M. D., Legendre, M and Verstrepen, K. J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet., 44, 445 - 477.

[12] Jansen, A., Gemayel, R. and Verstrepen, K. J. (2012) Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. Genome Dyn., 7, 108 - 125.

[13] Benson, G. (1999) Tandem Repeat Finder: a program to analyze DNA sequences. Nucleic Acid Res. , 27, 573 – 580.

[14] Sharma, D., Isac, B., Raghava, G., P., S. And Ramaswamy, R. (2004) Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformations. Bioinformatics, 20, 1405 -1412

[15] Mitchell, A. R., Gosden, J. R., Miller, D. A. (1985) A cloned sequence, p82H, of the alphoid repeated DNA family found at the centromeres of all human chromosomes. Chromosoma , 92, 369 – 377.

[16] Waye, J. S., Willard, H.F. (1987) Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. Nucleic Acid Res. 15, 7549 - 7569.

[17] Willard, H. F. (1958) Chromosome – specifc organization of human alpha satellite DNA. Am J Hum Genet., 37, 524-532.

[18] Ugarkovic, D. (2005) Functional elements residing within satellite DNAs. BMBO reports, 6, 1035 – 1039.

[19] Warburton, P. E., Willard, H. F. (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In: Jackson, M., Strachan, T., Dover, G. (eds) Human Genome Evolution. BIOS Scientific, Oxford, pp. 121 – 145.

[20] Alexandrov I., Kazakov, A., Tumeneva, I., Shepelev,, V., Yurov, Y (2001) Alpha satellite DNA of primates and new families. Chromosoma, 110, 253 – 266.

[21] Paar, V., Glunčić, M., Rosandić, M., Basar, I., Vlahović, I. (2011) Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees. Mol. Biol. Evol., 28, 1877 – 1892.

[22] Glunčić, M., Rosandić, M. Jelovina, D., Dekanić, D., Vlahović, I. and Paar, V. Global Repeat Map Method for Higher Order Repeat Alpha Satellites in Human and Chimpanzee Genomes (Build 37.2 Assembly). (2012) Croat. Chem. Acta, 85, 327 - 351.

[23] Paar, V., Glunčić, M., Basar, I., Rosandić, M., Paar, P., Cvitković, M. (2011) Large Tandem, Higher Order Repeats Contribute Substantially to Divergence Between Human and Chimpanzee Y Chromosomes. J Mol Evol, 72, 34-55.

[24] Lim, K. G., Kwoh, C. K., Hsu, L. Y. and Wirawan, A. (2012) Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. Brief Bioinform. doi: 10.1093/bib/bbs023.

[25] Tomilin, N. V. (2008) Regulation of mammalian gene expression by retroelements and non – coding tandem repeats. BioEssays, 30, 338-348.

[26] Reichwald, K., Lauber, C, Nanda, I., Kirschner, J., Hartmann, N., Schories†, Ulrike Gausmann, S., Taudien, S., Schilhabel, M. B., Szafranski, K., Glöckner, G., Schmid, M., Cellerino, A., Schartl, M., Englert, C. and Platzer, M. (2009) High tandem repeat content in the genome of the short-lived annual fish Nothobranchius furzeri: a new vertebrate model for aging research. Genome Biol., 10(2), R16.

[27] Mravinac, B. Plohl., M., Ugarković, Đ. (2004) Conserved patterns in the evolution of *Tribolium satellite* DNAs. Gene, 332, 169 – 177.

[28] Mravinac, B. Plohl, M., Ugarković, Đ. (2005) Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. J Mol Evol, 61, 542 – 550.

[29] Wang, S., Lorenzen, M. D., Beeman, R. W., Brown, S. J. (2008). Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biology, 9:R61)

[30] Plohl, M. (2010) Those mysterious sequences of satellite DNAs. Periodicum Biologorum, 112, 403 – 410.

[31] Feliciello, I., Chinali, G., Ugarkovic, Đ. (2011). Structure and population dynamics of the major satellite DNA in the red flour beetle *Tribolium castaneum*. Genetica, 139, 999–1008.

[32] Altschul1, S. F., Gish, W., Miller, W., Myers, E. W.,. Lipman, D. J. (1990) Basic local alignment search tool. 215, 403 – 410.

[33] Sanger, F. , Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. molec. Biol., 94, 441–448.

[34] Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. natn. Acad. Sci. USA, 74, 5463–5467.

[35] Griffin, H. G., Grihhin, A.M. (1993) DNA sequencing – recent innovations and future trends. Appl. Biochem. Biotechnol. 38, 147–159.

[36] Franca, L. T. C., Carrilho, E. and Kist, T.B. L. (2002) A review of DNA sequencing techniques. Quarterly Reviews of Biophysics, 35, 169 – 200.

[37] Maxam, A. M., Gilbert, W. (1977) A new method for sequencing DNA. Proc. natn. Acad. Sci. USA, 74, 560 – 564.

[38] Voss, H., Schwager, C., Wirkner, U., Sproat, B., Zimmermann, J. Rosenthal, A., Erfle, H., Stegemann, J. and Ansorge, W. (1989) Direct genomic fluorescent on-line sequencing and analysis using in vitro amplification of DNA. Nucleic Acids Res., 17, 2517 – 2527.

[39] Nyren, P., Lundin, A. (1985) Enzymatic method for continuos monitoring of inorganic pyrophosphate synthesis. Analyt. Biochem. 151, 504 – 509.

[40] Hyman, E. D. (1988). A new method of sequencing DNA. Analyt. Biochem. 174, 423 – 436.

[41] Ronaghi, M., Karamohamed,, S., Pettersson, B., Uhlén, M. and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Analyt. Biochem., 242, 84 – 89.

[42] Jett, J. H., Keller, R. A., Martin, J. C., Marrone, B. L. Moyzis., R. K., Ratliff, R. L., Seitzinger, N. K., Shera, E. B., and Stewart, C.C. (1989) High – speed DNA sequencing – an approach based upon fluorescence detection of single molecules. J. biomolec. Struct. Dyn, 7, 301 – 309.

[43] http://www.ddbj.nig.ac.jp/

[44] http://www.embl.org/

[45] http://www.ncbi.nlm.nih.gov/

[46] http://www.genome.gov/10001772

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

[48] http://report.nih.gov/NIHfactsheets/ViewFactSheet.aspx?csid=45&key=H#H

[49] http://www.genome.gov/11006946

[50] Richards, S. Tribolium Genome Sequencing Consortium. (2008) The genome of the model beetle and pest Tribolium castaneum. Nature, 452, 949 - 955.

[51] Druker, B. J., Sawyers, C. L., Kantarjian, H., Resta, D. J., Reese, S. F., Ford, J. M., et al. (2001) Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis

of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. N Engl J Med;344:1038-42.

[52] http://blast.ncbi.nlm.nih.gov/Blast.cgi

[53] http://www.ncbi.nlm.nih.gov/tools/epcr/

[54] http://www.ncbi.nlm.nih.gov/gorf/gorf.html

[56] http://www.ensembl.org/index.htm

[57] http://www.bioinformatics.org/combosa3d/

[58] http://www.bioinformatics.org/sewer/

[59] Pathak, D., Ali, S. (2012). Repetitive DNA: A Tool to Explore Animal Genomes/Transcriptomes, Functional Genomics, Dr. Germana Meroni (Ed.), ISBN: 978-953-51-0727-9, InTech, DOI: 10.5772/48259.

[60] Brown, T. A. (2002) Genomes, Oxford: Wiley-Liss; 2002. ISBN-10: 0-471-25046-5.

[61] Cordaux, R., Batzer, M. A. (2009) The impact of retrotransposons on human genome evolution. Nature Reviews Genetics, 10, 691 - 703.

[62] Muñoz-López, M., García-Pérez, J. L. (2010) DNA Transposons: Nature and Applications in Genomics. Curr Genomics 11, 115 – 128.

[63] Schaap, M., et al. (2013) Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. BMC Genomics, 14, 143.

[64] Georges et al. (1991). Characterization of a set of variable number of tandem repeat markers conserved in bovidae. Genomics.11, 24 - 32.

[65] Mravinac, B., Meštrović, N., Čavrak, V. V., Plohl, M. (2011) TCAGG, an alternative telomeric sequence in insects. CHROMOSOMA, 120, 367 – 376.

[66] Jeffreys, A.J., Neil, D. L., Neumann, R. (1998). Repeat instability At Humna Minisatellites Arising From Meiotic Recombination. Embo J., 17, 4147 - 4157.

[67] Legendre, M., Pochet, , N., Pak, T., Verstrepen, K.J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res., 17,1787 - 96.

[68] Ambrosini, A., Paul, S., Hu, S., Riethman, Human subtelomeric duplicon structure and organization. Genome Biology, 8, 151.

[69] Behura, S. K., Severson, D. W. (2013) Association of microsatellite pairs with segmental duplications in insect genomes. BMC Genomics, 14, 907.

[70] Catasti, P., Chen, X., Mariappan, S. V., Bradbury, E..M., Gupta, G. (1999) Dna repeats In The Human Geenome. Genetica, 106, 15 - 36.

[71] Martienssen, R. A., Colot, V., (2001) DNA Methylation And Epigenetic Inheritance In Plants And Filamentous Fungi. Science, 293,1070 – 1074.

[72] Zhang, L., Yuan, D., Yu, S., Li, Z., Cao, Y., Miao, Z., Qian, H., Tang, K. (2004) Preference of simple sequence repeats in coding and non-coding regions of Arabidopsis thaliana. Bioinformatics, 20, 1081 – 1086.

[73] Mirkin, S. M. (2007) Expandeble DNA repeats and human disease. Nature, 447, 932 - 940.

[74] Warby, S. C., Graham, R. K, Hayden, M. R. (1998) Huntington Disease. GeneReviews® [Internet].

[75] Demuth, J. P., Drury, D. W., Peters, M. L., David van Dyken, j., Priest, N. K., Wade, M. J. (2007) Genome – wide survey of *Tribolium castaneum* microsatellites and description of 509polymorphic markers. Molecular Ecology Notes, 7, 1189 – 1195.

[76] Charlesworth, B., Sniegowski, P., Stephan, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371, 215 - 220.

[77] Ugarkovic, D., Plohl, M., (2002) Variation In Satellite DNA Profiles - Causes And Effects. EMBO, 21, 5955 – 5959.

[78] Dover, G. A. (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet., 2, 159–165.

[79] http://www.ufrgs.br/imunovet/molecular_immunology/dnaturnover.html

[80] Smith, G. P. (1976) Science. Evolution of repeated DNA sequences by unequal crossover. 191, 528-35.

[81] http://www.web-books.com/MoBio/Free/Ch8D6.htm

[82] Stephan, W. (1986) Recombination and the evolution of satellite DNA. Genet Res., 47,167 - 74.

[83] Viguera, E., Canceill, D., Ehrlich, S. D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. EMBO J., 20, 2587 – 2595.

[84] Fire, A., Xu, S. Q. (1995) Rolling replication of short DNA circles. Proc. Natl. Acad. Sci. USA.,92, 4641 – 4645.

[85] King K, Jobst J, Hemleben V. (1995) Differential homogenization and amplification of two satellite DNAs in the genus *Cucurbita* (*Cucurbitaceae*). J Mol Evol., 41, 996 - 1005.

[86] Mestrović, N., Plohl, M., Mravinac, B., Ugarković, D. (1998) Evolution of satellite DNAs from the genus *Palorus*-experimental evidence for the "library" hypothesis. Mol Biol Evol.,15, 1062 - 8.

[87] Fry, K., Salser, W. (1977) Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. Cell., 12, 1069 - 84.

[88] Ugarković, D., Durajlija, S., Plohl, M. (1996) Evolution of *Tribolium madens* (Insecta, Coleoptera) satellite DNA through DNA inversion and insertion. J Mol Evol., 42, 350 -8.

[89] Abad JP, Carmena M, Baars S, Saunders RD, Glover DM, Ludeña P, Sentis C, Tyler-Smith C, Villasante A. (1992) Dodeca satellite: a conserved G+C-rich satellite from the centromeric heterochromatin of *Drosophila melanogaster*. Proc Natl Acad Sci U S A., 89, 4663 -7.

[90] Mravinac, B., Plohl, M., Mestrovic, N., Ugarkovic, D., (2002) Sequence of PRAT satellite DNA 'frozen' in some coleopteran species. Jmol Evol, 54, 774 – 783.

[91] YX, L., Kirby M. L. (2003) Coordinated and conserved expression of alphoid repeat and alphoid repeat – tagged coding sequences. Dev Dynamics, 228, 72 - 81.

[92] Henikoff, S., Ahmad, K., Malik, H. S. (2001) Review The centromere paradox: stable inheritance with rapidly evolving DNA. Science., 293, 1098 - 102.

[93] Rosandić, M., Paar, V., Glunčić, M., Basar, I., Pavin, N. (2003) Key-string Algorithm - Novel Approach to Computational Analysis of Repetitive Sequences in Human Centromeric DNA., CMJ, 44, 386 – 406.

[94] Rosandić, M., Glunčić, M., Paar, V. (2011) Start/stop Codon-like Trinucleotides (CLTs) and Extended Clusters as New Language of DNA. Croat. Chem. Acta, 84, 331 – 341.

[95] Rosandić, M., Glunčić, M., Paar, V. (2013). Extended start/stop codon like trinucleotides (CLTs) as regulators and "new language" in noncoding DNA. Bioinformatics and biological physics : proceedings of the scientific meeting, 191-215.

[96] Palomeque, T., Lorite, P. (2008) Satellite DNA in insects: a review . Heredity, 100, 564–573.

[97] King, L. M., Cummings, M. P. (1997). Satellite DNA repeat sequence variation is low in three species of burying beetles in the genus *Nicrophorus* (Coleoptera: *Silphidae*). Mol Biol Evol, 14, 1088 – 1095.

[98] Blanchetot, A. (1991). Genetic variability of a satellite sequence in the dipteran *Musca domestica*. EXS, 58, 106 – 112.

[99] Lorite, P., Carrillo, J. A., Aguilar, J. A., Palomeque, T. (2004). Isolation and characterization of two families of satellite DNA with repetitive units of 135 bp and 2.5 kb in the ant *Monomoriumsubopacum* (*Hymenoptera, Formicidae*). Cytogenet Genome Res, 105, 83 – 92.

[100] Yoshimura, A., Nakata, A., Mito, T., Noji, S. (2006). The characteristics of karyotype and telomeric satellite DNA sequences in the cricket, *Gryllus bimaculatus* (*Orthoptera, Gryllidae*). Cytogenet Genome Res.,112, 329 – 336.

[101] Bachmann, L., Sperlich, D. (1993). Gradual evolution of a specific satellite DNA family in *Drosophila ambigua, D. tristis,* and *D. obscura*. Mol Biol Evol, 10, 647 – 659.

[102] Plohl, M., Mestrovic, N., Bruvo, B., Ugarkovic, D. (1998). Similarity of structural features and evolution of satellite DNAs from *Palorus subdepressus (Coleoptera)* and related species. J Mol Evol, 46, 234 – 239.

[103] Bonaccorsi, S., Lohe, A. (1991) Fine mapping of satellite DNA sequences along the Y chromosome of *Drosophila melanogaster*: relationships between satellite sequences and fertility factors. Genetics, 129, 177 – 189.

[104] Pons, J., Bucur, R., Vogler, A. P. (2003). Higher-order repeats in the satellite DNA of the cave beetle *Pholeuon proserpinaeglaciale* (*Coleoptera: Cholevidae*). Hereditas, 139, 28 –34.

[105] Palomeque, T., Muñoz-López, M., Carrillo, J. A., Lorite, P. (2005) Characterization and evolutionary dynamics of a complex family of satellite DNA in the leaf beetle *Chrysolina carnifex* (*Coleoptera, Chrysomelidae*). Chromosome Res, 13, 795 – 807.

[106] Mravinac, B., Ugarkovic, D., Franjevic, D., Plohl, M. (2005). Long inversely oriented subunits form a complex monomer of *Tribolium brevicornis* satellite DNA. J Mol Evol, 60, 513 – 525.

[107] Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zabolan, M., Marks, D., Gaasterland, T. et al. (2003). The small RNA profile during *Drosophila melanogaster* development. Dev Cell, 5, 337 –350 .

[108] Usakin, L., Abad, J., Vagin, V. V., De Pablos, B., Villasante, A., Gvozdev, V. A. (2007) Transcription of the 1.688 satellite DNA family is under the control of RNA interference machinery in *Drosophila melanogaster* ovaries. Genetics, 176, 1343 – 1349.

[109] Rojas, A. A., Vázquez-Tello, A., Ferbeyre, G., Venanzetti, F., Bachmann, L., Paquin, B. et al. (2000) Hammerhead-mediated processing of satellite pDo500 family transcripts from *Dolichopoda* cave crickets. Nucleic Acids Res, 28, 4037 – 4043.

[110] Rouleux-Bonnin, F., Bigot, S., Bigot, Y. (2004) Structural and transcriptional features of *Bombus terrestris* satellite DNA and their potential involvement in the differentiation process. Genome, 47, 877 – 888.

[111] Heikkinen, E., Launonen, V., Muller, E., Bachmann, L. (1995) The pvB370 BamHI satellite DNA family of the *Drosophila virilis* group and its evolutionary relation to mobile dispersed genetic pDv elements. J Mol Evol, 41, 604 – 614.

[112] Sun, X., Wahlstromm J. M. , Karpen, G. H. (1997) Molecular structure of a functional Drosophila centromere. Cell, 9, 1007 – 1019.

[113] Stratikopoulos, E. E., Augustinos, A. A., Gariou-Papalexiou, A., Zacharopoulou, A., Mathiopoulos, K. D. (2002) Identification and partial characterization of a new Ceratitis capitata-specific 44-bp pericentromeric repeat. Chromosome Res, 10, 287 – 295.

[114] López-León, M. D., Vazquez, P., Hewitt, G. M., Camacho, J. P. (1995). Cloning and sequence analysis of an extremely homogeneous tandemly repeated DNA in the grasshopper *Eyprepocnemis plorans*. Heredity, 75, 370 – 375.

[115] Luchetti, A., Marini, M., Mantovani, B. (2006) Non-concerted evolution of the RET76 satellite DNA family in Reticulitermes taxa (Insecta, Isoptera). Genetica, 128, 123-132.

[116] Yoda, K., Ando, S., Okuda, A., Kikuchi, A., Okazaki, T. (1998) In vitro assembly of the CENP-B/alpha-satellite DNA/core histone complex: CENP-B causes nucleosome positioning. Genes Cells, 3, 533 – 548.

[117] Kipling, D., Warburton, P. B. (1997) Centromeres, CENP-B and Tigger too. Trends Genet, 13, 141 – 144.

[118] Henikoff, S., Malik, H. S. (2002). Centromeres: selfish drivers. Nature, 417, 227.

[119] Talbert, P. B., Bryson, T. D., Henikoff, S. (2004) Adaptive evolution of centromere proteins in plants and animals. J Biol, 3, 18.

[120] Lohe, A. R., Hilliker, A. J., Roberts, P. A. (1993) Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. Genetics, 134, 1149 – 1174.

[121] Landais, I., Chavigny, P., Castagnone, C., Pizzol, J., Abad, P., Vanlerberghe-Masutti, F. (2000) Characterization of a highly conserved satellite DNA from the parasitoid wasp *Trichogramma brassicae*. Gene, 255, 65 – 73.

[122] Sun, X., Le, H. D., Wahlstrom, J. M., Karpen, G. H. (2003) Sequence analysis of a functional *Drosophila centromere*. Genome Res, 13, 182 – 194.

[123] Hall, S. E., Kettler, G., Preuss, L. (2003) Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. Genome Res, 13, 195 –205.

[124] Hall, S. E., Luo, S., Hall, A. E., Preuss, D. (2005) Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. Genetics, 170, 1913 – 1927.

[125] Brown, S. J., Henry, J. K., Black, W. C.4th, Denell, R. E.(1990) Molecular genetic manipulation of the red flour beetle: genome organization and cloning of a ribosomal protein gene. Insect Biochem, 20, 185 - 193.

[126] Beeman, R. W., Thomson, M. S., Clark, J. M., DeCamillis, M. A., Brown, S. J., Denell, R. E. (1996) Woot, an active gypsy-class retrotransposon in the flour beetle, *Tribolium castaneum*, is associated with a recent mutation. Genetics, 143, 417 - 426.

[127] Juan, C., Petitpierre, E.(1989) C-banding and DNA content in seven species of *Tenebrionidae* (*Coleoptera*). Genome, 32, 834 – 839.

[128] Kim, H. S., Murphy, T., Xia, J., Caragea, D., Park, Y., Beeman, R. W., Lorenzen, M. D., Butcher, S., Manak, J. R.,Brown,S. J. (2010) BeetleBase in 2010: revisions to provide

comprehensive genomic information for *Tribolium castaneum*. Nucleic Acids Research, 38, 437 - 442.

[129] Brajković, J., Feliciello, I., Bruvo-Mađarić, B.,Ugarković, Đ. (2012) Satellite DNA-Like Elements Associated With Genes Within Euchromatin of the Beetle *Tribolium castaneum.* G3, 2, 931 - 941.

[130] Kolpakov, R., Kucherov, G. (2003) Finding aproxximate repetitions uder Hamming distance. Theoret. Comput. Sci., 33, 135 -156.

[131] Landau, G. M., Schmidt, J. P., Sokol, D. (2001) An algorithm for approximate tandem repeats. J. Comput. Biol., 8, 1 - 18.

[132] Sokol, D., Benson, G., Tojeira, J. (2006) Tandem repeats over the edit distance. Bioinform, 23, e30 - e35.

[133] Kao, M.Y (2008) Encyclopedia of algorithms. Springer Science+Buisiness Media LLC. ISBN: 978-0-387-30162-4.

[134] Kolpakov, R., Bana, G, Kucherov, G. (2003) Mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res., 31, 3672 -3678.

[135] Delgrange, O., Rivals, E.(2004) STAR: an algorithm to search for tandem approximate repeats. Bioinformatics; 20, 2812 - 2820.

[136] Myers, E. W., Miller, W. (1989) Approximate matching of regular expressions. Bulletin of Mathematical Biology, 51, 5 – 37.

[137] Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T., Kudo, Y., Mori, H., Kanaya, S. (2002) Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. Gene, 300, 203 - 211.

[138] Vaidyanathan, P. P., Yoon, B. J. (2004) The role of signal-processing concepts in genomics and proteomics. J. Franklin Inst., 341, 111-135.

[139] Gupta, R., Sarthi, D., Mittal, A., Singh, K. (2007) A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequences. EURASIP J. Bioinform. Syst. Biol.,1, 43596.

[140] Chechetkin, V. R. (2011) Spectral sum rules and search for periodicities in DNA sequences. Phys. Lett. A, 375, 1729 - 1732.

[141] Cristea, P. D. (2003) Large scale features in DNA genomic signals. Sign. Process., 83, 871 - 888.

[142] Voss, R.F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett., 68, 3805 - 3808.

[143] Li, W. (1997) The study of correlation structures of DNA sequences: a critical review. Comput. Chem., 21, 257 – 271.

[144] Herzel, H., Weiss, O. and Trifonov, E.N. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics, 15, 187–193.

[145] Ramaswamy, R., Ramachandran, S. (1999) Gene identification in bacterial and organellar genomes using GeneScan. Comput. Chem., 23, 165–174.

[146] Glunčić,M., Paar, V. (2012) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Research, 41, e17.

[147] Vlahović, I., Glunčić, M., Rosandić, M., Dekanić, K., Ugarković, Đ., Paar, V. (2013) Pronounced repeat and higher order repeat structure in genome of insect Tribolium castaneum. Bioinformatics and biological physics : proceedings of the scientific meeting. Vladimir Paar (ur.). Zagreb : Hrvatska akademija znanosti i umjetnosti, 191-215.

[148] http://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

[149] Needleman, S. B., Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443 – 53.

[150] Collins, J. F., Coulson, A. F. (1984) Applications of parallel processing algorithms for DNA sequence analysis. Nucleic Acids Research 12, 181 – 192.

[151] Smith, T. F., & Waterman, M. S. (1981) Identification of common molecular subsequences. Journal of Molecular Biology, 147, 195 – 197.

[152]http://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096#

[153] http://www.ncbi.nlm.nih.gov/genome/216?genome_assembly_id=59530

8. Životopis

Ime i prezime: Ines Vlahović

Datum rođenja: 15.10.1983.

Mjesto rođenja: Zagreb, Hrvatska

OBRAZOVANJE:

2008 – sada:	Sveučilišteu Zagrebu, Prirodoslovno-matematički fakultet, Poslijediplomski studij fizike, smjer Biofizika
2002 – 2008:	Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Fizički odsjek, smjer Profesor fizike i informatike
2006 – 2007:	Informatičko učilište Algebra, Zagreb, ECDL ispitivač
1998 – 2002:	I. Gimnazija, Zagreb (opći smjer)
1990 – 1998:	Osnovna škola „Ive Andrića“, Zagreb
1996 – 1998 :	Glazbena škola „Ivan Zajec“, Zagreb
1994 – 2002:	Djevojački zbor „Zvjezdice“, Zagreb
1991 – 2000:	Škola za strane jezike „Varšavska – Sova“

RADNO ISKUSTVO

2008 – sada:	Znanstveni novak – PMF, Fizički odsjek, Zagreb Držanje nastave iz područja fizike i informatike, rad na projektu MZOŠ-a 119-0982464-1253,
2006 – 2009:	Predavač na informatičkom učilištu „Algebra“
2013:	Izvršni urednik zbornika radova: Bioinformatika i biološka fizika

NAGRADE

05.06.2001.:	2 svjetska zlata u zborskom pjevanju „Langollen International Music Festival“
2002. -2007.	Državna stipendija za izvrsnost

PODRUČJA ZNANOSTI:

Biofizika, interdisciplinarne prirodne znanosti, bioinformatika.

ČLANSTVA U PROFESIONALNIM UDRUGAMA/DRUŠTVIMA

2011. - sada: Hrvatsko biofzičko društvo

ZNANSTVENA PUBLIKACIJA

Izvorni znanstveni radovi i pregledni radovi u CC časopisima

1. Glunčić, Matko; Rosandić, Marija; Jelovina, Denis; Dekanić, Krešimir; Vlahović, Ines; Paar, Vladimir. Global repeat map method for higher order repeat alpha satellites in human and chimpanzee genomes (build 37.2 assembly). // Croatica chemica acta. 85 (2012), 3; 327-351.

2. Paar, Vladimir; Glunčić, Matko; Rosandić, Marija; Basar, Ivan; Vlahović, Ines. Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. // Molecular biology and evolution. 28 (2011), 6; 1877-1892.

Drugi radovi u zbornicima skupova s recenzijom

1. Glunčić, M.; Paar, V.; Basar, I.; Vlahović, I.; Rosandić, M.; Dekanić, K.; Cvitković, M.; Jelovina D.;Paar, P.; Kelić, A.; Batista, J. Direct mapping of symbolic DNA sequence int frequency domain and identification of higher order repeats// Bioinformatics and biological physics: proceedings of the scientific meeting/Vladimir Paar (ur). Zagreb: Hrvatska akademija znanosti i umjetnosti, 2013.17-46 (predavanje, domaća recenzija, objavljeni rad).
2. Vlahović, I; Glunčić, M.; Rosandić, M.; Dekanić, K.; Ugarković, Đ.; Paar, V. Pronaunced repeat and higher order repeat structure in genome of insect *Tribolium castaneum*.// Bioinformatics and biological physics: proceedings of the scientific meeting/Vladimir Paar (ur). Zagreb: Hrvatska akademija znanosti i umjetnosti, 2013. 191 – 215 (predavanje, domaća recenzija, objavljeni rad).

Radovi u zbornicima skupova bez recenzije

1. Glunčić, M.; Vlahović, I. Nastavne metode prikaza gibanja u gravitacijskom polju uz pomoć računalnih aplikacija.// NASTAVE FIZIKE-POSTIGNUĆA I IZAZOVI, ZBORNIK RADOVA/Pećina, P. (ur.). Zagreb: Hrvatsko fizikalno društvo, 2011. 255-149 (predavanje, objavljeni rad).

Sažeci u zbornicima skupova

1. Ines Vlahović, Matko Glunčić, Marija Rosandić, Vladimir Paar. Pronalaženje struktura repeticija višeg reda u genomu insekta *Tribolium castaneum* koristeći metodu Global Repeat Map// Knjiga sažetaka, 8. Znanstveni sastanak Hrvatskog fizikalnog društva. 2013. 127 -127 (poster, sažetak, znanstveni).
2. Matko Glunčić, Marija Rosandić, Ines Vlahović, Mislav Cvitković, Vladimir Paar. Direktno preslikavanje simboličke sekvence u frekventnu domenu i primjena na istraživanje korelacija u genomima eukariota.// Kniga sažetaka, 7. Znanstveni sastanak Hrvatskog fizikalnog društva. 2011, (predavanje, sažetak, znanstveni)

Druge vrste radova

1. Vlahović, Ines; Glunčić, Matko; Ugarković, Đurđica; Paar, Vladimir. Finding tandem repeats in *Tribolium castaneum* using computational method Global Repeat Map. //Knjiga sažetaka – 11th Greta Pifat Mrzljak International School of Biophysics. Primošten, Croatia, 30.9-9.10.2012, p75.

Diplomski rad

1. Vlahović, Ines. Povezivanje nastave informatike i fizike na primjeru Excela./Zagreb: PMF, 25.09.2008., 57 str. Voditelj: Gorjana Jerbić-Zorc.