Usporedba metoda za sastavljanje i anotaciju mitohondrijskih i jezgrinih genoma na primjeru filogenije divokoza (Rupicapra spp.)

Tešija, Toni

Doctoral thesis / Disertacija

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Agriculture / Sveučilište u Zagrebu, Agronomski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:204:190684

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-06-21**

Repository / Repozitorij:

Repository Faculty of Agriculture University of Zagreb

Sveučilište u Zagrebu

AGRONOMSKI FAKULTET

Toni Tešija

USPOREDBA METODA ZA SASTAVLJANJE I ANOTACIJU MITOHONDRIJSKIH I JEZGRINIH GENOMA NA PRIMJERU FILOGENIJE DIVOKOZA *(Rupicapra* spp*.)*

DOKTORSKI RAD

Zagreb, 2022.

University of Zagreb

FACULTY OF AGRICULTURE

Toni Tešija

COMPARISON OF METHODS FOR ASSEMBLY AND ANNOTATION OF MITOCHONDRIAL AND NUCLEAR GENOMES FOR APPLICATION OF CHAMOIS (*Rupicapra* spp.) PHYLOGENY

DOCTORAL THESIS

Zagreb, 2022.

Sveučilište u Zagrebu

AGRONOMSKI FAKULTET

Toni Tešija

USPOREDBA METODA ZA SASTAVLJANJE I ANOTACIJU MITOHONDRIJSKIH I JEZGRINIH GENOMA NA PRIMJERU FILOGENIJE DIVOKOZA *(Rupicapra* spp*.)*

DOKTORSKI RAD

Mentor: doc. dr. sc. Toni Safner

Zagreb, 2022.

University of Zagreb

FACULTY OF AGRICULTURE

Toni Tešija

COMPARISON OF METHODS FOR ASSEMBLY AND ANNOTATION OF MITOCHONDRIAL AND NUCLEAR GENOMES FOR APPLICATION OF CHAMOIS (*Rupicapra* spp.) PHYLOGENY

DOCTORAL THESIS

Supervisor: Assist. Prof. Toni Safner, PhD

Zagreb, 2022.

Bibliografski podaci:

- **Znanstveno područje:** Biotehničke znanosti
- **Znanstveno polje:** Poljoprivreda
- **Znanstvena grana:** Genetika i oplemenjivanje bilja, životinja i mikroorganizama
- **Institucija:** Sveučilište u Zagrebu Agronomski fakultet, Zavod za oplemenjivanje bilja, genetiku i biometriku
- **Voditelj doktorskog rada:** Doc. dr. sc. Toni Safner
- **Broj stranica:** 116
- **Broj slika:** 13
- **Broj tablica:** 15
- **Broj priloga:** 0
- **Broj literaturnih referenci:** 290
- **Datum obrane doktorskog rada:** 20. 07. 2022.
- **Sastav povjerenstva za obranu doktorskog rada:**
  1. izv.prof.dr.sc. Nikica Šprem, Sveučilište u Zagrebu Agronomski fakultet
  2. izv.prof.dr.sc. Jelena Ramljak, Sveučilište u Zagrebu Agronomski fakultet
  3. izv.prof.dr.sc. Ana Galov, Sveučilište u Zagrebu Prirodoslovno-matematički fakultet

Rad je pohranjen u:

Nacionalnoj i sveučilišnoj knjižnici u Zagrebu, Ulica Hrvatske bratske zajednice 4 p.p. 550, Knjižnici Sveučilišta u Zagrebu Agronomskog fakulteta, Svetošimunska cesta 25, 10000 Zagreb.

Tema rada prihvaćena je na redovitoj sjednici Fakultetskog vijeća Sveučilišta u Zagrebu Agronomskog fakulteta, održanoj 19. siječnja 2021. te odobrena na 08. redovitoj sjednici Senata Sveučilišta u Zagrebu, održanoj 16. ožujka 2021.

SVEUČILIŠTE U ZAGREBU

AGRONOMSKI FAKULTET

IZJAVA O IZVORNOSTI

Ja, **Toni Tešija**, izjavljujem da sam samostalno izradio doktorski rad pod naslovom:

USPOREDBA METODA ZA SASTAVLJANJE I ANOTACIJU MITOHONDRIJSKIH I JEZGRINIH GENOMA NA PRIMJERU FILOGENIJE DIVOKOZA *(Rupicapra* spp*.)*

Svojim potpisom jamčim:

- da sam jedini autor ovog doktorskog rada;
- da je doktorski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onih koji su u njemu navedeni;
- da sam upoznat s odredbama Etičkog kodeksa Sveučilišta u Zagrebu (Čl. 19).

Zagreb, 2022

_____

Potpis doktoranda

Ocjena doktorskog rada

Doktorski rad je obranjen na Agronomskom fakultetu Sveučilišta u Zagrebu 20. srpnja 2022. pred povjerenstvom u sastavu:

1. izv.prof.dr.sc. Nikica Šprem, _____

Sveučilište u Zagrebu Agronomski fakultet

2. izv.prof.dr.sc. Jelena Ramljak, _____

Sveučilište u Zagrebu Agronomski fakultet

3. izv.prof.dr.sc. Ana Galov, _____

Sveučilište u Zagrebu Prirodoslovno- matematički fakultet

Informacije o mentoru:

doc. dr. sc. Toni Safner

Toni Safner je rođen 3.1.1974. godine u Zagrebu. Godine 1992. završio je XV Gimnaziju u Zagrebu, da bi iste godine upisao Agronomski fakultet Sveučilišta u Zagrebu (SuZ AFZ), na kojem je 1999. diplomirao, te magistrirao 2005. godine. U statusu znanstvenog novaka bio je zaposlen od 2001. do 2012. godine na Zavodu za oplemenjivanje bilja, genetiku i biometriku Agronomskog fakulteta u Zagrebu. Godine 2006. upisuje doktorski studij u postupku zajedničkog mentorstva (Cotutelle) na Sveučilištu u Zagrebu i L'Universite Joseph Fourier u Grenobleu, Francuska. Doktorski rad naslova „Spatial genetic analysis for delineating boundaries to gene flow“, pod mentorstvom dr. sc. Jerka Gunjače i dr. sc. Stephanie Manel obranio je 20. studenog 2009. godine na L'Universite Joseph Fourier u Grenobleu, Francuska, te dobio pravo na sjecanje akademskih stupnjeva doktora znanosti iz područja Chimie et sciences du vivant na L'Universite Joseph Fourier i doktora Biotehničkih znanosti, znanstveno polje Poljoprivreda na Sveučilištu u Zagrebu.

Od 2012. do 2015. godine je bio zaposlen kao stručni savjetnik u tvrtci IRES ekologija d.o.o. za zaštitu prirode i okoliša.

Od 15. lipnja 2016. godine izabran je u znanstveno zvanje znanstvenog suradnika, 7. rujna izabran je prvi puta u znanstveno-nastavno zvanje docenta, a od 14. rujna iste godine do danas zaposlen je kao docent na SuZ Agronomskom fakultetu.

Znanstveno se usavršavao pohađajući brojne specijalističke tečajeve u Hrvatskoj i inozemstvu, te tijekom boravaka na inozemnim znanstvenim institucijama (npr: 2002. Biotehniška fakulteta u Ljubljani, Slovenija; 2007. LECA, UJF, Grenoble, Francuska; 2018. Huazhong Agriculture University, Wuhan, Kina). Bio je član organizacijskog odbora međunarodnih znanstvenih skupova: XIII EUCARPIA Biometrics in Plant Breeding Section Meeting, 2006. i III International Rupicapra symposium, 2021. Bio je suradnik na više znanstvenih projekata, a trenutno je glavni istraživač na projektu Hrvatske zaklade za znanost: ”MedUng - Uloga lova i lovnog gospodarenja u širenju novonastalih populacija divljih papkara na Mediteranu” te suradnik na Horizon 2020 projektu “RESBIOS - RESponsible research and innovation grounding practices in BIOSciencies” i na Znanstvenom centru izvrsnosti za bioraznolikost i molekularno oplemenjivanje bilja (ZCI CroP-BioDiv). Kao nositelj ili suradnik sudjeluje u izvedbi nastavnog programa na studijima Sveučilišta u Zagrebu Agronomskog fakulteta. Objavio je više od 30 znanstvenih radova iz kategorije A1 koji su bili citirani više od 250 puta. Aktivno se služi engleskim, a pasivno ruskim, francuskim i njemačkim jezikom. Član je EUCARPiA-e. Dobitnik je nagrade za najbolji objavljeni istraživački znanstveni rad časopisa International journal of molecular sciences u 2015 godini.

Počevši od akademske godine 2001./02., sudjeluje u nastavi više modula iz područja biometrike. Bio je gostujući nastavnik na Sveučilištu u Sassariju, Sassari, Italija (2012.), Višoj tehničkoj školi u Zagrebu (2015.) i Huazhong Agriculture University, Wuhan, Kina (2018.).

Zahvala

Sažetak

Genomi su sve češće korišteni podaci u proučavanju biologije i evolucije organizama, a broj dostupnih genoma u Banci gena se u posljednih nekoliko godina udvostručio kao posljedica razvoja tehnologije sekvenciranja. Razvoj ovih tehnologija utjecao je na populariziranje područja genomike i to prvenstveno zbog značajnog pada cijene sekvenciranja.

Rekonstrukcija genoma provodi se u tri koraka: sekvenciranje, sastavljanje i anotacija, a za svaki korak postoji više različitih pristupa. Sastavljanje genoma je računalno i vremenski nazahtjevniji korak te je jedan od glavnih fokusa istraživanja u području genomike. Tri su trenutno dostupne metode za sastavljanje genoma (mapiranje, *de novo* i hibridna metoda), a odabir metode ovisi o nekoliko glavnih parametara koji uključuju: vrstu organizma koji se proučava, pokrivenost genomskih podataka, dostupnost referentne sekvence, broj uzoraka, dostupnost računalnog servera za provođenje analiza i sl. Prema tome, svaki genomski projekt je jedinstven i teško je odabrati samo jednu metodu koja će dati najbolje rezultate, pogotovo kada se proučavaju nemodelne vrste.

Divokoza (*Rupicapra* spp.) je zbog svoje rasprostranjenosti i predložene sistematike dobar model za proučavanje utjecaja povijesnih i evolucijskih događaja. U ovoj se disertaciji koristilo nekoliko metoda za sastavljanje i anotaciju mitohondrijskih i jezgrinih genoma divokoze, a dobiveni su se rezultati usporedili. Na temelju usporedbi rezultata metoda za sastavljanje mitohondrijske i jezgrine DNA, procijenile su se pogodnosti različitih metoda za sastavljanje i anotaciju genoma, uspoređen je utjecaj korištenja osam genoma divokozi srodnih vrsta kao referenci u metodi mapiranja te su se rekonstruirali filogenetski odnosi s ciljem boljeg razumijevanja povezanosti taksonomskih jedinica roda *Caprini* i vrste *Rupicapra*. Uz navedeno, testirana je točnost novosastavljenih genoma divokoze usporedbom izoliranih fragmenata introna s intronskim sekvencama divokoza dostupnih u Banci gena.

Rezultati ovog istraživanja pridonijet će boljem poznavanju raznolikosti i evolucije genoma divokoze, razjašnjavanju taksonomskih odnosa podvrsta, a sastavljeni genomi pružit će dobru referentnu osnovu za buduće populacijske i genomske analize divokoze i njenih srodnika.

Ključne riječi: genom, mtDNA, mitogenom, sastavljanje genoma, anotacija genoma, filogenija, divokoza, *Rupicapra*, planinski papkari, *Caprinae*

Extended Abstract

Comparison of methods for assembly and annotation of mitochondrial and nuclear genomes for application of chamois (*Rupicapra* spp.) phylogeny

The genome is a collection of all biological information necessary for the functioning of an organism, and in humans and animals consists of the mitochondrial genome (mtDNA) and the nuclear genome (nDNA). With the development of genomic technologies, the genome data are increasingly being used to study the biology and evolution of organisms, which is confirmed by the fact that the number of available genomes in the Gene bank has doubled in recent years. In addition, the development of these technologies has influenced the popularization of the field of genomics, mainly due to the significant decrease in the cost of sequencing. Genomic analysis can be used: to identify genes responsible for inherited diseases or adaptations to the environment, to study structural changes, to identify common conserved sites, to find genes specific to a group of organisms, etc. To perform any of these analyzes, the genome of the species under study must be reconstructed.

Reconstruction of a genome involves three steps: sequencing, assembly, and annotation. There are different approaches for each step, and the choice of method depends primarily on whether the reconstruction is of the mitochondrial or nuclear genome. Genome assembly is a computationally intensive and time-consuming step, and there are currently three available methods for genome assembly (mapping, *de novo*, and hybrid methods). The choice of method depends on several key parameters that include: the type of organism studied, the coverage of genomic data, the availability of reference sequences, the number of samples, the availability of a computer server for analysis, etc. Therefore, each genome project is unique and it is difficult to determine the method that will be most successful, especially when studying non-model species.

The chamois (*Rupicapra* spp.) is a good model for studying the effects of historical and evolutionary events because of its wide distribution and proposed systematics. In this dissertation, 12 completely sequenced chamois samples were used and their genome reconstruction was performed using different methods for assembling and annotating mtDNA and nDNA. Mapping and *de novo* methods were used for mtDNA assembly, while GeSeq and MITOS annotators were used for annotation. All sequences obtained with both assembly methods were compared and validated with the web application BLAST. In this process, each sequence was compared to the chamois mtDNA reference sequences. For nDNA assembly, eight available genomes of closely related species were used as references for the mapping method, followed by SNP calling procedure. Based on the filtered SNPs and references, 56 combinations were identified (newly assembled chamois genomes) that were validated and annotated using BUSCO tools. Then, three smaller sets of genes were defined from the common set of annotated genes, based on which distance matrices were calculated and the relationships were visualized using the multidimensional scaling (MDS) method. Newly assembled genomes generated by mapping chamois samples against the domestic goat genome were used to verify structure by comparing them with the chamois genome sequences (23 sets of introns) available in Gene bank. Phylogenetic analyses of mtDNA (maximum likelihood and Bayesian methods) were performed using a dataset containing 10 mtDNA sequences from this dissertation, 5 chamois mtDNA sequences from Gene bank, and two related sequences as outgroup. Phylogenetic analyses (maximum likelihood and Bayesian methods) of the genus *Caprinae* were performed on a data set containing 40 sequences of the genus *Caprinae* and 5 sequences of *Bovidae* as outgroup. The final phylogenetic analysis of chamois was performed using the program BEAST on a common

alignment of an intron data set consisting of 21 chamois sequences and three sequences representing an outgroup.

After performing two methods for reconstructing complete mtDNA sequences (mapping and *de novo*) and comparing the obtained mtDNA sequences, it was clearly determined that both methods were suitable for reconstruction. The *de novo* method proved to be the better choice because of its speed and simpler procedure. In addition, the *de novo* methods successfully isolated complete mtDNA sequences from samples that had failed quality control (Gams53, Gams85, OSIL-06). For this reason, all mtDNA analyses were performed on a larger number of samples. In other words, if only the mapping method had been used, these three samples could not have been included in the further analysis.

The mtDNA annotation tools MITOS and GeSEQ gave very similar results for all 10 samples, and variations in START and STOP codons were detected in four genes (ND1, ND2, ND3, ND5). The variations found refer to two or three bases found in the START and STOP codons. Their occurrence can be interpreted as a consequence of the different algorithms used by the annotators in the analyzes and as a consequence of the larger variation in these genes.

Phylogenetic analyses (maximum likelihood and Bayesian methods) performed on ten mtDNA sequences from chamois reconstructed in this dissertation, in combination with five mtDNA sequences from Gene bank, yielded identical phylogenetic trees for the genus *Rupicapra*. These results confirmed previous research on chamois in which they were divided into three mtDNA clusters (W, C, and E). Performed phylogenetic analyses (maximum likelihood and Bayesian method) on 40 mtDNA sequences of the genus *Caprinae* (including the four sequences obtained in this dissertation) and five sequences of the genus *Bovidae* also revealed the same topology as presented in previous studies.

From the comparison of the newly assembled chamois genomes with the genomes of related species, it was concluded that almost all of the related genomes used can serve as good references. Although all the species used were non-model species, the best results were obtained with the genomes of domestic goats and domestic sheep, which was to be expected since these species are extremely important species in agriculture and are often the focus of research. The number of genes found for most combinations of chamois and related species was very high, confirming that these genomes can be used for mapping processes. However, during the mapping processes it was found that some of the genomes used were of low quality, while some genomes were found to have irregularities in the information available in the Gene bank. This confirms once again that not all available genomes are of good quality. In other words, any sequence available in the Gene bank should be verified before use.

From the similarity analyses, it can be concluded that the relationships between all combinations depend primarily on the gene or genome fragments used for these analyses. Although the number of polymorphisms found had a greater impact on the results when single gene fragments were used, this number was negligible when longer portions of the genome (100 and 500 genes) were used, with differences between samples of approximately 1 %. In other words, larger distances were calculated between combinations from shorter alignments. The results of the MDS for a set of 100 and 500 genes clearly showed that samples from chamois samples mapped to different references were more similar to each other, while still exhibiting some differences in amino acid composition. Smaller differences between samples were found for combinations with domestic sheep and American mountain goat (about 1 % and 000,5 %, respectively).

The comparisons of the intron regions of the newly assembled chamois genomes with the introns available in the Gene bank suggest that the intron sequences obtained from the newly assembled genomes are of satisfactory quality and have been grouped with other chamois samples at the species and subspecies level.

The results of this research will contribute to a better knowledge of the diversity and evolution of the chamois genome, elucidate the taxonomic relationships among subspecies, and assembled genomes will provide a good reference base for future population and genome analyses of chamois and its relatives.

Keywords: genome, mtDNA, mitogenome, genome assembly, genome annotation, phylogeny, chamois, *Rupicapra*, mountain ungulates, *Caprinae*

SADRŽAJ

Popis kratica

Kratica	**Značenje**
DNA	Deoksiribonukleinska kiselina
RNA	ribonukleinska kiselina
nDNA	jezgrin (nuklearni) genom
mtDNA	mitohondrijski genom, mitohondrijska DNA, mitogenom
NGS	sekvenciranje sljedeće generacije
NCBI	engl. *National Center for Biotechnology Information*
ATP	adenozin trifosfat
PCG	protein-kodirajuće regije, protein-kodirajući geni
rRNA	ribosomska ribonukleinska kiselina
tRNA	transportna ribonukleinska kieslija
CR	kontrolna regija (D-loop)
SAM	engl. *Sequence Alignment Format*
BAM	engl. *Binary Alignment Map*
VCF	engl. *Variant Call Format*
BCF	engl. *Binary Variant Call Format*
SNP	engl. *Single Nucleotide Polymorphysm*
INDEL	insercije/delecije (*engl. indels*)
OLC	engl. *overlap-layout-consensus*
DBG	engl. *de Bruijn graph*
CYTB	citokrom b
BUSCO	engl. *Benchmarking Universal Single-Copy Orthologs*
PCR	lančana rekacija polimerazom
ORF	engl. *Open reading frames*
BLAST	engl. *Basic Local Alignment Search Tool*
MCMC	Markovljevi lanaci Monte Carlo
EST	engl. *expressed sequence tags*
cDNA	komplementarne DNA
QUAL	kvaliteta pronađenog SNP-a
DP	dubina sekvenciranja
MQ	kvaliteta mapiranja

Popis Tablica

Tablica 15. Rezultati dobiveni procesima mapiranja, pozivanja varijanti i BUSCO analize uzoraka divokoze mapiranih na referentni genom domaće koze.

Popis Slika

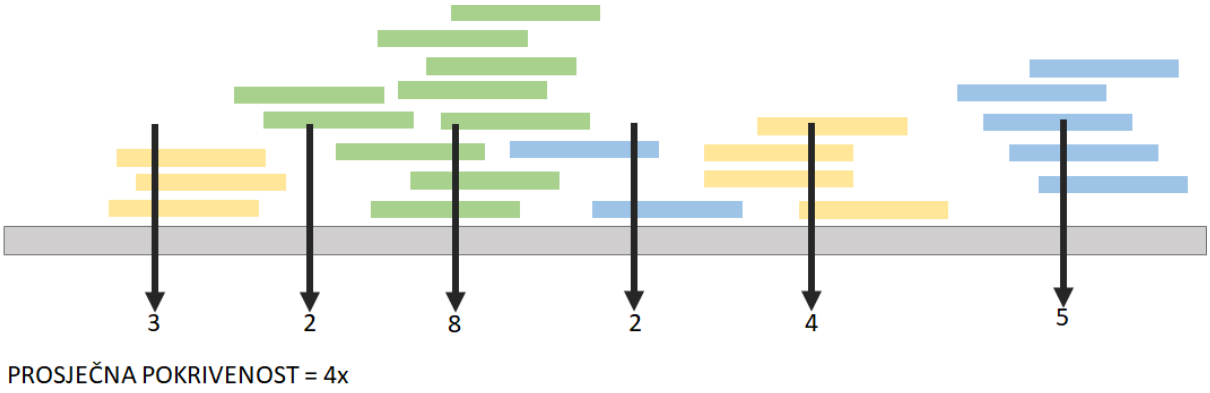
Bayesovske posteriorne vjerojatnosti. Označeni klasteri su uzorci iz ove disertacije: crvena – 4 uzorka sjeverne divokoze; zelena – 2 uzorka južne divokoze.

1 UVOD

Genom predstavlja skup svih bioloških informacija koje su neophodne za funkcioniranje nekog organizma. Kodiran je u obliku deoksiribonukleinske kiseline (DNA) i kod ljudi i životinja se sastoji od dva dijela: jezgrinog (nuklearnog, nDNA) genoma i mitohondrijskog (mitogenoma, mtDNA) genoma (Roth, 2019). Rekonstrukcijom genoma i njegovom anotacijom, mogu se identificirati geni i utvrditi genetska varijabilnost iz koje se potom traže specifični odgovori usko vezani za detekciju varijanti odgovornih za nasljedne bolesti ili za utvrđivanje varijanti među organizmima koje su nastale kao posljedica evolucijskih događaja i okolišnih utjecaja. The Human Genome Project prvi je veliki projekt sekvenciranja koji je započeo 1990. godine a završio 2001. Provedbom ovog projekta postavljeni su bitni temelji za daljnji razvoj ovog područja. U posljednjih desetak godina razvoj NGS tehnologija (engl. *Next Generation Sequencing*) pojednostavnio je i pojeftinio korištenje genoma u poljima biologije i biomedicine. Iz tog razloga, kompletni mtDNA i nDNA genomi sve su češće korišteni u proučavanju bioraznolikosti na što ukazuju podaci velikog broja sekvenciranih i deponiranih genoma u Banci gena (NCBI, engl. *GenBank*; https://www.ncbi.nlm.nih.gov/). Za detekciju gena, njihovih varijanti te ostalih strukturnih promjena u genomu nekog organizma kao i za njegovu usporedbu s drugim organizmima potrebno je najprije provesti rekonstrukciju genoma.

Rekonstrukcija mtDNA i nDNA uključuje tri ključna koraka: sekvenciranje, sastavljanje i anotaciju (Ekblom i Wolf, 2014). U procesu NGS sekvenciranja kompletne DNA, DNA iz uzorka se prvo izolira i polomi na stotine milijuna fragmenata (engl. *reads*). Uređaj za sekvenciranje potom zapisuje informacije o svakom sekvenciranom fragmentu (kvaliteta svake sekvencirane baze, redoslijed i broj nukleotida) u specijaliziranom formatu datoteka FASTQ. Programski alati za sastavljanje genoma potom čitaju FASTQ datoteke iz kojih uspoređuju kratke fragmente i slažu ih u jedinstveni slijed. Trenutno postoje tri metode za sastavljanje genoma: sastavljanje pomoću referentne sekvence (mapiranje), *de novo* metoda te hibridna metoda (Young i Gillung, 2020). U prvoj metodi programski alati slažu kratke fragmente duž referentne sekvence koja pripada bliskom srodniku, a potom pronalaze i uspoređuju sve strukturne razlike između referentne sekvence i složenih fragmenata. U *de novo* metodi programski alati pronalaze DNA fragmente koji se najefikasnije mogu složiti u jedinstveni niz po principu preklapanja (engl. *overlapping*). Ova metoda se najčešće koristi kada za genom koji se želi rekonstruirati ne postoji referentna sekvenca srodnika. Izbor

metode za sastavljanje prvenstveno ovisi o kvaliteti genomskih podataka (Ekblom i Wolf, 2014). Najbitnije svojstvo genomskih podataka je pokrivenost (engl. *coverage*, *x-coverage*) (Slika 1) ili prosječan broj koji prikazuje koliko je puta pojedina baza sekvencirana (Young i Gillung, 2020). Pokrivenost se izražava simbolom „x“ (puta), npr. 10x bi značilo da je svaka baza u prosjeku sekvencirana 10 puta. Pokrivenost genomskih podataka ima veliki utjecaj na rezultate u svim metodama sastavljanja genoma. Veća pokrivenost znači i manji stupanj pojave pogrešaka u cijelom procesu ali i veću cijenu sekvenciranja. Za genomske podatke koji imaju pokrivenost veću od 50x preporuča se korištenje *de novo* metode sastavljanja, dok se za podatke s malom pokrivenošću koriste metode temeljene na mapiranju. U novije vrijeme opisane su i nove, hibridne metode koje se mogu koristiti za podatke male pokrivenosti (Lischer i Shimizu, 2017; Buza i sur., 2015).



Slika 1. Shematski prikaz prosječne pokrivenosti genomskih podataka

Anotacija genoma postupak je kojim se biološke informacije povezuju s dijelovima genoma na način da programski alati prvo analiziraju strukturu i sastav sekvence te je uspoređuju s poznatim genomima srodnih vrsta (Dominguez Del Angel i sur., 2018). Anotacija jezgrinog genoma specifična je za vrstu s obzirom da je svaki genom različit dok je anotacija mtDNA jednostavniji proces zbog visoke sličnosti mtDNA sekvenci među organizmima. Osim toga, i proces anotacije mnogo je lakši ukoliko je dostupan anotirani referentni genom (Ekblom i Wolf, 2014).

Analizom genoma proučavaju se strukturne promjene, identificiraju se zajednička konzervirana mjesta te se pronalaze geni karakteristični za neku skupinu organizama. Međutim, takve analize su računalno i vremenski zahtjevni procesi koji ovise o puno faktora i

zbog toga se redovno razvijaju novi pristupi i algoritmi kojima bi se ovi procesi olakšali i ubrzali (Dominguez Del Angel i sur., 2018). Iako se gotovo svake godine razvijaju novi protokoli kojima se nastoji ubrzati svaki od ova tri koraka, većina novih protokola je nastala proučavanjem modelnih vrsta čiji su genomi visoke kvalitete i često je implementacija takvih protokola u istraživanju nemodelnih vrsta izazovan proces. Osim toga, svaki projekt je jedinstven i ovisi o velikom broju parametara zbog čega je gotovo nemoguće odabrati i slijediti samo jedan protokol već je potrebno koristiti kombinaciju više njih. Parametri koji se najčešće uzimaju u obzir prilikom odabira metoda su: vrsta organizma, broj uzoraka uključenih u projekt, kvaliteta i pokrivenost genomskih podataka, dostupnost računalnog servera za provođenje analiza i sl. Rekonstrukcija genoma nemodelnih vrsta još uvijek predstavlja veliki izazov u području genomike jer su referentni genomi i kvalitetne i točne anotacije dostupne u Banci gena za jako mali broj vrsta. Međutim, kombinacijom različitih metoda moguće je doći do točnih informacija o vrsti koja se proučava.

Jedna od takvih vrsta je divokoza (*Rupicapra* spp.) čiji genom do sada nije sastavljen, a u ovoj disertaciji će divokoza poslužiti kao pokazni primjer za usporedbu metoda za rekonstrukciju mtDNA i nDNA. Divokoza je papkar koji naseljava planinske predjele Euroazije. Prema morfološkim i bihevioralnim karakteristikama, dvije su poznate vrste divokoze: sjeverna divokoza (*Rupicapra rupicapra*) sa sedam podvrsta (*asiatica*, *balcanica*, *carpatica*, *cartusiana*, *caucasica*, *rupicapra*, *tatrica*) te južna divokoza (*Rupicapra pyrenaica*) s tri podvrste (*parva*, *pyrenaica*, *ornata*) (Corlatti i sur., 2022a). Prema podacima o brojnom stanju divokoza, niti jedna vrsta nije ugrožena, međutim, na razini podvrsta postoje razlozi za zabrinutost oko zaštite pojedinih podvrsta. Drugim riječima, na razini podvrste, divokozu možemo smatrati jednom od najugroženijih papkara na području Europe (Corlatti i sur., 2011). Osim toga, divokoza je zbog svoje rasprostranjenosti i predložene sistematike dobar model za proučavanje utjecaja povijesnih i evolucijskih događaja. Do sada su za analize molekularne raznolikosti divokoze korišteni neutralni markeri, dijelovi mtDNA (citokrom b, RNA geni, i kontrolna regija) te pojedini jezgrini geni (Rodríguez i sur., 2009; Rodríguez i sur., 2010; Pérez i sur., 2017; Safner i sur., 2019). Rezultati filogenetskih istraživanja temeljeni na mtDNA i analizi mikrosatelita grupiraju istraživane populacije u tri geografska područja (istočni, zapadni i centralni).

Filogenetska istraživanja temeljena na mitohondrijskim dijelovima DNA identificirala su tri geografske linije (istočna, zapadna i centralna) (Rodríguez i sur., 2009; Rodríguez i sur., 2010). Slični rezultati dobiveni su analizama mikrosatelita. Međutim, analiza mikrosatelita

pokazala je veću podudarnost s morfološkom klasifikacijom nego s rezultatima dobivenim analizom mtDNA (Rodríguez i sur., 2009; Rodríguez i sur., 2010). Osim toga, analize mitohondrijskih pseudogena i jezgrinih introna također se nisu u potpunosti poklapala s podjelom temeljenoj na mtDNA. Zbog boljeg uvida i razumijevanja filogenije divokoza kao i cijelog roda *Caprini* kojem pripadaju, u ovom će se istraživanju koristiti sastavljeni i anotirani mtDNA te dijelovi jezgrinog genoma. Općenito, filogenetske analize koristan su alat za opisivanje odnosa između gena, genoma, vrsta i drugih taksonomskih jedinica (Young i Gillung, 2020). Nukleotidni i aminokiselinski sljedovi gena često su slični unutar zajedničkih grupa, a male promjene u njihovom rasporedu omogućuju procjenu srodnosti i vremena divergencije putem kalibriranih molekularnih satova (Young i Gillung, 2020). Rekonstrukcija filogenetskih odnosa između vrsta ovisi o genetskim markerima koji se koristi u analizi, a analize zasnovane na jednom genu često daju ograničene zaključke. Novija istraživanja pokazala su da se odnosi između pojedinih vrsta i podvrsta mogu preciznije opisati korištenjem molekularnih markera koji sadrže veću količinu informacija (varijabilnih mjesta) kao npr. većih dijelova genoma, kompletnih sekvenci mtDNA ili čitavih genoma (Phillippe i Telford, 2006; Young i Gillung, 2020).

Korištenjem 12 uzoraka divokoze, u ovoj će se disertaciji koristiti metode za rekonstrukciju mtDNA i nDNA, a dobivene će se sekvence (kompletne mtDNA i dijelovi nDNA) usporediti i potom koristiti za rekonstrukciju filogenetskih odnosa roda *Rupicapra* i cijelog roda *Caprini*.

1.1 Hipoteze i ciljevi istraživanja

1.1.1 Hipoteze:

1. Mitohondrijski genomi divokoza mogu se, uz jednaku pouzdanost i točnost, sastaviti korištenjem različitih metoda.
2. Kombiniranjem metoda za sastavljanje i korištenjem genomskih podataka male pokrivenosti, dobivene sekvence jezgrinog genoma mogu se koristiti za usporedbu s genima srodnih vrsta i za pronalazak gena uključenih u metaboličke procese.
3. Rekonstrukcija filogenetskih odnosa temeljena na genomskim sekvencama jednoznačno će definirati odnose unutar roda *Caprini* i vrste *Rupicapra*.

1.1.2 Ciljevi:

1. Usporediti pouzdanost i točnost mitohondrijskih genoma divokoza sastavljenih različitim metodama za sastavljanje.
2. Korištenjem dviju metoda za sastavljanje sastaviti jezgrin genom divokoze iz podataka male pokrivenosti te usporediti dobivene sekvence s bazama ortolognih gena srodnih vrsta.
3. Rekonstruirati filogeniju unutar roda *Caprini* i vrste *Rupicapra* iz mtDNA i sekvenci jezgrinog genoma.

2 PREGLED DOSADAŠNJIH ISTRAŽIVANJA

2.1 Jezgrin genom (nDNA)

Jedno od najznačajnih otkrića u području genetike dogodilo se 1953. godine (Watson i Crick,1953) kada su objavljene bitne spoznaje o sastavu i strukturi DNA. Te iste godine Sanger i Thompson (1953a; 1953b) prvi su put, korištenjem parcijalne kromatografije, sekvencirali dva lanca proteina inzulina. Iako su proteini sekvencirani prije nukleotida, principi su bili vrlo slični i upravo se ovo smatra začetkom modernog DNA sekvenciranja. Alanin tRNA prva je sekvencirana RNA molekula koju su sekvencirali Holley i sur. (1965). Sljedećih deset godina sekvencirano je nekoliko DNA i RNA molekula (Sanger i sur., 1977a) da bi se razvojem Sangerovog sekvenciranja 1977. godine po prvu puta sekvencirao kompletni genom bakteriofaga φX174 duljine 5.368 nukleotida (Sanger i sur., 1977b). Potom je Staden (Staden, 1979) uveo termin tzv. shotgun sekvencirajne (engl. *shotgun sequencing*) u kojem se bakterijski vektori koriste za kloniranje DNA fragmenata koji se nakon nasumičnog paralelnog sekvenciranja sastavljaju metodom preklapanja. Svega nekoliko godina poslije, Sanger i sur. (1982) sekvencirali su i sastavili genom bakteriofaga λ veličine 58.502 nukleotida. U 80-ima je započela prava revolucija u području sekvenciranja genoma. Banka gena (Nacionalni centar za biotehnološke informacije, NCBI) osnovana je 1982. godine (Smith, 2013) s deponiranih pola milijuna nukleotida u bazi da bi do kraja tog desetljeća broj nukleotida porastao na 40 milijuna. U 90-ima su osnovane tvrtke koje su najviše doprinijele razvoju sekvenciranja (Illumina, Solexa, 454, Ion Torrent) i intenzivno radile na usavršavanju tehnologija s ciljem ubrzanja i povećanja točnosti cijelog procesa (Giani i sur., 2019). Početak 90-ih godina posebno je značajan zbog pokretanja velikog projekta The Human Genome Project kojim je započeto sekvenciranje ljudskog genoma. Prva genetska karta (engl. *genetic map*) ljudskog genoma pomoću koje se mogao identificirati i karakterizirati veći broj gena odgovornih za nasljedne bolesti objavljena je 1994. godine (Murray i sur.,1994), no u toj karti su sve informacije o genima dobivene iz manjeg dijela genoma, odnosno iz svega 50 milijuna nukleotida. Početkom 2000-ih, Celera tim je započeo s razvojem novih tehnologija sekvenciranja bez korištenja bakterijskih klonova, i uz korištenje novih računskih algoritama u procesu sekvenciranja što je rezultiralo sekvenciranim i sastavljenim genomom vinske mušice (*Drosophila melanogaster*) (Myers i sur., 2000). Potom je istom metodom 2000. godine sekvencirano dvije trećine ljudskog genoma, da bi 2001. godine bio objavljen kompletni ljudski genom (Venter i sur., 2001). Rezultati ovog projekta objavljeni su u znanstvenim časopisima Nature i Science 2001. godine (IHGSC, 2001; Venter i sur., 2001) a

dobivene i objavljene sekvence predstavljale su nacrt genoma (engl. *draft genome*) kojim se opisalo oko 83-84 % ljudskog genoma što je bilo dovoljno za postavljanje prve spoznaje o ljudskom genomu te njegovoj funkciji i organizaciji. Ustanovljeno je da ljudski genom sadrži oko 3.2 milijardi nukleotida podijeljenih u 24 linearne molekule koje predstavljaju kromosome (22 autosomna i 2 spolna kromosoma). Nacrtni ljudski genom potom se koristio kao glavna referenca za sastavljanje genoma drugih modelnih vrsta. Referentni ljudski genom sastavljen je iz više uzoraka (različitih genoma) budući da svaka jedinka ima jedinstvenu genomsku sekvencu (Brown, 2002). Do danas se referentni ljudski genom kontinuirano poboljšavao s ciljem stvaranja tzv. pan genoma (engl. *pan genome*) čovjeka kojim se nastoji obuhvatiti što veći broj varijacija prisutnih u ljudskoj populaciji (Schneider i sur., 2017).

Genom oblića *Caenorhabditis elegans* prvi je objavljeni životinjski jezgrin genom sastavljen 1998. godine (CESC, 1998), a sastojao se od 97 milijuna nukletidinih baza (9.7 Mb) koje kodiraju 19,000 gena. Od tada se svakim novim sekvenciranim genomom nastoji predočiti kompleksnost bioraznolikosti među živućim organizmima. Glavni fokus sekvenciranja bili su genomi modelnih vrsta te onih vrsta koje se najčešće koriste u biomedicinskim istraživanjima kao npr. kućni miš (*Mus musculus*) (MGSC, 2002), smeđi štakor (*Rattus norvegicus*) (RGSPC, 2004), čimpanza (*Pan troglodytes*) (CSAC, 2005), veliki panda (*Ailuropoda melanoleuca*) (Li i sur., 2010a), zebrica (*Danio rerio*) (Howe i sur., 2013) i dr. Dva su ključna događaja utjecala na revoluciju u području genomike: usavršavanje NGS tehnologija i pad cijene sekvenciranja. Kao posljedica navedenog, osnovan je veliki broj istraživačkih grupa tzv. konzorcija s ciljem sekvenciranja što većeg broja životinjskih vrsta (The Vertebrate Genome Project (Rhie i sur., 2021), The Bird 10K Project (Zhang i sur., 2015), The i5K Project (Evans i sur, 2013), The Earth BioGenome Project (Lewin i sur., 2018) itd). Do 2021. godine su u Banci gena dostupni genomi 3.278 vrsta životinja, što čini oko 0,2 % svih životinjskih vrsta dok je prosječna veličina sekvenciranih genoma oko 1 milijardu nukletodinih baza (1 Gb) (Hotaling i sur., 2021). Genomi sisavaca u prosjeku su veliki između 2.5 i 3.5 Gb i sadrže između 20.000 i 40.000 gena. Velika većina ovih gena kodira proteine uključene u ekspresiju, replikaciju i održavanje samog genoma dok manji dio kodira za proteine odgovorne za građu stanica (Brown, 2002). Do 2022. godine je u Banci gena bilo dostupno ukupno 2.075 genoma sisavaca, s tim da je 490 genoma definirano kao referentni (budući da je za neke vrste dostupno više genoma, a samo se najkvalitetniji opisuje kao referentni) dok svega189 genoma ima dostupne anotacije.

2.2 Mitohondrijska DNA (mtDNA, mitogenom)

Mitohondriji su stanični organeli odgovorni za proizvodnju energije u eukariotskim stanicama. U procesu oksidativne fosforilacije, gdje adenozin trifosfat (ATP) nastaje kao glavni produkt, mitohondriji oksidiraju metaboličke supstrate kako bi stvorili energiju i vodu (Boore, 1999; Wallace, 2007). Mitohondriji, za razliku od drugih organela, u svojoj jezgri imaju nekoliko kopija vlastite DNA molekule (Robin i Wong, 1988) koja je odgovorna za kodiranje ribosomskih i transportnih RNA te protein-kodirajućih regija (PCG). Nass i Nass (1963) su u svom istraživanju po prvi puta izolirali i opisali mtDNA proučavanjem mitohondrijskih vlakana iz kojih su, na temelju stabilizacije i bojanja, dokazali da se radi o DNA molekuli. Osamnaest godina kasnije, Anderson i sur. (1981) su objavili prvu kompletnu sekvencu mtDNA čovjeka. Ovim su radom utemeljene spoznaje o mtDNA te je po prvi puta prezentiran nukleotidni zapis svih 37 gena karakterističnih za mitogenom sisavaca. MtDNA je kratka (~16 kb) kružna molekula koja se sastoji od dva lanca koji kodiraju ukupno 37 gena: teški (H) lanac koji je bogat gvaninom i kodira 28 gena te laki (L) lanac koji je bogat citozinom i kodira 9 gena. Osim toga, mtDNA se dijeli na kodirajuće i nekodirajuće regije. Kodirajuće regije uključuju fragmente mtDNA koji kodiraju svih 37 gena: 13 PCG-a, 22 transportne RNA (tRNA) te dvije ribosomske RNA (rRNA). Ove su regije visoko konzervirane i imaju malu stopu mutacija u usporedbi s jezgrinim genomom (Wolstenholme, 1992). Najčešći uzroci mutacija u tim regijama povezani su s transportom jezgrinih gena (uključenih u funkcije mitohondrija) iz citolosola u mitohondrij (Patrushev i sur., 2014). S druge strane, nekodirajuće regije mtDNA su varijabilni dio mtDNA s velikom stopom mutacija, a uključuju tri dijela: kotrolnu regiju (CR, D-loop), O_L regiju (mjesto gdje započinje replikacija L lanca) te intergenske regije (regije koje se nalaze između pojedinih gena) (Taanman, 1999). Replikacija lanaca mtDNA odvija se na različitim lokacijama. Mjesto replikacije H lanca O_H nalazi se u CR dok se mjesto replikacije L lanca O_L nalazi između gena koji kodiraju tRNA za asparagin i cistin (Hixson i sur., 1986). U procesu replikacije, prvo se replicira H lanac. Prije nego se upari s novim sinteziranim L lancem, roditeljski H lanac izložen je oksidacijskim oštećenjima jer je samo djelomično zaštićen proteinima (Matosiuk i sur., 2014). Kao posljedica, dolazi do pojave velikog broja mutacija u mtDNA, posebno u hipervarijabilnim regijama 1 i 2 (Brown i Simpson, 1982; Miller i sur., 1996; Reyes i sur., 1998). Novonastale mutacije u CR često nisu letalne i mogu se prenijeti na potomstvo. S druge strane, broj novonastalih mutacija u kodirajućim regijama koji se prenese na potomstvo je nizak budući da su mutacije u tim regijama često letalne (Gupta i sur., 2015).

Promjene u regijama mtDNA najčešće su posljedica prilagodbe jedinki na okolišne utjecaje. Odnosno, tijekom dužeg vremenskog perioda, razvoj različitih grupa organizama i njihova prilagodba na okolišne utjecaje usko su bile vezane s promjenama u mtDNA. Novonastale mutacije su se kroz određeni vremenski period fiksirale, definirajući na taj način haplogrupe te manje haploptipove (Mishmar i sur., 2003). Upravo iz tog razloga, proučavanjem mtDNA može se pratiti tijek evolucijskih događaja karakterističnih za neku grupu organizama i spoznati njihove evolucijske odnose (Mereu i sur., 2008; Manee i sur., 2019; Prada i Boore, 2019). Osim toga, u procesu oplodnje očinski se mitohondriji uništavaju zbog čega se mtDNA nasljeđuje isključivo majčinskom linijom. Iz ovog razloga, u nasljeđivanju mtDNA izostavlja se proces rekombinacije (Clayton, 1992) što znači da bi mtDNA haplotipovi trebali biti zajednički svim jedinkama unutar majčinske linije (Hutchinson i sur., 1974; Gupta i sur., 2015).

Kako je već spomenuto, mitohondriji za razliku od drugih organela u svojoj jezgri imaju nekoliko kopija vlastite DNA molekule (Robin i Wong, 1988). Zbog toga vrijednost pokrivenosti u genomskim podacima neće biti jednaka za nDNA i mtDNA. Drugim riječima, prilikom sekvenciranja čitavih genoma, mtDNA fragmenti se češće pojavljuju među rezultirajućim fragmentima i s većim brojem kopija. Broj fragmenata mtDNA u genomskim uzorcima je veći od broja nDNA fragmenata zbog čega se mtDNA fragmenti mogu lakše izolirati i koristiti za rekonstrukciju cijele mtDNA (Al-Nakeeb i sur., 2017). Drugim riječima, ako je u nekom sekvenciranju nDNA pokrivenost 5x, onda će u istom rezultatu pokrivenost mtDNA biti veća i dovoljna za sastavljanje.

2.3 Sastavljanje genoma

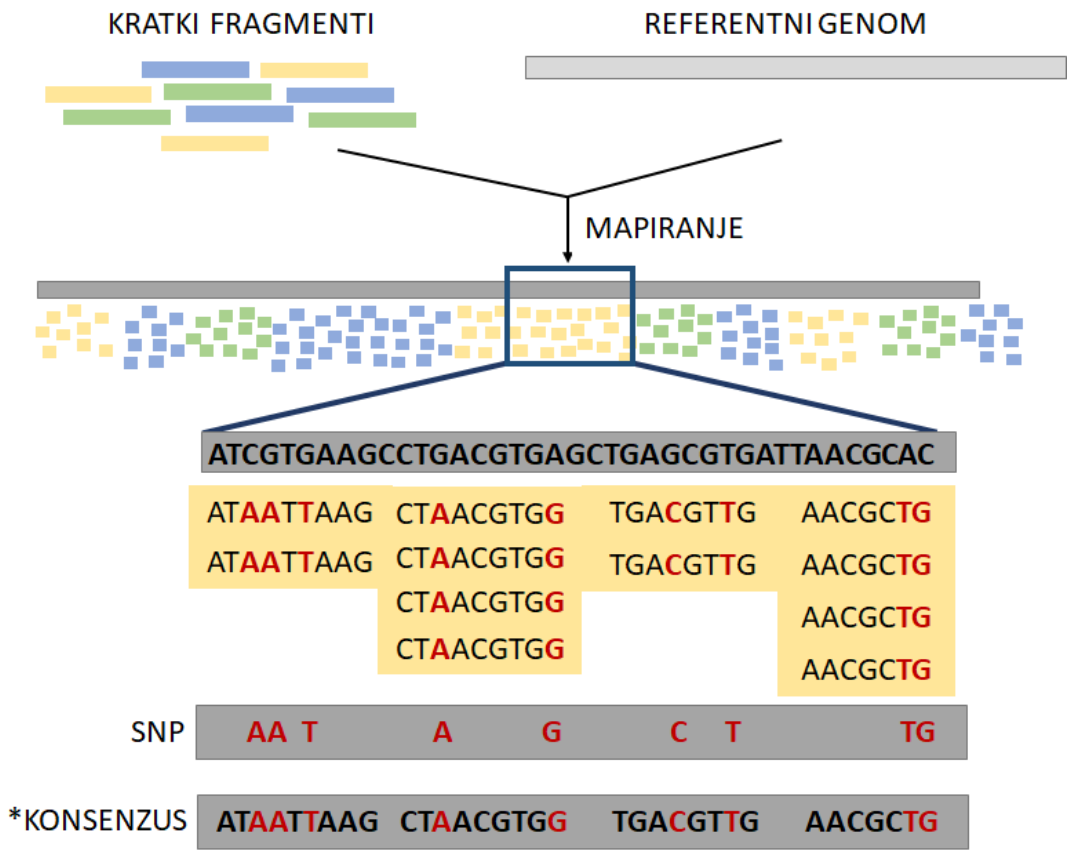
Sastavljanje genoma (engl. *genome assembly*) proces je u kojem se iz velikog broja sekvenciranih kratkih fragmenata nastoji rekonstruirati duga sekvenca korištenjem računalnih programa tzv. asemblera (engl. *assemblers*; programi koji sastavljaju genom metodom *de novo*) ili mapera (engl. *mappers*; programi koji sastavljaju genom metodom mapiranja na referentnu sekvencu) (Schatz i sur., 2010; Fuentes-Pardo i Ruzzante, 2017; Young i Gillung, 2020). U posljednih nekoliko godina aktualna je i treća, hibridna metoda sastavljanja genoma gdje se genomi nastoje rekonstuirati kombinacijom spomenute dvije metode (Buza i sur., 2015; Lischer i Shimizu, 2017; Young i Gillung, 2020). Odabir metode za sastavljanje genoma ovisi o nekoliko bitnih čimbenika, a najvažniji su: kvaliteta i količina genomskih podataka, broj uzoraka, pokrivenost i dostupnost referentne sekvence (Dominguez Del Angel i sur., 2018).

2.3.1 Sastavljanje genoma mapiranjem na referentni genom

Za sastavljanje mapiranjem koriste se računalni programi maperi koji uspoređuju svaki fragment s referentnom sekvencom mapirajući ga na odgovarajuće mjesto. Drugim riječima, cilj ove metode je prvo pronaći stvarnu lokaciju svakog fragmenta (iz seta od nekoliko stotina milijuna fragmenata) na referentnoj sekvenci, te potom detektirati sva varijabilna mjesta između fragmenata i referentne sekvence (Davey i sur., 2011; Pfeifer, 2017). U isto vrijeme maper za svaku pronađenu varijantu mora moći prepoznati radi li se zaista o stvarnoj varijanti ili o pogreški. Zbog toga je mapiranje računalno zahtjevan proces koji se razvojem metoda nastoji optimizirati. U rekonstrukciji genoma ova metoda se koristi kada je referentna sekvenca vrste koja je predmet istraživanja dostupna u Banci gena te kada pokrivenost genomskih podataka nije dovoljno velika za korištenje *de novo* metode (Martin i Wang, 2011; Ekblom i Wolf, 2014). Maperi se intezivno počinju razvijati od 2008. godine i njihov je razvoj uglavom vezan za razvoj NGS tehnologija (Fonseca i sur., 2012). Budući da maperi rade sa stotinama milijuna fragmenata, optimizacija cijelog procesa podrazumijeva adaptaciju na genomske podatke (tip podataka, duljina fragmenata i sl.), brzinu, točnost te smanjenje stupnja pogreške (Pfeifer, 2017). Testiranje više desetaka mapera opisano je u nekoliko preglednih radova (Li i Homer, 2010; Holtgrewe i sur., 2011; Fonseca i sur., 2012; Ruffalo i sur., 2012; Schbath i sur., 2012; Hatem i sur., 2013). Iz takvih analiza teško je odrediti koji je maper najpogodniji prvenstveno iz razloga što se testiranje mapera često provodi korištenjem simuliranih genomskih podataka, kratkih referentnih genoma (npr. bakterijski genom) ili genoma visoke kvalitete (npr. ljudski genom) što u praksi nije uvijek slučaj (npr. korištenje mapera u projektima gdje su genomski podaci ali i same reference upitne kvalitete) (Hatem i sur., 2013; Pfeifer, 2017). Kod odabira referentnog genoma preporučuje se korištenje reprezentativnog genoma visoke kvalitete. On može pripadati vrsti koja je predmet istraživanja ili se može koristiti genom srodne vrste (Gnerre i sur., 2009; Nevado i sur., 2014; Wang i sur., 2014; Fuentes-Pardo i Ruzzante, 2017). Kvalitetan genom nije uvijek dostupan, pogotovo kod nemodelnih vrsta čiji je referentni genom ujedno i reprezentativan genom za tu vrstu a rekonstruiran je sekvenciranjem samo jedne jedinke. Korištenje takvog genoma u analizi može uzrokovati pristranost i gubitak varijanti u uzorcima prilikom njihovog mapiranja na taj genom (Huang i sur., 2013; Nevado i sur., 2014; Smolka i sur., 2015; Gopalakrishnan i sur., 2017; Prasad i sur., 2021).

Metoda mapiranja se u pravilu sastoji od četiri koraka (Slika 2.): provjera ulaznih podataka (engl. *quality control and data preprocessing*), poravnanje kratkih očitanja na referencu (engl.

alignment, *mapping*), analiza rezultata poravnanja te pozivanje strukturnih varijanti i njihovo filtriranje (engl. *variant calling and filtering*) (Pfeifer, 2017).



Slika 2. Shematski prikaz metode mapiranja. Kratki DNA fragmenti se mapiraju na referentnu sekvencu. Potom se detektiraju sve razlike (najčešće SNP-ovi) koje se zapisuju u VCF formatu i koriste u daljnjim analizama. Dio slike označen zvjezdicom (*KONSENZUS) je korak koji se koristio u ovoj disertaciji, a odnosi se na pozivanja konsenzusne sekvence koja predstavlja kombinaciju reference i detektiranih (filtriranih) SNP-ova.

Pregledom ulaznih genomskih podataka i provjerom kvalitete ispravljaju se pogreške nastale u procesu sekvenciranja (adapteri, kontaminacije, duplicirani fragmenti i sl.) a najzastupljeniji alat kojim se provodi ovaj korak je FASTQC (Andrews, 2010). U drugom koraku kratki fragmenti se mapiraju duž referentnog genoma. Ovaj proces je računalno i vremenski najzahtjevniji i u njemu postoji najveći rizik za pojavu pogrešaka. Iz ovog razloga se

uglavnom koriste najzastupljeniji maperi kao kao što su BWA (Li i Durbin, 2009), SPAdess (Bankevich i sur., 2012), Bowtie (Langmead i sur., 2009), Bowtie2 (Langmead i Salzberg, 2012) i sl. Rezultate poravnatih fragmenata maperi pohranjuju u SAM (engl. *Sequence Alignment Format*) datoteku i u njenu binarnu verziju BAM (engl. *Binary Alignment Map*) u kojima se nalaze informacije o lokaciji, orijentaciji i kvaliteti za svaki fragment (Li i sur., 2009). Za ovaj korak također je razvijen set alata čiji je glavni zadatak manipulacija SAM/BAM datotekama među kojima su najpoznatiji SAMtools (Li i sur., 2009) i Picard (http://broadinstitute.github.io/picard) iz GATK paketa (McKenna i sur., 2010). Ovim korakom se, osim navedenih informacija, dobiju i statističke informacije koje pružaju bolji uvid u rezultate mapiranja (broj mapiranih fragmenata, broj nemapiranih fragmenata, broj lokacija na koje se ni jedan fragment nije mapirao i sl.). Osim toga, dostupno je nekoliko alata namijenjenih za vizualizaciju cijelog poravnanja s kojima se može vidjeti detaljnija slika rezultata mapiranja: Broad Institute's Integrative Genomics Viewer (Robinson i sur., 2011), Tablet (Milne i sur., 2010) i neki drugi.

Pronalazak strukturnih varijanti u genomu i njihova točna identifikacija posljednji je korak koji, uz sve navedeno (izbor mapera, izbor reference, kvaliteta ulaznih podataka i dr.), ovisi i o odabiru alata za provođenje ovog koraka, ali i filtriranja dobivenih rezultata (Altman i sur., 2012; Pfeifer, 2017). Pozivanje varijanti proces je pomoću kojeg specijalizirani programi, tzv. caller-i (engl. *callers*) pronalaze varijabilna mjesta između mapiranih fragmenata i reference koja se potom zapisuju u VCF datoteku (engl. *Variant Call Format*) ili njenu binarnu verziju (BCF) (Dancek i sur., 2011). Pronađene razlike po svojoj strukturi mogu biti SNP-ovi (engl. *Single Nucleotide Polymorphysm*), insercije ili delecije (engl. *indels*, indelsi) te varijante u broju kopija (engl. *copy number variation,* CNVs) (Nielsen i sur., 2011). SNP-ovi su najčešći oblik strukturnih razlika i npr. kod čovjeka čine 90 % svih genetičkih varijanti ili polimorfizama koji utječu na fenotipsku raznolikost (Collins i sur., 1998). Međutim, u procesu pozivanja varijanti postoji mogućnost identificiranja lažno pozitivnih (engl. *false-positives*) i to se događa kao posljedica pogrešaka nastalih tijekom mapiranja. Detekcija varijanti bazira se na identifikaciji polimorfizama na razini jedne baze što znači da se svaka pogreška u sekvenciranju može prepoznati kao netočni (lažno pozitivni) SNP (Yu i Sun, 2013; Pfeifer, 2017). Za detekciju strukturnih promjena postoji veliki broj dostupnih alata. Ne koriste svi alati isti algoritam i prema tome, neće svi alati dati jednake rezultate. Razvoj caller-a također je pratio razvoj NGS tehnologija ali i mapera s obzirom da su rezultati mapera zapravo ulazni podaci ovim alatima. Ovo je još jedan razlog zbog čega dobiveni rezultati neće biti jednaki te

će, uz definiranje velikog broja parametara, ovisiti i o kombinaciji mapera i caller-a. Više desetaka ovakvih kombinacija testirano je u nekoliko navrata na različitim vrstama podataka (Jia i sur., 2012; Liu i sur., 2013; Yu i Sun, 2013; O'Rawe i sur., 2013). Zaključak je bio sličan kao i u radovima koji su testirali različite mapere: performanse svakog pojedinog caller-a ili njega u kombinaciji s maperom ovise o puno parametara, ali prije svega o dizajnu samog istraživanja. Ne može se reći da je neka kombinacija bolja od neke druge u svim mogućim scenarijima i pri korištenju svih kriterija. Uz navedeno, pokrivenost genomskih podataka i u ovom koraku ima veliki utjecaj na rezultate. S većom pokrivenošću veća je i vjerojatnost da je pronađena varijanta stvarna (Yu i Sun, 2013; Pfeifer, 2017). Jedan od mogućih načina kojima se riješava problem male pokrivenosti je povećanje broja uzoraka te pozivanje varijanti iz svih uzoraka istovremeno (engl. *joint variant calling*) (Nielsen i sur., 2011; Yu i Sun, 2013). Međutim, prilikom korištenja podataka manje pokrivenosti maperi nailaze na još jedan problem koji se godinama nastoji riješiti, a to je sastavljanje ponavljajućih regija (engl. *repetitive elements*). Ponavljajuće regije predstavljaju duže ili kraće kopije fragmenata koje se nalaze na različitim lokacijama u genomu. Ponekad te regije mogu biti puno duže od sekvenciranih genomskih fragmenata zbog čega se pojedini fragmenti mogu mapirati na nekoliko različitih mjesta u genomu (Schatz i sur., 2010; Pfiefer, 2017). Ovaj problem posebno je prisutan u genomu sisavaca gdje su ponavljajuće regije prisutne između 25 i 50 % (Kazazian, 2004), a može ih se podijeliti na: dugačke isprekidane nuklearne elemente (engl. *long interspersed nuclear elements*), kratke isprekidane nuklearne elemente (engl. *short interspersed nuclear elements*), dugačka terminalna ponavljanja (engl. *long terminal repeats*) i jednostavna tandem ponavljanja (engl. *short tandem repeats*) (Cordaux i Batzer, 2009).

Neovisno radi li se sastavljanju nDNA ili mtDNA, procedura mapiranja je jako slična i uglavnom se koriste isti maperi i caller-i. Jedine bitne razlike su vrijeme, kompleksnost računanja i pohranjivanje rezultata (procesi za nDNA su puno kompliciraniji nego kod mtDNA). Prilikom sastavljanja mtDNA, fragmenti iz uzorka se mapiraju na referenetni mitohondrijski genom koji je kod sisavaca u prosjeku dug 16 kb, dok kod sastavljanja nDNA, maper treba pronaći lokaciju fragmentima duž referentne sekvence koja u prosjeku iznosi 3 Gb (Gupta i sur., 2015). Budući da je pokrivenost mtDNA fragmenata puno veća, a sama mtDNA je jako kratka sekvenca, proces pozivanja varijanti je puno jednostavniji. Postoji jedan alternativni pristup sastavljanja mtDNA koji se često koristi u komercijalnim programima za sastavljanje genoma kao što su Geneious Prime® (Kearse i sur., 2012) i CLC (CLCbio,

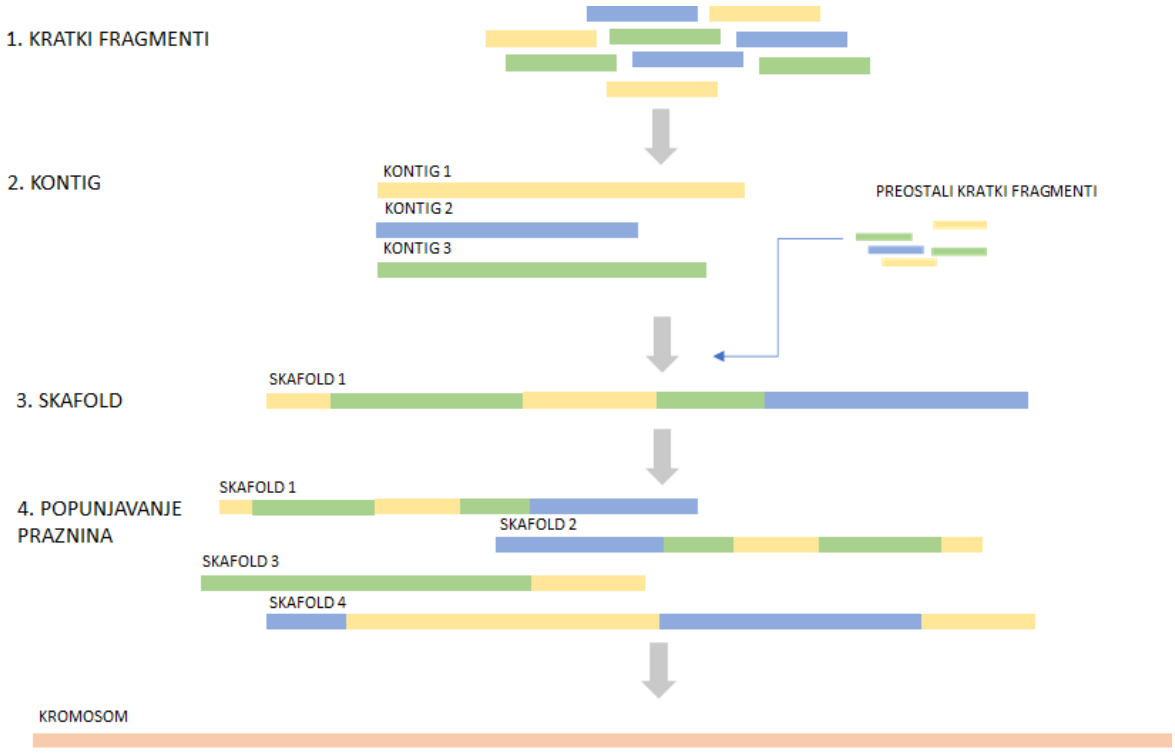
Aarhus, Denmark) (Miller i sur., 2012). U ovom se procesu ne dobiva VCF datoteka sa svim pronađenim varijantama nego se poziva cijela konsenzusna sekvenca (engl. *consensus sequence*) iz poravnanja koje sadrži konsenzuse svih nukleotida iz fragmenata i reference. Drugim riječima, ako je u referenci na određenoj lokaciji baza A, a na to mjesto se mapiralo 15 fragmenata koji na tom istom mjestu imaju bazu G, konsenzusna baza na ovoj poziciji će biti G jer je prisutnost baze G potvrđena 15x.

2.3.2 Sastavljanja genoma *de novo* metodom

Sastavljanje genoma *de novo* je metoda u kojoj se genomske sekvence sastavljaju iz velikog broja (kratkih ili dugih) DNA fragmenata bez prethodne informacije o ispravnom redoslijedu tih fragmenata. Prvi pokušaji rekonstrukcije genomske sekvence bazirali su se na iterativnom preklapanju genomskih fragmenata (Simpson i Pop, 2015). Računalni programi koji se koriste za sastavljanje sekvenci *de novo* metodom nazivaju se asembleri (engl. *assemblers*), a glavni im je cilj dobiti informaciju iz genomskih podataka slaganjem nukleotida u ispravan redoslijed (Li i sur., 2010b). Razvoj asemblera u zadnjih deset godina povezan je s razvojem NGS tehnologija. Prve generacije asemblera većinom su koristile OLC algoritam (engl. *overlap-layout-consensus*) kojeg je predstavio Staden (1979), a kasnije je implementiran u mnoge druge asemblere (Batzoglou i sur., 2002; Huang i Yang, 2005). Princip rada ovog algoritma sastoji se od tri koraka (Liao i sur., 2019). Prvo se pronalaze sva moguća preklapanja između fragmenata te se potom gradi nacrt koji sadrže sve složene fragmente u obliku grafa. U zadnjem se koraku stvara konsenzusni slijed na temelju grafa. Ovaj pristup uspješan je kod sastavljanja genomske sekvence iz malog broja dužih fragmenata, odnosno kod genomskih podataka dobivenih trećom generacijom sekvenciranja (dugi fragmenti). Međutim, korištenje ovog pristupa na podacima s velikim brojem kratkih fragmenta dugotrajan je proces zbog čega se javila potreba za razvojem algoritama specijaliziranih za veće količine genomskih podataka (Rice i sur., 2019). Asembleri druge generacije koriste algoritam DBG (engl. *de Bruijn graph*) koji je predstavljen 1995. godine (Idury i Waterman, 1995), a radi na principu da sve kratke fragmente prvo reže na manje dijelove, tzv. *k-mer*-ove koje potom koristi za izgradnju grafa i za slaganje genomskih sekvenci. Prvih nekoliko godina interes za korištenje i implementaciju ovog grafa u asemblere je bio vrlo nizak sve dok se tehnologije druge generacije nisu probile na tržište. Od tad, DBG algoritam se uspješno koristio, prvenstveno kod sastavljanja manjih genoma. Nedugo nakon toga, upotreba DBG algoritma proširila se njegovom implementacijom u najmodernije asemblere za sastavljanje velikih eukariotskih

genoma koji se i danas koriste: Velvet (Zerbino i Birney, 2008), ABySS (Simpson i sur., 2009), SOAPdenovo (Li i sur., 2010) i AllPath-LG (Gnerre i sur., 2011) i dr.

Sastavljanje genoma metodom *de novo* uglavnom se koristi prilikom rekonstruiranja genoma za vrste čiji genom nije sekvenciran i sastavljen. Kako je već spomenuto, ova metoda zahtjeva visoku kvalitetu genomskih podataka velike pokrivenosti, a izbor asemblera prvenstveno ovisi o vrsti i veličini sekvenciranih fragmenata. Cijeli proces DBG algoritma sastoji se od tri osnovna koraka (Compeau i sur., 2011; El-Metwally i sur., 2013; Simpson i Pop, 2015) (Slika 3.). U prvom se koraku kratki fragmenti, izrezani na k-merove, slažu u duže konsenzusne sekvence ili kontige (engl. *contigs*) koji se potom povezuju u još veće sekvence tzv. skafolde (engl. *scaffolds*). Potom slijedi korak koji se naziva popunjavanje praznina (engl. *gap-filling*). Praznine su mjesta koja su nastala spajanjem dvaju (ili više) skafolda i kojima fali informacija o nukleotidu na određenom mjestu, a popunjava se preostalim kratkim fragmentima koje asembler nije uspio sklopiti u prvom koraku. Popunjavanje praznina kao i povezivanje kontiga u skafolde asembler provodi iterativno s ciljem slaganja što većeg broja preostalih fragmenata (Tsai i sur., 2010).



Slika 3. Shematski prikaz *de novo* metode. Kratki DNA fragmenti se metodom preklapanja prvo sastavljaju u kontige. Potom se kontizi skupa s preostalim fragmentima sastavljaju u skafolde. Veliki broj dobivenih skafolda se potom uspoređuje čime se popunjavaju nedostajuća mjesta (praznine) kako bi se dobile još veće sekvence koje predstavljaju kromosome i njihove dijelove.

Sva tri koraka su izuzetno računalno i vremenski zahtjevna a različiti asembleri imaju različite pristupe u optimizaciji pojedinih koraka. Bitno je naglasiti da je prije provođenja navedenih koraka, potrebno provjeriti kvalitetu ulaznih podataka jednako kao i kod metode mapiranja. U novije vrijeme za metodu *de novo* razvijaju se asembleri specijalizirani za manipulaciju genomskih podataka treće generacije sekvenciranja. Puno duži fragmenti koji se dobiju ovom tehnologijom se koriste kao zasebni ulazni podaci za nove projekte ili kao dopuna kod već sastavljenih genoma s namjerom da popune praznine. U pravilu, asembleri koji isključivo rade s dugim fragmentima koriste OLC algoritam.

U *de novo* metodi, za razliku od mapiranja, procedure sastavljanja mtDNA i nDNA se znatno razlikuju. Kako je već spomenuto, jedan od glavnih parametara je velika pokrivenost koju genomski podaci moraju imati kako bi se iz njih sastavio genom velike točnosti. Najpopularniji asembleri u početku se se koristili i za sastavljanje mtDNA. Međutim, takvi asembleri sastavit će mtDNA niže kvalitete zbog čega će se dobivene sekvence morati dalje dorađivati kako bi

se popravila njihova kvaliteta. Razlog tome je što asembleri namjenjeni sastavljanju nDNA ne mogu dobro manipulirati regijama na koje će se sastaviti veliki broj mitohondrijskih fragmenata zbog iznimno velike pokrivenosti (Meng i sur., 2019). Osim toga, nDNA asembleri nisu namjenjeni sastavljanju kružnih sekvenci zbog čega lošije prepoznaju ponavljajuće regije u mtDNA (Dierckxsens i sur., 2017; Meng i sur., 2019). Kako bi se izbjegli navedeni problemi, intenzivno su se počeli razvijati specijalizirani *de novo* asembleri za sastavljanje mtDNA. S obzirom na malu pokrivenost genomskih podataka u korištenim uzorcima, u ovoj disertaciji će se *de novo* metoda koristiti samo za sastavljanje mtDNA, i zbog toga će samo ona biti opisana.

2.3.3 Sastavljanje mtDNA *de novo* metodom

MtDNA asemblere je lakše optimizirati u usporedbi s asemblerima za sastavljanje nDNA zbog velike sličnosti mtDNA sekvenci kod sisavaca (struktura, duljina). Neki od najčešće korištenih asemblera razvijenih za sastavljanje mtDNA su MitoZ (Meng i sur., 2019), MITObim (Hahn i sur., 2013), NOVOPlasty (Dierckxsens i sur., 2017) i Norgal (Al-Nakeeb i sur., 2017). MitoZ i Norgal koriste sličan algoritam kao i nDNA asembleri dok NOVOPlasty i MITObim koriste podvrstu *de novo* algortima „seed-and-extend“. Ovaj algoritam koristi tzv. seed sekvencu koja predstavlja startnu poziciju novog genoma. Seed sekvenca može biti kratki gen (najčešće CYTB) ili cijela referetna mtDNA koja pripada bliskom srodniku (Dierckxsens i sur., 2017; Alqahtani i Madoiu, 2020). Ovi asembleri koriste seed sekvencu kao ulaznu i prema njoj iterativno skeniraju start i stop pozicije tvoreći na taj način tip podataka koji se zove hash tablica (engl. *hash table*). Hash tablica se koristi za indeksiranje genoma te sadrži listu genomskih pozicija za svaki fragment (Wu, 2016). Slični fragmenti će se zajedno grupirati i na taj način stvoriti kružnu sekvencu. Sekvenca će biti kompletna kada se oba kraja preklope s najmanje 200 nukleotida nakon čega se stvara konsenzusna sekvenca (Dierckxsens i sur., 2017).

Uz navedene asemblere, učestalo se koriste komercijalni programi za sastavljanje mtDNA (Geneious Prime i CLC) koji su puno efikasniji u procesu sastavljanja mtDNA u odnosu na nDNA jer se mogu pokrenuti na osobnom računalu (Miller i sur., 2012; Caparroz i sur., 2015; Hill et al., 2017).

Nakon procesa sastavljanja, provodi se procjena kompletnosti dobivene sekvence s ciljem pronalaska pogrešaka (npr. krivo sastavljene ponavljajuće regije, kriva orijentacija i sl.). Osim toga, postoji mogućnost da dobivena sekvenca ne sadrži samo informacije o željenoj mtDNA

nego i drugim sekvencama koje su se potencijalno mogle greškom uklopiti u sastavljenu mtDNA (dijelovi nDNA ili dijelovi DNA koja pripada nekoj bakteriji i sl.). Zbog toga se dobivena konsenzusna sekvenca uspoređuje s mtDNA bliskog srodnika pomoću BLAST algoritma (Altschul i sur., 1990). BLAST će na temelju sličnosti potvrditi kompletnost nove sekvence te će kao rezultat dati listu sličnih sekvenci dostupnih u Banci gena. Za dodatnu provjeru moguće je provesti i metodu mapiranja tih istih fragmenata na novu dobivenu mtDNA sekvencu. Postoji i treći način provjere kompletnosti sekvence gdje se nova sekvenca poravna s nekoliko sekvenci koje pripadaju bliskom srodniku. Zajedničkom poravnanjem (engl. *alignment*) može se dodatno provjeriti postoje li mjesta u novoj sekvenci koja su se potencijalno krivo složila.

2.3.4 Hibridno sastavljanje genoma

Kako je već prethodno spomenuto, s povećanjem broja genomskim podataka, porasla je i potreba za pronalaženje metode kojom će se najlakše doći do točne informacije o genomu vrste koja se proučava. Postoji nekoliko vrsta hibridnog sastavljanja genoma (engl. *reference-guided assembly*) koje su zapravo kombinacija mapiranja i *de novo* i koriste se za sastavljanje nDNA. U većini slučajeva fragmenti se prvo sastave *de novo* u duže sekvence. Potom se tako dobivene sekvence mapiraju na genom srodne vrste na način da ih se posloži i orijentira te posloži u kromosome prema kromosomima iz referentne sekvence (Vezzi i sur., 2011; Bao i sur., 2014). Ovu metodu je moguće provesti i na genomskim podacima male pokrivenosti (Vezzi i sur., 2011; Card i sur, 2014). Zbog korištenja referentnog genoma potrebno je imati na umu da ova metoda ima i svoje nedostatke kao što su: pristranost rezultata te nemogućnost sastavljanja alternativnih regija u novom genomu jer iste nisu prisutne u referentnom. Uz navedeno, veliki problem predstavlja pokretanje prvog koraka, *de novo* metode, na podacima male pokrivenosti prilagodbom velikog broja parametara. Posljednih nekoliko godina intenzivno se radilo na rješavanju navedenih problema zbog čega je i za ovu metodu razvijeno i predloženo nekoliko programskih alata i metodologija (Tsai i sur., 2010; Vezzi i sur., 2011; Kim i sur., 2013; Bao i sur., 2014; Kolmogorov i sur., 2014; Wang i sur., 2014; Buza i sur., 2015; Tamazian i sur., 2016; Lischer i Shimizu, 2017; Kolmogorov i sur., 2018; Siddiki i sur, 2019). Međutim, većina programskih alata razvijena je i testirana na relativno jednostavnim bakterijskim genomima (Tamazian i sur., 2016). Osim toga, neki od predloženih alata traže, uz referentni genom, tzv. mate-pair fragmente (Kim i sur., 2013). Ovi fragmenti nastaju specijaliziranom tehnikom sekvenciranja istog uzorka gdje se uz normalne kratke fragmente, dodatno sekvenciraju veći genomski fragmenti (duljine oko

1000 parova baza (pb)) koji se potom koriste kao pomoć u *de novo* metodama. Međutim, često je slučaj da genomski podaci ne sadrže ovakav tip podataka zbog čega se neke od ovih metoda ne mogu koristiti.

2.4 Anotacija

Anotacija genoma postupak je kojim se dijelovi genoma povezuju s biološkim funkcijama. Za dobivanje točne informacije o genomu, novosastavljene sekvence potrebno je identificirati i anotirati. Proces anotacije također je vremenski i računalno zahtjevan proces čiji rezultati prije svega ovise o kvaliteti i kompletnosti novosastavljenog genoma (otprlike 90 % kompletnosti genoma smatra se dovoljnim za dobivanje zadovoljavajućih rezultata anotacije kod velikih genoma) (Yandell i Ence, 2012; Ejigu i Jung, 2020). Programski alati prvo analiziraju strukturu i sastav sekvence te je uspoređuju s poznatim genomima srodnih vrsta (Dominguez Del Angel i sur., 2018). Proces se sastoji od dva koraka: 1) strukturna anotacija kojom se identificiraju elementi u genomu (kodirajuća područja i genska struktura); 2) funkcionalna anotacija kojom se biološke informacije vežu na genomske elemente (Ekblom i Wolf, 2014). Strukturna anotacija podrazumijeva pronalaženje DNA struktura što uključuje egzone, introne, promotorna mjesta itd. U prvom koraku se provodi predikcija gena (engl. *gene prediction*) u kojem se prvo identificiraju kodirani geni za koje postoji velika vjerojatnost da će biti pronađeni u novoj sekvenci (Wang i sur., 2004). Ovim se procesom pronalaze lokacije i strukture gena u genomu tako da se nukleotidne sekvence prvo translatiraju te se potom pronalaze otvoreni okviri čitanja (engl. *Open Reading Frames*, ORF) (Dominguez Del Angel i sur., 2018). ORF je frakcija DNA molekule koja, kada se translatira u aminokiselinski zapis, ne sadrži stop kodon. Međutim, različita duljina i raspored introna skupa s alternativnim spajanjem egzona (engl. *alternative splicing*) uvelike otežava ovaj proces.

Postoje dvije metode predikcije: predikcija temeljena na sličnosti (engl. *similarity-based*, *homology-based*) i *ab initio* metoda. Prva metoda nastoji pronaći slične regije gena između sekvenci a temelji se na pretpostavci da su egzoni evolucijski konzerviraniji u usporedbi s intronima i drugim nefunkcionalnim regijama. Anotatori koji koriste ovu metodu predviđaju gene na način da se sekvence poravnavaju s već poznatim anotiranim sekvencama iz dostupnih baza podataka. Poznate anotirane sekvence mogu biti tzv. *expressed sequence tags* (EST) i komplementarne DNA (cDNA) (Ejigu i Jung, 2020). S druge strane, metoda *ab initio* koristi već poznatu strukturu gena kao predložak za detekciju gena u novoj sekvenci korištenjem dva tipa informacija: signalnih senzora (engl. *signal sensors*) i sadržajnih senzora

(engl. *content sensors*). U signalne senzore spadaju kratke sekvence karakteristične za svaki gen a to su mjesta spajanja (engl. *splice sites*), start i stop kodoni i sl. Sadržajni senzori koriste se u detekciji egzona na način da se kodirajuće regije mogu izdvojiti od nekodirajućih (Wang i sur., 2004). *Ab initio* metoda korištenjem matematičkih modela identificira gene zajedno s njihovim intron-egzon dijelovima te koristi informacije o genomu (npr. frekvencija kodona, duljina intron-egzon regija) s ciljem da prepozna gene u novoj sekvenci (Korf i sur., 2004). Korištenje ovakvih anotatora zahtjeva „treniranje“ podataka uzimajući u obzir informacije o specifičnosti organizma kojeg se anotira (frekvencija kodona, duljina i distribucija introna i egzona otd.) u kombinaciji s genskim modelima. Uz navedeno, za bolju točnost predikcije *ab inition* metoda također zahtjeva sekvence EST te sekvence poznatih proteina iz baza ali i RNA-seq podatke (sekvencirani genomski podaci RNA) (Ejigu i Jung, 2020).

Nakon strukturne anotacije, predviđenom proteinu i njegovim derivatima (geni, RNA i sl.) dodjeljuje se biološke informacije pomoću funkcionalne anotacije. Ranijih godina, glavni fokus u procesu funkcionalne anotacije bile su kodirajuće sekvence proteina. Međutim, posljednjih godina s detekcijom različitih transkripta, počeli su se razvijati anotatori koji u obzir uzimaju i varijante različitih gena. Njihov glavni cilj je identificirati i prioritetizirati stvarne varijante gena na temelju njihove funkcije (Cutting, 2014). Identifikacija ovakvih varijanti vrši se u BLAST alatu koji se implementiran u gotovo sve popularne anotatore. BLAST novu sekvencu nastoji poravnati na dostupne sekvence u Banci gena na temelju sličnosti. Na ovaj način se nastoje utvrditi evolucijski odnosi između sekvenci.

2.4.1 Anotacija mtDNA

S obzirom na jednostavnu strukturu i veliku sličnost mtDNA među sisavcima, proces njene anotacije je puno jednostavniji od anotacije nDNA. Zbog toga se posljednjih 15 godina intenzivno radi na automatizaciji cijelog procesa. Najčešće korišteni programski alati za anotaciju mtDNA su DOGMA (Wyman i sur., 2004); MITOS (Bernt i sur., 2013), GeSeq (Tillich i sur., 2017) i MitoZ (Meng i sur., 2019). DOGMA, MITOS i GeSeq su u obliku web aplikacija dok je MitoZ softwer otvorenog koda (engl. *open source software*). Većina navedenih alata kao ulaznu datoteku, uz sekvencu koju se želi anotirati, zahtjeva i sekvencu bliskog srodnika u FASTA formatu jer se dobivene sekvence uspoređuju s referentnom sekvencom, ali i sa sekvencama iz Banke gena. GeSeq ima dodatnu opciju kojom, uz FASTA datoteku srodnika, može prihvatiti i GenBank datoteku kao ulazni format koji sadrži

informacije o anotiranim regijama referentnog genoma. Najbrži alat za anotaciju mtDNA je GeSeq koji može anotirati nekoliko sekvenci istovremeno. Osim toga, GeSeq koristi BLAT algoritam koji, za razliku od BLAST-a, prilikom identificiranja kodirajućih i nekodirajućih regija, u obzir uzima i informacije o ORF-ovima čija je lista također dostupna u Banci gena. tRNA scan-SE (Lowe i Eddy, 1997; Lowe i Chan, 2016) program se koristi za anotaciju tRNA gena i on je implemenetiran u navedenim alatima (Cui i sur., 2007; Jiang i sur., 2013; Matosiuk i sur., 2014; Zhou i sur., 2019). Nakon procesa anotacije potrebno je provjeriti sve identificirane regije i usporediti ih s rezultatima anotacije srodnih vrsta (ukoliko su dostupni). S obzirom da indels-i mogu uzrokovati pojavu stop kodona usred kodirajućih regija, potrebno je provjeriti jesu li se stop kodoni ispravno anotirali provjerom njhovih lokacija i duljina (Bernt i sur., 2013). Izlazna datoteka alata za anotaciju u većini slučajeva je GFF format u kojem su zapisane informacije od strukturi i funkciji pronađenih regija. Neki programi također mogu zapisati informacije u GTF, BED i GenBank formatu (Dominguez Del Angel i sur., 2018)

2.4.2 Anotacija nDNA

Anotacija nDNA računski i vremenski je kompleksan proces budući da alati anotatori pronalaze strukturu i funkciju gena duž velike sekvence. Iako je ovaj proces kompliciraniji nego kod mtDNA, anotacija nDNA može se odraditi alatima za provođenje automatizirane anotacije (Yandell i Ence, 2012).

Jedan od najkorištenijih alata za automatsku anotaciju nDNA je BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão i sur., 2015; Manni i sur., 2021a; Manni i sur., 2021). Ranije verzije BUSCO alata većinom su korištene za procjenu kvalitete novosastavljenih genoma, a glavni genski prediktor bio je Augustus (Stanke i sur., 2006). Augustus u procesu genske predikcije koristi bazu ortolognih gena OrthoDB na temelju koje skenira i pronalazi gene u novosastavljenim sekvencama. Augustus zatim detektira pronađene sekvence koje potom prolaze proces translacije pomoću HMMER alata koji je također implementiran u BUSCO-u (Simão i sur., 2015). Međutim, za velike genome, Augustus-u je potrebno nekoliko dana da provede anotaciju, što povećava ukupno vrijeme potrebno za provedbu analiza. Zbog toga se u najnovijoj verziji BUSCO-a (verzija 5) kao prediktor koristi MetaEuk koji ubrazava cijeli proces te poboljšava analizu velikih genoma (Manni i sur., 2021a; Manni i sur., 2021b). Osim toga, MetaEuk omogućuje procjenu velike količine podataka (genomskih, metagenomskih i transkriptomskih), uključujući eukariotske, prokariotske i virusne podatke (Manni i sur. , 2021a; Manni i sur., 2021b). U novije se vrijeme

pokazalo da je razvojem BUSCO-a omogućeno korištenje BUSCO gena u filogenomske svrhe što je testirano u nekoliko navrata na genomima insekata (Zhang i sur., 2019.; Dias i sur., 2020. ; Sun i sur., 2020.; Wang i sur., 2021.) i kvasaca (Shen i sur., 2016a; Shen i sur., 2020).

2.5 Filogenija

Filogenija opisuje odnose između gena, genoma, vrsta i drugih taksonomskih jedinica na temelju fenotipskih i molekularnih podataka te prikazuje rezultat u obliku filogenetskog stabla (Young i Gillung, 2020). Drugim riječima, na temelju pretpostavki o zajedničkom pretku filogenijom se nastoje prikazati evolucijski odnosi između taksonomskih jedinica te opisati događaje uključene u njihovu evoluciju. Zaključci su se u filogeniji dugo donosili na temelju fenotipskih (morfološki, fiziološki, behavioralni) podataka sve do kraja 70tih godina kada su otkriveni PCR (lančana rekacija polimerazom, engl. *Polymerase Chain Reaction*) i Sangerovo sekvenciranje (Brocchieri, 2000). Razvojem PCR-a, rekonstrukcija filogenije se najviše temeljila na podacima jednog ili nekolicine nuklearnih i mitohondrijskih gena tipično generiranih PCR-om (dijelovi DNA i RNA proteina,16S i 18S) (Fox i sur., 1980; Woese, 1987; Field i sur., 1988; Kapli i sur., 2020; Young i Gillung, 2020). Zbog korištenja tih istih gena iz različitih izvora podataka često je dolazilo do nepodudarnosti zaključaka u znanstvenim radovima (Brocchieri, 2001).

Načelo homologije bazira se na činjenici da dva gena dijele zajedničko prodrijetlo i sličnu biološku funkciju. Drugim riječim, na temelju homologije može se pretpostaviti da, ako se zna funkcija nekog gena, onda se može pretpostaviti da i njemu sličan gen ima sličnu biološku funkciju (Fitch, 1970; Sjölander, 2004). Ovo se načelo intenzivno koristi u genomici, pogotovo kod novosastavljenih sekvenci genoma. U pravilu se nove sekvence uspoređuju s dostupnim bazama sekvenci, gena ili proteina (Sjölander, 2004). Ortologija je vrsta homologije i podrazumijeva da su dva gena ortologna ako se pojavljuju u različitim vrstama, a potekli su od zajedničkog pretka procesom specijacije. S druge strane, paralogni geni se odnose na dva gena koja imaju zajedničkog pretka, a nastali su u vrstama procesom duplikacije. Ortolozi imaju veću vjerojatnost da imaju sličniju funkciju i zbog toga se koriste u genomskim analizama u kojima se informacija o proteinu iz jedne vrste koristi kao funkcionalna anotacija ortolognog proteina u drugoj vrsti (Hulsen i sur., 2006). Young i Gillung (2020) su u svom preglednom radu opisali da postoje dvije metode koje se koriste kod predikcije ortologa: 1) metoda temeljenje na sličnosti ili grafovima (engl. *similarity-or graph-based methods*) koje

pronalaze lokuse koje potom raspoređuje u grupe ortologa prema načelu ortologije (Altenhoff i Dessimoz, 2009); 2) metode temeljenje na filogeniji (engl. *phylogeny-based methods*) kojima je cilj identificirati ortologe na temelju najniže povezanosti gena od posljednjeg zajedničkog pretka (Gabaldón, 2008). Iako su obje metode vremenski i računalno iznimno zahtjevne, metode bazirane na filogeniji zbog svoje specifičnosti imaju veću moć detekcije ortologa među udaljenijim jedinicama (Altenhoff i Dessimoz, 2009).

Iako su filogenetska istraživanja jedna su od najučestalijih analiza u poljima biologije, Magee (2014) te Young i Gillung (2020) svojim su se radovima dotakli problematike reproducibilnosti filogenetskih analiza. Magee (2014) smatra da je otplike 60 % objavljenih filogenetskih analiza gotovo nemoguće replicirati. Glavni razlog je nedostupnost podataka korištenih u rekonstrukciji te nedovoljno informativan opis korištenih metoda. S druge strane, filogenetske analize osjetljive su na pogreške i pristranost koji se u većini slučajeva događaju kao posljedica metode a riječe kao nedostatak podataka (Kumar i sur., 2012) što se također treba uzeti u obzir prilikom usporedbe. Genomi i genomske sekvence u usporedbi s pojedinačnim nuklearnim genima mogu dati više točnih informacija u filogenetskim istraživanjima (Phillippe i Telford, 2006; Young i Gillung, 2020). S razvojem NGS tehnologija porastao je i broj dostupnih genomskih sekvenci zbog čega su se počele razvijati novije metode ze rekonstrukciju filogenije. Provedba filogenetskih analiza bazira se na poravnanju nukleotidnih ili aminokiselinskih sljedova iz kojeg se rekonstruira filogenetsko stablo. U pravilu ne postoji stvarno stablo koje prikazuje stvarne evolucijske događaje, nego se prema nekoj hipotezi evolucijskih događaja generira stablo (engl. *inferred tree*) na temelju dostupnih molekularnih podataka i odabranog evolucijskog modela. Broj genomskih sekvenci koje se koriste u filogenetskim istraživanjima u konstantnom je porastu. Međutim, to ne znači da se mogu opisati i prikazati odnosi svih vrsta životinja u drvu života (engl. *tree of life*). Zbog toga se preporuča detaljna selekcija lokusa koji bi mogli biti ključni za rješavanje odnosa između grupa (Young i Gillung, 2020). S druge strane, provedba takve selekcije je otežana i često se postavlja pitanje korištenja kodirajućih ili nekodirajućih regija, konzerviranih ili visoko varijabilnih lokusa, dužine poravnjanja i slično (Betancur-R i sur., 2014; Chen i sur., 2017). Korištenje genomskih podataka zbog toga može utjecati na nepodudarnost koja se događa kao posljedica korištenja gena koji imaju kompleksne evolucijske procese ili kao posljedica pogrešaka u metodologiji (Betancur-R i sur., 2014). Osim toga, kod korištenja genomskih podataka potrebno je posebno obratiti pažnju na pokrivenost genomskih uzoraka te na odabir metode za sastavljanje genoma na temelju koje će se izolirati genomske sekvence i dalje

koristiti u filogenetskim analizama (Young and Gillung, 2020). Stoga je izrazito bitno prethodno odrediti koja će se kombinacija setova podataka i metoda koristiti. Svaki tip podataka ima svoje prednosti i mane te Chen i sur. (2017), kao jedno od potencijalnih rješenja, preporučuju provedbu analiza na nekoliko različitih tipova podataka i njihovu usporedbu na temelju koje bi se trebao donijeti zaključak.

Moderne filogenetske (ili filogenomske) analize se u pravilu baziraju na informaciji dobivenoj iz većih genomskih fragmenata što umanjuje stupanj pogreške ali i limitirajuće informacije koje se mogu dobiti korištenjem kraćih sekvenci, npr. pojedinačnih gena. Međutim, korištenje genomskih podataka ima i svoje nedostatke. Prvi je cijena sekvenciranja. Sangerovo sekvenciranje manje genske regije neusporedivo je jeftinije i može se provesti na puno većem broju uzoraka što može biti ključno u ovim analizama (Rosenberg i Kumar, 2003; Nabhan i Sarkar, 2012). S druge strane, sekvenciranje cijelog genoma skuplji je i dugotrajniji proces koji se unaprijed treba detaljno isplanirati. U takvim se analizama, zbog visoke cijene, koristi manji broj uzoraka nego kod klasičnih filogenetskih analiza baziranih na manjim dijelovima. Osim toga, genomski podaci zahtjevaju puno veći prostor za pohranu te su računalno i vremenski puno zahtjevniji (npr. genomski podaci sekvencirani s pokrivenošću od 100x mogu zauzimati prostor od nekoliko stotina gigabajta) (Ekblom i Wolf, 2014). Uzimajući u obzir ovu problematiku, Heath i sur. (2008) su zaključili da povećanje ovakvih uzoraka u većoj mjeri može poboljšati točnost filogenetskih analiza. Osim toga, Young i Gillung (2020) kao jedan od potencijalnih rješenja predlažu da se prije početka filogenetske analize pretraže sve dostupne baze koje mogu sadržavati genomske podatke vrste koja se želi proučiti i na taj način proširi broj uzoraka.

2.6 Divokoza (*Rupicapra* spp.)

Divokoza (*Rupicapra* spp.) (Slika 4.) je planinski papkar iz porodice šupljorožaca (*Bovidae, Caprinae*) i nastanjuju planinske predjele Europe i Bliskog Istoka na visinama između 500 i 3.000 m nadmorske visine. Životni prostor, nepovoljni klimatski uvjeti te čovjekove aktivnosti najviše su utjecali na geografsku rasprostranjenost divokoze. Divokoze nastanjuju planinske masive Pirinejskog poluotoka, Apeninskog poluotoka, Balkanskog poluotoka, Alpa, Karpata, Anatolije i Kavkaza (Tosi i Pedrotti, 2003; Aulagnier i sur., 2008). Uz navedeno, divokoza nastanjuju masive Novog Zelanda gdje je introducirana početkom 20. stoljeća (Forsyth, 2005). Prema modernoj taksonomiji te na temelju morfoloških svojstava i geografske distribucije divokoza se dijeli u dvije vrste: sjevernu (*Rupicapra rupicapra*) i južnu (*Rupicapra*

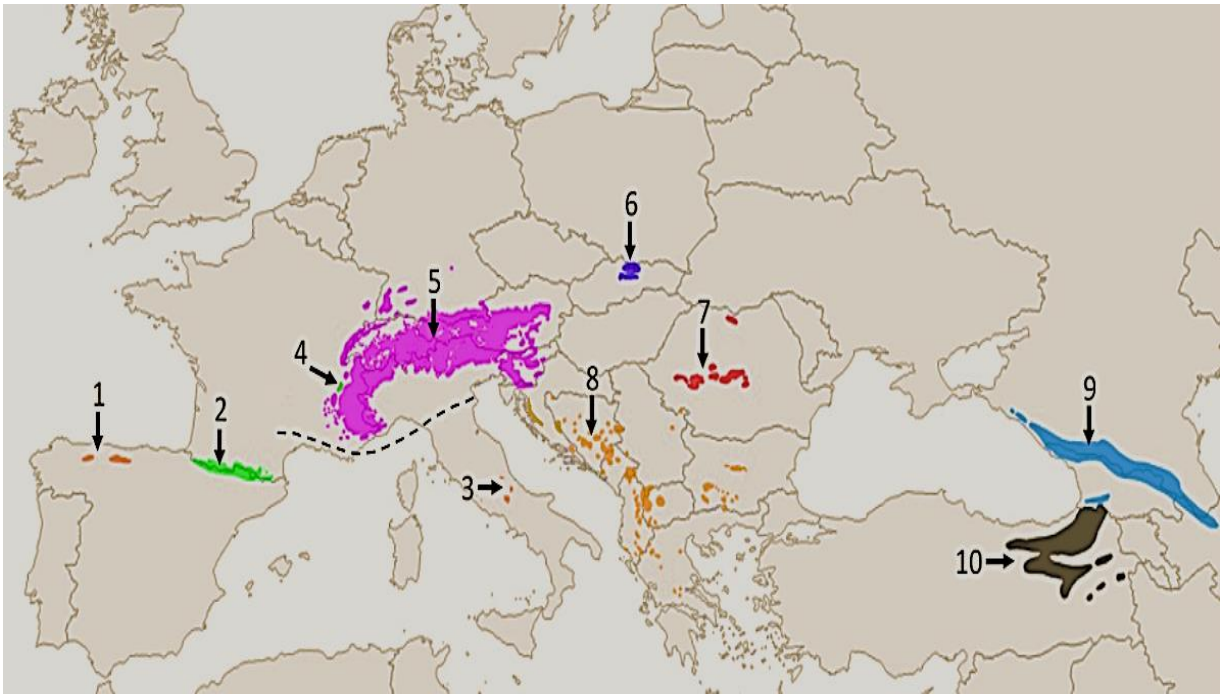
pyrenaica) (Grubb, 2004). Prema Corlatti i sur. (2011) brojno stanje sjeverne divokoze je između 250.000 i 350.000 jedinki, a južne oko 36.000 jedinki, dok je prema IUCN Crvenoj listi (Anderwald i sur., 2020. https://iucnredlist.org, pristupljeno 4. prosinca, 2021.) brojno stanje sjeverne divokoze 300.000, a južne 50.000. Prema podacima o brojnom stanju divokoza, ni jedna vrsta nije ugrožena, međutim, na razini podvrsta postoje razlozi za zabrinutost oko očuvanja pojedinih podvrsta. Ukoliko se podvrste uzmu u obzir, divokozu možemo smatrati jednom od najugroženijih papkara na području Europe (Corlatti i sur., 2022a).

Slika 4. Balkanska divokoza (*R. r. balcanica*) na Biokovu. (Foto: Krešimir Kavčić).

2.6.1 Pregled filogenetskih istraživanja divokoze (*Rupicapra* spp.)

Predak današnjih divokoza, *Procamptoceras*, pojavio se tijekom Miocena u području Azije, a krajem Miocena i početkom Pleistocena se počeo širiti prema Europi (divokoza) i Sjevernoj Americi (američka planinska koza) (Corlatti i sur., 2011). Današnji oblici divokoza pojavili su se tijekom srednjeg i kasnog Pleistocena u području zapadne Euroazije (Kurtén, 1968). Prema arheološkim ostacima iz posljednjeg ledenog doba, utvrđena je prisutnost dviju vrsta divokoze *R. pyrenaica* i *R. rupicapra*. *R. pyrenaica* se još u to vrijeme razdvojila u dvije geografske skupine: na Pirinejskom poluotoku i u centralnom dijelu Apeninskog područja, dok se *R. rupicapra* zadržala u hladnijim područjima centrale i istočne Europe. Na temelju ovih podataka postavljen je prva hipoteza koja objašnjava razlike u morfološkoj građi i bihevioralnim karakteristikama u modernim populacijama divokoza (Masini i Lovari, 1988).

Iako je divokoza prilagođena na stjenoviti i strmoviti teren, udaljenost i fragmentiranost planinskih masiva Euroazije utjecale su na distribuciju, rascjepkanost te izoliranost populacija. Osim toga, s obzirom da je divokoza atraktivna lovna vrsta, pretjerani lov i krivolov utjecali su na nestanak divokoze s pojedinih područja zbog čega su vršene introdukcije, reintrodukcije i translokacije ove vrste s ciljem obnavljanja populacija (Zemanová i sur., 2011; Apollonio i sur., 2014; Zemanová i sur., 2014). Prema geografskim obilježjima (Slika 5.) ali i vanjskim karakteristikama, obje su vrste podijeljene na podvrste: *Rupicapra rupicapra* sa sedam podvrsta (*asiatica*, *balcanica*, *carpatica*, *cartusiana*, *caucasica*, *rupicapra*, *tatrica*) i *Rupicapra pyrenaica* s tri podvrste (*parva*, *pyrenaica*, *ornata*) (Corlatti i sur., 2011). Prema IUCN-ovoj Crvenoj listi (Anderwald i sur., 2020; pristupljeno u svibnju, 2021) podvrste sjeverne divokoze nastanjuju sljedeća područja: *R. r. asiatica* naseljava sjeveroistočna planinska područja Crnog mora, istočne dijelove Anatolije (Turska) te sjeverozapadnu regiju u Gruziji s veličinom populacije između 500 i 750 jedinki. *R. r. balcanica* naseljava planinska područja Hrvatske (Dinara, Biokovo), Bosne i Hercegovine, Srbije, Kosova, Crne Gore, Sjeverne Makedonije, Albanije, Bugarske, Grčke s veličinom populacije oko 10.000 jedinki. *R. r. carpatica* s nekoliko manjih populacija nastanjuje planinske predjele Transilvanije i Karpata u Rumunjskoj s ukupnom veličinom populacije oko 8.000 jedinki. *R. r. cartusiana* nastanjuje usko područje planinskog masiva Chartreuse u Francuskoj s veličinom populacije oko 1.500 jedinki. *R. r. caucasica* nastanjuje područje masiva Kavkaz na području Rusije, Gruzije i Azerbajdžana s veličinom populacije oko 9.000 jedinki. *R. r. rupicapra* najbrojnija je podvrsta divokoze a naseljava šire područe Alpi te područje Slovenije i Hrvatske (Velebit) s grupo procijenjenom veličinom populacije na nešto manje od 300.000 jedinki. *R. r. tatrica* nastanjuje usko područje planinskog masiva Tatre između Poljske i Slovačke s veličinom populacije oko 1.350 jedinki. Prema IUCN-ovoj Crvenoj listi (Herrero i sur., 2020) podvrste južne divokoze nastanjuju sljedeća područja: *R. p. pyrenaica* nastanjuje Pirinejski masiv na području Andore, Španjolske Francuske s veličinom populacije oko 30.000 jedinki. *R. p. parva* nastanjuje područje Kantabrijskog gorja u Španjolskoj s veličinom populacije oko 16.000 jedinki. *R. p. ornata* nastanjuje nekoliko manjih područja Apenina u centralnoj Italiji s veličinom populacije oko 2.500 jedinki.



Slika 5. Geografska rasprostranjenost divokoze: Južna divokoza (*R. pyrenaica*): (1) *parva*, (2) *pyrenaica*, (3) *ornata*. Sjeverna divokoza (*R. rupicapra*): (4) *cartusiana*, (5) *rupicapra*, (6) *tatrica*, (7) *carpatica*, (8) *balcanica*, (9) *caucasica*, (10) *asiatica*. Isprekidana linija označava granicu između dvije vrste (Corlatti i sur., 2022a, prilagođena slika).

Postavljanje hipoteza o podrijetlu i povezanosti vrsta divokoze česte su teme istraživanja prošlog ali i ovog stoljeća. Međutim, još uvijek nije postignut dogovor između znanstvenika o broju podvrsta divokoze. U proučavanju odnosa među vrstama i podvrstama divokoze korišten je veliki broj molekularnih markera: alozimi (Nascetti i sur., 1985), minisateliti (Pérez i sur., 1996), geni glavnog sustava tkivne podudarnosti (Schaschl i sur., 2012; Alvarez-Busto i sur., 2007), mtDNA geni (Hammer i sur., 1995; Crestanello i sur., 2009; Pérez i sur., 2014; Šprem i Bužan, 2016), mtDNA kontrolna regija (Šprem i Bužan, 2016; Rezić i sur., 2022), mikrosateliti (Soglia i sur., 2010; Rodríguez i sur., 2009; 2010; Papaioannou i sur., 2019; Rezić i sur., 2022), Y-kromosom (Pérez i sur., 2011), introni (Pérez i sur., 2017) , SNP-ovi (Leugger i sur., 2022). Na temelju informacija dobivenih analizom svih navedenih markera divokoze se podijeljene u dvije vrste, tri mitohondrijska klastera (istočna, centralna i zapadna) te tri klastera dobivena iz mikrosatelita koji se ne poklapaju u potpunosti s mitohondrijskim klasterima. Osim toga, ova taksnomija ne podudara se s taksonomijom dobivenoj na temelju

morfoloških, bihevioralnih i geografskih podataka prema kojima sjeverna divokoza ima sedam podvrsta, a južna tri.

3 MATERIJALI I METODE

3.1 Uzorkovanje i sekvenciranje

U sklopu ove disertacije sakupljeno je dvanaest uzoraka od dvije vrste divokoza. Osam uzoraka pripadalo je sjevernoj divokozi, od kojih su dva bila podvrste tatranske (*Rupicapra rupicapra tatrica*), dva balkanske *(Rupicapra rupicapra balcanica*), četiri alpske (*Rupicapra rupicapra rupicapra*) i dva uzorka potencijalnih hibrida (*R. r. balcanica* x *R. r. rupicapra*). Preostala dva uzorka pripadala su južnoj divokozi, podvrsti *Rupicapra pyrenaica pyrenaica*. Svi su uzorci prikupljeni tijekom redovnih službenih sezona lova u skladu sa zakonskom regulativom pojedine države i prirodnim uginućem (npr. lavine). Genomska DNA ekstrahirana je iz uzoraka tkiva standardnom metodom koja koristi fenol i kloroform. Fragmenti DNA su pripremljeni i obrađeni prema Illumina protokolu za pripremu uzorka DNA. Sekvenciranje genoma provedeno je na platformi Illumina HiSeq 2500 s obostranim sekvenciranjem fragmenata (engl. *paired-end*) i duljinom čitanja od 100 pb za knjižnice od 350 pb. Prosječna veličina svakog fragmenta bila je 150 pb. Za svaki uzorak dobivene su dvije FASTQ datoteke koje u sebi sadrže informacije o svim sekvenciranim fragmentima u oba smjera (R1 *forward* i R2 *reverse*). Informacije o genomskim uzorcima prikazane su u tablici 1.

Tablica 1. Informacije o sirovim genomskim podacima.

Uzorak	ID uzorka	Podvrsta	Lokacija	Broj fragmenata
1.	Gams7	*R. r. balcanica*	Hrvatska	216.319.578
2.	Gams65	*R. r. balcanica*	Hrvatska	6.836.052
3.	Gams21	*R. r. rupicapra*	Hrvatska	222.981.802
4.	Gams85	*R. r. rupicapra*	Hrvatska	244.129.826
5.	Osil-06	*R. r. rupicapra*	Slovenija	196.999.200
6.	LM 2/07	*R. r. rupicapra*	Slovenija	193.843.460
7.	Gams108	*R. p. pyrenaica*	Španjolska	225.995.072
8.	Gams109	*R. p. pyrenaica*	Španjolska	278.609.626
9.	Gams53	*R. r.* hibrid	Hrvatska	235.807.388
10.	Gams57	*R. r.* hibrid	Hrvatska	261.340.198
11.	B532	*R. r. tatrica*	Slovačka	241.085.630
12.	B539	*R. r. tatrica*	Slovačka	189.729.312

3.2 Kontrola kvalitete genomskih podataka

Kontrola kvalitete svih uzoraka provodila se u programima FastQC (Andrews, 2010) i Trimmomatic (Bolger i sur., 2014). Ulazni podaci za FastQC su dvije FASTQ datoteke (R1 i R2) za svaki uzorak. FastQC je na temelju informacija iz R1 i R2 datoteka izračunao broj fragmenata, sastav i odnos baza, zastupljenost ponavljajućih regija te zastupljenost adapter sekvenci. FastQC je za većinu uzoraka dao izvještaj da se radi o podacima dobre kvalitete uz dodatnu potrebu rezanja (engl. *trimming*) adapter sekvenci i uklanjanja dupliciranih fragmenata. Za 4 od 12 uzoraka FastQC je pronašao kontaminacije i neispravnosti na temelju broja i zastupljenosti fragmenata (kontaminacije su pronađene u R2 FASTQ datotekama). Zbog nejednakog broja fragmenata u R1 i R2 datotekama, 4 uzorka su uklonjena iz budućih analiza, prvenstveno jer ih niti jedan programski alat neće moći učitati. Uz navedeno, jedan uzorak (Gams 65) uklonjen je iz analize zbog iznimno male pokrivenosti (manje od 1x). Međutim, u procesu sastavljanja mtDNA *de novo* metodom, tri od navednih pet uzoraka koji nisu prošli kontrolu mogla su se koristiti. Za dodatnu provjeru ovih 5 uzoraka korišten je fastp program (Chen i sur., 2018) koji je dao iste rezultate kao i FastQC. Trimmomatic alat je potom korišten za rezanje sekvenci kraćih od 150 pb te za uklanjanje adapter sekvenci i dupliciranih sekvenci iz uzoraka koji su prošli kontrolu.

3.3 Sastavljanje mtDNA metodom mapiranja

Za svaki uzorak koji je prošao kontrolu kvalitete, R1 i R2 FASTQ datoteke korištene su kao ulazni podaci za maper BWA (verzija 0.7.17). Reprezentativne sekvence mtDNA sjeverne (NCBI pristupni broj: FY207539) i južne divokoze (NCBI pristupni broj: FJ207538.1) korištene su kao referentne sekvence na koje će BWA mapirati genomske fragmente iz uzoraka. Referentne mtDNA divokoze su u FASTA formatu i sadrže svoj nukleotidni zapis u jednoj liniji. Cijeli protokol sastoji se od devet glavnih koraka koji su provedeni zasebno za svaku kombinaciju uzorka i referentne sekvence:

1. BWA indeksiranje referentnog genoma (fukcija *index*): BWA je datoteku referentnog genoma u FASTA formatu prvo skenirao te potom indeksirao. Ovim su se korakom očitala sva mjesta u referentnoj datoteci na koja će se poravnavati genomski fragmenti
2. BWA mapiranje genomskih fragmenata na referentni genom (funkcija *bwa-mem*): Ulazni podaci su R1 i R2 FASTQ datoteke i indeksirani referentni genom u FASTA formatu. Izlazna datoteka u SAM formatu

3. SAMTools konvertiranje i sortiranje (funkcija *sort* i *view*): SAM datoteka se konvertirala u BAM format i potom se provelo sortiranje BAM datoteke
4. SAMTools manipulacija BAM datoteka (funkcija *fixmate* i *markdup*): U sortiranoj BAM datoteci uklonili su se svi duplicirani i PCR fragmenti nakon čega se BAM datoteka ponovno sortirala
5. SAMTools kontrola kvalitete QC (funkcija *view -q 20*): Uklonili su se svi fragmenti koji imaju kvalitetu mapiranja ispod 20 (engl. *mapping quality*)
6. SAMTools indeksiranje (funkcija *index* i *faidx*): sa SAMtools alatom, sortirana i filtrirana BAM datoteka te datoteka referentnog genoma su se indeksirale.
7. BCFtools pozivanje strukturnih varijanti (funkcija *mpileup* i *call*): na temelju poravnanja provelo se pozivanje svih strukturnih varijanti pronađenih između genomskih fragmenata i reference. Indeksirane BAM i FASTA datoteke korištene su kao ulazne, dok je izlazna datoteka bila u binarnom BCF formatu
8. BCFtools konvertiranje BCF u VCF datoteku i provjera osnovne statistike sirovih podataka (funkcija *convert* i *stats*): ovim su se korakom dobile informacije o vrsti i broju pronađenih varijanti
9. BCFtools indeksiranje VCF datoteke i pozivanje konsenzusne sekvence (funkcije *tabix* i *consensus*): Filtrirana datoteka VCF prvo se kompresirala i indeksirala. Potom se pozvala nova, konsenzusna sekvenca koja je predstavlja kombinaciju referentne sekvence i svih prisutnih strukturnih varijanti iz VCF datoteke. Izlazna datoteka novog genoma koja se koristila u daljnjim analizama bila je u FASTA formatu

Rezultirajuće sekvence zapisane u FASTA datotekama korištene su u daljnim koracima za provjeru kvalitete i validaciju.

3.4 Sastavljanje mtDNA metodom *de novo*

Za provedbu sastavljanja mtDNA metodom *de novo*, R1 i R2 FASTQ datoteke svakog uzorka korištene su kao ulazni podaci za asembler NOVOPlasty (verzija 4.3). CYTB geni iz reprezentativne sekvence mtDNA sjeverne i južne divokoze korišteni su kao seed sekvence, tj. kao startna pozicija na koju se slažu genomski fragmenti. Prije procesa sastavljanja bilo je potrebno podesiti sve parametre u config.txt datoteci koja se nalazi u instalacijskom direktoriju NOVOPlasty alata. Podesili su se slijedeći parametri:

1. Parametrom „Type“=mito definiralo se da će se u ovom koraku sastaviti mtDNA budući da se ovaj program može koristiti i za sastavljanje kloroplastne DNA.

2. Parametrom „Genome Range“=14000-15000 definirala se očekivana duljina nove sekvence.
3. Parametrom „Seed input“=./cytb.fasta definirala se putanja do CYTB FASTA datoteke koja će se koristit kao seed sekvenca.
4. Pod kategorijom „Dataset 1“ gdje se alatu daju informacije o genomskim podacima definirani su sljedeći parametri: „Read Lenght“=150; „Insert size“=300; „Platform“=Illumina; „Single/Paired“=PE;
5. Parametrom „Forward reads“ i „Reverse reads“ definirale su se putanje do R1 i R2 FASTQ datoteka

Preostali parametri nisu se mijenjali. NovoPlasty je također korišten za sastavljanje mtDNA iz tri neispravna uzorka koja nisu prošla kontrolu kvalitete. Sve novosastavljenje sekvence zapisane u FASTA datotekama korištene su u daljnim koracima za provjeru kvalitete i validaciju.

3.5 Validacija i usporedba sastavljenih sekvenci mtDNA

Provjera kvalitete i validacija kompletnosti novosastavljenih sekvenci provjerila se u tri osnovna koraka. Za svih sedam uzoraka, koji su bili ispravni, dobiveno je 14 novosastavljenih mitohondrijskih sekvenci pomoću dvije metode sastavljanja. Metodom *de novo* dobivene su tri sekvence mtDNA iz neispravnih uzoraka. Za svaku novu sekvencu zasebno je provedena homologna pretraga prema sličnosti u BLAST web aplikaciji. Na ovaj se način svaka nova mtDNA sekvenca usporedila s dostupnim mtDNA sekvencama u Banci gena. Ovim se korakom provjerilo sadrže li nove sekvence sve regije tipične za mtDNA sisavaca. Nakon toga se par sekvenci dobivenih za isti uzorak korištenjem dvije različite metode zasebno poravnalo s referentnom sekvencom divokoze s ciljem pronalaženja strukturnih razlika, osobito u CR regiji. Sva poravnanja provela su se u programu MEGA X (Kumar i sur., 2018). Na temelju svih navedenih usporedbi, za svaki uzorak je odabrana po jedna mtDNA sekvenca koja je bila korištena u procesu anotacije.

3.6 Anotacija mtDNA

Proces anotacije 10 mtDNA sekvenci proveden je u GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html) i MITOS (http://mitos.bioinf.uni-leipzig.de/) web aplikacijama. Za anotaciju 22 tRNA gena korišten je tRNA-SE (Lowe i Eddy, 1997; Lowe i Chan, 2016) program koji je implementiran u svim navedenim anotatorima. Rezultat anotacije za svaku sekvencu zapisan je u GenBank formatu u kojem su definirana sva START i STOP mjesta

svih PCG gena, tRNA i rRNA. Rezultati dvaju anotatora su se usporedili. Budući da je proces anotacije mtDNA automatiziran, uvijek postoji mogućnost da se START i STOP pozicije pojedinih gena identificiraju na različitim pozicijama (često budu pomaknuti za par baznih mjesta). Zbog toga se za validaciju rezultata anotacije napravilo zajedničko poravnjanje svih mtDNA s dvije reference. Iz zajedničkog poravnanja su se izdvojili PCG geni te su se njihove anotirane START i STOP pozicije usporedile s rezultatima anotacije referentne mtDNA sjeverne divokoze dostupne u Banci gena. Navedenim koracima provjerila se struktura svih novih sekvenci prije filogenetskih analiza.

3.7 Filogenetske analize mtDNA

Nakon validacije svih rezultata anotacije, izrađena su dva seta podataka koja su bila korištena za filogenetske analize. Prvi set prikazan je u Tablici 2. a sastojao se od deset novosastavljenih mtDNA divokoze, dvije referentne sekvence, četiri sekvence mtDNA divokoze (*R. p. pyrenaica, R. r. cartusiana, R. p. ornata*, *R. r. pyrenaica*) te dvije sekvence koje su bile korištene kao uljezi (engl. *outgroup*) u filogenetskim analizama (*Ammotragus lervia*, *Arabitragus jayakari*).

Tablica 2. Popis korištenih mtDNA sekvenci u rekonstrukciji filogenije roda *Rupicapra.* Pet mtDNA sekvenci preuzeto je iz Banke gena. Ostale su sastavljene u sklopu ove disertacije.

Uzorak	Vrsta (podvrsta)	Duljina (pb)	Referenca
FJ207539	*R. rupicapra (reference)*	16.434	Hassanin i sur., (2009)
FJ207538	*R. pyrenaica (reference)*	16.438	Hassanin i sur., (2009)
B532	*R. r. tatrica*	16.434	disertacija
B539	*R. r. tatrica*	16.434	disertacija
Gams7	*R.r. balcanica*	16.432	disertacija
Gams21	*R. r. rupicapra*	16.432	disertacija
Gams53	*R. r. rupicapra x R. r. balcanica*	16.432	disertacija
Gams57	*R. r. rupicapra x R. r. balcanica*	16.433	disertacija
Gams85	*R. r. rupicapra*	16.433	disertacija
Gams108	*R. p. pyrenaica*	16.438	disertacija
Gams109	*R. p. pyrenaica*	16.438	disertacija
OSIL	*R. r. rupicapra*	16.433	disertacija
KJ184175	*R. r. cartusiana*	16.398	disertacija
KJ184173	*R. p. ornata*	16.399	Pérez i sur., (2014)
KJ184174	*R. p. pyrenaica*	16.398	Pérez i sur., (2014)
NC_009510	*Ammotragus lervia*	16.530	Mereu i sur., (2008)
NC_020621	*Arabitragus jakari*	16.457	Hassanin i sur., (2009)

Prije provedbe filogenetskih analiza, za analizu sekvenci divokoza korišten je DNAsp program (verzija 6.12.03) (Rozas i sur., 2017) za definiranje broja haplotipova, raznolikosti haplotipova (h), raznolikosti nukleotida (π), broj polimorfnih mjesta (S) te za Fu-ovu statistiku (Fs).

Drugi set podataka prikazan je u Tablici 3., a sastojao se od 40 sekvenci mtDNA roda *Caprini* te od 5 sekvenci mtDNA roda *Bovidae* koje su korištene kao uljezi. Šest sekvenci divokoza (jedna za svaku podvrstu) korišteno je u ovoj analizi.

Tablica 3. Popis korištenih mtDNA sekvenci u rekonstrukciji filogenije roda Caprinae. Četiri mtDNA sekvence dobivene su u sklopu ove disertacije. Ostale su preuzete iz Banke gena.

NCBI pristupni broj	Vrsta/podvrsta	duljina (pb)	Referenca
NC_004069	*Muntiacus reevesi*	16.354	Zhang i sur., 2002
NC_006853	*Bos taurus*	16.338	Chung i Ha., 2005
NC_049568	*Bubalus bubalis*	16.358	Verma i sur., 2020
NC_023543	*Damaliscus lunatus*	16.419	Steiner i sur., 2013
NC_020627	*Damaliscus pygargus*	16.385	Hassanin i sur., 2009
NC_007441	*Pantholopos hodgsonii*	16.498	Xu i sur., 2005
NC_009510	*Ammotragus lervia*	16.530	Mereu i sur., 2008
NC_020621	*Arabitragus jayakari*	16.457	Hassanin i sur., 2009
Gams109	*Rupicapra pyrenaica pyrenaica*	16.438	disertacija
KJ184175	*Rupicapra rupicapra cartusiana*	16.398	Pérez i sur., 2014
KJ184173	*Rupicapra pyrenaica ornata*	16.399	Pérez i sur., 2014
Gams7	*Rupicapra rupicapra balcanica*	16.432	disertacija
Gams21	*Rupicapra rupicapra rupicapra*	16.432	disertacija
B532	*Rupicapra rupicapra tatrica*	16.434	disertacija
NC_020630	*Oreamnos americanus*	16.604	Hassanin i sur., 2009
NC_039431	*Ovis nivicola lydekkeri*	16.471	Dotsev i sur., 2018
MH094035	*Ovis canadensis*	16.466	Davenport i sur., 2018
NC_039432	*Ovis dalli*	16.464	Dotsev i sur., 2018
NC_001941	*Ovis aries*	16.616	Hiendleder i sur., 1998
KX609626	*Ovis ammon darwini*	16.618	Mao i sur., 2016
MN564883	*Ovis ammon ammon*	16.612	Wand i sur., 2020
JX101654	*Ovis ammon hodgsoni*	16.688	Jiang i sur., 2013
NC_043930	*Budorcas taxicolor taxicolor*	16.584	Kumar i sur., 2019
NC_013069	*Budorcas taxicolor*	16.667	Wu i sur., 2009
NC_039686	*Budorcas taxicolor tibetana*	16.665	Zhou i sur., 2018
FJ207537	*Pseudois nayaur*	16.737	Hassanin i sur., 2009
KP998469	*Pseudois nayaur szechuanensis*	16.741	Liu i sur., 2015
KR059226	*Capra aegagrus*	16.639	Colli i sur.,2015
NC_005044	*Capra hircus*	16.643	Hassanin i sur., 2010
NC_020622	*Capra falconieri*	16.640	Hassanin i sur., 2009
NC_020683	*Capra caucasica*	16.624	Hassanin i sur., 2012
NC_020625	*Capra pyrenaica*	16.561	Hassanin i sur., 2009
NC_020623	*Capra ibex*	16.716	Hassanin i sur., 2009
NC_020624	*Capra nubiana*	16.705	Hassanin i sur., 2009
NC_020626	*Capra sibirica*	16.583	Hassanin i sur., 2009
NC_020628	*Hemitragus jemlahicus*	16.712	Hassanin i sur., 2009

Nastavak Tablice 3.

NCBI pristupni broj	Vrsta/podvrsta	duljina (pb)	Referenca
NC_020631	*Ovibos moschatus*	16.431	Hassanin i sur., 2009
NC_021381	*Naemorhedus goral*	16.555	Yang i sur., 2013
NC_013751	*Naemorhedus caudatus*	16.519	Jang i Hwang, 2010
NC_020723	*Naemorhedus griseus*	16.448	Hassanin i sur., 2012
NC_020722	*Naemorhedus baileyi*	16.448	Hassanin i sur., 2012
KT345703	*Capricornis thar jamrachi*	16.444	Zhang i sur., 2015
NC_020629	*Capricornis sumatraensis*	16.441	Hassanin i sur., 2009
NC_010640	*Capricornis swinhoei*	16.524	Lee i sur., 2008
NC_012096	*Capricornis crispus*	16.453	Yasue i su., 2009

Za oba seta podataka napravljeno je poravnanje iz kojeg su se uklonila sva mjesta koja su nastala greškom u procesu sastavljanja ili mjesta čiji dijelovi predstavljaju nDNA pseudogene. Nadalje, kako bi se izbjegli pogrešni zaključci u filogenetskim analizama, isključena su i sva mjesta u zajedničkom poravnanju koja su predstavljala insercije, delecije, nedefinirane praznine te tandem ponavljanja prisutna u kontrolnoj regiji porodice *Caprinae*.

Dvije metode su korištene za rekonstrukciju filogenetskih odnosa: maksimalna vjerodostojnost (engl. *maximum likelihood*) u programu IQ-TREE (Kalyaanamoorthy i sur., 2017) te Bayesovska metoda u programu MrBayes 3.2.7 (Ronquist i sur., 2012). IQ-TREE program prvo je korišten za odabir modela za oba seta podataka te je provedena rekonstrukcija sa zadanim parametrima uz par izmjena što uključuje parametre genetski kod - vertebrate mitochondrial; substitucijski model - TIM2+I+G4; bootstrap analiza - ultrafast bootstrap analysis UFBoot2 (Hoang i sur., 2018). Bayesovska metoda za set podataka *Caprinae* provedena je korištenjem GTR+I+G4 modela i simulacije Markovljevih lanaca Monte Carlo (MCMC, engl. *Markov Chain Monte Carlo*) za 1.2 milijuna generacija s tim da su stabla uzorkovana svakih 500 generacija dok se nije postigla standardna devijacija odvojenih frekvencija od 0.011. Za *Rupicapra* set podataka korišten je model GTR+G4 i MCMC simulacija za 1.7 milijuna generacija (uzorkovanje svakih 200 generacija dok se nije postigla standardna devijacija odvojenih frekvencija 0.006). Za obje analize rezultati su se provjeravali prema ispisu probabilističkih log grafova (Huelsenbeck i sur., 2001). Za svaku iteraciju (engl. *run*), 25% početnih stabala bilo je odbačeno (engl. *burn-in*). Vizualizacija stabala provedena je pomoću R paketa ggtree (Yu i sur., 2017).

3.8 Sastavljanje nDNA metodom mapiranja

Za svaki uzorak koji je prošao kontrolu kvalitete, R1 i R2 FASTQ datoteke korištene su kao ulazni podaci za maper BWA (verzija 0.7.17). Proces mapiranja proveden je u dvije faze. U prvoj se fazi provodilo mapiranje genomskih uzoraka divokoze na osam odabranih genoma divokozi srodnih vrsta s ciljem usporedbi razlika genomskih sekvenci dobivenih mapiranjem istih uzoraka na različite referentne genome. Osim toga, ovim su se korakom testirale kvalitete dostupnih referentnih genoma budući da svi pripadaju nemodelnim vrstama. U drugoj se fazi kao referenca koristio referentni genom domaće koze kako bi se provjerila struktura novosastavljenih genoma usporedbom s genomskih sekvencama divokoze dostupnim u Banci gena. Genom domaće koze je najveće kvalitete među domaćim životinjama te sadrži isti broj kromosoma kao i divokoza (29 autosomalnih i jedan spolni). Odabrani genomi bliskih srodnika divokoze koji su se koristili kao reference za mapiranje bili su: domaća koza (*Capra hircus*), divlja koza (*Capra aegagrus*), grivasti skakač (*Ammotragus lervia*), američka planinska koza (*Oreamnos americanus*), baral (*Pseudois nayaur*), alpski kozorog (*Capra ibex*), sibirski kozorog (*Capra sibirica*) te domaća ovca (*Ovis aries*). Referentni genomi su zapisani u FASTA formatu u nekoliko stotina ili tisuća skafolda. Kod referentnih genoma koji su sastavljeni na razini kromosoma prvih 26 - 30 skafolda (ovisno o broju kromosoma) predstavljaju krosmosome. Za svaki skafold u FASTA datoteci zapisano je ime sekvence te pripadajući nukleotidni zapis. Informacije o referentnim genomima prikazane su u Tablici 4.

Tablica 4. Osnovne informacije o preuzetim referentnim genomima iz Banke gena.

Vrsta	**Reprezentativni genom i NCBI pristupni broj**	**Razina sastavljenog genoma**	**Godina objave**	**Broj skafolda**
Capra hircus	ARS1; GCA_001704415.1	Kromosom (30)	Bickhart i sur., 2016.	29.907
Capra ibex	IBX; GCA_006410555.1	Skafold	Chen i sur., 2019.	55.914
Capra sibirica	ASM318261v2; GCA_003182615.2	Skafold	Northwest A&F University, 2018.	85.609
Capra aegagrus	CapAeg_1.0; GCA_000978405.1	Kromosom (30)	Dong i sur., 2015.	89.498
Ammotragus lervia	ALER1.0; GCA_002201775.1	Skafold	Chen i sur., 2019.	77.320
Oreamnos americanus	ASM975805v1; GCA_009758055.1	Skafold	Martchenko i sur., 2020.	3.212
Ovis aries	Oar_rambouillet_v1.0; GCA_002742125.1	Kromosom (26)	Baylor College, 2017.	2.641
Pseudois nayaur	ASM318257v1; GCA_003182575.1	Skafold	Chen i sur., 2019.	214.772

Protokol mapiranja sličan je opisanom protokolu za sastavljanje mtDNA uz nekoliko dodatnih koraka koji se odnose na filtriranje varijanti. Protokol se sastoji od 12 glavnih koraka koji su provedeni zasebno za svaku kombinaciju uzorka i referentne sekvence:

1. BWA indeksiranje referentnog genoma (fukcija *index*): Ulazna datoteka u FASTA formatu prvo se skenirala te potom indeksirala. Ovim su se korakom očitala sva mjesta u referentnoj datoteci na koja će se poravnati genomski fragmenti
2. BWA mapiranje genomskih fragmenata na referentni genom (funkcija *bwa-mem*): Ulazni podaci su R1 i R2 FASTQ datoteke i indeksirani referentni genom u FASTA formatu. Izlazna je datoteka u SAM formatu
3. SAMTools konvertiranje i sortiranje (funkcija *sort* i *view*): SAM datoteka se konvertirala u BAM format i potom se provelo sortiranje BAM datoteke
4. SAMTools manipulacija BAM datoteka (funkcija *fixmate* i *markdup*): U sortiranoj BAM datoteci uklonili su se svi duplicirani i PCR fragmenti nakon čega se BAM datoteka ponovno sortirala
5. SAMTools kontrola kvalitete QC (funkcija *view -q 30*): Uklonili su se svi fragmenati koji imaju kvalitetu mapiranja ispod 30
6. SAMTools indeksiranje (funkcija *index* i *faidx*): sortirana i filtrirana BAM datoteka te datoteka referenetnog genoma su se indeksirala
7. BCFtools pozivanje strukturnih varijanti (funkcija *mpileup* i *call*): na temelju poravnanja provelo se pozivanje svih strukturnih varijanti pronađenih između genomskih fragmenata i reference. Indeksirane BAM i FASTA datoteke korištene su kao ulazne, dok je izlazna datoteka bila u binarnom BCF formatu
8. BCFtools konvertiranje BCF u VCF datoteku i provjera osnovne statistike sirovih podataka (funkcija *convert* i *stats*): ovim su se korakom dobile informacije o vrsti i broju pronađenih varijanti
9. BCFtools zadržavanje samo SNP-ova (funkcija *view*): Ovim su se korakom uklonile sve varijante (indels, SNV, CSV) koje se neće koristi u analizama. Izlazna VCF datoteka sadržavala je samo SNP mjesta
10. BCFtools zadržavanje samo bi-alelnih SNP-ova (funkcija *view*): Ovim su se korakom uklonili multi-alelni SNP-ovi i zadržali samo bi-alelni SNP-ovi
11. BCFtools filtriranje SNP-ova (funkcija *view QUAL, DP, MQ*): Ovim se korakom provelo filtriranje dobivenih SNP-ova prema tri glavna kriterija. Odbacili su se SNP-ovi čija je vrijednost 'QUAL' (kvaliteta pronađenog SNP; engl. *quality score*) manja od 15,

vrijednost 'DP' (dubina sekvenciranja, engl. *sequencing depth*) manja od 5 i čija vrijednost 'MQ' (kvaliteta mapiranja; engl. *mapping quality*) je manja od 20

12. BCFtools indeksiranje VCF datoteke i pozivanje konsenzusne sekvence (funkcije *tabix* i *consensus*): Filtrirana datoteka VCF prvo se kompresirala i indeksirala. Potom se pozvala nova, konsenzusna sekvenca koja je predstavlja kombinaciju referentne sekvence i svih prisutnih SNP-ova iz VCF datoteke. Izlazna datoteka novog genoma koja se koristila u daljnjim analizama bila je u FASTA formatu

Konsenzusne sekvence zapisane u FASTA datotekama korištene su u daljnjim koracima za validaciju i anotaciju.

3.9 Validacija i anotacija dobivenih nDNA sekvenci

Sedam mapiranih uzoraka na osam različitih referenci rezultiralo je s ukupno 56 kombinacija konsenzusnih sekvenci. Validacija svake kombinacije (BAM format) prvo se provela pomoću alata SAMtools i funkcije stat gdje su se dobile informacije o proporciji mapiranih fragmenata na svaku referentnu sekvencu. Ovom su se funkcijom dobile i informacije o broju poravnatih fragmenata, broju fragmenata koji su se poravnali na više lokacija duž genoma te o fragmentima koji se nisu poravnali. Alat Tablet je korišten za vizualnu provjeru svih kombinacija. Potom je korišten alat BUSCO (verzija 5) koji je u ovom slučaju korišten za validaciju, ali i za automatsku anotaciju gena. Pomoću BUSCO-a su se pretraživale sve prisutne strukture gena u konsenzusnim sekvencama na temelju baze gena iz referentnog *Cetartiodactyla* podatkovog seta ortologa (odb10, https://busco.ezlab.org/). *Cetartiodactyla* podatkovni set gena sadrži ukupno 13.335 konzerviranih ortologa. Prema tome, BUSCO je prvo proveo automatsku anotaciju te je na temelju pronađenih gena napravio statistički izvještaj prema kojem se vršila validacija novosastavljenih sekvenci. Prema provedenim analizama, BUSCO je prvo dao rezultate validacije koja se odnosi na kompletnost testiranog genoma (u postocima), dok su drugi rezultati izlazne datoteke koje sadrže aminokiselinski zapis svakog pronađenog tj. anotiranog gena i to: 1) geni zastupljeni s jednom kopijom (engl. *single-copy*), 2) geni koji nedostaju (engl. *missing*), 3) duplicirani geni (engl. *duplicated*), 4) fragmentirani geni (engl. *fragmented*).

3.10 Analize sličnosti uzoraka i referenci

Glavni cilj ovih analiza bio je provjeriti razlike između sekvenci istih uzoraka mapiranih na različitim referencama različite kvalitete korištenjem manjih frakcija genoma (nasumično odabrani genski fragmenti zapisani u obliku aminokiselina). Na primjer, uzorak1 mapiran je

na svih 8 različitih referenci te se očekuje da će dobivene sekvence kombinacija tog uzorka sa svim referencama biti sličnije jedna drugoj (uzorak1*ref1, uzorak1*ref2, uzorak1*ref3, itd.) nego što će sekvence uzorka1 biti slične sekvencama drugih uzoraka mapiranih na istu referencu (uzorak2*ref1, uzorak3*ref1, uzorak4*ref1, itd). U idealnom slučaju, ako referenca nema utjecaja na rezultat mapiranja, sve sekvence istog uzorka mapiranog na različite reference biti će potpuno jednake.

Nakon BUSCO anotacije i validacije nDNA sekvenci, proveden je protokol za provođenje analiza sličnosti koji se sastojao se od nekoliko koraka:

1. Iz zajedničkog seta BUSCO gena (pronađenih u svim konsenzusnim sekvencama) napravljen je prvi set podataka koji se sastojao od 10 nasumično odabranih gena:
    a. Za svaki gen napravljeno je poravnanje koje se sastojalo od 64 sekvence tog gena (56 dobivenih iz konsenzusnih sekvenci i 8 iz referentnih genoma).
    b. Iz zajedničkog poravnanja svakog gena izračunata je matrica genetskih udaljenosti između aminokiselinskih sekvenci svake kombinacije uzorak*referenca, procijenjenih kao [broj različitih aminokiselina u genu]/[ukupan broj aminokiselina u genu] (izračunato je ukupno 10 matrica udaljenosti – za svaki gen po jedna).
    c. Iz tih 10 matrica udaljenosti se konstruirala zajednička matrica s aritmetičkim sredinama i rasponima udaljenosti iz svih matrica pojedinačnih gena za sve kombinacije uzoraka i referenci.
2. Iz zajedničkog seta BUSCO gena (pronađenih u svim konsenzusnim sekvencama) napravljen je drugi set podataka koji se sastojao od 100 nasumično odabranih gena:
    a. 100 gena nasumično je raspoređeno u 10 setova (10 setova po 10 gena).
    b. Za svaki set napravljeno je poravnanje koje se sastojalo od 64 sekvence te je izračunata matrica genetske udaljenosti, ukupno 10 matrica (na isti način kao i u koraku 1.b.)
    c. Iz tih 10 matrica udaljenosti se konstruirala zajednička matrica s aritmetičkim sredinama i rasponima udaljenosti iz svih matrica za sve kombinacije uzoraka i referenci.
    d. Napravljeno je novo poravnanje svih 100 gena iz kojeg se izračunala nova genetska matrica udaljenosti (isto kao i u koraku 1.b.).

e. Nova matrica udaljenosti koristila se za grafički prikaz odnosa svih kombinacija u dvodimenzionalnom prostoru korištenjem metode multidimenzionalnog skaliranja (MDS).

3. Iz zajedničkog seta pronađenih BUSCO gena (pronađenih u svim konsenzusnim sekvencama) napravljen je treći set podataka koji se sastojao od 500 nasumično odabranih gena:
    a. Napravljeno je poravnanje od 500 gena iz kojeg se izračunala genetska matrica udaljenosti (isto kao i u koraku 1.b.).
    b. Nova matrica udaljenosti koristila se za grafički prikaz odnosa svih kombinacija u dvodimenzionalnom prostoru korištenjem MDS metode.

Računanje svih matrica udaljenosti (parne p udaljenosti – udio aminokiselina koje se razlikuju između dvije sekvence) (Nei i Zhang, 2006) provedeno je u R programu (https://www.R-project.org/) i u paketima vegan (Oksanen i sur., 2020) i bios2mds (Pele i sur., 2020). Sve prisutne praznine u zajedničkim poravnanjima kao i nedostajuće regije u pojedinim sekvencama su bile isključene tijekom procesa računanja matrica udaljenosti. Spomenuti setovi gena u aminokiselinskom zapisu izolirali su se iz svake kombinacije uzorak*referenca dok su se sva poravnanja provela korištenjem MAFFT alata.

3.11 Usporedba nDNA konsenzusnih sekvenci s dostupnim genomskim sekvencama u Banci gena

Nakon provedene analize sličnosti uzoraka i referenci, za daljnje usporedbe konsenzusnih sekvenci s dostupnim genomskim sekvencama divokoze u Banci gena su korištene genomske sekvence uzoraka dobivene mapiranjem na referentni genom domaće koze. Za procjenu kvalitete i kompletnosti i za provjeru strukture konsenzusnih sekvenci, pretražena je Banka gena s ciljem pronalaska svih dostupnih genomskih sekvenci divokoza. Uz mitohondrijske sekvence divokoze, pronađena su 23 seta podataka intronskih sekvenci koje su u ranijim istraživanjima korišteni (i pohranjeni u Banci gena) za rekonstrukciju filogenije divokoza (Pérez i sur., 2017), a isti su korišteni u ovoj disertaciji za usporedbu i procjenu strukture novih genoma. Za usporedbu dostupnih sekvenci bilo je potrebno u novim genomima detektirati pozicije introna pronađenih u Banci gena. Svaki intronski set sastojao se od 14 sekvenci divokoza i obuhvaćao je svih 10 podvrsta: *R. p. pyrenaica* n = 2, *R. p. parva* n = 2, *R. r. ornata* n = 1, *R. r. cartusiana* n = 2, *R. r. rupicapra* n = 2, *R. r. tatrica* n = 1,

R. r. capratica n = 1, *R. r. balcanica* n = 1, *R. r. asiatica* n = 1, *R. r. caucasica* n = 1). Informacije o intronima nalaze se u Tablici 5.

Tablica 5. Informacije o preuzetim intronskim sekvencama. Stupac INTRON predstavlja rednu poziciju tog introna u genu. U stupcu POZICIJA, slovo „c“ označava da se radi o komplementarnoj sekvenci.

GEN	PUNO IME GENA	INTRON	DULJINA	KROMOSOM	POZICIJA GENA	POZICIJA INTRONA
TRAPPC10	trafficking protein particle complex subunit 10	9.	558	1	143654313:143733893	46755:47312
CLCA1	chloride channel accessory 1	12.	847	3	63650547:63687394	32775:33621
LRGUK	leucine rich repeats and guanylate kinase domain containing	14.	713	4	c:22286874:22389982	66430-67154
SEL1L3	SEL1L family member 3	20.	830	6	c:45659049:45769026	99758:100383
COPE	COPI coat complex subunit epsilon	6.	1024	7	c:104443052:104460725	13627:14699
ABCA1	ATP binding cassette subfamily A member 1	49.	627	8	c:94415525:94547054	128165:128791
HDAC2	histone deacetylase 2	13.	652-658	9	23767599:23801380	28619:29285
PABPN1	poly(A) binding protein nuclear 1	3.	704	10	79846418:79853276	2888:3591
SPTBN1	spectrin beta, non-erythrocytic 1	31.	681	11	36787139:36997846	200214:200894
ATP12A	ATPase H+/K+ transporting non-gastric alpha2 subunit	14.	660	12	50085029:50106629	13563:14222
GAD2	glutamate decarboxylase 2	1.	665	13	26030133:26095874	306:971
AZIN1	antizyme inhibitor 1	8.	568	14	c:20561181:20591914	25413:25998
LYVE1	lymphatic vessel endothelial hyaluronan receptor 1	5.	538	15	c:40252430:40268023	13634:14171
PTGS2	prostaglandin-endoperoxide synthase 2	3.	684-685	16	c:66345155:66353195	2003:2687
FGB	fibrinogen beta chain	8.	580	17	67973450:67981196	6528:7122

Nastavak Tablice 5.

GEN	PUNO IME GENA	INTRON	DULJINA	KROMOSOM	POZICIJA GENA	POZICIJA INTRONA
GGA3	golgi associated, gamma adaptin ear containing, protein 3	4.	611	19	55453234:55467347	7184:7794
PNN	pinin, desmosome	1.	654	21	48567124:48578454	3832:4485
SCN5A	sodium voltage-gated channel alpha subunit 5	26.	647	22	c:11837245:11940676	93372:94019
RIOK3	RIO kinase 3	6.	657	24	c:33468447:33489208	8632:9288
CARHSP1	calcium regulated heat stable protein 1	2.	660	25	c: 7567320:7580497	8172:8842
TUFM	Tu translation elongation factor, mitochondrial	9.	756	25	258755917:25879954	2904:3668
ZFYVE27	zinc finger FYVE-type containing 27	6.	670	26	c: 32707254:32730127	12530:13199
KLC2	kinesin light chain 2	11.	418	29	44716666:44726433	7240:7658

Kako bi se usporedila struktura novosastavljenih genoma s dostupnim sekvencama divokoze, morale su se izdvojiti identične regije introna iz novih genoma. Svaki genom divokoze bio je pohranjen u FASTA formatu i sadržavao je 29.907 sekvenci, isto kao i referenca koze. Potom je u svim FASTA datotekama zadržano samo 29 sekvenci koje predstavljaju 29 kromosoma (bez spolnog kromosoma) koristeći seqtk alat (Shen i sur., 2016b). Nakon toga su se kromosomske sekvence odvojile u pojedinačne FASTA datoteke korištenjem alata bioawk (https://github.com/lh3/bioawk). Budući da je svaki od 23 seta introna sadržavao samo jedan intron koji je dio gena, i budući da jedan gen može sastojati od više introna, bilo je potrebno izolirati cijele sekvence svih 23 gena čiji su introni korišteni u analizi (točne pozicije introna nisu poznate). Prvo su pronađene regije svakog gena u genomu koze te su se, na temelju informacija o položaju (start i stop pozicija, položaj i smjer), izolirale cijele regije gena iz genoma divokoza pomoću alata seqkit. Potom su se provela poravnavanja cijelih regija gena sa svakim intronskim skupom (ukupno 23 poravnjanja) pomoću softvera MEGA X. Iz svakog poravnanja su odrezane regije koje ne pripadaju tom intronu a potom su se sekvence svakog introna (23) pohranile u zasebne FASTA datoteke. Ista procedura dobivanja sekvenci introna provedena je na genomima koze, ovce i goveda (*Bos taurus*; pristupni broj: GCA_002263795.2) koje su korištene kao uljezi u filogenetskim analizama. Sve 23 intronske

sekvence potom spojene u finalno poravnjanje koje se sastojalo od 14 divokoza iz Pérez i sur. (2017), 7 divokoza iz ove disertacije, te od sekvenci koze, ovce i goveda. Poravnanje se sastojalo od 14.980 nukleotidnih baza.

S dobivenim sekvencama introna glavni cilj nije bio proučiti filogeniju divokoza, već usporediti sekvence izvučene iz rekonstruiranih genoma s intronskim sekvencama dostupnim u Banci gena. S pretpostavkom da je moguće izolirati specifične intronske regije iz rekonstruiranih genoma i usporediti ih s dostupnim sekvencama, provedene su filogenetske analize u BEAST2 programu (Bouckaert i sur., 2014). Slijedeći metodologiju Pérez i sur. (2017), rekonstruirano je filogenetsko stablo iz cijelog poravnanja i ispitano je pozicioniranje novih genoma divokoza na stablu u odnosu na ostale uzorke korištene u radu.

3.12 Hibridno sastavljanje genome

Za provođenje prvog koraka hibridne metode, korištena su dva alata za *de novo* sastavljanje: Abyss (verzija 2.2.4) i SPAdess (verzija 3.15.4). Oba alata pokreću se kodom (u kojem se definiraju vrijednosti parametara) te se na taj se način pozivaju skripte integrirane u instalacijskom direktoriju ovih programa.

Pokretanje *de novo* metode u alatu Abyss izveden je s nekoliko različitih parametara:

1. abyss-pe - osnovni parametar kojim se definiralo da su ulazni podaci dvije FASTQ datoteke.
2. np (np=8, np=10, np=20) - parametar kojim se definiralo tzv. MPI način rada (engl. *Message Passing Interface*) i kojim se odredio broj procesa koji se koristio za paralelno računanje
3. k (30, k=51, k=64) - parametar kojim su se definirale veličine kmer-ova.
4. B (B=15, B=20) - parametar kojim se definirao tzv. Bloom filter mode kojim se nastojala rasteriti ukupna memorija prilikom pokretanja cijelog procesa
5. Lib (lib = 'pe1 pe2' pe1 - parametar kojim se definirao broj ulaznih datoteka te njihova putanja (pe1=' ./gams57_R1.fq.gz' pe2= './gams57_R1.fq.gz')

Primjer koda za Abyss:

abyss-pe k=30 B=15 s=200 name=Rupicapra in=' /home/ ttesija/ de_novo/ gams57_R1.fq.gz/ home/ ttesija/ de_novo/ gams57_R2.fq.gz'

Pokretanje *de novo* metode u alatu SPAdess izveden je s nekoliko različitih parametara:

1. spades.py - osnovni parametar kojim se definiralo da se u ovom koraku alat SPAdess koristi za *de novo* metodu (budući da ovaj program može raditi i druge analize).
2. -1, -2 - parametar kojim su se definirale putanje do FASTQ datoteka
3. t (-t 20, -t 8) - parametar kojim se definirao broj procesora koji će se koristiti prilikom pokretanja metode
4. -m (500) - parametar kojim se definiralo ograničenje memorije u gigabajtima Gb. Bez definiranja ovog parametra, SPAdes prekida proces ukoliko memorija dosegne postavljenu granicu od 250 Gb
5. --sc - parametar kojim se omogućilo pokretanje metoda sa zadanim veličinama kmer-ova 21, 33 i 55
6. -o - parametar kojim se definirala putanja direktorija u kojem će se pohraniti rezultati

Primjer koda za pokretanje SPAdess:

spades.py -1 gams57_R1.fq.gz -2 gams57_R2.fq.gz -t 8 -m 500 --sc -o /home /ttesija /de_novo /proba2

4 REZULTATI

4.1 Kontrola kvalitete genomskih podataka

Provedena je kontrola kvalitete za svih 12 uzoraka divokoze. FastQC alat je detektirao greške u četiri uzorka (Gams53, Gams85, OSIL-06, LM 2/07), točnije u R2 datotekama. Kod sva četiri uzorka, redoslijed sekvenci u R1 i R2 datotekama nije bio jednak zbog čega ih većina alata nije mogla očitati niti koristiti u svojim analizama. Izuzetak je NOVOPlasty alat koji je uspješno izolirao mitohondrijske sekvence iz tri uzorka. Uz navedena četiri uzorka, jedan uzorak (Gams65) je, zbog niske kvalitete i malog broja fragmenata, isključen iz daljnjih analiza. Trimmomatic alat je potom korišten za filtriranje broja fragmenata u svakom uzorku pomoću kojeg su se uklonili svi fragmenti koji su kraći od 150 pb, oni koji sadržavaju adapter sekvence te svi duplicirani fragmenti.

Nakon svih provedenih analiza za čišćenje podataka, sedam se uzoraka koristilo u svim daljnjim analizama za rekonstrukciju mtNDA i nDNA, dok su tri uzorka korištena isključivo u *de novo* metodi sastavljanja mtDNA. Informacije o svim uzorcima te njihov broj fragmenata prije i nakon čišćenja prikazani su u Tablici 6.

Tablica 6. Informacije o korištenim genomskim uzorcima nakon procesa čišćenja. Uzorci naznačeni sa zvjezdicom nisu prošli kontrolu kvalitete, ali su korišteni u de novo metodi za sastavljanje mtDNA.

Uzorak	ID uzorka	Podvrsta	Lokacija	Broj fragmenata	Broj fragmenata nakon čišćenja	Udio odbačenih fragmenata (%)
1.	B532	*R. r. tatrica*	Slovačka	241.085.630	238.968.476	0,87
2.	B539	*R. r. tatrica*	Slovačka	189.729.312	188.248.712	0,78
3.	Gams7	*R. r. balcanica*	Hrvatska	216.319.578	212.864.034	1,58
4.	Gams21	*R. r. rupicapra*	Hrvatska	222.981.802	219.682.804	1,48
5.	Gams57	*R. r.* hibrid	Hrvatska	261.340.198	257.903.832	1,31
6.	Gams108	*R. p. pyrenaica*	Španjolska	225.995.072	223.046.622	1,30
7.	Gams109	*R. p. pyrenaica*	Španjolska	278.609.626	274.569.732	1,45
8.*	Gams85	*R. r. rupicapra*	Hrvatska	244.129.826	-	-
9.*	OSIL-06	*R. r. rupicapra*	Slovenija	196.999.200	-	-
10.*	Gams53	*R. r.* hibrid	Hrvatska	235.807.388	-	-

4.2 Usporedba mtDNA sekvenci dobivenih metodom mapiranja i metodom *de novo*

Metoda mapiranja provedena je na sedam uzoraka korištenjem dvije referentne sekvence (pet uzoraka mapirano na genom sjeverne divokoze, dva uzorka mapirana na genom južne divokoze). Budući da se genomski podaci sastoje od velikog broja fragmenata, a referentne sekvence su duljine oko 16.000 pb, samo se mali udio fragmenata mapirao (između 0,01 i 1,55 %) što odgovara količini fragmenata mtDNA u genomskim uzorcima (Tablica 7.). S druge strane, metoda *de novo* provedena je na 10 uzoraka, a CYTB sekvence sjeverne i južne divokoze (duljina 1.143 pb) korištene su kao seed sekvence. U Tablici 7. prikazane su dobivene vrijednosti korištenjem dviju različitih metoda sastavljanja mtDNA. U posljednjem

stupcu prikazan je broj pronađenih varijanti nakon usporede para sekvenci dobivenih iz istog uzorka korištenjem različitih metoda.

Tablica 7. Rezultati metoda za sastavljanje mtDNA korištenjem mapiranja i *de novo* metode. Uzorci označeni sa zvjezdicom korišteni su u metodi *de novo*. Rezultati BLAST analize sličnosti računate su za mtDNA sekvence dobivene *de novo* metodom. U posljednjem stupcu podebljani brojevi predstavljaju broj polimorfizama pronađenih između sekvenci.

Uzorak	**Metoda mapiranjem**		**Metoda *de novo***				**Broj varijabilnih mjesta između parova sekvenci dobivenih različitim metodama**
	Prosječna pokrivenost (*x*)	**Postotak mapiranih fragmenata (%)**	**Prosječna pokrivenost (*x*)**	**Postotak mapiranih fragmenata (%)**	**Duljina novih mtDNA sekvenci (pb)**	**BLAST analiza sličnosti (%)**	
B532	9.150	0,57	11.234	0,51	16.434	98,99	**2**
B539	13.660	1,13	18.018	1,03	16.434	98,99	0
Gams7	6.355	0,38	6.474	0,33	16.432	99,03	0
Gams21	21.168	1,55	27.612	1,35	16.432	99,03	0
Gams57	8.393	0,43	8.859	0,37	16.433	99,04	0
Gams108	230	0,01	174	0,01	16.438	99,89	**268**
Gams109	1.282	0,06	1.174	0,05	16.438	99,73	0
Gams53*			6.739	0,31	16.432	99,03	
Gams85*			4.681	0,21	16.433	99,17	
OSIL-06*			3.905	0,22	16.433	99,17	

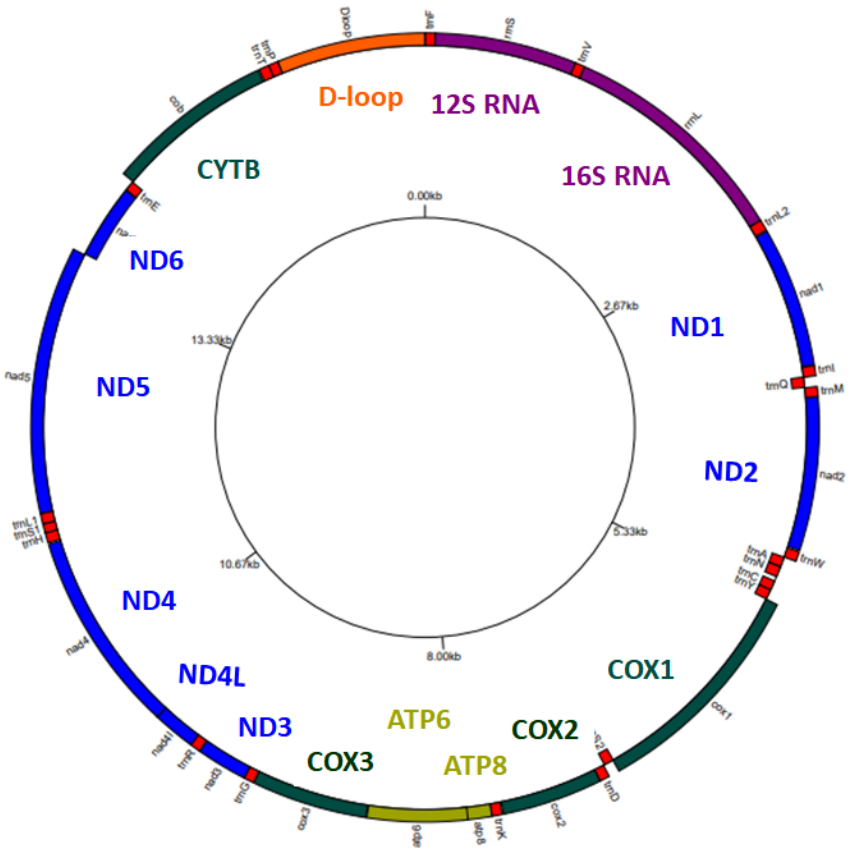
Obje metode za sastavljanje mtDNA rezultirale su identičnim sekvencama kod pet uzoraka, dok su usporedbom sekvenci kod dva uzorka pronađeni polimorfizmi. Sekvence dobivene iz uzorka B532 razlikovale su se u dvije baze, dok su se sekvence dobivene iz uzorka Gams108 razlikovale u 268 baza. Prosječna pokrivenost i postotak mapiranih fragmenata vrlo su slični među gotovo svim uzorcima, osim kod uzorka Gams108 koji i u ovim vrijednostima značajno odskače u odnosu na vrijednosti drugih uzoraka. Razlike u prosječnoj pokrivenosti i postotku mapiranih fragmenata između dvije metode očekivane su s obzirom da sam broj mtDNA fragmenata unutar genomskih podataka jako varira (različita tkiva imaju drugačiji broj mitohondrijskih stanica) (Al-Nakeeb i sur., 2017). Međutim, kada se usporede sekvence iz

istog uzorka, vrijednosti parametara su jako slične za obje metode. Uzrok velike varijabilnosti unutar uzorka Gams108 može biti kontaminacija (bakterijske DNA, virusne DNA i sl.) koju alati za provjeru kvalitete nisu prepoznali. Drugi razlog može biti vrsta tkiva iz kojeg se uzorkovala genomska DNA. Drugim riječima, manja koncentracija mtDNA u uzorku rezultirala je smanjenim udjelom mapiranih fragmenata. Za određivanje kompletnosti, obje sekvence uzorka Gams108 su provjerene pomoću BLAST-a. BLAST pretraga je potvrdila da je sekvenca iz uzorka Gams108 dobivena mapiranjem 98,30 % slična s referentnom sekvencom južne divokoze (FJ207538.1), dok je ista sekvenca dobivena *de novo* metodom 99,89 % slična s referentnom sekvencom (Tablica 7.).

Iako su obje metode rezultirale gotovo identičnim sekvencama, vremenski period potreban za provođenje ovih metoda jako se razlikuje. Metoda mapiranja računalno je i vremenski zahtjevnija i sastoji se od više koraka. S druge strane, korištenje NOVOPlasty alata je puno brže te je za pokretanje programa dovoljno izmjeniti putanje do direktorija u kojem se nalaze uzorci (ostali se parametri ne mijenjaju). Osim toga, NOVOPlasty je uspješno sastavio mtDNA iz tri neispravna uzorka (Gams53, Gams85, OSIL-06). Prema tome, 10 sekvenci dobivenih *de novo* metodom odabrane su za provođenje preostalih analiza (anotacija i filogenija). BLAST analiza provedena na 10 (*de novo*) sekvenci (Tablica 7.) rezultirala je s visokim postotkom sličnosti između svih NOVOPlasty sekvenci i njihovih referenci iz Banke Gena. Najveći postotak sličnosti zabilježen je kod uzorka Gams108 (99,89 %) dok su sekvence dobivene iz neispravnih uzoraka (Gams53, Gams85, OSIL-06) rezultirale većim postotkom sličnosti u odnosu na uzorke koji su prošli kvalitetu. Ovim rezultatom se potvrdilo da se alatom NOVOPlasty mogu dobiti kvalitetne sekvence i iz uzoraka loše kvalitete.

4.3 Anotacija mtDNA

MITOS i GeSeq web aplikacije korištene su za provođenje automatske anotacije svih mtDNA. Obje aplikacije uspješno su anotirale sva mjesta u mtDNA uključujući 13 PCG-a, 2 rRNA, 22 tRNA te CR. Na Slici 6. prikazana je struktura organizacije mtDNA karakteristična za rod *Rupicapra* a istaknuti su PCG, rRNA i CR. Geni obojeni istom bojom pripadaju istoj genskoj skupini. Gen NAD6 i 7 tRNA gena nalaze se na komplementarnom lancu što je karakteristično za mtDNA sisavaca.



Slika 6. Prikaz strukture i organizacije mtDNA roda *Rupicapra*. Na slici je označeno 13 gena koji kodiraju za proteine (od kojih se samo ND6 gen nalazi na lakom L lancu), dva RNA gena (12S RNA, 16S RNA) te kontrolna regija (D-loop).

U Tablici 8. prikazani su rezultati MITOS i GeSEQ anotacije za PCG gene, a prikazana su samo tri uzorka, kao primjeri, budući da su rezultati isti za sve uzorke. Iako su oba anotatora pronašla sve kodirajuće regije u mtDNA, definirani START i STOP kodoni nisu identični za sve pronađene gene. Geni kojima su anotatori predvidjeli drugačije START i STOP kodone podebljani su u tablici (ND1, ND2, ND3, ND5). Ovakav rezultat može biti posljedica veće varijabilnosti u ovim regijama.

Tablica 8. Rezultati alata MITOS i GeSeq za tri mtDNA dobiveni iz uzoraka B532, B539 i Gams7. Podebljani brojevi označavaju gene (ND1, ND2, ND3 i ND5) čije se START i STOP pozicije razlikuju.

PCG	B532		B539		GAMS7	
	MITOS	**GeSeq**	**MITOS**	**GeSeq**	**MITOS**	**GeSeq**
ATP6	7932-8612	7932-8612	7932-8612	7932-8612	7931-8611	7931-8611
ATP8	7771-7971	7771-7971	7771-7971	7771-7971	7770-7970	7770-7970
COX1	5328-6872	5328-6872	5328-6872	5328-6872	5327-6871	5327-6871
COX2	7015-7698	7015-7698	7015-7698	7015-7698	7014-7697	7014-7697
COX3	8612-9395	8612-9395	8612-9395	8612-9395	8611-9394	8611-9394
CYTB	14155-15297	14155-15297	14155-15297	14155-15297	14154-15296	14154-15296
ND1	**2741-3697**	**2741-3696**	**2741-3697**	**2741-3696**	**2740-3696**	**2740-3695**
ND2	**3906-4949**	**3906-4947**	**3906-4949**	**3906-4947**	**3905-4948**	**3905-4946**
ND3	**9474-9821**	**9465-9810**	**9474-9821**	**9465-9810**	**9473-9820**	**9464-9809**
ND4	10171-11548	10171-11548	10171-11548	10171-11548	10170-11547	10170-11547
ND4L	9881-10177	9881-10177	9881-10177	9881-10177	9880-10176	9880-10176
ND5	**11741-13570**	**11750-13570**	**11741-13570**	**11750-13570**	**11740-13569**	**11749-13569**
ND6	13554-14081	13554-14081	13554-14081	13554-14081	13553-14080	13553-14080

Za validaciju rezultata anotacije, korištena je dostupna anotacija referentnog mitogenoma sjeverne divokoze u Banci gena. Prvo je provedeno zajedničko poravnjanje 12 mtDNA sekvenci (10 sekvenci i dvije referentne). Iz zajedničkog poravnjanja su se izolirali svi geni čije su se START i STOP pozicije usporedile s dostupnim anotacijama iz Banke gena. Informacije o točnim pozicijama gena kao i njihova duljina te START i STOP kodoni prikazani su u Tablici 9. Pronađeni START (ATG i ATA) i STOP kodoni (ATA, T) karakterisični su za mitohondijsku sekvencu sisavaca (Xiufeng i Árnason, 1994; Taanman, 1999; Gupta i sur., 2015). Sekvence kodirajućih regija PCG-a činile su 69,3 % (11395 pb) ukupne mtDNA.

Tablica 9. Prikaz duljina i pozicija te START i STOP kodona 13 protein-kodirajućih regija mtDNA.

PCG	POZICIJA	DULJINA	START	STOP
ATP6	7937-8616	680	ATG	ATA
ATP8	7776-7976	200	ATG	ATA
COX1	5333-6877	1544	ATG	ATA
COX2	7020-7702	683	ATG	ATA
COX3	8617-9400	784	ATG	T
CYTB	14160-15300	1141	ATG	ATA
ND1	2746-3702	955	ATG	T
ND2	3911-4952	1042	ATA	T
ND3	9470-9815	346	ATA	T
ND4L	9886-10180	295	ATG	T
ND4	10176-11553	1378	ATG	T
ND5	11755-13573	1819	ATA	T
ND6	13559-14086	528	ATG	T

4.4 Filogenetske analize mtDNA

Za provedbu filogenetskih analiza, izrađena su dva seta podataka. Prvi set se sastojao od poravnanja 15 sekvenci mtDNA divokoza (deset iz ove disertacije, dvije referentne te tri sekvence dostupne u Banci gena) dok se drugi set sastojao od poravnanja 40 dostupnih sekvenci mtDNA porodice *Caprini* te pet uzoraka iz porodice *Bovidae*.

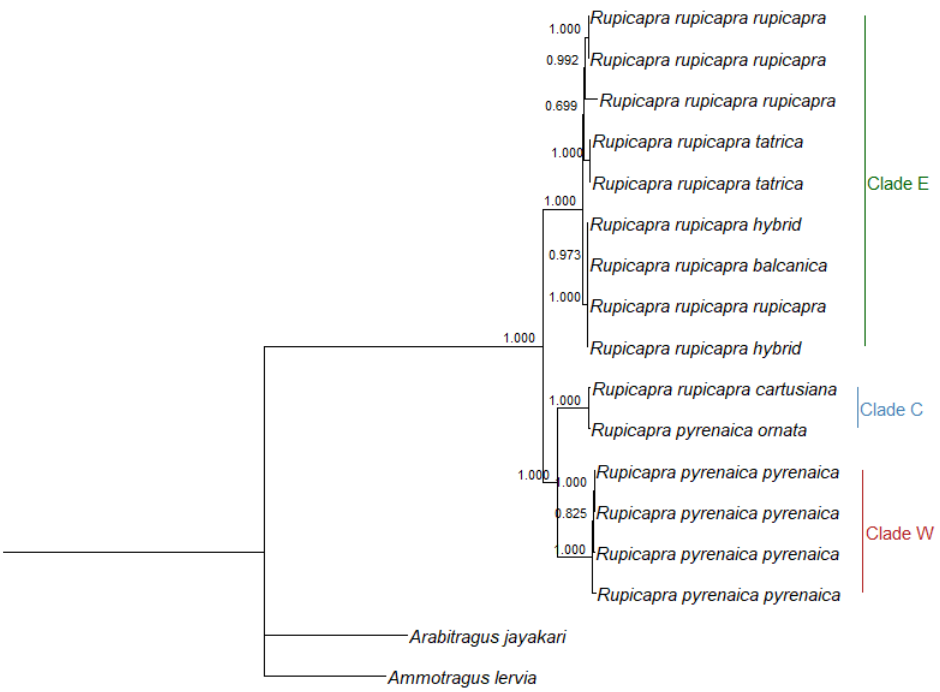
Zajedničko poravnanje sekvenci divokoza sastojalo se od 16.433 nukleotida. DNAsp alat korišten za procjenu opisnih parametara skupina (vrsta i podvrsta) poravnatih sekvenci a rezultati su prikazani u Tablici 10.

Tablica 10. Varijabilnost mtDNA sekvenci. N uzoraka = broj uzoraka, N Haplotipova = broj haplotipova, S = broj polimorfnih mjesta, π = nukleotidna raznolikost, h = haplotipna raznolikost, SD = standardna devijacija, Fs = Fu-ova Fs statistika. FJ207539, FJ207538 (Hassanin i sur. (2009)); KJ184175, KJ184173, KJ184174 (Pérez i sur. (2014)).

Podvrste	**N Uzoraka**	**N Haplotipovi**	**S**	**π (SD)**	**h (SD)**	**Fs**	**Naziv sekvence**
R. rupicapra (referenca)	1	1					FJ207539
R. r. rupicapra	3	2		0,003 (0,001)	0,667 (0,314)	7,456	Gams21 Gams85 OSIL
R.r. balcanica	1	1					Gams7
R. r. rupicapra x R. r. balcanica	2	2	1	0,000	1,000 (0,500)	0,00	Gams53 Gams57
R.r. tatrica	2	1					B532 B539
R. r. cartusiana	1	1					KJ184175
***R. rupicapra* UKUPNO**	**10**	**6**	**679**	**0,01 (0,004)**	**0,889 (0,075)**	**13,486**	
R. pyrenaica (referenca)	1	1					FJ207538
R. p. ornata	1	1					KJ184173
R. p. pyrenaica	3	3	54	0,002 (0,0007)	1,000 (0,727)	2,467	KJ184174 Gams108 Gams109
***R. pyrenaica* UKUPNO**	**5**	**5**	**479**	**0,012 (0,006)**	**1,00 (0,1269)**	**2,955**	
UKUPNO	**15**	**11**	**930**	**0,021 (0,003)**	**0,952 (0,040)**	**11,944**	

Među 15 mtDNA sekvenci divokoza zabilježena je visoka haplotipna (h=0,952, SD=0,040) i nukleotidna raznolikost (π=0,021, SD=0,003) s velikom varijabilnosti prisutnom između dvije vrste, ali i među podvrstama. Velika varijabilnost između *R. rupicapra* i *R. pyrenaica* uočena je u PCG regijama, pri čemu je *R. rupicapra* općenito pokazala nižu raznolikost, unatoč većem broju uzoraka.

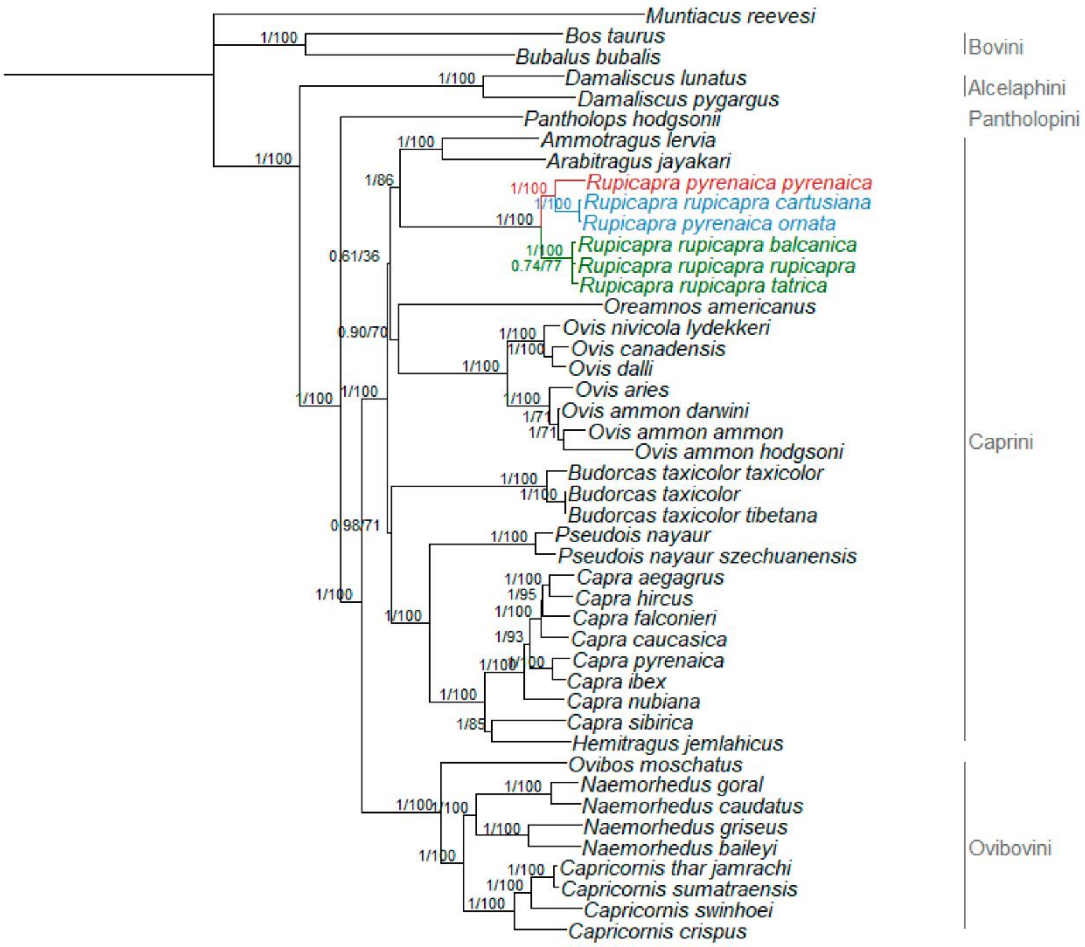
Korištene metode (maksimalna vjerodostojnost i Bayesovska metoda) u rekonstrukciji filogenije rezultirale su identičnim filogenetskim stablima za rod *Rupicapra*. Filogenetsko stablo roda *Rupicapra* dobiveno Bayesovskom metodom prikazano je na Slici 7.



Slika 7. Ukorijenjeno filogenetsko stablo dobiveno Bayesovskom metodom za rod *Rupicapra*. Iznad čvorova prikazane su Bayesovske posteriorne vjerojatnosti. Označeni klasteri: crvena – klaster W; plava – klaster C, zelena – klaster E.

Na stablu su jasno definirana tri klastera (E – istočni, C – centralni, W – zapadni) s tim da su zapadni i centralni klaster međusobno sličniji (distance između grupa W:C = 0,024; W:E = 0,032 ; C:E = 0,030). Unutar istočnog klastera, *R. r. balcanica* (uključujući i uzorke hibrida) je pokazala najveću stopu diferencijacije u odnosu na ostale sekvence, dok su *R. r. tatrica* i *R. r. rupicapra* bile u sestrinskom odnosu.

Filogenetska analiza roda *Caprinae* provedena je na setu od 40 mtDNA sekvenci roda *Caprinae* te od 5 sekvenci roda *Bovidae*. Obje metode (maksimalna vjerodostojnost i Bayesovska metoda) korištene za rekonstrukciju filogenetskih odnosa rezultirale su istom topologijom. Na Slici 8. prikazano je filogenetsko stablo dobiveno Bayesovskom metodom.

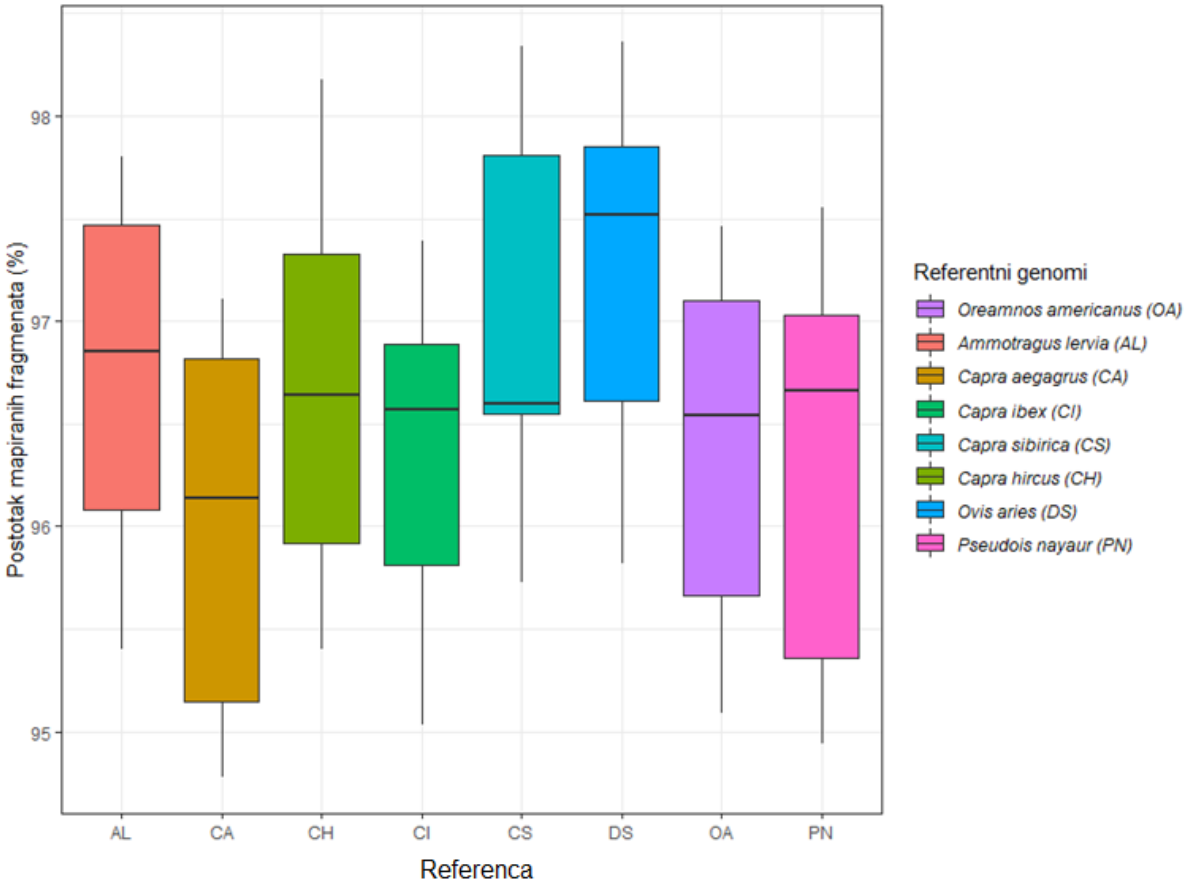


Slika 8. Ukorijenjeno filogenetsko stablo dobiveno Bayesovskom metodom za rod *Caprinae*. Iznad čvorova prikazane su Bayesovske posteriorne vjerojatnosti te vrijednosti za maksimalnu vjerodostojnost (npr. 1/100). Obojene grane predstavljaju klastere: crvena – klaster W; plava – klaster C, zelena – klaster E.

Filogenetsko stablo *Caprinae* rezultiralo je identičnom toplogijom kao i stablo iz Hassanin i sur. (2009) te Pérez i sur. (2014). Na stablu se jasno može definirati da su rodovi *Bovini* i *Caprini* monofiletični te je unutar roda *Ovicaprini* jasno podržano odvajanje *Caprini* i *Ovibovini*. Osim toga, vidljivo je da je rod *Damaliscus* u sestrinskom odnosu s rodom *Ovicaprini* te je vidljiva bazalna divergencija roda *Pantholopini* od *Ovicaprini*. Unutar roda *Caprinae*, *Capra* i *Hemitragus* su formirali tzv. goat-like klaster koji su u sestrinskom odnosu s rodovima *Pseudois* i *Budorcas*. *Oremanos* se grupirao s ostalim ovcama u tzv. sheep-like klasteru dok su se *Ammotragus* i *Arabitragus* bili u sestrinskom odnosu s rodom *Rupicapra*, koji je monofiletičan. Ovi rezultati objavljeni su u radu „A Mother's Story, Mitogenome Relationships in the Genus Rupicapra" (Iacolina i sur., 2021).

4.5 Provjera kompletnosti i anotacija dobivenih nDNA sekvenci

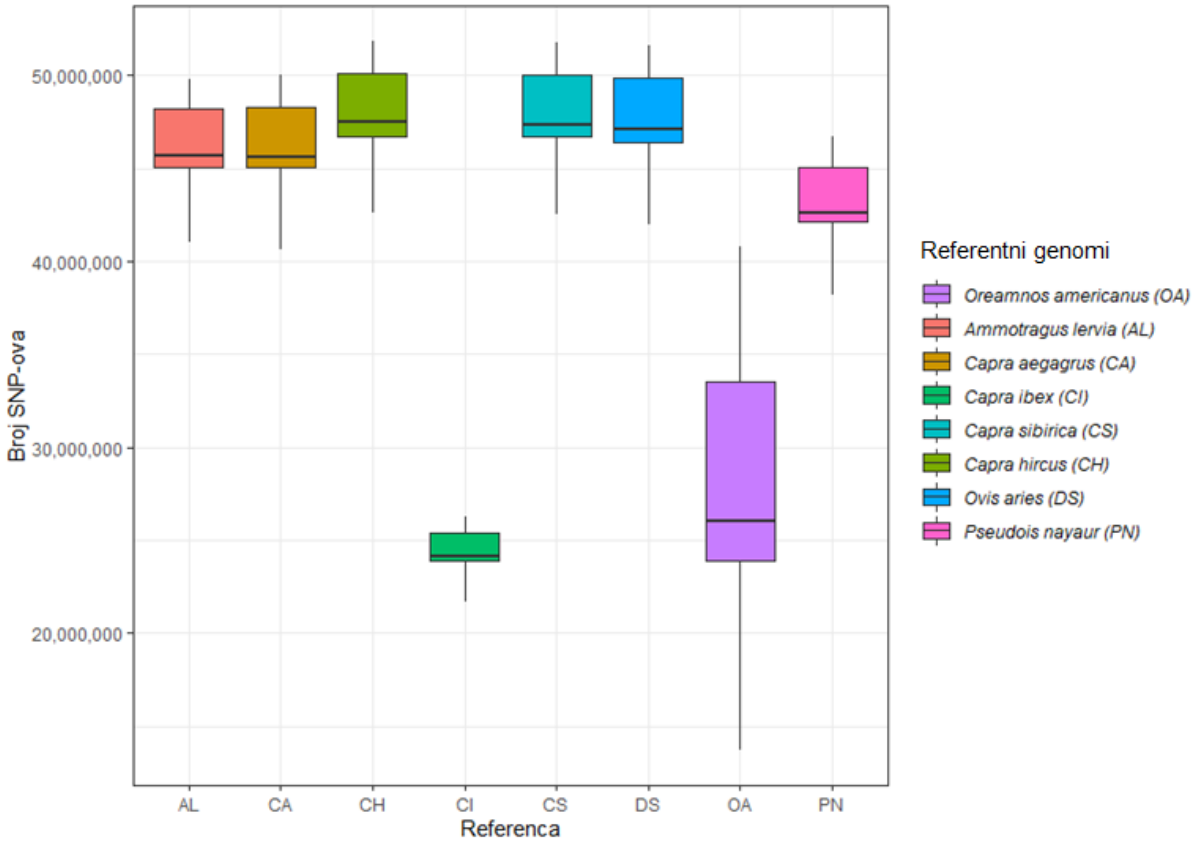
U prvoj fazi mapiranja, osam genoma iz Banke gena je korišteno kao referentne sekvence na koje se mapiralo sedam uzoraka divokoze, čineći ukupno 56 kombinacija (uzorak x referenca). Prosječna pokrivenost mapiranih uzoraka iznosila je između 8x i 14x. Vrijednosti proporcija, odnosno udijeli svih mapiranih fragmenata iz uzoraka na različite reference prikazani su na Slici 9.



Slika 9. Udio mapiranih fragmenata divokoze (7 uzoraka) na osam referentnih sekvenci, iskazan u postotcima. x os označava reference, y os označava vrijednosti proporcija.

Proporcije mapiranih fragmenata svih uzoraka na različite referentne genome su bile vrlo slične, u rasponu od 95 do 98 %. Najveću proporciju imali su uzorci mapirani na genom ovce koja je, u odnosu na preostale vrste, genetski najudaljenija od divokoze. S druge strane, najniži postotak mapiranja imali su uzorci mapirani na genom divlje koze koja je najsrodnija

divokozi. Nakon validacije i filtriranja mapiranih fragmenata u BAM datotekama, proveden je korak pozivanja SNP-ova, a rezultati nakon njihovog filtriranja prikazani su na Slici 10.



Slika 10. Broj detektiranih SNP-ova pronađenih između uzoraka divokoza i referenci nakon filtriranja. Na x osi nalaze se reference dok se na y osi nalaze brojevi SNP-ova nakon filtriranja.

Broj detektiranih, nefiltriranih SNP-ova kretao se između 28.572.067 i 60.429.454. Nakon procesa filtriranja prema zadanim kriterijima (QUAL > 15; DP > 5; MQ > 20), broj SNP-ova kretao se između 13.740.102 i 51.814.829 (Slika 10.). Broj pronađenih SNP-ova između uzoraka i referentnih genoma bio je vrlo sličan za šest referenci, dok je taj broj bio znatno manji za reference alpskog kozoroga (*C. ibex*) i američke planinske koze (*O. americanus*). Posljedica manjeg broja pronađenih SNP-ova kod ove dvije reference može biti uzrokovan metodom mapiranja ili se može pripisati njihovoj pristranosti. Filtrirani SNP-ovi u kombinaciji s referentnim sekvencama potom su korišteni za pozivanje 56 konsenzusnih sekvenci zapisanih u FASTA formatu.

BUSCO alat korišten je za provjeru kompletnosti i anotaciju 56 konsenzusnih sekvenci, ali i za 8 korištenih referenci kako bi se provjerila njihova kvaliteta i kompletnost. Svaka sekvenca zasebno je analizirana pomoću BUSCO-a s ciljem pronalaženja svih prisutnih struktura gena. Rezultati BUSCO analize za provjeru kompletnosti prikazani su u Tablici 11.

Tablica 11. Rezultati BUSCO analize za procjenu kompletnosti novosastavljenih genoma izraženi u postotcima (%). U redcima se nalaze uzorci divokoze i sama referenca.

	Oreamnos americanus	***Ammotragus lervia***	***Capra aegagrus***	***Capra hircus***	***Capra ibex***	***Capra sibirica***	***Ovis aries***	***Pseudois nayaur***
Referenca	86,8	92,6	86,7	93,8	50,3	93,8	93,0	59,0
B532	86,7	92,5	86,6	93,7	50,2	93,7	93,0	59,0
B539	86,7	92,5	86,5	93,6	50,1	93,6	93,0	58,9
Gams7	86,8	92,6	86,6	93,7	50,1	93,7	92,9	58,9
Gams21	86,7	92,6	86,6	93,8	50,1	93,7	92,9	59,1
Gams57	86,7	92,6	86,6	93,8	50,1	93,7	93,0	59,0
Gams108	86,7	92,7	86,7	93,8	50,1	93,7	93,1	59,0
Gams109	86,7	92,6	86,6	93,7	50,1	93,6	93,0	59,0

Rezultati kompletnosti sekvenci kretali su se između 50 i 93 %. Uzorci mapirani na četiri sekvence imali su kompletnost veću od 90 % što se smatra velikim postotkom (Jauhal i Newcomb, 2021). U gotovo svim slučajevima, referentne sekvence iz Banke gena su imale najveće BUSCO vrijednosti u usporedbi s kombinacijama tih referenci i uzoraka. Iznimke su domaća koza i ovca kod kojih je većina kombinacija uzoraka i reference imala iste BUSCO vrijednosti kao i referenca što može biti posljedica visoke kvalitete referentnog genoma ovih dviju vrsta. S druge strane, rezultati kompletnosti kombinacija uzoraka divokoze s alpskim kozorogom i baralom bili su iznimno malih vrijednosti (između 50 i 59 %) što ukazuje na jako lošu kvalitetu te reference čija je kompletnost također bila niska. Ovo je dobar primjer koji pokazuje da se reference, bez obzira na informacije iz Banke gena, uvijek treba testirati na kompletnost.

U procesu anotacije sekvenci, BUSCO je na temelju referenetnog seta ortologa dao rezultate o broju pronađenih gena te izlazne FASTA datoteke u kojima su pronađene strukture gena zapisane u aminokiselinskom obliku. Rezultati BUSCO anotacije prikazane su u Tablici 12.

Tablica 12. Broj pronađenih BUSCO gena u novosastavljenim genomima. Podebljani brojevi predstavljaju broj gena koji su pronađeni u svim kombinacijama uzoraka i referenci dok broj označen sa zvjezdicom predstavlja zajednički broj gena pronađenih u svim kombinacijama nakon uklanjanja svih genoma dobiveni mapiranjem uzoraka na referentni genom vrste *Capra ibex*.

	Oreamnos americanus	***Ammotragus lervia***	***Capra aegagrus***	***Capra hircus***	***Capra ibex***	***Capra sibirica***	***Ovis aries***	***Pseudois nayaur***
Referenca	11.571	12.352	11.559	12.504	6.702	12.510	12.397	7.866
B532	11.561	12.340	11.552	12.492	6.689	12.500	12.402	7.864
B539	11.560	12.338	11.541	12.486	6.684	12.486	12.400	7.856
Gams7	11.574	12.343	11.554	12.492	6.686	12.491	12.394	7.848
Gams21	11.567	12.337	11.555	12.502	6.688	12.496	12.387	7.878
Gams57	11.568	12.344	11.558	12.503	6.685	12.493	12.407	7.862
Gams108	11.568	12.358	11.563	12.508	6.687	12.501	12.410	7.868
Gams109	11.565	12.347	11.547	12.489	6.686	12.485	12.409	7.862
Zajednički geni	11.323	12.365	11.313	12.282	6.542	12.287	12.183	7.665
Broj zajedničkih gena u svim kombinacijama uzoraka i referenci								**3.093**
Broj zajedničkih gena nakon uklanjanja svih kombinacija uzoraka s referencom *Capra ibex*								**5.739***

Broj anotiranih BUSCO gena odgovara rezultatima kompletnosti iz Tablice 11., ali su u ovom slučaju, rezultati iskazani u broju pronađenih BUSCO gena. Od ukupno 13.335 gena dostupnih u referentnom ortolognom setu, BUSCO je za svaku kombinaciju tražio maksimalni broj prisutnih sastavljenih genskih sekvenci. I u ovom slučaju najveći broj pronađenih gena imali su uzorci mapirani na reference domaće koze, sibirskog kozoroga te domaće ovce. Budući da je glavni cilj bio usporediti sve uzorke kako bi se analizirali odnosi između mapiranih uzoraka na različite reference, bilo je potrebno detektirati sve zajedničke gene koji su prisutni u svim kombinacijama. U posljednjem retku u Tablici 12. navedeni su zajednički geni pronađeni u svim kombinacijama (npr. za američku planinsku kozu, broj pronađenih gena varirao je među uzorcima, dok je 11.323 zajedničkih gena pronađeno u svim kombinacijama uzoraka i te reference). Na kraju posljednjeg retka, broj 3.093 prikazuje ukupni broj gena prisutnih u svim kombinacijama i referencama. Međutim, prije provedbe daljnjih analiza, inicijalnim poravnanjem nekoliko dobivenih gena se ustanovilo da su nasumično odabrani i poravnati fragmenti bili identični kod svih uzoraka mapiranih na reference alpskog te sibirskog kozoroga. Dodatnom provjerom ovih sekvenci ustanovilo se da su navedene referentne sekvence identične iako imaju različite informacije te drugačiji pristupni broj u Banci gena. Budući da su kombinacije s referencom sibirskog kozoroga

rezultirale većim brojem pronađenih gena, ista je zadržana u daljnim analizama dok su se referenca alpskog kozoroga i njene konsenzusne sekvence isključile iz daljnjih analiza. Nakon isključivanja ove reference, ukupno je bilo 5.739 gena prisutnih u svim kombinacijama iz kojih su slučajnim odabirom kreirana tri podatkovna seta gena.

4.6 Analize sličnosti uzoraka i referenci

Statističke analize sličnosti provedene su na tri seta gena. U svim analizama korištena su poravnanja (ovisno o odabranom setu) koja su se, nakon isključenja sekvenci alpskog kozoroga, sastojala od 56 sekvence (46 iz konsenzusnih sekvenci i 8 iz sekvenci iz referentnim genomima) a rezultati su prikazani prema koracima opisanim u protokolu u poglavlju Materijali i metode:

1. Zajednički set 10 nasumično odabranih BUSCO gena:
    a. Zajedničko poravnanje sastojalo se od 56 sekvenci, a prosječna duljina poravnanja bila je 923,5 aminokiselinskih baza.
    c. Zajednička matrica genetskih udaljenosti (izračunata iz 10 matrica udaljenosti za svih 10 setova) prikazana je u Tablici 13. Prema rezultatima se vidi da su genetske udaljenosti zabilježene kod prvog seta puno veće (u odnosu na udaljenosti iz drugog seta, Tablica 14.) što je jasno budući da su sve matrice izračunate iz puno kraćih poravnanja. Sličnosti između sekvenci jednog poravnatog gena znatno su varirale te na primjer, za pojedine gene udaljenosti su bile vrlo male ili nisu postojale. Osim toga, najveće udaljenosti su pronađene usporedbom uzoraka s referencama. Drugim riječima, svaki uzorak divokoze je bio sličniji drugim uzorcima divokoze nego bilo kojoj referenci. Par iznimki vidljivo je kod npr. uzorka B532 koji je bio sličniji uzorku B539 nego što je bio sekvencama istog uzorka (B532) mapiranog na različite reference ili kod uzorka Gams21 koji je bio sličniji uzorku Gams57 nego što je bio sekvencama uzorka Gams21 mapiranog na različite reference. Ovakav rezultat može se tumačiti kao posljedica korištenja kratkih fragmenata visoke sličnosti s jako malim brojem polimorfizama.

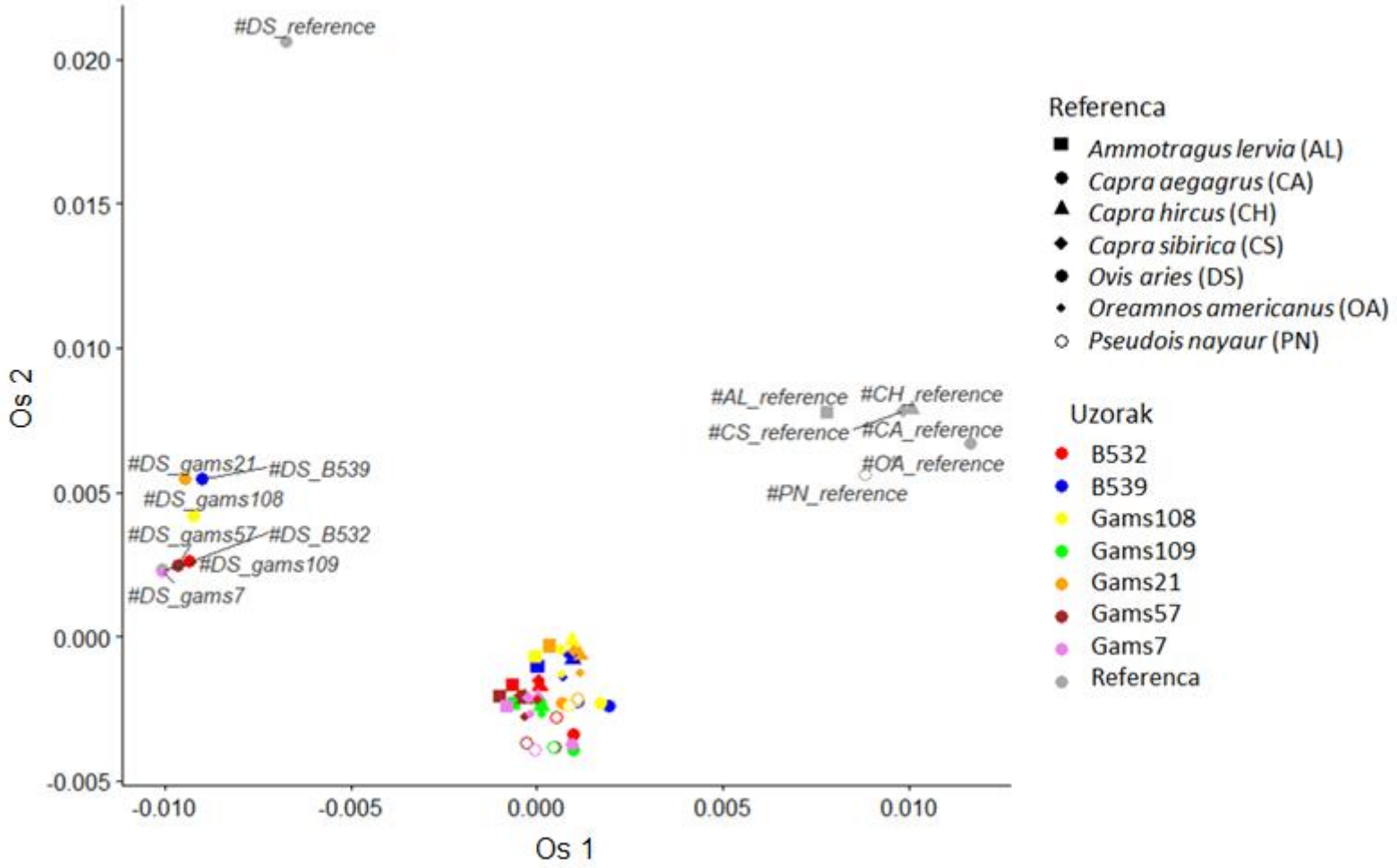
Tablica 13. Zajednička genetska matrica (konstruirana iz 10 genetskih matrica udaljenosti) iz seta od 10 BUSCO gena. Svaki stupac predstavlja isti uzorak mapiran na 8 različitih referentnih genoma. Prvi brojevi predstavljaju aritmetičke sredine, a drugi raspone na temelju 10 matrica udaljenosti izračunatih za svaki gen.

	B532	**B539**	**Gams108**	**Gams109**	**Gams21**	**Gams57**	**Gams7**	**Reference**
B532	0,0169 (0-0,0986)							
B539	0,0167 (0-0,0892)	0,0172 (0-0,1035)						
Gams108	0,0205 (0-0,0964)	0,0211 (0-0,0974)	0,0186 (0-0,1090)					
Gams109	0,0195 (0-0,0927)	0,0202 (0-0,0938)	0,0199 (0-0,1008)	0,0174 (0-0,1025)				
Gams21	0,0178 (0-0,0876)	0,0181 (0-0,0887)	0,0220 (0-0,0959)	0,0212 (0-0,1025)	0,0192 (0-0,1015)			
Gams57	0,0168 (0-0,0874)	0,0173 (0-0,0886)	0,0213 (0-0,0959)	0,0202 (0-0,0922)	0,0181 (0-0,0871)	0,0178 (0-0,1013)		
Gams7	0,0157 (0-0,0892)	0,0166 (0-0,0902)	0,0206 (0-0,0975)	0,0195 (0-0,0938)	0,0178 (0-0,0886)	0,0164 (0-0,0886)	0,0167 (0-0,1025)	
Reference	0,0232 (0-0,0786)	0,0222 (0-0,0776)	0,0260 (0-0,0867)	0,0255 (0-0,0823)	0,0233 (0-0,0779)	0,0232 (0-0,0780)	0,0235 (0-0,0769)	0,0201 (0-0,0616)

2. Zajednički set 100 nasumično odabranih BUSCO gena:
    a. Zajedničko poravnanje sastojalo se od 56 sekvenci, a prosječna duljina poravnanja svih 10 poravnanja bila je 12.182.3 aminokiselinskih baza.
    c. Zajednička matrica genetskih udaljenosti (izračunata iz 10 matrica udaljenosti za svih 10 setova) prikazana je u Tablici 14. U usporedbi s rezultatima prvog seta (Tablica 13.), može se zaključiti da su u ovom setu najmanje genetske udaljenosti veće od 0 što je očekivano budući da se u većim poravnanjima očekuje veći broj polimorfizama. Osim toga, svaki uzorak je sebi bio najsličniji (npr. sve sekvence dobivene mapiranjem uzorka Gams21 na različite reference su si bile najsličnije), a najveće su udaljenosti pronađene usporedbom uzoraka s referencama. Drugim riječima, svaki uzorak divokoze je bio sličniji drugim uzorcima divokoze nego bilo kojoj referenci.
    d. Spajanjem sekvenci svih 100 gena u jednu sekvencu rezultiralo je poravnanjem duljine od 136.459 aminokiseliskih baza.
    e. Nova genetska matrica udaljenosti izračunata iz zajedničkog poravnanja (136.459) korištena je za grafički prikaz svih kombinacija uzoraka i referenci u dvodimenzionalnom prostoru, a rezultat je prikazan na Slici 11. Prema rezultatima MDS analize jasno je vidljiva velika sličnost uzoraka divokoza mapiranih na šest različitih referenci. Štoviše, svaki uzorak divokoze, neovisno o referenci na koju je mapiran, bio je bliži drugim uzorcima divokoze nego svojoj referenci. Iznimka su uzorci divokoze mapirani na genom domaće ovce koji su bili udaljeniji od drugih uzoraka ali i od vlastite reference. Međutim, jasno je vidljivo da su svi uzorci bili proporcionalno udaljeni od svojih referenci, uključujući i uzorke mapirane na genom domaće ovce. Osim toga, razlike između uzoraka mapiranih na genom ovce i ostalih uzoraka bila je oko 0,01 što znači da, iz ukupnog poravnanja od 136.459 aminokiselinskih baza, u sekvencama ovih uzoraka bilo je prisutno oko 1300 polimorfizama što predstavlja svega 1 % ukupnog poravnanja. S druge strane, razlike između tih divokoza i reference domaće ovce bile su nešto veće (oko 1,5 %).

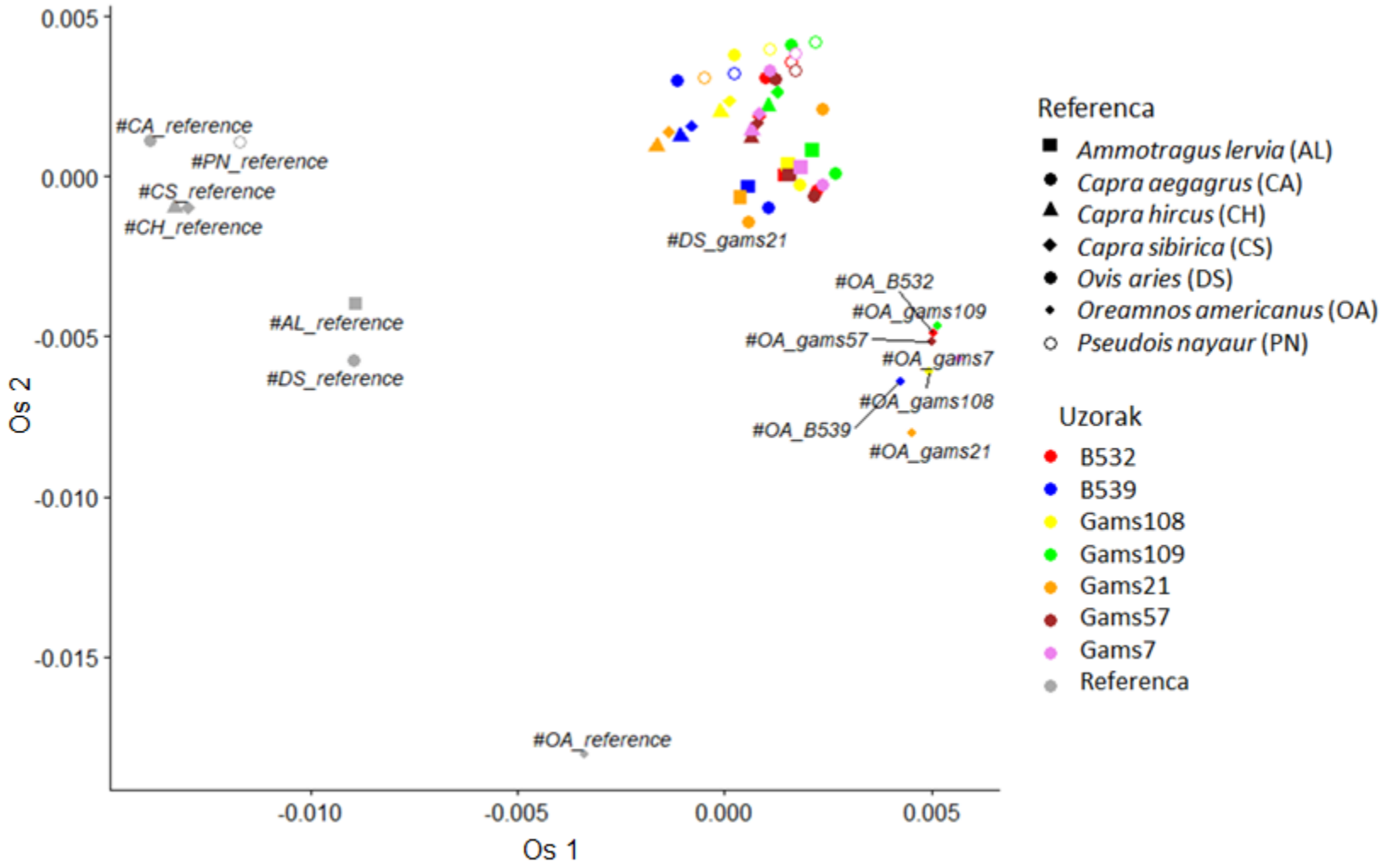
Tablica 14. Zajednička genetska matrica (konstruirana iz 10 genetskih matrica udaljenosti) iz seta od 100 BUSCO gena. Svaki stupac predstavlja isti uzorak mapiran na 8 različitih referentnih genoma. Prvi brojevi predstavljaju aritmetičke sredine, a drugi raspone na temelju 10 matrica udaljenosti izračunatih za svaki set od 10 gena.

	B532	**B539**	**Gams108**	**Gams109**	**Gams21**	**Gams57**	**Gams7**	**Reference**
B532	0,0076 (0,0020-0,0157)							
B539	0,0097 (0,0039-0,0165)	0,0093 (0,0043-0,0162)						
Gams108	0,0107 (0,0063-0,0189)	0,0116 (0,0067-0,0197)	0,0082 (0,0043-0,0172)					
Gams109	0,0114 (0,0042-0,0188)	0,0123 (0,0055-0,0199)	0,0104 (0,0049-0,0185)	0,0081 (0,0023-0,0164)				
Gams21	0,0117 (0,0059-0,0184)	0,0125 (0,0065-0,0191)	0,0125 (0,0070-0,0213)	0,0135 (0,0063-0,0216)	0,0109 (0,0045-0,0193)			
Gams57	0,0093 (0,0050-0,0169)	0,0105 (0,0057-0,0175)	0,0110 (0,0057-0,0204)	0,0115 (0,0061-0,0202)	0,0113 (0,0055-0,0191)	0,0081 (0,0040-0,0164)		
Gams7	0,0098 (0,0047-0,0155)	0,0112 (0,0055-0,0174)	0,0120 (0,0062-0,0200)	0,0125 (0,0064-0,0195)	0,0124 (0,0060-0,0191)	0,0097 (0,0030-0,0168)	0,0085 (0,0030-0,0159)	
Reference	0,0203 (0,0118-0,0261)	0,0196 (0,0114-0,0251)	0,0199 (0,0117-0,0270)	0,0212 (0,0134-0,0278)	0,0201 (0,0115-0,0259)	0,0207 (0,0129-0,0270)	0,0210 (0,0127-0,0271)	0,0180 (0,0086-0,0246)



Slika 11. MDS grafički prikaz svih kombinacija uzoraka i referenci u dvodimenzionalnom prostoru (Os 1 - Dimenzija 1, Os 2 - Dimenzija 2). Rezultat je dobiven na temelju genetske matrice udaljenosti izračunatoj na poravnanju 136.459 aminokiselinskih baza (100 spojenih BUSCO gena). Reference su označene znakovima, a uzorci divokoza bojama.

3. Zajednički set 500 nasumično odabranih BUSCO gena:
    a. Poravnanje od 500 gena sastojalo se od 308.675 aminokiselinskih baza.
    b. Genetska matrica udaljenosti izračunata iz zajedničkog poravnanja (308.675) korištena je za grafički prikaz svih kombinacija u dvodimenzionalnom prostoru, a rezultat je prikazan na Slici 12. U rezultatima MDS analize odnosi uzoraka bili su nešto drugačiji ali vrlo slični onima sa Slike 8. Na temelju poravnanja od 500 BUSCO gena, uzorci divokoze mapiranih na šest sekvenci opet su bili vrlo slični s tim da su se ovoga puta odvojili uzorci mapirani na američku planinsku kozu. Udaljenosti između tih uzoraka i drugih grupiranih uzoraka su bile oko 0,005 što je otprilike oko 1.500 polimorfizama u ukupnom poravnanju od 308.675 aminokiselinskih baza. Također, i u ovom su slučaju svi uzorci bili proporcionalno udaljeni od svojih referenci



Slika 12. MDS grafički prikaz svih kombinacija uzoraka i referenci u dvodimenzionalnom prostoru (Os 1 - Dimenzija 1, Os 2 - Dimenzija 2). Rezultat je dobiven na temelju genetske matrice udaljenosti izračunatoj na poravnanju 308.675 aminokiselinskih baza (500 spojenih BUSCO gena). Reference su označene znakovima, a uzorci divokoza bojama.

4.7 Usporedba novosastavljenih sekvenci divokoza s dostupnim genomskim sekvencama iz Banke gena

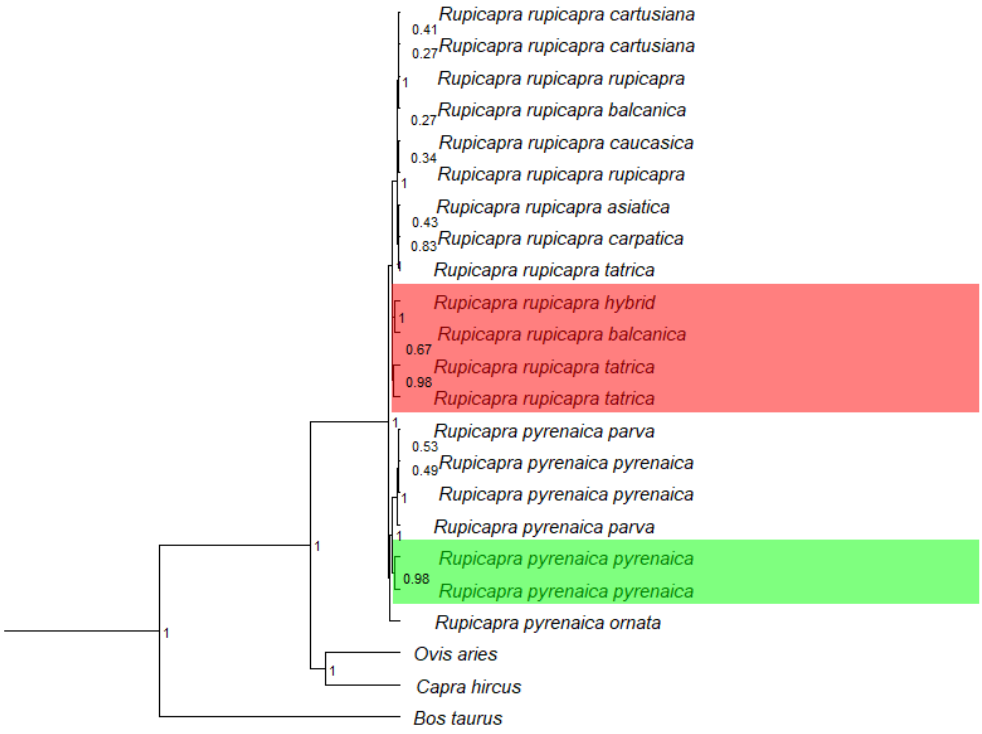
Na temelju ranije prikazanih rezultata, uzorci divokoze mapirani na referentni genom domaće koze izabrani su za provedbu evaluacije i usporedbe dobivenih fragmenata s dostupnim sekvencama divokoze iz Banke gena. Nakon odbacivanja dupliciranih fragmenata te genomskih fragmenata niske kvalitete broj mapiranih fragmenata divokoze na referentni genom domaće koze kretao se od 166.926.451 do 244.675.965. Drugim riječima, udjeli mapiranja svakog uzorka bili su slični i kretali su se od 95 do 98% (Tablica 15.). Procedura pozivanja SNP-ova dala je između 56.000.000 i 57.000.000 sirovih SNP-ova. Nakon filtriranja zadržano je između 42.500.000 i 51.700.000 SNP-ova (Tablica 15).

Tablica 15. Rezultati dobiveni procesima mapiranja, pozivanja varijanti i BUSCO analize uzoraka divokoze mapiranih na referentni genom domaće koze.

Uzorak	Podvrsta	Lokacija	Proporcija mapiranja (%)	Broj sirovih SNP-ova	Broj filtriranih SNP-ova	Broj BUSCO gena (od 13.335) (%)
B532	*R. r. tatrica*	Slovačka	96,64	56.869.372	46.991.911	95,51
B539	*R. r. tatrica*	Slovačka	96,36	56.333.438	42.590.829	95,46
GAMS7	*R. r. balcanica*	Hrvatska	95,47	56.457.896	47.458.499	95,50
GAMS21	*R. r. rupicapra*	Hrvatska	97,25	56.889.057	46.489.008	95,54
GAMS57	*R. r. rup/balc*	Hrvatska	98,18	56.838.238	51.520.107	95,55
GAMS108	*R. p. pyrenaica*	Španjolska	97,41	57.023.898	48.698.661	95,61
GAMS109	*R. p. pyrenaica*	Španjolska	95,40	57.451.530	51.814.829	95,53

Novosastavljeni genomi divokoze su evaluirani u dva koraka. Kvaliteta i kompletnost genoma provjerena je pomoću BUSCO-a, dok je pouzdanost i struktura provjerena usporedbom s dostupnim sekvencama divokoze. BUSCO rezultati bili su visoki za sve uzorke, u rasponu između 95,46 % (12.736) i 95,61 % (12.750) pronađenih gena (Tablica 13.). Za usporedbu strukture novih genoma divokoze, prvo se reproducirala rekonstrukcija filogenetske analize na izvornom poravnanju (duljine 15.382 pb) iz rada Pérez i sur. (2017.). Njihovo inicijalno poravnanje sastojalo se od 14 sekvenci divokoze i jedne sekvence ovce. Potom je provedena ista analiza dodavanjem 7 uzoraka iz ove disertacije te dvije sekvence (goveda i koze) koje su korištene kao uljezi. U zajedničkom poravnanju pronađeno je 178 varijabilnih mjesta među 21 divokoze. Jedan od uzoraka (Gams21) pokazao je veliku varijabilnost s 53 varijabilna

mjesta specifična samo za ovaj uzorak zbog čega je isti isključen iz daljnjih analiza. Nakon dodavanja sekvenci koze i goveda te nakon uklanjanja svih indel pozicija, konačno poravnanje se sastojalo od 14.980 pb za 23 sekvence sa 123 varijabilnih mjesta (unutar divokoza). Filogenetsko stablo dobiveno BEAST analizom prikazano je na Slici 13.



Slika 13. Ukorijenjeno filogenetsko stablo dobiveno Bayesovskom metodom za rod *Rupicapra* dobiveno na poravnanju intronskih sekvenci. Iznad čvorova prikazane su Bayesovske posteriorne vjerojatnosti. Označeni klasteri su uzorci iz ove disertacije: crvena – 4 uzorka sjeverne divokoze; zelena – 2 uzorka južne divokoze.

Šest sekvenci divokoza iz ove disertacije se grupiralo s ostatkom divokoza, uz visoke vrijednosti bootstrapa. Drugim riječima, iako je genom koze korišten kao referenca, uzorci se nisu grupirali s kozom nego su po svojoj genomskoj strukturi bili bliži drugim divokozama. Osim toga, uzorci su se na razini vrste te na razini podvrste grupirali sukladno taksonomiji i očekivanjima. Dva uzorka (označena zelenom bojom) koja su pripadala južnim divokozama grupirala su se s ostalim južnim divokozama, dok su se 4 uzorka sjeverne divokoze (označena crvenom bojom) grupirala s ostalim podvrstama sjeverne divokoze. Međutim, unutar glavnih klastera uzorci su formirali manje podklastere, što može biti posljedica korištenja različitih tehnologija sekvenciranja u izvornom i našem istraživanju.

4.8 Sastavljanje nDNA hibridnom metodom

Budući da je ova metoda znatno kompleksnija u usporedbi s metodom mapiranja, korišten je samo jedan uzorak (Gams57). Za provođenje prvog koraka ove metode korištena su dva alata za sastavljanje genoma *de novo*: Abyss i SPAdess. Oba alata su pokrenula iz koda u kojim su se zadale određene vrijednosti parametara s ciljem ubrzanja i optimizacije cijelog procesa. Za svaki alat pokrenuto je nekoliko varijacija koda a sve su rezultirale naglim prekidom procesa zbog nedostatka računalne memorije, uobičajeno peti ili šesti dan nakon pokretanja. Čak i uz dodavanje parametara kojim se alatu dopušta korištenje duplo većeg prostora nije bilo dovoljno za provedbu ovog procesa.

5 RASPRAVA

5.1 Usporedba metoda za sastavljanje i anotaciju mtDNA

MtDNA jedan je od najčešće korištenih alata za proučavanje evolucije i filogenije na što ukazuje veliki broj deponiranih mtDNA sekvenci u Banci gena (38.500 sekvenci sisavaca, pristupljeno 27.4.2022) (Gordon i sur., 1998; Dierckxsens i sur., 2016; Song i sur., 2016). Kako je već spomenuto, proces sastavljanja mtDNA jednostavniji je ukoliko se koristi referenca iste vrste. Kada referenca nije dostupna, koristi se referenca srodne vrste. Međutim treba imati na umu da prilikom korištenja srodne reference, rezultat može biti pristran ukoliko je ta referenca genetski udaljenija od vrste koja se proučava. Ne koriste se sve mtDNA sekvence kao reference, pa tako npr. u Banci gena od 25.041 dostupnih mtDNA čovjeka, samo se jedna koristi kao referenca. Međutim, za većinu vrsta samo je mali broj mtDNA sekvenci dostupan i njihove reference nisu uvijek visoke kvalitete niti imaju dostupnu anotaciju (Prada i Boore, 2019). Štoviše, mogu sadržavati pogreške nastale procesom sastavljanja ili anotacije (Hassanin i sur., 2010). Na ovaj problem ukazalo je nekoliko autora (Hassanin i sur., 2010; Smith, 2016; Prada i Boore, 2019) tvrdeći da veliki broj dostupnih mtDNA sekvenci iz Banke gena sadrže pogreške u anotaciji. Osim toga, čest je slučaj da su takve sekvence dobivene u sklopu istraživanja u kojima je proces rekonstrukcije mtDNA oskudno opisan (Hu i sur., 2015; Hu i sur., 2016; Hill i sur., 2017; Mao i sur., 2017; Pramod i sur. 2018; Davenport i sur., 2018). Na primjer, Hu i sur. (2014) su u svom radu sastavili mtDNA divlje svinje (*Sus celebensis*) u kojem su definirali strukturu nukleotida te START i STOP pozicije svih gena. Međutim, naveli su da je za sastavljanje korišten samo CLC Genomic Workbench alat bez definiranih parametara, dok alat za anotaciju nisu spomenuli. Slično je i u radu o mtDNA europske divlje svinje (*Sus scrofa scrofa*) od istog autora (Hu i sur., 2015). Pramod i sur. (2018) u svom radu o mtDNA sekvenci indijskog goveda (*Bos indicus*) također su izostavili informacije o korištenim alatima za rekonstrukciju kao i autori rada o sastavljenoj sekvenci mtDNA američkog muflona (*Ovis canadensis*) (Davenport i sur., 2018). U radu o argali ovci (*Ovis ammon darwini*) (Mao i sur. (2017) opis metoda je detaljniji, ali i dalje nedostaju konkretne informacije o korištenim parametrima u alatima za sastavljanje mtDNA. Glavni problem je što se dobivene sekvence iz takvih radova deponiraju u Banku gena i često predstavljaju referentnu sekvencu za svoju vrstu. Međutim, postavlja se pitanje o kvaliteti tih sekvenci kao i o posljedicama njihovog korištenja u daljnim analizama ukoliko sadrže pogreške (Hassanin i sur., 2010). Prada i Boore (2019) su u svom radu istraživali strukture i anotacije 304 kompletnih mtDNA sekvenci (koje pokrivaju 29 taksonomskih

redova) dostupnih u Banci gena. Uspoređivanjem nukleotidnih regija, autori su detektirali promjene za koje se uspostavilo da su pogreške koje su nastale u procesu sastavljanja. Slično istraživanje proveli su Hassanin i sur. (2010) u kojem su sastavili mtDNA domaće koze i usporedili je s referentom sekvencom domaće koze (koju su prethodno objavili Parma i sur. (2003)) s još četiri dostupne sekvence. Usporedbom sekvenci zaključili su da su Parma i sur. (2003) sekvencirali svega 44.5 % ukupne mtDNA koze (koja je tada predstavljala referentnu sekvencu) dok su preostale regije pripadale jezgrinom genomu. Sličnom se problematikom bavio i Smith (2016) te je u svom radu zaključio kako su istraživanja mtDNA postala ponavljajuća i da takva istraživanja općenito pate od nedostatka testiranja hipoteza. Zbog navedene problematike, neki autori isključivo preporučuju korištenje *de novo* metode s obzirom na dostupnost velikog broja *de novo* alata (Dierckxsens i sur., 2016).

Iz tog razloga su u ovoj disertaciji korištene obje metoda za sastavljanje mtDNA. Na ovaj način se lakše moglo utvrditi postoje li greške između dvije sekvence dobivene iz istog uzorka korištenjem različitih metoda. Kako je gore već navedeno, prije provođenja metode mapiranja, bilo je potrebno ispitati dostupne referentne sekvence koje će se koristiti u analizama. U ovoj su se disertaciji za potrebe sastavljanja mtDNA metodom mapiranja koristile dvije referentne sekvence divokoza: sjeverna i južna, obje sastavljene u radu Hassanin i sur. (2009). Iako su obje reference zadovoljavajuće kvalitete, referenca južne divokoze nema anotaciju što u ovom slučaju nije predstavljalo problem jer se koristila anotacija sjeverne divokoze. Nedostatak anotacije u Banci gena uobičajena je pojava čime se otežava sam proces validacije i usporedbe rezultata. Međutim, u takvim slučajevima korištenje referentnih sekvenci bliskih srodnika i njihove anotacije može pružiti alternativno rješenje. Za provođenje metode *de novo* pomoću NOVOPlasty alata kao seed sekvence korištene su genske regije CYTB gena koje su također pripadale istim referencama divokoza.

Nakon provođenja dviju metoda za sastavljanje te uspoređivanjem dobivenih sekvenci mtDNA jasno se moglo zaključiti da su obje metode za sastavljanje pogodne u procesima rekonstrukcije mtDNA. Za pet korištenih uzoraka, ove dvije metode rezultirale su identičnim sekvencama. Razlike između sekvenci zabilježene su samo kod uzoraka B532 (dvije varijante) te GAMS108 (268 varijanti) što se može definirati kao posljedica kontaminacije tog uzorka ili kao posljedica manje koncentracije mtDNA u uzorku (Al-Nakeeb i sur., 2017). *De novo* metoda provedena pomoću NOVOPlasty alata, osim što je brža, rezultirala je sastavljenim sekvencama iz uzoraka koji nisu prošli kontrolu kvalitete. Ovi rezultati ukazuju da se u provedbi ovih metoda ne treba uvijek ograničavati lošom kvalitetom uzoraka.

Rezultati dobiveni iz tih uzoraka uzeti su s oprezom, a dodatnim provjerama i usporedbom s ostalim sekvencama se utvrdilo da su zadovoljavajuće kvalitete i da se mogu koristiti u daljnjim analizama (što je u slučaju malog broja uzoraka od velike važnosti). Kad se u obzir uzmu vrijeme, pohrana i brzina provođenja ovih metoda, *de novo* metoda je opet u prednosti. Mapiranje se sastoji od nekoliko koraka u kojima se lako dođe do greške prilikom provedbe svakog koraka, dok se pokretanje NOVOPlasty programa svodi na pozivanje skripte u kojoj se definiraju svi potrebni parametri. Cijeli postupak rekonstrukcije sekvence mtDNA metodom mapiranja traje u prosjeku nekoliko dana s tim da se u prvih nekoliko koraka stvaraju velike datoteke koje zauzimaju radni prostor (npr. prosječna veličina datoteke SAM 100 GB i BAM 20 GB) iz kojih se tek u posljednjim koracimu izoliraju mtDNA sekvence. S druge strane, prosječni proces sastavljanja pomoću NOVOPlasty alata traje oko 45 minuta dok su rezultati prezentirani u tri datoteke zanemarivih veličina. Nakon detaljne usporedbe svih sekvenci, sekvence mtDNA dobivene *de novo* metodom korištene su u daljnjim analizama.

Rezultati anotacije dobiveni su korištenjem MITOS i GeSEQ anotatora. I u ovom procesu, oba anotatora su dali vrlo slične rezultate (Tablica 8). Varijacije u START i STOP kodonima su očekivane s obzirom da anotatori koriste drugačije algoritme. Neovisno o rezultatima, preporuča se provođenje validacije i usporedbe rezultata anotacije s anotacijom dostupne reference iz Banke gena (Tešija i Safner, 2020). Na ovaj se način uspoređuju START i STOP kodoni te se mogu detektirati prisutne greške u novonastalim sekvencama (Bernt i sur., 2013).

5.2 Korištenje mtDNA u filogenetskim analizama papkara

Rod *Cetartiodactyla*, kojem pripada porodica *Caprini*, jedna je od najraznovrsnijih skupina među postojećim vrstama a njihova predložena taksonomija bazirana je većinom na analizama različitih dijelova mtDNA. 90-ih godina rod *Cetartiodactyla* nije postojao nego se zvao *Artiodactyla* i uključivao je je samo papkare. Dijelovi mtDNA koji su se najčešće koristili u analizama ovog roda su dva rRNA gena (16S i 12S) (Miyamoto i sur., 1989; Allard i sur., 1992; Gatesy i sur., 1992), cijeli gen CYTB ili samo njegovi dijelovi (Irwin i sur., 1991; Chikuni i sur., 1995; Groves i Shields, 1996; Groves i Shields, 1997; Hassanin i sur., 1998; Randi i sur, 1998; Hassanin i Douzery, 1999) te dijelovi kontrolne regije (Douzery i Randi, 1997; Polziehn i Strobeck, 2002). Molekularne analize temeljenje na jednom genu uspješno su smjestile većinu vrsta u taksnomske redove i plemena, ali odnosi pojedinih grupa i vrsta nisu se mogli uspostaviti (Gentry, 1990).

Samo jedan od mnogih primjera su vrste saiga (*Saigra tatarica*) i tibetanska antilopa (*Panthalops hodgsoni*) za koje se smatralo da pripadaju porodici *Caprinae* i o čijoj se filogentskoj pozadini raspravljalo godinama. Grubb (1993) i McKenna i Bell (1997) predložili su da bi ove vrste trebale pripadati prodici *Bovidae*, dok je Gentry (1992) smatrao da su dio porodice *Caprinae*. Potom su Hassanin i sur.(1998) proveli filogenetsku analizu ovih dviju vrsta i ostalih 18 vrsta *Caprinae* koristeći kompletne regije CYTB gena. Njihov rezultat dodatno je zakomplicirao taksonomiju roda *Caprinae* predloživši da bi se vrste saiga i tibetanska antilopa trebale isključiti iz roda *Caprini*. Osim toga, njihovim se istraživanjem nisu mogli definirati odnosi pojedinih vrsta *Caprinae* kao što su divokoza, američka planinska koza i grivasti skakač. Uz navedeno, došli su do zaključka da su rezultati analiza dobivenih na dijelu jednog gena limitirajuće i da bi se to moglo izbjeći dodavanjem većeg broja uzoraka i korištenjem većeg broja mitohondrijskih gena.

Rod *Caprinae* nije jedini čija se rezolucija godinama nastojala riješiti. Rasprave su se vodile i oko roda *Cervidae* (jeleni) i *Bovidae* (šupljorošci) (Irwin i sur., 1991; Chikuni i sur.,1995; Grubb, 1993; Randi i sur., 1998; Kuwayama i Ozawa,2000), a povezivanje roda *Cetacea* (kitovi) skupa s ostalim papkarima i tvoreći na taj način novi rod *Cetartiodactyla*, dodatno je zakompliciralo cijelu taksonomiju (Montgelard i sur., 1997; Ursing i Arnason, 1998; Nikaido i sur., 1999; Ursing i sur., 2000). Međutim, već krajem 90-ih godina bilo je jasno da je usporedba i korištenje većih dijelova mtDNA sekvenci puno informativnije nego kod korištenja pojedinih mtDNA gena, prvenstveno jer veći broj polimorfnih mjesta može ukazati na značajne evolucijske promjene u mtDNA (Ingman i sur., 2000; Ursing i sur., 2000; Boore i sur., 2005; Miller i sur., 2012; Kim i sur., 2014). Ursing i Arson (1998) te Ursing i sur. (2000) među prvima su prikazali prednosti korištenja kompletnih mtDNA i to korištenjem 12 PCG-a (ND6 se često isključuje iz analiza budući da se jedini nalazi na lakom lancu mtDNA) u kombinaciji s CR regijom u filogenetskim istraživanjima porodice *Cetartiodactyla*. Razlog zbog kojeg se preporuča korištenje samo PCG-a je visoka razina konzerviranosti i manja sklonost grešakama budući da su takve regije jako slične za sve sisavce. S druge strane, ovakvim se pristupom smanjuje broj pogrešnih polimorfnih mjesta (koja su često pristuna u kontrolnoj regiji) što utječe na rezoluciju filogenetskih analiza (Gibson i sur., 2005; Mereu i sur., 2008; Miller i sur., 2012; Douglas i sur., 2011; Jiang i sur., 2013; Świsłocka i sur., 2020). Zhou i sur. (2019) su u svom radu koristili sekvence koje su se sastojale od 13 PCG-a i 2 rRNA gena ali bez korištenja START i STOP kodona. S druge strane, Hassanin i sur. (2009), Matosiuk i sur. (2014) te Mohandesan i sur. (2017) su napravili zasebne setove podataka za

svaku gensku obitelj (ND, ATP, COI, CYTB) kako bi procjenili odnose supstitucija. Hassanin i sur. (2009) su dokazali da su najveće razlike pronađene u ATP genskoj obitelji dok su preostale tri bile više konzervirane. S druge strane, Matosiuk i sur. (2014) su najveći broj varijanti pronašli u genskoj obitelji ND dok su Mohandesan i sur. (2017) najveći broj varijanti pronašli u CYTB genu.

U isto vrijeme provedena su istraživanja kojima su autori htjeli dokazati da segmenti mtDNA mogu dati jednako dobre rezultate kao i kompletne mtDNA sekvence. Zurano i sur. (2019) u svom su radu koristili dva različita pristupa za testiranje pouzdanosti korištenja parcijalnih i kompletnih sekvenci mtDNA. Filogenetske analize proveli su na dva različita seta: prvi se sastojao od kompletnih mtDNA sekvenci od 225 vrsta, dok je drugi set sadržavao 93 vrste koje su imali bar jedan dostupan gen. Analize provedene na oba seta rezultirale su vrlo sličnim topologijama s malim razlikama u procjenama vremena divergencija pojedinih vrsta. Ovim su istraživanjem potvrdili da manji setovi mtDNA mogu dati jako slične rezultate kao i kompletne mtDNA sekvence. Do sličnog zaključka došli su Jiang i sur. (2013) koji su koristili četiri seta podataka s ciljem proučavanja filogenetskih odnosa vrsta divljih ovaca (*O. musimon*, *O. vignei*, *O. ammon hodgsoni*). Korištena četiri seta podataka uključivala su: 1) kompletne sekvence mtDNA; 2) spojene sekvence manjih regija (kontrola regija i PCG); 3) CYTB regija; 4) kontrolna regija. Zaključili su da su sve četiri analize prikazale identičnu topologiju. Slične rezultate dobili su Naseem i sur. (2020) gdje su zaključili da se COI gen može efektivno koristiti u procesima identifikacije vrsta.

U sklopu ove disertacije provedena je rekonstrukcija filogenije za rodove *Caprinae* i *Rupicapra*. Filogenetska stabla (maksimalna vjerodostojnost i Bayesovska metoda) *Caprinae* rezultiralo je identičnom topologijom kao i stablo iz Hassanin i sur. (2009) te Pérez i sur., (2014). Rezultati filogenetskih analiza iz oba rada ali i u ovoj disertaciji dobiveni su korištenjem kompletnih mtDNA sekvenci. Na stablu dobivenom o sklopu ove disertacije jasno se može definirati da su rodovi *Bovini* i *Caprini* monofiletični te je unutar roda *Ovicaprini* jasno podržano odvajanje *Caprini* i *Ovibovini* te da su unutar roda *Caprinae*, *Capra* i *Hemitragus* formirali tzv. goat-like klaster. S druge strane, usporedbom rezultata filogenetske analize za rod *Rupicapra* s prethodnim istraživanjima, može se zaključiti da će, bez obzira na korišteni segment mtDNA, topologija biti ista. Međutim, korištenje većeg broja uzoraka s jednakim brojem podvrsta vjerojatno bi dalo bolji uvid u evoluciju mtDNA sekvence.

Zaključno, bez obzira što se za većinu vrsta može dobiti vrlo slična toplogija korištenjem pojedinačnih gena i kompletnih mtDNA sekvenci, bitno je imati na umu da analize provedene na pojedinačnim genima neće uvijek biti dovoljne za određivanje filogenetskih odnosa nekih vrsta papkara. Glavni razlog je ograničena raznolikost mtDNA kod tih vrsta i u tom se slučaju preporuča korištenje većih segmenata mtDNA (Arif i sur., 2012; Świsłocka i sur., 2020; Corlatti i sur., 2022b).

5.3 Filogenetske analize mtDNA roda Rupicapra

Filogenetskim se analizama nastoje opisati odnosi vrsta i podvrsta divokoza ali i njihovi odnosi s drugim vrstama roda *Caprinae*. Corlatti i sur. (2011) u preglednom su radu naveli nekoliko hipoteza o taksonomiji divokoza postavljenih tijekom prve polovice 20. stoljeća ali i povijesni pregled molekularnih istraživanja provedenih na divokozi: na temelju morfoloških podataka divokoza se u nekoliko navrata definirala kao jedna vrsta (Lydekker, 1913; Couturier, 1938), dvije (Neumann,1899) ili čak tri (Camerano, 1914) odvojene vrste. Međutim, Dolan (1963) je ponovno potvrdio hipotezu u kojoj se divokoza smatra kao jedna vrsta sve dok se nisu pojavili nova istraživanja koja su na temelju morfoloških i bihevioralnih karakteristika objasnile podjelu divokoze na dvije vrste (Lovari i Scala, 1980; Lovari, 1985). Potom su za proučavanje odnosa divokoza korišteni molekularni markeri: alozimi (Nascetti i sur., 1985), minisateliti (Pérez i sur., 1996), geni glavnog sustava tkivne podudarnosti (MHC) (Schaschl i sur., 2012; Alvarez-Busto i sur., 2007), dijelovi mitohondrijske DNA (Hammer i sur., 1995). Prikaz filogenetskih odnosa između vrsta i podvrsta najviše ovise o korištenim molekularnim markerima budući da različiti geni mogu imati drugačije modele transmisije te različite stope povijesne evolucije (Tajima, 1983). Osim toga, hibridizacija ima drugačiju stopu utjecaja na markere što može uzrokovati različite rezultate.

Rodríguez i sur. (2009; 2010) su sekvencirali 1.700 parova baza manjih dijelova mtDNA i mikrosatelita kako bi detaljno proučili odnose između vrsta. Rezultati temeljeni na podacima mtDNA svrstali su divokozu u tri majčinske linije: zapadna, centralna i istočna. Slični rezultati dobiveni su analizama mikrosatelita. Međutim, analiza mikrosatelita pokazala je veću podudarnost s morfološkom klasifikacijom nego s rezultatima dobivenim analizom mtDNA. Autori su rezultate temeljene na podacima mtDNA objasnili kao posljedicu hibridizacije između ove dvije vrste koja se dogodila tijekom migracija sredinom Pleistocena. Međutim, utjecaj hibridizacije nije pronađen korištenjem mikrosatelita. Materinsko nasljeđivanje mtDNA može utjecati na rezultate filogeografske strukture zbog toga što se ženke divokoze češće

zadržavaju oko područja gdje su okoćene, što može utjecati na prostornu razdiobu haplotipova mtDNA, pogotovo kod malih i izoliranih populacija (Corlatti i sur., 2011). S druge strane, Pérez i sur. (2011) proučavanjem Y-kromosoma divokoze zaključili su kako očinska linija divokoza ne potječe iz Azije, nego iz Mediterana, točnije s Pirinejskog poluotoka. U svom su istraživanju koristili 87 uzoraka (40 od *R. pyrenaica* i 47 od *R. rupicapra*) koji su prikupljeni s gotovo cijelog područja u kojem divokoze obitavaju. Ovo je bio još jedan dokaz kako rezultati dobiveni korištenjem molekularnih markera variraju te zbog toga mogu imati utjecaj na donošenje različitih zaključaka. Crestanello i sur. (2009) su također sekvencirali 1500 parova baza mtDNA. Rezultati njihove analize pokazali su da se jedan haplotip iz područja Alpi grupirao s haplotipovima s Pirinejskog poluotoka. Autori su ovaj fenomen prepisali čovjekovoj intervenciji gdje su divokoze, ne uzimajući u obzir genetsku strukturu, vrlo vjerojatno translocirane s područja Pirinejskog poluotoka u Alpe prije 150 godina kao posljedica razmjene lovne divljači. Međutim, ova hipoteza nije objasnila kako je moguće da se pirinejski haplotip brzo raspodijelio u tako velikoj populaciji u samo 150 godina (Corlatii i sur., 2011) i kako je moguće da tako velik stupanj hibridizacije nije ostavio traga na mikrosatelitima. Populacije iz Apenina i masiva Chartreuse su integrirane i jako slične bez obzira na njihov taksonomski status. *R. p. ornata* ima izuzetno nisku razinu varijabilnosti u usporedbi s ostalim podvrstama. Glavni razlog za to je dugotrajna izoliranost i fragmentiranosti manjih. S druge strane, populacije (*R. r. cartusiana*) iz Chartreuse nose mitohondrijski haplotip centralne majčinske genetske linije (Rodriguez i sur. 2010), dok prema analizama mikrosatelita pripadaju istočnoj liniji. Međutim, prema morfometrijskim podacima *R. r. cartusiana* dijeli sličnosti s obje vrste, a pretpostavlja se da je uzrok tome hibridizacija koja se vjerojatno dogodila krajem posljednjeg ledenog doba u zapadnim Alpama kada su *R. rupicapra* i *R. pyrenaica* bile u kontaktu (Lovari i Scala, 1980). Slično istraživanje o hibridizaciji provedeno je na populacijama divokoza u Hrvatskoj (Šprem i Bužan, 2016). Autori su korištenjem mikrosatelita i dijelova mtDNA ustanovili postojanje dvaju haplotipova: alpski i balkanski. Uz to, dokazano je da su translokacije jedinki iz drugih geografskih područja utjecale na stvaranje kontaktne zone između ove dvije populacije što može loše utjecati na očuvanje podvrste.

Do sada su u samo dva rada (Hassanin i sur., 2009; Pérez i sur., 2014) kompletne sekvence mtDNA divokoza korištene za rekonstrukciju filogenije. Sve sekvence dobivene su korištenjem 23 početnice objavljene u Hassanin i sur. (2009) a dobivenih pet sekvenci (iz oba rada) jedine su kompletne mtDNA sekvence divokoze dostupne u Banci gena. U sklopu ove

disertacije je iz kompletnih NGS genomskih podataka rekonstruirano 10 mtDNA sekvenci divokoza koje su deponirane u Banku gena (pristupni brojevi: MW588895.1-MW588903.1). Korištenjem tih 10 novosastavljenih mitogenoma u kombinaciji s pet sekvenci iz Banke gena, provedena je rekonstrukcija filogenije a dobiveni su rezultati prikazali iste odnose divokoza podijelivši ih u tri klastera kao i u prethodnim istraživanjima (Rodríguez i sur., 2009; 2010; Pérez i sur., 2011; 2014). I u ovoj analizi, centralni klaster tvore *R. r. cartusiana* i *R. p. ornata*, podvrste koje pripadaju i sjevernoj i južnoj divokozi, a taj je klaster prema morfološkim i molekularnim podacima, puno bliži zapadnom klasteru (Lovari i sur., 1980; Pérez i sur., 2014) što je još jednom potvrdilo nepodudarnost sistematike roda *Rupicapra*. Nadalje, na razini vrsta, zanimljivo je da je detektirana veća varijabilnost pronađena kod južnih divokoza koje imaju duplo manji broj uzorka što otvara dodatno pitanje o povijesti populacija obje vrste. U tom slučaju, od velike bi pomoći bila usporedba modernih i drevnih uzoraka koja bi potencijalno moga ukazati na složenu evolucijske povijesti divokoza što bi s druge strane zahtjevalo doprinose iz drugih područja kao što su paleontologija, morfologija i ponašanje (Corlatti i sur., 2022b).

5.4 Usporedba novosastavljenih nDNA sekvenci dobivenih metodom mapiranja

Mapiranje je najučestalija metoda u procesima sastavljanja jezgrinih genoma te u njihovim komparativnim analizama. Noviteti u tehnologijama sekvenciranja ali i u razvoju bioinformatičkih softvera uglavnom su potaknuti istraživanjima ljudskog genoma te biomedicinskim istraživanjima. Korištenje tih noviteta nastoji se primijeniti i u drugim znanstvenim disciplinama, prvenstveno u poljima ekologije, konzervacijske i populacijske genetike. Međutim, kako je već spomenuto, primjena novih tehnologija na specifične genomske podatke i dobivanje točnih informacija iz njih može predstavljati izazov budući da većina novih metoda nije primjenjiva na nemodelne vrste (McCormack i sur., 2013; McMahon i sur., 2014). Ovo je posebno izazovno kod istraživačkih grupa koje u većini slučajeva nisu nužno specijalizirane za bioinformatiku, a koriste DNA sekvence kao glavni alat u svojim istraživanjima i često se oslanjaju na preporučene protokole za sastavljanje genoma temeljene na znanstvenoj literaturi (Kalbfleisch i Heaton, 2014; Galla i sur., 2018). U takvim istraživanjima često je glavni cilj izolirati fragmente genoma i proučiti njihovu biološki pozadinu (npr. identificirati gene ključne za adaptaciju ili gene odgovorne za evoluciju i očuvanje neke vrste). Bourgeois i Warren (2021) su u svom preglednom radu ukazali na ovaj problem. Zbog nedostatka komunikacije između znanstvenih polja te nedostataka softvera

prilagođenih korisniku, postaje sve teže svim potencijalnim korisnicima pratiti novitete u tim područjima, pogotovo kada je u pitanju odabir najprikladnije metode za njihove potrebe (i podatke). Uz sve navedeno, zaključili su da prilagodba novih metoda na genomske podatke predstavlja jedan od najvećih izazova u ovom polju. Ovisno o cilju istraživanja u genomici, glavni ulazni podaci su genomski fragmenti DNA iz kojih se nastoji dobiti točna informacija. Različite varijacije metoda mapiranja, iako jako korisne, neće uvijek dati konkretno rješenje ili odgovoriti na specifično pitanje, pogotovo zbog činjenice da čak i genomi blisko srodnih vrsta mogu međusobno jako varirati (Ekblom i Wolf, 2014; Fuentes-Pardo i Ruzzante, 2017; Valiante-Mullor i sur., 2021). S obzirom da su genomi sisavaca kompleksni, jako je malo dostupnih radova o testiranju referentnih genoma, pogotovo kod nemodelnih vrsta. Uzevši u obzir sve navedeno, u ovoj se disertaciji provela metoda mapiranja upravo s ciljem testiranja i usporedbe nekolicine srodnih referentnih genoma. Budući da referentni genom divokoze nije dostupan, genomski uzorci divokoze su bili kombinirani s nekoliko različitih referenci. Iako korištenje srodnih sekvenci može uzrokovati pristranost rezultata, i dalje je moguće dobiti točnu informaciju o vrsti koja se istražuje (Sousa i Hey, 2013; Gopalakrishnan i sur., 2017). Rezultati ove disertacije još su jednom potvrdili da genom srodne vrste može poslužiti kao dobra referenca kod proučavanja nemodelnih vrsta. Isti zaključci su donešeni u nekoliko istraživanja: uzorci ovce mapirani na genome goveda i ovce (Kalbfleisch i Heaton, 2014); uzorci vuka mapirani na genome psa (pasmine bokser) i vuka (Gopalakrishnan i sur., 2017); uzorci beluge mapirani na genome nekoliko vrsta dupina i kitova (Prasad i sur., 2021). Svaki mapirani uzorak divokoze na različite reference rezultirao je visokom proporcijom mapiranih fragmenata (95-98 %). Ovaj rezultat je bio očekivan s obzirom da rezultati mapiranja u velikoj mjeri ovise o genetskoj udaljenosti između vrsta (Bohling, 2020). Međutim, zanimljivo je da je najveću proporciju mapiranih fragmenata imala domaća ovca koja je, od svih korištenih referenci, genetski najudaljenija od divokoze. Razlog tome može biti sama kvaliteta genoma domaće ovce koja, skupa s domaćom kozom, ima najkvalitetniji i najpotpuniji genom u Banci gena a da ne pripada modelnim vrstama (Kalbfleisch i Heaton, 2014). Osim toga, moguće je da su u genomu ovce prisutne genomske regije (koje nedostaju u drugim genomima) na koje su se mapirali fragmenti divokoze što je rezultiralo većom proporcijom.

Varijabilnost u broju detektriranih SNP-ova također ovisi o genetskoj udaljenosti uzoraka i referentnog genoma (Farrer i sur., 2013; Kalbfleisch i Heaton, 2014). Broj nefiltriranih SNP-ova u kombinacijama uzoraka divokoze i referenci kretao se između 28.000.000 i 60.000.000 dok je nakon filtriranja zadržano između 13.740.102 i 51.814.829 SNP-ova. Kalbfleisch i

Heaton (2014) su u svom istraživanju potvrdili hipotezu da je moguće koristit visokokvalitetan genom srodne vrste u procesima mapiranja. U svom su radu mapirali genomske podatke ovce na dvije različite reference visoke kvalitete, ovcu i govedo. Proporcija mapiranih fragmenata ovce na genom goveda iznosila oko 76 %, dok je broj detektiranih SNP-ova bio je oko 83 milijuna SNP-ova. Na temelju ovih rezultata, autori su zaključili da bi taj broj bilo dovoljan ako bi se ti SNP-ovi dalje koristili u komparativnoj genomici ili studijama povezanih s bolestima. U radu Ghanatsaman i sur. (2020) detektirano je između 7 i 10 milijuna SNP-ova dobivenih mapiranjem tri uzorka psa i tri uzorka vuka na referentni genom psa. Galla i sur. (2018) su u svom radu dokazali da srodni genom korišten kao referenca može rezultirati zadovoljavajućim brojem SNP-ova na temelju kojih se može procijeniti raznolikost ugroženih vrsta. U svom su radu proučavali genom ptice, vrste *Himantopus novaezelandiae* čije su genomske podatke mapirali na četiri reference srodnih vrsta. Na temelju rezultata, zaključili su da ovakav pristup ima veliku prednost u projektima koji korištenjem genomskih podataka i s ograničenim resursima mogu brže djelovati po pitanju očuvanja kritično ugroženih vrsta. Kod većine projekata koji koriste metodu mapiranja, velika se pažnja pridodaje koraku detekcije SNP-ova, prvenstveno zbog toga što se filtrirani SNP-ovi koriste kao glavni ulazni podaci za daljnje analize. To često podrazumijeva provođenje nekoliko koraka u filtriranju SNP-ova (indetifikacija indels-a, identifikacija homozigotnih i heterozigotnih mjesta, engl. *base recalibration*, engl. *indes realignment*, engl. *indels identification*). U ovoj disertaciji SNP-ovi su se koristili za stvaranje konsenzusnih sekvenci koje su se potom koristile kao ulazni podaci. Iz tog razloga, provedeno je filtriranje SNP-ova koristeći tri glavna kriterija: kvaliteta pronađenih SNP-ova (QUAL>15), dubina sekvenciranja (DP>5), kvaliteta mapiranja (MQ>20). Povećanjem vrijednosti ovih kriterija, broj SNP-ova bi bio puno manji zbog manje pokrivenosti uzoraka čime bi se mogle zanemariti (izgubiti) stvarne varijante. S druge strane, smanjenjem tih vrijednosti, broj SNP-ova bi bio veći jer, npr. vrijednost DP>3 značila bi da su tri fragmenta dovoljna da se očita SNP koji bi bio upitne kvalitete. Slični kriteriji za detekciju SNP-ova kao i pozivanje konsenzusnih sekvenci korišteni su u radu Valiante-Mullor i sur. (2021) u kojem su autori testirali utjecaj bakterijskih referentnih genoma na rezultate mapiranja. Od 565 dostupnih genoma iz Banke gena i inicijalnim poravnanjem, zadržali su 28 genoma koje su koristili kao reference. Potom su za svaku vrstu odabrali 20 uzoraka koje su mapirali na svaki od pet genoma. Broj detektiranih SNP-ova u velikoj mjeri je ovisio o kombinaciji uzoraka i referenci ali svi SNP-ovi su bili filtrirani prema nešto strožim kriterijima u

usporedi s ovima iz disertacije (QUAL>40, DP>10, MQ>30) što je logično budući da su imali kvalitetnije genomske podatke.

Dobivene konsenzusne sekvence potom su se koristile u BUSCO analizama. Kako je već opisano, BUSCO alat osmišljen je s ciljem procjene kvalitete novosastavljenih genoma, a novom se verzijom proširila njegova primjena u filogeniji, većinom kod insekata (Zhang i sur., 2019; Dias i sur., 2020; Sun i sur., 2020; Wang i sur., 2021) te kvasaca (Shen i sur., 2016a; Shen i sur., 2020). BUSCO analize procijenile su vrlo visoku kvalitetu genoma svih konsenzusnih sekvenci (između 88 i 95 %) osim kod kombinacija uzoraka s baralom (59 %) i alpskim kozorogom (50 %). Ovi rezultati su očekivani s obzirom da su BUSCO vrijednosti njihovih referentnih genoma vrlo slične. Drugim riječima, geni koji su odsutni u referentnom genomu, također će biti odsutni u novosastavljnim sekvencama. BUSCO vrijednosti u ovoj disertaciji bile su slične BUSCO vrijednostima u radovima u kojima su sastavljene reference: BUSCO vrijednost za vrstu *Ammotragus lervia* u ovoj disertaciji bila je oko 92,6 % dok je u radu Chen i sur. (2019) bila 93 %; BUSCO vrijednost za *Oreamnos americanus* bila je oko 86,7 % dok je u radu Martchenko i sur. (2020) bila nešto veća (93 %); BUSCO vrijednost za *Ovis aries* bila je oko 93 %, ista kao u radu Davenport i sur. (2022); BUSCO vrijednost za vrstu *Capra hircus* bila je 93,8 %, isto kao i u radu Li i sur. (2021). Za vrste *Capra sibirica* i *Capra aegagrus* nema dostupnih rezultata iz BUSCO analize. BUSCO vrijednosti za vrste *Capra ibex* (50,1 %) i *Pseudois nayaur* (59,8 %) nisu se poklapale s vrijednostima dobivenih u radu Chen i sur. (2019) u kojem su ovi genomi sastavljeni. Osim toga, prilikom inicijalnog poravnanja kraćih BUSCO sekvenci u ovoj disertaciji, primijećeno je da su sekvence dobivene kombinacijom uzoraka divokoze te alpskog (*C. ibex*) i sibiriskog kozora (*C. sibirica*) potpuno identične. Nakon pretraživanja dodatnih informacija o ovim genomima, uočene su nepravilnosti u radu (Chen i sur., 2019) u kojem su autori obje vrste definirali kao jednu dok su u Banci gena dostupne reference za obje vrste. Reprezentativni genom alpskog kozoroga (IBX; pristupni broj: GCA_006410555.1) sastavili su Chen i sur. (2019) (s još 43 genoma drugih vrsta preživača). Reprezentativni genom sibirskog kozoroga (ASM318261v2; pristupni broj: GCA_003182615.2) dostupan je u Banci gena, bez referentnog rada u kojem je opisana metoda, ali s informacijom da ga je u Banku gena deponirala ista institucija kao i alpskog kozoroga. U Banci gena, pod poljem BioProject, Sequence Read Archive (SRA) nalaze se informacije direktno vezane za korišteni uzorak. Pregledom tih informacija za sibirskog kozoroga uočeno je da uzorci sadrže informacije u kojima se spominju obje vrste kozoroga što je nelogično budući da bi sve informacije o nekom uzorku trebale pripadati isključivo

jednoj vrsti. Iz tog razloga je alpski kozorog uklonjen iz daljnih analiza budući da je sama referenca loše kvalitete. Ovo je još jedan dokaz koji ukazuje da sekvence iz Banke gena mogu sadržavati netočne i nepotpune informacije.

Od 5.739 zajedničkih BUSCO gena pronađenih u svim kombinacijama uzoraka i referenci, nasumično odabrani geni su se koristili u tri podatkovna seta: 10 pojedinačnih gena, 100 gena i 500 gena. Očekivano je da odnosi između istih uzoraka mapiranih na različite reference neće biti identični budući da su za analize odabrani različiti podatkovni setovi. Za prvi (10 gena) i drugi (10x10 gena) set nasumično odabranih gena zasebno je izračunato deset matrica udaljenosti iz kojih su se napravila dvije zajedničke matrice koje su sadržavale prosječne vrijednosti i raspone (za svih deset matrica). Iz zajedničke matrice udaljenosti prvog seta moglo se zaključiti da odnosi između uzoraka nisu ujednačeni. U svim kombinacijama, minimalne vrijednosti bile su jednake 0, što proizlazi iz činjenice da su sekvence barem jednog gena bile identične. Međutim, pronađene nelogičnosti u odnosima između uzoraka (npr. uzorak B532 je sličniji uzorku B539 nego sam sebi) mogu se definirati kao posljedica korištenja kratkih sekvenci u izračunima gdje i najmanje promjene (npr. nekoliko desetaka polimorfizama) mogu imati veliki utjecaj na rezultat. Međutim, dodavanjem većeg broja gena, kao što je slučaj u drugom setu (10x10), odnosi među uzorcima imaju više smisla jer su svi odnosi računati na puno većim fragmentima (udio pronađenih polimorfizama je puno manji). Prema tome se jasno moglo da su uzorci divokoza međusobno sličniji bez obzira na korištenu referencu. Kao potvrda ovih rezultata, provedene metode MDS-a na drugom (100) i trećem setu (500) gena također su prikazale vrlo slične odnose između uzoraka s tim da su i ovdje postojale male razlike kod uzoraka mapiranih na domaću ovcu i američku planinsku kozu. Jasno se moglo zaključiti da su si uzorci divokoze međusobno sličniji u odnosu na reference na koje su se mapirali, dok su postojeće razlike unutar uzoraka jako male (oko 1 % za drugi i 000,5 % za treći set). Nadalje, svi uzorci su bili gotovo jednako udaljeni od svojih referenci, s očekivano nekoliko iznimki. Ovo je još jedan dokaz da metode mapiranja mogu uspješno sastaviti vrlo slične genomske fragmente iz sekvenciranih uzoraka bez obzira na odabrani referentni genom. Do sličnog zaključka došli su Gopalakrishnan i sur. (2017) u svom radu o utjecaju referenci vuka i psa na mapiranje. Analize koje su proveli potvrdile su da su rezultati vrlo slični bez obzira na korištenu referencu i da su razlike vrlo male. Naravno, te su razlike uglavnom rijetke varijante u genomu koje se mogu detaljnije proučiti radi preciznije procjene. Isto vrijedi i u ovom slučaju gdje bi za detaljniju analizu

genoma bilo poželjno proučiti svaki korišteni fragment (što uključuje informacije o genima koji su se koristili u analizi).

5.5 Usporedba nDNA konsenzusnih sekvenci s dostupnim genomskim sekvencama divokoza u Banci gena

Kako je već prethodno spomenuto, u procesima mapiranja referenca može imati značajan utjecaj na rezultate. Analizama sličnosti dokazan je utjecaj referenci na rezultate ali u puno manjoj mjeri nego što je to bilo očekivano. Međutim, treba imati na umu da bi se svaka korištena referenca treba detaljno proučiti prije nego se počne koristiti u analizama. Na temelju rezultata mapiranja svih uzoraka na različite reference, kombinacije uzoraka divokoze i genoma domaće koze pokazale su se kao najbolji izbor zbog najveće kvalitete referentnog ali i novosastavljenih genoma, bliske srodnosti domaće koze i divokoze te zbog istog broja kromosoma. Glavni cilj u ovim analizama je bio izolirati kratke regije iz konsenzusnih sekvenci i usporediti ih s istim sekvencama divokoze dostupnih u Banci gena. Pretragom Banke gena ustanovljeno je da najveći dio pronađenih genomskih sekvenci divokoze pripadaju mtDNA (PCG, CR). Uz njih, pronađen je podatkovni set intronskih regija koji je bio korišten u provedbi filogenetske analize divokoza (Pérez i sur., 2017). Cijela studija bazirala se na rekonstrukciji filogenije divokoza korištenjem različitih intronskih sekvenci koje se nalaze na različitim kromosomima te usporedbi rezultata s onima dobivenim analizom mtDNA sekvenci, drugih nuklearnih markera i kromosoma Y. Glavni cilj je bio procijeniti evolucijsku povijest divokoza iz genomskih komponenti introna. Selekcija introna provedena je na temelju prethodnih istraživanja (Hassanin i sur., 2012; Igea i sur., 2010; Hailer i sur., 2012) te je po jedan intron odabran iz svakog kromosoma (nakon provedene amplifikacij ukupno 23 introna od 30). U ovoj disertaciji su te iste intronske sekvence uspješno izolirane iz konsenzusnih sekvenci i napravljeno je poravnanje koje se sastojalo od 14 sekvenci divokoza iz Pérez i sur. (2017), šest sekvenci divokoza iz konsenzusnih genoma te tri sekvence koje su predstavljale uljeze: govedo, koza i ovca. Rezultati filogenije potvrdili su da su intronske sekvence dobivene iz konsenzusnih genoma zadovoljavajuće kvalitete i da su se grupirale s preuzetim uzorcima divokoze iz Banke gane na razini vrste i podvrste. Osim toga, znatno su se razlikovale od sekvenci koje su korištene kao uljezi. Ovom se analizom dokazalo da se na ovaj način može doći do vrijednih zaključaka, čak i iz malog broja uzoraka niske pokrivenosti, pogotovo ako je referenca visoke kvalitete dostupna u Banci gena. Osim toga, provedena analiza bi mogla biti korisna za buduće studije koje koriste slične genomske podataka niske pokrivenost. Unatoč činjenici da će u budućnosti postojati više visokokvalitetnih referentnih

genoma, ovaj pristup se još uvijek može koristiti za dobivanje dijelova genoma koji se mogu koristit za određeni znanstveni interes uz nisku cijenu.

5.6 Hibridno sastavljanje genoma

Kako je već spomenuto, u hibridnim metodama sastavljanja najčešće se provodi *de novo* metoda kao prvi korak a dobiveni se rezultati potom mapiraju na dostupnu referencu. Nekoliko radova koristilo je nekoliko različitih varijacija ove metode (Vezzi i sur., 2011; Kim i sur., 2013; Bao i sur., 2014; Kolmogorov i sur., 2014; Wang i sur., 2014; Buza i sur., 2015; Tamazian i sur., 2016; Lischer i Shimizu, 2017; Kolmogorov i sur., 2018; Siddiki i sur, 2019). Nekoliko specijaliziranih alata razvijeno je za ovu metodu u kojima su ulazne datoteke kontig sekvence (sastavljene *de novo*) te referentni genom (Kim i sur., 2013; Bao i sur., 2014; Buza i sur., 2015) ili više njih (Kolmogorov i sur., 2014). Međutim, u takvim radovima većinom nisu naglašeni nedostaci ovih metoda kao ni izazovi s kojima se susreću korisnici. Jedan od najvećih izazova je provedba prvog koraka, *de novo* metode. Općenito, ova metoda zahtjeva visoke performanse računala na kojem se provode analize a količina memorije potrebna za ovu metodu najviše ovisi o kvaliteti i tipu genomskih podataka te o vrsti genoma koji se želi sastaviti (Tsai i sur., 2010; Kim i sur., 2013; Bao i sur., 2014). Osim toga, bitno je napomenuti da je većina ovih metoda testirana na bakterijskim genomima ili simuliranim genomskim podacima (Kolmogorov i sur., 2014; Buza i sur., 2015). Kako je već spomenuto, genomski podaci koji su dobiveni NGS tehnologijama često se sastoje od nekoliko milijuna kratkih fragmenata (veličine 150-500 pb). Prilikom korištenje tih fragmenata, svaki fragment se uspoređuje s milijunima drugih fragmenata zbog čega je cijeli proces računalno intenzivan i zahtjeva puno radne memorije. Upravo iz ovih razloga, u sklopu ove disertacije hibridno sastavljanje genoma je bilo neuspješno, a glavni razlog je nedostatak memorije uzrokovan usporedbom milijuna kratkih fragmenata. Cijeli proces bio bi manje zahtjevnji kada bi se koristili duži fragmenti (Wang i sur., 2014). Uz navedeno, za veliki broj opisanih protokola hibridne metode, potrebno je dodatno sekvencirati genomske podatke kojim bi se nadopunile nepotpune informacije u sastavljenim genomima (Assefa i sur., 2009; Soderlund i sur., 2011; Bosi i sur., 2015). Zaključno, može se reći da ova metoda, iako pruža alternativno riješenje u procesima sastavljanja genoma, relativno je nova i zbog toga se još uvijek treba poraditi na njenoj optimizaciji.

6 ZAKLJUČCI

- Nakon provođenja dviju metoda za rekonstrukciju kompletnih sekvenci mtDNA (mapiranje i *de novo*) te usporedbom dobivenih sekvenci mtDNA jasno se moglo zaključiti da su obje metode bile pogodne za rekonstrukciju. Metoda *de novo* zbog brzine i jednostavnijeg postupka pokazala se kao puno bolji izbor. Osim toga, *de novo* metodom su uspješno rekonstruirane kompletne mtDNA sekvence iz uzoraka koji nisu prošli kontrolu kvalitete (Gams53, Gams85, OSIL-06). Na ovaj način sve analize mtDNA provedene su na većem broju uzoraka. Drugim riječima, ukoliko bi se koristila samo metoda mapiranjem, navedena tri uzorka se ne bi mogla koristiti u daljnim analizama.
- Programski alati za anotaciju mtDNA MITOS i GeSEQ dali su vrlo slične rezultate za svih 10 uzoraka, a varijacije u START i STOP kodonima detektirane su kod četiri gena (ND1, ND2, ND3, ND5). Pronađene varijacije odnose se na dvije ili tri baze u START i STOP kodonima, a njihova se pojava može tumačiti kao posljedica drugačijih algoritama koje anotatori koriste u analizama i zbog prisutnosti veće varijacije u ovim genima.
- Provedene filogenetske analize (maksimalna vjerodostojnost i Bayesovska) na deset mtDNA sekvenci divokoza rekonstruiranih u ovoj disertaciji u kombinaciji s pet mtDNA sekvenci iz Banke gena rezultirale su identičnim filogenetskim stablima za rod *Rupicapra*. Ovim su se rezultatima potvrdila prethodna istraživanja o divokozi podijelivši je u tri mtDNA klastera (W, C i E). Osim toga, među sekvencama je zabilježena visoka haplotipna i nukleotidna varijabilnost na razini vrsta i podvrsta. Velika varijabilnost također je bila prisutna u PCG regijama s tim da je *R. rupicapra* imala nižu raznolikost unatoč većem broju uzoraka, dok je na razini podvrste, *R. r. balcanica* (uključujući i jedan hibrid) pokazala najveću stopu diferencijacije.
- Provedene filogenetske analize (maksimalna vjerodostojnost i Bayesovska) na 40 mtDNA sekvenci roda *Caprinae* (uključujući četiri sekvence dobivene u ovoj disertaciji) i pet sekvenci roda *Bovidae* rezultirale su istom topologijom. Na filogenetskom stablu cijelog roda *Caprinae* prikazana su bila tri klastera (*Pantholopini*, *Caprini*, *Ovibovini*) čime su se potvrdili rezultati dobiveni u prethodnim istraživanjima.
- Na temelju usporedbe novosastavljenih genoma divokoze s genomima srodnih vrsta moglo se zaključiti da gotovo svi korišteni srodni genomi mogu poslužiti kao dobra referenca. Međutim, iako su sve korištene vrste nemodelne, najbolji rezultati dobiveni

su korištenjem genoma domaće koze i ovce što je i očekivano budući da se radi o izrazito bitnim vrstama u agronomiji koje su često u fokusu istraživanja. Broj pronađenih gena za većinu kombinacija divokoze i srodnih vrsta bio je jako visok čime se potvrđuje da se ti genomi mogu koristiti u procesima mapiranja. Međutim, provedenim procesima mapiranja ustanovljeno je da su neki od korištenih genoma niske kvalitete dok su nekim genomima pronađene nepravilnosti u informacijama koje su dostupne u Banci gena. Ovim se još jednom potvrdilo da nisu svi dostupni genomi dobre kvalitete. Drugim riječima, svaka sekvenca dostupna u Banci gena bi se prije korištenja trebala provjeriti.

- Iz analiza sličnosti moglo se zaključiti da odnosi između svih kombinacija prije svega ovise o fragmentima gena ili genoma koji se koriste u ovim analizama. Iako je broj pronađenih polimorfizama imao veći utjecaj na rezultate kod korištenja pojedinačnih fragmenata gena, taj broj je bio zanemariv kada se koriste duži dijelovi genoma (100 i 500 gena) s razlikama unutar uzoraka od 1 %. Drugim riječima, veće udaljenosti između kombinacija su izračunate kod kraćih poravnanja. Prema rezultatima MDS-a na setu od 100 i 500 gena jasno se moglo zaključiti da su uzorci divokoza mapirani na različite reference međusobno sličniji s tim da su pronađene razlike kod uzoraka mapiranih na domaću ovcu i američku planinsku kozu iznosile oko 1,5 % (100 gena) i 000,5 % (500 gena).
- Prema provedenim usporedbama intronskih regija iz novosastavljenih genoma divokoze, introna dostupnih u Banci gena te tri referentna genoma, moglo se zaključiti da su intronske sekvence dobivene iz konsenzusnih genoma zadovoljavajuće kvalitete. Šest sekvenci divokoza iz ove disertacije se grupiralo s ostatkom divokoza, uz visoke vrijednosti *bootstrapa*. Drugim riječima, iako je genom koze korišten kao referenca, uzorci se nisu grupirali s kozom nego su po svojoj genomskoj strukturi bili bliži drugim divokozama. Osim toga, uzorci su se na razini vrste te na razini podvrste grupirali sukladno taksonomiji i očekivanjima.
- Sve provedene metode za hibridno sastavljanje genoma kao i kombinacija dvaju alata za *de novo* sastavljanje rezultirale su naglim prekidom procesa zbog nedostaka računalne memorije, uobičajeno peti ili šeti dan nakon pokretanja. S obzirom da genomski podaci iz ove disertacije sadrže kratke DNA fragmente niske pokrivenosti, ovakav rezultat je očekivan.

7 LITERATURA

1. Allard M.W., Miyamoto M.M., Jarecki L., Kraus F., Tennant M.R. (1992). DNA systematics and evolution of the artiodactyl family *Bovidae*. Proc Natl Acad Sci U S A 89 (9): 3972–3976. doi:10.1073/pnas.89.9.3972
2. Al-Nakeeb K., Petersen T.N., Sicheritz-Pontén T. (2017). Norgal: Extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. BMC Bioinformatics 18 (1): 1–7. doi:10.1186/s12859-017-1927-y
3. Alqahtani F., Măndoiu I.I. (2020). Statistical Mitogenome Assembly with RepeaTs. J Comput Biol 27 (0): 1–15. doi:10.1089/cmb.2019.0505
4. Altenhoff A.M., Dessimoz C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5 (1). doi:10.1371/journal.pcbi.1000262
5. Altmann A., Weber P., Bader D., Preuß M., Binder E.B., Müller-Myhsok B. (2012). A beginners guide to SNP calling from high-Throughput DNA-sequencing data. Hum Genet 131 (10): 1541–1554. doi:10.1007/s00439-012-1213-z
6. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990). Basic local alignment search tool. J Mol Biol 215 (3): 403–410. doi:10.1016/S0022-2836(05)80360-2
7. Alvarez-Busto J., García-Etxebarria K., Herrero J., Garin I., Jugo B.M. (2007). Diversity and evolution of the Mhc-DRB1 gene in the two endemic Iberian subspecies of Pyrenean chamois, *Rupicapra pyrenaica*. Heredity (Edinb) 99 (4): 406–413. doi:10.1038/sj.hdy.6801016
8. Anderson S., Bankier A.T., Barrell B.G., De Bruijn M.H.L., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H., Smith A.J.H., Staden R., Young I.G. (1981). Sequence and organization of the human mitochondrial genome. Nature 290 (5806): 457–465. doi:10.1038/290457a0
9. Anderwald P., Ambarli H., Avramov S., Ciach M., Corlatti L., Farkas A., Jovanovic M., Papaioannou H., Peters W., Sarasa M., Šprem N., Weinberg P., Willisch C. (2021). *Rupicapra rupicapra* (amended version of 2020 assessment). The IUCN

Red List of Threatened Species 2020. https://www.iucnredlist.org/species/39255/195863093 [pristupljeno 4. prosinca, 2021]

10. Andrews S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
11. Apollonio M., Scandura M., Šprem N. (2014). Reintroductions as a management tool for European Ungulates. In Behaviour and Management of European Ungulates. Whittles Publishing, Scotland, UK
12. Arif I.A., Bakir M.A., Khan H.A. (2012). Inferring the phylogeny of *Bovidae* using mitochondrial DNA sequences: Resolving power of individual genes relative to complete genomes. Evol Bioinforma 2012 (8): 139–150. doi:10.4137/EBO.S8897
13. Assefa S., Keane T.M., Otto T.D., Newbold C., Berriman M. (2009). ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25 (15): 1968–1969. doi:10.1093/bioinformatics/btp347
14. Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A. V., Sirotkin A. V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19 (5): 455–477. doi:10.1089/cmb.2012.0021
15. Bao E., Jiang T., Girke T. (2014). AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references. Bioinformatics 30 (12): 319–328. doi:10.1093/bioinformatics/btu291
16. Batzoglou S., Jaffe D.B., Stanley K., Butler J., Gnerre S., Mauceli E., Berger B., Mesirov J.P., Lander E.S. (2002). ARACHNE: a whole-genome shotgun assembler. Genome Res 12(1):177-89. doi: 10.1101/gr.208902
17. Bernt M., Donath A., Jühling F., Externbrink F., Florentz C., Fritzsch G., Pütz J., Middendorf M., Stadler P.F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecul Phylogenet Evolution 69 (2): 313–319. doi:10.1016/j.ympev.2012.08.023
18. Betancur-R R., Naylor G.J.P., Ortí G. (2014). Conserved genes, sampling error, and phylogenomic inference. Syst Biol 63 (2): 257–262. doi:10.1093/sysbio/syt073
19. Bohling J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. Ecol Evol 10 (14): 7585–7601. doi:10.1002/ece3.6483

20. Bickhart D.M., Rosen B.D., Koren S., Sayre B.L., Hastie A.R., Chan S., Lee J., Lam E.T., Liachko I., Sullivan S.T., Burton J.N., Huson H.J., Nystrom J.C., Kelley C.M., Hutchison J.L., Zhou Y., Sun J., Crisà A., Ponce De León F.A., Schwartz J.C., Hammond J.A., Waldbieser G.C., Schroeder S.G., Liu G.E., Dunham M.J., Shendure J., Sonstegard T.S., Phillippy A.M., Van Tassell C.P., Smith T.P.L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet 49 (4): 643–650. doi:10.1038/ng.3802
21. Bolger A.M., Lohse M., Usadel B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30 (15): 2114–2120. doi:10.1093/bioinformatics/btu170
22. Boore J.L. (1999). Animal mitochondrial genomes. Nucleic Acids Res 27 (8): 1767–1780. doi:10.1093/nar/27.8.1767
23. Boore J.L., Macey J.R., Medina M. (2005). Sequencing and comparing whole mitochondrial genomes of animals. Methods Enzymol 395: 311–348. doi:10.1016/S0076-6879(05)95019-2
24. Bosi E., Donati B., Galardini M., Brunetti S., Sagot M.F., Lió P., Crescenzi P., Fani R., Fondi M. (2015). MeDuSa: A multi-draft based scaffolder. Bioinformatics 31 (15): 2443–2451. doi:10.1093/bioinformatics/btv171
25. Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10 (4): 1–6. doi:10.1371/journal.pcbi.1003537
26. Bourgeois Y.X.C., Warren B.H. (2021). An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. Mol Ecol 30 (23): 6036–6071. doi:10.1111/mec.15989
27. Brocchieri L. (2001). Phylogenetic inferences from molecular sequences: Review and critique. Theor Popul Biol 59 (1): 27–40. doi:10.1006/tpbi.2000.1485
28. Brown G.G., Simpson M. V. (1982). Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc Natl Acad Sci U S A 79 (10 I): 3246–3250. doi:10.1073/pnas.79.10.3246
29. Brown T.A. (2002). Genomes. 2nd ed. Oxford: Wiley-Liss; 2002. PMID: 20821850

30. Buza K., Wilczynski B., Dojer N. (2015). RECORD: Reference-assisted genome assembly for closely related genomes. Int J Genomics 2015. doi:10.1155/2015/563482
31. C. elegans Sequencing Consortium (CeSC). (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282(5396):2012-8. doi: 10.1126/science.282.5396.2012
32. Camerano L. (1914). Ricerche intorno ai camosci (Parte Ia, Iia, IIIa). Memorie della Reale Accademia delle Scienze di Torino (Cl Sci Fis Mat Nat) 64: 1–82, 64: 1–88, 65: 1–82
33. Caparroz R., Mantellatto A.M.B., Bertioli D.J., Figueiredo M.G., Duarte J.M.B. (2015). Characterization of the complete mitochondrial genome and a set of polymorphic microsatellite markers through next-generation sequencing for the brown brocket deer *Mazama gouazoubira*. Genet Mol Biol 38 (3): 338–345. doi:10.1590/S1415-475738320140344
34. Card D.C., Schield D.R., Reyes-Velasco J., Fujita M.K., Andrew A.L., Oyler-McCance S.J., Fike J.A., Tomback D.F., Ruggiero R.P., Castoe T.A. (2014). Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. PLoS One 9 (9). doi:10.1371/journal.pone.0106649
35. Chen L., Qiu Q., Jiang Y., Wang K., Lin Z., Li Z., Bibi F., Yang Y., Wang J., Nie W., Su W., Liu G., Li Q., Fu W., Pan X., Liu C., Yang J., Zhang Chenzhou, Yin Y., Wang Yu, Zhao Y., Zhang Chen, Wang Z., Qin Y., Liu W., Wang B., Ren Y., Zhang R., Zeng Y., Da Fonseca R.R., Wei B., Li R., Wan W., Zhao R., Zhu W., Wang Yutao, Duan S., Gao Y., Zhang Y.E., Chen C., Hvilsom C., Epps C.W., Chemnick L.G., Dong Y., Mirarab S., Siegismund H.R., Ryder O.A., Gilbert M.T.P., Lewin H.A., Zhang G., Heller R., Wang W. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science (80-) 364 (6446). doi:10.1126/science.aav6202
36. Chen M.Y., Liang D., Zhang P. (2017). Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. Genome Biol Evol 9 (8): 1998–2012. doi:10.1093/gbe/evx147
37. Chen S., Zhou Y., Chen Y., Gu J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34 (17): i884–i890. doi:10.1093/bioinformatics/bty560

38. Chikuni K., Mori Y., Tabata T., Saito M., Monma M., Kosugiyama M. (1995). Molecular phylogeny based on the κ-casein and cytochrome b sequences in the mammalian suborder *Ruminantia*. J Mol Evol 41 (6): 859–866. doi:10.1007/BF00173165
39. Clayton D.A. (1992). Structure and Function of the Mitochondrial Genome. J Inherit Metab Dis 15 (4): 439–447. doi:10.1007/BF01799602
40. Colli L., Lancioni H., Cardinali I., Olivieri A., Capodiferro M.R., Pellecchia M., Rzepus M., Zamani W., Naderi S., Gandini F., Vahidi S.M.F., Agha S., Randi E., Battaglia V., Sardina M.T., Portolano B., Rezaei H.R., Lymberakis P., Boyer F., Coissac E., Pompanon F., Taberlet P., Ajmone Marsan P., Achilli A. (2015). Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. BMC Genomics 16 (1): 1–12. doi:10.1186/s12864-015-2342-2
41. Collins F.S., Brooks L.D., Chakravarti A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8 (12): 1229–1231. doi:10.1101/gr.8.12.1229
42. Compeau P.E.C., Pevzner P.A., Tesler G. (2011). How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29 (11): 987–991. doi:10.1038/nbt.2023
43. Cordaux R., Batzer M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet 10 (10): 691–703. doi:10.1038/nrg2640
44. Corlatti L., Herrero J., Ferretti F., Anderwald P., García-González R., Hammer S., Nores C., Rossi L., Lovari S. (2022a). Northern Chamois, *Rupicapra rupicapra* (Linnaeus, 1758) and Southern Chamois, *Rupicapra pyrenaica* Bonaparte, 1845.). In: Handbook of the Mammals of Europe - Terrestrial *Cetartiodactyla* (Corlatti L., Zachos F., eds.), Springer Nature
45. Corlatti L., Iacolina L., Safner T., Apollonio M., Buzan E., Ferretti F., Hammer S.E., Herrero J., Rossi L., Serrano E., Arnal M.C., Brivio F., Chirichella R., Cotza A., Crestanello B., Espunyes J., Fernández de Luco D., Friedrich S., Gačić D., Grassi L., Grignolio S., Hauffe H.C., Kavčić K., Kinser A., Lioce F., Malagnino A., Miller C., Peters W., Pokorny B., Reiner R., Rezić A., Stipoljev S., Tešija T., Yankov Y., Zwijacz-Kozica T., Šprem N. (2022b). Past, present and future of chamois science. Wildlife Biol 1–13. doi:10.1002/wlb3.01025
46. Corlatti L., Lorenzini R., Lovari S. (2011). The conservation of the chamois *Rupicapra* spp. Mamm Rev 41 (2): 163–174. doi:10.1111/j.1365-2907.2011.00187.x
47. Couturier M. (1938). Le Chamois. Arthaud, Grenoble, France

48. Crestanello B., Pecchioli E., Vernesi C., Mona S., Martínková N., Janiga M., Hauffe H.C., Bertorelle G. (2009). The genetic impact of translocations and habitat fragmentation in chamois (*Rupicapra*) spp. J Hered 100 (6): 691–708. doi:10.1093/jhered/esp053
49. Cui P., Ji R., Ding F., Qi D., Gao H., Meng H., Yu J., Hu S., Zhang H. (2007). A Complete Mitochondrial Genome Sequence of the Wild Two Humped Camel (*Camelus bactrianus ferus*): an Evolutionary History of Camelidae. BMC Genomics 8: 241. doi:10.1186/1471-2164-8-241
50. Cutting G.R. (2014). Annotating DNA Variants Is the Next Major Goal for Human Genetics. The Amer Jour. of Hum. Genet. 94(1)5-10. doi.org/10.1016/j.ajhg.2013.12.008
51. Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. (2011). The variant call format and VCFtools. Bioinformatics 27 (15): 2156–2158. doi:10.1093/ bioinformatics/btr330
52. Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. (2011). The variant call format and VCFtools. Bioinformatics 27 (15): 2156–2158. doi:10.1093/bioinformatics/btr330
53. Davenport K.M., Bickhart D.M., Worley K., Murali S.C., Salavati M., Clark E.L., Cockett N.E., Heaton M.P., Smith T.P.L., Murdoch B.M., Rosen B.D. (2022). An improved ovine reference genome assembly to facilitate in-depth functional annotation of the sheep genome. Gigascience 11: 1–9. doi:10.1093/gigascience/giab096
54. Davenport K.M., Duan M., Hunter S.S., New D.D., Fagnan M.W., Highland M.A., Murdoch B.M. (2018). Complete mitochondrial genome sequence of bighorn sheep. Genome Announc 6 (23): 4–5. doi:10.1128/genomeA.00464-18
55. Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12 (7): 499–510. doi:10.1038/nrg3012
56. Dias G.R., Dupim E.G., Vanderlinde T., Mello B., Carvalho A.B. (2020). A phylogenomic study of *Steganinae* fruit flies (*Diptera*: *Drosophilidae*): strong gene tree heterogeneity and evidence for monophyly. BMC Evol Biol 20 (1): 1–12. doi:10.1186/s12862-020-01703-7

57. Dierckxsens N., Mardulyn P., Smits G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45 (4). doi:10.1093/nar/gkw955
58. Dolan J.M. (1963). Beitrag zur systematischen Gliederung des Tribus Rupicaprini Simpson, 1945. Journal of Zoological Systematics and Evolutionary Research 1: 311–407
59. Dominguez Del Angel V., Hjerde E., Sterck L., Capella-Gutierrez S., Notredame C., Vinnere Pettersson O., Amselem J., Bouri L., Bocs S., Klopp C., Gibrat J.F., Vlasova A., Leskosek B.L., Soler L., Binzer-Panchal M. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research 7. doi:10.12688/f1000research.13598.1
60. Dong Y., Zhang X., Xie M., Arefnezhad B., Wang Z., Wang Wenliang, Feng S., Huang G., Guan R., Shen W., Bunch R., McCulloch R., Li Q., Li B., Zhang G., Xu X., Kijas J.W., Salekdeh G.H., Wang Wen, Jiang Y. (2015). Reference genome of wild goat (capra aegagrus) and sequencing of goat breeds provide insight into genic basis of goat domestication. BMC Genomics 16 (1): 1–11. doi:10.1186/s12864-015-1606-1
61. Dotsev A. V., Kunz E., Shakhin A. V., Petrov S.N., Kostyunina O. V., Okhlopkov I.M., Deniskova T.E., Barbato M., Bagirov V.A., Medvedev D.G., Krebs S., Brem G., Medugorac I., Zinovieva N.A. (2019). The first complete mitochondrial genomes of snow sheep (*Ovis nivicola*) and thinhorn sheep (*Ovis dalli*) and their phylogenetic implications for the genus Ovis. Mitochondrial DNA Part B Resour 4 (1): 1332–1333. doi:10.1080/23802359.2018.1535849
62. Douglas K.C., Halbert N.D., Kolenda C., Childers C., Hunter D.L., Derr J.N. (2011). Complete mitochondrial DNA sequence analysis of *Bison bison* and bison-cattle hybrids: Function and phylogeny. Mitochondrion 11 (1): 166–175. doi:10.1016/j.mito.2010.09.005
63. Douzery E., Randi E. (1997). The mitochondrial control region of *Cervidae*: Evolutionary patterns and phylogenetic content. Mol Biol Evol 14 (11): 1154–116 doi:10.1093/oxfordjournals.molbev.a025725
64. Ejigu G.F., Jung J. (2020). Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology (Basel) 9 (9): 1–27. doi:10.3390/biology9090295

65. Ekblom R., Wolf J.B.W. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7 (9): 1026–1042. doi:10.1111/eva.12178
66. El-Metwally S., Hamza T., Zakaria M., Helmy M. (2013). Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PLoS Comput Biol 9 (12). doi:10.1371/journal.pcbi.1003345
67. Evans J.D., Brown S.J., Hackett K.J.J., Robinson G., Richards S., Lawson D., Elsik C., Coddington J., Edwards O., Emrich S., Gabaldon T., Goldsmith M., Hanes G., Misof B., Muñoz-Torres M., Niehuis O., Papanicolaou A., Pfrender M., Poelchau M., Purcell-Miramontes M., Robertson H.M., Ryder O., Tagu D., Torres T., Zdobnov E., Zhang G., Zhou X. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104 (5): 595–600. doi:10.1093/jhered/est050
68. Farrer R.A., Henk D.A., MacLean D., Studholme D.J., Fisher M.C. (2013). Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. Sci Rep 3: 1–6. doi:10.1038/srep01512
69. Field K.G., Olsen G.J., Lane D.J., Giovannoni S.J., Ghiselin M.T., Raff E.C., Pace N.R., Raff R.A. (1988). Molecular phylogeny of the animal kingdom. Science (80-) 239 (4841): 748–753. doi:10.1126/science.3277277
70. Field K.G., Olsen G.J., Lane D.J., Giovannoni S.J., Ghiselin M.T., Raff E.C., Pace N.R., Raff R.A. (1988). Molecular phylogeny of the animal kingdom. Science 239(4841 Pt 1):748-53. doi: 10.1126/science.3277277. PMID: 3277277
71. Fitch W.M. (1970). Distinguishing homologous from analogous proteins. Syst Zool 19 (2): 99–113. doi:10.2307/2412448
72. Fitch W.M. (1970). Distinguishing Homologous from Analogous Proteins. System. Biol 19(2): 99-113. doi:https://doi.org/10.2307/2412448
73. Fonseca N.A., Rung J., Brazma A., Marioni J.C. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics 28 (24): 3169–3177. doi:10.1093/bioinformatics/bts605
74. Forsyth DM. (2005). Chamois. In: King CM (ed.) The Handbook of New Zealand Mammals, 351–360. Oxford University Press, Aukland, New Zealand.
75. Fox G.E., Stackebrandt E., Hespell R.B., Gibson J., Maniloff J., Dyer T.A., Wolfe R.S., Balch W.E., Tanner R.S., Magrum L.J., Zablen L.B. The phylogeny of prokaryotes. Science. doi:1980;209:457–63

76. Fuentes-Pardo A.P., Ruzzante D.E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol 26 (20): 5369–5406. doi:10.1111/mec.14264
77. Gabaldón T. (2008). Large-scale assignment of orthology: back to phylogenetics? Genome Biol 9 (10): 235. doi:10.1186/gb-2008-9-10-235
78. Galla S.J., Forsdick N.J., Brown L., Hoeppner M.P., Knapp M., Maloney R.F., Moraga R., Santure A.W., Steeves T.E. (2019). Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. Genes (Basel) 10 (1). doi:10.3390/genes10010009
79. Gatesy J., Yelon D., Desalle R., Vrba E.S. (1992). Phylogeny of the *Bovidae* (*Artiodactyla*, *Mammalia*), Based on Mitochondrial Ribosomal DNA Sequences. Mol Biol. 9(3):433-46. doi: 10.1093/oxfordjournals.molbev.a040734.
80. Gentry A.W. (1992). The subfamilies and tribes of the family *Bovidae*. Mamm Rev 22 (1): 1–32. doi:10.1111/j.1365-2907.1992.tb00116.x
81. Ghanatsaman Z.A., Wang G.D., Asadollahpour N.H., Asadi Fozi M., Peng M.S., Esmailizadeh A., Zhang Y.P. (2020). Whole genome resequencing of the Iranian native dogs and wolves to unravel variome during dog domestication. BMC Genomics 21 (1): 1–11. doi:10.1186/s12864-020-6619-8
82. Giani A.M., Gallo G.R., Gianfranceschi L., Formenti G. (2019). Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J 18: 9–19. doi:10.1016/j.csbj.2019.11.002
83. Gibson A., Gowri-Shankar V., Higgs P.G., Rattray M. (2005). A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol 22 (2): 251–264. doi:10.1093/molbev/msi012
84. Gnerre S., Lander E.S., Lindblad-Toh K., Jaffe D.B. (2009). Assisted assembly: How to improve a de novo genome assembly by using related species. Genome Biol 10 (8). doi:10.1186/gb-2009-10-8-r88
85. Gnerre S., MacCallum I., Przybylski D., Ribeiro F.J., Burton J.N., Walker B.J., Sharpe T., Hall G., Shea T.P., Sykes S., Berlin A.M., Aird D., Costello M., Daza R., Williams L., Nicol R., Gnirke A., Nusbaum C., Lander E.S., Jaffe D.B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci 108 (4): 1513–1518. doi:10.1073/pnas.1017351108

86. Gopalakrishnan S., Samaniego Castruita J.A., Sinding M.H.S., Kuderna L.F.K., Räikkönen J., Petersen B., Sicheritz-Ponten T., Larson G., Orlando L., Marques-Bonet T., Hansen A.J., Dalén L., Gilbert M.T.P. (2017). The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. BMC Genomics 18 (1): 1–11. doi:10.1186/s12864-017-3883-3
87. Gordon D., Abajian C., Green P. (1998). Consed: A graphical tool for sequence finishing. Genome Res 8 (3): 195–202. doi:10.1101/gr.8.3.195
88. Groves P., Shields G.F. (1996). Phylogenetics of the *Caprinae* based on cytochrome b sequence. Mol Phylogenet Evol 5 (3): 467–476. doi:10.1006/mpev.1996.0043
89. Groves P., Shields G.F. (1997). Cytochrome B Sequences Suggest Convergent Evolution of the Asian Takin and Arctic Muskox. Mol Phylogenet Evol 8 (3): 363–374. doi:10.1006/mpev.1997.0423
90. Gupta A., Bhardwaj A. (2015). Mitochondrial DNA - a Tool for Phylogenetic and Biodiversity Search in Equines. J Biodivers Endanger Species 01 (s1). doi:10.4172/2332-2543.s1-006
91. Hahn C., Bachmann L., Chevreux B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. Nucleic Acids Res 41 (13). doi:10.1093/nar/gkt371
92. Hailer F., Kutschera V.E., Hallström B.M., Klassert D., Fain S.R., Leonard J.A., Arnason U., Janke A. (2012). Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science (80) 336 (6079): 344–347. doi:10.1126/science.1216424
93. Hammer S., Nadlinger K., Hartl G.B. (1995). Mitochondrial DNA differentiation in chamois (genus *Rupicapra*): implications for taxonomy, conservation, and management 145–155
94. Hassanin A., Bonillo C., Nguyen B.X., Cruaud C. (2010). Comparisons between Mitochondrial Genomes of Domestic Goat (Capra hircus) Reveal the Presence of Numts and Multiple Sequencing Errors. Mitochondrial DNA 21 (3–4): 68–76. doi:10.3109/19401736.2010.49 0583
95. Hassanin A., Delsuc F., Ropiquet A., Hammer C., Jansen Van Vuuren B., Matthee C., Ruiz-Garcia M., Catzeflis F., Areskoug V., Nguyen T.T., Couloux A. (2012). Pattern and timing of diversification of *Cetartiodactyla* (*Mammalia*, *Laurasiatheria*), as

revealed by a comprehensive analysis of mitochondrial genomes. Comptes Rendus - Biol 335 (1): 32–50. doi:10.1016/j.crvi.2011.11.002

96. Hassanin A., Douzery E.J.P. (1999). Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of *Bovidae*. Proc R Soc B Biol Sci 266 (1422): 893–900. doi:10.1098/rspb.1999.0720
97. Hassanin A., Pasquet E., Vigne J.D. (1998). Molecular systematics of the subfamily *Caprinae* (*artiodactyla*, *bovidae*) as determined from cytochrome b sequences. J Mamm Evol 5 (3): 217–236. doi:10.1023/A:1020560412929
98. Hassanin A., Ropiquet A., Couloux A., Cruaud C. (2009). Evolution of the mitochondrial genome in mammals living at high altitude: New insights from a study of the tribe Caprini (*Bovidae*, *Antilopinae*). J Mol Evol 68 (4): 293–310. doi:10.1007/s00239-009-9208-7
99. Hatem A., Bozdağ D., Toland A.E., Çatalyürek Ü. V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics 14 (1). doi:10.1186/1471-2105-14-184
100. Heath T.A., Hedtke S.M., Hillis D.M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46 (3): 239–257. doi:10.3724/SP.J.1002.2008.08016
101. Herrero J., Lovari S., Nores C., Toigo C. 2020. *Rupicapra pyrenaica*. The IUCN Red List of Threatened Species 2020. *Rupicapra pyrenaica*. The IUCN Red List of Threatened Species 2020. Available at: https://www.iucnredlist.org/species/19771/171131310 [pristupljeno 4. prosinca, 2021]
102. Hiendleder S., Lewalski H., Wassmuth R., Janke A. (1998). The complete mitochondrial DNA sequence of the domestic sheep (*Ovis aries*) and comparison with the other major ovine haplotype. J Mol Evol 47 (4): 441–448. doi:10.1007/PL00006401
103. Hill E., Linacre A.M.T., Toop S., Murphy N.P., Strugnell J.M. (2017). The complete mitochondrial genome of *Axis porcinus* (*Mammalia*: *Cervidae*) from Victoria, Australia, using MiSeq sequencing. Mitochondrial DNA Part B Resour 2 (2): 453–454. doi:10.1080/23802359.2017.1357451
104. Hixson J.E., Wong T.W., Clayton D.A. (1986). Both the conserved stem-loop and divergent 5'-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. J Biol Chem 261 (5): 2384–2390

105. Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. Mol Biol Evol 35 (2): 518–522. doi:10.5281/zenodo.854445
106. Holley R.W., Apgar J., Everett G.A., Madison J.T., Marquisee M., Merrill S.H., Penswick J.R., Zamir A. Structure of a ribonucleic acid. Science. 1965 Mar 19;147(3664):1462-5. doi: 10.1126/science.147.3664.1462. PMID: 14263761
107. Holtgrewe M., Emde A.K., Weese D., Reinert K. (2011). A novel and well-defined benchmarking method for second generation read mapping. BMC Bioinformatics 12. doi:10.1186/1471-2105-12-210
108. Hotaling S., Kelley J.L., Frandsen P.B. (2021). Toward a genome sequence for every animal: Where are we now? Proc Natl Acad Sci 118 (52): 1–8. doi:10.1073/pnas.2109019118
109. Howe K., Clark M.D., Torroja C.F., Torrance J., Berthelot C., Muffato M., Collins J.E., Humphray S., McLaren K., Matthews L., McLaren S., Sealy I., Caccamo M., Churcher C., Scott C., Barrett J.C., Koch R., Rauch G.J., White S., Chow W., Kilian B., Quintais L.T., Guerra-Assunção J.A., Zhou Y., Gu Y., Yen J., Vogel J.H., Eyre T., Redmond S., Banerjee R., Chi J., Fu B., Langley E., Maguire S.F., Laird G.K., Lloyd D., Kenyon E., Donaldson S., Sehra H., Almeida-King J., Loveland J., Trevanion S., Jones M., Quail M., Willey D., Hunt A., Burton J., Sims S., McLay K., Plumb B., Davis J., Clee C., Oliver K., Clark R., Riddle C., Eliott D., Threadgold G., Harden G., Ware D., Mortimer B., Kerry G., Heath P., Phillimore B., Tracey A., Corby N., Dunn M., Johnson C., Wood J., Clark S., Pelan S., Griffiths G., Smith M., Glithero R., Howden P., Barker N., Stevens C., Harley J., Holt K., Panagiotidis G., Lovell J., Beasley H., Henderson C., Gordon D., Auger et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature 496 (7446): 498–503. doi:10.1038/nature12111
110. Hu X. Di, Li K., Gao L.Z. (2016). The complete mitochondrial genome of Celebes wild boar, *sus celebensis* (*Cetartiodactyla*: *Suina*: *Suidae*), and comparative mitochondrial genomics of the Sus species. Mitochondrial DNA 27 (2): 1476–1477. doi:10.3109/19401736.2014.953099
111. Hu X. Di, Yang X.T., En-Yang. (2015). The complete mitochondrial genome of European wild boar, *Sus scrofa scrofa*. Mitochondrial DNA 27 (5): 3244–3245. doi:10.3109/19401736.2015.1007366

112. Huang L., Popic V., Batzoglou S. (2013). Short read alignment with populations of genomes. Bioinformatics 29 (13). doi:10.1093/bioinformatics/btt215
113. Huang X., Yang S.P. (2005). Generating a genome assembly with PCAP. Curr Protoc Bioinformatics Chapter 11: 1–23. doi:10.1002/0471250953.bi1103s11
114. Huelsenbeck J.P., Ronquist F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17 (8): 754–755. doi:10.1093/bioinformatics/17.8.754
115. Hulsen T., Huynen M.A., de Vlieg J., Groenen P.M.A. (2006). Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7 (4). doi:10.1186/gb-2006-7-4-r31
116. Hutchinson C.A., Newbold J.E., Potter S.S., Edgell M.H. (1974). Maternal Inheritance of Mammalian Mitochondrial DNA. Nature 251(5475): 536-538. doi: 10.1038/251536a0
117. Iacolina L., Buzan E., Safner T., Bašić N., Geric U., Tesija T., Lazar P., Arnal M.C., Chen J., Han J., Šprem N. (2021). A mother's story, mitogenome relationships in the genus *Rupicapra*. Animals 11 (4): 1–11. doi:10.3390/ani11041065
118. Idury R.M., Waterman M.S. (1995). A new algorithm for DNA sequence assembly. J Comput Biol. 2(2):291-306. doi: 10.1089/cmb.1995.2.291
119. Igea J., Juste J., Castresana J. (2010). Novel intron markers to study the phylogeny of closely related mammalian species. BMC Evol Biol 10 (1): 369. doi:10.1186/1471-2148-10-369
120. Ingman M., Kaessmann H., Pääbo S., Gyllensten U. (2000). Mitochondrial genome variation and the origin of modem humans. Nature 408 (6813): 708–713. doi:10.1038/35047064
121. International Human Genome Sequencing Consortium (IHGSC). (2001). Nature 412 (6846): 565–566. doi:10.1038/35087627
122. Irwin D.M., Kocher T.D., Wilson A.C. (1991). Evolution of the Cytb Gene of Mammals.Pdf 128–144
123. Jang K.H., Hwang U.W. (2010). Complete mitochondrial genome of the Korean goral *Naemorhaedus caudatus* (*Ruminantia*, *Bovidae*, *Antilopinae*) and conserved domains in the control region of Caprini. Mitochondrial DNA 21 (3–4): 62–64. doi:10.3109/19401736.2010.490833

124. Jauhal A.A., Newcomb R.D. (2021). Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour 21 (5): 1416–1421. doi:10.1111/1755-0998.13364
125. Jia P., Li F., Xia J., Chen H., Ji H., Pao W., Zhao Z. (2012). Consensus rules in variant detection from next-generation sequencing data. PLoS One 7 (6). doi:10.1371/journal.pone.0038470
126. Jiang L., Wang G., Tan S., Gong S., Yang M., Peng Q., Peng R., Zou F. (2013). The complete mitochondrial genome sequence analysis of Tibetan argali (*Ovis ammon hodgsoni*): Implications of Tibetan argali and Gansu argali as the same subspecies. Gene 521 (1): 24–31. doi:10.1016/j.gene.2013.03.049
127. Kalbfleisch T., Heaton M.P. (2014). Mapping whole genome shotgun sequence and variant calling in mammalian species without their reference genomes. F1000Research 2: 1–12. doi:10.12688/f1000research.2-244.v2
128. Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., Von Haeseler A., Jermiin L.S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods 14 (6): 587–589. doi:10.1038/nmeth.4285
129. Kapli P., Yang Z., Telford M.J. (2020). Phylogenetic tree building in the genomic age. Nat Rev Genet 21 (7): 428–444. doi:10.1038/s41576-020-0233-0
130. Kazazian H.H. (2004). Mobile Elements: Drivers of Genome Evolution. Science (80-) 303 (5664): 1626–1632. doi:10.1126/science.1089670
131. Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647–1649. doi:10.1093/bioinformatics/bts199
132. Kim J., Larkin D.M., Cai Q., Asan, Zhang Y., Ge R.L., Auvil L., Capitanu B., Zhang G., Lewin H.A., Ma J. (2013). Reference-assisted chromosome assembly. Proc Natl Acad Sci U S A 110 (5): 1785–1790. doi:10.1073/pnas.1220349110
133. Kim K.M., Park J.H., Bhattacharya D., Yoon H.S. (2014). Applications of next-generation sequencing to unravelling the evolutionary history of algae. Int J Syst Evol Microbiol 64 (PART 2): 333–345. doi:10.1099/ijs.0.054221-0
134. Kolmogorov M., Armstrong J., Raney B.J., Streeter I., Dunn M., Yang F., Odom D., Flicek P., Keane T.M., Thybert D., Paten B., Pham S. (2018). Chromosome assembly

of large and complex genomes using multiple references. Genome Res 28 (11): 1720–1732. doi:10.1101/gr.236273.118

135. Kolmogorov M., Raney B., Paten B., Pham S. (2014). Ragout - A reference-assisted assembly tool for bacterial genomes. Bioinformatics 30 (12): 302–309. doi:10.1093/bioinformatics/btu280
136. Korf I. (2004). Gene Finding in Novel Genomes. BMC Bioinformatics 5: 1–9. doi:10.1186/1471-2105-5-59
137. Kumar A., Gautam K.B., Singh B., Yadav P., Gopi G.V., Gupta S.K. (2019). Sequencing and characterization of the complete mitochondrial genome of Mishmi takin (*Budorcas taxicolor taxicolor*) and comparison with the other *Caprinae* species. Int J Biol Macromol 137: 87–94. doi:10.1016/j.ijbiomac.2019.06.201
138. Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. (2012). Statistics and truth in phylogenomics. Mol Biol Evol 29 (2): 457–472. doi:10.1093/molbev/msr202
139. Kumar S., Stecher G., Li M., Knyaz C., Tamura K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35 (6): 1547–1549. doi:10.1093/molbev/msy096
140. Kurtén B. (1968). Pleistocene Mammals of Europe. Weidenfeld & Nicholson, London, UK
141. Kuwayama R., Ozawa T. (2000). Phylogenetic relationships among European red deer, wapiti, and sika deer inferred from mitochondrial DNA sequences. Mol Phylogenet Evol 15 (1): 115–123. doi:10.1006/mpev.1999.0731
142. Langmead B., Salzberg S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (4): 357–359. doi:10.1038/nmeth.1923
143. Langmead B., Trapnell C., Pop M., Salzberg S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10 (3). doi:10.1186/gb-2009-10-3-r25
144. Leugger F., Broquet T., Karger D.N., Rioux D., Buzan E., Corlatti L., Crestanello B., Curt-Grand-Gaudin N., Hauffe H.C., Rolečková B., Šprem N., Tissot N., Tissot S., Valterová R., Yannic G., Pellissier L. (2022). Dispersal and habitat dynamics shape the genetic structure of the Northern chamois in the Alps. J Biogeogr (November 2021): 1–14. doi:10.1111/jbi.14363

145. Lewin H.A., Robinson G.E., Kress W.J., Baker W.J., Coddington J., Crandall K.A., Durbin R., Edwards S. V., Forest F., Gilbert M.T.P., Goldstein M.M., Grigoriev I. V., Hackett K.J., Haussler D., Jarvis E.D., Johnson W.E., Patrinos A., Richards S., Castilla-Rubio J.C., Van Sluys M.A., Soltis P.S., Xu X., Yang H., Zhang G. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A 115 (17): 4325–4333. doi:10.1073/pnas.1720115115

146. Li H., Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (14): 1754–1760. doi:10.1093/bioinformatics/btp324

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

148. Li H., Homer N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11 (5): 473–483. doi:10.1093/bib/bbq015

149. Li R., Fan W., Tian G., Zhu H., He L., Cai J., Huang Q., Cai Q., Li B., Bai Y., Zhang Zhihe, Zhang Y., Wang W., Li Jun, Wei F., Li H., Jian M., Li Jianwen, Zhang Zhaolei, Nielsen R., Li Dawei, Gu W., Yang Z., Xuan Z., Ryder O.A., Leung F.C.C., Zhou Y., Cao J., Sun X., Fu Y., Fang X., Guo X., Wang B., Hou R., Shen F., Mu B., Ni P., Lin R., Qian W., Wang G., Yu C., Nie W., Wang Jinhuan, Wu Z., Liang H., Min J., Wu Q., Cheng S., Ruan J., Wang M., Shi Z., Wen M., Liu B., Ren X., et al. (2010a). The sequence and de novo assembly of the giant panda genome. Nature 463 (7279): 311–317. doi:10.1038/nature08696

150. Li R., Yang P., Dai X., Asadollahpour N.H., Fang W., Yang Z., Cai Y., Zheng Z., Wang X., Jiang Y. (2021). A near complete genome for goat genetic and genomic research. Genet Sel Evol 53 (1): 1–17. doi:10.1186/s12711-021-00668-5

151. Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li Shengting, Shan G., Kristiansen K., Li Songgang, Yang H., Wang Jian, Wang Jun. (2010b). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20 (2): 265–272. doi:10.1101/gr.097261.109

152. Liao X., Li M., Zou Y., Wu F.X., Yi-Pan, Wang J. (2019). Current challenges and solutions of de novo assembly. Quant Biol 7 (2): 90–109. doi:10.1007/s40484-019-0166-9

153. Lischer H.E.L., Shimizu K.K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18 (1): 1–12. doi:10.1186/s12859-017-1911-6
154. Liu X., Han S., Wang Z., Gelernter J., Yang B.Z. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. PLoS One 8 (9): 1–11. doi:10.1371/journal.pone.0075619
155. Lovari S. (1985). Behavioural repertoire of the Abruzzo chamois *Rupicapra pyrenaica ornata*. Saugetierkundliche Mitteilungen 32: 113–136
156. Lovari S., Scala C. (1980). Revision of rupicapra genus. I. a statistical re-evaluation of couturier's data on the morphometry of six chamois subspecies. Bolletino di Zool 47 (1–2): 113–124. doi:10.1080/11250008009440328
157. Lowe T.M., Chan P.P. (2016). tRNAscan-SE On-line: Integrating Search and Context for Analysis of Transfer RNA Genes. Nucleic Acids Res 44 (W1): W54–W57. doi:10.1093/nar/gkw413
158. Lowe T.M., Eddy S.R. (1997). TRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Res 25 (5): 955–964. doi:10.1093/nar/25.5.0955
159. Lydekker R. (1913). Catalogue of the Ungulate Mammals of British Museum (Natural History). British Museum, London, UK
160. Magee A.F., May M.R., Moore B.R. (2014). The dawn of open access to phylogenetic data. PLoS One 9 (10). doi:10.1371/journal.pone.0110268
161. Manee M.M., Alshehri M.A., Binghadir S.A., Aldhafer S.H., Alswailem R.M., Algarni A.T., AL-Shomrani B.M., AL-Fageeh M.B. (2019). Comparative analysis of camelid mitochondrial genomes. J Genet 98 (3). doi:10.1007/s12041-019-1134-x
162. Manni M., Berkeley M.R., Seppey M., Simão F.A., Zdobnov E.M. (2021a). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol 38 (10): 4647–4654. doi:10.1093/molbev/msab199
163. Manni M., Berkeley M.R., Seppey M., Zdobnov E.M. (2021b). BUSCO: Assessing Genomic Data Quality and Beyond. Curr Protoc 1 (12): 1–41. doi:10.1002/cpz1.323
164. Mao H., Liu H., Ma G., Yang Q., Guo X., Lamaocao Z. (2017). The complete mitochondrial genome of *Ovis ammon darwini* (*Artiodactyla*: *Bovidae*). Conserv Genet Resour 9 (1): 59–62. doi:10.1007/s12686-016-0620-1

165. Martchenko D., Chikhi R., Shafer A.B.A. (2020). Genome assembly and analysis of the north American mountain goat (*Oreamnos americanus*) reveals species-level responses to extreme environments. G3 Genes, Genomes, Genet 10 (2): 437–442. doi:10.1534/g3.119.400747
166. Martin J.A., Wang Z. (2011). Next-generation transcriptome assembly. Nat Rev Genet 12 (10): 671–682. doi:10.1038/nrg3068
167. Masini F., Lovari S. (1988). Systematics, phylogenetic relationships, and dispersal of the chamois (*Rupicapra* spp.). Quat Res 30 (3): 339–349. doi:10.1016/0033-5894(88)90009-9
168. Matosiuk M., Sheremetyeva I.N., Sheremetyev I.S., Saveljev A.P., Borkowska A. (2014). Evolutionary neutrality of mtDNA introgression: Evidence from complete mitogenome analysis in roe deer. J Evol Biol 27 (11): 2483–2494. doi:10.1111/jeb.12491
169. McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol 66 (2): 526–538. doi:10.1016/j.ympev.2011.12.007
170. McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297-303. doi: 10.1101/gr.107524.110
171. McMahon B.J., Teeling E.C., Höglund J. (2014). How and why should we implement genomics into conservation? Evol Appl 7 (9): 999–1007. doi:10.1111/eva.12193
172. Meng G., Li Y., Yang C., Liu S. (2019). MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res 47 (11). doi:10.1093/nar/gkz173
173. Mereu P., Palici Di Suni M., Manca L., Masala B. (2008). Complete nucleotide mtDNA sequence of Barbary sheep (*Ammotragus lervia*). DNA Seq - J DNA Seq Mapp 19 (3): 241–245. doi:10.1080/10425170701550599
174. Miller J.M., Malenfant R.M., Moore S.S., Coltman D.W. (2012). Short reads, circular genome: Skimming solid sequence to construct the bighorn sheep mitochondrial genome. J Hered 103 (1): 140–146. doi:10.1093/jhered/esr104

175. Miller K.W.P., Dawson J.L., Hagelberg E. (1996). A concordance of nucleotide substitutions in the first and second hypervariable segments of the human mtDNA control region. Int J Legal Med 109 (3): 107–113. doi:10.1007/BF01369668
176. Milne I., Bayer M., Cardle L., Shaw P., Stephen G., Wright F., Marshall D. (2009). Tablet-next generation sequence assembly visualization. Bioinformatics 26 (3): 401–402. doi:10.1093/bioinformatics/btp666
177. Mishmar D., Ruiz-Pesini E., Golik P., Macaulay V., Clark A.G., Hosseini S., Brandon M., Easleyf K., Chen E., Brown M.D., Sukernik R.I., Olckers A., Wallace D.C. (2003). Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100 (1): 171–176. doi:10.1073/pnas.0136972100
178. Miyamoto M.M., Tanhauser S.M., Laipis P.J. (1989). Systematic Relationships in the Artiodactyl Tribe Bovini (Family *Bovidae*), as Determined from Mitochondrial DNA Sequences. Syst Zool 38 (4): 342. doi:10.2307/2992400
179. Mohandesan E., Fitak R.R., Corander J., Yadamsuren A., Chuluunbat B., Abdelhadi O., Raziq A., Nagy P., Stalder G., Walzer C., Faye B., Burger P.A. (2017). Mitogenome Sequencing in the Genus *Camelus* Reveals Evidence for Purifying Selection and Long-term Divergence between Wild and Domestic Bactrian Camels. Sci Rep 7 (1): 1–12. doi:10.1038/s41598-017-08995-8
180. Montgelard C., Catzeflis F.M., Douzery E. (1997). Phylogenetic relationships of Artiodactyls and Cetaceans as deduced from the comparison of cytochrome b and 12S rRNA mitochondrial sequences. Mol Biol Evol 14 (5): 550–559. doi:10.1093/oxfordjournals.molbev.a025792
181. Mouse Genome Sequencing Consortium (MGSC). (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420 (6915): 520–562. doi: 10.1038/nature01262
182. Murray J.C., Buetow K.H., Weber J.L., Ludwigsen S., Scherpbier-Heddema T., Manion F., Quillen J., Sheffield V.C., Sunden S., Duyk G.M., Weissenbach j., Gyapay G., Dib C., Morrissette J., Lathrop G.M., Vignal A., White R., Matsunami N., Gerken S., Melis R., Albertsen H., Plaetke R., Odelberg S., Ward D., Dausset J., Cohen D., Cann H. (1994). A comprehensive human linkage map with centimorgan density. Science 265(5181):2049-54. doi: 10.1126/science.8091227
183. Myers E.W., Sutton G.G., Delcher A.L., Dew I.M., Fasulo D.P., Flanigan M.J., Kravitz S.A., Mobarry C.M., Reinert K.H., Remington K.A., Anson E.L., Bolanos R.A., Chou

H.H., Jordan C.M., Halpern A.L., Lonardi S., Beasley E.M., Brandon R.C., Chen L., Dunn P.J., Lai Z., Liang Y., Nusskern D.R., Zhan M., Zhang Q., Zheng X., Rubin G.M., Adams M.D., Venter J.C. (2000). A whole-genome assembly of *Drosophila*. Science 287(5461):2196-204. doi: 10.1126/science.287.5461.2196. PMID: 10731133

184. Nabhan A.R., Sarkar I.N. (2012). The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy. Brief Bioinform 13 (1): 122–134. doi:10.1093/bib/bbr014
185. Nascetti G., Lovari S., Lanfranchi P., Berducou C., Mattiucci S., Rossi L., Bullini L. (1985). Revision of Rupicapra genus. III. Electrophoretic studies demonstrating species distinction of chamois populations of the Alps from those of the Apennines and Pyrenees. In: Lovari S. (ed.) Biology and Management of Mountain Ungulates, 56–62. Croom-Helm, London, UK
186. Naseem A., Batool S., Abbas F. i. (2020). Utility of mitochondrial COI gene for identification of wild ungulate species of conservational importance from Pakistan. Mitochondrial DNA Part B Resour 5 (2): 1924–1928. doi:10.1080/23802359.2020.1756476
187. Nass M., Nass S.A. (1963). Intramitochondrial fibers with dna characteristics : I. Fixation and Electron Staining Reactions. J Cell Biol (1963) 19 (3): 593–611. https://doi.org/10.1083/jcb.19.3.593
188. Nei M., Zhang J. (2006). Evolutionary Distance: Estimation. eLS 1–4. doi:10.1038/npg.els.0005108
189. Neumann O. (1899). Die Gemse der Abruzzen. Annali del Museo Civico di Storia Naturale di Genova 2: 40–44
190. Nevado B., Ramos-Onsins S.E., Perez-Enciso M. (2014). Resequencing studies of non-model organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics. Mol Ecol 23 (7): 1764–1779. doi:10.1111/mec.12693
191. Nielsen R., Paul J.S., Albrechtsen A., Song Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12 (6): 443–451. doi:10.1038/nrg2986
192. Nikaido M., Rooney A.P., Okada N. (1999). Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements:

Hippopotamuses are the closest extant relatives of whales. Proc Natl Acad Sci U S A 96 (18): 10261–10266. doi:10.1073/pnas.96.18.10261

193. O'Rawe J., Jiang T., Sun G., Wu Y., Wang W., Hu J., Bodily P., Tian L., Hakonarson H., Johnson W.E., Wei Z., Wang K., Lyon G.J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med 5 (3). doi:10.1186/gm432
194. Oksanen J., Blanchet F.G., Friendly M., Kindt R., Legendre P., McGlinn D., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P., Stevens M.H.H., Szoecs E., Wagner. H. (2020). vegan: Community Ecology Package. R package, version 2.5-7. https://CRAN.R-project.org/package=vegan
195. Papaioannou H., Fernández M., Pérez T., Domínguez A. (2019). Genetic variability and population structure of chamois in Greece (*Rupicapra rupicapra balcanica*). Conserv Genet 20 (4): 939–945. doi:10.1007/s10592-019-01177-1
196. Parma P., Feligini M., Greppi G., Giuseppe E. (2003). The complete nucleotide sequence of goat (*Capra hircus*) mitochondrial genome goat mitochondrial genome. DNA Seq - J DNA Seq Mapp 14 (3): 199–203. doi:10.1080/1042517031000089487
197. Patrushev M. V., Kamenski P.A., Mazunin I.O. (2014). Mutations in mitochondrial DNA and approaches for their correction. Biochem 79 (11): 1151–1160. doi:10.1134/S0006297914110029
198. Pele J., Becu J.M., Baubaker R.B., Abdi H., Chabbert M. (2020). Bios2mds: From Biological Sequences to Multidimensional Scaling. R package version 1.2.3. https://CRAN.R-project.org/package=bios2mds
199. Pérez T., Albornoz J., Garcia-Vazquez E., Domínguez A. (1996). Application of DNA fingerprinting to population study of chamois (*Rupicapra rupicapra*). Biochem Genet 34 (7–8): 313–320. doi:10.1007/BF02399950
200. Pérez T., Fernández M., Hammer S.E., Domínguez A. (2017). Multilocus intron trees reveal extensive male-biased homogenization of ancient populations of chamois (*Rupicapra* spp.) across Europe during late Pleistocene. PLoS One 12 (2): 1–21. doi:10.1371/journal.pone.0170392
201. Pérez T., González I., Essler S.E., Fernández M., Domínguez A. (2014). The shared mitochondrial genome of *Rupicapra pyrenaica ornata* and *Rupicapra rupicapra cartusiana*: Old remains of a common past. Mol Phylogenet Evol 79 (1): 375–379. doi:10.1016/j.ympev.2014.07.004

202. Pérez T., Hammer S.E., Albornoz J., Domínguez A. (2011). Y-chromosome phylogeny in the evolutionary net of chamois (genus *Rupicapra*). BMC Evol Biol 11 (1): 1–12. doi:10.1186/1471-2148-11-272
203. Pfeifer S.P. (2017). From next-generation resequencing reads to a high-quality variant data set. Heredity (Edinb) 118 (2): 111–124. doi:10.1038/hdy.2016.102
204. Philippe H., Telford M.J. (2006). Large-scale sequencing and the new animal phylogeny. Trends Ecol Evol 21 (11): 614–620. doi:10.1016/j.tree.2006.08.004
205. Polziehn R.O., Strobeck C. (2002). A phylogenetic comparison of red deer and wapiti using mitochondrial DNA. Mol Phylogenet Evol 22 (3): 342–356. doi:10.1006/mpev.2001.1065
206. Prada C.F., Boore J.L. (2019). Gene annotation errors are common in the mammalian mitochondrial genomes database. BMC Genomics 20 (1): 1–8. doi:10.1186/s12864-019-5447-1
207. Pramod R.K., Velayutham D., Sajesh P.K., Beena P.S., Zachariah Anil, Zachariah Arun, Chandramohan B., Sujith S.S., Santhosh S., Iype S., Ganapathi P., Dhinoth Kumar B., Gupta R., Thomas G. (2018). The complete mitochondrial genome of Indian cattle (*Bos indicus*). Mitochondrial DNA Part B Resour 3 (1): 207–208. doi:10.1080/23802359.2018.1437836
208. Prasad A., Lorenzen E.D., Westbury M. V. (2021). Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. Mol Ecol Resour (March): 1–11. doi:10.1111/1755-0998.13457
209. R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/
210. Randi E., Mucci N., Pierpaoli M., Douzery E. (1998). New phylogenetic perspectives on the Cervidae (*Artiodactyla*) are provided by the mitochondrial cytochrome b gene. Proc R Soc Lond 265 (January): 793–801
211. Rat Genome Sequencing Project Consortium (RGSPC). (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428 (6982): 493–521
212. Reyes A., Gissi C., Pesole G., Saccone C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol 15 (8): 957–966. doi:10.1093/oxfordjournals.molbev.a026011

213. Rezić A., Iacolina L., Bužan E., Safner T., Bego F., Gačić D., Maletić V., Markov G., Milošević D., Papaioannou H., Šprem N. (2022). The Balkan chamois, an archipelago or a peninsula? Insights from nuclear and mitochondrial DNA. Conserv Genet (0123456789). doi:10.1007/s10592-022-01434-w

214. Rhie A., McCarthy S.A., Fedrigo O., Damas J., Formenti G., Koren S., Uliano-Silva M., Chow W., Fungtammasan A., Kim J., Lee C., Ko B.J., Chaisson M., Gedman G.L., Cantin L.J., Thibaud-Nissen F., Haggerty L., Bista I., Smith M., Haase B., Mountcastle J., Winkler S., Paez S., Howard J., Vernes S.C., Lama T.M., Grutzner F., Warren W.C., Balakrishnan C.N., Burt D., George J.M., Biegler M.T., Iorns D., Digby A., Eason D., Robertson B., Edwards T., Wilkinson M., Turner G., Meyer A., Kautt A.F., Franchini P., Detrich H.W., Svardal H., Wagner M., Naylor G.J.P., Pippel M., Malinsky M., Mooney M., Simbirsky M., Hannigan B.T., Pesout T., Houck M., Misuraca A., Kingan S.B., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature 592 (7856): 737–746. doi:10.1038/s41586-021-03451-0

215. Rice E.S., Green R.E. (2019). New Approaches for Genome Assembly and Scaffolding. Annu Rev Anim Biosci 7: 17–40. doi:10.1146/annurev-animal-020518-115344

216. Robin E.D., Wong R. (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol 136 (3): 507–513. doi:10.1002/jcp.1041360316

217. Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G., Mesirov J.P. (2011). Integrative Genome Viewer. Nat Biotechnol 29 (1): 24–6. doi:10.1038/nbt.1754.Integrative

218. Rodríguez F., Hammer S., Pérez T., Suchentrunk F., Lorenzini R., Michallet J., Martinkova N., Albornoz J., Domínguez A. (2009). Cytochrome b phylogeography of chamois (*Rupicapra* spp.). Population contractions, expansions and hybridizations governed the diversification of the genus. J Hered 100 (1): 47–55. doi:10.1093/jhered/esn074

219. Rodríguez F., Pérez T., Hammer S.E., Albornoz J., Domínguez A. (2010). Integrating phylogeographic patterns of microsatellite and mtDNA divergence to infer the evolutionary history of chamois (genus *Rupicapra*). BMC Evol Biol 10 (1). doi:10.1186/1471-2148-10-222

220. Ronquist F., Teslenko M., Van Der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61 (3): 539–542. doi:10.1093/sysbio/sys029
221. Rosenberg M.S., Kumar S. (2003). Taxon Sampling, Bioinformatics, and Phylogenomics. Syst Biol 52 (1): 119–124. doi:10.1080/10635150309344
222. Roth S.C. (2019). What is genomic medicine? J Med Libr Assoc 107 (3): 442–448. doi:10.5195/jmla.2019.604
223. Rozas J., Ferrer-Mata A., Sanchez-DelBarrio J.C., Guirao-Rico S., Librado P., Ramos-Onsins S.E., Sanchez-Gracia A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34 (12): 3299–3302. doi:10.1093/molbev/msx248
224. Ruffalo M., Koyutürk M., Ray S., LaFramboise T. (2012). Accurate estimation of short read mapping quality for next-generation genome sequencing. Bioinformatics 28 (18): 349–355. doi:10.1093/bioinformatics/bts408
225. Safner T., Buzan E., Rezic A., Šprem N. (2019). Small-scale spatial genetic structure of Alpine chamois (*Rupicapra rupicapra*) in Northern Dinarides. Eur J Wildl Res 65 (2). doi:10.1007/s10344-019-1259-5
226. Sanger F., Coulson A.R., Hong G.F., Hill C., Petersen G.B. (1982). Nucleotide sequence of bacteriophage λ DNA. J Mol Biol 162 (4): 729–773
227. Sanger F., Nicklen S., Coulson A.R. (1977a). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74(12):5463-7. doi:10.1073/pnas.74.12.5463
228. Sanger F., Thompson E.O. (1953a). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J 53 (3): 353–366. doi:10.1042/bj0530353
229. Sanger F., Thompson E.O. (1953b). The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem J 53 (3): 366–374. doi:10.1042/bj0530366
230. Sanger F.,Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes J.C., Hutchison C.A., Slocombe P.M., Smith M. (1977b). Nucleotide sequence of bacteriophage φ X174 DNA. Nature 265, 687–695. https://doi.org/10.1038/265687a0

231. Schaschl H., Suchentrunk F., Morris D.L., Slimen H. Ben, Smith S., Arnold W. (2012). Sex-specific selection for MHC variability in Alpine chamois. BMC Evol Biol 12 (1). doi:10.1186/1471-2148-12-20

232. Schatz M.C., Delcher A.L., Salzberg S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome Res 20 (9): 1165–1173. doi:10.1101/gr.101360.109

233. Schbath S., Martin V., Zytnicki M., Fayolle J., Loux V., Gibrat J.F. (2012). Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. J Comput Biol 19 (6): 796–813. doi:10.1089/cmb.2012.0022

234. Schneider V.A., Graves-Lindsay T., Howe K., Bouk N., Chen H.C., Kitts P.A., Murphy T.D., Pruitt K.D., Thibaud-Nissen F., Albracht D., Fulton R.S., Kremitzki M., Magrini V., Markovic C., McGrath S., Steinberg K.M., Auger K., Chow W., Collins J., Harden G., Hubbard T., Pelan S., Simpson J.T., Threadgold G., Torrance J., Wood J.M., Clarke L., Koren S., Boitano M., Peluso P., Li H., Chin C.S., Phillippy A.M., Durbin R., Wilson R.K., Flicek P., Eichler E.E., Church D.M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 27 (5): 849–864. doi:10.1101/gr.213611.116

235. Shen W., Le S., Li Y., Hu F. (2016b). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11 (10): 1–10. doi:10.1371/journal.pone.0163962

236. Shen X.X., Steenwyk J.L., LaBella A.L., Opulente D.A., Zhou X., Kominek J., Li Y., Groenewald M., Hittinger C.T., Rokas A. (2020). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum *Ascomycota*. Sci Adv 6 (45). doi:10.1126/SCIADV.ABD0079

237. Shen X.X., Zhou X., Kominek J., Kurtzman C.P., Hittinger C.T., Rokas A. (2016a). Reconstructing the backbone of the saccharomycotina yeast phylogeny using genome-scale data. G3 Genes, Genomes, Genet 6 (12): 3927–3939. doi:10.1534/g3.116.034744

238. Siddiki A., Alam M., Shawrob K., Rahman A.H., Hossain M.A., Mollah A., Islam M.S., Khan M. (2019). First report of reference guided genome assembly of Black Bengal goat (*Capra hircus*). bioRxiv 603266. doi:10.1101/603266

239. Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E. V., Zdobnov E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19): 3210–3212. doi:10.1093/bioinformatics/btv351

240. Simpson J.T., Pop M. (2015). The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet 16: 153–172. doi:10.1146/annurev-genom-090314-050032

241. Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. (2009). ABySS: A parallel assembler for short read sequence data. Genome Res 19 (6): 1117–1123. doi:10.1101/gr.089532.108

242. Sjölander K. (2004). Phylogenomic inference of protein molecular function: Advances and challenges. Bioinformatics 20 (2): 170–179. doi:10.1093/bioinformatics/bth021

243. Smith D.R. (2016). The past, present and future of mitochondrial genomics: Have we sequenced enough mtDNAs? Brief Funct Genomics 15 (1): 47–54. doi:10.1093/bfgp/elv027

244. Smith K. (2013). A Brief History of NCBI's Formation and Growth. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK148949/

245. Smolka M., Rescheneder P., Schatz M.C., von Haeseler A., Sedlazeck F.J. (2015). Teaser: Individualized benchmarking and optimization of read mapping results for NGS data. Genome Biol 16 (1): 1–10. doi:10.1186/s13059-015-0803-1

246. Soderlund C., Bomhoff M., Nelson W.M. (2011). SyMAP v3.4: A turnkey synteny system with application to plant genomes. Nucleic Acids Res 39 (10). doi:10.1093/nar/gkr123

247. Soglia D., Rossi L., Cauvin E., Citterio C., Ferroglio E., Maione S., Meneguz P.G., Spalenza V., Rasero R., Sacchi P. (2010). Population genetic structure of Alpine chamois (*Rupicapra r. rupicapra*) in the Italian Alps. Eur J Wildl Res 56 (6): 845–854. doi:10.1007/s10344-010-0382-0

248. Song H.J., Lee J.M., Graf L., Rho M., Qiu H., Bhattacharya D., Yoon H.S. (2016). A novice's guide to analyzing NGS-derived organelle and metagenome data. Algae 31 (2): 137–154. doi:10.4490/algae.2016.31.6.5

249. Sousa V., Hey J. (2013). Understanding the origin of species with genome-scale data: Modelling gene flow. Nat Rev Genet 14 (6): 404–414. doi:10.1038/nrg3446

250. Staden R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Res, 6(7), 2601–2610. doi:https://doi.org/10.1093/nar/6.7.2601
251. Stanke M., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B. (2006). AUGUSTUS: A b initio prediction of alternative transcripts. Nucleic Acids Res 34 (WEB. SERV. ISS.): 435–439. doi:10.1093/nar/gkl200
252. Steiner C.C., Charter S.J., Houck M.L., Ryder O.A. (2014). Molecular phylogeny and chromosomal evolution of *alcelaphini* (*antilopinae*). J Hered 105 (3): 324–333. doi:10.1093/jhered/esu004
253. Sun X., Ding Y., Orr M.C., Zhang F. (2020). Streamlining universal single-copy orthologue and ultraconserved element design: A case study in Collembola. Mol Ecol Resour 20 (3): 706–717. doi:10.1111/1755-0998.13146
254. Šprem N., Bužan E. (2016). The genetic impact of chamois management in the dinarides. J Wildl Manag 80(5),783-793. doi: https://doi.org/10.1002/jwmg.21081
255. Świsłocka, M., Matosiuk, M., Ratkiewicz, M., Borkowska, A., Czajkowska, M., Mackiewicz, P. (2020). Phylogeny and diversity of moose (*Alces alces*, *Cervidae*, *Mammalia*) revealed by complete mitochondrial genomes. Hystrix, Ital Jour of Mammal 31(1), 1-9. https://doi.org/10.4404/hystrix-00252-2019
256. Taanman J.W. (1999). The mitochondrial genome: Structure, transcription, translation and replication. Biochim Biophys Acta - Bioenerg 1410 (2): 103–123. doi:10.1016/S0005-2728(98)00161-3
257. Tajima F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics 105 (2): 437–460. doi:10.1093/genetics/105.2.437
258. Tamazian G., Dobrynin P., Krasheninnikova K., Komissarov A., Koepfli K.P., O'Brien S.J. (2016). Chromosomer: A reference-based genome arrangement tool for producing draft chromosome sequences. Gigascience 5 (1): 1–11. doi:10.1186/s13742-016-0141-6
259. Tešija T., Safner T. (2021). Analyses of wild ungulates mitogenome. Agric Conspec Sci 86 (1): 1–12
260. The Chimpanzee Sequencing and Analysis Consortium (CSAC). (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437 (7055): 69–87. doi:10.1038/nature04072

261. Tillich M., Lehwark P., Pellizzer T., Ulbricht-Jones E.S., Fischer A., Bock R., Greiner S. (2017). GeSeq - Versatile and Accurate Annotation of Organelle Genomes. Nucleic Acids Res 45 (W1): W6–W11. doi:10.1093/nar/gkx391

262. Tsai I.J., Otto T.D., Berriman M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol 11 (4). doi:10.1186/gb-2010-11-4-r41

263. Ursing B.M., Arnason U. (1998). Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. Proc R Soc B Biol Sci 265 (1412): 2251–2255. doi:10.1098/rspb.1998.0567

264. Ursing B.M., Slack K.E., Arnason U. (2000). Subordinal artiodactyl relationships in the light of phylogenetic analysis of 12 mitochondrial protein-coding genes. Zool Scr 29 (2): 83–88. doi:10.1046/j.1463-6409.2000.00037.x

265. Valiente-Mullor C., Beamud B., Ansari I., Frances-Cuesta C., Garcia-Gonzalez N., Mejia L., Ruiz-Hueso P., Gonzalez-Candelas F. (2021). One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput Biol 17 (1). doi:10.1371/JOURNAL.PCBI.1008678

266. Venter J., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., Nelson C., Broder S., Clark A.G., Nadeau J., McKusick V.A., Zinder N., Levine A.J., Roberts R.J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K. (2001). The sequence of the human genome. Science 80-291 (5507): 1304–1351. doi:10.1126/science.1058040

267. Vezzi F., Cattonaro F., Policriti A. (2011). e-RGA: enhanced Reference Guided Assembly of Complex Genomes. EMBnet.journal 17 (1): 46. doi:10.14806/ej.17.1.208

268. Wallace D.C. (2007). Why Do We Still Have a Maternally Inherited Mitochondrial DNA? Insights from Evolutionary Medicine. Annu Rev Biochem 76 (1): 781–821. doi:10.1146/annurev.biochem.76.081205.150955

269. Wang B., Ekblom R., Bunikis I., Siitari H., Höglund J. (2014). Whole genome sequencing of the black grouse (Tetrao tetrix): Reference guided assembly suggests faster-Z and MHC evolution. BMC Genomics 15 (1). doi:10.1186/1471-2164-15-180

270. Wang Z., Chen Y., Li Y. (2004). A Brief Review of Computational Gene Prediction Methods. Genomics, Proteomics Bioinforma / Beijing Genomics Inst 2 (4): 216–221. doi:10.1016/S1672-0229(04)02028-5
271. Wang, Y., Zhang, R., Ma, Y., Li, J., Fan, F., Liu, X., & Yang, D. (2021). Low-Coverage Whole Genomes Reveal the Higher Phylogeny of Green Lacewings. Insects, 12(10), 857. https://doi.org/10.3390/insects12100857
272. Watson, J., Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyr Nucl Acid. Nature 171, 737–738 (1953). https://doi.org/10.1038/171737a0
273. Woese C.R. (1987) Bacterial evolution. Microbiol Rev. 51(2):221-71. doi:10.1128/mr.51.2.221-271.1987
274. Wolstenholme D.R. (1992). Animal Mitochondria1 DNA: Structure and Evolution. Intern Rev of Cyt 141:173-216. doi.org/10.1016/S0074-7696(08)62066-5
275. Wu T.D. (2016). Bitpacking techniques for indexing genomes: I. Hash tables. Algorithms Mol Biol 11 (1): 1–13. doi:10.1186/S13015-016-0069-5
276. Wyman S.K., Jansen R.K., Boore J.L. (2004). Automatic Annotation of Organellar Genomes with DOGMA. Bioinformatics 20 (17): 3252–3255. doi:10.1093/bioinformatics/bth352
277. Xiufeng X., Árnason Ú. (1994). The complete mitochondrial DNA sequence of the horse, Equus caballus: extensive heteroplasmy of the control region. Gene 148 (2): 357–362. doi:10.1016/0378-1119(94)90713-7
278. Xu S.Q., Yang Y.Z., Zhou J., Jing G.E., Chen Y.T., Wang Jun, Yang H.M., Wang Jian, Yu J., Zheng X.G., Ge R.L. (2005). A mitochondrial genome sequence of the Tibetan antelope (*Pantholops hodgsonii*). Genomics, Proteomics Bioinforma 3 (1): 5–17. doi:10.1016/S1672-0229(05)03003-2
279. Yandell M., Ence D. (2012). A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13 (5): 329–342. doi:10.1038/nrg3174
280. Yang C., Xiang C., Qi W., Xia S., Tu F., Zhang X., Moermond T., Yue B. (2013). Phylogenetic analyses and improved resolution of the family *Bovidae* based on complete mitochondrial genomes. Biochem Syst Ecol 48: 136–143. doi:10.1016/j.bse.2012.12.005
281. Young A.D., Gillung J.P. (2020). Phylogenomics - principles, opportunities and pitfalls of big-data phylogenetics. Syst Entomol 45 (2): 225–247. doi:10.1111/syen.12406

282. Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T.Y. (2017). Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol Evol 8 (1): 28–36. doi:10.1111/2041-210X.12628
283. Yu X., Sun S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics 14 (1): 1. doi:10.1186/1471-2105-14-274
284. Zemanová B., Hájková P., Hájek B., Martínková N., Mikulíček P., Zima J., Bryja J. (2015). Extremely low genetic variation in endangered Tatra chamois and evidence for hybridization with an introduced Alpine population. Conserv Genet 16 (3): 729–741. doi:10.1007/s10592-015-0696-2
285. Zemanová, B., P. Hájková, J. Bryja, J. Zima Jr., A. Hájková, and J. Zima. 2011. (2011). Development of multiplex microsatellite sets for noninvasive population genetic study of the endangered Tatra chamois. Folia Zool 60 (1): 70–80. doi:10.25225/fozo.v60.i1.a11.2011
286. Zerbino D.R., Birney E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18 (5): 821–829. doi:10.1101/gr.074492.107
287. Zhang F., Ding Y., Zhu C.D., Zhou X., Orr M.C., Scheu S., Luan Y.X. (2019). Phylogenomics from low-coverage whole-genome sequencing. Methods Ecol Evol 10 (4): 507–517. doi:10.1111/2041-210X.13145
288. Zhang, G. Bird sequencing project takes off. (2015). Nature 522,34. doi:https://doi.org/10.1038/522034d
289. Zhou M., Yu J., Li B., Ouyang B., Yang J. (2019). The complete mitochondrial genome of *Budorcas taxicolor tibetana* (*Artiodactyla*: *Bovidae*) and comparison with other *Caprinae* species: Insight into the phylogeny of the genus *Budorcas*. Int J Biol Macromol 121: 223–232. doi:10.1016/j.ijbiomac.2018.10.020
290. Zurano J.P., Magalhães F.M., Asato A.E., Silva G., Bidau C.J., Mesquita D.O., Costa G.C. (2019). *Cetartiodactyla*: Updating a time-calibrated molecular phylogeny. Mol Phylogenet Evol 133 (December 2018): 256–262. doi:10.1016/j.ympev.2018.12.015

8 ŽIVOTOPIS AUTORA

Toni Tešija rođen je u 23. veljače, 1993. godine u Splitu. Diplomirao je na 2017. na Agronomskom fakultetu, smjer Genetika i oplemenjivanje životinja. Od ožujka 2019. godine zaposlen je kao asistent na Agronomskom fakultetu u sklopu HRZZ projekta „Razvoj karijera mladih istraživača“. Sudjelovao je na projektu HRZZ-a pod nazivom „DNA kao dokaz o distribuciji i vitalnosti ugrožene Balkanske divokoze“ čiji je voditelj izv.prof.dr.sc. Nikica Šprem, te je član istraživačke grupe na aktivnom HRZZ-ovom projektu „Uloga lova i lovnog gospodarenja u širenju novonastalih populacija divljih papkara na Mediteranu“ čiji je voditelj doc.dr.sc. Toni Safner. Znanstveno se usavršavao na više specijalističkh tečajeva i radionica u zemlji i inozemstvu te je sudjelovao je na nekoliko znanstvenih i stručnih konferencija.

Popis znanstvenih radova

1. Corlatti L., Iacolina L., Safner T., Apollonio M., Buzan E., Ferretti F., Hammer S.E., Herrero J., Rossi L., Serrano E., Arnal M.C., Brivio F., Chirichella R., Cotza A., Crestanello B., Espunyes J., Fernández de Luco D., Friedrich S., Gačić D., Grassi L., Grignolio S., Hauffe H.C., Kavčić K., Kinser A., Lioce F., Malagnino A., Miller C., Peters W., Pokorny B., Reiner R., Rezić A., Stipoljev S., **Tešija T.**, Yankov Y., Zwijacz-Kozica T., Šprem N. (2022). Past, present and future of chamois science. Wildlife Biol 1–13. doi:10.1002/wlb3.01025
2. Iacolina L., Buzan E., Safner T., Bašić N., Geric U., **Tešija T.**, Lazar P., Arnal M.C., Chen J., Han J., Šprem N. (2021). A mother’s story, mitogenome relationships in the genus rupicapra. Animals 11 (4): 1–11. doi:10.3390/ani11041065
3. **Tešija T.**, Safner T. (2021). Analyses of wild ungulates mitogenome. Agric Consp Scient 86 (1): 1–12
4. Šalamon D., Furdić P., **Tešija T.**, Džidić A. (2019). Genetic parameters for the external udder morphology in commercial farms of Istrian sheep from Croatia. Jour of Central Euro agricul, 1; 68-73 doi:/10.5513/JCEA01/20.1.2462