

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA
ANALIZA MIKROSATELITOV V KLOOROPLASTNEM
GENOMU LAŠKEGA SMILJA
(*Helichrysum italicum* (Roth) G. Don)

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

**ANALIZA MIKROSATELITOV V KLOROPLASTNEM
GENOMU LAŠKEGA SMILJA
(*Helichrysum italicum* (Roth) G. Don)**

(Microsatellite analysis in the chloroplast genome of immortelle (*Helichrysum italicum* (Roth) G. Don))

Ime in priimek: Zala Žužek

Študijski program: Bioinformatika

Mentor: doc. dr. Matjaž Hladnik

Somentor: izr. prof. dr. Dunja Bandelj

Koper, september 2020

Ključna dokumentacijska informacija

Ime in PRIIMEK: Zala ŽUŽEK

Naslov zaključne naloge: Analiza mikrosatelitov v kloroplastnem genomu laškega smilja (*Helichrysum italicum* (Roth) G. Don)

Kraj: Koper

Leto: 2020

Število listov: 44

Število slik: 10

Število tabel: 6

Število prilog: 1

Št. strani prilog: 3

Število referenc: 40

Mentor: doc. dr. Matjaž Hladnik

Somentor: izr. prof. dr. Dunja Bandelj

Ključne besede: kloroplastna DNA, mikrosatelitski markerji, Laški smilj, GMATA, krait, MISA

Izvleček:

Mikrosatelitski markerji so prepoznani kot učinkoviti za identifikacijo rastlinskih vrst in filogenetske raziskave. Namen naloge je bil identificirati in karakterizirati mikrosatelitske lokuse (SSR) v kloroplastni DNA laškega smilja ter razvoj začetnih oligonukleotidov za pomnoževanje identificiranih lokusov. V nalogi je predstavljen pregled študij na temo mikrosatelitskih markerjev iz kloroplastne DNA pri genetskih raziskavah kot npr. filogeografija, populacijske študije, identifikacija vrst, itd. Prav tako je predstavljen tudi pregled prisotnosti mikrosatelitov v kloroplastni DNA drugih rastlinskih vrst. S pomočjo bioinformacijskih orodij smo poiskali mikrosatelitske lokuse na kloroplastnem genomu laškega smilja ter razvili začetne oligonukleotide za nadaljnje testiranje informativnosti in morebitno uporabo kot molekularni markerji za genetske analize. Za identifikacijo SSR v kloroplastnem genomu smilja smo uporabili bioinformacijska orodja GMATA, MISA ter Krait ter rezultate primerjali. Dodatno smo na kloroplastni genom nalegli odčitke, ki so jih pridobili raziskovalci UP FAMNIT pri sekvenciranju genomske DNA laškega smilja in s tem preverili morebitno prisotnost mutacij oz. razlik v številu ponovitev motiva SSR, kar bi lahko bila posledica heteroplazmije.

Key document information

Name and SURNAME: Zala Žužek

Title of the final project paper: Microsatellite analysis in the chloroplast genome of immortelle (*Helichrysum italicum* (Roth) G. Don)

Place: Koper

Year: 2020

Number of pages: 44

Number of figures: 10

Number of tables: 6

Number of appendix: 1 Number of appendix pages: 3

Number of references: 40

Mentor: Assist. Prof. Matjaž Hladnik, PhD

Co-Mentor: Assoc. Prof. Dunja Bandelj, PhD

Keywords: chloroplast DNA, microsatellite markers, Immortelle, GMATA, Krait, MISA

Abstract:

Microsatellite markers are recognized as one of the effective methods for plant species identification and phylogenetic research. This thesis presents an overview of studies on the topic of microsatellite markers from chloroplast DNA used for genetic research such as phylogeography, population studies, species identification, etc. An overview of the presence of microsatellites in the chloroplast DNA of other plant species is also presented. Using bioinformatics tools, we searched for microsatellite loci on the chloroplast genome of immortelle and developed primers for the amplification and further evaluation of loci to determine if they are suitable for use as molecular markers for genetic analysis. To identify SSR loci in the chloroplast genome, we used bioinformatics tools GMATA, MISA and Krait, and compared the results. In addition, sequences obtained from high-throughput sequencing of immortelle total DNA by UP FAMNIT researchers were aligned on the reference genome to identify mutations or differences in the number of SSR motif repeats, which could be due to the heteroplasmy.

ZAHVALA

Iskreno se zahvaljujem mentorju doc. dr. Matjažu Hladniku, somentorici izr. prof. dr. Dunji Bandelj, za vso spodbudo, podporo in pomoč pri izdelavi diplomske naloge in na celotni študijski poti.

Posebno se zahvaljujem družini in prijateljem, ki so mi ob izdelavi zaključne naloge stali ob strani in me spodbujali.

KAZALO VSEBINE

1	UVOD.....	1
2	PREGLED LITERATURE.....	3
2.1	Laški smilj.....	3
2.1.1	Izvor imena.....	3
2.1.2	Opis morfoloških značilnosti.....	3
2.1.3	Taksonomska uvrstitev in geografska razširjenost.....	4
2.1.4	Uporaba.....	4
2.1.5	Pomen smilja v pridelavi.....	5
2.2	Kloroplastna DNA.....	6
2.2.1	Sestava.....	6
2.2.2	Pomen za genetske analize.....	8
2.3	Ponovljiva zaporedja.....	9
2.3.1	Tandemska zaporedja.....	9
2.4	MIKROSATELITI.....	9
2.4.1	Klasifikacija mikrosatelitov.....	10
2.4.2	Mikrosateliti v kloroplastni DNA.....	10
3	METODE DELA IN REZULTATI.....	12
3.1	Metode dela.....	12
3.1.1	Identifikacija mikrosatelitskih regij v kloroplastni DNA laškega smilja.....	12
3.1.2	GMATA.....	12
3.1.3	Krait.....	13
3.1.4	MISA.....	14
3.1.5	Naleganje odčitkov.....	14
3.1.6	Identifikacija variabilnih mest.....	15
3.2	REZULTATI IN DISKUSIJA.....	16
3.2.1	Identifikacija mikrosatelitov.....	16
3.2.3	Začetni oligonukleotidi za namnoževanje lokusov s popolnimi mikrosateliti.....	21
3.2.3	PRIMERJAVA REZULTATOV IN ORODIJ.....	23
3.2.4	Genetska variabilnost kloroplastne DNA znotraj rastline.....	24
3.2.5	Analiza raznolikosti v kloroplastnem genomu.....	25
4	ZAKLJUČEK.....	26
5	LITERATURA IN VIRI.....	27

KAZALO PREGLEDNIC

Tabela 1: Podvrste in razširjenost <i>H. italicum</i> (Herrando-Moraira in sod. 2016).....	4
Tabela 2: Seznam mikrosatelitov, identificiranih z orodjem GMATA	17
Tabela 3: VNTR ponovitve	19
Tabela 4: Začetni oligonukleotidi za pomnoževanje 12 mikrosatelitskih lokusov.....	21
Tabela 5: Primerjava identificiranih SSR orodij	23
Tabela 6: Primerjava rezultatov identificiranih mononukleotidnih SSR z orodjema Krait in MISA.....	23

KAZALO SLIK IN GRAFIKONOV

Slika 1: Socvetje laškega smilja, vir: Ninčević in sod. (2019).....	3
Slika 2: Naravno rastišče laškega smilja na otoku Cres (Ninčević in sod. 2019)	5
Slika 3: Nasad laškega smilja (Ninčević in sod. 2019).....	6
Slika 4: Genska mapa kloroplastnega genoma laškega smilja	7
Slika 5: Privzete nastavitve v orodju Krait.....	14
Slika 6: Delež SSR glede na dolžino in nukleotidno zaporedje osnovnega motiva mikrosatelitov	18
Slika 7: Porazdelitev SSR glede na dolžino vseh SSR	18
Slika 8: JBrowse pregledovalnik	20
Slika 9: Primer določitve priležnih mest začetnih oligonukleotidov	22
Slika 10: Analiza raznolikosti v kloroplastnem genomu s pregledovalnikom Tablet	25

KAZALO PRILOG

Priloga A: Identifikacija nepopolnih SSR v orodju Krait

SEZNAM KRATIC

bp - bazni par

kbp - kilo baznih parov

°C - stopinja Celzija

DNA - deoksiribonukleinska kislina

cpDNA - kloroplastna deoksiribonukleinska kislina

PCR - verižna reakcija s polimerazo (angl. Polymerase Chain Reaction)

RNA - ribonukleinska kislina

rRNA - ribosomska ribonukleinska kislina

tRNA - prenašalna ribonukleinska kislina

RUBISCO - ribuloza-1,5-bifosfat karboksilaza/oksigenaza

SNP - polimorfizem posameznega nukleotida (angl. Single Nucleotide Polymorphism)

1 UVOD

Genetsko variabilnost lahko proučujemo z uporabo molekularnih tehnik, s katerimi neposredno analiziramo genom proučevanega organizma. V zadnjih dvajsetih letih je razvoj tehnologije za analizo DNA omogočil uporabo tovrstnih tehnik na številnih področjih, med drugim tudi pri proučevanju genetske variabilnosti prosto rastočih in kulturnih rastlin, genetske strukture populacij, identifikaciji rastlinskih vrst, itd. V ta namen so bili razviti različni molekularni markerji kot npr. AFLP (angl. amplified fragment length polymorphism) oz. polimorfizem dolžine namnoženih fragmentov, RAPD (angl. random amplification of polymorphic DNA) oz. tehnika naključno namnožene polimorfne DNA, RFLP (angl. restriction fragment length polymorphism) oz. polimorfizem dolžine restrikcijskih fragmentov, mikrosatelitski markerji (angl. simple sequence repeats, SSR), markerji SNP (angl. single nucleotide polymorphism) oziroma polimorfizem posameznega nukleotida, itd.

Mikrosateliti so enostavna tandemska ponovljiva zaporedja za katera je značilno, da so sestavljena iz krajših nukleotidnih zaporedij (do 250 bp), kjer se osnovni motiv ponavlja en za drugim. Prav zaradi tandemskih ponovitev, ki so pri podvajanju DNA bolj podvržene napakam, je za te lokuse značilno večje število alelov in so zato primerni tudi za razvoj molekularnih markerjev (Eisse 1999). Mikrosatelitski markerji se uporabljajo v genetskih raziskavah za študije raznolikosti, filogenetske analize sorodnosti, izdelavo genetskih kart in za genotipizacijo (Provan in sod. 2001).

Vsestranska uporabnost mikrosatelitov je možna zaradi visoke pogostosti pojavljanja v evkariontskih genomih, kodominantnosti, hipervariabilnosti, robustnosti in visoke informacijske vrednosti (Bandelj 2006).

Laški smilj (*Helichrysum italicum* (Roth) G. Don) je zdravilna rastlina s pozitivnimi farmakološkimi učinki. Uvrščamo jo v družino Asteraceae, pleme *Gnaphalieae* in je razširjena po celotnem Sredozemlju. Vrsta *H. italicum* je na podlagi morfoloških značilnosti, geografske razširjenosti in genetske variabilnosti razdeljena na štiri podvrste in sicer, *italicum*, *siculum*, *microphyllum*, *tyrrhenicum* (Herrando-Moraira in sod. 2016). Raziskovalci proučujejo genetsko variabilnost vrste, kar je izjemnega pomena za ohranjanje naravnih populacij, kot tudi za potrebe identifikacije genotipov za uporabo v farmacevtski industriji zaradi zdravilnih učinkov eteričnega olja, ki ga pridobivajo iz te rastline.

V okviru naloge smo analizirali mikrosatelite v kloroplastnem genomu laškega smilja (*Helichrysum italicum* (Roth) G. Don) ter pregledali študije, ki obravnavajo prisotnost

mikrosatelitov v kloroplastni DNA drugih vrst, molekulske markerje v kloroplastni DNA in primerjavo učinkovitosti le teh z mikrosatelitskimi markerji pri genetskih raziskavah kot npr. filogeografske študije, populacijske študije, identifikacija vrst, itd.

Za pomnoževanje mikrosatelitskih lokusov v verižni reakciji s polimerazo PCR (angl. Polymerase Chain Reaction) smo razvili tudi začetne oligonukleotide z uporabo različnih orodij.

Za identifikacijo SSR lokusov v kloroplastnem genomu smilja smo uporabili različna bioinformacijska orodja in sicer GMATA, MISA ter Krait ter rezultate med seboj primerjali. Zaradi različnih algoritmov za identifikacijo mikrosatelitov se namreč lahko pojavijo razlike pri rezultatih. Začetne oligonukleotide smo razvili s spletnim orodjem Primer3Web in z orodjem Krait. Z naleganjem odčitkov, ki so jih pridobili raziskovalci UP FAMNIT pri sekvenciranju genomske DNA laškega smilja (Hladnik in sod. 2019) na kloroplastni genom pa smo preverili morebitno prisotnost alelov z različnim številom ponovitev motiva mikrosatelita, kar bi lahko bila posledica heteroplazmije.

2 PREGLED LITERATURE

2.1 Laški smilj

2.1.1 Izvor imena

Ime rodu *Helichrysum* izvira iz grških besed "helios" in "chryos", kar v prevodu pomeni zlato sonce. Ime se navezuje na svetlo rumeno barvo socvetja (slika 1), kar je značilnost vrste *H. italicum*. Že od antičnih časov je ime povezano z brezčasnostjo in nesmrtnostjo (Perrini in sod. 2009). Stari Grki so ga zato uporabljali za okrasitev templjev, posvečenih bogovom. Imenovali so ga "amarantos" ali "večno mlad", saj so se očitno dobro zavedali njegove sposobnosti ohranjanja mladostnega videza kože. V nekaterih jezikih ime vrste še danes nosi prevod za nesmrtnost; "immortelle" v francoščini, "eternal" v angleščini. Zaradi vsestranskih bioloških aktivnosti in uporabe, se je zanimanje za znanstvene raziskave v zadnjih nekaj desetletjih povečalo.



Slika 1: Socvetje laškega smilja, vir: Ninčević in sod. (2019)

2.1.2 Opis morfoloških značilnosti

Laški smilj je aromatičen grm, visok 30-70 cm (Galbany-Casals in sod. 2011). Zanj je značilno, da ima močan in obstojen vonj, podoben curryju (Appendino in sod. 2007) ter rumene cvetove. Rastlina cveti med majem in junijem. Ima več vegetativnih stebel, ki so pokončna, listnata in dlakava vzdolž celotne dolžine. Stebla so dolga od 20 do 50 cm. Cvetočna stebela so podolgovata, trpka, dlakava, na zgornji strani zelenkasta ter na spodnji srebrnkasta (Bianchini in sod. 2009).

2.1.3 Taksonomska uvrstitev in geografska razširjenost

Vrsta je prilagojena za rast v sušnih (uvrščamo jo med kserofite), peščenih in kamnitih rastiščih na območju Sredozemlja. Ta lastnost omogoča rast laškega smilja do nadmorske višine vse do 2200 m (Galbany-Casals in sod. 2011).

Laški smilj uvrščamo v družino *Asteraceae*, pleme *Gnaphalieae* in jo lahko nadalje razdelimo na štiri podvrste, ki so razširjene v različnih regijah Sredozemlja (Herrando-Moraira in sod. 2016).

Tabela 1: Podvrste in razširjenost *H. italicum* (Herrando-Moraira in sod. 2016)

Vrsta	Razširjenost
<i>Helichrysum italicum</i> (Roth) G. Don subsp. <i>italicum</i>	Sredozemlje
<i>Helichrysum italicum</i> subsp. <i>microphyllum</i> (Willd.) Nyman	Kreta in Ciper
<i>Helichrysum italicum</i> subsp. <i>tyrrhenicum</i>	Balearski otoki (Majorka in Dragonera), Sardinija, Korzika
<i>Helichrysum italicum</i> subsp. <i>siculum</i> (Jord. & Fourr.) Galbany & al.	Sicilija

2.1.4 Uporaba

Laški smilj je dragocena zdravilna rastlina, ki se je v ljudskem zdravilstvu uporabljala že desetletja kot diuretik, za lajšanje kašlja, zdravljenju alergij, revmatizma, zaprtja in zdravljenju številnih drugih bolezni (Chinou in sod. 1996). Danes je najbolj znana zaradi učinkov pri obnavljanju in pomlajevanju kože, saj predstavlja dragocen vir naravnih bioaktivnih spojin in je tako prisotna v mnogih kozmetičnih izdelkih. Posebej je koristna pri zmanjševanju starostnih gub in pri celjenju ran. Zmanjšuje brazgotine, še posebej pa je primeren za odstranjevanje hematomov ali modric na telesu. Zaradi vsebnosti kariofilena ima protivnetni učinek, prav tako deluje protimikrobno in pomirjevalno. Zaradi značilnega vonja je pomembna tudi v industriji parfumov (Hellivan 2009). Cvetovi in listi rastline so najpogosteje uporabljeni deli za zdravljenje zdravstvenih motenj, kot so alergije, prehlad, kašelj, motnje kože, jeter in žolčnika, vnetja, okužbe in nespečnosti (Antunes in sod. 2014). Laški smilj je bogat z eteričnim oljem, ki se uporablja za različne medicinske namene. Da bi potrdili blagodejno delovanje ekstraktov, je bilo izvedenih več znanstvenih

raziskav. Rigano in sod. (2013) so z *in vivo* ter *in vitro* poskusi raziskovali učinek etanolnega izvlečka *H. italicum* in potrdili zdravilne učinke tradicionalne uporabe cvetov laškega smilja za zdravljenje črevesnih obolenj. Nostro in sod. (2001) so proučevali vpliv ekstrakta dietiletra na rast bakterij *Staphylococcus aureus* in ugotovili, da je ekstrakt deloval zaviralno na rast *S. aureus*, kar potrjuje protimikrobno delovanje spojin laškega smilja. V naslednji študiji so Staver in sod. (2018) na podlagi dobljenih rezultatov uporabe eteričnega olja v celičnih linijah Hela in MCD-7 zaznali omejeno širjenje rakavih celic, kar je rezultat apoptoze. Rezultati nakazujejo potencialno uporabo eteričnega olja laškega smilja kot vira alternativnih naravnih antibiotikov ali zdravil proti raku ter za preprečevanje in zdravljenje različnih bolezni.

2.1.5 Pomen smilja v pridelavi

Zaradi velikega povpraševanja po eteričnem olju in povečanega izkoriščanja divjih populacij *H. italicum*, se je gojenje nasadov močno razširilo v sredozemskih državah. Na podlagi številnih znanstvenih študij je bilo opravljenih več kemijskih analiz eteričnega olja. O gojenju vrste, kar vključuje zahteve glede setve, pridelave sadik, gostote rastlin, obiranje, bolezni in škodljivce in zatiranje plevelov je malo znanega. Številne raziskave se usmerjajo v genetsko strukturo in raznolikost te vrste, ki so izjemnega pomena za ohranjanje naravnih populacij in razvoja različnih strategij gojenja rastline (Ninčević in sod. 2019).



Slika 2: Naravno rastišče laškega smilja na otoku Cres (Ninčević in sod. 2019)



Slika 3: Nasad laškega smilja (Ninčević in sod. 2019)

2.2 Kloroplastna DNA

2.2.1 Sestava

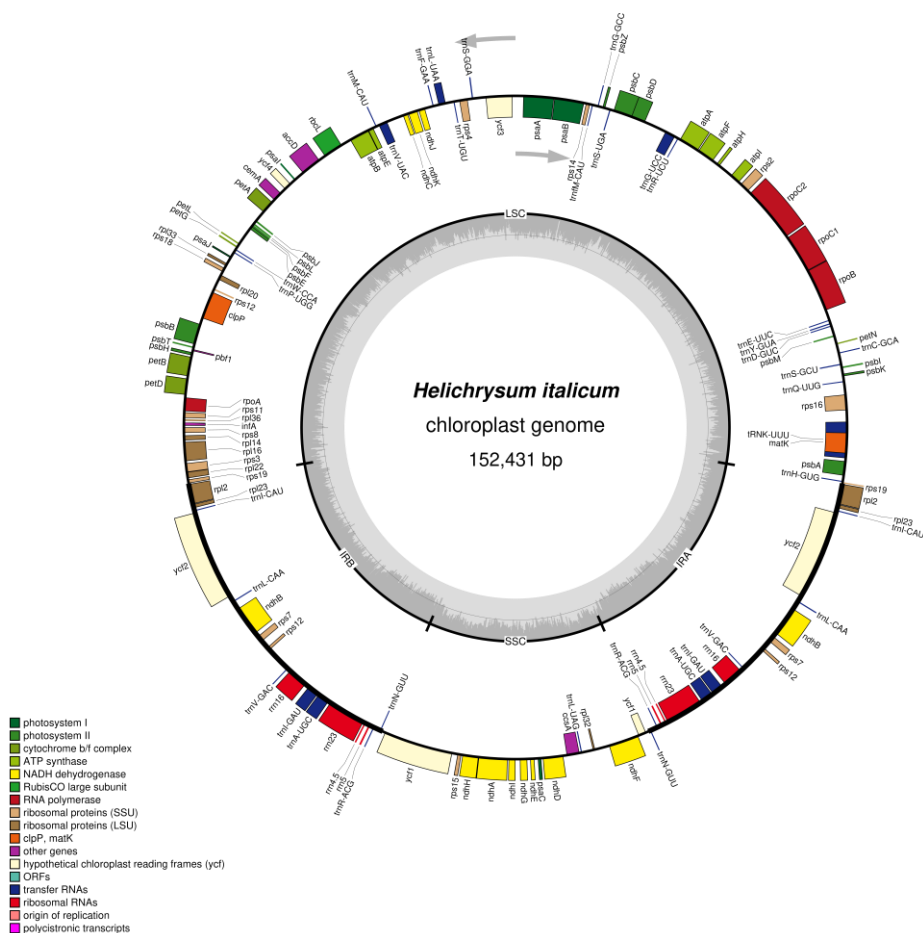
Pri večini rastlinskih vrst kloroplastni genom vsebuje dve inverzni oz. nasprotno orientirani ponovitvi (angl. inverted repeats ali IR), ki ločujeta daljšo kopijo (angl. large single copy ali LSC) od krajše kopije (angl. short single copy ali SSC). Velikost kloroplastne DNA se pri skoraj vseh vrstah nahaja v intervalu med 120 kbp – 160 kbp, z nekaj izjemami znotraj skupine zelenih alg, medtem ko imajo kopenske rastline kloroplasten genom velikosti od 120 - 271 kbp (Petit in Vandramin, 2007). Razlike v dolžini cpDNA se pojavljajo zaradi daljših ali krajših IR regij, ki povzročajo spremembe v razliki dolžine tudi do 70 kbp (Sugiura 1992).

Ločimo tri funkcijske kategorije, vključno s kodirajočimi geni, introni in medgenskimi presledki. Slednja dva ne kodirata proteinov in ju tako lahko uporabimo za raziskovanje rastlinskih filogenij, molekularne evolucije in populacijske genetike. Kloroplastni genomi nekaterih vrst sestojijo iz dveh ali več krožnih molekul DNA (npr. kloroplastni genom rjave alge *Pylaiella littoralis* ima dve krožni molekuli DNA v velikosti 133 kb in 58 kb). Kloroplastni genomi imajo v povprečju 120 genov.

V prvo skupino genov kloroplastnega genoma sodijo geni za rRNA za sestavo 70S ribosomov. Le-ti se namreč razlikujejo od 80S ribosomov, ki se nahajajo v citoplazmi. Naslednja skupina so geni za tRNA, za katere je značilno, da imajo v kloroplastnem genomu kar 20-40 genov tRNA, kar pomeni, da cpDNA vsebuje zapis za vse prenašalne RNA, potrebne za sintezo proteinov v kloroplastu. V zadnjo skupino uvrščamo gene za sintezo ribosomskih proteinov. Kloroplastni ribosom sestavlja 60 različnih proteinov, tretjina teh pa naj bi bila sintetizirana na podlagi zapisa v cpDNA. Prav tako kloroplastna DNA vsebuje zapis za različne transkripcijske faktorje ter gene za sintezo podenot RNA

polimeraze. Naslednja velika skupina so geni fotosinteznega sistema, kjer cpDNA vsebuje gene za sintezo nekaterih komponent kompleksa fotosistema I, fotosistema II, citokrom b/f kompleksa in kompleksa ATP sintaze. Kot zadnja skupina genov cpDNA so *ndh* geni. Število kopij kloroplastnega genoma je v celicah lista 10 000 ali več (Sugiura 1992).

Hladnik in sod. (2019) so določili celotno nukleotidno zaporedje kloroplastnega genoma *H. italicum* iz območja Severnega Jadrana in prišli do sledečih zaključkov; velikost genoma je 152 431 bp, ki vsebuje 85 kodirajočih genov, 36 genov za tRNA, 8 genov za rRNA ter 2 delna gena (*ycf1* in *rps19*). Delež posameznih nukleotidov v kloroplastnem genomu je sledeč: 31,3 % A, 18,3 % C, 18,8 % G in 31,5 % T. Dolžine LSC, SSC in IR regij so bile 84.445 bp, 18.440 bp in 24.773 bp (slika 4).



Slika 4: Genska mapa kloroplastnega genoma laškega smilja. Geni na notranjem krogu so prepisani v smeri urinega kazalca, tisti na zunanem krogu se prepisujejo v nasprotni smeri urinega kazalca. Različne funkcije genov so prikazane v različnih barvah. Notranji krog označuje IR regiji (IRa in IRb), ki ločujeta SSC in LSC kopiji (izrisano z aplikacijo OGDRAW).

2.2.2 Pomen za genetske analize

Kloroplast je najbolj pogosto analiziran rastlinski genom za raziskave molekularne evolucije in sistematike (Palmer in sod. 1988). Za kloroplastno DNA je značilno citoplazemsko dedovanje in podvajanje s kloniranjem. Zato je primerna za razvoj markerjev za identifikacijo vrst, analizo raznolikosti in variabilnosti rastlin in filogenetske študije.

Na primer, Dong-Hyuk in sod. (2017) so primerjali kloroplastna genoma vrst *Anaphalis sinica* in *Leontopodium leirolepis* iz plemena *Gnaphalieae* s kloroplastnimi genomi drugih vrst iz družine *Asteraceae*. Zaradi razlik med kloroplastnimi genomi (izguba trnT-GGU) so opredelili pleme *Gnaphalieae* kot samostojno skupino (trnT-GGU gen je zaradi tega koristen filogenetski marker za raziskovanje filogenije in evolucije plemena *Gnaphalieae*). Iz razlik med genomi kloroplastov je mogoče proučevati tudi evolucijsko ločitev vrste znotraj *Asteraceae*.

Proučevanje kloroplastne DNA je pomembno tudi s stališča žlahtnjenja rastlin. Npr. Amiryousefi in sod. (2018) povzemajo, da imajo herbicidi, kot sta PSI in zaviralci PSII, tarčne gene v genomu kloroplasta, zato je razumevanje kloroplastnega genoma ključnega pomena pri raziskovanju odpornosti na herbicide in razvoju novih metod. Enako velja tudi za lastnosti, s katerimi se rastline prilagajajo na stresne dejavnike, itd.

Uporabnost kloroplastne DNA za identifikacijo vrst so pokazali Rebollo in sod. (2017), ki so vzpostavili učinkovito metodo za preverjanje pristnosti oljčnega olja na osnovi kloroplastnega gena *trnL* (UAA). Priležna mesta začetnih oligonukleotidov so bila zasnovana na območju introna gena tRNA *trnL* (UAA). Za identifikacijo vrste v predelanih živilih je kloroplastna DNA bolj primerna, ker je prisotna v večjem številu kopij.

Pri proučevanju izvora žlahtne vinske trte so Arroyo-Garcia in sod. (2006) analizirali variabilnost klorotipov *Vitis vinifera* ssp. *sylvestris* in *Vitis vinifera* ssp. *sativa* s celotnega območja razširjenosti vrste (Iberijski polotok, Srednja Evropa, Severna Afrika, Apeninski polotok, Balkanski polotok, Vzhodna Evropa in Bližnji vzhod). Genetsko analizo so izvedli s kloroplastnimi mikrosatelitskimi markerji in ugotovili, da naj bi obstajala vsaj dva pomembna centra izvora, in sicer na Bližnjem vzhodu ter v zahodnem Sredozemlju. Do

omenjenih ugotovitev so prišli na osnovi rezultatov, da je več kot 70 % sort iz Iberskega polotoka imelo klorotip zahodnih populacij *Vitis vinifera* ssp. *sylvestris*.

2.3 Ponovljiva zaporedja

Velik del človeškega genoma in drugih evkariontov sestavljajo ponovljiva zaporedja. Te delimo v dve kategoriji, in sicer ločimo razpršena zaporedja in tandemska zaporedja. Za razpršena zaporedja (ang. interspersed repeats) je značilno, da se ista zaporedja pojavljajo razpršeno po celotnem genomu, medtem ko se pri tandemsko ponovljivih zaporedjih (ang. tandemly repeated DNA) enako nukleotidno zaporedje ponavlja en za drugim (Brown 2018).

2.3.1 Tandemska zaporedja

Poznamo različne tipe satelitne DNA, in sicer jih ločimo glede na dolžino osnovnega motiva na tri skupine (Brown, 2018):

- Mikrosateliti (angl. simple sequence repeats, SSR) ali kratke tandemske ponovitve (angl. short tandem repeats, STR): osnovni motiv je krajši, na lokusu tvori ponovitve skupne dolžine do 150 bp, s ponavljajočim motivom do 13 baznih parov ali manj,
- Minisateliti ali regije DNA s spremenljivim številom tandemskih ponovitev (angl. variable number of tandem repeats, VNTRs): na posameznem lokusu tvori ponovitve do dolžine 20 kbp, s ponavljajočim motivom do 25 baznih parov,
- Sateliti: predstavljajo najdaljše zaporedje, sestavljeni iz motiva dolžine do 200 bp.

2.4 MIKROSATELITI

Za mikrosatelite je značilno, da so lokusno specifični in je detekcija alelov oz. fragmentov namnoženih v reakciji PCR in ločenih z elektroforezo nezahtevna. Mikrosateliti so kodominantni markerji, kar omogoča detekcijo različnih alelov oziroma ugotavljanje homozigotnosti in heterozigotnosti. SSR regije so visoko polimorfne, saj imajo različno število ponovitev osnovnega motiva. Posledica tega so številne mutacije, predvsem insercije in delecije (Eisse 1999). Prisotnost različnih alelov mikrosatelitskih lokusov na kloroplastni DNA je mogoča in je posledica heteroplazmije. Prisotnost heteroplazmije so naprimer potrdili Sun in sod. (2019) pri vrsti *Gentiana tongolensis*. Vzrok tega ni popolnoma jasen, vendar predvidevajo, da se je heteroplazmija pojavila zaradi

dvostarševskega dedovanja. Kloroplastni genomi so ponavadi podedovani po materini strani in so haploidni v večini kritosemenk.

2.4.1 Klasifikacija mikrosatelitov

Bandelj (2006) povzema način delitve mikrosatelitov glede na tip ponovitve osnovnega motiva in sicer na naslednjih šest razredov:

- popoln mikrosatelit (angl. pure microsatellite) je sestavljen samo iz enega motiva baz, ki se tandemsko ponavlja in ni prekinjen z nobeno drugo bazo, npr. (AC)₁₄
- sestavljen mikrosatelit (angl. compound microsatellite) je sestavljen iz vsaj dveh različnih osnovnih motivov ponovitve, npr. (CT)₂₂(CA)₆
- popoln in prekinjen oz. nepopolni mikrosatelit (angl. interrupted pure microsatellite) ima osnoven motiv prekinjen z insercijo enega nukleotida ali več baznih parov, ki se ne ujemajo z osnovnim motivom, npr. (CA)₄TA(CA)₇
- sestavljen in prekinjen mikrosatelit (angl. interrupted compound microsatellite) ima poleg vsaj dveh različnih osnovnih motivov še krajšo insercijo baznih parov, ki se razlikujejo od osnovnih motivov, npr. (AC)₁₄AGAA(AG)₁₂
- kompleksen mikrosatelit (angl. complex microsatellite) je širši izraz za popolne in sestavljene mikrosatelite, ki nastanejo zaradi insercij baz, ki predstavljajo kratko ponovitev, npr. (TC)₄(T)₆(CT)₄CTCC(TCC)₆

2.4.2 Mikrosateliti v kloroplastni DNA

Dong-Hyuk in sod. (2017) so analizirali SSR v kloroplastnem genomu *A. sinica* in *L. leiolepis* in našli 63 SSR v kloroplastnem genomu *A. sinica* (38 mono-, 7 di-, 9 tri- in 9 tetranukleotidov) ter 81 SSR v genomu *L. leiolepis* (44 mono-, 12 di-, 9 tri- in 16 tetranukleotidov).

Amiryousefi in sod. (2018) so sekvencirali celoten genom kloroplasta *Solanum dulcamara*, ki je velik 155 580 bp, sestavljen iz 112 genov, od katerih je 81 protein kodirajočih genov, 27 tRNA in štiri rRNA. Z uporabo orodja MISA so identificirali 374 SSR ponovitev v genomu, od katerih je bilo 253 mono-, 40 di-, 70 tri-, 10 tetra- in en je bil pentanukleotid. Odkrili so, da je bilo več SSR v regijah LSC in SSC v primerjavi z IR. Identificirali so tudi 25 večjih ponovitev (nad 10 bp). Najbolj variabilna regija je bila trnE-UUC — trnT-GGU IGS.

Shen in sod. (2018) so določili celoten kloroplastni genom zdravilne rastline *Aster tataricus*, ki se uporablja za zdravljenje oz. lajšanje kašlja. Genom *A. tataricus* kloroplasta je velik 152.992 bp in vsebuje obrnjeni ponovitvi (IRa in IRb, vsaka dolžine 24.850 bp). V kloroplastnem genomu *A. tataricus* so odkrili 115 genov, vključno z 81 kodirajočimi geni proteinov, 4 gene za ribosomske RNA in 30 genov za prenašalno RNA. Rezultate so primerjali še z genomi drugih dveh vrst *Aster altaicus* in *Aster spathulifolius*, in ugotovili, da je genom *A. tataricus* največji, medtem ko so si genomi po sestavi podobni. Najmanj razlik so ugotovili med kloroplastno DNA vrst *A. tataricus* in *A. altaicus*. V kloroplastnem genomu *A. tataricus* so našli 70 SSR ponovitev, v genomu *A. altaicus* 58 SSR ponovitev in v *A. spathulifolius* 36 SSR ponovitev, med katerimi je bilo največ mononukleotidnih ponovitev.

George in sod. (2015) so analizirali mikrosatelite v kloroplastnem genomu nižje in višje uvrščenih rastlin. Pri tem so identificirali enostavne in sestavljene mikrosatelite v 164 kloroplastnih genomih številnih rastlin. V raziskavi so ugotovili, da: a) je v vseh genomih prisotno večje število mononukleotidnih SSR v primerjavi z di-, tri-, tetra-, penta- in heksanukleotidnimi ponovitvami; b) nižje uvrščene rastline, kot na primer alge, kažejo večje razlike v številu, gostoti in porazdelitvi mikrosatelitskih ponovitev v primerjavi s kritosemenkami; c) daljših SSR ponovitev ni prisotnih v kodirajočih regijah pri večini rastlin. Pri kritosemenkah so npr. identificirali od 516 (*Acorus americanus*) do 892 (*Arabidopsis thaliana*) mononukleotidnih (daljših od 6 bp) SSR ponovitev. Pri genomih alg se je število mononukleotidnih SSR (daljših od 6 bp) gibalo od 136 (*Eutreptiella gymstica*) do 3342 (*Stigeoclonium helveticum*).

Sonah in sod. (2011) so pregledali zaporedja šestih kloroplastnih genomov iz rodu *Vigna* za identifikacijo popolnih SSR. Skupno je bilo določenih 21, 24 in 25 kloroplastnih SSR ponovitev v vrstah *Vigna angularis*, *Vigna radiata* in *Vigna unguiculata*. Dolžina lokusov SSR je bila odvisna od različnih enot ponovitve od 12 do 24 bp. Najpogostejše so bile mononukleotidne ponovitve, ki so jim sledile dinukleotidne ponovitve pa vse do pentanukleotidnih ponovitev. Heksanukleotidnih ponovitev v genomih kloroplasta iz rodu *Vigna* niso identificirali. Molekulski markerji, ki so primerni za analizo pri različnih vrstah so uporabni za kartiranje genomov in primerjalno kartiranje med vrstami iz rodu *Vigna*.

3 METODE DELA IN REZULTATI

3.1 Metode dela

3.1.1 Identifikacija mikrosatelitskih regij v kloroplastni DNA laškega smilja

Za identifikacijo mikrosatelitskih regij v kloroplastni DNA smo uporabili nukleotidno zaporedje kloroplastnega genoma, ki so ga objavili Hladnik in sod. (2019) in je dosegljivo v podatkovni zbirki GenBank z akcesijsko oznako NC_041458. Celotno zaporedje kloroplastnega genoma je dolžine 152.431 bp. Kloroplastni genom smilja smo analizirali s tremi bioinformacijskimi orodji za identifikacijo mikrosatelitskih regij, in sicer GMATA v. 3 (Wang in Wang 2016), MISA v. 3 (Beier in sod. 2017) ter Krait v. 2 (Du in sod. 2018).

3.1.2 GMATA

Programski paket GMATA je namenjen za identifikacijo SSR lokusov in razvoj začetnih oligonukleotidov. V programu lahko nastavimo parametre za iskanje mikrosatelitov in sicer dolžino motiva ter minimalno število ponovitev le tega. Za analizo smo uporabili privzete nastavitve in sicer dolžina mikrosatelitskega motiva od 2 do 6, število ponovitev motiva pa vsaj petkrat (npr. ATATATATAT). Program ima razvit tudi grafični uporabniški vmesnik, vendar smo analizo izvedli s pomočjo ukazov v ukazni vrstici operacijskega sistema Ubuntu. Program je na voljo na spletnem portalu Github (<https://github.com/XuwenWangUGA/GMATA>).

Za identifikacijo mikrosatelitov smo uporabili naslednji ukaz:

```
$ perl gmat.pl -i <pot do fasta datoteke> -r 5 -m 2 -x 6 -s 0
```

Parameter `-r` pomeni minimalno število ponovitev osnovnega motiva, `-m` najkrajšo dolžino osnovnega motiva ter `-x` najdaljšo dolžino osnovnega motiva. S parametrom `-s` pa lahko v preglednico z identificiranimi SSR vključimo še zaporedje mikrosatelita in robne regije. Z oznako `1` pridobimo originalno DNA zaporedje z malimi črkami ter ponavljajoče se zaporedje v velikimi črkami. Po zaključenem prvem koraku analize program oblikuje preglednico s šestimi stolpci, in sicer ime zaporedja v katerem je bil identificiran mikrosatelit (v našem primeru je to bilo samo zaporedje kloroplastnega genoma), dolžina zaporedja, prvo mesto mikrosatelita, zadnje mesto mikrosatelita in osnovni motiv mikrosatelita.

Orodje GMATA omogoča tudi grafični prikaz rezultatov, ki ga lahko pridobimo z ukazom:

```
$ perl ssrfig.pl -i <pot do datoteke s končnico .ssr.sat2> -  
f1 10 -f2 10 -f3 20
```

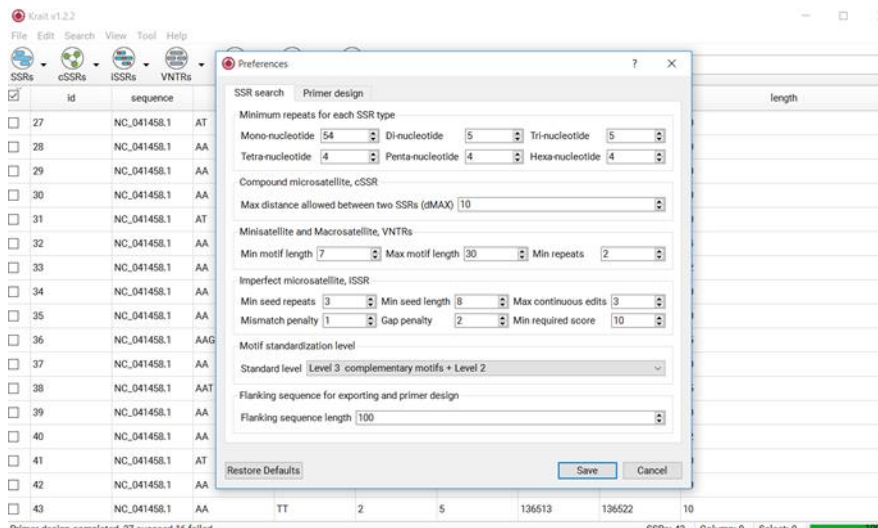
V zgornjem ukazu parameter `-f1` določi do katere dolžine oz. k-mere želimo z diagramom prikazati delež pogostosti pojavljanja, `-f2` določi koliko najpogosteje prisotnih osnovnih motivov želimo prikazati z diagramom, `-f3` parameter pa določi število združenih motivov (glede na komplementarnost).

Za vizualni pregled porazdelitve mikrosatelitov v genomu s pomočjo genomskega brskalnika JBrowse smo z orodjem GMATA izdelali tudi datoteko z obliko zapisa GFF3 (angl. General feature format), ki vsebuje podatke o začetnem in končnem mestu mikrosatelitskih lokusov, na kateri verigi DNA je identificirano zaporedje prisotno, itd. Ukaz za pridobitev gff3 datoteke:

```
$ perl ssr2gff3.pl -i <pot do datoteke z rezultati,  
pridobljene v prvem koraku s končnico .fa.ssr>
```

3.1.3 Krait

Krait je naslednje orodje, ki smo ga uporabili za identifikacijo mikrosatelitov. Omogoča identifikacijo popolnih, nepopolnih in sestavljenih SSR kot tudi regije DNA s spremenljivim številom tandemskih ponovitev (VNTR). Začetne oligonukleotide oblikuje z uporabo Primer3 (Untergasser in sod. 2012), ki je integriran v Krait. Za identifikacijo dinukleotidnih mikrosatelitov smo določili minimalno 5 ponovitev osnovnega motiva, zato da so bile nastavitve poenotene z orodjem GMATA. Začetne oligonukleotide smo določili z uporabo privzetih nastavitvev. Določili smo le zeleno dolžino produkta PCR in sicer od 100 do 400.



Slika 5: Privzete nastavitve v orodju Krait. Slika prikazuje uporabo privzetih nastavitvev v orodju Krait, kjer smo določili najmanjšo in najdaljšo število ponovitev za vsako SSR zaporedje.

3.1.4 MISA

MISA omogoča identifikacijo popolnih in sestavljenih mikrosatelitov. Za analizo smo določili dolžino mikrosatelitskega motiva od 2 do 6 in minimalno 5 ponovitev osnovnega motiva. Začetni oligonukleotidi za namnoževanje lokusov SSR so bili zasnovani z uporabo spletnega orodja Primer3Web (<http://primer3.ut.ee/>).

3.1.5 Naleganje odčitkov

Morebitno prisotnost razlik v številu ponovitev na mikrosatelitskih lokusih smo preverili z naleganjem odčitkov DNA kloroplastnega genoma z orodjem BWA v. 0.7.17 (Li in Durbin 2009). Odčitke kloroplastne DNA so predhodno pridobili Hladnik in sod. (2019) s sekvenciranjem DNA laškega smilja (v nadaljevanju podatkovni_set.fastq). Število vseh zaporedij v podatkovnem setu je bilo 17 025 150 (odčitki z jedrne DNA in kloroplastne DNA).

Naleganje odčitkov na referenčni genom kloroplasta smo smo izvedli z naslednjim ukazom:

Izdelava indeksa pred naleganjem z referenčnim genomom kloroplasta:

```
$ bwa index -a is \ # izbor algoritma za izdelavo FM-  
indeksa (primeren za krajše referenčne genome)  
-p BWA_ \ # predpona za izhodne datoteke  
NC_041458.fasta \ # referenčni genom kloroplastne DNA
```

Naleganje odčitkov :

```
$ bwa mem \ # izbor BWA-MEM algoritma za naleganje  
-t 30 \ # število niti  
podatkovni_set.fastq > BWA_rezultat.sam
```

Kot rezultat smo dobili datoteko v formatu SAM (angl. Sequence alignment map), ki jo je potrebno za nadaljnjo analizo pretvoriti v format BAM (angl. Binary alignment map). V datoteki SAM je zapisano na katerem mestu v referenčnem genomu je nalegel posamezen odčitek. Za pretvorbo datoteke SAM v BAM smo uporabili ukaz:

```
$ samtools view -S -b BWA_rezultat.sam > BWA_rezultat.bam
```

Parameter -S je potreben v primeru uporabe datoteke SAM, -b pa določi BAM format za izhodno datoteko.

3.1.6 Identifikacija variabilnih mest

3.1.6.1 Ročni pregled poravnave na mestih z mikrosateliti z orodjem Tablet

Datoteko BAM smo uporabili za ročno pregledovanje nalegih odčitkov na mikrosatelitske lokuse z orodjem Tablet (verzija 1.19.09.03, <https://doi.org/10.1093/bib/bbs012>). Pred tem smo odčitke v datoteki BAM najprej razporedili po vrstnem redu glede na prileganje na referenčni genom (angl. sorted BAM). Orodje Tablet poleg datoteke BAM potrebuje še indeks (datoteka s končnico ».bai«) in referenčno zaporedje. Razvrščanje odčitkov in izdelavo indeksa smo izvedli z naslednjimi ukazi:

- razporeditev nalegih odčitkov po vrstnem redu glede na referenčni genom

```
$ samtools sort -@ 10 BWA_rezultat.bam  
BWA_rezultat_sorted.bam
```

S parametrom `-@` določimo število niti.

- Izdelava indeksa

```
$ samtools index BWA_rezultat_sorted.bam
```

3.1.6.2 Identifikacija morebitnih razlik v genomu kloroplasta z orodjem bcftools

Orodje bcftools je namenjeno za identifikacijo alelnih variant oz. različic lokusov.

Analizo smo izvedli z naslednjim ukazom:

```
$ bcftools mpileup -Ou -f NC_041458.fasta BWA_rezultat_sorted.bam |  
bcftools call -mv -Ob --ploidy 1 -o calls.bcf
```

V zgornjem ukazu parameter `mpileup` izračuna verjetnost genotipa na vsakem mestu v genomu, `-f` določa vhodno datoteko (v našem primeru referenčno zaporedje kloroplastnega genoma) pomeni preskočitev mest, kjer stolpci ne vsebujejo sortiranih nizov. Parameter `call` na osnovi rezultatov analize z `mpileup` parametrom `v` v datoteko BCF izpiše različice (SNP, insercije, delecije) lokusov. V drugem delu ukaza smo s parametrom `-m` izbrali privzeto metodo za določitev možnih genotipov, s parametrom `-v` smo določili, da v izhodni datoteki BCF želimo le podatek za polimorfna mesta

Kot rezultat smo pridobili datoteko v binarni obliki (BCF), ki jo je za lažje pregledovanje potrebno pretvoriti v obliko VCF. Datoteko v obliki VCF lahko odpremo z urejevalnikom besedil ali npr. z MS Excel.

Pretvorbo iz zapisa bcf v vcf smo izvedli z ukazom `bcftools view`:

```
$ bcftools view calls.bcf > calls.vcf
```

3.2 REZULTATI IN DISKUSIJA

3.2.1 Identifikacija mikrosatelitov

3.2.1.1 GMATA

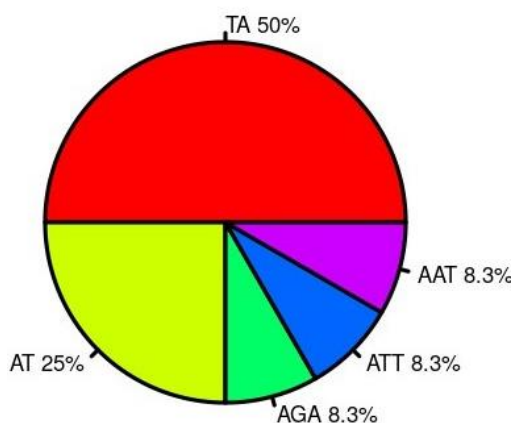
Z orodjem GMATA smo v kloroplastnem genomu laškega smilja identificirali 12 mikrosatelitskih regij (tabela 2, slika 6 in 7). Od tega 3 mikrosatelite s trinukleotidnim motivom (ATT, AGA, AAT) ter 9 mikrosatelitov z dolžino osnovnega motiva 2 bp (TA in

AT). Število identificiranih mikrosatelitov je manjše v primerjavi s številom SSR lokusov pri vrstah *L. leiolepis* in *A. sinica* iz družine *Asteraceae*. Pri *L. leiolepis* je bilo najdenih 25 (brez mononukleotidnih) lokusov SSR od tega 7 di-, 9 tri- in 9 tetranukleotidnih ponovitev), medtem ko so pri vrsti *A. sinica* določili 37 lokusov SSR (12 di-, 9 tri- in 16 tetranukleotidov).

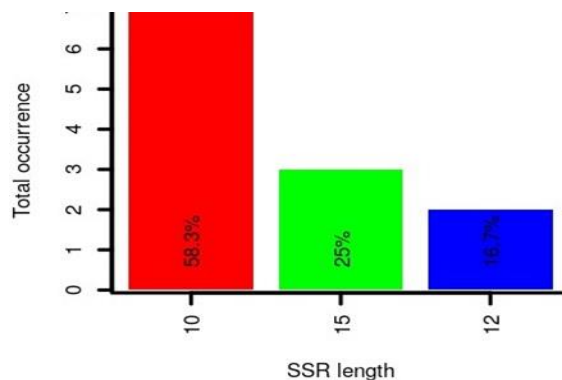
Tabela 2: Seznam mikrosatelitov, identificiranih z orodjem GMATA

	Zap. Št. SSR	začetek	konec	ponovitev	motiv	Dolžina
	1	2063	2077	5	ATT	15
	2	18669	18678	5	TA	10
	3	19665	19674	5	AT	10
	4	36185	36194	5	AT	10
	5	41628	41639	6	TA	12
	6	42072	42083	6	AT	12
	7	46438	46447	5	TA	10
	8	67942	67951	5	TA	10
	9	77234	77243	5	TA	10
	10	109291	109305	5	AGA	15
	11	111479	111493	5	AAT	15
	12	123336	123345	5	TA	10

V tabeli 2 je za vsak identificiran lokus SSR zapisana dolžina mikrosatelita, začetek in konec ponovitve glede na referenčni genom, število ponovitev motiva ter zaporedje motiva. V krožnem diagramu je prikazan delež identificiranih SSR lokusov glede na zaporedje osnovnega motiva, motiv TA predstavlja polovico vseh SSR sekvenc. Sledi motiv AT s četrtinskim deležem in trinukleotidni motivi. Glede na rezultate in primerjave z drugimi študijami, je motiv TA pri večini najbolj pogost.



Slika 6: Delež SSR glede na dolžino in nukleotidno zaporedje osnovnega motiva mikrosatelitov



Slika 7: Porazdelitev SSR glede na dolžino vseh SSR

3.2.1.2 Krait

Z orodjem Krait smo prav tako pridobili isti seznam 12 popolnih mikrosatelitov, kot je prikazan v tabeli 2. V primerjavi z orodjem GMATA ima orodje Krait možnost iskanja sestavljenih mikrosatelitov, vendar jih v genomu kloroplasta ni identificiral. Pridobili pa smo 162 nepopolnih SSR (66 mononukleotidnih, 28 dinukleotidnih, 53 trinukleotidnih ter 15 štirinukleotidnih). Za vsak nepopolni SSR je v prilogi A podano število substitucij, insercij in delecij.

V tabeli 3 so navedene ponovitve VNTR. Z orodjem Krait smo identificirali 32 VNTR zaporedij v dolžini od 14 do 54 ter od števila ponovitev od 2 do 3. Dolžina motiva, ki se ponavlja pa je od 7 do 26.

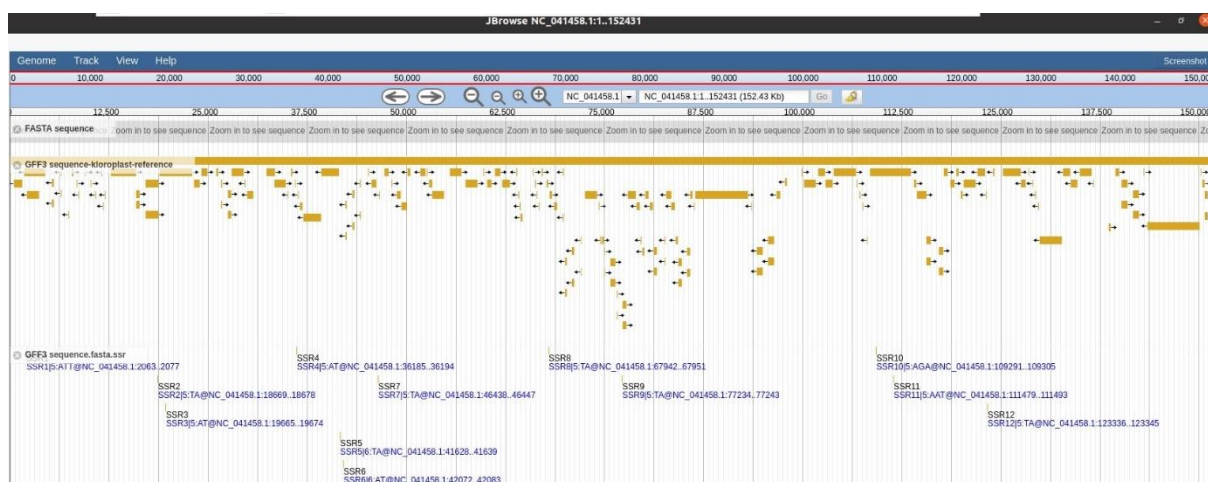
Tabela 3: VNTR ponovitve

Id	Motiv	Dolžina motiva	Ponovitev	Začetek	Konec	Dolžina
1	CTACATA	7	2	1948	1961	14
2	GGAGAATA	8	2	4936	4951	16
3	TAGGAAAGGAATCCATATAAAATAAA	26	2	5616	5667	52
4	AAAAAATT	8	2	21161	21176	16
5	TAATAAA	7	2	25325	25338	14
6	GATATTA	7	2	25630	25643	14
7	GAGTTATAGT	10	2	31225	31244	20
8	TTGTAAAA	8	2	35286	35301	16
9	ATTTTTTT	8	2	44086	44101	16
10	TCTCTAAT	8	2	45408	45423	16
11	TCCTATA	7	2	46487	46500	14
12	CTAATT	7	2	48693	48706	14
13	CTTTATT	7	2	50985	50998	14
14	CACTAGA	7	2	51469	51482	14
15	GAAAATC	7	2	59114	59127	14
16	TTTGGTATA	9	2	59898	59915	18
17	TAGAACG	7	2	63396	63409	14
18	AAAATAG	7	2	65561	65574	14
19	CTCTTTT	7	2	67284	67297	14
20	TAAAAAA	7	2	74357	74370	14
21	TTATACAT	8	2	84024	84039	16
22	GGAACATTT	9	2	88785	88802	18
23	ATTGATGCTAGTGACGAT	18	3	91321	91374	54
24	CCAGTTAG	8	2	94208	94223	16
25	TAGTTAA	7	2	99350	99363	14
26	TCAAGAACG	9	2	103105	103122	18
27	GAGAATT	7	2	113397	113410	14
28	ATTA ACT	7	2	137311	137324	14
29	AATTAAGAA	9	2	137671	137688	18
30	ATAATCT	7	2	142528	142541	14
31	TCGTC ACTAGCATCAATA	18	3	145302	145355	54
32	AAATGTTCC	9	2	147873	147890	18

3.2.1.3 MISA

Tudi z orodjem MISA smo pridobili enako število mikrosatelitskih lokusov kot pri orodjih GMATA in Krait (tabela 2). Tako kot Krait ima tudi MISA možnost identifikacije sestavljenih mikrosatelitov. Pri orodjih GMATA in Krait smo z nastavitvami omejili iskanje mikrosatelitov z dolžino motiva vsaj dveh bp. Ker pa smo v literaturi zasledili, da so v kloroplastnih genomih pogosto prisotni mononukleotidni lokusi, bi želeli predstaviti podatek, da smo z orodjem MISA identificirali 31 mononukleotidnih SSR ponovitev (tabela 6).

3.2.2 Lokacije popolnih SSR v genskih in medgenskih regijah



Slika 8: JBrowse pregledovalnik

Z Jbrowse pregledovalnikom smo preverili ali so SSR v genskih ali medgenskih regijah. Ugotovili smo, da SSR1 (trinukleotidno zaporedje z osnovnim motivom ATT) leži v regiji gena *trnK-UUU*. SSR2 (dinukleotidni mikrosatelit z osnovnim motivom TA) leži v genu *rpoC1*, SSR3 (dinukleotidni mikrosatelit z osnovnim motivom AT) pa leži v regiji gena *rpoC2*. Mikrosateliti SSR4, SSR5, SSR6, SSR7 in SSR8 ležijo v medgenskih regijah. SSR9 (dinukleotidni mikrosatelit z osnovnim motivom TA) leži v regiji introna gena *petD*. SSR10 (trinukleotidni mikrosatelit z osnovnim motivom AGA) in SSR11 ležita v regiji gena *ycfI*, zadnja SSR12 ponovitev pa se nahaja v medgenski regiji.

Iz rezultatov lahko sklepamo, da trinukleotidna zaporedja z motivi ATT, AGA in AAT ležijo v kodirajočih regijah genov. Prav tako ležita v kodirajočih regijah dinukleotidni zaporedji z motivoma TA in AT. Vse ostale SSR ponovitve pa najdemo v medgenskih regijah. Mikrosateliti SSR1 do SSR9 ležijo v kopiji LSC, SSR10 do SSR12 pa v SSC.

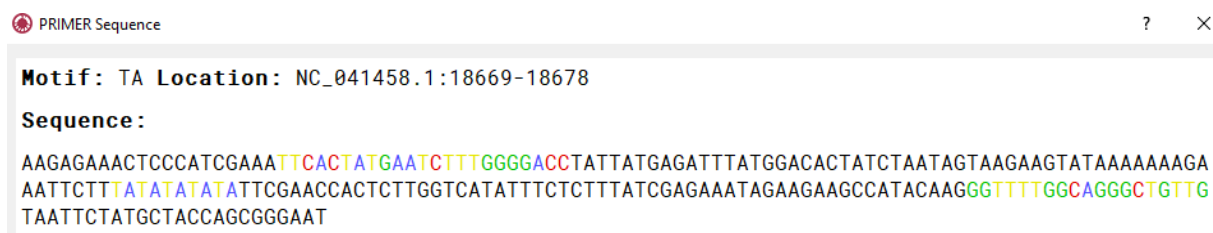
3.2.3 Začetni oligonukleotidi za namnoževanje lokusov s popolnimi mikrosateliti

V tabeli 4 so navedeni začetni oligonukleotidi za pomnoževanje 12 mikrosatelitskih lokusov, ki so bili določeni s Primer3Web in Krait.

Tabela 4: Začetni oligonukleotidi za pomnoževanje 12 mikrosatelitskih lokusov

Zap. št.	Uporabljeno orodje za razvoj primerjev	Smerni začetni oligonukleotid	Tm	GC	Protismerni začetni oligonukleotid	Tm	GC	Dolžina produkta PCR
SSR 1	Primer3Web	ACACGGCTTCCCTCTGTAT	58.72	50	TCGCGAAGGTTAICTAGAAAGGA	58.45	45.45	328
SSR 2	Krait	TTCACATGAATCTTTGGGGACC	58.14	43.48	CAAACAGCCCTGCCAAAACC	59.93	57.89	166
	Primer3Web	CAACGGTATTTGCTGCAAG	59.02	50	AACCCGGTGAGACTCGAATT	59.0	50	244
SSR 3	Krait	ATGGAATGATGGCGGACAGG	60.18	55	TCTATTGACAAAGCCCAACCCC	59.99	52.38	128
	Primer3Web	CGAAATGGAATGATGGCGGA	58.69	50	AGCTTCTCCCAATTGACCCA	59.02	50	261
SSR 4	Primer3Web	CCCGCCCAAAATGAAGTCA	58.35	55	CACTACACTATGACGGCTAACT	57.37	45.45	250
	Krait	ACACATACAAAGAGACCCGCC	60.34	52.38	AACACTTGCCCGGATCG	60.05	61.11	168
SSR 5	Primer3Web	ACATACAAAGAGACCCGCCA	59.02	50	TAGATCGAACACACTTGCCCC	59.46	55	174
SSR 6	Primer3Web	CGCGATCGGGCCTTAGTTAT	60.04	55	TCCCTTCTAAGCGTCTTATTCA	57.33	39.13	236
	Primer3Web	ACGGACATAAAGACTCCTTC	58.12	47.6	TTTCGCAAAAAGCCATCGAATT	57.7	38.1	288
SSR 7	Krait	AGATCGAAGCGGAGTACCTGC	59.62	55	TCCCTAAAATCCTAATTCAGCCG	58.29	43.48	163
	Primer3Web	CGAGTACCTGCGCTTATC	57.96	55	CCCTAAAATCCTAATTCAGCCGT	58.54	43.48	153
SSR 8	Primer3Web	CAGAAAAGCAAATTGATCGTGA	57	42.9	TTTTCAATCAATCCGGCTCGAG	58.06	50	250
	Primer3Web	AGAAACTGAACAACCTCGACA	57.46	42.9	TGGCTAATCCATCGGTTATAAT	57.23	39.13	209
SSR 9	Primer3Web	GGACAAAATTATTCACAGGCCGA	58.99	45.5	TGTGACTCCGGCTTACTATTCT	57.6	39.13	448
	Krait	GCTTGCTCGAACTCATAATACCC	59.5	47.83	TCGGGAAATTCCTAGTTGC	58.2	47.62	171
SSR 10	Primer3Web	GCGTGTCTACCGAATTCACC	59	55	TCTCGGGAAGGGCAAAATAAGA	59.03	43.48	533

Primer prikaza mikrosatelita in priležnih mest za začetne oligonukleotide je na sliki 9.



Slika 9: Primer določitve priležnih mest začetnih oligonukleotidov. Z rumeno modro barvo označen mikrosatelit (ponovitev TA motiva), medtem ko so barvne regije priležna mesta za začetne oligonukleotide.

3.2.3 PRIMERJAVA REZULTATOV IN ORODIJ

Kot omenjeno smo v kloroplastnem genomu laškega smilja z vsemi tremi orodji določili 12 SSR regij. Ker pa imajo orodja možnosti iskanja tudi drugih skupin SSR (poleg popolnih mikrosateliov), smo v preglednici 5 predstavili rezultate vseh treh orodij. V tabeli 8 je pregled mononukleotidnih SSR.

Tabela 5: Primerjava identificiranih SSR orodij

Identificirani SSR	Parametri	GMATA	Krait	MISA
Popolni SSR	Število vseh	12	12 (+43 mononuk.)	12 (+43 mononukl)
	dinukleotidi	9	9	9
	trinukleotidi	3	3	3
Nepopolni SSR	Število vseh		162	
	dinukleotidi		28	
	trinukleotidi		53	
	štirinukleotidi		15	
VNTR			32	

Iz literature smo ugotovili, da so mononukleotidni SSR večinoma sestavljeni iz A oz. T (Dong-Hyuk in sod. 2017), kar velja tudi za laški smilj. V preglednici je seznam mononukleotidnih SSR identificiranih s Krait in orodjem MISA.

Tabela 6: Primerjava rezultatov identificiranih mononukleotidnih SSR z orodjema Krait in MISA

ID mikrosatelita	Motiv	Začetek	Konec	Krait	MISA	Dolžina
1	T	12701	12712	x	x	12
2	T	26467	26479	x	x	13
3	T	65136	65148	x	x	13
4	T	77988	78001	x	x	14
5	T	80676	80688	x	x	13
6	T	120246	120257	x	x	12
7	T	9230	9240		x	11
8	TT	9230	9239	x		10
9	A	9419	9428	x	x	10
10	A	13389	13398	x	x	10

11	T	20902	20911	x	x	10
12	TT	26029	26038	x		10
13	AA	29779	29788	x		10
14	TT	46540	46549	x		10
15	TT	77988	78001	x		14
16	T	26629	26638		x	10
17	T	27974	27984	x	x	11
18	T	34914	34923	x	x	10
19	A	35298	35308	x	x	11
20	T	43108	43118	x	x	11
21	A	44157	44166	x	x	10
22	A	46937	46947	x	x	11
23	T	55153	55162	x	x	10
24	T	59846	59855	x	x	10
25	A	66986	66996	x	x	11
26	A	70325	70334	x	x	10
27	A	71230	71239	x	x	10
28	T	71615	71625	x	x	11
29	T	81178	81188	x	x	11
30	T	84344	84353	x	x	10
31	A	109860	109869	x	x	10
32	A	112183	112192	x	x	10
33	A	123481	123490	x	x	10
34	T	136513	136522	x	x	10

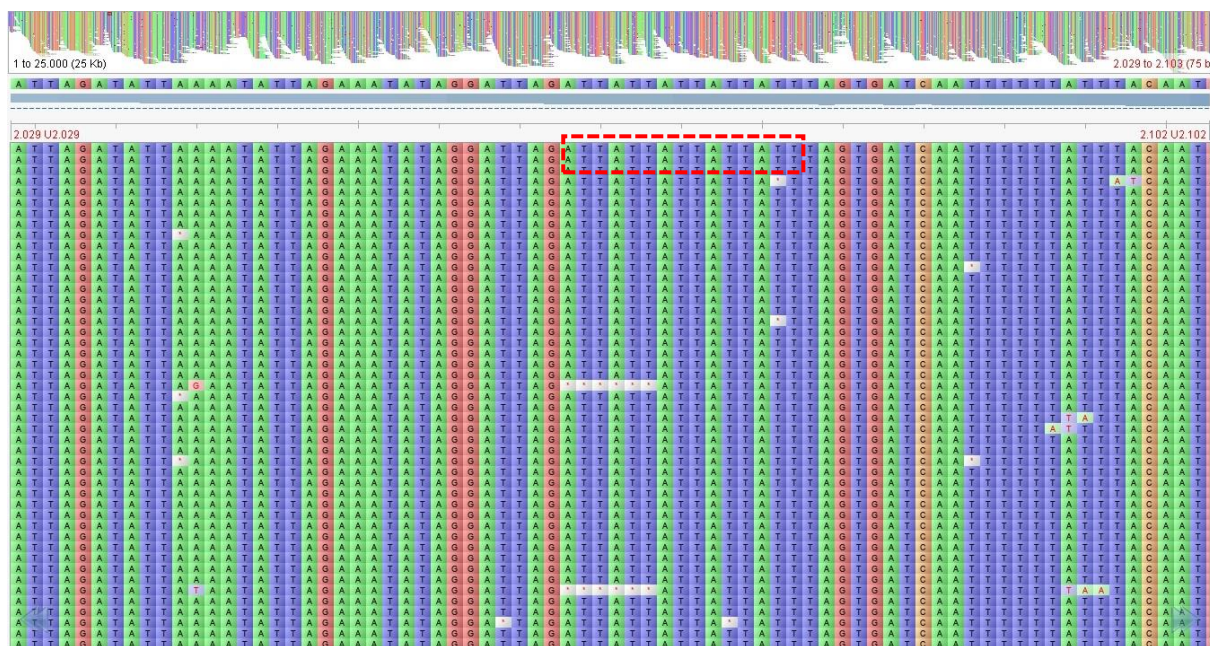
3.2.4 Genetska variabilnost kloroplastne DNA znotraj rastline

Razlogov zakaj je mogoče zaznati variabilnost v odčitkih iz kloroplastne DNA iste rastline je več (Wolfe in Randle 2004). Lahko je zaradi variabilnosti v rastlini, torej biološkega izvora, ali pa je vir variabilnost tehnična (t.j. namnoževanje za pripravo knjižnice za sekvenciranje in določanje zaporedja s sekvenatorjem). Pri izolaciji DNA iz rastline dobimo tako jedrno, kloroplastno in mitohondrijsko DNA. Obstaja možnost, da so med temi genomi dovolj podobne regije, da pri naleganju odčitek iz jedrne DNA prileže na genom kloroplasta. Scarcelli in sod. (2015) so testirali različna bioinformacijska orodja za identifikacijo variabilnih mest kloroplastne DNA. Odčitke, ki so nalegli na kloroplastni

genom, so uporabili za naleganje tudi na mitohondrijski in jedrni genom. V študiji so ugotovili, da je pojav polimorfizma v genomu verjetno rezultat kloroplastne DNA, prenesene v mitohondrijski in jedrni genom. Kljub dejstvu, da je kloroplastni genom haploiden, so našli veliko število posameznih polimorfnih položajev, in sicer so analizirali napake na devetih različnih SNP mestih. Heteroplazmija je definirana kot prisotnost neidentičnih molekul kloroplasta v celici ali organizmu. Heteroplazmija se lahko pojavi zaradi mutacij med kopijami kloroplasta, kot tudi zaradi dvostranskega dedovanja kloroplastov v nekaterih organizmih. Tudi če so kloroplasti večinoma podedovani enostransko (materinsko v večini kritosemenk), dvostarševski prenos kloroplasta opazimo med cvetočimi rastlinami in je včasih lahko precej pogost (Corriveau in Coleman 1988).

3.2.5 Analiza raznolikosti v kloroplastnem genomu

Na genom kloroplastne DNA je naleglo 550 000 odčitkov. Pri ročnem pregledu nalegih odčitkov na genom kloroplasta smo sicer opazili manjše število odčitkov z manjšim številom ponovitev osnovnega motiva (Slika 10), vendar je bilo število takih odčitkov premajhno, da bi jih kot morebitni drugi alel identificirali tudi z orodjem bcftools. Z orodjem bcftools namreč nismo uspeli identificirati nobenega variabilnega mesta. Krajši odčitki bi lahko nastali zaradi napak pri pomnoževanju DNA med pripravo predloge za sekvenciranje. Ali je v posamezni rastlini res prisotnen samo en alel za vsak kloroplastni mikrosatelitski lokus, bo mogoče preveriti tudi z analizo dolžine produktov PCR.



Slika 10: Analiza raznolikosti v kloroplastnem genomu s pregledovalnikom Tablet. Na sliki je v rdečem pravokotniku označen SSR lokus s ponavljajočim motivom ATT.

4 ZAKLJUČEK

V zaključni nalogi smo identificirali lokuse SSR v kloroplastnem genomu laškega smilja in razvili začetne oligonukleotide za namnoževanje le-teh. Pri tem smo uporabljali različna bioinformacijska orodja, in sicer GMATA, MISA in Krait.

Vsa orodja so se izkazala kot enakovredna pri iskanju pravih mikrosatelitov. Ugotovili smo tudi, da imajo nekatera orodja npr. Krait v program vgrajenih več funkcij za analizo različnih vrst mikrosatelitskih ponovitev (sestavljene in nepopolne SSR, VNTR) v primerjavi z ostalima dvema orodjema. Prav tako imata Krait in GMATA možnost dodatne analize, ki analizira razvite začetne oligonukleotide. Kljub temu, da smo se osredotočili na mikrosatelite z dolžino osnovnega motiva vsaj dveh nukleotidov, smo ugotovili, da je prisotnih tudi večje število mononukleotidnih ponovitev (A/T ponovitvi), ki so v primerjavi z ostalimi ponovitvami najpogostejši v genomu. Na drugem mestu po zastopanosti so dinukleotidi (kjer je bilo najdenih največ TA ponovitev), nato trinukleotidi (kjer sta se pojavili ATT in AGA zaporedji) in tetranukleotidi.

Priležna mesta za določitev začetnih oligonukleotidov smo analizirali s pomočjo orodij Krait in spletnega orodja Primer3Web. Tako Krait kot Primer3Web temeljita na primer3 programu. Pri tem smo se osredotočili na vseh dvanajst identificiranih popolnih mikrosatelitskih lokusov.

V orodju Tablet smo pri ročnem pregledu nalegih odčitkov na genom kloroplasta, opazili manjše število odčitkov z manjšim številom osnovnega motiva, vendar zaradi majhnega števila odčitkov, ne moremo z zagotovostjo trditi, da je variabilnost biološkega izvora. Različnih genotipov nismo uspeli potrditi tudi z orodjem bcftools.

Rezultat o zastopanosti SSR zaporedij in razviti začetni oligonukleotidi bodo v nadaljevanju uporabljeni za proučitev ali lahko na osnovi mikrosatelitskih markerjev ločimo različne kloroptipe. S tem bi pripomogli k identifikaciji vrst in podvrst, k proučevanju variabilnosti populacij ter k proučevanju genetske variabilnosti.

5 LITERATURA IN VIRI

Amiryousefi, A., Hyvönen, J., Poczai, P. (2018). The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLOS ONE* 13(4): e0196069. <https://doi.org/10.1371/journal.pone.0196069>

Antunes Viegas, D., Palmeira-de-Oliveira, A., Salgueiro L, Martinez-de-Oliveira, J, Palmeira-de-Oliveira, R. (2014). *Helichrysum italicum*: from traditional use to scientific data. *J Ethnopharmacol.*;151(1):54-65. doi:10.1016/j.jep.2013.11.005

Appendino, G., Ottino, M., Marquez, N., Bianchi, F., Giana, A., Ballero, M., Sterner, O., Fiebich, B.L., Munoz, E., (2007). Arzanol, an anti-inflammatory and anti-HIV-1 phloroglucinol alpha-Pyrone from *Helichrysum italicum* ssp. *microphyllum*. *Journal of Natural Products* 70, 608-612.

Bandelj Mavsar, D. (2006). Mikrosatelitski markerji in njihova uporabnost v oljkarstvu. *Series Historia Naturalis* 16(2): 209–222.

Beier, S., Thiel, T., Münch, T., Scholz, U., Mascher M. (2017). MISA-web: a web server for microsatellite prediction, *Bioinformatics*, Volume 33, Issue 16, Pages 2583–2585, <https://doi.org/10.1093/bioinformatics/btx198>

Bianchini, A., Santoni, F., Paolini, J., Bernardini, A.F., Mouillot, D., Costa, J. (2009). Partitioning the relative contributions of inorganic plant composition and soil characteristics to the quality of *Helichrysum italicum* subsp. *italicum* (Roth) G. Don fil. essential oil. *Chemistry & Biodiversity* 6, 1014-1033.

Birky, C.W., Jr, Fuerst, P., Maruhama, T. (1989). Organelle gene diversity under migration, mutation and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics*, 121, pp. 613-627

Brown, T.A. (2018). *Eukaryotic Nuclear Genomes*. *Genomes* 4. Garland Science: 176-178

Buels, R., Yao, E., Diesh, C.M. in sod. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 17, 66. <https://doi.org/10.1186/s13059-016-0924-1>

Chambers, G. K., MacAvoy, E. S. (2000). Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology Part B* 126: 455-576.

Chen, H., Wang, L., Wang, S., Liu, C., Blair, M.W., Cheng, X. (2015). Transcriptome sequencing of mung bean (*Vigna radiata* L.) genes and the identification of EST-SSR markers, *PLoS ONE*, 10, pp. 1-15

Cheng, J., Zhao, Z., Li, B., Qin, C., Wu, Z., Trejo-Saavedra D.L., in sod. (2016). A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in capsicum, *Scientific Reports*, 6, p. 18919

Chinou, I.B., Roussis, V., Perdetzoglou, D., Loukis, A., (1996). Chemical and biological studies on two *Helichrysum* species of Greek origin. *Planta Med.* 62, 377-379.

Hellivan, P.J., (2009). Immortelle's sustainable resurgence. *Perfum. Flavor.* 34, 34-40.

Corriveau, J.L, Coleman, A.W. (1988). Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 Angiosperm species. *American Journal of Botany*, 75, 1443–1458

Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *Plos One* 7, e35071, <https://doi.org/10.1371/journal.pone.0035071> (2012).

Dong-Hyuk L., Won-Bum C., Byoung-Hee C., Jung-Hyun L. (2017). Characterization of two complete chloroplast genomes in the tribe Gnaphalieae (Asteraceae): Gene loss or pseudogenization of trnT-GGU and implications for phylogenetic relationships. *Hortic Sci Technol.* 35:769–783.

Du, L., Zhang, C., Liu, Q., Zhang, X., Yue, B. (2018). Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design, *Bioinformatics*, Volume 34, Issue 4, Pages 681–683, <https://doi.org/10.1093/bioinformatics/btx665>

Eisse, J.A. (1999). Mechanistic basis for microsatellite instability. V. Microsatellites – evolution and applications. Goldstein D.B., Schlotterer C. (ur.). Oxford, Oxford University Press: 34-48

Galbany-Casals, M., Blanco-Moreno, J.M., Garcia-Jacas, N., Breitwieser, I., Smissen, R.D., (2011). Genetic variation in Mediterranean *Helichrysum italicum* (Asteraceae; Gnaphalieae): do disjunct populations of subsp. *microphyllum* have a common origin? *Plant Biology* 13, 678-687.

George, B., Bhatt, B. S., Awasthi, M., George, B., & Singh, A. K. (2015). Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Current Genetics*, 61(4), 665–677

Herrando-Moraira, S., Blanco-Moreno, J.M., Sáez, L., Galbany-Casals, M., (2016). Re-evaluation of the *Helichrysum italicum* complex (Compositae: Gnaphalieae): a new species from Majorca (Balearic Islands). *Collect Bot.* 35:e009.

Hladnik, M., Baruca Arbeiter, A., Knap, T., Jakše, J. & Bandelj, D. (2019). The complete chloroplast genome of *Helichrysum italicum* (Roth) G. Don (Asteraceae), Mitochondrial DNA Part B, 4:1, 1036-1037, DOI: 10.1080/23802359.2019.1580156

Levinson, G., Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203–221

Li, H., Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, Volume 25, Issue 14, Pages 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324>

Lohse, M., Drechsel, O. in Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* 52: 267– 274.

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D., Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data, *Briefings in Bioinformatics*, Volume 14, Issue 2, Pages 193–202, <https://doi.org/10.1093/bib/bbs012>

Ninčević, T., Grdiša, M., Šatović, Z., & Jug-Dujaković, M. (2019). *Helichrysum italicum* (Roth) G. Don: Taxonomy, biological activity, biochemical and genetic diversity. *Industrial Crops and Products*, 138, 111487.

Nostro, A., Bisignano, G., Angela Cannatelli, M., Crisafi, G., Paola Germano, M., Alonzo, V., (2001). Effects of *Helichrysum italicum* extract on growth and enzymatic activity of *Staphylococcus aureus*. *International Journal of Antimicrobial Agents* 17, 517-520.

Palmer, J.D., & Thomson, W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, 29(2), 537-550.

Palmer, Jeffrey D., et al. (1988). "Chloroplast DNA Variation and Plant Phylogeny." *Annals of the Missouri Botanical Garden*, vol. 75, no. 4, pp. 1180–1206. JSTOR, www.jstor.org/stable/2399279.

Perrini, R., Morone-Fortunato, I., Lorusso, E., Avato, P., (2009). Glands, essential oils and in vitro establishment of *Helichrysum italicum* (Roth) G. Don ssp. *microphyllum* (Willd.) Nyman. *Industrial Crops and Products* 29, 395-403.

Petit, R.J., Vendramin, G.G. (2007). Plant phylogeography based on organelle genes: an introduction. In: *Phylogeography of southern European refugia* (eds Weiss S & Ferrand N), pp. 23–97. Springer, Dordrecht.

Rigano, D., Formisano, C., Senatore, F. In sod. (2013). Intestinal antispasmodic effects of *Helichrysum italicum* (Roth) Don ssp. *italicum* and chemical identification of the active ingredients. *J Ethnopharmacol.*;150(3):901-906. doi:10.1016/j.jep.2013.09.034

Scarcelli, N., Mariac, C., Couvreur, T. L. P., Faye, A., Richard, D., Sabot, F., Berthouly-Salazar, C., Vigouroux, Y. (2015). Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it?, *Molecular ecology resources*, <https://doi.org/10.1111/1755-0998.12462f>

Sonah, H., Deshmukh, R.K., Sharma, A., Singh, V.P., Gupta, D.K., Gacche, R.N. in sod. (2011). Genome-Wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium*, *PLoS ONE*, 6, pp. 1-9

Staver, Mladenka & Gobin, Ivana & Ratkaj, Ivana & Petrović, Marinko & Vulinovic, Andrea & Dinarina-Sablic, Marija & Broznic, Dalibor. (2018). In vitro Antiproliferative and Antimicrobial Activity of the Essential Oil from the Flowers and Leaves of *Helichrysum italicum* (Roth) G. Don Growing in Central Dalmatia (Croatia). *Journal of Essential Oil Bearing Plants*. 21. 77-91. 10.1080/0972060X.2018.1433071.

Untergasser, A., Cutcutache, I., Koressaar, T., in sod. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Res.*;40(15):e115. doi:10.1093/nar/gks596

Wang, X. & Wang, L. (2016). GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *Frontiers in Plant Science* 7, 1350, <https://doi.org/10.3389/fpls.2016.01350>

Wolfe, A.D., Randle, C.P. (2004). Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: implications for plant molecular systematics. *Systematic Botany*, 29, 1011–1020

PRILOGE

Priloga A: Identifikacija nepopolnih SSR v orodju Krait

ID	Motiv	Začetek	Konec	Dolžina	Ujemanje	Substitucija	Insercija	Delecija
1	AAAT	124	165	42	33	6	3	1
2	ATT	2054	2077	24	21	3	0	1
3	AAG	2941	3019	79	49	24	6	1
4	ATTT	4183	4227	45	33	10	2	1
5	AATA	4948	4982	35	30	5	0	2
6	ATC	6865	6875	11	11	0	0	0
7	ATA	7439	7448	10	10	0	0	0
8	TGA	8017	8026	10	10	0	0	0
9	AT	9801	9817	17	14	3	0	0
10	ATA	10472	10496	25	20	1	4	0
11	GAA	17446	17458	13	13	0	0	1
12	TA	18650	18679	30	23	7	0	1
13	AT	19665	19675	11	11	0	0	0
14	TAA	23122	23163	42	31	9	2	0
15	TAA	23994	24018	25	20	3	2	1
16	GTT	24805	24824	20	17	3	0	0
17	TA	26545	26594	50	34	13	3	1
18	TAT	30849	30917	69	55	9	5	5
19	TTAT	30914	30960	47	35	9	3	1
20	ATTT	30962	30992	31	27	2	2	0
21	TCT	31020	31054	35	24	8	3	0
22	TA	31045	31099	55	46	9	0	9
23	TAA	31153	31267	115	82	23	10	4
24	ATA	31664	31706	43	30	11	2	1
25	TAA	32182	32215	34	26	7	1	0
26	AGG	33984	33994	11	11	0	0	0
27	GA	35057	35074	18	15	3	0	1
28	AT	35888	35936	49	33	14	2	0
29	TAA	36067	36086	20	16	3	1	0
30	ATT	36161	36161	36	31	2	3	1
31	AT	36183	36208	26	24	2	0	5
32	ATG	38438	38447	10	10	0	0	0
33	TA	41617	41641	25	21	4	0	1
34	AT	42026	42116	91	63	28	0	4

35	ATTT	44940	44963	24	18	6	0	1
36	CTT	46255	46269	15	13	2	0	0
37	TA	46436	46457	22	16	6	0	0
38	AAT	46813	46825	13	13	0	0	0
39	AT	47914	47975	62	41	15	6	2
40	ATA	48309	48331	23	18	2	3	0
41	ATTA	50923	50940	18	16	2	0	0
42	TTA	51825	51843	19	15	4	0	0
43	ATA	54754	54795	42	32	8	2	2
44	AAG	55278	55295	18	14	4	0	0
45	AT	55304	55327	24	19	3	2	1
46	TTG	55467	55476	10	10	0	0	0
47	TAGA	57081	57105	25	21	2	1	1
48	ATT	57299	57346	48	43	1	4	1
49	TAT	58782	58791	10	10	0	0	0
50	TCAA	59337	59366	30	25	4	1	1
51	TTTA	59532	59566	35	25	10	0	0
52	TA	59589	59642	54	44	8	2	4
53	AAT	59641	59692	52	40	10	2	2
54	AAT	60902	60923	22	20	2	0	2
55	AT	62175	62189	15	14	0	1	1
56	TA	63672	63697	26	22	3	1	2
57	TTA	63744	63771	28	21	6	1	1
58	ATA	66881	66896	16	14	2	0	0
59	TA	67942	67965	24	21	3	0	2
60	TATT	67989	68025	37	26	7	4	0
61	AAC	68241	68275	35	24	10	1	1
62	GAT	69903	69917	15	13	1	1	0
63	TA	71354	71382	29	26	3	0	1
64	GATT	71667	71687	21	17	4	0	0
65	TAA	71825	71868	44	30	12	2	0
66	TCT	73921	73934	14	13	1	0	1
67	TTTC	74077	74088	12	12	0	0	0
68	AAT	75500	75543	44	38	3	3	2
69	TA	77232	77254	23	18	5	0	0
70	TA	78155	78169	15	14	1	0	1
71	ATA	81828	81844	17	15	0	2	0
72	TTTC	82414	82452	39	31	7	1	2

73	AT	82749	82814	66	45	21	0	6
74	TA	83993	84020	28	26	2	0	6
75	TTA	84235	84245	11	11	0	0	0
76	TA	84914	84948	35	27	7	1	1
77	GA	86562	86584	23	18	5	0	0
78	CTT	90308	90336	29	21	7	1	0
79	AG	94580	94619	40	28	12	0	2
80	AGA	96327	96336	10	10	0	0	0
81	AAT	108848	108858	11	11	0	0	0
82	AGA	109284	109320	37	30	6	1	0
83	ATT	114017	114055	39	26	11	2	0
84	AAT	117087	117108	22	17	3	2	0
85	AGA	117369	117378	10	10	0	0	0
86	TA	118626	118656	31	22	8	1	0
87	TATT	121624	121635	12	12	0	0	0
88	TA	123331	123346	16	14	1	1	0
89	TTA	123426	123470	45	36	7	2	0
90	ATT	124945	125008	64	50	12	2	1
91	CTT	125348	125360	13	12	1	0	0
92	TTA	127813	127823	11	11	0	0	0
93	TTC	140339	140348	10	10	0	0	0
94	CT	1420956	142095	40	28	12	0	2
95	AAG	146339	146367	29	21	7	1	0
96	TC	150091	150113	23	18	5	0	0