SZENT ISTVÁN EGYETEM– KAPOSVÁRI CAMPUS
AGRÁR-ÉS KÖRNYEZETTUDOMÁNYI KAR

Állattenyésztés-technológia és Menedzsment Intézeti Tanszék

A doktori iskola vezetője
PROF. DR. SZABÓ ANDRÁS
MTA doktora

Témavezető:
PROF. DR. HORN PÉTER
MTA rendes tagja

Társ-témavezető:
PROF. DR. OROSZ LÁSZLÓ
MTA rendes tagja

A GÍMSZARVAS KROMOSZÓMÁK DNS SZEKVENCIÁJA, A CERELA1.0 GÍMSZARVAS GENOM ÖSSZEÁLLÍTÁSA

Készítette:
BANA ÁGNES NÓRA

KAPOSVÁR
2020

TARTALOMJEGYZÉK

1 RÖVIDÍTÉSEK ÉS SZAKKIFEJEZÉSEK JEGYZÉKE

AFLP	Amplified Fragment Length Polymorphism
Alignment	Összeillesztett szekvenciák vagy felillesztett szekvencia
Assembly	Bioinformatikai módszerekkel összerakott pszeudogenom
BAC	Bacterial Artificial Chromosome (egyfajta DNS vektor)
Backbone	MMSc, RGSc, és IRGSc scaffoldok helyes sorrendje alapján képzett kromoszónkénti szekvencia gerinc
bp	Bázispár, (előtagok: K-kilo, M-mega, G-giga)
Bt	*Bos taurus* kromoszózávma (utána a megfelelő szám)
BTA1	*Bos taurus* 1. kromoszóma (átvéve Bonnet és mtsai., 2001)
CCD	Charge-coupled Device kamera, analóg jelet továbbító kamera
cDNS	Komplementer dezoxiribonukleinsav szál
Ce	*Cervus elaphus* kromoszóma (utána a megfelelő szám)
CerEla1.0	Kromoszómába rendezett gímszarvas referencia genom
ChIP-Seq	Chromatin immunoprecipitation (ChIP) assays with sequencing, transzkripciós faktorok kötőhelyeinek azonosítása a DNS szálon
chr	Chromosome/kromoszóma
CIC pontszám	CIC: Conseil International de la Chasse et du Conservation du Gibier, trófeabírálat egységesítésére létrehozott pontozási rendszerben az agancs minőségét meghatározó mutatókra adott összpontszám
cM	Centimorgan, genetikai térképegység
CNP7	Szikaszarvas (*Cervus nippon*) 7. kromoszóma (átvéve Bonnet és mtsai., 2001)
CNP25	Szikaszarvas (*Cervus nippon*) 25. kromoszóma (átvéve Bonnet és mtsai., 2001)
contig	Kontig, kontinuus DNS szekvencia szakasz

DDBJ	DNA Data Bank of Japan, japán bioinformatikai adatbázis
DNS	Dezoxiribonukleinsav
ddNTP	didezoxinukleotid-trifoszfát
dNTP	Deoxinukleotid-trifoszfát
EDTA	Etilén-diamin-tetraecetsav
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive, nukleotid adatbázis
ENSEMBL	Online genom adatbázis és böngésző (európai)
EST	Expressed sequence tag, génkifejeződési-marker, cDNS darabja
Flanking régió	A kitüntetett szekvencia melleti régió.
Gap	Ismeretlen szekvencia, vagy szekvenciák közti rés az assemblyben
GFSc	Gap filling/réskitöltő scaffold
GO	Gene Ontology database/gén ontológia adatbázis
GWAS	Genome Wide Association Study/Genom asszociációs vizsgálatok
HGP	Human Genome Project
HTP	High-throughput/nagy áteresztőképesség szekvenálás
INDEL	IN-inszerció/beépülés, DEL-deléció/kivágódás
Insert	Read párok tagjai közötti rész
IRGSc	Inter reference genes scaffolds – nem UCSC referencia, rRNS, tRNS, miRNS géneket tartalmazó scaffoldok
LSU	Large subunit ribosomal ribonucleic acid, rRNS nagy alegysége
miRBase	Mikro RNS adatbázis
miRNS	Mikro RNS
MMSc	Mapmarker scaffolds/géntérképpont scaffoldok
mRNS	Hírvivő (messenger) RNS
N50	Szekvencia összeszerelés minőségét mutató statisztikai mérőszám
NCBI	National Center for Biotechnology Information, Biotechnológiai szervezet és online adatbázis (USA)

NGS Next-generation sequencing/Új generációs szekvenálás

NHGRI National Human Genome Research Institute (USA)

ORF Open-reading frame, nyitott leolvasási keret kódoló DNS-nél

PacBio Pacific Biosciences vállalat, hosszú read szekvenálás

PCR Polymerase Chain Reaction/polimeráz láncreakció

Phred érték DNS-szekvenálás során leolvasott nukleotidbázisok azonosításának minőségét adja meg.

Pszeudogenom Több szekvencia közös, összeilleszkedő részei alapján in silico létrehozott, a valóságban nem létező mesterséges genom, egyaránt állhat contigokból, scaffoldokból és kromoszómákból is.

Pszeudokromoszóma Több szekvencia közös, összeilleszkedő részei alapján in silico létrehozott, a valóságban nem létező mesterséges kromoszóma, a leghosszabb scaffold.

PubMed NCBI publikációk gyűjteménye (USA)

Query Kereső szekvencia a blast illesztéseknél

RBG R-bands after BrdUrd (5-bromo-2' deoxyuridine) and Giemsa, kromoszóma festési eljárás

Read A DNS szekvenálás során leolvasott nukleotidbázisok

Referencia genom A haploid genom pszeudoszekvenciáit tartalmazó konszenzus szekvenciák összessége, a valóságban nem létező mesterséges genom. A többi azonos vagy rokon fajból származó szekvenciát ehhez viszonyítjuk.

RFLP Restriction Fragment Length Polymorphism, a populáció egyedei különböző restrikciós enzimkötőhellyel rendelkeznek, ami eltérő restrikciós hasítással nyert fragment méretet eredményez közöttük.

RFLV Restriction Endonuclease Fragment Length Variation, Rekombinációs géntérképeknél a kapcsoltsági csoportokon

	található marker típus, amely az eltérő endonukleáz kötőhelyekből adódó fragment hosszúsági eltérést mutatja.
RGSc	Reference gene containing scaffolds/referencia gén scaffoldok
RNA-Seq	RNA sequencing/RNS szekvenálás
rRNS	Riboszómális RNS
Scaffold	Szkaffold, contigok összekapcsolásával képzett szekvencia váz
SILVAdb	Riboszómális RNS adatbázis
snRNS	Small nuclear RNA/kis sejtmagi RNS
SNP	Single Nucleotide Polymorphism/egypontos nukleotid polimorfizmus, meghatározott allél frekvenciával (1% fölötti) rendelkezik a populációban.
SNV	Single-nucleotide variant/egypontos nukleotid variáns, nem jelöl allél gyakoriságot a populációban.
SSU	Small subunit ribosomal ribonucleic acid, rRNS kis alegysége
STR	Short Tandem Repeat/rövid tandem ismétlődés
Subject	Az a szekvencia, amire keresőszekvenciát illesztenek.
Suffix	Az informatikában használt kifejezés, amely az adatbázisokban a mezők azonosítására szolgáló rövid nevet vagy betűcsoportot jelenti.
Szekvenálás	Bázispárok leolvasása, sorrendjük meghatározása
Szkript	(Script) Valamilyen programozási nyelven íródott utasítássor, rövid program.
Treshold	Küszöbérték blast illesztésnél
tRNS	Szállító (transzfer) RNS
UCSC	University of California, Santa Cruz, genom adatbázis, böngésző
Uniprot	The Universal Protein Resource, bioinformatikai fehérje adatbázis
UTR	Untranslated region/nem transzlálódó génrégió
WGS	Whole Genome Sequencing/teljes genom szekvenálás

2 BIOINFORMATIKAI PROGRAMOK JEGYZÉKE

ALLPATHS-LG Sequence Assemble Programme, Read illesztőprogram

ANNOVAR Genetikai variánsokat annotáló szoftver

AUGUSTUS Eukarióta gén prediktáló szoftver

BARRNAP BAsicRapid Ribosomal RNA Predictor/rRNS gén kereső program

BAM Sequence Alignment/Map (SAM) formátumú fájl bináris verziója

BED Browser Extensible Data (Fájl: megadja a gének lokalizációját)

BLAST Basic Local Alignment Search Tool, lokális DNS/fehérje szekvencia illesztőprogram (variációi: BLAST2, WU-BLAST, PHI-BLAST, PSI-BLAST, MegaBLAST, BLASTZ)

BLAT Blast Like Alignment Tool, DNS/protein szekvencia illesztőprogram

BLOSUM BLOcks SUbstitution Matrix, fehérje szubsztitúciós mátrix

BWA Burrows-Wheeler Aligner, DNS szekvencia illesztőprogram

BWT Burrows-Wheeler-transzformáció, blokkrendező algoritmus

EMBOSS Bioinformatikai program csomag

EXONERATE Páros szekvencia összehasonlító program

FASTA DNS/fehérje szekvencia illesztőprogram, szekvencia fájl kiterjesztés

GFF Gene-finding format, táblázatos fájl formátum

HMM Hidden Markov Model, statisztikai modell a bioinformatikában

HSP High-scoring segment pairs, rövid szekvencia szakaszpárok (blast)

InterProScan Bioinformatikai programcsomag, fehérje funkcionális analízis

JASPAR Database of transcription factor binding profiles

LASTZ Large-Scale Genome Alignment Tool, szekvencia illesztőprogram

MAKER Genom annotáló programcsomag

MSP Maximal-scoring segment pairs, maximális szekvencia szakaszpár

MUMmer Maximal Unique Matcher, szekvencia illesztőprogram

PAM Point accepted mutation, fehérje szubsztitúciós mátrix

RepeatMasker Repetitív szekvencia azonosító és maszkoló program

RepeatModeler Transzpozon azonosító program

RepeatRunner Repetitív elemeket azonosító program

RepeatScout Repetitív elemeket azonosító program

RECON Repetitív elemeket azonosító program

SAM Sequence Alignment/Map formátumú, readeket tartalmazó fájl

SAMtools Read szekvenciákat tartalmazó fájlokat kezelő program

SEQMAN Genomi szekvencia vizualizációs, szerkesztő program

SNAP Protein szekvencia alapján annotáló program

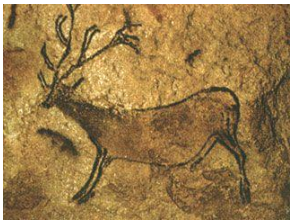
tRNAscan-SE Transzfer RNS kereső program

TRANSFAC TRANScription FACtor database, transzkripciós faktor adatbázis

VCF Variant call format, genetikai variánsokat tartalmazó fáj

3 BEVEZETÉS

A gímszarvas csaknem 20000 éve fontos elemét jelenti az emberi kultúrának. Ikonikus lény, amelyet évezredek óta kultikus tisztelet és csodálat övez. Alakját számos jégkorszak-kori barlang falán megcsodálhatjuk. A Világörökség részét képező Lascaux-i barlangban 15-17000 éves festett műalkotásokra bukkantak, amelyeken nagy gonddal kidolgozott szarvas rajzok láthatók (1. ábra).



1. ábra. Lascaux-i barlangrajz (Forrás: d'Huy, 2011).

A gímszarvas irodalmi szereplő az ókor óta, versek és mitológiák hőse. Zeneművek, képzőművészeti, film és fotóművészeti alkotások, egyházi énekek témájaként szolgál. A csodaszarvas a hun-magyar mondakörben és a magyar népi hagyományokban is ismert Isten által küldött mitikus vezérállat.

A kulturális jelentőségen kívül kiemelkedő társadalmi, természeti és gazdasági haszonnal rendelkezik (Milner és mtsai., 2006, Andersen és mtsai., 2010, Burbaitė és Csányi, 2010). Holarktikus elterjedésű faj, amelynek egyedei nagyszámban megtalálhatók Eurázsia, Észak-Amerika és bizonyos észak-afrikai területek erdőségeiben és pusztáin. A magyarországi fauna rekord méretű trófeáival (Szálka, Gamás, Lenti, Pusztakovácsi, Gemenc-Karapancsa, Vajszló, Lábod) emelkedik ki az európai populációk közül. A gemenci szarvas állomány nagy jelentőséggel bír a természetbarát turizmus és a trófea vadászat terén. Világszerte jelennek meg szarvasfarmok, ahonnan a tenyésztési eljárásoknak köszönhetően jó minőségű húskészítmények és magas CIC

pontszámú trófeák kerülhetnek ki (Horn, 2004, Andersen és mtsai., 2010, Sonkoly és mtsai., 2013). Magyarországon, Bőszénfán a Kaposvári Egyetem által létrehozott Szarvasfarm nemzetközi hírű központ. A modern vadgazdálkodás és állattenyésztési eljárások feltétele a genetikai úton történő egyedazonosítás, valamint az apasági és az anyasági vizsgálat, amelyet szarvasfélék esetében jelenleg mikroszatellita, illetve mitokondriális markerek használatával oldanak meg (Hartl és mtsai., 2003, Feulner és mtsai., 2004, Fajardo és mtsai., 2007, Zsolnai és mtsai., 2009, Szabolcsi és mtsai., 2014, Olivieri és mtsai., 2014). A gímszarvas genomjának ismerete nem csupán az orvostudományban tölthet be fontos szerepet. Jó modell szervezet lehet az emberi oszteoporózis genetikai sajátságainak megértéséhez, mivel az élővilág leggyorsabb csontfejlődése a gímszarvas éves agancs ciklusa során következik be (Stéger és mtsai., 2010).

A gímszarvas genom program egy 21 éves történet, amely a Mezőgazdasági Biotechnológiai Központ (MBK-NAIK, Gödöllő), a Kaposvári Egyetem Állattenyésztési, a Bőszénfai Szarvasfarm, az ELTE Genetikai, és a SOTE 1. Belgyógyászati Klinika MSc és PhD programjai összefogásával, kizárólagosan hazai erőforrásokból jöhetett létre. A CerEla1.0 az első nyilvánosan hozzáférhető gímszarvas genom szekvencia a világon, ami egyben Magyarország első emlős genom programja is. Célkitűzései nagyban hasonlítanak a Humán Genom szekvencia első verziójának elkészítéséhez. Az 1998-ban induló program koncepcióját Orosz László dolgozta ki, az állattenyésztési és vadgazdálkodási hasznosítást Horn Péter, a klinikai irányt Lakatos Péter jegyezte. E genom program megvalósulásának feltétele a klasszikus és molekuláris genetikai analízisek megvalósítása, a géntechnológiák, az emlős molekuláris sejt, szövet és szervbiológiai fejlődés, a bioinformatika, és a nagyérzékenységű statisztikai elemzések ismerete és célzott használata volt.

4 IRODALMI ÁTTEKINTÉS

4.1 A GÍMSZARVAS ÁLTALÁNOS JELLEMZŐI

A gímszarvas a párosujjú patások (*Artiodactyla*) rendjébe tartozik, ahol a *Pecora* fejcsapnyúlványos csoporton belül három családot különíthetünk el: a tülkösszarvúakat (*Bovidae*), a pézsmaszarvasokat (*Moschidae*) és a szarvasféléket (*Cervidae*). A szarvasfélék mintegy 27 millió éve, elsőként váltak el a két másik családtól és az eocénban jelentek meg Európában. A *Cervidae* családba 2 alcsalád tartozik: az óvilági szarvasformák (*Cervinae*) például a gímszarvasok és az újvilági szarvasformák (*Capreolinae*) például az őzek. A hátsó láb külső lábközépcsont fejlettsége szerint az összes újvilági faj és az óvilágiak közül az őzek és a vízőzek a telemetacarpaliak, míg az óvilági fajok (az őzek és a vízőzek kivételével) a plesiometacarpaliak csoportjába rendeződnek. A *Cervidae* családba jelenleg 55 fajt sorolnak be a kutatók, közülük a Schomburgk-szarvast (*Rucervus schomburgki*) gyakorlatilag kihaltnak tekintik. A gímszarvas *Cervidae* család, *Cervinae* alcsaládjának a tagja (Gilbert és mtsai., 2006, Agnarsson és May-Collado, 2008). A fajon belül sokféle változat megtalálható, amelyek elterjedési területe Észak-Amerika, Észak-Afrika, Európa, Közép-Ázsia, de betelepített fajként jelen vannak Dél-Amerikában, Ausztráliában és Új-Zélandon is. E változatok (vapiti, maral, tibeti szarvas, Thorold szarvasa, sárgafarkú szarvas, kasmírszarvas, és „európai" gímszarvas) agancsuk fejlődésében némi eltérést mutatnak. Az „európai" gímszarvas agancsa „koronás jellegű", vagyis a 4. 5. ág a többi ág síkjából ered úgy, hogy a nagyobb számú ág mellett a végső villa elágazásnál három azonos helyből sarjadzó ágak nőnek ki (https://mek.oszk.hu/03400/03408/html/200.html). Ez az agancs típus a legtöbb szarvasféléhez hasonlóan minden évben újra épül és plusz ágakkal gyarapszik, amíg el nem éri a végleges formáját, vagyis a klasszikus szemág,

jégág, középág, korona szerkezetet, ami minimum 10-es agancsot jelent (1. melléklet).

A gímszarvas testfelépítése karcsú, erőteljes és izmos egyben. A Kárpát-medencei populációban a hossza 2,1 m-t is lehet, vállmagassága 1,5 m. A súly a teheneknél 90-110 kg között mozog, míg egy jól megtermett bika a 300 kg-t is elérheti. A felnőtt állatok jellegzetes vörösesbarna színűek „rőt vad, rot hirsch, red deer" a farukon fehér tükör található. A borjak bundája fehér pöttyökkel tűzdelt. A bikák sörényt és agancsot viselnek, a feromonok kiválasztását a szem alatt található két szemgödör üreg teszi lehetővé. Feje hosszúkás. Lábai karcsúak, a fattyú csülkök nem érintik a földet. Évente kétszer vedlik, téli ruhája szürkésebb. A szarvasok rudlikban élnek, a női egyedekből álló rudlit a vezértehén vezeti. Az októberi szarvasbőgés idején a gímszarvas bika igyekszik, minél több tehenet maga köré gyűjteni, és megvédeni a kihívó bikákkal szemben. A hét nyolcévesnél idősebb gímszarvas bikák, már rendelkezhetnek olyan kifejlett 12-es vagy annál nagyobb osztású aganccsal, amelynek a látványa is elég, hogy távol tartsa a többi bikát a háremtől. Amikor a „display" (agancs mutogatás) önmagában nem elegendő, akkor kerül sor a bikák közötti küzdelemre. A vemhesség 33-34 hétig tart. A gímszarvas borjak májusban vagy júniusban születnek meg. Egy tehén rendszerint egy borjat ellik, aki megpróbálja azonnal követni az anyját. A szoptatás időszaka a következő szaporodási idényig (rigyetésig) tart. A gímszarvas kérődző állat, az őzzel ellentétben nem pákosztos étrendű, viszont nem is annyira igénytelen, mint a dámvad. A friss hajtásokon kívül a zsenge fák kérgeit is szívesen elrágja, ezzel nem kevés vadkárt okoz.

Körülbelül 6000 emlős faj él a világon, ezek közül azonban csak öt nagytestű fajt háziasított egészen az ember. Még további kilenc faj (dromedár, kétpúpú teve, láma, alpaka, szamár, rénszarvas, ázsiai bivaly, jak, banteng) domesztikációja nem teljesen történt meg, illetve csupán helyi jellegű maradt

(Diamond, 2000). Természetesen akadnak még egyéb speciális esetek is, mint például az afrikai jávorantiloppal történt próbálkozás. A szarvasfélék esetében a rénszarvas félnomád tartása Lappföldön és a szibériai tundrán közismert. Szibériában és Kínában a gímszarvas egyes alfajait (maral, szikaszarvas, pettyes szarvas) fogságban tartják, hogy a szarvasbikák barkás agancsából úgynevezett pantokrint (roboráló afrodiziákumot) készítsenek.

4.2 SZARVASFARMOK MAGYARORSZÁGON ÉS ÚJ-ZÉLANDON

A Kaposvári Egyetem és a Bőszénfai Vadgazdálkodási Tájközpont együttműködése 1991-92 között kezdődött meg, lényegében egy kutatási programról van szó, amelyhez szorosan kapcsolódik a takarmánytermesztés, az erdő és vadkerti gazdálkodás, a vadászat, a húsfeldolgozás és a turizmus. Az 1300 hektár nagyságú területen mintegy 1500 gímszarvas, 900 vaddisznó, 200 dámvad, 150 muflon, 50 őz és sokféle háziállat és egzotikus vad él. A vadfarmok a zárttéri állattartás legintenzívebb formáját jelentik. Bőszénfa jó ökológiai adottságú térségben helyezkedik el és a Dél-dunántúli gímszarvas állomány kiváló genetikai adottságú, jó növekedési eréllyel és agancsfelrakási képességgel rendelkezik (2. ábra). A gímszarvasok csoportosan tarthatók, jól társíthatók más fajokkal, ellenállóak a betegségekkel szemben és kiválóan hasznosítják a Bőszénfán rendelkezésükre álló legelőterületet. Természetesen itt is szerepet kap a vadhús előállítás, a sport vadászat, ami leggyakrabban a nagy értéket képviselő agancs trófeákért történik.

2. ábra. 7 éves bőszénfai gímszarvas bika (Crot.3016). A CerEla1.0 genom program kulcsszereplője.

Itt található Közép-Európa legnagyobb tenyészállománya, amely fontos magyar génbázis is egyben. A tenyészállatok exportjának célországai Spanyolország, Németország, Lengyelország, Horvátország, és Szlovénia. Olyan előremutató állattartási és tenyésztési technológiákat alkalmaznak itt a Kaposvári Egyetem szakembereinek a bevonásával, mint például a mesterséges szarvastej előállítás, nyomelem és ásványi anyag kimutatása szőrmintából, a vadon befogott szarvasok TBC mentesítése, diagnosztikai eljárások kifejlesztése stressztűrő képességre és a modern szaporodás biológiai módszerek. Ezek közé a szaporítási módszerek közé tartozik a mesterséges megtermékenyítés, az 1996-ban kidolgozott embriótranszplantáció és transzfer, a posztmortem szarvasbikákból történő sperma levétel, és az inszeminációs technikák. A Kaposvári Egyetemen lehetőség nyílik élőállatok test összetételének a vizsgálatára CT MRI digitális képalkotó berendezések használatával. A Kaposváron 1998-ban megrendezett Szarvas biológiai Világkongresszusra készült el a gímszarvas CT és MRI anatómiai atlasz (Horn, 1998).

A Kaposvári Egyetem az 1970-es évek óta áll kapcsolatban az új-zélandi szarvasfarmokkal. Az új-zélandi szarvastartás, az ország több mint 250 ezer

km^2 nagyságú dús legelőire alapoz. A Lincoln Egyetemen farmszerű tartás módszerek oktatása bekerült az állattenyésztési tanmenetbe és állattenyésztési kutató központok (Invermay Agricultural Centre) alakultak meg. Az új-zélandi állományt jó genetikai tulajdonságú, kapitális agancsú szarvasbikák importjával igyekeztek feljavítani, ezért 1980-ban több magyar szarvasbikát (a leghíresebbek: Kapos, Magyar, László) is behoztak az országba. A nagymértékű állami támogatás és a korszerű és tervszerű tenyésztési eljárásoknak köszönhetően az új-zélandi mezőgazdaság nagyon jelentős, önálló, nemzeti ágazatává, „iparággá" fejlődött (Horn, 2004). 2000-re már több, mint kétmillió gímszarvas nevelkedett új-zélandi szarvasfarmokon, ami a Föld tenyésztett szarvas populációjának a felét teszi ki (Deer Farmer, 2000).

4.3 KÜLÖNBÖZŐ SZARVASFAJOK KROMOSZÓMÁLIS EVOLÚCIÓJA

A kromoszóma szám jellemző az adott fajra, de nem zárja ki, hogy egymástól evolúciósan távol eső fajok kromoszóma száma megegyezzen, például a kínai muntyákszarvas (*Muntiacus reevesi*) és az ember (*Homo sapiens*) ugyanúgy 23 pár kromoszómával rendelkezik a diploid sejtben (23-mal a haploid sejtekben). Különös érdekesség, hogy a muntyákszarvas indiai fajában a 23 haploid kromoszóma 6 kromoszómában fúzionálódik. A fajképződés során általában kiterjedt kromoszómális átrendeződések történhetnek, például deléciók, inverziók, transzlokációk, duplikációk, kromoszómák centrális fúziója (Robertsoniális transzlokáció), tandem fúziók (Robertsoniális fúzió), tandem és centrális hasadások.

A törzsfejlődés folyamán az emlősökre jellemző kromoszómális hasadások kromoszómaszám növekedéshez vezettek, ami elősegítette a robbanásszerű fajképződést. A párosujjú patások (*Artiodactylia*) és ezen belül a szarvasfélék (*Cervidae*) kariotípusa egy ősi 2n=20 karioípusból alakulhatott ki a

kromoszómális hasadásoknak köszönhetően (Todd, 1975, 2000). Az első *Artiodactylia*-k paleontológiai maradványai a közép-eocén rétegből kerültek elő 53 millió évvel ezelőttről. A kérődző alrendnek a megjelenése is ehhez az időszakhoz köthető. A kérődzők kicsi mindenevő, erdőlakó élőlények voltak. A kromoszómaszámuk kezdetben megegyezhetett a többi *Cetartiodactylia*-éval (2n=48) (Kulemzina és mtsai., 2011). A párosujjú patásokon belül a homlokcsaposok, vagyis a *Pecora*-k megjelenése és gyors elterjedése a korai miocén időszakra tehető (körülbelül 30 millió évvel ezelőtt). Három mai kérődző faj és két külcsoport (ember, egér) homeológ kromoszómális régióinak az összevetése alapján az ősi *Pecora*-k kromoszóma száma a diploid genomra nézve a nemi kromoszómákkal együtt 29 pár lehetett (Slate és mtsai., 2002).

A homlokcsaposok csoportja öt ma is élő családra bontható fel: zsiráffélék (*Giraffidae*), pézsmaszarvasfélék (*Moschidae*), szarvasfélék (*Cervidae*), villásszarvúantilop-félék (*Antilocapridae*), és tülkösszarvúak (*Bovidae*). Ezek a családok viszonylag kevés kromoszómális tulajdonságban mutatnak egyezést, aminek az lehet az oka, hogy a közös őstől való elválásuk közel azonos időben történt. A szétválás után mindegyik ág függetlenül fejlődött és halmozott fel egyéni, apomorf kromoszómális átrendeződéseket (Kulemzina és mtsai., 2009).

A *Bovidae*-k közös ősében történt egy kromoszómális hasadás, így a diploid kromoszómaszám 60-ra változott, ez jellemző a mai háziasított szarvasmarhára (*Bos taurus*), de a juhban (*Ovis aries*) három fúzió és egy transzlokáció következtében a diploid kromoszóma szám 54-re csökkent (Slate és mtsai., 2002).

A *Moshidae* (pézsmaszarvas) családba egyetlen nemzettség tartozik, amely egyaránt mutat *Bovidae* és *Cervidae* jellegeket, emiatt taxonomiai hovatartozását heves viták előzték meg és sokáig a szarvasfélék alcsaládjaként

tartották számon (Wang és mtsai., 1993). A legújabb molekuláris filogenetikai elemzések alapján önálló fajként a *Bovidae*-k testvércsoportjának tekinthetők (Hassanin és Douzery, 2003).

A *Cervidae*-k közös őse 68 akrocentrikus autoszómával, 1 akrocentrikus X kromoszómával és egy kisméretű szubmetacentrikus Y kromoszómával rendelkezett.

A szarvasfélék családjában 7-9 millió évvel ezelőtt alakultak ki a szarvasformák (*Cervinae*) közép-Ázsiában. A *Cervinae* alcsaládon belül kétféle kariotípus jelent meg. Az egyik a *Cervini*-kre jellemző 2n=68, amely minimum két Robertsoniális transzlokációnak köszönhetően két metacentrikus autoszómát is tartalmaz az akrocentrikusak mellett (a diploid sejtekre nézve). E képlet jellemző fajai például a gímszarvas (*Cervus elaphus*), a disznószarvas (*Axis porcinus),* a szikaszarvas *(Cervus nippon),* a dámvad (*Dama dama*), és a Dávid-szarvas (*Elaphurus davidianus*). A gímszarvas Y kromoszómája az őséhez hasonlóan szubmetacentrikus maradt (2. melléklet) (Fontana és Rubini, 1990). A számbárszarvas (*Cervus unicolor)* széles elterjedésű faj, a különböző elszigetelt populációiban eltérő a kromoszómaszám, mivel evolúciója során további metacentrikus fúziók történtek. Nagyon hasonló genetikai események zajlottak le a szikaszarvas esetében is, ahol a földrajzi izoláció jóvoltából sokféle alfaj alakult ki Ázsiában és a környező sziget csoportokon. A *C. nippon nippon* alfaj kromoszómális Giemsa festése megmutatta, hogy a két legnagyobb méretű akrocentrikus kromoszómájukon rögszerű képletek, szatelliták, erősen kondenzált DNS szekvenciák ülnek (Bonnet és mtsai., 2001, Fontana és Rubini, 1990).

A másik *Cervinae* kariotípus a *Muntiacini* ág. A ma élő muntyákszarvasokon mind az agancsot, mind a jelentős nagyságú szemfogat megtaláljuk, ami a modernebb szarvas alakokon már nem látható. A muntyákok kromoszómaszáma nagyon különböző fajonként az eltérő számú tandem

fúziók miatt. A kínai muntyákszarvasok (*Muntiacus reevesi*) 2n =46, míg az indiai muntyákszarvas ünők (*Muntiacus muntjak vaginalis*) 2n=6 kromoszómával rendelkeznek, ami a legkisebb kromoszómaszám az emlősök között (Fontana és Rubini, 1990). Az utóbbi alfajnál ez az állapot a sokszoros lineáris tandem fúziók és a repetitív tandem transzlokációk miatt alakulhatott ki (Fredga, 1971, Tsipouri és mtsai., 2008). A *M. muntjak vaginalis* és az *M. reevesi* közötti erős kariotípusos különbségek ellenére a két faj hibridizálódhat egymással, amelynek eredményeképpen életképes, ámde steril utódok születhetnek, mivel a spermatogenezis korai profázisa gátolt (Fontana és Rubini, 1990) (3. melléklet).

Az őzformák (*Capreolinae*, régebben *Odocoileinae*) alcsalád képviselői 6-8 millió évvel ezelőtt jelentek meg. Ebben az időszakban a Tethys-óceán teljesen eltűnt, átadva helyét a dús vegetációjú, nagy kiterjedésű legelőknek (Heffelfinger, 2006). E típus egyik jellegzetes képviselőjeként az öszvérszarvast (*Odocoileus heraionus*) lehet említeni, ahol 2n=70 a kromoszómák száma. Az őzformák evolúciója során egyaránt előfordult Robertsoniális transzlokáció és pericentrikus inverzió. Az autoszómák között akadnak akrocentrikusak, metacentrikusak és szubmetacentrikusak is. Megjegyzendő azonban, hogy a szarvasformák és az őzformák metacentrikus kromoszómái nem feleltethetők meg egymásnak. Az őzformák kromoszómája szubmetacentrikus (Fontana és Rubini, 1990). A Capreolinae alcsalád három nagyobb nemzetségbe sorolható rénszarvas-rokonúak (*Rangiferini*), őz-rokonúak (*Capreolini*), és jávorszarvas-rokonúak (*Alceini*).

Az őzek összes autoszómája akrocentrikus (2n=70). Az európai őz (*Capreolus capreolus*) szubmetacentrikus és a gímszarvas (*C. elaphus*) akrocentrikus X kromoszóma G-sávmintázatának összehasonlító vizsgálata alapján megállapították, hogy egy közös ős akrocentrikus X kromoszómájában pericentrikus inverzió következett be az őzeknél, azonban a *Cervinae* és a

Hydropotinae alcsaládokban változatlan maradt az X kromoszóma (Rubini és Fontana, 1988). A centromeron heterokromatin elvesztése is igazolja az őz kromoszóma átrendeződés primitív természetét a *Capreolini* egész nemzettségen belül. Az ázsiai őz (*C. capreolus pygargus* alfajcsoport) különböző populációiban változó számú B kromoszómát találtak (Sokolov és mtsai., 1978, Neitzel, 1987). A B kromoszómák, olyan kis méretű kromoszómák, amelyek hagyományosan nem tartoznak egy adott faj kromoszóma szerelvényébe. Sokszor heterokromatikusak, sejtosztódáskor megoszlásuk egyenlőtlen. Nincsenek hatással vagy alig vannak hatással a fenotípusra és a vitalitásra. Az európai és a keleti őz szomatikus és kariotípusos különbségei miatt a tudósok egy része két különálló fajként ismeri el őket (Groves és Grubb, 1987).

4.4 GÉNTÉRKÉPEK

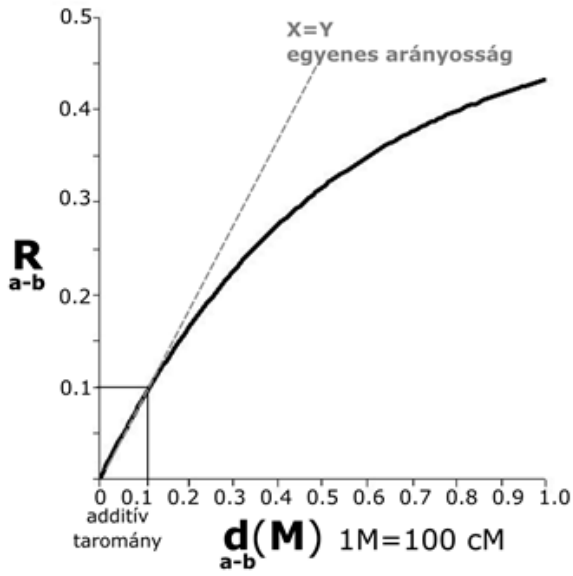
A géntérképezés célja, hogy meghatározzuk a gének vagy a markerek, azaz az ismert DNS szekvenciák egymástól való távolságát és kromoszómális elhelyezkedését genom szekvenálása nélkül. A kapott értékek nagyvonalakban megadják a markerek pozícióit és alkalmasak lehetnek arra, hogy egy ismert marker helye alapján meghatározzák egy eddig ismeretlen marker vagy gén helyzetét, amennyiben az ismert és az ismertlen markerek között kapcsoltsági viszony áll fenn. A kapcsoltsági viszonyra az együtt öröklődő fenotípusos tulajdonságok alapján tudunk következtetni. A kapcsoltsági rekombináns géntérképekkel ellentétben a fizikai térképek a teljes referencia genom hosszúságához képest határozzák meg a markerek közötti fizikai távolságokat. Egy genom teljes fizikai térképének elkészítéséhez teljes genomszekvenálás szükséges, ahol a távolságok bázispárban adhatók meg.

A kapcsoltsági térképezés, vagyis a linkage mapping módszere abból indul ki, hogy két vagy több marker a meiózis folyamán nem feltétlenül követi a Mendel

III. törvénye szerinti szabad/független kombinálódást, mivel egymással fizikailag is össze vannak kötve a DNS molekulán. E markerek egymással összekapcsoltan, együtt mozognak az őket hordozó kromoszómával és így kerülnek be az ivarsejtbe. Először az ecetmuslicában (*Drosophila melanogaster)* írták le a jelenséget. A w/w^+ (szemszín fenotípusok: fehér, vadtelt piros) és az m/m^+ (apró szárny, normális-vad szárny fenotípusok) allélpárok a criss-cross szabály (az X kromoszóma generációról generációra más nemű egyedbe jut „a fiú az anyjától örökli, és a lányának adja") szerint öröklődnek, egymástól függetlenül nem kombinálódnak, és ugyanazon az X kromoszómán találhatók. Megjegyzendő azonban, hogy a crossing-over mechanizmus újra kombinálhatja az allélpárokat. Szintén az ecetmuslicához köthető a világ első kapcsoltsági géntérképe, amelyet Alfred Henry Sturtevant készített el (Sturtevant, 1913). A homológ kromoszómapárok tagjai között a meiózis profázisának I. szakaszában figyelhető meg a crossing-over következménye a kiazma, ami tulajdonképpen a homológ kromoszómapár karjai közötti átkereszteződést (kialakul a Holiday szerkezet) és kromoszómális szakasz cseréjét jelenti, ami a dupla szálú DNS-ek/kromoszómák törésével és újraegyesülésével írható le egyszerűbben („törés-újraegyesülés modell"). A folyamat eredményeképpen új allélkombinációjú (rekombináns) kromoszómák jönnek létre a gamétákban. Kialakulnak az úgynevezett parentális kombinációk: P1, P2 (a kiindulási heterozigótát megjelenítő kombinációk, amelyek változatlanul adódtak tovább a gamétákba) és rekombináns kombinációk: R1, R2 (a heterozigóta meiózisa során képződő új kombinációk, amelyek a gamétákba jutnak). A cisz és transz heterozigóta egyenértékű, azaz azonos R értéket ad. Az R érték nem más, mint a rekombinációs gyakoriság, ami tulajdonképpen a rekombináns gaméták összessége osztva az összes gaméta számával. A gaméták genetikai milyenségét a P és R kombinációkat klasszikusan tesztelő keresztezéssel

mutatjuk ki („test-cross”). Általánosan az is elmondható, hogy egyetlen meiózist megjelenítő tetrádban két lókusz közötti rekombináció átlagosan 50%-nyi rekombináns kromoszómát eredményez. Azaz R az összes tetrádra vetített rekombináció gyakorisága attól függ, hogy hány tetrádot érint crossing-over a meiotikusan osztódó sejthalmazban (4. melléklet).

A klasszikus géntérképezés alapját a crossing-over törvények írják le, mint például a kétpontos vagy hárompontos elemzés, amelyek lehetővé teszik a genetikai rekombináció/crossing over kimutatását. A crossing-over és a rekombináció egymással összefüggő jelenségek, de nem azonosak. A rekombinációs gyakoriság (R) kimutatása irányított keresztezésekkel történhet, ahol az utódokból következtethetünk vissza a P_1, P_2, R_1, R_2 gaméták gyakoriságára (tesztelő keresztezés, „teszt-cross”). A rekombináció valószínűsége megnő két kapcsolt marker vagy gén között a genetikai távolság növekedésével, ez azonban nem mutat egyenes arányosságot, mivel csak a két gén között fellépő páratlan számú crossing-over vezet rekombinációhoz. A gének közötti távolság és a gének közötti rekombináció gyakorisága összefüggését a genetikai térképezési függvény, a Haldane géntérképezési függvény írja le (Haldane, 1919) (3. ábra).



3. ábra. Ra-b= ½ (1-e-2d), ahol R az a és b pontok/gének között a rekombináció gyakorisága, d az a és b pontok/gének távolsága, úgynevezett Morgan (M) egységekben kifejezve.

1 Morgan (M), annak a két pontnak/génnek a távolsága, amelyek között 1, a crossing-overek átlagos gyakorisága, 1 centimorgan (cM) távolság pedig 1% átlagos crossing-over gyakoriságnak felel meg. A genetikai térképeken 1 cM egy genetikai térképegységnek felel meg. A M, cM mérőszámai relatív, statisztikus, valószínűségi értéket jelentenek, nem metrikusak. Elméletben az a és b pontok közötti fizikai távolság (mondjuk Mbp-ban, bp-ban kifejezve) egyenesen arányos a crossing-overek átlagos gyakoriságával (M, cM érték). A genetikai analízis során a crossing-overek önmagukban nem érzékelhetők, hanem kísérleti úton például irányított keresztezés („teszt-cross”) eredményeképpen a rekombináns utódokból tudunk visszakövetkeztetni rájuk. A gének és markerek a genetikai távolságon kívül, egyéb tulajdonságaikban, például a sorrendjükben is összefüggést mutatnak a rekombinációs gyakoriságokon alapuló kapcsoltsági térképeken és a szekvenálás eredményeképpen összeállított fizikai térképeken, vagyis a géntérképi pontok

és a kromoszómán található lókuszok egymással megfeleltethetők, azonos sorrendben követik egymást. Ezt a jelenséget a géntérkép és a kromoszóma kolinearitásának nevezzük. A két pont cM és a valós fizikai távolsága közötti megfelelés, azonban nem egységes a fajok között, hanem az adott fajra jellemző crossing-over mechanizmus/enzimatikus apparátus hevességétől függ. Az emlősökben, beleértve ebbe a gímszarvast és az embert is, közelítőleg 1 millió bázispár hosszú DNS szakasz (1 Mbp) 1 cM géntérképi távolsággal egyezik meg, de bizonyos élőlényekben akár 20-50-szer kisebb is lehet az értéke (például *Drosophila melonagaster, Caenorhabditis elegans),* a mikrobiális világban még kisebb. A Haldane géntérképezési függvény értelmében a crossing-over véletlen, autonóm esemény, amely a kromoszóma mentén bárhol azonos valószínűséggel léphet fel. Ez azonban csak az idealizált állapot, mert a génkonverzió, a crossing-overek interferenciája és a rekombinációs „forró és hideg" pontok jelenléte az egészen rövid DNS szakaszokon jelentősen torzíthatja a rekombinációs gyakorisági értékeket. Minél nagyobb a fizikai távolság például egy kromoszóma két távoli pontja esetén, annál valószínűbb, hogy a crossing-overek száma hűen tükrözi a valós fizikai távolságot (bázispárban mért hossz). A kisebb területen tapasztalt lokális egyenetlenségek kioltják egymást (Kosambi, 1943). A gímszarvasok esetében is megfigyeltek egy már ismert, az emlősökre jellemző általánosságot, jelenséget, miszerint a meiózisban a rekombináció hevesebb a nőivarú egyedekben, mint a hímekben (Johnston és mtsai., 2017).

A gímszarvas géntérkép elkészülését látszólag megnehezítette, hogy ennek a fajnak nincsen háziasított változata és emiatt beltenyésztett fajtái sincsenek. A háziállatok fajtái gyakran jellegzetes fenotípusos tulajdonságokkal rendelkeznek, aminek az az oka, hogy a beltenyésztés és szelekció során az előnyös tulajdonságokért felelős homozigóta szakaszok fennmaradtak a vonalakban. A különböző fajták keresztezésével rekombinációs elemzés

végezhető. Egy másik lehetőség a három generációs családok elemzésével megvalósított DNS diagnosztika (például mikroszatelliták és SNP-k stb.). Ilyen például az ember vagy a szarvasmarha nagysűrűségű géntérképe, melyeket ennek megfelelően készítettek el. A gímszarvasoknál egészen más alternatíva adódott a géntérkép elkészítéséhez az un. „interspecifikus back-cross" módszer. A szarvasfélék egyes fajai közel állnak egymáshoz genetikailag. Ilyen közel álló faj a Dávid-szarvas (*Elaphurus davidianus*) másnéven milu, eredetileg Kínában őshonos szarvasféle. A boxer lázadások idején Kínában gyakorlatilag kipusztult, azonban ekkora az európai állatkertekbe is kerültek miluk. Külső megjelenésében meglehetősen furcsa, teve nyaka, marha patája, szamár farka és szarvas agancsa van, ezért is nevezik kínaiul sibuxiang-nak. Az agancsa eltér a gímszarvasétól, a főág elágazás nélkül hátrafelé nyúlik, míg az elülső ágon több elágazás is található, fordított gímszarvas agancs hatást kelt. A szarvasfélék közeli fajai keresztezhetők egymással, és fertilis fajhibrideket hoznak létre (Biedrzycka és mtsai., 2012). Ezt láthatjuk a milu és a gímszarvas interspecifikus keresztezésénél, ahol a Haldane szabállyal (Az emlős fajhibridek hímjei, XY, a „heterogamétás szex", általában sterilek, a női egyedek XX, a „homogamétás szex" ugyanakkor fertilisek) ellentétben nem csak a nőivarú, hanem a hím és női ivadékok egyaránt termékenyek. Bár a két faj megjelenésében és betegség rezisztencia szempontjából meglehetősen különbözik egymástól, a kromoszóma számuk megegyező (2n=68) és kariotípusuk is hasonló. Nagyon valószínű, hogy a kromoszómáikban a génsorrendek megegyeznek, szinténikusak.

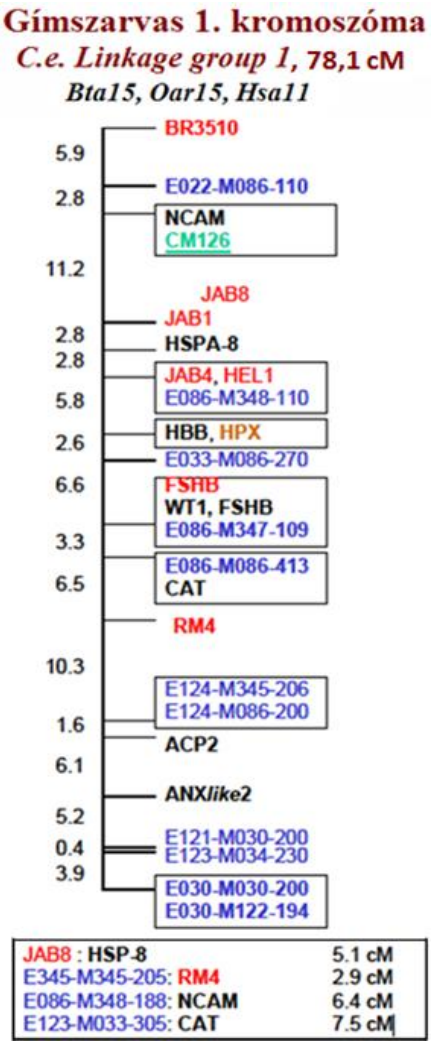
Az evolúciós törzsfán történt szétválásuk azonban nem maradt nyom nélkül. A genomjukban nagyszámú fajspecifikus mikroszatellita található. A mikroszatelliták egymás után elhelyezkedő rövid ismétlődésekből álló DNS szekvenciák. Átlagosan 1-6 bázispár hosszúságú repetitív egységekből épülnek fel. E markerek allélikus variációi jól kimutathatók DNS diagnosztikai

eljárásokkal, ha az allélok hossza különbözik (elektroforézis) és felhasználhatók a rekombinációs gyakoriságok meghatározásához és a géntérképezéshez.

Két új-zélandi szarvas farmon és az Invermay Agricultural Centre-ben 1989 és 1995 között 7 darab F1 milu (*Elaphurus davidianus*) és gímszarvas (*Cervus elaphus*) hibrid szarvasbikát kereszteztek 267 gímszarvas ünővel, amelynek eredményeképpen 351 back-cross utód született. A (gímszarvas-milu) hibrid hímek nagyszámú ivarsejtet tudnak produkálni, ami elegendő mennyiséget jelentett a farmokon nevelt gímszarvas ünők mesterséges megtermékenyítéséhez. Megvalósult a fajhibrid visszakeresztezés a kiinduló fajok egyikével (hibrid hímek x gímszarvas ünők). A mesterséges megtermékenyítésből született (F2) szarvasborjak adták a térképező populációt a géntérkép elkészítéséhez (5. melléklet). A borjúkból vért vettek és DNS-t izoláltak. Meghatározták a DNS marker variációk között a rekombinációk gyakoriságait, ebből cM távolságokat számoltak és így tudták megszerkeszteni a gímszarvas autoszómák géntérképét. Az X kromoszóma térképezéséhez az F2 back-cross nemzedék ünőit keresztezték gímszarvas hímekkel. Az ünők egyik X kromoszómája ui. gím, a másik X Dávid szarvas eredetű volt (criss-cross öröklés!), ezáltal rekombináns X kromoszómák is keletkezhetnek. Egy további keresztezéskor az F2 ünők az X kromoszómáikat a fiaiknak (F3!) adják tovább. Az F3 nemzedék szarvas bikáinak X kromoszómái vizsgálata alapján szerkesztették meg a gímszarvas X kromoszóma géntérképét (Slate és mtsai., 2002, Tate és mtsai., 1995) (6. melléklet).

A géntérképen összesen 714 markert sikerült meghatározniuk. Ezekből azonban csak 621-t tudtak pontosan meghatározott helyzetű térképpontként a kapcsolódási csoportokon bejelölni. A markerek különböző emlősök ortológ szekvenciáiból (gímszarvas, őszvérszarvas, virginai fehérfarkú szarvas, juh,

karibu, vapiti, kecske, disznó, majom, ember, patkány, nyúl, illetve nagyrészt szarvasmarhából) származtak. A kapcsoltsági csoportokat MapMaker V3.0 Genetic Linkage Map program készítette LOD analízis használatával. Minden olyan genetikai variáció használható genetikai markernek, amellyel jellemezni lehet egy lókuszt, azaz a homológ kromoszómákon levő, allélikus lókuszok között különbséget lehet tenni. A markerpontok két főbb típusba sorolhatók. Az I. típusú markerek nagymértékben konzerváltak a fajok között, de változatlanok egy fajon belül, a fajok elkülönítésére szolgálnak. A kapcsoltsági csoportok és az evolúciós elágazások meghatározásához jól alkalmazhatóak. Ide tartoznak a proteinek, az EST szekvenciák és az RFLV szekvenciák. A II. típusúak viszont polimorfak egy fajon belül is és kevésbé szélesen konzerváltak a fajok között. A mikroszatelliták és az AFLP markerek ebbe a csoportba tartoznak. A géntérkép pontjai 34 kapcsoltsági csoportba rendeződtek a gímszarvas haploid kromoszóma számának megfelelően (Slate és mtsai., 2002). Az X és az Y kromoszóma homológ szakasza azonos linkage groupnak felelt meg. Az ilyen módon elkészült géntérkép teljes hosszúsága 2532 cM. A munkám során ezeket a kapcsoltsági csoportokat használtam referencia géntérképként a kettős referencia vezérelt illesztéshez. A képen az 1. gímszarvas kapcsoltsági csoport látható (4. ábra) (Slate és mtsai., 2002).



4. ábra. Gímszarvas 1. kapcsoltsági csoportja (Slate és mtsai., 2002). A fejlécben a szarvasmarha, a juh és az ember megfelelő ortológ kromoszóma (számjelzés szerinti) olvasható le, továbbá a linkage group/kapcsoltsági csoport cM hossz értéke. A baloldalon lévő számok az egyes markerek közötti cM-ban megadott távolságok. A különböző színek a markerek típusait különítik el (kék-AFLP, piros-mikroszatellita, zöld-EST, fekete-RFLV/gén, barna-protein). Az alsó boxban feltüntetett markerekről csak az biztos, hogy szomszédosak, vagyis nem tudták megmondani a pontos pozíciójukat a géntérképen, ezért az egyik szomszédos markertől való távolságukat adták meg.

4.5 GENOM SZEKVENÁLÁS, ÖSSZESZERELÉS ÉS GENOM PROGRAMOK

A DNS szekvenálás lényegében azt jelenti, hogy meghatározzuk egy DNS molekula nukleotid sorrendjét. Az 1970-es évek végén jelentek meg az úgynevezett első generációs szekvenálási technikák. Az első ilyen klasszikus módszer a Sanger és Coulson nevével fémjelzett +/- szekvenálás, amellyel először olvasták le egy vírus, a phi X174 bakteriofág teljes genomját (Sanger és Coulson, 1975).

A második módszer a Maxam- és Gilbert-féle kémiai hasítás (Maxam és Gilbert, 1977) amely, hagyományos kémiai ismereteket használ fel. Az eljárás során a DNS szálak 5'-végét radioaktív 32P-tal jelölik meg, egyszálúsítják, majd négyféle kémiai kezelésnek vetik alá, eltörik. A különböző méretű DNS fragmentumokat denaturáló poliakrilamid gélelektroforézissel választják el egymástól, és autoradiográfiával azonosítják őket.

A három egymástól módszertanilag különböző technológia közül az utóbbi, a legegyszűrebben kivitelezhető Sanger-féle láncterminációs szekvenálás (Sanger és mtsai., 1977) terjedt el széles körben. Bizonyos laborok a pontossága miatt a mai napig használják a klasszikus Sanger-féle láncterminációs szekvenálást. Először a DNS szál hő denaturációját végzik majd a templát szálhoz szekvenáló primereket hibridizálnak. Következőleg négy egymással párhuzamosan futó szekvenálási reakció közeget állítanak be, amelyek mindegyikébe kerül templát DNS szál, primer, DNS polimeráz és azonos mennyiségben négyféle dezoxinukleotid-trifoszfát (dNTP). Mind a négy reakcióhoz egyszerre csak egyféle didezoxinukleotid-trifoszfát (ddNTP) molekulát kevernek hozzá. A DNS polimerizáció során komplementer dNTP vagy ddNTP épülhet be az új szálba, a nukleotidok koncentrációjának arányától függően. Amennyiben dNTP épül be a szálba a szintézis tovább folytatódik, mert a dNTP 3' végén -H csoport található. Amikor ddNTP

illesztődik be akkor a dNTP 3’ végén lévő reaktív -OH csoport miatt a szintézis leáll. Ezt nevezik a láncterminációs reakciónak (Dovichi és Zhang, 2000). Az újonnan keletkezett, különböző hosszúságú DNS szálak szétválasztása régen poliakrilamid gélen történt, ma azonban kapilláris gél elektroforézissel zajlik. Az automata fluoreszcens szekvenálásnak köszönhetően a láncszintézis csupán egy PCR készülékben megy végbe. A DNS fragmentumok detektálására pedig négyféle fluoreszcens festéket, fluoroflórt használnak (7. melléklet).

A "*de novo* szekvenálás" kifejezés azokra az eljárásokra utal, amelyeknél nem történt meg a DNS szekvenciájának előzetes meghatározása. Ilyenkor a genomi DNS-t restrikciós enzimekkel vagy különböző mechanikai eljárásokkal véletlenszerűen apró darabokra tördelik fel. A kis DNS fragmentumokat plazmid vektorba klónozzák be és a recipiens baktérium sejtekben amplifikálják. Az egyes bakteriális klónokból származó DNS nukleotid sorrendjét meghatározzák, majd a hosszabb szekvencia részeket (contigokat, szupercontigokat) az átfedő DNS-régiók alapján elektronikus úton állítják össze. Az összeszerelt szekvenciában lévő rések feltöltését „genom vagy primer séta (genome walking)” technológiával is elvégezhetik. A genom séta esetében 40-200 Kbp-os DNS darabokból készítik el a szekvenáló könyvtárat. A könyvtárak klónjait a végüktől a középső részük felé haladva, minden fordulóban új specifikus primerekkel szekvenálják meg. A klónokat nagy összefüggő szekvenciákba rendezik és a köztük található réseket az előbb ismertetett módszert alkalmazva töltik fel. Az 1990-es évektől a nagyobb genomi részek leolvasása is lehetségessé vált a „shotgun” eljárásnak köszönhetően, ilyen volt például a *Haemophilus influenzae* nevű baktérium teljes genomjának a szekvenálása (Fleischmann és mtsai., 1995). Ebben az esetben a klónokat két irányból szekvenálják meg olyan módon, hogy a megszekvenált DNS többszörösen fedje le a teljes genomot, így a contigok

közötti rések/gap-ek száma jelentősen lecsökkenthető. A contigok közötti kisszámú gap részt PCR segítségével igyekeznek „eltüntetni”.

A *de novo* szekvenálási eljárások közé tartoznak az úgynevezett új generációs szekvenálási (next-generation sequencing, „NGS”) technológiák. Ezek a Sanger-féle módszerhez képest jelentős időbeli megtakarítást tesznek lehetővé, így jelentősen lecsökkentik a szekvenálási költségeket (8. melléklet). Továbbá nem csupán genomi DNS szekvenálást lehet végezni velük, hanem RNA-Seq és ChIP-Seq készítésére is alkalmasak. Sokféle módszer tartozik ide, amelyek mindegyikéről elmondható, hogy a DNS amplifikációja mindig valamilyen PCR reakciónak köszönhető. A nagyteljesítményű számítógépeken futó illesztőprogramok a szekvenálás során leolvasott egy-néhány száz bázispár körüli szekvenciákból, vagyis a readekből az egymással azonos részleteik alapján klasztereket hozhatnak létre. A klasztereket az egymással átfedő readjeik hosszú „kontinuus” szekvenciákká contigokká kötik össze. Az egymás után következő contigokból pedig még hosszabb szekvenciák, scaffoldok képezhetők (Barta és mtsai., 2016). Az NGS technológiákhoz tartoznak a nagy áteresztőképességű (high-throughput, „HTP”) módszerek. Ilyenkor sok minta szekvenálása zajlik egyidőben, párhuzamosan (9. melléklet). Ennek köszönhetően egy emlős genomja néhány nap esetleg pár óra alatt leolvashatóvá válik. Hátrányuk a Sanger szekvenáláshoz képest, hogy több olvasási hibát ejtenek és leolvasott readek hossza nem haladja meg a néhány száz bázispárt.

A harmadik generációs nanoporus technikánál nincs szükség PCR-re, mert a szekvenálás valós időben történik (real-time). A DNS lánc elektroforézis hatására 1 nanométer átmérőjű membránpóruson megy át, ezzel bázisonként megváltoztatja a membrán elektromos potenciálját. A változást nagy érzékenységű detektor észleli és a szekvenátor készülék ez alapján adja meg a DNS szekvenciáját (Niedringhaus és mtsai., 2011).

Az első lehetőség egy hosszabb DNS szekvenciákból felépülő úgynevezett assembly (contigok, scaffoldok) készítésére az „Align-then-assemble” módszer, ami azt jelenti, hogy a teljes genomi DNS szekvenálása során keletkezett readeket egy referencia genomra illesztik fel. Ez lehet ugyanannak a fajnak vagy egy közel rokon fajnak a genom szekvenciája. Majd a felilleszkedő, jó minőségű readekből contigokat és scaffoldokat (assembly) generálnak. A második esetben „Assemble-then-align” stratégiát követnek, ilyenkor a readekből először *de novo* assemblyt hoznak létre és a keletkezett contigokat rakják rá a referencia genomra (Wang és mtsai., 2014).

A referencia genom mindig haploid és úgynevezett konszenzus pszeudoszekvenciákból épül fel, vagyis több szekvencia közös, összeilleszkedő részei alapján in silico létrehozott, a valóságban nem létező, mesterséges genom szekvencia. A referencia genomok a jelentősebb bioinformatikai, genomikai portálokon elérhetők, letölthetők. A megbízhatóságukat szakemberek ellenőrzik és hagyják jóvá. A kutatók hozzájuk viszonyítják, illesztik saját szekvenciáikat. Mivel folyamatosan frissítik ezeket, így a genomi koordinátáik verziónkként eltérhetnek egymástól. Egy assembly minőségét sokféleképpen megadhatjuk az egyik legelterjedtebb mérőszám az N50. Ez egy súlyozott medián érték, amely azt a scaffold hosszúságot adja meg, aminél hosszabb scaffoldok az összes assembly felét teszik ki.

A genom assemblyk és referenciák nem használhatók annotáció nélkül. A genom annotálása során biológiai információt rendelnek a DNS szekvenciához, illetve annak különböző régióihoz, génjeihez. Ez a biológiai információ lehet strukturális, amikor például egy gén kromoszómális pozícióját adják meg, vagy funkcionális, ilyenkor az adott kromoszóma rész/gén biokémiai folyamatokban való szerepét (kölcsönhatások, reguláció) írják le.

A nagy genom programokat időben megelőzték az egyes gének szekvenciáinak leírásai. Azonban az egyre modernebb és olcsóbb szekvenálási eljárások elterjedésével lehetőség nyílt az élőlények teljes genomjának a megismerésére. A genomika legnagyszabásúbb programja az emberi teljes genom szekvenálása 1986-ban kezdődött az amerikai Energiaiügyi Minisztérium (Department of Energy, DOE) és a NIH-el (National Institute of Health) társulásával. A szervezőmunka után hivatalosan 1990-ben indult el a program. A későbbiekben a szekvenáláshoz a a „shotgun sequencing" módszert használták, ami felgyorsította a folyamatot. További fejlődést eredményezett a Sanger szekvenálásban, hogy négyféle fluorescens festékkel jelölték meg a termékeket és emiatt egyszerre tudták őket futtatni kapilláris elektroforézissel. Az Applied Biosystem 96 kapillárisos szekvenáló automatát fejlesztett ki erre a célra. A bakteriális mesterséges kromoszóma vektor használata nagyobb genom darabok (200 Kbp) klónozását tette lehetővé (https://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011_0025_bio_4/ch35.html). A bioinformatika tudományág nélkülözhetetlenné vált a humán genomszekvenciák értelmezésében és használatában. 2001-ben készült el a humán draft (sok hibát és ismeretlen bázist tartalmazó) genom (Lander és mtsai., 2001, International Human Genome Sequencing Consortium, 2001), majd ezután a HGP 2003-ban közölte az első, jó minőségű 8-9-szeres lefedettségű humán genom szekvenciát. A teljes munka összesen több mint 3 milliárd dollárba került (Hood és Rowen, 2013). A humán genom program robbanásszerű fejlődést hozott az új generációs szekvenálási módszerekben. Napjainkban elegendő nyolc munkaóra egy ember teljes genomjának a szekvenálásához, ami maga után vonja nagy, rendkívüli fontosságú szekvenálási programok például az 1000 genom projekt (2500 különböző etnikumú ember teljes genom szekvenciájának a létrehozása), vagy a Genome 10k projekt (10 ezer gerinces faj megszekvenálása) megvalósulását (Koepfli

és mtsai., 2015). Az NCBI-ra feltöltött genomok száma a rohamosan fejlődő szekvenálási eljárások és az erősen csökkenő költségek miatt a jövőben még inkább növekedni fog (10. melléklet).

2018. január 23.-án a Világgazdasági Fórum partnerségre lépett több genomikai intézettel és nem mindennapi bejelentést tett a svájci Davos-ban. A konferencián kihirdették a Föld Bio-Genome Project (EBP) céljait, amely lényegében a világ minden eukarióta élőlényének (összes növény, állat és egysejtű szervezet) a teljes genom szekvenálása és egy fenntartható biogazdaság létrehozása. Reményeik szerint ezek az intézkedések fontos szerepet játszanak a negyedik ipari forradalom kibontakozásában és 20000 veszélyeztetett faj megmentésében. A projekt előreláthatóan 10 évet és 4,7 milliárd dollárt vesz majd igénybe (Lewin és mtsai., 2018).

Az orvosi, a genetikai kutatásokon és az ökológiai hasznosításon kívül a mezőgazdaságban is nagy jelentőséggel bírnak a haszonállat WGS programok (11/a. és 11/b melléklet).

Sok ország igyekszik a saját természetvédelmileg és/vagy nemzeti identitás szempontjából mérvadó fajai genetikai tulajdonságainak meghatározására. Ausztráliában már 2007-ben kromoszómákba rendezték a kacsacsőrű emlős (*Ornithorhynchus anatinus*) teljes DNS-ét (Warren és mtsai., 2008). Kínában, 2010-ben az Illumina új-generációs szekvenálási platformot alkalmazták az óriás panda (*Ailuropoda melanoleuca*) SOAPdenovo scaffold assembly összeállításához (Li és mtsai., 2010). Honduras nemzeti állatának a fehérfarkú szarvasnak (*Odocoileus virginianus)* a texasi alfaját 2011-ben szekvenálták először (Seabury és mtsai., 2011). A szarvasfélék közül leközölték és az NCBI-ban online elérhetővé tették például az őz (*Capreolus capreolus*), a milu (*Elaphurus davidianus*), a disznószarvas (*Axis porcinus*), a rénszarvas (*Rangifer tarandus*), és az öszvérszarvas (*Odocoileus hemionus*) scaffold assemblyjét, és a magyar gímszarvas kromoszómába rendezett genom

feltöltését követően a Yarkand gímszarvas (*Cervus hanglu yarkandensis*), az indiai muntyákszarvas (*Muntiacus muntjak*) és a kínai muntyákszarvas (*Muntiacus reevesi*) kromoszóma szekvenciáit. A magyarországi gímszarvas genom program nem csupán azért bír nagy jelentőséggel, mert egy fontos nemzeti jelképünkről szól, hanem mert ez Magyarország első igazi, nemzetközileg elismert genom programja, továbbá a szarvasfélék közül a magyar gímszarvas az első a világon, amely kromoszómákba rendezett genommal képviseli magát az NCBI nyilvános adatbázisában.

4.6 BIOINFORMATIKAI HÁTTÉR, SZEKVENCIA ILLESZTÉS, GENOM ANNOTÁCIÓ

A szekvenáló laborok óriási mértékben ontják magukból a nagy mennyiségű szekvencia adatokat. Három, egymással átfedő és szoros kapcsolatban lévő új tudományág foglalkozik ezen adatok kezelésével. A számítógépes biológia egy olyan mérnöki terület, amely informatikai eszközöket, szoftvereket, alkalmazásokat készít a bioinformatika számára. A bioinformatika az élettudomány területéről származó adatokkal dolgozik, megfigyelő, leíró analitikus jellegű tudomány. Informatikai eszközöket és módszereket használ biológiai adatok gyűjtésére, elemzésére, tárolására, biológiai folyamatok megismerésére és ezek alapján biológiai modellek elkészítésére. A klasszikus bioinformatika tárgykörébe tartozik a szekvencia illesztés, a statisztikai analízis (gén hosszúság, CG arány), a genom annotáció (ORF, gén predikció, promóter analízis), szekvencia adatbázisok használata, szekvencia keresés, makromolekulák háromdimenziós modellezése és a fehérjék közötti kapcsolatok feltárása. A modern bioinformatika, a humán teljes genom referencia szekvencia elkészítése után született meg, főként NGS adatokkal foglalkozik. Idetartozik a komparatív genomika, transzkriptomika, proteomika és a rendszer biológia.

Az adatbázisok hasonló minőségű, struktúrált adatok számítógépes tárolására szolgálnak. Fontos kitétel, hogy könnyedén lehessen adatokat keresni és szerkeszteni. Az adatforrások egyben adatbázisok is, biztosítják az adatok online hozzáférhetőségét és ingyenes, közösségi használatát. Az adatbázisok és adatforrások tartalmát szigorú szabályok szerint ellenőrzik és karbantartják. Lényegi tartalmuk alapján lehetnek elsődlegesek, másodlagosak vagy származtatott adatbázisok, harmadlagos, interakciós adatbázisok és úgy nevezett nem szekvenciás adatbázisok. Az elsődleges adatbázisok tárolják a kísérletekből, szekvenálásokból, elsődleges bioinformatikai feldolgozásból származó adatokat, mint DNS, RNS szekvenciákat, a molekulák strukturális információit. Ilyen adatbázisok például az EMBL-Bank (Kulikova és mtsai., 2007), az NCBI GenBank (Benson és mtsai., 2010) és a DDBJ (Sugawara és mtsai., 2008). Léteznek olyan adatbázisok, amelyek az adott DNS szekvencia valamilyen sejtbiológiai jellemzőjét veszik alapul, a SILVAdb-ben a különböző élőlényekből származó rRNS-ek kis és nagy alegység szekvenciái találhatók (Quast és mtsai., 2013), a miRBase pedig nyers és érett mikroRNS szekvenciákat tárol (Griffiths-Jones és mtsai., 2008). A nukleotidokat protein szekvenciává fordítják le a másodlagos, fehérje adatbázisok ilyen például az Uniprot (UniProt Consortium, 2007). A származtatott adatbázisok tárolhatnak transzkripciós faktor kötőhely motívumokat és regulációs adatokat JASPAR (Sandelin és mtsai., 2004), és a TRANSFAC (Matys és mtsai., 2006). A harmadlagos adatbázisok az entitások közötti kapcsolatrendszer leírását adják meg (szabályozás, metabolizmus, szignalizáció, és kötési interakciók). A gének funkció szerinti csoportosításával a GO foglalkozik (Gene Ontology Consortium, 2006). A nem szekvencia típusú adatbázisokban találhatók az irodalmi források, a folyóiratok és a publikációk (NCBI PubMed).

A szekvencia illesztések a szekvenciák közötti hasonlóság kimutatására szolgálnak. A hasonlóság alapján azt feltételezhetjük, hogy a szekvenciák

homológok egymással. Ezek a programok lényegében két vagy több szekvenciát a leginkább hasonló pozícióik összerendezésével illesztenek egymáshoz úgy, hogy a nem illeszkedő részekhez gapeket (hézag, rés) szúrnak be. Az illesztések helyességét pontozással írják le helyettesítési mátrixból vett pontértékek alapján. Azonossági mátrixokat alkalmaznak a nukleinsav szekvenciákra, ilyen esetekben az egyezés pontértéke 1, az eltérésé 0. Természetesen más pontszámokat is adhatnak ezek a programok, ha például hangsúlyozni szeretnék a tranzíció és transzverzió közötti különbséget. A Dayhoff-féle mutációs vagy PAM mátrixok a szelekció által jóváhagyott mutációk alapján számolják ki az evolúciós távolságokat a szekvenciák között (Dayhoff és mtsai., 1979). A BLOSUM mátrixok (BLOcks SUbstitution Matrix) a BLOCKS adatbázison alapulnak, amelyekben fehérje családok többszörösen, gap nélkül összerendezett szekvenciablokkjai találhatók, így klaszterek jönnek létre szekvencia hasonlóság szerint (Henikoff és Henikoff, 1992). A szekvencia illesztések közül mindig azt fogadjuk el optimálisnak, amely az adott pontozómátrix szerint a legnagyobb pontértéket adja. A hasonlóságokat és egyezéseket plusz ponttal jutalmazza, a különbségeket (mismatch), hézagokat pedig negatív pontértékkel bünteti. Két szekvencia összevetése esetén beszélhetünk globális és lokális illesztésről. A globális, Needleman–Wunsch algoritmus a két szekvenciát teljes hosszában, míg a lokális Smith–Waterman algoritmus csak a szekvenciák hasonló régióit illeszti egymáshoz. Minél hosszabb és/vagy több szekvenciát illesztünk annál gyorsabb és nagy problémamegoldó képességű programozási lépésekre van szükség. A dinamikus programozás a komplex feladatokat kisebb alfeladatokra bontja. A progresszív illesztéseknél megjelenik a lokális minimum problémája, mely az algoritmus mohó (greedy) természetéből fakad, vagyis az illesztési folyamatban korábban elkövetett hibákat már nem lehet

utólag kijavítani. A legtöbb szekvencia illesztőprogram fasta formátumú fájlokat kezel.

Az ALLPATHS-LG egy olyan genom összeszerelő program, amely rövid read szekvenciákból mindenféle méretű (akár emlős) teljes genom assembly összeállítását teszi lehetővé (Gnerre és mtsai., 2011). Kifejezetten Illumina szekvenálásból származó 100 bp-os vagy annál nagyobb hosszúságú readek kezelésére tervezték, de PacBio adatokkal is jól boldogul. A működéséhez elengedhetetlen, hogy legalább két, páros (paired) read könyvtár álljon rendelkezésre, amelyekből az egyik 180 bp körüli, a másik körülbelül 3000 bp hosszú inserteket tartalmazzon, és mindkét könyvtárra minimum 45-szörös átlagos lefedettség legyen jellemző. A beimportálandó fájlok lehetnek bam, fastq, fasta, fastb formátumúak. A kimenetként megkapjuk az assemblyt, efasta és fasta kiterjesztéssel. Az ALLPATHS-LG, abban különbözik más, hagyományos assembler-ektől, hogy a polimorfizmusokból, repetitív szakaszokból és az ambivalens szekvenciákból eredő információkat is megőrzi, mivel gráfokat hoz létre.

Stephen Altschul és munkatársai 1990-ben fejlesztették ki az NCBI-nál a BLAST-ot (Altschul és mtsai., 1990). Könnyű, akár online felületeken való használatának köszönhetően rendkívül népszerűvé vált napjainkra, bár a bioinformatikusok leginkább Linux platformokon futtatják ezeket. Heurisztikus algoritmus, amely más szekvencia illesztőknél például a FASTA-nal (Lipman és Pearson, 1985) jóval gyorsabb, de kevésbé érzékenyebb. Éppen ezért nem garantálja az optimális illesztést, viszont kiválóan megtalálja a hasonló szekvenciákat. Statisztikai módszerek alkalmazásával becsüli meg a találatok szignifikanciáját. Az első verziója a BLAST hézagmentes illesztést tett lehetővé, a második verziók (BLAST2: NCBI_BLAST, WU-BLAST) már a hézagos illesztésre is képesek. A program nagy gyorsasággal vet össze úgynevezett kereső „query” szekvenciákat adatbázisokkal „subject”. Első

lépésként az alacsony komplexitású vagy ismétlődő szekvencia részeket távolítja el, „maszkolja ki” a kereső szekvenciából, a fehérjéknél a SEG a nukleinsavaknál a DUST paraméterek bekapcsolásával. Ezután a kereső szekvenciákból meghatározott hosszúságú (W) szavakat, úgynevezett szomszédos szólistákat generál. Az ilyen módon elkészített szólistákból csak azokat tartja meg, amelyek az alkalmazott pontozómátrix alapján a T küszöbérték (threshold) felett találhatók és ezért magas pontszámokat értek el. A hasonlósági pontozó mátrixok közül PAM-ot vagy BLOSUM-ot használ az aminosavaknál, és BLAST merőszámot a DNS-nél. Ezután tökéletes egyezéseket keres az adatbázis szekvenciák és a küszöbérték feletti szomszédos szavak között, és az illeszkedéseket pontozza. Így alakulnak ki az úgynevezett szegmens párok. Ezt követően történik meg a találatok kiterjesztése upstream és downstream irányban, így kapjuk meg a HSP-ket („High-scoring segment pairs”), amelyek összpontszáma, „össz-score” értéke nagyobb, mint a küszöbérték. Vagyis az illesztés addig halad tovább a találati szekvencia két végén, amíg a HSP halmozódott össz-score értéke növekedni tud, illetve a szekvencia végéig. Amikortól a szekvencia szegmens mérete elérte azt a nagyságot, ahonnan az össz-score érték nem tud tovább nőni, onnantól maximális szegmens párnak nevezzük (MSP). Az MSP mérőszáma mindig magasabb, mint a küszöbérték. Legutolsó lépésként a Smith-Waterman algoritmus egyik verziójának a segítségével az alkalmazás megkeresi a legjobb lokális illesztéseket a kereső-szekvencia és a találat között, és meghatározza a rések helyzetét. Kimenetként a beállított paraméterektől függően kaphatunk grafikus és táblázatos jellegű fájlokat, amelyek a két szekvencia maximális szegmens párjait mutatják be. Ezekben ábrázolják, vagy leírják az azonos és a hasonló nukleotidok vagy aminosavak helyzetét és számukat. A két legfőbb statisztikai mutatóból következtetni lehet az illesztés minőségére. Az E-érték (E-value) a szignifikanciát mutatja, vagyis a véletlen hasonlóság mértékéhez

viszonyított várható értéket (expectation) adja meg. Azt írja le mi lehet annak a valószínűsége, hogy az adott adatbázisban a találati hasonlóság a véletlen műve legyen. Tehát az E-érték minél kisebb, annál nagyobb fokú szignifikanciát jelez. A „score" (S) érték hasonlósági pontozómátrix használatával kapott, azonos, hasonló nukleotidok/aminosavak és rés pozíciók alapján adható összpontszám. Mivel az S pontérték (nyers pontérték) jelentése nem egyértelmű, ezért normalizálják, és így kapjuk meg az alignment vagy bit score S' értéket. A Bit score kiszámításának képlete: S'=(λS-lnK)/ln2, ahol K és λ statisztikai paraméterek: Az adatbázis méretétől és a pontozórendszertől függenek, és az S pedig egy empírikusan megállapított cut-off érték. A BLAST algoritmus több típusát különíthetjük el az adatbázis és a kereső-szekvencia egymáshoz való viszonya szerint (12. melléklet). Távoli hasonlóságok keresésére alkalmazzák protein szekvenciák esetén a PSI-BLAST-ot (Position-Specific Iterated BLAST) (Altschul, 1997) és a PHI-BLAST-ot (Pattern Hit Initiated BLAST) (Zhang és mtsai., 1998). Az elsőnél iteratív illesztés történik és a találatok alapján pozícióspecifikus pontozómátrixok jönnek létre. A másodiknál kereső-szekvenciaként reguláris kifejezéseket használhatunk. A MegaBLAST-ot nagy méretű adatbázisoknál érdemes futtatni, ilyenkor hosszabb szegmens párok generálódnak, ami miatt az érzékenység csökken (Morgulis és mtsai., 2008). A BLAST egyik rendkívül gyors alternatívája a BLAT, amely ugyanakkor jóval kevésbé érzékeny algoritmus. A BLAT („Blast Like Alignment Tool"), a BLAST-tal ellentétben nem lineáris keresést végez, hanem az adatbázist indexelt k-mer-ekre osztja, így gyorsabban talál egyezéseket (Kent, 2002). Az olyan online adatbázisok, mint a UCSC szívesen alkalmazzák ezt az algoritmust. A BLAST program család BLASTZ tagja hosszú szekvenciák illesztésére alkalmas (Schwartz és mtsai., 2003).

A LASTZ páros, akár kromoszóma hosszúságú, különböző élőlényekből származó szekvenciák illesztésére használt automatizált program (Harris,

2007). A parancsainak szintaktikája kompatibilis a BLASTZ-jével. Jól működik Macintosh OS X, Linux és Unix platformokon. Maga a program, pipeline (csővezeték) felépítésű, és Python programozási nyelven írták. Tulajdonképpen az eljárás során itt is résmentes HSP-k, HSP láncolatok jönnek létre az adatbázis és a kereső-szekvencia szegmenspárjaiból, de plusz műveletként különféle szűréseket hajt végre a program.

A BWA (Burrows-Wheeler Aligner) egy szoftvercsomag, amelyet arra terveztek, hogy alacsony divergenciájú szekvenciákat térképezzen kiterjedt genomiális részekhez. A működése a Burrows–Wheeler-transzformáción (BWT, blokkrendező algoritmus) nyugszik. Ez valójában egy adattömörítő eljárás, ahol a BWT-vel tömörített sztring (karakterlánc) egyik karaktere sem változtat értéket, csupán a karakterek permutációja történik meg (Li és Durbin, 2009). Ha az eredeti sztring számos, gyakran előforduló rész sztringet tartalmaz, akkor az átalakítás következményeként keletkező új sztring számos helyen fog azonos betűből álló sorokat magában foglalni. Az ismétlődő karakterek, mintázatok teszik lehetővé magát a tömörítést. A BWA három alapvető algoritmusból épül fel: BWA-backtrack, BWA-SW és BWA-MEM. Az első algoritmust Illumina szekvenciákra tervezték 100 bp-ig, míg a másik kettőt a 70 bp-tól 1 Mbp-ig tartományba esőkre. A BWA-SW az olyan heurisztikus programcsomagokkal ellentétben, mint a BLAST, képes az összes lehetséges optimális illeszkedést megtalálni az által, hogy indexeléssel gyorsabbá teszi a Smith-Waterman algoritmust. A BWA-SW és a BWA-MEM hasonló funkciókat kínál. A legfrissebb BWA-MEM a kiváló minőségű lekérdezésekhez ajánlott, mivel gyorsabb és pontosabb és a teljesítménye is jobb. Minden algoritmus esetében a BWA először elvégzi a referencia genom indexelését, amely lehetővé teszi a tömörítési eljárást.

A MUMmer egy nyílt forráskódú, Linux rendszeren működő szoftvercsomag. Hosszú DNS vagy aminosav szekvenciák gyors összeillesztésére szolgál,

vagyis teljes genomok összehasonlítását is lehetővé teszi. Legnagyobb részt C és C++ program nyelven íródott. A minták tökéletes egyezését utótag „suffix” fa adatstruktúra segítségével keresi meg. (A suffix az informatikában használt kifejezés, amely az adatbázisokban a mezők azonosítására szolgáló rövid nevet vagy betűcsoportot jelenti. Keresésnél egy szó vagy kifejezés mögé írva biztosíthatjuk, hogy csak a megadott adatmezőben keressen a gép.)

A MUMmer (Maximal Unique Matcher) korábbi verzióinak célja az volt, hogy a két bemeneti szekvencia közötti minimális hossz maximális és egyedi pontosságú egyezését megtalálja (MUM szakaszok) (Delcher és mtsai., 1999). A munkám során használt verzió (3.0) teljes egészében figyelmen kívül hagyhatja az egyediséget, de megadhatjuk kapcsolóként a parancssorban (Kurtz és mtsai., 2004). A leggyakrabban használt MUMmer csővezetékek „pipeline”-ok a NUCmer, PROmer, run-mummer1 és run-mummer3 (Delcher és mtsai., 2002). A fenti parancsok három fő műveleti szakaszra bonthatók. Az első rész a két bemeneti szekvencia közötti maximális, pontos egyezést adja meg, a második szakasz ezeket az egyezéseket olyan csoportokba osztja, amelyek megbízhatóan összeilleszkedő, horgonypontokat hoznak létre, a harmadik végső szakasz kiterjeszti ezeket a lehorgonyzott szekvenciákat és klasztereket készít belőlük. A NUCmer (NUCleotide MUMmer) egy felhasználó barát Perl szkript pipeline, amely a nagyon hasonló, többszörös DNS szekvenciák illesztésére szolgál. A PROmer is Perl alapú, a NUCmer-hez nagyon hasonló alkalmazás, ugyanakkor megengedi a bemeneti DNS szekvenciák különbözőségét. A run-mummer1 és run-mummer3 cshell-ben íródott pipeline-ok.

A MAKER Perl programozási nyelven készült pipeline-okból álló genom annotáló program. A MAKER UNIX, Linux és OS X operációs rendszereken működik. Minimális erőforrás igényű és kevés bemeneti fájlt is jól hasznosít. Egyaránt alkalmas prokarióta és eukarióta élőlények *de novo* és homológia

alapján történő gén meghatározására, illetve ezen eredmények összefésülésére. Nagy pontossággal képes leírni a gén exon-intron szerkezetét, doménjait és a különböző motívumokat. Azonosítja a genomiális snRNS-ket, repetitív szakaszokat és transzpozonokat. Működése során, az egymásra épülő műveleti fázisokban különböző programokat hív be. Az egyes lépések kimenetei alapján alakítja át gén predikciós algoritmusait. A számítási fázisban RepeatMasker program lágy maszkolást (a repetitív nukleotidokat kisbetűs karakterekkel írja le) végez az alacsony komplexitású ismétléseknél (Korf és mtsai., 2003). Saját belső transzpozon könyvtárával BLASTX alkalmazással hasonlítja össze a genomi mobilis elemeket. A felhasználó által is jóváhagyott és feltöltött különböző élőlényekből származó EST, mRNS, cDNS, és transzkriptom keresőszekvenciákat felilleszti („blastolja") a célgenomra. Érdemes a legjobban annotált genomok (például ember) és egy közel rokon faj EST szekvenciáit belevenni a vizsgálatba. Az blastolás eredményeképpen kapott kódoló régiók, csupán a valóság durva megközelítését tükrözik, ezért a „splicing" helyek (exon összeillesztő helyek) meghatározásához a MAKER egy újabb programot az Exonerate-et hívja be (Slater és Birney, 2005). A szűrési és klaszterezési fázisban zajlanak a pontszámok és a százalékban megadott azonosságok szerint történő szekvencia igazítások, eltávolítások, és az átfedő szakaszok találatainak egyazon adatcsoportba rendezése. Ezután következik a BLAST illesztések pontosítási lépése. Majd a szintézis fázisában az eddig kapott eredményeket fésüli össze a program, pontszámokkal látja el a nukleotidokat pozíciójuk, keresőszekvenciához való hasonlóságuk alapján. Ilyen módon keletkezik egy becslés az UTR-k, intergénikus, kódoló régiók, exonok és intronok helyzetére vonatkozóan, amit továbbít a SNAP program számára (Cantarel és mtsai., 2008). Ezen információk alapján a SNAP módosítja belső Hidden Markov Model (HMM) által készített modellt. Amikor a MAKER nem talál a kereső EST és protein szekvenciák BLAST illesztésével

semmit akkor a SNAP elsődleges ab initio gén-predikcióját használja fel. A végső annotációs fázisban a program az elkészült SNAP gén meghatározásokat minden EST-vel és mRNS-el szemben leellenőrzi, és megadja az exonok, 5' és 3' UTR-ek pontos genomi koordinátáit. A kimeneti fájlok táblázatos formában (GFF3 formátum) tartalmazzák az annotációkat.

A tRNAscan-SE program online felületen és linux platformon is használható, amely genomiális tRNS-t kódoló szekvenciák keresésére és annotálására szolgál (Schattner és mtsai., 2005).

A genom-böngészők a teljes genom vizualizációját teszik lehetővé. A genom DNS szekvenciája egy kétdimenziós. bázispárok szerint felosztott számegyenesen jelenik meg scaffold, contig és akár nukleotid szinten. A számegyeneseken a genom annotáció is láthatóvá tehető. A legtöbb nagyobb online elérhető genom adatbázis rendelkezik böngészővel (ENSEMBL, UCSC és NCBI). Ezekre az egyszerű felhasználók is feltölthetik saját szekvencia adatukat.

A bioinformatikai munka során használt programokat az 1. táblázatban foglalom össze.

1. táblázat. Bioinformatikai munka programjai.

Név	Cél
ALLPATHS-LG	Gímszarvas readek scaffoldokká szerelése.
BLAST	Térképpont markerek illesztése szarvasmarha referencia genomhoz, és gímszarvas scaffoldokhoz.
MegaBLAST	Gímszarvas rRNS, tRNS és miRNS géneket tartalmazó scaffoldok azonosítása (nem referencia géneket tartalmazó scaffoldok).
LASTZ	Scaffoldok illesztése szarvasmarha referencia genomra.
BWA-MEM	
MUMmer	
EMBOSS seqret	Szarvasmarha kromoszómák átalakítása a gímszarvas kapcsoltsági géntérképnek megfelelően.
RepeatMasker	Szarvasmarha és gímszarvas alacsony komplexitású repetitív szekvenciáinak kimaszkolása.
SeqMan	Gímszarvas contigok újrafűzése új scaffoldokká. Az allpath-al rosszul összefűzött scaffoldok esetében.
MAKER	Gének azonosítása a gímszarvas genomon.
tRNAscan-SE	TRNS kódoló gének azonosítása a gímszarvas genomon.
Barrnap	Gímszarvas RNS 5S alegységének azonosítása.
SAMtools	Gímszarvas genetikai variánsok keresése.
BCFtools	Gímszarvas VCF (variant-calling format) fájlok létrehozása.
ANNOVAR	Gímszarvas aminosav funkcionális változásokat okozó variánsok annotálása

4.7 SZARVASMARHA REFERENCIA GENOM

A szarvasmarha (*Bos taurus*) rendkívül fontos mezőgazdasági haszonállat. Nagy jelentőséggel bír a hús- és a tejiparban egyaránt. Fontos modellállata számos orvosi kutatásnak és szekvenált pszeudogenomja révén az összehasonlító genomikának is (https://www.ncbi.nlm.nih.gov/genome/?term=txid9913[orgn]).

Ez idáig 13 teljes, kromoszómába rendezett szarvasmarha referencia pszeudogenom szekvencia található meg az NCBI oldalán. A legrégebbi ezek közül a 2009-es UMD Bos_taurus 2.0, a legújabb pedig a 2018-as ARS-UCD1.2 referencia. Megjegyzendő azonban, hogy ezek nagyrésze 3

különböző kutatóintézet által elkészített háromféle assembly egymásra épülő, továbbfejlesztett változata. Jelenleg a legfrisebb verziók: Bos_taurus_UMD_3.1.1, Btau_5.0.1, ARS-UCD1.2 (Merchant és mtsai., 2014, Bovine Genome Sequencing and Analysis Consortium, 2009). Munkám során eleinte a Baylor Btau_4.6.1 változattal, végül pedig a Btau_5.0.1-el dolgoztam. Mindkét pszeudogenomot a Cattle Genome Sequencing International Consortium készítette el. A Btau_5.0.1 2,7 Gbp hosszúságú szekvenciája 29 haploid autoszómába, 1 X és 1 Y kromoszómába rendeződik. Az unplaced kategóriába kerülnek azok a scaffoldok, amelyek kromoszómális helyzete nem meghatározható. Más szekvenált gímszarvas rokon fajokkal összehasonlítva a szarvasmarha a legjobban annotált genommal rendelkezik. Ezenfelül kutatócsoportunk korábbi cikkeiben (Gyurján és mtsai., 2007, Stéger és mtsai., 2010) megállapította, hogy a bemutatott „zoo-cloning" szekvenáló eljárás eredménye alapján a szarvasmarha és a gímszarvas gének kódoló régiója 98-100 % körüli azonosságot mutat, és több mint 90%-os egyezés figyelhető meg a gének 5' regulatórikus promóter régiójában. Ezen okok miatt a szarvasmarha genom rendkívül hiteles templátot biztosít a gímszarvas scaffoldok sorba rendezéséhez, és génjeinek feltárásához.

A két faj pszeudogenomjának összehasonlításán alapuló vizsgálatok, jól működnek a gén szekvenciák esetében, viszont a gímszarvas kromoszómák centromeron helyzetének meghatározásánál nem alkalmazhatók. A szarvasmarha referencia pszeudogenom szekvenciákban nincsenek megadva a centromeronok pozíciói. A genom szekvenálási módszerek általános elterjedése előtt a *B. taurus* és a *C. elaphus* kariotípusát (hasonlóan a többi szarvasféléhez) klasszikus citogenetikai módszerekkel tárták fel (Gustavsson és Sundt, 1968, Fontana és Rubini, 1990, Bonnet és mtsai., 2001). A korábbi kromoszóma citológiai vizsgálatok centromeron festéssel és in situ DNS hibridizációk alapján határozták meg a centromeronok és a hozzájuk közel eső

gének helyzetét (Ma és mtsai., 1996, Everts-van der Wind és mtsai., 2004, Everts-van der Wind és mtsai., 2005). A szarvasmarha 58 akrocentrikus (A kromoszóma) autoszómával rendelkezik és a nemi kromoszómái szubmetacentrikusak (SM kromoszóma) (Di Meo és mtsai., 2005). A gímszarvas kromoszómák felépítése meglehetősen „primitív" jelleget mutat, mivel szinte az összes akrocentrikus (A kromoszóma). A Ce5 metacentrikus (M kromoszóma) a *Cervinae* ág létrejöttekor két akrocentrikus kromoszóma Robertsoniális transzlokációjával jött létre. A *Pecora* leszármazási vonalak (*Bovidae, Cervidae*) kariogramjai alapján, az evolúciójuk során nagyon gyakran történhetett tandem összekapcsolódás úgynevezett Robertsoniális fúzió és hasadás, azaz amikor 2 akrocentrikusból 1 akrocentrikus kromoszóma képződik és vice versa. A szikaszarvas (*Cervus nippon)* a gímszarvas legközelebbi rokonának, alfajának tekinthető, ezért a szikaszarvas és a szarvasmarha RBG sávmintázatát (Giemza festéssel nyert mintázat továbbfejlesztett változata) vizsgáló, citogenetikai komparatív elemzések eredményei (Bonnet és mtsai., 2001) jól hasznosíthatók a gímszarvas tekintetében is. E tanulmány alapján a *B. taurus* és a *C. nippon* kromoszómális elrendeződése sok pontban hasonlóságot, homológiát mutat. Egymással egy az egyben megfeleltethető kromoszómák, összeolvadások és szétválások figyelhetők meg.

5 A DISSZERTÁCIÓ CÉLKITŰZÉSEI

A fő célom az volt, hogy a Gímszarvas/Csodaszarvas Genom Program keretén belül létrehozzam a „Gímszarvas Referencia Genom szekvenciát (CerEla1.0), ami a világ első gímszarvas, és egyben a Magyarországon készült első, teljes, emlős referencia genom szekvenciája. A munkám során a következő részcélokat tűztem ki.

1. Megfeleltetni egymással a rendelkezésre álló gímszarvas genetikai térképet (Slate és mtsai., 2002) és a *de novo* gímszarvas szekvencia vázakat (genom assemblyt/scaffoldokat).
2. Egyeztetni egymással a gímszarvas genetikai térkép pontjait (DNS markereit) és a szarvasmarha referencia genom ortológ szekvenciáit.
3. Azonosítani a szarvasmarha referencia gének ortológ szekvenciáit a gímszarvas scaffoldokon.
4. Összehasonlítani és felilleszteni a *de novo* gímszarvas scaffoldokat a szarvasmarha referencia genomra az evolúciós változásokat figyelembe véve.
5. Megkeresni a gímszarvas genom fehérje kódoló génjeit, repetitív szekvenciáit, transzfer RNS, riboszómális RNS és mikro RNS szekvenciáit.
6. Meghatározni a gímszarvas kromoszómák centromeron pozícióit.
7. Az egyedi azonosítóval ellátott gímszarvas teljes genom feltöltése és online elérhetővé, letölthetővé tétele az NCBI szerverén.

6 A GÍMSZARVAS GENOM ÖSSZEÁLLÍTÁS ELŐZMÉNYEI

6.1 MINTAGYŰJTÉS, DNS IZOLÁLÁS

A gímszarvas genom program megvalósulása több hazai egyetem, kutatócsoport, és kutató együttműködésének a következménye. Emiatt fontosnak tartom, hogy a bioinformatikai munkát megelőző fázisokról is szót ejtsek.

A minta a Kaposvári Egyetem Vadgazdálkodási Tájközpont Bőszénfai Szarvas farmján, egy természet közeli körülmények között élő 7 éves, kapitális gímszarvasbikából (fülszáma: Crot. N.o. 3016) származott. (A kapitális agancsos egyed definíció szerint rendkívül erős agancsú szarvasbika, akinek trófeája legalább 170 CIC pontot érne el, vagyis legalább bronzérmes lenne.) A Crot3016 szarvasbika aranyérmes trófeát fejlesztett, ami a 210 CIC ponthatárt jelentősen meghaladta volna hullott agancsai alapján, és a 240 CIC pontot is megközelítette.

A 3x10 ml vér levételét állatgyógyászati gyakorlat szerint, a Magyar Állatgyógyászati Kamara előírásait figyelembe véve (243/1998, XII.31) állatorvos végezte el. A minta EDTA pufferbe került és a felhasználásig -20°C-on tároltuk. A teljes genomi DNS kivonása a vérmintából történt Duplicα Prep automata the Duplicα Prep Automated DNA/RNA Extraction System (EuroClone S.p.A., Olaszország) kitt alkalmazásával.

6.2 ILLUMINA SZEKVENÁLÁS

Az izolált DNS mintákat elküldtük az Aros Applied Biotechnology nevű cégnek (Aarhus, Dánia), és a teljes genomi DNS-t szekvenáltattuk Illumina HiSeq2000 szekvenáló platformon. Az Illumina szekvenálás folyamata során

a teljes genomi DNS-t 200-500 bp-os darabokra törik szét. A dupla szálú DNS két végét úgy javítják ki, hogy a ragadós végek eltűnjenek. A 3’ végekhez egy adenint kapcsolnak, amihez ezután egy timin túlnyúló véggel rendelkező adapter DNS-t ligálnak, majd a DNS-t NaOH-os kezelésssel egyszálúsítják. Az egyszálú DNS-ket egy „flow cell”-re vagyis szekvenáló lemezre viszik fel. Minden flow cell-ben 8 cső, vagyis „lane” fut végig. A csövekben a DNS fragment mindkét végén lévő adapterrel komplementer szekvenciák horgonyozódnak ki. Ezekhez hibridizáltatják a DNS szálakat. Itt történik az első PCR reakció, aminek a végén a kapott dupla szálú DNS templát szálát (amit eredetileg a horgonyra kötődnek) formamiddal leválasztják és lemossák. Az egyszálú DNS elhajlik, és a szabad végével egy másik lehorgonyzott primerrel kapcsolódik és végbemegy az úgynevezett hídamplifikáció (bridge amplification). Ilyenkor az átírt szálak egymással vagy egy másik primerrel hibridizálhatnak és a PCR folyamat révén duplikálódnak. A keletkező PCR termékek klaszterekbe rendeződnek. A PCR állandó hőmérsékleten (60°C) megy végbe a megfelelő puffer és denaturálószer jelenlétében. A reagenseket minden egyes ciklus végén lemossák. Az utolsó PCR és a végső denaturáció után jön a szekvenálás, ami szintén szintézissel történik. A szekvenálás során a templáthoz egy primert ligálnak, majd a reakcióhoz egyszerre adják hozzá a különböző fluorescens jelölésekkel ellátott 4 nukleotidot. Amikor a szintézis során egy nukleotid beépül a keletkező DNS szálba, akkor a rá jellemző, felvillanó fluorescens fényt CCD kamera észleli. A szabadon maradt nukleotidokat, a kémiai úton eltávolított fluorofórt és a 3' blokkolót lemossák. A szekvenálási folyamat során először az első primertől a fragment szekvencia felé történik meg a körülbelül 100 bázispáros leolvasás, így keletkezik az első read. Ezt követi az első index szekvencia, a második index szekvencia és második read szintézis. A megszintetizálódott read termék lemosása után, az első index szekvenciát is megszintetizálják, leolvassák, lemossák, majd a

fragment szál lehajlik a kihorgonyzott komplementer oligóhoz és így lehetővé teszi a második index megszekvenálását, ezután következik a keletkezett rövid termék eltávolítása és azt követően hídban lévő fragmenthez komplementer szál szintetizálódik. Utána szét választják a szálakat, az eredeti szálat lemossák és a reverz szálról a második primerrel kezdődően elkészül a read párja. A több milliónyi, 100 bázispáros read jellemző az adott fragmentre, és az egyedi index szekvenciák segítségével azonosítható. Az azonos readek klaszterekbe rendezhetők. A klasztereket forward és reverz readek alkotják. Az előbbiekben vázolt folyamatban úgynevezett paired-end read szekvenciák jöttek létre. A mate pair szekvenciák sokkal nagyobb közbenső részt, vagyis insertet határolnak, mint a pair end readek, így a későbbiekben sokat segíthetnek a contigok összeillesztésénél, és a scaffoldok generálásánál (https://www.eurofinsgenomics.co.in/en/eurofins-genomics/product-faqs/next-generation-sequencing/general-technical-questions/what-is-the-principal-of-the-illumina-sequencing-technology.aspX) (13. melléklet).

A mate pair szekvenálásnál 2-5 Kbp nagyságú DNS fragmentumokat hoznak létre, amelyek végét biotinilálják majd a szálat cirkularizálják és végül kisebb darabokra tördelik. Az új fragmentumok közül néhány tartalmazza mindkét biotinilált mate pair szegmenst insert szekvencia nélkül. A további lépések a paired-end szekvenáláshoz hasonlók (Bentley és mtsai., 2008).

Az illumina szekvenálás eredményeként keletkezett readek 101 bázispár hosszúságúak lettek, a paired-end szekvenciáknál 500 bázispáros, a mate pair szekvenciáknál pedig 1,2 kilobázispár nagyságú volt az insert méret. A readekből összesen 4 paired-end és 2 mate pair szekvencia könyvtár készült el. A vegyes könyvtárak (paired-end és mate pair együtt) létrehozása igen előnyös a *de novo* assembly összeállítása során, mert kevés gappal rendelkező nagyobb méretű contig és scaffold szekvenciák is generálhatók belőlük. A paired-end és a mate pair szekvencia könyvtárban nyers szekvenciaként összesen több

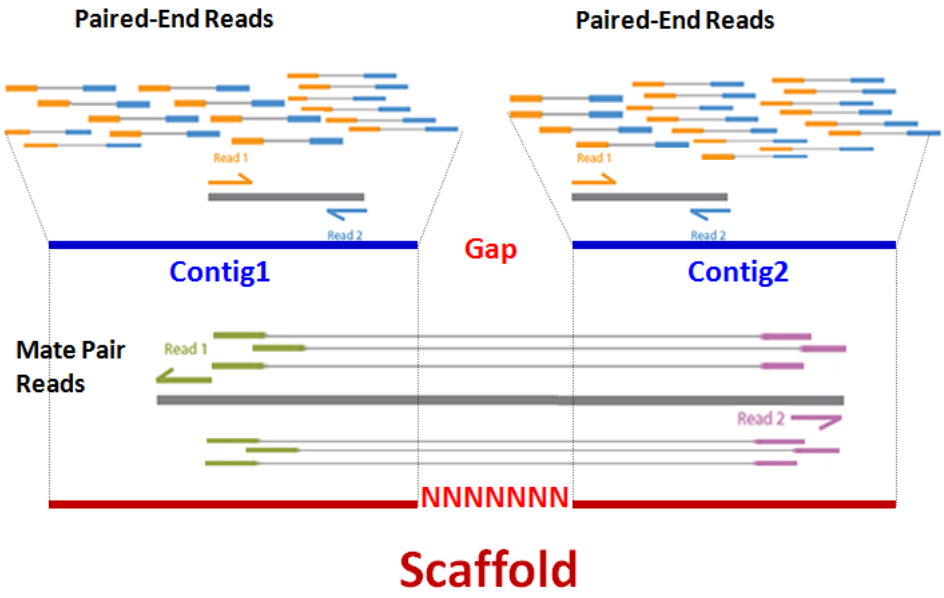
mint 2 milliárd read szekvencia található, ami 223 milliárd bázispár hosszú DNS-nek felel meg. Ez a gímszarvas haploid genomra vonatkoztatva 74-szeres átlagos lefedettséget jelent (2. táblázat).

2. táblázat. Gímszarvas read szekvencia statisztikák.

Read szekvencia könyvtárak		**Readek száma**	**Bázisok száma (Gbp)**	**Insert méret**	**Read méret**
Paired-end	Ce_PE_s1_1	202009674	20,40	300-500 bp	101 bp
	Ce_PE_s1_2	204044499	20,61		
	Ce_PE_s2_1	184790536	18,66		
	Ce_PE_s2_2	188274662	19,02		
Mate pair	Ce_MP_1	163020536	16,47	1,2 Kbp	
	Ce_MP_2	160624627	16,22		
Összesen	1x	1102764534	111,38	-	-
	2x	**2205529068**	**222,76**	-	-

6.3 CONTIGOK ÉS SCAFFOLDOK ÖSSZESZERELÉSE

A contig és a scaffold assemblyt, vagyis a readek hosszú, egybefüggő szekvenciákká szerelését Dr. Barta Endre végezte. Az összeszerelés során a Broad Institute ALLPATHS-LG programmal, alap beállítású paraméterekkel dolgozott. Az egymást átfedő paired-end readek révén összefüggő, kontinuus szekvenciák, úgynevezett contigok jöttek létre. Bizonyos részeken nincsenek átfedő readek, ilyenkor a mate pair read párok tagjai (melyek egymástól 2-5 Kbp távolságra találhatók) adták meg az azonos scaffoldra eső contigokat, vagyis a scaffoldok mate pair read párokkal összefűzött contigokból álltak össze. Néhány helyen azonban az ALLPATHS-LG nem talált átfedő readeket, ilyenkor ezeket a contigok közötti ismeretlen bázisokat N-nel jelölte (gap régiók) (5. ábra).



5. ábra. A contig és scaffold assembly *de novo* létrehozásának sematikus rajza.

A program az SGI UV 1000 gépen (1152 Intel CPU mag és 6 TB memória) futott le. Összesen 437412 darab contig keletkezett, amelyek N50 értéke 7,5 Kbp nagyságú. A contigokból összefűzött scaffoldok száma 34724, N50 értékük 430 Kbp (a teljes scaffoldokra nézve) hosszúságú. A legkisebb scaffold mérete 613, a legnagyobbé pedig 4487240 bp (14. melléklet) volt.

7 ANYAG ÉS MÓDSZER

A kutatási munkám fő irányvonalát a gímszarvas referencia genom kromoszóma szintű összeállítása, annotációja, centromeron pozícióinak meghatározása és evolúciós viszonyainak feltárása képezte. A társszerzős cikkemben leírt gímszarvas mitokondriális genom összeállításában főként irodalmazási és bioinformatikai ellenőrzési feladatokat láttam el. A munkahelyi védésen elhangzott opponensi javaslatok alapján a mitokondriális genom elkészítésének folyamatát ebben a doktori disszertációban nem fejtem ki.

7.1 FELHASZNÁLT SZÁMÍTÓGÉPEK, ADATBÁZISOK ÉS GÉNTÉRKÉP

A munkám elvégzéséhez szükséges programokat Linux platformokon, nagyrészt a gödöllői Mezőgazdasági Biotechnológiai Kutatóintézet szerverein futtattam le. Ezenfelül igénybe vettem a Debreceni Egyetem Genomi Medicina és Bioinformatikai Szolgáltató Laboratóriumának ngs szerverét is. A számítógépek fő jellemzőit a következő táblázatban foglalom össze: (3. táblázat).

3. táblázat. Felhasznált gépek fő jellemzői.

Szerver neve	**z800**	**emboss**	**genome**	**ngs**
Helyszín	Gödöllő	Gödöllő	Gödöllő	Debrecen
Operációs rendszer	Ubuntu	Ubuntu	Ubuntu	Red Hat linux
CPU	8,00	4,00	8,00	40+88+24+7×12
Memória (GB)	74,21	16,43	49,46	3×128+256+6x48
Külső tárhely	-	-	-	2×12 darab 2TB

A kutatás során a kevésbé ismert kettős referencia vezérelt genom összeállítás módszerét alkalmaztam. Ez azt jelenti, hogy a rendelkezésre álló gímszarvas kapcsoltsági géntérképét (Slate és mtsai., 2002) és egy közel rokon faj, a

szarvasmarha referencia genom szekvenciáját egyaránt felhasználtam a scaffoldok sorrendjének, orientációjának megállapítására, hogy ily módon gímszarvas kromoszómába tudjam rendezni azokat. A rokon fajok közül a házi juh genom (ISGC Oar_v3.1), kevéssbe jól annotált, ezért a későbbiekben csupán a kérdéses illesztéseket tisztáztam vele (International Sheep Genomics Consortium, 2010). A szarvasmarha referencia genomok közül több változattal (Baylor Btau_4.6.1, Btau_5.0.1) is dolgoztam, de végül a Btau_5.0.1 referenciára kapott eredményeket használtam fel, illetve ezzel a referenciával egyeztettem össze a korábbi megoldásaimat. Első lépésként a gímszarvas géntérkép pontjainak DNS szekvenciáit név és azonosító alapján irodalmi adatokból és online elérhető Ensembl, UCSC, Uniprot, NCBI, és ENA adatbázisokból gyűjtöttem ki.

7.2 GÍMSZARVAS GENOM ÖSSZEÁLLÍTÁSA

7.2.1 *MMSc keresés*

A munkám második lépéseként a generálódott scaffoldokat gímszarvas térképpontokhoz és szarvasmarha genom szekvenciákhoz rendeltem hozzá. Az összes gímszarvas scaffoldból és a szarvasmarha referencia genomból makeblastdb paranccsal külön-külön BLAST könyvtárakat készítettem. A markereket úgynevezett query vagyis kereső multifasta fájlokba fűztem. A mikroszatellitákat háromféle módon gyűjtöttem össze: 1) Csupán a mikroszatellita forward és reverse primer szekvenciáit írtam ki. 2) A mikroszatellitákat primereikkel együtt adtam meg. 3) A mikroszatellitákat a primereikkel és 500 bp-os flanking régióval raktam be a keresőszekvenciákba. Az RFLV-ként azonosított gének teljes szekvenciáját és cDNS-eit egyaránt alkalmaztam. Az EST markereket csak önmagukban használtam fel. Az illesztés a szarvasmarha és a gímszarvas-scaffold BLAST könyvtárakra

mindenesetben a blastn (Altschul és mtsai., 1990; Korf és mtsai., 2003) paranccsal a következő kapcsolókkal történt.: -evalue 1e-10; -outfmt 6; -best_hit_score_edge 0.05;
-best_hit_overhang 0.25; -perc_identity 50; -max_target_seqs 3. A Slate-féle térképponti markereket tartalmazó multifasta fájlba a 10 darab DeerPlex mikroszatellitát (DeerPlex: Szabolcsi és mtsai., 2014) is becsatoltam. Az RFLV marker gének cDNS és az EST marekerek szarvasmarha genomon kapott BLAST találatait tblastx programmra írt szkripttel ellenőriztem le. Az utolsó marker csoportot a fehérjék alkották, emiatt egy másik multifasta fájlba kerültek. Ebben az esetben a keresőszekvenciák aminosavak, ezért tblastn-t alkalmaztam a blastn helyett, és az E-értéket 1e-5-re csökkentettem. Az AFLP markerekhez nem találtam szekvenciákat az adatbázisokban, így ezekkel a továbbiakban nem foglalkoztam. A BLAST eredményeket tovább szűrtem, sorba rendeztem (parancs: sort) és az átfedő pozíciók alapján összevontam (paraméter: --merge). Így kaptam meg a legjobb BLAST találatot minden egyes markerre nézve. Az ilyen módon kihalászott scaffoldokat „mapmarker vagyis térképpont” scaffoldoknak (MMSc) neveztem el. A géntérképi pontok kiterjedt (scaffoldnyi) szekvencia környezetbe kerültek. Továbbá a gímszarvas térkép pontok megfelelőit azonosítottam a szarvasmarha teljes genom szekvenciában blastolással. Majd az MMSc-ket is ráillesztettem MUMmer3.0 és BWA 0.7.10-r789 programokkal megerősítésképpen. Azaz az összehasonlító géntérképezési elvet használtam a gímszarvas felől szarvasmarha felé.

7.2.2 *RGSc keresés*

A Slate-féle géntérkép már megfeleltette egymással a gímszarvas kapcsoltsági csoportokat és a szarvasmarha ortológ kromoszómákat. Az MMSc-ék illeszkedési pozíciói szépen visszaigazolták ezeket a feltevéseket. Ezenfelül a

gímszarvas géntérkép pontjai hosszú szakaszokon azonos sorrendben, kolineárisan helyezkedtek el a szarvasmarha genomban is, azaz kiterjedt szinténiákat, lokális kapcsoltságokat tapasztaltam. Mivel a gímszarvas referencia térképi pontok szekvenciája csaknem megegyezik a két fajban, továbbá ezen pontok sorrendje azonos, valamint a térközök/térkép szegmensek hosszai is arányosak, és az egyes ortológ gének szekvenciái pedig csaknem azonosak, ezért ezeket az ortológ szarvasmarha géneket használtam a gímszarvas scaffoldok halászatára a továbbiakban, vagyis az összehasonlító géntérképezési elvet alkalmaztam a szarvasmarha felől a gímszarvas felé. A szarvasmarha gének pozícióit UCSC online adatbázisból (Tools/Table browser menü pont) szedtem ki kromoszómánként. (A munkám kezdetén 2012-ben a Baylor Btau_4.6.1 assembly volt a legfrissebb elérhető szarvasmarha referencia genom, ebből kifolyólag csupán Baylor Btau_4.6.1 genom USCS Refseq génjei álltak rendelkezésre, kezdetben tehát ezeket használtam. A Btau_5.0.1 assemblyt csak 2015-ben tették elérhetővé, ekkor nyílt lehetőség a két referencia genom assembly génjeinek az összehasonlítására is.) A kapott szöveges fájlokat bed formátumúvá alakítottam át. A szarvasmarha referencia genomból bedtools getfasta paranccsal kivettem a bed pozícióknak megfelelő szekvenciákat, amelyekből aztán multifasta queryt hoztam létre (15.-16. melléklet). A keresőszekvenciákra a gímszarvas scaffold BLAST könyvtárral szemben megablastot futtattam le. A 3 legjobb találatot kértem le és táblázatos formátumot állítottam be a következő paraméterekkel -v 3 -b 3 -m 9. Az ilyen módon kihalászott scaffoldokat reference gene containing scaffolds-nak (RGSc) neveztem el. Az RGSc-kből óriás excel táblákat hoztam létre, amelyekben a név, score, E-érték, hossz, és orientáció adatokat adtam meg. Az egy génhez tartozó 3 scaffold közül a legmagasabb score és a legalacsonyabb e-value értékűeket fogadtam el. Megerősítésképpen az RGSc-k elejét és végét,

vagyis körülbelül 1000-1000 bp nagyságú részeket visszaillesztettem a szarvasmarha referencia szekvenciára.

7.2.3 *A szarvasmarha referencia genom átalakítása*

A két faj közötti evolúciós távolság miatt eltér a kromoszóma szám, valamint kromoszóma átrendeződések figyelhetők meg. Így kénytelen voltam a cél szekvenciákat (Btau_5.0.1.fasta) a gímszarvas kapcsoltsági térképnek megfelelően módosítani. 19 autoszóma és az X, Y kromoszómák egyértelműen egyeznek egymással a két rokon fajban. Viszont 2 gímszarvas kromoszómánál a markerek alapján kettéhasadások láthatók és emiatt 4 szarvasmarha kromoszómával feleltethetők meg. Ettől eltérően 6 esetben 2 gímszarvas kromoszóma 1 szarvasmarha kromoszómával ortológ. A Bt1 (*Bos taurus* kromoszóma 1) és a Ce19 és Ce31 (*Cervus elaphus* kromoszóma 19 és 31) összeolvadása olyan szempontból különleges, hogy a Ce19-en belül történt egy transzlokáció, ami egy 8 markert magában foglaló rész és egy 3 markeres rész felcserélődését jelentette az eredeti sorrend megtartásával. Ezen evolúciós változások pontosabb meghatározására a centromeron fejezetben térek ki. Néhányszor a kromoszómán belül, két szomszédos marker között történt meg a genomi részek megfordulása, ezekkel az apróbb transzverziókkal azonban csak a munkám legvégén foglalkoztam. Előfordult még az is, hogy Slate-féle téképpontok sorrendjével ellentétes irányban helyezkedtek el a markerek szarvasmarha referencia genomon, ilyenkor a Slate kapcsoltsági térképet tekintettem mérvadónak (17. melléklet). A szarvasmarha referencia kromoszómák módosításának első lépése a kemény maszkolás (hard-masking), ami lényegében az alacsony komplexitású genomi helyek és repetitív szekvenciák nukleotidjainak az „eltüntetését” N vagy X karakterre történő cseréjét jelenti. A kis komplexitású genomi szekvenciák vagy a repetitív elemek, a hosszabb kereső szekvenciák (scaffoldok) illesztésekor

hamis eredményeket produkálhatnak. A program futattásához a -species artiodactyl paramétert használtam. A maszkolt cél-szekvencia referenciát az EMBOSS (Rice és mtsai., 2000) programcsomaghoz tartozó seqretsplit paranccsal kromoszómákra szedtem szét. A cél-szekvencia kromoszómák tördelésénél és forgatásánál az EMBOSS seqret programot futtattam. A Mbp hosszakat a bizonytalan kromoszómához tartozó szekvencia részeknél a markerpontok közti cM távolság arányoknak megfelelően adtam meg.

7.2.4 *RGSc és IRGSc illesztés*

Először megkerestem a gímszarvas genetikai térkép minden egyes kapcsoltsági csoportjához tartozó, markerpontok közti szegmensekkel átfedésbe hozható szarvasmarha genomi régiók génjeit. A génszekvenciák letöltése az online elérhető UCSC genomikai adatbázisból történt. A továbbiakban ezekkel a kereső génekkel, mint „csalikkal" kihalásztam azokat a „préda" scaffoldokat, amelyek tartalmazták a csali génnel ortológ gímszarvas szekvenciákat. Az ilyen scaffoldokat „referencia gén scaffoldoknak" (RGSc-knek) neveztem el. Annál bizonyosabb volt egy RGSc helyzete minél több, azonos sorrendben következő szarvasmarha gén ortológ szekvenciát foglalt magában. A folyamat eredményeként a markerpontok közötti részeket feltöltöttem RGSc-kkel. A felillesztéséhez MUMmer3.0, és BWA 0.7.10-r789 bioinformatikai szoftver csomagokat használtam. Az RGSc-kre kapott eredményektől függetlenül egy LASTZ_32 (Harris, 2007) illesztőprogram is lefutott a szarvasmarha referencia és az összes gímszarvas scaffold között. A LASTZ_32 kimeneti fájl nagyjából meghatározta, hogy melyik szarvasmarha kromoszómára kell tenni a scaffoldokat, azonban nem helyezte el az összes scaffoldot, pontatlanul adta meg a scaffoldok sorrendjét és irányultságát, emellett a bonyolult (vagyis valamilyen struktúrális átrendeződésen átesett) és az X Y kromoszómák feldolgozása elmaradt.

A nem referencia gének (USCS Refseq genes-en kívüli gének), rRNS, tRNS, miRNS géneket tartalmazó scaffoldok vagyis az úgynevezett inter reference genes scaffolds (IRGSc) felillesztéséhez a MUMmer3.0, és a BWA 0.7.10-r789 verzióját futtattam le. Ennél a munkafolyamatnál az RGSc-k és az IRGSc-k kromoszómális lokalizációját, sorrendjét és irányultságát határoztam meg a gímszarvas linkage group-oknak megfelelően széttördelt, és hard repeatmaszkolt szarvasmarha kromoszómákon. A BWA program futtatását megelőzte a referencia kromoszómák indexelése. A BWA -t a mem opcióval és a -t 6 -a -O 20,20 -L 100,100 paramétereket megadva alkalmaztam. A keletkezett sam kiterjesztésű fájlt SAMtools-szal sorba rendeztem (sort parancs) és indexeltem, valamint bedtools bamtobed paranccsal 12 oszlopos bed formátumú fájlokká alakítottam át (18. melléklet). A MUMmer-t a nucmer opcióval -mum -coords -c 100 kapcsolókkal futtattam le. A keletkezett coords fájlokból bed fájlokat generáltam (19. melléklet). A kétféle program eredményfájljait három saját készítésű Bash szkripttel dolgoztam fel, lényegében awk, sed, bedtools merge és sort parancsokkal. Az azonos nevű és irányultságú, 100000000 bp-nyi szekvencia részeket átfedő scaffold részleteket egybevontam és hat oszlopból álló bed fájlokká konvertáltam át. A szűrés során arra törekedtem, hogy egy scaffold csupán egyszer a legjobb, legnagyobb alignment részt leíró pozícióban fordulhasson elő, így jött létre a kromoszómánkénti scaffold vagyis a gerinc (backbone) (20/a, 20/b, 20/c. melléklet). A kimaradt, fel nem illeszkedő scaffoldokra lefuttattam egy megablast illesztést a következő kapcsolók megadásával:
-best_hit_score_edge 0.05 -best_hit_overhang 0.25 -evalue 1e-20 -perc_identity 50 -outfmt 6 -max_target_seqs 1 -num_threads 6 -max_hsps_per_subject 3.

7.2.5 *GFSc-k feltűzése*

A szarvasított (gímszarvas kapcsoltsági csoportoknak megfelelően széttördelt) szarvasmarha referencia genomot egyaránt kemény maszkoltam (hard masking) az alacsony komplexitású repetitív szekvenciákra (RepeatMasker) és a már elhelyezett backbone pozíciókra (bedtools maskfasta). Ezután a megmaradt réseket a már ismertett BWA és MUMmer szoftverek alkalmazásával feltöltöttem a kimaradt 1999 bp feletti scaffoldokkal (GFSc, gap filling scaffolds).

7.2.6 *Helytelenül illeszkedő scaffoldok*

A scaffoldok (99,6%, 23491/23593) nagy része egyértelműen igazodik a szarvasmarha genomjához, azonban 102 scaffold lokalizációja továbbra is kétértelműnek bizonyult. Ezekből 70-et páronként 35 új scaffolddá fésültem össze SeqMan (Swindell és mtsai., 1997) program segítségével. A maradék scaffoldokat contigokra bontottam, így 2582 egyedi contigot kaptam. Az újonnan létrejött szekvencia elemeket újra felillesztettem, kézzel leellenőriztem és ez alapján a megfelelő helyekre szúrtam be a gímszarvas pszeudokromoszómán.

7.2.7 *Genomi részek beforgatása*

A szarvasmarha referencia kromoszómáin megtalált marker szekvenciák sorrendje az esetek többségében követi az ortológ kapcsoltsági csoportok térképpontjainak sorrendjét, de néhányszor két szomszédos marker megcserélődését figyeltem meg. Az 1 cM-nál és az annál nagyobb távolság esetén elképzelhetőnek tartottam a valós inverziót, azonban 1 cM alatt megmaradt a szarvasmarhának megfelelő rangsor. Forgatáskor a markereken túli scaffoldok megtartják az eredeti lokalizációjukat. Ha azonos scaffoldon

található az inverz marker pár meg kellett vizsgálni a scaffold génsorrendjét szarvasra és szarvasmarhára vonatkoztatva. Az egyetlen szomszéddal rendelkező markerek beillesztésénél figyelembe vettem a távolságot, amikor inverziót észleltem kihagytam a markert. A bizonytalan szekvenciájú, általában like végű nem illeszkedő markerekkel szintén nem foglalkoztam, kivéve DIA1like (DIA1-szerű) fehérjét. A lehető legkevesebb inverziót feltételeztem.

7.2.8 *Scaffoldok összefűzése*

A scaffoldokat és contigokat megfelelő sorrendben és orientációban helyeztem el egy táblázatban. Majd egy saját Bash szkripttel fűztem össze, olyan módon, hogy az NCBI követelményeinek megfelelően 100 bp N karakter ékelődjön közéjük (21/a., 21/b. melléklet). Habár az én Bash szkriptem is megoldotta ezt a feladatot, ennek ellenére a munkacsoport vezetője a kollégám Nagy Tibor scaffold összefűző Python szkriptje mellett döntött a Python nyelv sajátságaiból eredő átláthatóság és alkalmazhatóság miatt.

7.3 GÍMSZARVAS GENOM ANNOTÁCIÓJA

7.3.1 *Repetitív szekvenciák, rRNS, miRNS, tRNS azonosítása*

A repetitív szekvenciákat a RepeatMasker Open-4.0. (Smit és mtsai., 2013-2015) nevű szoftverrel azonosítottam a következő beállítások megadásával: -a -inv -gff -X -poly -cut -s -gccalc -par 10 -species artiodactyl. Az emlős riboszomális RNS kis és nagy alegység (SSU, LSU) szekvenciáit a SILVA123 referencia adatbázisból szedtem ki.

A prekurzor formátumú mikroRNS-ek a miRBase adatbázisból (Kozomara és mtsai., 2014) származtak, amelyekből 21 emlős miRNS szekvenciát

használtam fel a későbbiekben. Az rRNS-ekből és külön a miRNS-ekből kereső-szekvencia fájlokat hoztam létre, amelyeket blastn paranccsal illesztettem a gímszarvas genomunkhoz.
Az LSU rRNS elhelyzés beállításai: -evalue 1e-5 -perc_identity 85 -word_size 50 -outfmt 7, az SSU rRNS-nél a word_size-ot 40-re csökkentettem le.
A miRNS illesztés paraméterei: -evalue 1e-3 -perc_identity 90 -word_size 70 -outfmt 7. A riboszómális RNS 5S egységét Barrnap 0.6 (https://github.com/tseemann/barrnap, v06) program segítségével kerestem meg.
A transzfer RNS-eket a tRNAscan-SE-1.3.1 szoftverrel találtam meg. A SINE maszkolt tRNS-t és pszeudogéneket nem vettem figyelembe. A parancsban a -H -f -m -o opciókat adtam meg (22. melléklet).

7.3.2 *Protein-kódoló gének keresése*

A gímszarvas protein-kódoló gének és az SNV-k annotációját Nyiri Anna kolléganőmmel együtt végeztük el. A fehérje kódoló gének annotációjához a MAKER 2.31.8 gén annotációs programot használtuk (Cantarel és mtsai., 2008, Campbell és mtsai., 2014). A MAKER program gén azonosítása azért is fontos, mert rámutat a rokon fajok közötti ortológ gének ko-lineáris elhelyezkedésére és a genom összerakás minőségére. Továbbá az új, annotált genom génjei alapvetők a további genetikai, bioinformatikai kutatásokban. A MAKER tulajdonképpen egy olyan parancssorozat, amely folyamatosan hív be és alkalmaz különböző bioinformatikai programokat.
Első lépésként identifikálja és maszkolja az ismétlődő elemeket a genomban a RepeatMasker-open-4.0.5 és a RepeatRunner programok segítségével. Azonfelül egy repetitív szekvenciák maszkolására alkalmas, saját fajspecifikus könyvtárat hoz létre a RepeatModeler 1.0.4 a RECON 1.08, és a RepeatScout

1.0.5 szoftverek alkalmazásával. Amíg a RepeatMasker azonosítja az ismert repetitív szekvenciákat, addig a RepeatModeler (Smit és mtsai., 2013-2015) újakat jelez előre.

A következő lépésben a MAKER létező, más fajokból származó EST, mRNS és fehérje szekvenciák felhasználásával úgynevezett evidence-based vagyis bizonyítékokon alapuló, kezdeti génmodelleket alkot. Az EST modell konstrukciójához szarvasmarhából származó cDNS-t használtunk (ftp://ftp.ensembl.org/pub/release-88/fasta/bos_taurus/cdna/). Letöltöttük a szikaszarvas (*Cervus nippon*) nyers RNS readjeit (Yao és mtsai., 2012), majd ezekből TrinityRNAseq (https://github.com/trinityrnaseq/trinityrnaseq) (Grabherr és mtsai., 2011) parancssorozattal szikaszarvas transzkriptomot generáltunk, amely ilyen formában alkalmassá vált az RNS modell készítésére (Jia és mtsai., 2016). A protein modellek építéséhez több faj teljes fehérje készletét vettük igénybe. Az ember (ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/pep/), a szarvasmarha (ftp://ftp.ensembl.org/pub/release-88/fasta/bos_taurus/pep/) és a juh (ftp://ftp.ensembl.org/pub/release-88/fasta/ovis_aries/pep/) protein szekvenciák az ENSEMBL adatbázisából származtak (International Human Genome Sequencing Consortium, 2004, Merchant és mtsai., 2014, Bovine Genome Sequencing and Analysis Consortium, 2009, International Sheep Genomics Consortium, 2010).

A RepeatMasker a következő repetitív szekvencia maszkolásnál több könyvtárat is kezelt. A RepBase adatbázis artiodactyl modellszervezetre jellemző ismétlődő szekvenciákat és a korábban létrehozott saját fajspecifikus könyvtárat együttesen alkalmazta a MAKER belső adatbázisával, amely transzponálható elemeket is tartalmazott. Ezt követően BLAST algoritmus illesztette fel az EST, az mRNS és a fehérje szekvenciákat a könnyű maszkolt (soft-masking) gímszarvas genomra. A könnyű maszkolás során az alacsony

komplexitású és repetitív genomi régiók nukleotidjait nagybetűről kicsire írja át a program.

Ezután az Exonerate szoftver (Slater és Birney, 2005., https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) finomította a BLAST illesztések pontosságát.

A nem bizonyítékokon alapuló, vagyis az *ab-inito* gén predikciót a SNAP (Korf, 2004), (Campbell és mtsai., 2014) és az AUGUSTUS (Stanke és mtsai., 2006) programok végezték el. A SNAP által előrejelzett gének készletén a MAKER újra lefutatta a SNAP-et.

A végleges gén annotáció elkészítéséhez a bizonyítékokon alapuló és *ab-initio* gén jóslatokat a MAKER összevetette egymással, és az egymásnak jól megfeleltethetőket kiszűrte, majd a végeredményt optimalizálta. A MAKER által létrehozott génmodell gff3 formátumú, ennélfogva könnyen feltölthető különböző genomi böngészőkbe.

A fehérje funkciók és a protein-kódoló gének azonosítása az InterProScan (Jones és mtsai., 2014) szoftverrel történt. A MAKER különféle RNS-kódoló géneket is észlelt, amiket aztán összevetettünk az előző fejezetekben leírt programokkal kapott RNS-ekel (23. melléklet).

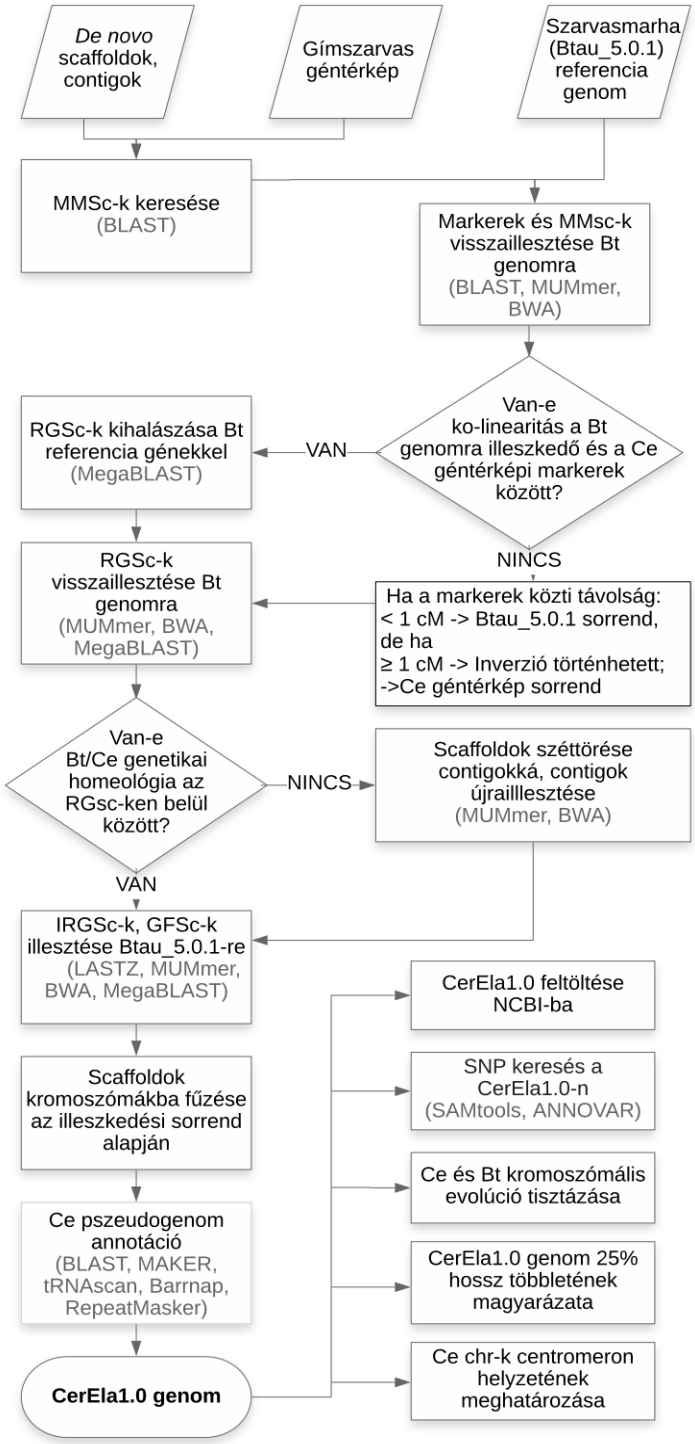
7.3.3 *Genetikai variánsok kimutatása*

A paired-end és a mate pair könyvtárak readjeit bwa-mem (verzió: 0.7.10-r789) program segítségével illesztettük fel a referencia genomra. Az úgynevezett „unmapped”, vagyis nem térképezhető readeket nem használtuk fel a további vizsgálatokra. Ebben a munkafázisban a SAMtools és a BCFtools SNV kereső programok paramétereit állítottam be.

A genetikai variánsok azonosítása SAMtools szoftverrel (Li és mtsai., 2009, Molnár és mtsai., 2014) (Verzió: 0.1.19-44428cd) a "mpileup -D -S -E -uf" parancssori opciók megadásával történt meg.

A vcf fájlokat a BCFtools program a "view -bvcg" paraméterek alkalmazásával hozta létre. A vcfutils.pl (https://github.com/lh3/samtools/blob/master/bcftools/vcfutils.pl) szkript "varFilter -D 188" opcióval azokat a változatokat szűrte ki, ahol a lefedettség legfeljebb 188 volt (legalább háromszoros átlagos genomi lefedettség esetén). A továbbiakban Nyiri Anna kolléganőm házilag készített Perl szkriptjével szűrte le azokat a változatokat, amelyek Phred minőségi pontszáma meghaladta a 30-at. A Phred minőségi pontszám az automatizált DNS-szekvenálás során leolvasott nukleotidbázisok azonosításának minőségét adja meg. A logaritmikusan kapcsolódik a bázis-hívó hiba valószínűségekhez például, ha a Phred 30-as minőségi pontszámot rendel a bázishoz, az esélye annak, hogy ezt a bázist helytelenül határoztuk meg az 1 az 1000-hez.
Az aminosav funkcionális változásokat okozó variánsok annotációját ANNOVAR szoftverrel (Wang és mtsai., 2010) végeztük el olyan módon, hogy létrehoztunk egy saját gén-definíciós adatbázist a MAKER-eredetű gff3 fájl segítségével. A genetikai variánsok annotációjához a table_annovar.pl (https://annovar.openbioinformatics.org/en/latest/user-guide/startup/).
szkriptet futtattuk le (24. melléklet).

Az anyag és módszer fejezetben kifejtett teljes munkafolyamatot a 6. ábrán foglalom össze.



6. ábra. A munkafolyamat bemutatása. A rövidítések a „Rövidítések és szakkifejezések jegyzékében” megtalálhatók, a munkafázisokhoz használt programok nevei szürke színűek.

8 EREDMÉNYEK ÉS MEGVITATÁSUK

8.1 GÍMSZARVAS REFENCIA GENOM SZEKVENCIA (CERELA1.0) ÖSSZEÁLLÍTÁSA

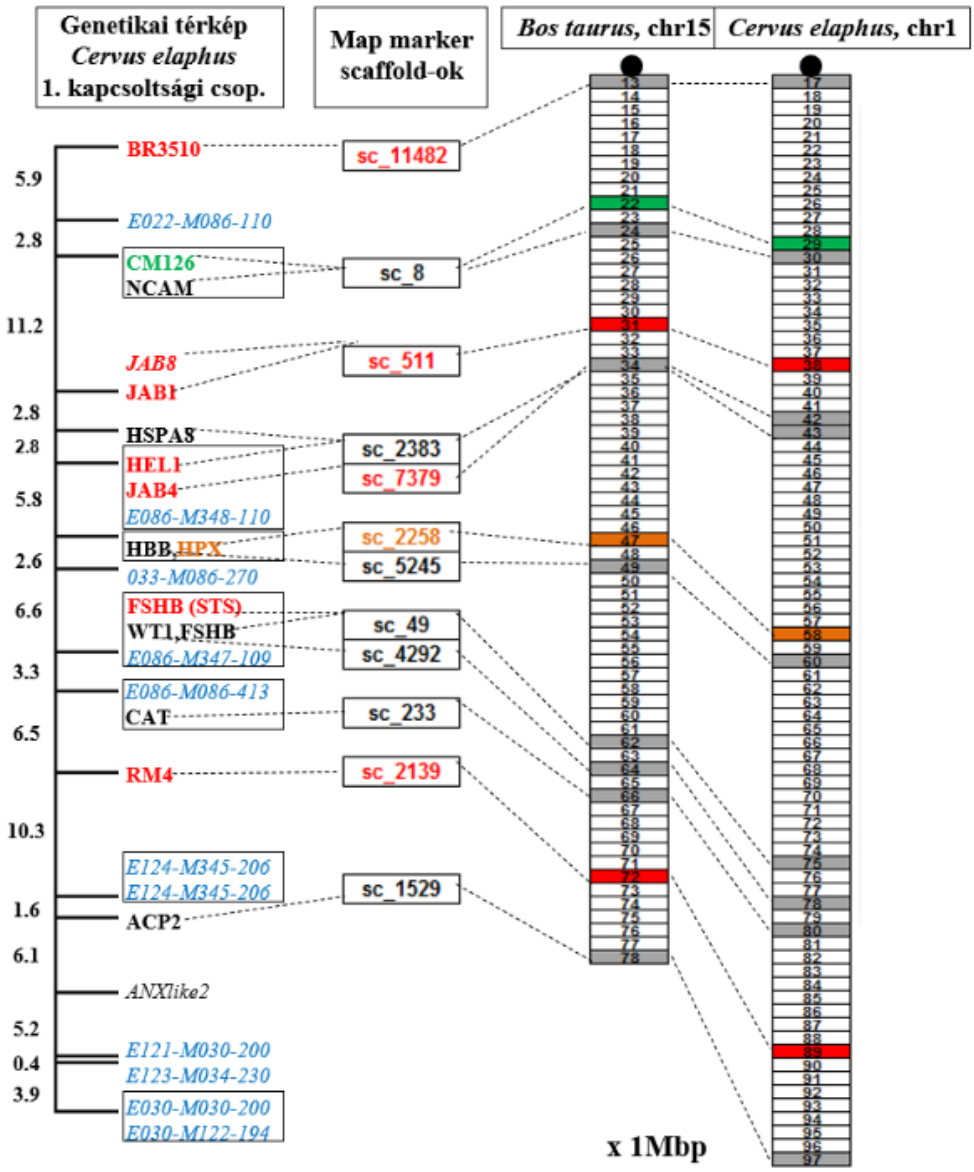
8.1.1 *Readek, contigok, scaffoldok*

A gímszarvas teljes genom szekvenálása során úgynevezett paired-end read, és mate pair read szekvenciák jöttek létre. Együttvéve 4 paired-end és 2 mate pair szekvencia könyvtár készült el, amelyekben összesen körülbelül 2,2 x 10^9 read található, amely 222,7 Gbp DNS-nek felel meg. Az elkészült *C. elaphus* haploid genomra (33 autoszóma és 1 nemi kromoszóma) nézve a readek átlagos lefedettsége 62-szeres. A *de novo* assembly készítést az ALLPATHS-LG program hajtotta végre, melynek eredményeképpen 437412, összesen 1,95 Gbp hosszú contig jött létre. A contigokat a program nagyobb szekvencia egységekké, scaffoldokká fűzte össze. A scaffoldok száma 34724, összhosszúságuk pedig a gapeket is figyelembe véve 3,4 Gbp. Ez a genom méret nagyon hasonló más emlős fajok például két filogenetikailag közelálló faj a szarvasmarha (*Bos taurus*) vagy a juh (*Ovis aries*), 3 Gbp körüli értékéhez.

8.1.2 *Térképpont vagyis "mapmarker" scaffoldok (MMSc-k)*

A gímszarvas genetikai térképe a *C. elaphus x Elaphurus davidianus* közötti interspecifikus keresztezés és back-cross F2 populáció adatai alapján készült el (Slate és mtsai., 2002). A géntérkép az összes kapcsoltsági csoportra nézve 2532 cM hosszúságú és 5,7 cM-os átlagsűrűséggel rendelkezik. Slate és munkatársai a haploid kromoszómaszámmal összhangban a genetikai térképet 34 kapcsoltsági csoportra osztották fel úgy, hogy az X és az Y kromoszómákat összevonták. Az általuk meghatározott 621 genetikai marker, majdnem 90%-

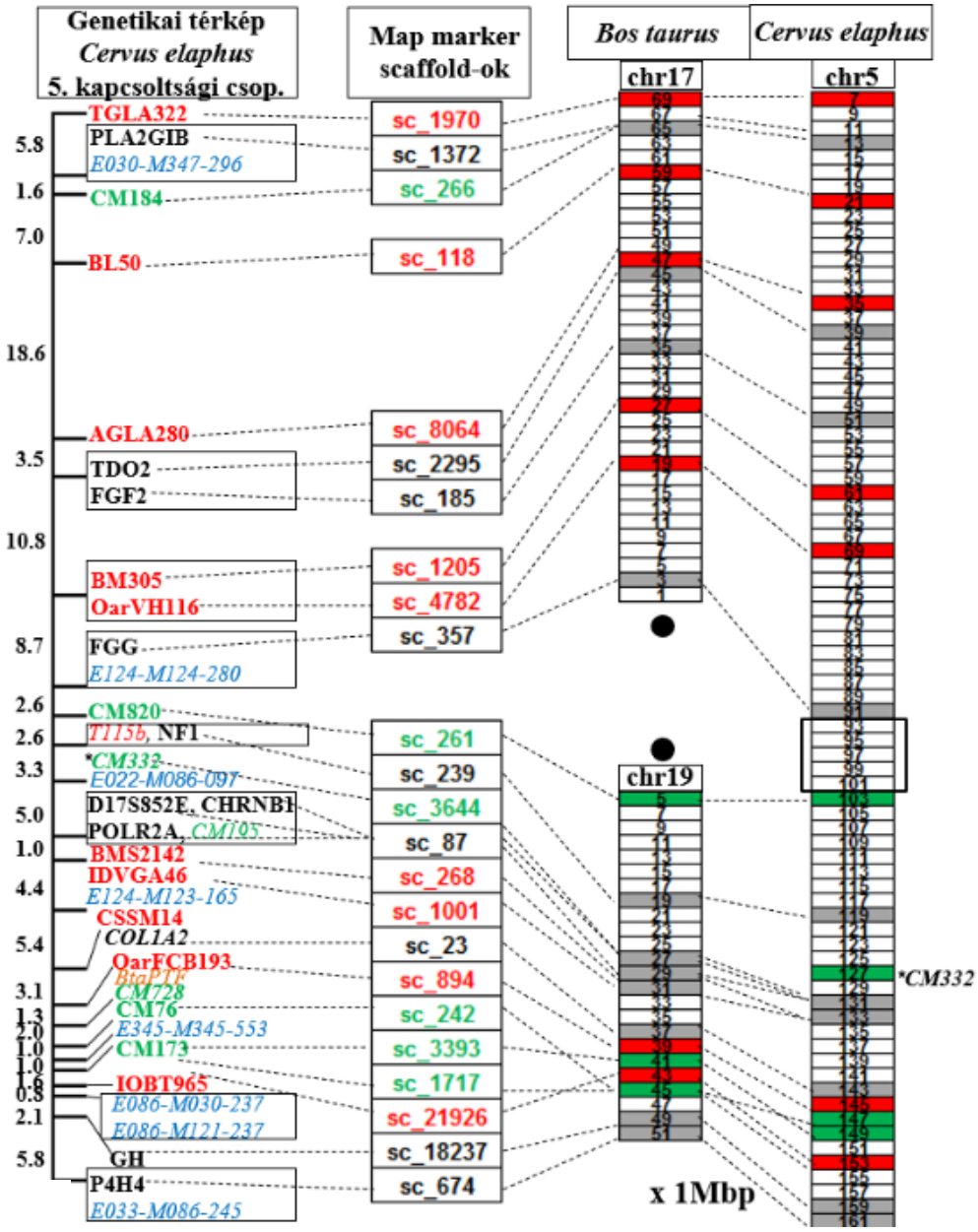
ban a párosujjú patásokból származtatható és viszonylag kevés a főemlős és a rágcsáló marker köztük. A 621 genetikai markert online genomikai adatbázisokból kerestem ki. Az AFLP-ék kivételével az összes típushoz (EST, RFLV, STS, protein) találtam szekvenciákat. Ekképpen 365 marker szekvenciáját sikeresen azonosítottam és helyeztem el a genetikai térképen. A 365 markert hordozó scaffoldokat "mapmarker" (MMSc) scaffoldoknak neveztem el. Az MMSc-k így meghatározott pozícióba kerültek a kapcsoltsági géntérképen (7. ábra).



7. ábra. Gímszarvas 1. kromoszóma. Az első oszlopban a gímszarvas géntérkép 1. kapcsoltsági csoportja látható, (marker színkód: piros-STR, fekete-RFLV/gén, zöld-EST, barna-protein, kék-AFLP, a szaggatott vonalak az illeszkedő/ortológ szekvenciákhoz/régiókhoz mutatnak) (Slate és mtsai., 2002), a 2. oszlopban a gímszarvas mapmarker scaffoldok (MMSc) találhatók, a 3. oszlopban a szarvasmarha ortológ pszeudokromoszóma (chr15) Mbp-ra osztott számegyenesét tüntettem fel, a 4. oszlopban 1. gímszarvas pszeudokromoszóma Mbp-onként felosztott számegyenese helyezkedik el. A centromeronokat fekete telt körrel ábrázoltam.

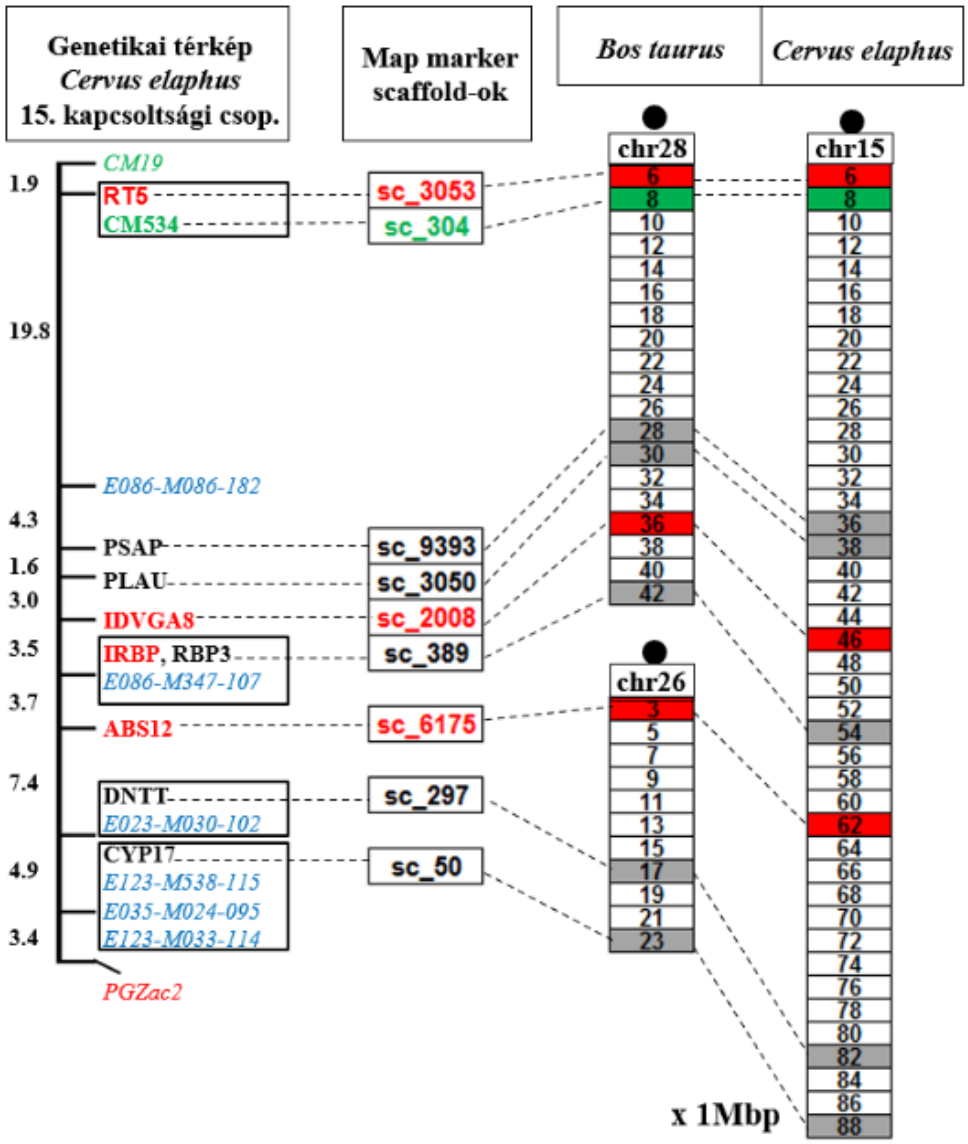
8.1.3 *A Cervus elaphus genetikai térkép és a Bos taurus genom közötti kolinearitás*

A *B. taurus* referencia genom két változatával (NCBI Btau_4.6.1, Btau_5.0.1) szemben illesztettem a 365 *C. elaphus* marker szekvenciáit illetve térképpont scaffoldjait (MMSc). Ennek megfelelően sikeresen azonosítottam a *B. taurus* genomban a 365 gímszarvas MMSc-vel hasonlóságot mutató, ortológ szekvenciákat. Ezek pozíciói a kromoszómákon végig ko-lineárisan helyezkedtek el mindkét fajban. Ezenfelül az MMSc-ben található gének sorrendje is azonos volt a „bovin" génekével, vagyis "intra-scaffold és intra-contig" szinten is érvényesült a szinténia. A gímszarvas kromoszómák többsége (19 autoszóma és X, Y) közvetlenül párba állítható egy homeológ szarvasmarha kromoszómával (7. ábra). A szarvasmarhában evolúció során megörződött a 6 akrocentrikus ősi kromoszóma, és ezek hasadásával a szarvasfélék evolúciója alatt keletkezett 12 akrocentrikus (szarvas) kromoszóma. Ez azt jelenti, hogy a szarvas evolúciós vonalon új centromeronok jöttek létre. Ezzel ellentétben a szarvas evolúció során, valamikor a *Cervinae* ág leválásánál két ősi *Pecora* kromoszóma centrálisan fúzionált egymással (Robertsoniális transzlokáció), ami a szarvasformákra, így a gímszarvasra is jellemző metacentrikus kromoszómát alakította ki. Ezáltal a Ce5 két karja ekvivalens két szarvasmarha akrocentrikus kromoszómával (Bt19 és Bt17) (8. ábra).



8. ábra. Gímszarvas 5. kromoszóma. Az oszlopok megfelelnek a 7. ábrán leírtakkal. Ez esetben a szarvasmarha 17. és 19. kromoszómái a gímszarvas 5. kromoszómájával ortológok (Robertsoniális transzlokáció). A centromeronokat fekete telt körrel jelöltem a szarvasmarha kromoszómákon, és üres négyzettel a gímszarvas chr5-ön. A CM332 markert csak a gímszarvas chr5-ön találtam meg a szarvasmarha chr17-en és chr19-en nem.

Az akrocentrikus Ce15 centromeronja azonos a Bt28-éval. A kromoszómális eltérés a Bt28 és Bt26 Robertsoniális fúziójával (tandem fúzió) írható le, ahol a Ce15 centromeron közeli rész a Bt28-al, a távoli rész pedig a Bt26-al ekvivalens (9. ábra) (Bonnet és mtsai., 2001 Frohlich és mtsai., 2017).

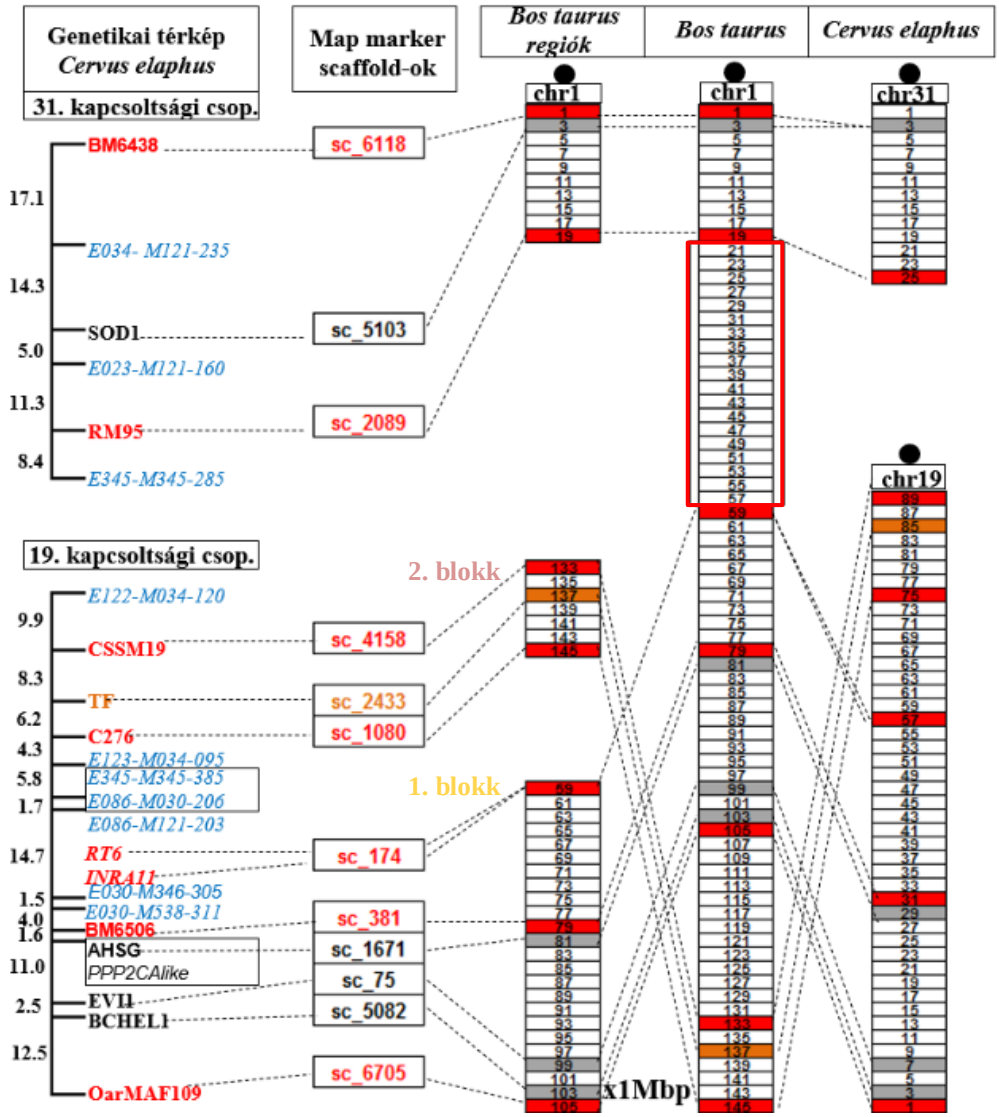


9. ábra. Gímszarvas 15. kromoszóma. Az oszlopok megfelelnek a 7. ábra leírásaival. A Robertsoniális fúzióval egyesült a Bt28 és a Bt26 a Ce15 két karjával egyezik meg.

Megfigyelhető két összetettebb evolúciós átrendeződés is. Az egyik komplex esetben az akrocentrikus Bt1 két akrocentrikus, ősi *Pecora* kromoszóma hasadásával létrejött gímszarvas kromoszómával egyenlő (Ce19, Ce31), amelyek közül a Ce19-ben lejátszódott egy hasadás és egy transzlokáció. A 31. gímszarvas kapcsoltsági csoport géntérképi pontjainak megfelelői jól azonosíthatók a szarvasmarha referencia genom (Btau_5.0.1) chr1 kromoszóma 1 bp-tól 19 Mbp-ig terjedő szakaszán. Azonban a 19. kapcsoltsági csoport markerei a Btau_5.0.1-re való illeszkedésük alapján két nagyobb blokkba rendeződnek. A géntérkép szerinti első blokk markereinek ortológ szekvenciái (OarMAF109, BCHEL1, EVI1, AHSG, BM6506, INRA11, RT6) fordított sorrendben a szarvasmarha genom 59 Mbp-tól 105 Mbp-ig terjedő szakaszán azonosíthatók. Míg a második blokk markerei (CSSM19, TF, C276) szintén fordított sorrendben 133 Mbp-tól 146 Mbp-ig tartó részen helyezkednek el. A két blokk belső elrendezése és egymáshoz való viszonya, egyedül transzlokációval magyarázható (4. táblázat, 10. ábra). A kérdéses hovatartozású szarvasmarha szekvencia részeket a géntérképi cM távolságok és az 1 cM~1Mbp arányt figyelembe véve osztottam el a kromoszómák között és azokon belül, azonban ez egy durva becslést jelent, így a térképpontokon kívüli részek helyzete bizonytalan. Minden gímszarvas kromoszómánál a Slate-féle térkép határozta meg a scaffoldok sorrendjét, ezért néhány esetben, mint például a Ce19-nél a centromerontól ellentétes irányban kezdődik a szekvenciák számozása.

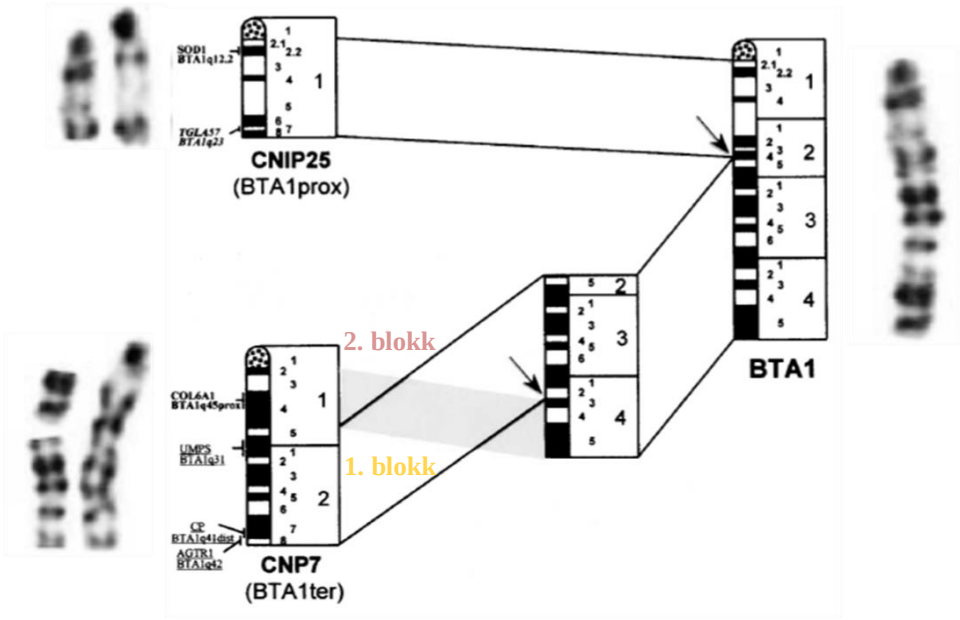
4. táblázat. Gímszarvas 19. kapcsoltsági csoport markerei. Egy-egy marker illeszkedésénél és elhelyezésénél a kezdeti és a végpont közötti szekvencia középértékét adtam meg. A táblázat blokkjait a 10. és a 11. ábrán is jelöltem.

Blokk	C.e. LG19 markerek		Btau_5.0.1; Bt1	CerEla1.0; Ce19
	Név	Távolság (cM)	Illeszkedés (bp)	Elhelyezés (bp)
1.	OarMAF109	0,0	105101342	157114
	BCHEL1	12,5	102720911	2984723
	EVI1	15,0	99447407	7087139
	AHSG	26,0	81497935	29517318
	BM6506	27,6	79961061	31418975
	INRA11	-	59698369	56696386
	RT6	-	59486881	57014285
2.	C276	59,6	145457584	75619922
	TF	65,8	137203056	85634407
	CSSM19	74,1	133510317	90353433



10. ábra. Gímszarvas 31. és 19. kromoszóma, hasadás, transzlokáció. A Bt1-hez képest az ortológ ősi szarvas kromoszómában történt egy hasadás és egy transzlokáció, így alakult ki a Ce31 és Ce19. Az oszlopok megfelelnek a 7. ábrán leírtakkal. A centromeronokat fekete telt körrel ábrázoltam, az üres piros keretes négyzet a kérdéses *B. taurus* régiót jelöli, nem lehet biztosan tudni, hogy melyik gímszarvas kromoszómához tartozik. A 31. kapcsoltsági térkép ezen az ábrán az átláthatóság kedvéért fordított helyzetű az eredeti Slate géntérképhez képest (Slate és mtsai., 2002).

A Ce19-nél tapasztalható bioinformatikailag bizonyított megállapításokat tökéletesen visszaigazolják a szakirodalomban felelhető citológiai összehasonlító vizsgálatok eredményei (Frohlich és mtsai., 2017). A szikaszarvas a gímszarvas legközelebbi rokona, mindkét irányban kereszteződnek és az utódok mindkét neme termékeny (McDevitt és mtsai., 2009). Mivel a gímszarvas és a szikaszarvas közeli rokonok, ezért a szikaszarvas és szarvasmarha kromoszómális összehasonlító elemzések megbízható támpontot adnak a gímszarvas esetén is. E tanulmányok alapján a *C. nippon* és a *B. taurus* kromoszómái a sávtérképek szerint is megfeleltethetők egymásnak. A Ce19 ekvivalens CNP7 (szikaszarvas 7. kromoszóma) és a Bt1 (az eredeti ábrán BTA1) kromoszómák összevetése a 11. ábrán látható. A CNP7 centromeronhoz legközelebbi marker génje a COL6A1 a Ce19-en 71 Mbp-nál, a Bt1-en 147 Mbp-nál lokalizálódik, ami a Slate-féle géntérkép 2. blokkjának markereihez közeli pozíciót jelent. Ettől a pozíciótól disztálisan helyezkedik el a következő marker az UMPS, ami 43 Mbp-nál található a gímszarvas 19. kromoszóman és 69 Mbp-nál a Bt1-en, ezáltal az 1. blokk markereivel hozható átfedésbe. A Ce19 centromeron a 71 Mbp-tól nagyobb számú szekvencia részeknél található meg.



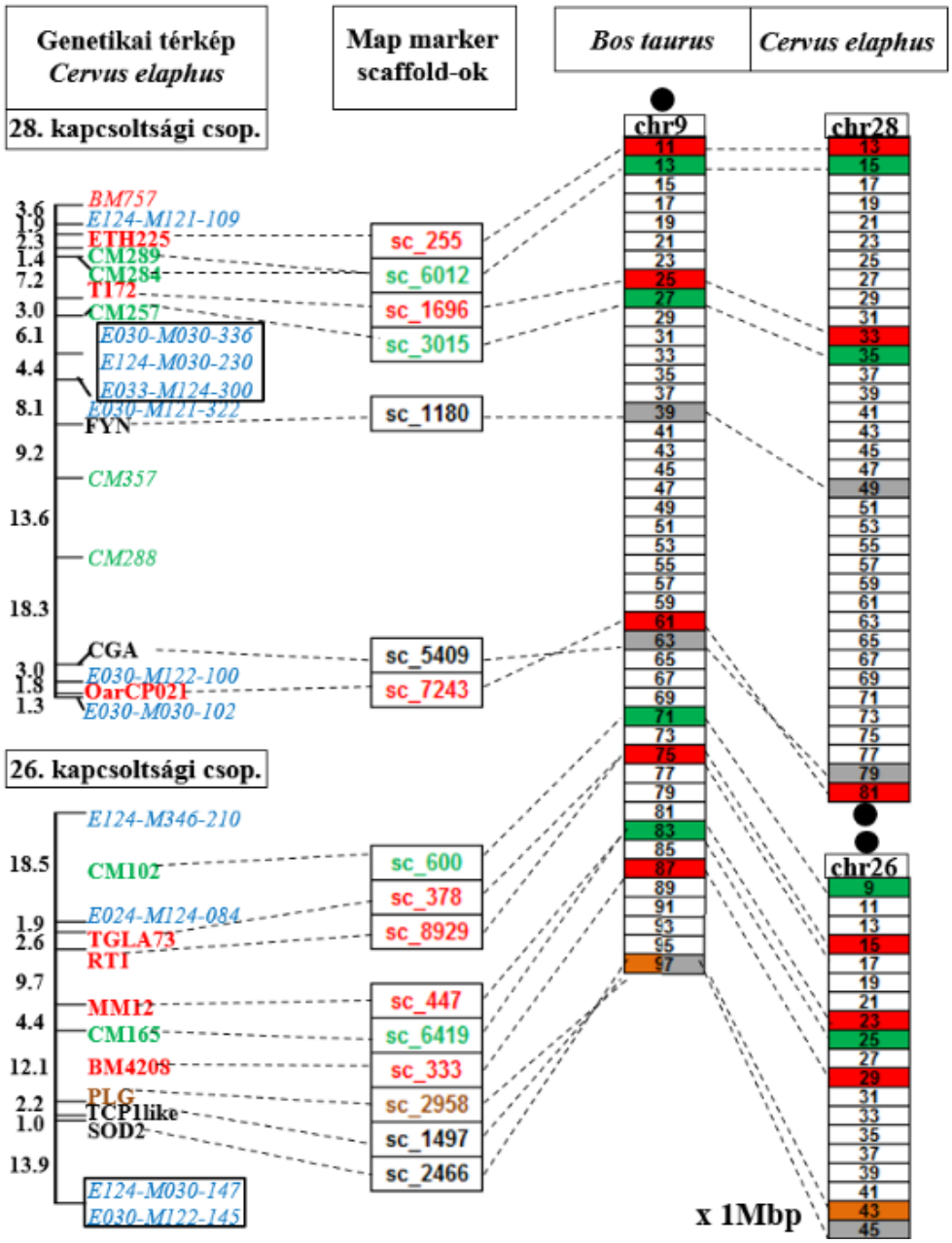
11. ábra. Komparatív citogenetikai analízis a BTA1 és CNP25-CNP7 között (Bonnet és mtsai., 2001). CNP/CNP: *Cervus nippon*, szikaszarvas, BTA: *Bos taurus*, szarvasmarha, CNIP25: szikaszarvas 25-ös kromoszómája, amely megfelel a gímszarvas 31. kromoszómájának és a szarvasmarha 1. kromoszóma (BTA1) centromeron proximális felének. CNP7: A szikaszarvas 7. kromoszómája, a gímszarvas 19. kromoszómájával egyezik meg és a kromoszómán belüli transzlokációs átrendeződés után megfeleltethető a szarvasmarha 1. kromoszóma (BTA1) centromeron disztális felével. Megjegyzendő, hogy ezen az ábrán megtartottam a cikk (Bonnet és mtsai., 2001) ábra eredeti elnevezéseit.

A másik komplex esetben a Bt9-el megfeleltethető két akrocentrikus Ce26 és Ce28 szarvas kromoszómák közül az egyikben, a Ce28-ban paracentrikus inverziót feltételeztem (12. ábra) egy korábbi tanulmány alapján, amely a szarvasmarha és a szikaszarvas közötti komparatív citogenetikai analízist írja le (Bonnet és mtsai., 2001). A bizonyítást nehézkessé teszi, hogy az inverzió töréspontjai az általam figyelembe vett Slate-féle géntérkép markerein kívül esnek a kromoszóma mindkét végén és a markerek sorrendje megegyezik a két fajban (5. táblázat). Emiatt a következő megállapításokat vettem alapul: Olyan akrocentrikus kromoszómákról van szó, ahol: (i) az összehasonlító

kromoszóma citológia mutatja az inverziót (Bonnet és mtsai., 2001), (ii) két gímszarvas akrocentrikus kromoszóma (Ce28, Ce26) keletkezett Bt9 akrocentrikus kromoszóma hasadásából úgy, hogy a Ce28 a Bt9 proximális darabjával feleltethető meg, azaz a Ce28 és a Bt9 centromeronok szinténikusak. A hasadást követően történhetett a Ce28-ban egy nagy méretű inverzió, ami az akrocentrikus kromoszómából kiindulva csak paracentrikus lehetett. Ilyenkor előfordulhat, hogy a letört centromeron és telomér a helyén marad és a köztük lévő rész megfordulva visszaforr, vagy éppen fordítva a törött cetromeron és a telomer vándorol el a közbenső rész két ellentétes végére, ahol aztán visszakapcsolódnak.

5. táblázat. LG28 géntérképi pontok illeszkedése. Az illeszkedési és elhelyezési szekvenciák középértékét adtam meg.

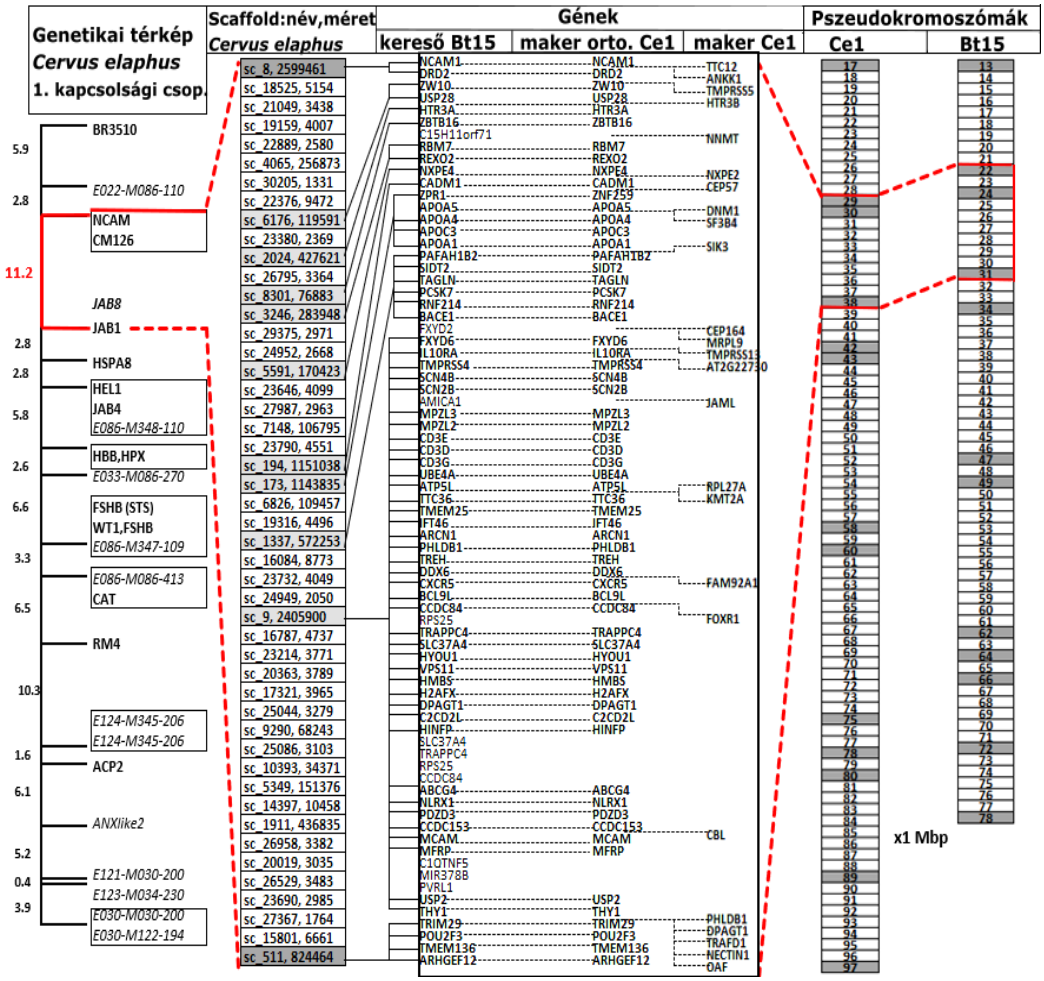
C.e. LG28 markerek		**Btau_5.0.1; Bt9**	**CerEla1.0; Ce28**
Név	**Távolság (cM)**	**Illeszkedés (bp)**	**Elhelyezés (bp)**
ETH225	0	10908161	13060454
CM289	2,3	13280621	16121761
CM284	3,7	13280203	16122207
T172	10,9	26115967	32871123
CM257	13,9	27795743	35082592
FYN	32,5	39293242	50068167
CGA	73,6	63979272	78512490
OarCP021	78,4	61067665	81944836



12. ábra. Gímszarvas 28. és 26. kromoszóma, hasadás, inverzió. A gímszarvas 28. kromoszómája ekvivalens a szarvasmarha 9. kromoszóma centromeron proximális részével. Ebben a részben egy hatalmas paracentrikus inverzió is történt. A gímszarvas 26. kromoszómája megfeleltethető a szarvasmarha 9. kromoszóma centromeron disztális felével. Az oszlopok azonosak a 7. ábrán leírtakkal. A centromeronokat fekete telt kör mutatja.

8.1.4 *C. elaphus genetikai térkép feltöltése, B. taurus referencia gén vezérelt scaffold keresés (RGSc-s)*

A gének szekvenciája nagyon hasonló, a térképpont markerek a két fajban a kromoszómákon kolinerárisan helyezkednek el és az MMSc-ken található gének is kiterjedt szinténiákat mutatnak a két rokon faj között, ezért kézenfekvővé vált, hogy a szarvasmarha genomot használjam cél-szekvenciaként a további gímszarvas scaffoldok kromoszómális sorba rendezéséhez. Lényegében a gímszarvas genetikai térkép szegmensekre igazított bovin génsorrendjét követve összesen 6013 scaffoldot lokalizáltam. A Gímszarvas és szarvasmarha gének közötti azonosságot a 13. ábra mutatja.



13. ábra. Gímszarvas és szarvasmarha gének közötti kolinearitás. 1. oszlop géntérképi markerek, 2. oszlop a pirossal kijelölt géntérkép szegmenshez (NCAM,CM16-JAB1 markerek határolják) illeszkedő scaffoldok azonosító száma és mérete vesszővel elválasztva, szürkével jelöltek a referencia géneket hordozók (RGSc) nincsenek beszínezve a referencia géneket nem hordozók (IRGSc, GFSc) , 3. oszlop erre a régióra illeszkedő kereső szarvasmarha gének, 4. oszlop a maker által annotált ortológ gímszarvas gének, 5. a maker által annotált új gímszarvas gének, 6. oszlop az 1 gímszarvas kromoszóma Mbp-ban megadott számegyenese (sötétszürkével jelöltek a markereket hordozó részek, pirossal a vizsgát régió), 7. oszlop az ortológ szarvasmarha 15 kromoszóma.

8.1.5 *Az RGSc-k közötti rések feltöltése (IRGSc)*

A következő munkafázisban az MMSc-RGSc és az RGSc-RGSc közötti réseket töltöttem fel az eddig el nem helyezett scaffoldokkal. Ezek feltehetően nem tartalmazták az UCSC által megadott Refseq protein-kódoló géneket, de nagy valószínűséggel rendelkeztek másmilyen fehérje kódoló, például nem Refseq, rRNS, tRNS, és miRNS génekkel. Ilyenkor megfordult a csali/préda viszony. Itt már a gímszarvas scaffoldok váltak olyan keresőszekvenciákká, amelyeket a szarvasmarha célgenomra illesztettem rá. A kereső, vagyis a csali scaffoldokat," inter referencia gén scaffold"-oknak (IRGSc) neveztem el. Ellenőrzésképpen megvizsgáltam a kapott génsorrendeket a két fajban. A tapasztalt szinténiák meggyőzőnek bizonyultak (13. ábra).

Az MMSc-ket, RGSc-ket és IRGSc-ket együttvéve 13748 scaffoldot tudtam elhelyezni, azonban 15205 2 Kbp-nál hosszabb scaffold lokalizációja továbbra is kérdéses maradt. A probléma megoldásaként kimaszkoltam a már meglévő 13748 *Bos taurus* genomi helyet. Az ilyen módon „maszkolt" referenciára illesztettem fel a megmaradt 2 Kbp-nál nagyobb szekvenciákat. Ennélfogva még 9845 új, úgynevezett réskitöltő, vagyis „gap filling" scaffoldhoz (GFSc) találtam szarvasmarha kromoszóma részeket. A munkafolyamat során sikerült azonosítani összesen 23593 scaffold szarvasmarha régióját. A szekvenciák nagyrészénél, pontosabban 99,6%-ánál egyértelmű, kölcsönös megfeleltetést tapasztaltam a két faj között, de 102 esetében elmondható volt, hogy akár több, egymástól elkülönülő kromoszómális szakasszal mutattak ortológiát vagy ezzel ellenkezőleg, azonos pozíciókat fedtek le. Megoldandó a problémát önálló contigokra bontottam a scaffoldokat vagy összefésültem őket egy azonos scaffoldban. A keletkezett 2582 egyedi contigot és az új 35 scaffoldot ebben a formában illesztettem újra a referencia genomhoz azért, hogy feloldjam e néhány scaffold esetében megfigyelt, belső ellentmondást és

ezáltal a legpontosabb gímszarvas pszeudokromoszómális helyet találjam meg a számukra.

A pszeudokromoszóma készítés végeredményeként létrehoztam a haploid gímszarvas (*C. elaphus hippelaphus*) referencia genomot, amely CerEla1.0 nevet kapta. A CerEla1.0 genom 23491 plusz 35 scaffoldból (MMSc-k, RGSc-k, IRGSc-k, GFSc-k plusz a 35 új) és 2582 önálló contigból tevődik össze, azaz összegezve 26108 szekvencia elemet tartalmaz, ami 3,4 Gbp genom hosszúságot ad ki.

Az „unplaced" kategóriába került 11444 scaffold, mivel nem tudtam az új gímszarvas referencia kromoszómákon elhelyezni őket. Az unplaced szekvenciák összesen 52989442 bp-t tesznek ki, ami a teljes genom 1,6 %-a.

8.1.6 *Scaffoldokon és contigokon belüli szinténiák*

Nagy jelentőséggel bírtak azok a scaffoldok, amelyek több gént, vagyis lehetséges kapcsolt, szinténikus genetikai elemet tartalmaztak. A Btau_5.0.1 genommal a gímszarvas kromoszómákat összevetve kiderült, hogy az intra-scaffoldikus gének a szarvasmarha ortológ génekkel megeggyezően lokális kapcsoltságokat mutattak. Minden több génes gímszarvas scaffold (3422) és ortológ bovin kromoszómális szegment esetében észleltem ezeket a kiterjedt szinténiákat. A bizonytalan helyzetekben egy másik közel rokon faj, a juh genomját is bevontam a vizsgálatba. Elmondható tehát, hogy a szinténia és a gének sorrendje nem csupán térképpont markerekre kromoszómális szinten igazolható, hanem szubkromoszómálisan, a scaffoldokon belül is megnyilvánul. (Példák a 13. ábrában sc sc-8, sc-9, sc-511).

8.2 A CERELA1.0 GENOM ANNOTÁCIÓJA

A gímszarvas gének identifikációja a szarvasmarha, juh és emberi transzkriptom és proteom illetve a *de novo* létrehozott szikaszarvas transzkriptom felhasználásával, a MAKER program alkalmazásával történt. A „pipeline”-okat futtató MAKER 19368 fehérje kódoló gént azonosított. Az identikus gének sorrendje túlnyomórészt megegyezett a szarvasmarha és a gímszarvas genomban, vagyis ezen gén mintaszettekben is valós szinténiák figyelhetők meg a két rokon faj között. A 6. táblázat tartalmazza a kromoszómákra vonatkozó lényegesebb információkat. Látható, hogy például a gímszarvas 1. kromoszóma egy az egyben megfeleltethető a szikaszarvas, a szarvasmarha és a juh ortológ kromoszómákkal, ugyanakkor például az emberi 11. kromoszóma a gímszarvas 1. és 2. kromoszómájának „fúziójával” írható le. A gének kromoszómális sűrűségére vonatkozó információkat a 14. ábrán foglaltam össze.

14. ábra. A gímszarvas gének Mbp-kénti eloszlása az egyes gap nélküli kromoszómákon.

A teljes referencia genomot feltöltöttük az NCBI-ba, ahol genom böngészővel megjeleníthetők az annotált gének (25. melléklet).

Ezenfelül sikerült meghatározni 589 rRNS kódoló gén (LSU, SSU) helyzetét, amely 983 Kbp hosszúságú szekvenciát fedett le, ami a pszeudokromoszóma összhossz 0,0029%-át tette ki. Az 5s rRNS génekből 1029-et (96 Kbp, a CerEla1.0 0,0028%-a), a tRNS-ekéből 2096-ot (128 Kbp, a CerEla1.0 0,0038%-a) a mikro RNS-ekéből pedig 264-et (27,7 Kbp, a CerEla1.0 0,0008%-a) lokalizáltam. A kapott eredmények jól egyeznek az egyéb emlős fajokban leírtakkal. A MAKER annotációs lépések részét képezi a genom repeatmaszkolás, ettől függetlenül önmagában is elindítottuk a RepeatMasker programot, amely együttesen 769492957 bp hosszúságú repetitív szakaszt ismert fel, ami a teljes genom 22,73% -át adja ki (26. melléklet).

6. táblázat. A gímszarvas kromoszómák legfontosabb adatai. Az adatokat átvettem: a, (Slate és mtsai., 2002) és (Szabolcsi és mtsai., 2008), b, (Bonnet és mtsai., 2001)

Gím-szarvas chr	**Teljes hossz (Mbp)**	**Hossz -gap (Mbp)**	**chr[a] hossz (cM)**	**Protein kódoló gének**	**Mar-kerek**	**Szika-szarvas[b] chr**	**Szarvas-marha[a] chr**	**Juh[a] chr**	**Ember[a] chr**
1	104,50	60,73	78,1	622	16	11	15	15	11
2	63,26	35,77	74,9	541	11	29	29	21	11
3	88,46	48,94	57	530	12	27	5	3	12
4	81,20	48,91	98,4	1035	11	1p	18	14	19
5	178,03	108,83	99	1698	26	2	17, 19	17, 11	4, 12, 17
6	73,11	38,60	79	333	13	22	6	6	4
7	66,84	38,71	48,2	492	10	23	23	20	6
8	55,92	34,46	54,8	455	11	18	2	2	1, 2
9	141,95	83,05	84,7	1025	15	5	7	5	5, 19
10	55,94	33,20	52,2	702	16	30	25	24	7, 16
11	140,39	82,43	64,3	910	16	8	11	3	2, 9
12	127,78	75,24	116,4	794	20	16	10	7	14, 15
13	89,79	52,03	112,5	472	8	21	21	18	14, 15
14	103,59	59,60	80,2	587	13	12	16	12	1
15	125,27	74,13	53,8	712	10	9	26, 28	22, 25	1, 10
16	62,95	35,72	55,6	327	6	32	8	2	8, 9
17	79,72	43,06	58,7	246	11	16	6	6	4
18	152,66	87,31	97,2	626	11	4	4	4	7
19	127,24	73,00	83,9	619	10	7	1	1	3, 21
20	149,34	92,56	102,6	1132	17	3	3	1	1
21	107,36	61,63	76,8	409	7	13	14	9	8
22	63,92	36,20	80,2	520	7	19	5	3	22
23	109,47	65,62	74,3	687	13	1q	13	13	10, 20
24	78,16	47,34	76,3	519	11	26	22	19	3
25	96,54	52,76	58,5	289	7	20	20	16	5
26	55,10	31,31	66,3	240	9	31	9	8	6
27	84,64	47,07	63,1	321	6	24	24	23	18
28	82,07	44,57	85,1	245	8	17	9	8	6, 9
29	80,17	45,64	59,9	307	7	15	8	2	9
30	117,80	63,65	100,2	382	8	10	12	10	13
31	75,46	38,50	56,1	193	3	25	1	1	21
32	60,01	32,24	46,3	196	6	28	27	26	4, 8
33	121,43	67,20	119	463	6	12	2, 22	2	2, 3
X	181,54	86,34	18,4	716	3	X	X	X	X
Y	4,03	1,81	-	23	1	Y	Y	Y	Y

8.3 CERELA1.0 KROMOSZÓMÁK CENTROMERON POZÍCIÓI

A citogenetikai tanulmányokkal ellentétben a gímszarvas genetikai térkép nem jelezte a centromeronok pozícióját. A kapcsoltsági csoportok pontjait itt nem lehet egységesen, a centromeronok helyzetéhez orientálni. Ilyen módon 6 szarvasmarha kromoszómával (Bt5, Bt6, Bt2, Bt8, Bt1, Bt9) párosan megfeleltethető 12 gímszarvas kromoszómán (Ce3-22, Ce6-17, Ce8-33, Ce16-29, Ce19-31, Ce26-28) is nagy biztonsággal meg lehetett adni a centromeronok helyzetét.

Az akrocentrikus Ce19, a Bt1 disztális felével, a szintén akrocentrikus Ce31 viszont a Bt1 proximális szakaszával ortológ. A helyzet azonban összetettebb ennél, hiszen a Ce19 kromoszóma ősében egy jól definiálható törés és transzlokáció (az alsó és a felső szegmens helyet cserélt) is lejátszódott az evolúció során. Ezt a folyamatot tökéletesen jól alátámasztják az egymással összhangban lévő citológiai és genomikai bizonyítékok (11. ábra).

A Ce28 és a Ce26 akrocentrikusak, a Bt9 akrocentrikus szarvasmarha kromoszóma két karjával azonosak. Ennek az az oka, hogy a szarvasfélék ősében ez a kromoszóma kettéhasadt. A Ce26 centromérájának nincs megfelelője a Bt9 szekvenciáiban, a Ce28-éval ellentétben. A Ce28 „örökölte meg" a Bt9 centromeronját, de a sávmintázatok és a genetikai térképpont sorrend összevetése alapján a gímszarvasban le kellet játszódnia egy paracentrikus inverziónak a fajképződés folyamán. Ennek következtében a kromoszóma összes ortológ markereit érintő, nagyméretű szegmense egészben megfordult, míg a centromeronhoz legközelebbi szekvenciák a helyükön maradtak. Két olyan esetet tártam fel, amikor egy gímszarvas kromoszóma két szarvasmarha kromoszómával ortológ.

A Ce5 metacentrikusságát citológiai és genomikai bizonyítékok egyaránt igazolták (8. ábra, Bonnet és mtsai., 2001). A Ce5 egyik karja a Bt17-el a másik a Bt19-el ortológ.

A Ce15 akrocentrikus, centromeronja megegyezik a Bt28-éval. Ez a gímszarvas kromoszóma egy olyan Robertsoniális fúzióval magyarázható, amelyben a centromeron közeli rész a Bt28-al, a távoli rész pedig a Bt26-al ekvivalens (tandem fúzió, 9. ábra) (Bonnet és mtsai., 2001) (7. táblázat).

7. táblázat. *Cervus elaphus* pszeudokromoszóma centromeron jellemzői.

C.e. chr	**Marker pozíciók (Mbp) centromeron**		**Ortológ B.t. chr**		
	proximális	**disztális**	**B.t. chr**	**C ***	**struktúra/illeszkedések**
1 A	BR3510 17149645	ACP2 97596143	15 A	15	*---Ce1 *---Bt15
2 A	TGLA86 60822183	KRN1 463507	29 A	29	*---Ce2 *---Bt29
3 A	ILSTS42 4768942	OarMAF23 66970424	5 A	5	*---Ce3 (*---Ce22) *--------Bt5----------
4 A	INRA38 11079846	JP23 76570752	18 A	18	*---Ce4 *---Bt18
q 5 M p	FGG 92136286 CM820 102829424	TGLA322 7052020 P4H4 161668917	17 A 19 A	f * 17/19	Ce5q----*----Ce5p Bt17---* *---Bt19
6 A	PDGFRA 62582512	PDE6B 2029781	6 A	n.c.*	(*---Ce17) *---Ce6 *----------Bt6-------
7 A	Bta_BoLaDIB 9457097	Delta14 49425282	23 A	23	*---Ce7 *---Bt23
8 A	TGLA226 52007152	ALPL 5210075	2 A	n.c.*	(*---Ce33) *---Ce8 *----------Bt2-------
9 A	RM12 6623210	ILSTS6 1213442281	7 A	7	*---Ce9 *---Bt7
10 A	HBA 229713	CM21 50800499	25 A	25	*---Ce10 *---Bt25
11 A	SLC8A1 109429674	ASS 9405010	11 A	11	*---Ce11 *---Bt11
12 A	BM3033 124818172	BMS2614 7220242	10 A	10	*---Ce12 *---Bt10
13 A	IGF1R 8595779	T193 79944215	21 A	21	*---Ce13 *---Bt21
14 A	CR2 6509868	BM1706 87172047	16 A	16	*---Ce14 *---Bt16
15 A	RT5 6402305	CYP17 89468482	28 A 26 A	28	*--------Ce15------ *---Bt28 *---Bt26
16 A	LPL 58006661	ORM1 10424686	8 A	n.c.*	(*---Ce29*---Ce16 *---Bt8----------
17 A	ILSTS93 886652	SPP1 47385900	6 A	6	*---Ce17 *---Ce6 *----------Bt6------

7. táblázat folytatása. *Cervus elaphus* pszeudokromoszóma centromeron jellemzői.

C.e. chr	Marker pozíciók (Mbp) centromeron		Ortológ B.t. chr		
	proximális	disztális	B.t. chr	C *	struktúra/illeszkedések
18 A	RM188	OarHH064	4 A	4	*---Ce18
	22742514	146986314			*---Bt4
19 A	CSSM19	OarMAF109	1 A	n.c.*	(*---Ce31) *---Ce19
	90353433	157115			*----------Bt1---------
20 A	INRA6	TGLA127	3 A	3	*---Ce20
	11820085	131656106			*---Bt3
21 A	CSSM66	BM2934	14 A	14	*---Ce21
	6097957	87929022			*---Bt14
22 A	TEXAN15	ACO2	5 A	n.c.*	(*---Ce3) *---Ce22
	59941417	10015108			*----------Bt5--------
23 A	PLC154	OarMAF18	13 A	13	*---Ce23
	2389036	102873535			*---Bt13
24 A	HUJ175	HIS-H1	22 A	22	*---Ce24
	27402800	70501351			*---Bt22
25 A	BM1225	BM4107	20 A	20	*---Ce25
	7980198	54541102			*---Bt20
26 A	CM102	SOD2	9 A	n.c.*	(Ce28--i--*) *---Ce26
	9035997	44533880			*----------Bt9---------
27 A	JP38	PAI2	24 A	24	*---Ce27
	34310557	84583784			*---Bt24
28 A	OarCP021	ETH225	9 A	9(i)	Ce28--i--* (*---Ce26)
	81944836	13060454			*----------Bt9----------
29 A	UWCA47	CM100	8 A	8	*---Ce29(*---Ce16)
	1242245	76016626			*----------Bt8-------
30 A	RM178	ILSTS33	12 A	12	*---Ce30
	21709783	108535018			*---Bt12
31 A	BM6438	RM95	1 A	1	*---Ce31 (*---Ce19)
	2858627	25506844			*----------Bt1--------
32 A	BM6526	BM203	27 A	27	*---Ce32
	15764812	55918160			*---Bt27
33 A	INRA40	INHBB	2 A	2	*---Ce33 (*---Ce8)
	2090945	93576890			*----------Bt2--------
X A	B9	T27b	X SM	n.c.*	*-------CeX
	11882131	51457707			--BtX--*----

A/SM: akrocentrikus/szubmetacentrikus kromoszóma, p/q: rövid/hosszú kromoszóma karok, C.e./B.t:, *C. elaphus/B. taurus*, i: inverzió, *: centromeron, c *: C.e. kromoszóma megfeleltethető B.t. kromoszómának, (*): a centromeron az akrocentrikus kromoszóma bármelyik végén lehet, f *: kondenzált centromeronok, n.c. *: nincs megfeleltethető centromeron a B.t. kromoszómán, * ----- * -----> * ---------- két akrocentrikus kromoszóma tandem fúziója, -------* *-----> -------*------ két akrocentrikus kromoszóma központosított fúziója (Robertsoniális transzlokáció).

8.4 KROMOSZÓMA ÁTRENDEZŐDÉSEK

Az inverziók, a transzlokációk, a kromoszómális fúziók és a törések áthelyezik és elkülönítik a térképponti markereket és új környezetet hoznak létre az átrendeződések töréspontjai körül. A gímszarvas-szarvasmarha rokoni kapcsolatát tekintve ezek az "evolúció által létrehozott" átrendezések új szomszédos szekvencia régiók kialakulásához vezettek. A CerEla1.0 genomban 26 darab szarvas-bovin átrendezést észleltem, amely 18 inverziót, 2 transzlokációt, és 6 kromoszómális fúziót és hasadást foglalt magába (8. táblázat).

Abban a 6 esetben, amikor 1 akrocentrikus szarvasmarha kromoszóma két akrocentrikus gímszarvas kromoszómával feleltethető meg az elülső Ce kromoszóma utolsó szegmensét és a hátulsó Ce kromoszóma első szegmensét nem lehetett egyértelműen meghatározni a szekvencia adatok (scaffoldok and contigok) és a szarvasmarha genom alapján. Az ortológ szarvasmarha régióhoz illeszkedő scaffoldokat és contigokat úgy osztottam fel a két gímszarvas kromoszóma között, hogy figyelembe vettem a gímszarvas genetikai térképeken található rekombinációs távolságokat és ezeknek megfelelően, arányosan, illetve a szarvasmarhában tapasztalt sorrend szerint rendeztem el őket. Az ilyen típusú kombinált DNS szekvenciák a CerEla1.0 genom 5 %-át (0,166 Gbp) teszik ki (8/a. táblázat).

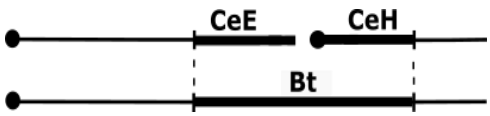
Az inverziók 54 gímszarvas-szarvasmarha szekvencia váltás („switch pont”) pontot eredményeztek, amelyek a túlnyúló, „flanking” régióikkal együtvéve tulajdonképpen megegyeztek a szomszédos MMSc-k által határolt, összesen 462,9 Mbp szekvenciával. Ez a CerEla1.0 genom 13,5 %-át jelenti (8/b. táblázat). Ezek a megfordított szekvencia szegmens határok azonban csak becsültek, konkrét helyüket ezidáig még nem sikerült megállapítanom.

A CerEla1.0 jelenlegi formájában a „switch” pontok „flanking” régiói a szarvasmarha szekvencia sorrendjét követő scaffoldokból és contigokból

állnak. Azonban bízunk abban, hogy a jövőbeli összehasonlítások más emlős genomokkal és a gímszarvas nagyobb sűrűségű genetikai térképével való összevetés a „switch” szegmensek tekintetében is precízebb gímszarvas genomot eredményezhet.

Összességében elmondható, hogy a CerEla1.0 81,5% -ában a gímszarvas gének az MMSc közötti szegmensekben a szarvasmarha ortológ gének sorrendjét követik, az MMSc-ken belül pedig egyértelműen a gímszarvas génsorrend érvényesült. A CerEla1.0 szekvencia 18,5%-ban (kromoszómális hasadások/fúziók, flanking részek inverziói) gímszarvas és szarvasmarha génjeinek szinténikus blokkjai kombinálódtak.

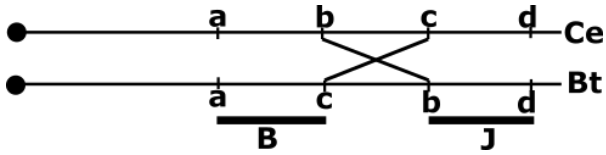
8/a. táblázat. CerEla1.0 genomi átrendeződései, topológia: kromoszóma fúzió/hasadás.



			∑közös (Mbp)	Közös Mbp-k a Ce chr-ák között (Mbp)	
Bt1	Ce31	Ce19	49,8	Ce31: 49,8	Ce19: <1
Bt2	Ce33	Ce8	31,6	Ce33: 27,8	Ce8: 3,7
Bt5	Ce3	Ce22	24,8	Ce3: 21,4	Ce22: 3,4
Bt6	Ce17	Ce6	42,0	Ce17: 32,0	Ce6: 10,0
Bt8	Ce29	Ce16	8,9	Ce29: 4,0	Ce16: 4,9
Bt9	Ce28	Ce26	8,9	Ce28: 0,002	Ce26: 8,9
∑			166,0		

8/a. táblázat. A kromoszómák Robertsoniális transzlokációja. CeE, CeH: Az elülső (E) gímszarvas kromoszóma disztális része és a hátulsó (H) gímszarvas kromoszóma proximális része, amelyek homológ régiói az ortológ szarvasmarha kromoszómán egymás mellett találhatók.

8/b. táblázat. CerEla1.0 genomi átrendeződései, inverziók túlnyúló „flanking” szegmensei.



	Flanking szegmensek (Mbp)	
	B	**J**
Ce4	0,8	4,2
	1,4	4,6
Ce5	2,0	5,5
Ce6	1,9	14,7
	4,8	16,3
Ce7	7,7	9,2
	0,9	16,7
Ce8	17,	1,2
Ce11	22,	13,6
Ce12	14,	0,1
	0,1	7,4
Ce14	1,7	11,6
	7,6	14,4
Ce18	6,5	10,1
	4,7	5,5
Ce20	4,8	1,4
Ce23	18,	9,3
	0,3	6,3
	1,9	6,0
Ce24	2,3	5,1
Ce26	13,	10,2
Ce27	34,	14,9
Ce28	1,9	16,3
	28,	0,002
Ce29	3,1	34,6
Ce32	13,	4,8
Ce33	2,1	1,2
∑	21	245,2

8/b. táblázat. Inverzió. Az egyszerűsített kromoszóma illusztrációiban megfordult részek, vagyis azok a genomiális régiók, ahol a gímszarvas-szarvasmarha marker szekvenciák egymással fordított helyzetben találhatók meg. B, J: Az inverziót határoló bal és jobb oldali túlnyúló régiók.

8.5 SNP MINTÁZAT/HETEROZIGOCITÁS VIZSGÁLAT A CERELA1.0 GENOM MENTÉN

A gímszarvas teljes genom szekvenálása során a DNS mindkét kópiáját leolvastuk, így a későbbiekben lehetőségünk nyílt rá, hogy megtaláljuk és azonosítsuk az egyed heterozigóta autoszómális variációit. Az eredeti, nyers readeket a teljes gímszarvas pszeudogenom szekvenciához illesztettük fel, ezután meghatároztuk az SNV-ket és a kis indeleket. Végül 2807458 SNV-t és 364689 indelt azonosítottunk. A MAKER annotációs pipeline eredmény fájljai alapján további a heterozigóta SNV-ket annotáltunk. Ily módon összesen 17700 nem-szinonim és 14252 szinonim SNV-t találtunk meg.

9 KÖVETKEZTETÉSEK ÉS JAVASLATOK

9.1 A GÍMSZARVAS GENOM ÖSSZEÁLLÍTÁS ÉRTÉKELÉSE

Az értekezés témája kiemelkedően fontos a hazai genomikai kutatás szempontjából, hiszen Magyarország első, valódi, nemzetközileg is elismert teljes referencia genom szekvencia előállításáról szól. Nagymértékben hozzájárul a gímszarvas populációgenetikai profiljának megismeréséhez, ezáltal lehetőséget teremt új mikroszatellita és SNP markerek tervezésére. A gímszarvas referencia genom segítheti e nagyvad evolúcióbiológiájának, leszármazási vonalainak a megismerését, és a farmtenyésztési eljárások fejlesztését. A gímszarvas gének szerkezetének és azok promóter szekvenciáinak ismerete alapján például feltárhatjuk a kapitális aganccsal rendelkező egyedek genetikai hátterét. Feltérképezhetők lesznek a csont-agancs metabolizmusban és tumor képződésben résztvevő gének működései és orvosbiológiai vonatkozásai. A disszertációban ismertetett bioinformatikai módszerek megfelelő mintául szolgálhatnak más mezőgazdasági vagy környezetvédelmi szempontból értékes faj referencia genomjának a létrehozásához.

A bioinformatikai munkát megelőzte a mintagyűjtés, a DNS izolálás és a teljes genomi DNS szekvenálása. A DNS mintavétel a Kaposvári Egyetem Vadgazdálkodási Tájközpont Bőszénfai Szarvas farmján élő 7 éves, kapitális gímszarvas bika véréből történt az állatvédelmi törvények figyelembevételével. A DNS izolálás eredményeképpen keletkezett nagy tisztaságú DNS-t a dániai Aros Applied Biotechnology cég szekvenálta meg Illumina HiSeq2000 készülékkel, aminek következményeként nagy mennyiségű, további vizsgálatokra alkalmas bioinformatikai nyers adat 2 milliárd read szekvencia jött létre. A későbbi populációgenetikai vizsgálatok szempontjából érdemes lenne több, különböző magyarországi populációból

származó körülbelül 150-300 gímszarvas egyedre kiterjeszteni a mintavételt és a teljes genom szekvenálást. A szekvenálási módszer tekintetében ajánlatos az Illumina technológiánál maradni, hiszen számos haszonállat és egyéb faj (házi macska, szarvasmarha) teljes referencia genom *de novo* előállításakor bizonyult jó választásnak.

A Broad Institute ALLPATHS-LG programjával a read szekvenciákból contigokat és scaffoldokat állítottunk össze. A folyamat eredményeképpen contigok, majd ezek összeillesztéséből scaffoldok keletkeztek. A scaffoldok száma 34724 volt. Viszonylag sok 2000 bp alatti, kisméretű és kevés nagy méretű scaffold jött létre. Amelyekből a kromoszómába rendezés szempontjából a 8 Kbp fölöttiek bizonyultak a legjobban használhatónak. A scaffoldok N50 érték-e 430 Kbp a teljes hosszukra nézve, és N50=265 Kbp értéket kaptunk a gap-mentes scaffoldok esetében. Ezen adatok alapján is valószínű, hogy a szükségesnél több kitöltő, N karaktert rakott be a program az összeszerelés során. E tény akkor vált kézzelfoghatóvá, amikor az elkészült teljes, kromoszómába rendezett gímszarvas genom fizikai és a cM távolságainak összevetését végeztem el. A referenciaként használt *C. elaphus* géntérkép hosszúsága 2532 cM (Slate és mtsai., 2002), míg az „összeszerelt" CerEla1.0 referencia genom összes hosszúsága 3,4 Gbp, vagyis 1 cM 1,34 Mbp-nak felel meg. Ez az érték szignifikánsan nagyobb, mint az általánosan használt megközelítő 1cM/1Mbp érték (a „hüvelykujj mérték/zsinórmérték, angolul thumb role") vagy a szarvasmarha genomra megállapított 0,8 Mbp/1cM (Arias és mtsai., 2009). Érdemes megemlíteni, hogy a szarvas és a „bovin" térképezési rendszer szignifikánsan különbözött egymástól, mivel a gímszarvas térképet interspecifikus „outbred" nemzedékből hozták létre. Egy közelmúltban megjelent nagy sűrűségű, *C. elaphus* genetikai térképen a cM/Mbp arányt 1,04 értékűnek becsülték meg a leszármazási (pedigree), azaz 3 generáció családfákra épülő térképezési rendszer alapján (Johnston és mtsai.,

2017). A gímszarvas CerEla1.0 referencia genom látszólagosan 25%-kal hosszabbnak adódik, mint a szarvasmarha, Btau_5.0.1 (NCBI, Science) referencia genom. Megjegyzendő, hogy a Btau_5.0.1 genomból csak a „gap nélküli", 2, 7 Gbp hosszúságú DNS szekvenciákat használtam fel templátként. A gímszarvas genom 0,7 Gbp-nyi „többlet" hosszúságának az okát keresve elvégeztem a CerEla1.0 és Btau_5.0.1 pszedogenom ortológ szegmenseinek összehasonlítását kisebb genomiális távolságokon. Ezeket a mindkét fajban egymással szinténikus szegmenseket a szomszédos gímszarvas genetikai markerek határozták meg. A 27. mellékletben látható, hogy néhány kivételtől eltekintve a gímszarvas és a szarvasmarha pszeudogenomok mentén a gímszarvas szegmensek egyenletesen 1,25-ször hosszabbak, mint a szarvasmarha szegmensek (Kivétel ez alól a Ce11 kromoszóma, ahol ez az arány 2,2). Ez azt is jelenti, hogy a CerEla1.0 scaffoldok arányosan 1,25-ször hosszabbak, mint a velük megfeleltethető szarvasmarha genomi régiók.

Összevetve a 25 %-os (0,7 Gbp) extrahosszúsággal rendelkező CerEla1.0 pszeudogenomot (3,4 Gbp) a Btau_5.0.1 pszeudogenommal (2,7 Gbp) megállapítható, hogy túl sok gaprégiót („NNN") épített be az ALLPATHS-LG program a contigok közé a scaffoldokban. Feltéve, hogy a *B. taurus* és *C. elaphus* genomjai méretüket tekintve lényegében megegyeznek egymással, továbbá figyelembe véve, hogy a CerEla1.0-ban található a contigok össz DNS-szekvenciái 1,9 Gbp-t tesznek ki feltételezhető, hogy a szarvasmarha genommal arányosítható 0,8 Gbp gaprégió helyett 1,5 Gbp-nyi gap került a scaffoldokba összesen. Ezek eloszlása azonban a fizikai távolságoknak megfelelően arányos. Figyelembe véve ezen megfontolásokat a CerEla1.0 contigjai 70 %-ban fedik le a Btau_5.0.1 genomot. (1,9/2,7, ahol 1,9 GBp contig, 2,7 Gbp CerEla1.0 genom hossz). A szarvas szarvasmarha szinténikus szegmens hosszaknál tapasztalt 1,25 arány a Ce11 kromoszóma esetén lényegesen több, 2,2 volt (27. melléklet), ennek magyarázatára nincs javaslat.

Más haszonállatok legfrissebb verziójú genomja lényegesen kevesebb, hosszabb és kisebb mennyiségű gapet tartalmazó scaffoldból épülnek fel, bár megjegyzendő, hogy ezek első verziói, sokkal rosszabb minőséget (kevesebb ismert nukleotidot, és kódoló régiót) mutattak, mint az első *de novo* gímszarvas referencia genom. A gímszarvas scaffoldokban a térképponti markerek szegmensei és gének sorrendje szinténiát mutatott a közel rokon faj szarvasmarha ortológ szegmenseivel, vagyis a scaffoldok lényeges, nukleotid szekvenciái mégis jónak bizonyultak, voltaképpen formai eltérésről és nem tartalmi tévesztésről van szó. További gímszarvas egyedek assembly készítésekor megfontolandó más ALLPATHS-LG paraméterek vagy újabb programok (DISCOVAR, Love és mtsai., 2016) használata.

A scaffoldok kromoszómába rendezése Jon Slate-féle gímszarvas genetikai marker térkép és a szarvasmarha referencia genom felhasználásával történt meg. A gímszarvas genetikai térképén 621 EST, RFLV, STS, protein genetikai markert határoztak meg, ám ezek szekvenciáját irodalmi adatokból és bioinformatikai adatbázisokból kellett kigyűjteni. A problémát az jelentette, hogy 229 AFLP-hez sehol sem lehetett szekvenciát találni, tehát ezek kiestek a további vizsgálatokból, és a többi kategóriánál is előfordult, hogy ismeretlen maradt egy-egy marker szekvenciája. Ezen okok miatt 365 markerpont DNS szekvenciáját azonosítottam és a továbbiakban kereső-szekvenciaként BLAST programmal illesztettem a scaffoldokhoz és a szarvasmarha referencia genomhoz. Az ilyen módon megtalált scaffoldokat térképpont vagyis "mapmarker" scaffoldoknak (MMSc-k) neveztem el. A vizsgálathoz több referencia genom verziót is használtam, bár az utolsó, akkortájt legfrissebb is elegendő lett volna (Btau_5.0.1).

A gímszarvas géntérkép pontjai hosszú szakaszokon azonos sorrendben, ko-lineárisan helyezkedtek el a szarvasmarha genomban és a gímszarvas scaffoldokban egyaránt, emiatt az összehasonlító géntérképezési elvet

felhasználva a szarvasmarha ortológ génekkel kihalászott gímszarvas scaffoldokkal töltöttem fel a marker pontok térképközeit. Ilyen módon 13748 valamilyen genetikai elemet tartalmazó scaffoldot tudtam elhelyezni.

Ez idáig BLAST parancsok különböző formáit építettem be a saját kézzel írt szkriptekbe. A későbbiekben ellenőrzésképpen a géneket tartalmazó scaffoldokat és a többi, 2 Kbp-nál hosszabb, referencia gént nem tartalmazó 15205 scaffoldot illesztettem fel a szarvasmarha referencia genomra LASTZ MUMmer és BWA programok alkalmazásával. Ezeket neveztem el inter-referencia gén scaffoldoknak (IRGSc).

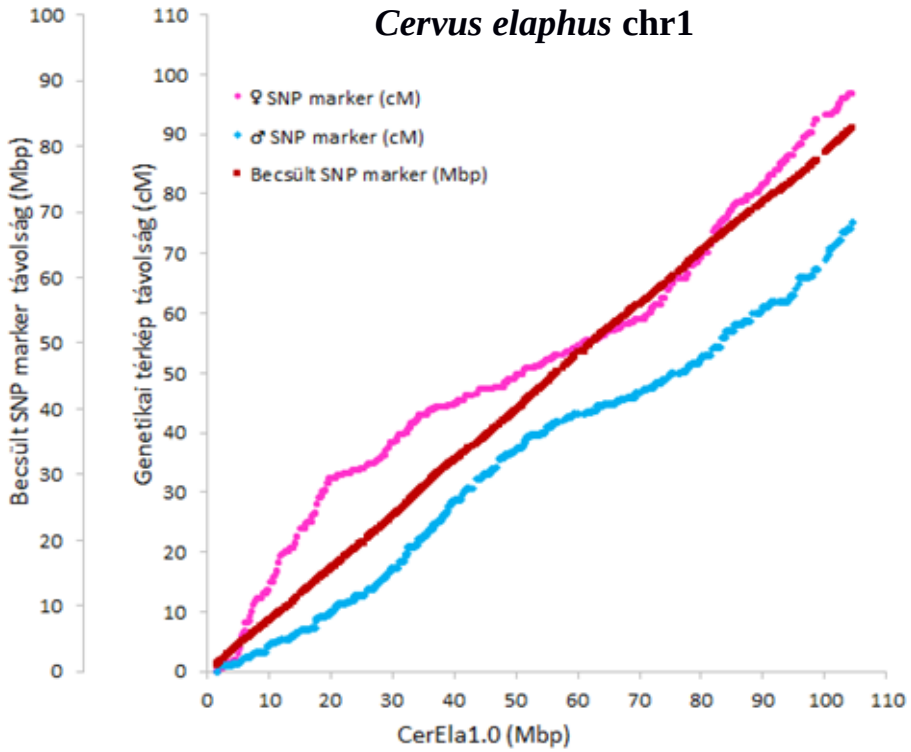
9.2 A KROMOSZÓMÁBA RENDEZETT GÍMSZARVAS GENOM VALIDÁLÁSA

A CerEla1.0 gímszarvas genomtól teljesen függetlenül, egy angol-újzélandi kutatócsoport által meghatározott, több mint 38 ezer nagyrészt gímszarvas kisebb részt szikaszarvas SNP rekombinációs pont elemzését tartalmazó nagy sűrűségű gímszarvas genetikai térkép is elkészült (Johnston és mtsai., 2017, Brauning és mtsai., 2015). Ez a géntérkép nagyszámú, 38083 SNP marker szekvenciát tartalmaz, amely egy komparatív vizsgálattal validálja a magyar gímszarvas referencia genomot, a CerEla1.0-t.

Az összehasonlító munka első lépéseként kikerestem és letöltöttem az összes angol-újzélandi SNP markert a 150-150 bázispáros szárnyi (flanking) régióikkal. Ezeket a markereket, mint kereső szekvenciákat a mi gímszarvas genomunkra illesztettem fel BLASTN paranccsal. Kiválasztottam a legjobb találatokat, amelyek eredményeiből táblázatokat generáltam minden egyes autoszómára olyan módon, hogy látható legyen a mi gímszarvas kromoszómális pozíciók megabázispárban megadva és a kereső angol-újzélandi SNP marker megabázis és cM pozíciója. A kapott eredményekből pont diagramokat szerkesztettem excel programmal. Az adatok és ábrák

alapján jól láthatók a lehetséges töréspontok és kromoszóma átrendeződések. A 38083 pontból 32408-at tudtam elhelyezni az elvártnak megfelelő kromoszómán, vagyis az összes pont 85%-át. Ezen pontoknál 33 esetben tapasztaltam több pontból álló rész nagyobb inverzióját. A fordulások legvalószínűbb oka a Slate-féle 621 marker és a Johnston-féle SNP marker genetikai távolság alapú géntérképek között fennálló különbség, illetve, hogy két fordított helyzetű, szomszédos Slate marker esetében nem lehettem biztos abban mekkora flanking részek tartoznak az invertált szakaszokhoz. Továbbá a gímszarvas referencia genom elkészítésénél a Slate térkép egyes szomszédos markerpontjai fordított helyzetűnek tűntek a szarvasmarha referencia genomhoz képest is, ami 54 gímszarvas-szarvasmarha inverziót, összesen 462,9 Mbp szekvenciát adott. A CerEla1.0 létrehozásához a Slate-féle géntérkép állt a rendelkezésre, tehát a scaffoldok elhelyezése e sorrendiségnek megfelelően történt. Az új-zélandi SNP pontos grafikus ábrákon is világosan látszik, hogy a megfordult részek a Slate szomszédos markerek régióját érintik. A felillesztett pontokból csupán 604 adott kiugró értéket, ami kevesebb, mint az elhelyezett pontok 1,9 %-a. 1019 SNP -t más kromoszómán lokalizált a BLAST program, e pontokat a nem megtalált kategóriához soroltam. Ezen pontok diszkrepanciáját az okozhatta, hogy a gímszarvas pszeudokromoszóma készítéshez használt közel rokon faj, a szarvasmarha kromoszómaszáma nem egyezik meg a gímszarvaséval. 19 autoszóma és a nemi kromoszómák egymással megfeleltethetők a két fajban, ugyanakkor 2 gímszarvas kromoszóma 4 szarvasmarha kromoszómával és 12 gímszarvas 6 szarvasmarha kromoszómával ortológ. Ezen 12 kromoszómánál a genom összeállítás során nem lehetett biztosan tudni, hogy az utolsó Slate-féle genetikai térképpontokon túl a két gímszarvas kromoszóma miként „osztozik” az ortológ szarvasmarha kromoszóma „elosztandó” szegmensén, azaz mely része tartozik az egyik, illetve a másik gímszarvas kromoszómához. Eme

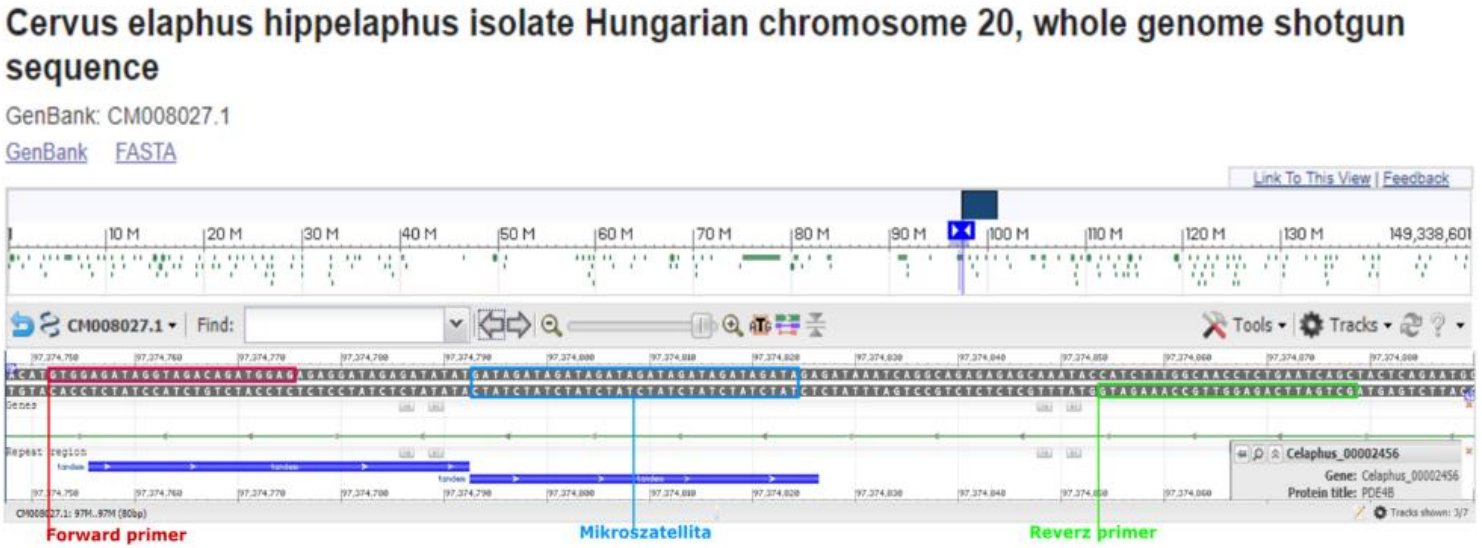
bizonytalan részek körülbelül 0,166 Gbp hosszú DNS szekvenciát tesznek ki. Összességében azonban elmondható, hogy a kapott eredmények meggyőző bizonyítékot szolgáltatnak a CerEla1.0 gímszarvas genom helyességének igazolásához. A következő 15. ábrán az 1-es gímszarvas kromoszóma látható. A korreláció egyértelműen visszaigazolja, hogy a CerEla1.0 összeállítása megfelel a fizikai valóságnak.



15. ábra. Az 1. gímszarvas kromoszómára felillesztett Johnston SNP markerek. x tengely: Az angol-újzélandi SNP géntérképi marker pontok lokalizációja a CerEla1.0 genomon (1. kromoszóma) Mbp megadva. y1 tengely: A sorrendjüket megtartó, angol-újzélandi SNP markerek saját kapcsoltsági géntérképük alapján cM-ban meghatározott géntérképi távolságai láthatók. Rózsaszín körök az ünőkre, a kék gyémánt a gímszarvas bikákra vonatkozik. y2 tengely: A sorrendjüket megtartó, angol-újzélandi markerek Mbp-ra átszámított értékei vannak feltüntetve, melynek jelölői vörösesbarna négyzetek. A Johnston-féle 1. kapcsoltsági csoport (kromoszóma) géntérképe 1158 pontot tartalmaz, amelyből 994-et lokalizáltam a *C. elaphus* 1. kromoszómán.

A Johnston-féle 5. kapcsoltsági géntérkép 2039 pontot tartalmaz, amelyből 1816-et helyeztem el a 5. gímszarvas kromoszómán. A Ce5 a Bt17 és a Bt19 Robertsoniális transzlokációjával írható le, a két szarvasmarha kromoszómának megfelelő kromoszóma kar közötti hyatus, egyben a gímszarvas 5 kromoszóma centromeron pozícióját is jelenti. Ez a fúzió igazoltan megjelenik mindhárom pontábránál a 28. mellékletben.

A nagysűrűségű SNP genetikai térképen kívül a munka kezdetétől fogva rendelkezésre állt a rendkívüli érzékenységű, 10 tetranukleotid mikroszatellita lokuszra épülő „parentage controll kit", a DeerPlex (Szabolcsi és mtsai., 2008). A DeerPlex azonosító ereje 1 a 30 trillióban (10^{-16} nagyságrend) (Szabolcsi és mtsai., 2014). A vizsgálat során sikerült ezeket a mikroszatellitákat a CerEla1.0 genomon azonosítani, kiterjedt DNS szekvencia környezetben elhelyezni.



16. ábra. Az NCBI genom böngészője, amely a T108b mikroszatellita lokalizációját mutatja a gímszarvas referencia genomon, a 20 kromoszómán.

A gímszarvas genom MAKER programmal történő annotációja során 19368 protein-kódoló gént találtunk, amely teljesen megfelel a többi emlős fajban is leírt 19000 és 21000 közé eső fehérje kódoló gén számnak és a kérődzők ismert génjeinek 90 %-át lefedi. A gímszarvas gének és szarvasmarha ortológjaik sorrendje kolinearitást mutat a CerEla1.0 és a Btau_5.0.1 genomban. A két közel rokon faj között tapasztalt genetikai szinténiák szintén igazolják a gímszarvas genom helyességét. Meghatároztuk a gének részletes szerkezetét (exon, intron szerkezete, 5' és 3' UTR régióik) evolúciós rokonságait (hasonló gének más fajokban), funkcióit (a sejten, szerveken, szöveteken belüli, fejlődésben betöltött szerepük). Feltártuk az ismétlődő, repetitív szekvenciákat, LTR elemeket, transzfer RNS, kis RNS és riboszóma RNS géneket, és mindezen genetikai elemek helyzetét a kromoszómákon. Azonosítottunk 2,8 millió heterozigóta pont változatot (SNV) és 365 ezer apró deléciót és inszerciót a gímszarvas genomban.

9.3 A GÍMSZARVAS REFERENCIA GENOM TOVÁBBI HASZNOSÍTÁSA

A CerEla1.0 felhasználható SNP és/vagy mikroszatellita alapú egyedazonosításhoz, populációgenetikai szintű vizsgálatokhoz, leszármazási vonalak és evolúciós viszonyok feltárásához. Jelenleg publikálás alatt van egy cikkünk, amelyben új XY kromoszómás mikroszatellita markerek fejlesztéséről írunk. Ezek a nemi kromoszómás markerek nagyban megkönnyítik az apai és az anyai leszármazási vonalak feltárását, lehetővé teszik az objektív és egyértelmű egyedazonosítást, a tudatosabb és megbízhatóbb állatnemesítést és tenyésztést, valamint megbízható eszközt jelenthetnek a nagymértékű vadorzás visszaszorításában.

A CerEla1.0 genom lehetőséget teremt a gazdaságilag fontos tulajdonságokat meghatározó gének azonosítása (húsminőség, agancs méret), illetve egy

alkalmas DNS diagnosztikai készlet kifejlesztésére. Ehhez 100-120, kellően szelektált szarvas genomját kellene szekvenálni és illeszteni a CerEla1.0-hoz. Ily módon számos kapcsolt DNS markerre tudnánk szűrni. Ezek a marker kitek alkalmasak lennének a leszármazási viszonyok (szülők) feltárására (tenyészállat) és segítségükkel már születéskor előre jelezhetővé válna a későbbi fenotípus (például húsformák, minőség, későbbi kapitális agancs reménye).

További orvosbiológiai témájú cikkek alapjául szolgálhatnak a gímszarvas csont és agancs metabolizmus gének promóter szekvenciáinak az elemzései. Az éves agancs ciklusuk során a gímszarvas bikák produkálják az élővilág leggyorsabb és legnagyobb mértékű csont növekedését. Februártól májusig lehullatják csontos agancsukat, ezután 100-120 napig tart az új agancs épülése. Ilyenkor nagymennyiségű ásványi anyagot kell táplálkozással biztosítani és eljuttatni az akár 14-17 kg-os csontszervbe. A tápanyag kalcium bevitel kevés, ezért vázelemekből-szegycsontból, bordákból, egyes csigolyákból-transzportálják a kalciumot és foszfátot az agancs mineralizációjához, ezzel csontsűrűség csökkenést, azaz fiziológiás csontritkulást indukálva az érintett vázcsontokban. Később az ásványi sók a döhérség kondíciójavító ideje alatt a dús vegetációból táplálkozás útján visszapótlódnak, azaz a csontsűrűség csökkenés regenerálódik a csontvázban (Kierdorf és mtsai., 2009). Kutatócsoportunk előzetes vizsgálatai szerint 8 a csontfejlődésben fontos szerepet játszó gén esetében a barkás agancs csontos részében sokkal nagyobb 10-szeres 30-szoros is lehet a génexpresszió, mint a vázcsontokban. Ezen gének közül a COL1A1 gén 1 és 5 kb-os promóter régiójában több runx2 transzkripciós faktor kötőhely található a gímszarvasban, mint az emberben vagy a szarvasmarhában, mert ezek a gének aktívabban működnek a gímszarvasban (Stéger és mtsai., 2010, Gu és mtsai., 2008). A jövőben szeretnénk megvizsgálni, hogy vajon a többi csontfejlődéssel kapcsolatba

hozható gén esetében is több runx2 vagy osx transzkripciós faktor kötőhely motívum található-e, vagy esetleg valamilyen SNP, vagy INDEL eltérés okozhat-e bármilyen osteogenezissel vagy osteoporozissal összefüggésbe hozható jelenséget. A runx2 és az osx transzkripciós faktorok kötőhelyein kívül egyéb konzervált csontgén promóter motívumok is érdekesek lehetnek, tehát érdemes lenne többszörös szekvencia illesztéseket végezni közel rokon, vagy csontfejlődés szempontjából figyelemre méltó fajok és a gímszarvas ezen DNS szakaszai között. Az osteoporosis genomikai vizsgálatainak mintájára elvégezhetünk egyéb orvosi biológiai kutatásokat a szervfejlődés/regeneráció, robusztus szövet gyarapodás/tumor biológia témakörében.

A gímszarvas genom összeállításának módszertana remek útmutatást jelenthet más mezőgazdaságilag, orvosbiológiailag vagy természetvédelmi szempontból fontos fajok kromoszómába rendezett referencia genomjának elkészítéséhez. Maga a szekvencia templátként szolgálhat olyan közel rokon faj teljes pszeudogenomjának létrehozásához, mint például a dámvad és az őz. Mivel a referencia genom elérhető az NCBI online adatbázisában, így nemzetközileg is jelentős mértékben hozzájárul a gímszarvas és egyéb kérődzők genetikai vizsgálatához, más szarvas fajok genomjának megismeréséhez.

10 ÚJ TUDOMÁNYOS EREDMÉNYEK

A doktori kutatásom szervesen kapcsolódott a gímszarvas (*Cervus elaphus*) genom programhoz, amely a világ első elismert gímszarvas referencia genom összeállítása. Feladataim döntően a bioinformatika területére koncentrálódtak. A szerteágazó projekten belül a munkámhoz kapcsolódó új tudományos eredményeim az alábbiak:

1. A tudományos vizsgálatban használt kettős referencia vezérelt illesztéssel nagyobb pontossággal lehetett pszeudokromoszómákba rendezni a rövidebb gímszarvas DNS szekvenciákat (*de novo* scaffoldokat), mint az általánosan elterjedt egyetlen referenciára történő illesztés esetében. Referenciaként egy gímszarvas közeli rokon faj, a szarvasmarha (*Bos taurus*) jól meghatározott és annotált referencia genomja, továbbá a Slate-féle 34 kapcsoltsági csoportból és 621 marker pontból álló gímszarvas genetikai térkép szolgált. A marker pontokból azonban csak 361-et tudtam felhasználni az illesztésekhez.
2. A bioinformatikai munka során egy újonnan bevezetett algoritmus alapján dolgoztam (6. ábra.):
    a) Megfeleltettem egymással a gímszarvas genetikai térkép markereket és a *de novo* gímszarvas scaffoldokat.
    b) Megkerestem a gímszarvas géntérképi markerekhez nagyon hasonló szekvenciákat a szarvasmarha referencia genomon.
    c) Azonosítottam a szarvasmarha referencia gének ortológ szekvenciáit a gímszarvas scaffoldokon.
    d) Felillesztettem a *de novo* gímszarvas scaffoldokat az evolúciós változásoknak megfelelően átalakított szarvasmarha referencia genomra.

3. Minden több gént hordozó gímszarvas scaffold esetében látható, hogy az intra-scaffoldikus génjeik kiterjedt szinténiákat mutattak a szarvasmarha ortológ génekkel. Ebből következően a gének sorrendje nem csupán a térképpont markerek által határolt szegmensekben és kromoszómális szinten, hanem a scaffoldokon belül is megegyezett.
4. A megfigyelt kapcsoltságok alapján elmondható, hogy a gímszarvas genom 81,5%-ában a gének az MMSc közötti szegmensekben a szarvasmarha ortológ gének sorrendjét, az MMSc-ken belül pedig a gímszarvas génsorrendet követik. A CerEla1.0 maradék 18,5%-ában, ahol a kromoszómális hasadások, fúziók és inverziók történtek a szarvasmarha és a gímszarvas génjeinek szinténikus blokkjai kombinálódtak.
5. Saját szkriptek készültek a különböző munkafázisokhoz.:
    a) A BWA és MUMmer illesztőprogramok eredményfájljainak értelmezését és feldolgozását saját Bash szkriptekkel végeztem (20/a, 20/b, 20/c melléklet).
    b) A meghatározott helyezetű scaffoldok kromoszómákká fűzéséhez Bash szkriptet írtam. A program a scaffoldok közötti rést az NCBI genomjainak mintájára 100 bázispárnyi N karakterrel töltötte fel (21/a, 21/b melléklet). (Végül azonban egy kollégám által készített Python szkript futtatásával létrehozott kromoszómákat töltöttük fel az NCBI-ba.)
6. A világon elsőként történt meg a gímszarvas gének szerkezetének és lokalizációjának meghatározása. Elkészült a kromoszómákba rendezett gímszarvas referencia genom teljes annotációja, beleértve ebbe a fehérje kódoló gének, a repetitív, a riboszómális RNS, a transzfer RNS és a mikro RNS szekvenciákat is.

7. A szarvasfélék kromoszómáinak sávtérképei és a szarvasmarha kromoszómák centromeron pozíciói alapján elhelyeztem a gímszarvas kromoszómák centromeronjait.
8. A bioinformatikai vizsgálat igazolta a gímszarvas kromoszómák feltételezett evolúciós változásait: 6 hasadás, 1 Robertsoniális transzlokáció és 1 Robertsoniális fúzió történt a szarvasmarha kromoszómákhoz képest. Bizonyítást nyert, hogy a Bt1 centromeron disztális karjával ekvivalens Ce19-ben lejátszódott egy hasadás és egy transzlokáció. A kromoszóma sávtérképek és géntérképi markerek helyzete alapján a Ce28-ban valószínűleg egy paracentrikus inverziós esemény zajlott le.
9. Összességében bioinformatikai kutatómunkámmal érdemben hozzájárultam a világ első, nemzetközileg elismert kromoszómákba rendezett, teljes gímszarvas referencia genom összeállításához. A gímszarvas referencia genomot (CerEla1.0) feltöltöttük a szigorú szabályrendszerrel működő NCBI genomikai adatbázis és böngésző weboldalára. Így mindenki számára ingyenesen online elérhetővé, letölthetővé (MKHE00000000.1) tettük a teljes genom szekvenciát. Ennek köszönhetően 12 tudományos cikk született az elmúlt 3 évben, amelyben hivatkoznak a CerEla1.0 genomra.

11 ÖSSZEFOGLALÁS ÉS TOVÁBBI TERVEK

A gímszarvas/Csodaszarvas genom projekt keretében egy bőszénfai kapitális gímszarvas bika teljes genomját szekvenáltattuk új-generációs technikával Illumina HiSeq 2000 szekvenáló platformon. A *de novo* összeállított scaffoldokból álló assemblyt kettős referencia vezérelt illesztést alkalmazva bioinformatikai programok felhasználásával kromoszómákba rendeztem. A teljes gímszarvas kromoszómába rendezett genom, CerEla1.0 összeállítása a gímszarvas rekombinációs géntérképe és a szarvasmarha referencia genom, vagyis a két referencia sorrendjének megfelelően, és a köztük lévő kolinearitást figyelembe véve történt meg. A gímszarvas az európai és a kárpát-medencei megafauna legértékesebb tagja. A CerEla1.0 online elérhető referencia genom szekvenciája a gímszarvast értékes modellállattá teheti a GWA-vizsgálatok (teljes genom asszociációs vizsgálatok) számára. Eme doktori disszertáció kiemelkedően fontos lehet egy teljesen új emlős referencia genom szekvencia létrehozása és annotációja szempontjából. A gímszarvas pszeudokromoszóma szett a molekuláris markereket tartalmazó gímszarvas kapcsoltsági genetikai térképén és egy közel rokon faj, a szintén párosujjú patásokhoz és fejcsap nyúlványosokhoz (*Pecora*) tartozó szarvasmarha jól annotált referencia genom szekvenciáján alapul. A CerEla1.0 26108 szekvencia elemből (scaffoldokból és contigokból) épül fel, 33 auto és egy X és Y pszeudokromoszómát tartalmaz, amely összesen 3,4 Gbp (1,9 Gbp gap-ek nélkül) hosszúságú DNS szekvenciát tesz ki. Ez azt jelenti, hogy a *C. elaphus* genom gyakorlatilag a teljes *B. taurus* genomot lefedi, továbbá a szarvasmarha ortológ gének 90%-át találtuk meg a gímszarvas pszeudokromoszómák mentén. A disszertáció ismerteti a szarvas (*Cervidae*) és a tulokformák (*Bovinae*) kromoszómái közötti kapcsolatot, vagyis megállapítja hol találhatóak egyezések, hol történtek fúziók, hasadások, inverziók és transzlokációk e fajok evolúciója

során. A gímszarvas gének olyan csoportokat, tömböket alkotnak a pszeudokromoszómákban, amelyek a szarvas (*Cervidae*) és a *Bovinae* DNS szekvencia szegmensek mozaikjainak tekinthetők. Azonban a „mapmarker" scaffoldok (MMSc-ek) között és a scaffoldokon belül, vagyis intra-scaffold szinten is lényegében csak a gímszarvas génsorrendnek megfelelő szinténiák találhatók meg. A CerEla1.0 genom hosszúsága 3,4 Gbp, mivel a scaffoldok *de novo* összeszerelése során az az ALLPATHS-LG program a szükségesnél több gap-et (N karakterrel jelölt részek a genomban) illesztett a contigok közé, emiatt a gímszarvas genom 1,25-szer hosszabb lett, mint a szarvasmarha genomja, amely 2,7 Gbp méretű Btau_5.0.1 genom verzióban. Ennek ellenére a teljes gímszarvas genomban és a scaffoldokban tapasztalt fizikai távolságok korrelációt mutatnak egymással. E jelenségnek az az oka, hogy a gímszarvas genomban található körülbelül 0,7 Gbp-nyi többlet „N szekvenciát" arányosan helyezte el az illesztőprogram a contigok közé, a scaffoldok generálása közben. A gímszarvas genom szekvencia (CerEla1.0) hitelességét igazolja:

1. Sikeresen azonosítottam a Slate-féle genetikai térkép 365 markerpontját, és a Deerplex 10 mikroszatellita szekvenciáját a gímszarvas pszeudokromoszómák mentén és a „Bovin" genomban egyaránt. A szarvasmarha és a gímszarvas genomi szegmenseit összehasonlítva kolinearitást tapasztaltam az ortológ kromoszómákban, valamint kromoszóma átrendeződéseket (fúziókat, hasadásokat és transzlokációkat) észleltem a két faj között. A citológiailag kimutatott kromoszóma átrendeződések megfeleltethetők voltak a szekvencia adatoknak.
2. Azokban az esetekben (3422), amikor egy scaffold két vagy több gént hordozott a gének sorrendje megegyezett a gímszarvas scaffoldok-ban és az ortológ szarvasmarha kromoszóma szegmenseiben.
3. A korábbi cikkekben feltárt agancsfejlődéssel és fiziológiás osteoporosissal kapcsolatba hozható összes klónozott gén helyzetét (Molnár és mtsai., 2007,

Gyurján és mtsai., 2007, Borsy és mtsai., 2009, Stéger és mtsai., 2010) megtaláltam a scaffoldokban és a pszeudokromoszómákon.

Habár *C. elaphus* genom szarvas-bovin sorrend alapján, mozaikosan épül fel, a két faj ortológ genom régiói között tapasztalt korrelációk alapján megállapítható, hogy a scaffoldok és contigok szinte teljesen lefedik az egész gímszarvas genomot. Ezt támasztja alá, hogy a bovin protein-kódoló gének 90%-ának az ortológját sikeresen azonosítottuk a gímszarvas genomban (19368 gén CerEla1.0-ban és 21514 Btau_5.0.1-ban az NCBI legfrissebb adatai alapján).

Terveink között szerepel a CerEla1.0 referencia genom szekvenciáinak és annotációjának folyamatos frissítése, bővítése esetleg más, újabb bioinformatikai programok, program verziók alkalmazásával. A közeljövőben reményeink szerint publikálni fogunk egy cikket, amelynek legfőbb célkitűzése a CerEla1.0 hitelességének igazolása. A cikkben ismertetni szándékozunk a CerEla1.0 kromoszómáinak összevetését egy angol-újzélandi kutatócsoport által leírt 38083 nagyrészt gímszarvas kisebb részt szikaszarvas SNP rekombinációs pontból felépülő gímszarvas genetikai térképpel (Johnston és mtsai., 2017). A CerEla1.0 szekvencia hozzásegíthet a *Cervidae* család vagy nagyobb taxonómiai csoportok például *Pecora*-k, kérődzők evolúciójának a megértéséhez, alkalmazható lehet hazai szarvasfélék (dámvad (*Dama dama*) vagy őz (*Capreolus capreolus*)), bioinformatikai jellegű genomikai kutatásában. A gímszarvas genom szekvenciái felhasználásával új kromoszóma specifikus mikroszatellita szetteket fejleszthetünk ki. Nagyszámú heterozigóta pontot ($2{,}8\times10^6$, SNV $3{,}6\times10^5$ indel) identifikáltunk a genomban, e vizsgálat több egyedre történő bővítésével lehetőség nyílhat az egész országot érintő gímszarvas populáció genetikai összefüggéseinek feltárására. Az egész genomra kiterjedő SNP és mikroszatellita elemzések nagymértékben megkönnyíthetik az egyedazonosítást, az apai és anyai vonalak nyomon

követését, a beltenyésztettség mértékének megállapítását, és fényt deríthetnek a genetikai introgressziókra. A bűnügyiazonosítás, vagy az allél összetétel meghatározása hasznosítható eszközt jelenthet a vadgazdálkodás területén. GWA vizsgálatokkal feltérképezhetjük a rekord trófeák genetikai hátterét. Számos orvosi kutatási területen alkalmazható és hasznosítható a gímszarvas és genomja, mint modellállat (például csont- és csontritkulás-kutatás, szervfejlődés, regeneráció, valamint a sejt és szöveti proliferáció /tumor biológia).

12 SUMMARY

In my Ph.D. thesis, the genome of the red deer *Cervus elaphus* was sequenced with the Illumina New Generation Sequencing technology. Guided by two references, i.e., by the co-linearity of the recombination map of the red deer *Cervus elaphus* and by the bovine reference genome sequence, and the whole genome as CerEla1.0 were successfully assembled. The sequences in CerEla1.0 were assorted in the pseudochromosome complement of the red deer, one of the most valuable members of the European megafauna, especially in the Carpathian Basin. CerEla1.0 made available deer, a so-far non-model animal, for Genome-Wide Association Studies. This study demonstrates the power, both in the annotation of a new mammalian genome and in the pseudochromosome set assembly, of the combination of the genetic map, based on molecular markers, of the target species (the deer) with the existing genome reference of a taxonomically related species (*Bos taurus*) in the present case, both belonging to ruminants Artiodactyla and *Pecora*. This deer genome (CerEla1.0) was assembled from 26,108 sequence elements (scaffolds and contigs), assorted in 33 auto plus X and Y pseudochromosomes with a total length of 3.4 Gbp (1.9 Gbp without gaps). This means that the scaffolds of the draft deer genome covered virtually the whole *B. taurus* genome, the contig sequences covered 70% of the same, and 90% of the bovine orthologous genes were identified along the deer pseudochromosomes. The relationships between the deer and bovine (and also of several ovine) chromosomes, i.e., congruencies, fusions, fissions, and inversions, were identified in a “semi-fine” scale (3. table, 22. supplement, 6., 7., 9., 10. figures). In its present state, the gene array in the red deer pseudochromosomes is a mosaic of deer and bovine segments. The order of the “mapmarker” scaffolds (MMSc-s) and the intrascaffold syntenies represent the valid deer arrays. The length of the

CerEla1.0 genome is 3.4 Gbp: due to the somewhat long inserted intercontig NNNs, it is 1.25-fold longer than the bovine genome, 2.7 Gbp in Btau_5.0.1. All along the two genomes, at sub-genomic, sub-chromosomal, and scaffoldic levels, the physical distances are directly correlated with each other in the same ratio of 1.25 (as shown in 23. supplement). The approximately 0.7 Gb (virtual) surplus length for the deer genome can be accounted for by the ALLPATHSLG program characteristics (i.e., insertion of NNN tracts in the scaffolds proportionally between the contigs). The validity of this deer genome sequence and pseudochromosome complement (CerEla1.0) is supported by:

1. Slate's 365 genetic mapmarkers and the 10 Deerplex STRs were also identified in the bovine genome, and the deer and bovine arrays were co-linear along entire chromosomes as well as in chromosome rearrangements (fusions, fissions, and translocations) relative to each other, i.e., between deer and cattle. It is worth noting that no reciprocal translocations were found in deer vs. bovine relation. 2. In all those cases (3422), when two or more genes were carried in a deer scaffold, the syntenies were identical in the deer scaffolds and the corresponding bovine genome segments.

3. All genes cloned previously (related to antler development and cyclic physiological osteoporosis of deer stag, Molnár, et al., 2007; Gyurján, et al., 2007; Borsy, et al., 2009; Stéger, et al., 2010) as well as the STR loci previously developed for multiplex PCR analyses (DeerPlex, Szabolcsi, et al., 2014) were recognized in the scaffold/contig sequences and localized in the pseudochromosomes.

Although the *C. elaphus* genome CerEla1.0 was arranged in a mosaic deer-bovine order, the fair correlations for the orthologous deer and bovine genomic regions indicated that the deer scaffolds and contigs covered nearly the whole deer genome. This was supported by the fact that some 90% of the deer orthologs of the bovine protein-coding genes were identified in deer (19368

for CerEla1.0 vs. 21427 Btau_5.0.1). Possible further studies: The reference genome CerEla1.0 of the red deer (*Cervus elaphus hippelaphus*) and its annotation, in accordance with fresh data from other programs, is under continuous monitoring and updating. The sequence data of the SNP-based map markers have been available recently (Johnston, et al., 2017). Confirmation of CerEla1.0 has become possible using the approach described in this work. The CerEla1.0 sequence and the high density SNP based genetic map by Johnston et al (2017) correlated with high accuracy. With few acceptions of 32408 SNPs, on average, nearly 1000 per chromosomes aligned in perfect order on the CerEla1.0 sequence. The sequence and the pseudochromosome complement of CerEla1.0 may provide a basis and a rich source for broader interests, including, among others: conservation genetics, refined evolution, and population studies within the family *Cervidae* [e.g., fallow deer (*Dama dama*) or roe deer (*Capreolus capreolus*)] as well as in a wider range of ruminants and *Pecora*. CerEla1.0 also provides a source for developing chromosome-specific microsatellite sets. This work is under progress already. A large number of SNP/heterozygotic sites were identified (2.8 × 106 SNVs, 3.6 × 105 indels and 30785 SNPs) and aligned to the deer pseudochromosomes. CerEla1.0 is a leading basis for future genome-wide SNP and microsatellite studies, which may shed light on inbreeding/outbreeding, which may help in the identification of gene introgressions and of descents for autosomal, maternal, and paternal lineages. Forensic identification, or definition of allelic compositions underlying phenotypes important, for example, in game management could also be possible areas of utilization. The exploration of the genetic secret of record antlers becomes possible by genome-wide association studies. Applications and utilization in several fields of medical research (e.g., bone and osteoporosis

research, organ development and regeneration, and robust tissue proliferation/tumor biology) are also feasible.

13 KÖSZÖNETNYILVÁNÍTÁS

Ezúton szeretném megköszönni a segítséget témavezetőimnek, Dr. Horn Péter és Dr. Orosz László Akadémikus uraknak a disszertáció elkészítése során nyújtott értékes tanácsokért és türelemért.

Külön köszönettel tartozom volt kollégáimnak Nyiri Annának, Dr. Nagy Tibornak és Dr. Barta Endrének a Mezőgazdasági Genomikai és Bioinformatikai Csoport vezetőjének a bioinformatikai munkám segítésért, és a sok hasznos tanácsért, tapasztalatért, amiket megosztottak velem. Hálával tartozom Dr. Frank Krisztiánnak és Dr. Stéger Viktornak a NAIK MBK Alkalmazott Vad és Haszonállat Genomikai Csoport vezetőjének és intézetigazgató helyettesnek a DNS minták előkészítésért és a laboratóriumi munkák elvégzésért. Szeretném még megköszönni a kutatócsoport minden tagjának az elmúlt években kapott segítséget.

Köszönöm Nagy Jánosnak a Kaposvári Egyetem Vadgazdálkodási Tájközpont vezetőjének, hogy lehetővé tette a gímszarvas bikából történő mintavételt. Szeretném megköszönni Dr. Sugár Lászlónak a szakmai tapasztalatok megosztását. Köszönettel tartozom az Állattenyésztési Tudományok Doktori Iskola vezetőjének Dr. Szabó András Professzor úrnak, és az Egyetem rektorának Dr. Kovács Melinda Akadémikus asszonynak, hogy lehetővé tette számomra, hogy a Doktori Iskola keretein belül az érdeklődési körömnek és a képességeimnek megfelelő kutatómunkát végezhessek.

Hálás köszönetet mondok a férjemnek Madár Gábornak, kisfiamnak Madár Sámuelnek és az édesanyámnak Bana Istvánnénak, a szeretetteljes biztatásért és tanulmányaim türelmes támogatásáért.

A doktori értekezés elkészítését az EFOP-3.6.3-VEKOP-16-2017-00005 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

14 IRODALOMJEGYZÉK

Agnarsson, I., & May-Collado, L. J. (2008). The phylogeny of Cetartiodactyla: the importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. *Molecular phylogenetics and evolution*, *48*(3), 964–985. https://doi.org/10.1016/j.ympev.2008.05.046

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. doi: 10.1093/nar/25.17.3389

Andersen, R., Apollonio, M., & Putman, R. (2010). *European ungulates and their management in the 21st century*. Cambridge: Cambridge University Press.

Arias, J. A., Keehan, M., Fisher, P., Coppieters, W., & Spelman, R. (2009). A high density linkage map of the bovine genome. *BMC genetics*, *10*, 18. https://doi.org/10.1186/1471-2156-10-18

Barta, E., Bánfalvi, Zs., Havelda, Z., Hiripi, L., Jeney, Zs., Kiss, J., Kolics, B., Marincs, F., Silhavy, D., Stéger, V., & Várallyay, É. (2016) Agricultural genomics: an overview of the Next Generation Sequencing projects at the NARIC-Agricultural Biotechnology Institute in Gödöllő. *Hungarian Agricultural Research*, *25*, 10–21.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2010). GenBank. *Nucleic Acids Research*, *38*(Database issue), D46–D51. https://doi.org/10.1093/nar/gkp1024

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Biedrzycka, A., Solarz, W., & Okarma, H. (2012). Hybridization between native and introduced species of deer in Eastern Europe. *Journal of Mammalogy*, *93*(5), 1331–1341. https://doi.org/10.1644/11-MAMM-A-022.1

Bonnet, A., Thévenon, S., Claro, F., Gautier, M., & Hayes, H. (2001). Cytogenetic comparison between Vietnamese sika deer and cattle: R-banded karyotypes and FISH mapping. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *9*(8), 673–687. https://doi.org/10.1023/a:1012908508488

Borsy, A., Podani, J., Stéger, V., Balla, B., Horváth, A., Kósa, J. P., Gyurján, I., Jr, Molnár, A., Szabolcsi, Z., Szabó, L., Jakó, E., Zomborszky, Z., Nagy, J., Semsey, S., Vellai, T., Lakatos, P., & Orosz, L. (2009). Identifying novel genes involved in both deer physiological and human pathological osteoporosis. *Molecular genetics and genomics: MGG*, *281*(3), 301–313. https://doi.org/10.1007/s00438-008-0413-7

Bovine Genome Sequencing and Analysis Consortium, Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., Eichler, E. E., Elnitski, L., Guigó, R., Hamernik, D. L., Kappes, S. M., Lewin, H. A., Lynn, D. J., Nicholas, F. W., Reymond, A., Rijnkels, M., Skow, L. C., Zdobnov, E. M., … Zhao, F. Q. (2009).

The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science (New York, N.Y.)*, *324*(5926), 522–528. https://doi.org/10.1126/science.1169588

Brauning, R., Fisher, P. J., Mcculloch, A. F., Smithies, R. J., Ward, J. F., Bixley, M. J., . . . Mcewan, J. C. (2015). Utilization of high throughput genome sequencing technology for large scale single nucleotide polymorphism discovery in red deer and Canadian elk. *bioRxiv* doi: 10.1101/027318

Burbaitė, L., & Csányi, S. (2010). Red deer population and harvest changes in Europe. *Acta Zoologica Lituanica, 20*(4), 179-188. doi:10.2478/v10043-010-0038-z

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current protocols in bioinformatics*, *48*, 4.11.1–4.11.39. https://doi.org/10.1002/0471250953.bi0411s48

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Dayhoff, M. O., Schwartz, R., & Orcutt, B. C. (1979). A model of Evolutionary Change in Proteins. In *Atlas of protein sequence and structure* (3rd ed., Vol. 5, pp. 345-358). Washington, D.C.: National Biomedical Research

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic acids research*, *27*(11), 2369–2376. https://doi.org/10.1093/nar/27.11.2369

Delcher, A. L., Phillippy, A., Carlton, J., & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic*

acids research, *30*(11), 2478–2483. https://doi.org/10.1093/nar/30.11.2478

Deer Farmer (2000). Industry Manual of New Zealand. *Herald Communications* Ltd. Timaru. 2000.

Di Meo, G. P., Perucatti, A., Floriot, S., Incarnato, D., Rullo, R., Caputi Jambrenghi, A., Ferretti, L., Vonghia, G., Cribiu, E., Eggen, A., & Iannuzzi, L. (2005). Chromosome evolution and improved cytogenetic maps of the Y chromosome in cattle, zebu, river buffalo, sheep and goat. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *13*(4), 349–355. https://doi.org/10.1007/s10577-005-2688-4

Diamond, J. (2010). *Háborúk, járványok, technikák: A társadalmak fátumai*. Budapest: Typotex.

Dovichi, N. J., & Zhang, J. (2000). How Capillary Electrophoresis Sequenced the Human Genome. *Angewandte Chemie International Edition, 39*(24), 4463-4468. https://doi.org/10.1002/1521-3773(20001215)39:24<4463::AID-ANIE4463>3.0.CO;2-8

Everts-van der Wind, A., Kata, S. R., Band, M. R., Rebeiz, M., Larkin, D. M., Everts, R. E., Green, C. A., Liu, L., Natarajan, S., Goldammer, T., Lee, J. H., McKay, S., Womack, J. E., & Lewin, H. A. (2004). A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome research*, *14*(7), 1424–1437. https://doi.org/10.1101/gr.2554404

Everts-van der Wind, A., Larkin, D. M., Green, C. A., Elliott, J. S., Olmstead, C. A., Chiu, R., Schein, J. E., Marra, M. A., Womack, J. E., & Lewin, H. A. (2005). A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proceedings of the National Academy of Sciences of the*

United States of America, *102*(51), 18526–18531. https://doi.org/10.1073/pnas.0509285102

Fajardo, V., González, I., López-Calleja, I., Martín, I., Rojas, M., Hernández, P. E., García, T., & Martín, R. (2007). Identification of meats from red deer (Cervus elaphus), fallow deer (Dama dama), and roe deer (Capreolus capreolus) using polymerase chain reaction targeting specific sequences from the mitochondrial 12S rRNA gene. *Meat science*, *76*(2), 234–240. https://doi.org/10.1016/j.meatsci.2006.11.004

Feulner, P. G., Bielfeldt, W., Zachos, F. E., Bradvarovic, J., Eckert, I., & Hartl, G. B. (2004). Mitochondrial DNA and microsatellite analyses of the genetic status of the presumed subspecies Cervus elaphus montanus (Carpathian red deer). *Heredity*, *93*(3), 299–306. https://doi.org/10.1038/sj.hdy.6800504

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, *269*(5223), 496–512. https://doi.org/10.1126/science.7542800

Fontana, F., & Rubini, M. (1990). Chromosomal evolution in Cervidae. *Bio Systems*, *24*(2), 157–174. https://doi.org/10.1016/0303-2647(90)90008-O

Fredga K. (1971). Idiogram and fluorescence pattern of the chromosomes of the Indian muntjac. *Hereditas*, *68*(2), 332–337. https://doi.org/10.1111/j.1601-5223.1971.tb02411.x

Frohlich, J., Kubickova, S., Musilova, P., Cernohorska, H., Muskova, H., Vodicka, R., & Rubes, J. (2017). Karyotype relationships among

selected deer species and cattle revealed by bovine FISH probes. *PloS one*, *12*(11), e0187559. https://doi.org/10.1371/journal.pone.0187559

Gene Ontology Consortium (2006). The Gene Ontology (GO) project in 2006. *Nucleic acids research*, *34*(Database issue), D322–D326. https://doi.org/10.1093/nar/gkj021

Gilbert, C., Ropiquet, A., & Hassanin, A. (2006). Mitochondrial and nuclear phylogenies of Cervidae (Mammalia, Ruminantia): Systematics, morphology, and biogeography. *Molecular phylogenetics and evolution*, *40*(1), 101–117. https://doi.org/10.1016/j.ympev.2006.02.017

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. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1513–1518. https://doi.org/10.1073/pnas.1017351108

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic acids*

research, *36*(Database issue), D154–D158. https://doi.org/10.1093/nar/gkm952

Groves, C. P., Grubb, P. (1987) Relationships of living deer, in: *Biology and Management of the Cervidae*, C.M. Wemmer (Ed.) (Smithsonian Institution Press, Washington), pp. 21- 59.

Gu, L., Mo, E., Sun, B., Sung*, C., Jia, X., Zhu, X., & Fang, Z. (2008). Analysis of gene expression in four parts of the red-deer antler using DNA chip microarray technology. *Animal Biology*, *58*(1), 67-90. https://doi.org/10.1163/157075608X303654

Gustavsson, I., & Sundt, C. O. (1968). Karyotypes in five species of deer (Alces alces L., Capreolus capreolus L., Cervus elaphus L., Cervus nippon nippon Temm. and Dama dama L.). *Hereditas*, *60*(1), 233–248. https://doi.org/10.1111/j.1601-5223.1968.tb02204.x

Gyurján, I., Jr, Molnár, A., Borsy, A., Stéger, V., Hackler, L., Jr, Zomborszky, Z., Papp, P., Duda, E., Deák, F., Lakatos, P., Puskás, L. G., & Orosz, L. (2007). Gene expression dynamics in deer antler: mesenchymal differentiation toward chondrogenesis. *Molecular genetics and genomics: MGG*, *277*(3), 221–235. https://doi.org/10.1007/s00438-006-0190-0

d'Huy, J. (2011). La distribution des animaux à Lascaux reflèterait leur distribution naturelle. *Bulletin de la Société Historique et Archéologique du Périgord CXXXVIII*, 493–502.

Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*, *8*, 299–309.

Harris, R.S. (2007). *Improved pairwise alignment of genomic DNA*. (Ph.D. Thesis) The Pennsylvania State University.

Hartl, G. B., Zachos, F., & Nadlinger, K. (2003). Genetic diversity in European red deer (Cervus elaphus L.): anthropogenic influences on natural populations. *Comptes rendus biologies, 326 Suppl 1,* S37–S42. https://doi.org/10.1016/s1631-0691(03)00025-8

Hassanin, A., & Douzery, E. J. (2003). Molecular and morphological phylogenies of ruminantia and the alternative position of the moschidae. *Systematic biology*, *52*(2), 206–228. https://doi.org/10.1080/10635150390192726

Heffelfinger, J. (2006). *Deer of the Southwest: A complete guide to the natural history, biology, and management of southwestern mule deer and white-tailed deer* (1st ed., pp. 1-57). College Station: Texas A & M University Press., ISBN 978-1-58544-515-8

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hood, L., & Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. *Genome medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

Horn, P. (1998). *Cross sectional CT and MR anatomy atlas of red deer*. Kaposvar: Diagnostics Centre, Pannon Agricultural University.

Horn, P. (2004). A gímszarvastenyésztés mint új állattenyésztési ágazat- Az elsőháziasított nagytestű emlős faj ötezer év óta. *Magyar Tudomány*, 110(4):453-460.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

International Sheep Genomics Consortium, Archibald, A. L., Cockett, N. E., Dalrymple, B. P., Faraut, T., Kijas, J. W., Maddox, J. F., McEwan, J. C., Hutton Oddy, V., Raadsma, H. W., Wade, C., Wang, J., Wang, W., & Xun, X. (2010). The sheep genome reference sequence: a work in progress. *Animal genetics*, *41*(5), 449–453. https://doi.org/10.1111/j.1365-2052.2010.02100.x

Jia, B. Y., Ba, H. X., Wang, G. W., Yang, Y., Cui, X. Z., Peng, Y. H., Zheng, J. J., Xing, X. M., & Yang, F. H. (2016). Transcriptome analysis of sika deer in China. *Molecular genetics and genomics: MGG*, *291*(5), 1941–1953. https://doi.org/10.1007/s00438-016-1231-y

Johnston, S. E., Huisman, J., Ellis, P. A., & Pemberton, J. M. (2017). A High-Density Linkage Map Reveals Sexual Dimorphism in Recombination Landscapes in Red Deer (*Cervus elaphus*). *G3 (Bethesda, Md.)*, *7*(8), 2859–2870. https://doi.org/10.1534/g3.117.044198

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kierdorf, U., Li, C., & Price, J. S. (2009). Improbable appendages: Deer antler renewal as a unique case of mammalian regeneration. *Seminars in cell & developmental biology*, *20*(5), 535–542. https://doi.org/10.1016/j.semcdb.2008.11.011

Koepfli, K. P., Paten, B., Genome 10K Community of Scientists, & O'Brien, S. J. (2015). The Genome 10K Project: a way forward. *Annual review*

of animal biosciences, *3*, 57–111. https://doi.org/10.1146/annurev-animal-090414-014900

Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Korf, I., Bedell, J., & Yandell, M. (2003). *BLAST:*. Beijing: O'Reilly.

Kosambi, D. D. (1943). The Estimation Of Map Distances From Recombination Values. *Annals of Eugenics*, *12*(1), 172-175. doi:10.1111/j.1469-1809.1943.tb02321.x

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, *42*(Database issue), D68–D73. https://doi.org/10.1093/nar/gkt1181

Kulemzina, A. I., Trifonov, V. A., Perelman, P. L., Rubtsova, N. V., Volobuev, V., Ferguson-Smith, M. A., Stanyon, R., Yang, F., & Graphodatsky, A. S. (2009). Cross-species chromosome painting in Cetartiodactyla: reconstructing the karyotype evolution in key phylogenetic lineages. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *17*(3), 419–436. https://doi.org/10.1007/s10577-009-9032-3

Kulemzina, A. I., Yang, F., Trifonov, V. A., Ryder, O. A., Ferguson-Smith, M. A., & Graphodatsky, A. S. (2011). Chromosome painting in Tragulidae facilitates the reconstruction of Ruminantia ancestral karyotype. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *19*(4), 531–539. https://doi.org/10.1007/s10577-011-9201-z

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz,

C., Lee, C., Leinonen, R., Lin, Q., … Apweiler, R. (2007). EMBL Nucleotide Sequence Database in 2006. *Nucleic acids research*, *35*(Database issue), D16–D20. https://doi.org/10.1093/nar/gkl913

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, *5*(2), R12. https://doi.org/10.1186/gb-2004-5-2-r12

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., … International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

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., 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., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., … Wang, J. (2010). The sequence and de novo assembly of the giant panda genome. *Nature*, *463*(7279), 311–317. https://doi.org/10.1038/nature08696

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science (New York, N.Y.)*, *227*(4693), 1435–1441. https://doi.org/10.1126/science.2983426

Love, M. I., Hogenesch, J. B., & Irizarry, R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology*, *34*(12), 1287–1291. https://doi.org/10.1038/nbt.3682

Ma, R. Z., Beever, J. E., Da, Y., Green, C. A., Russ, I., Park, C., Heyen, D. W., Everts, R. E., Fisher, S. R., Overton, K. M., Teale, A. J., Kemp, S. J., Hines, H. C., Guérin, G., & Lewin, H. A. (1996). A male linkage map of the cattle (Bos taurus) genome. *The Journal of heredity*, *87*(4), 261–271. https://doi.org/10.1093/oxfordjournals.jhered.a022999

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., & Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids*

research, *34*(Database issue), D108–D110. https://doi.org/10.1093/nar/gkj143

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564. https://doi.org/10.1073/pnas.74.2.560

Mcdevitt, A. D., Edwards, C. J., O'Toole, P., O'Sullivan, P., O'Reilly, C., & Carden, R. F. (2009). Genetic structure of, and hybridisation between, red (Cervus elaphus) and sika (Cervus nippon) deer in Ireland. *Mammalian Biology*, *74*(4), 263-273. doi: 10.1016/j.mambio.2009.03.015

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, *2*, e675. https://doi.org/10.7717/peerj.675

Milner, J. M., Bonenfant, C., Mysterud, A., Gaillard, J., Csányi, S., & Stenseth, N. C. (2006). Temporal and spatial development of red deer harvesting in Europe: Biological and cultural factors. *Journal of Applied Ecology*, *43*(4), 721-734. doi: 10.1111/j.1365-2664.2006.01183.x

Molnár, A., Gyurján, I., Korpos, E., Borsy, A., Stéger, V., Buzás, Z., Kiss, I., Zomborszky, Z., Papp, P., Deák, F., & Orosz, L. (2007). Identification of differentially expressed genes in the developing antler of red deer Cervus elaphus. *Molecular genetics and genomics: MGG*, *277*(3), 237–248. https://doi.org/10.1007/s00438-006-0193-x

Molnár, J., Nagy, T., Stéger, V., Tóth, G., Marincs, F., & Barta, E. (2014). Genome sequencing and analysis of Mangalica, a fatty local pig of Hungary. *BMC genomics*, *15*(1), 761. https://doi.org/10.1186/1471-2164-15-761

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics (Oxford, England)*, *24*(16), 1757–1764. https://doi.org/10.1093/bioinformatics/btn322

Neitzel, H. (1987). Chromosome Evolution of Cervidae: Karyotypic and Molecular Aspects. *Cytogenetics,* 90-112.

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Analytical chemistry*, *83*(12), 4327–4341. https://doi.org/10.1021/ac2010857

Olivieri, C., Marota, I., Rizzi, E., Ermini, L., Fusco, L., Pietrelli, A., De Bellis, G., Rollo, F., & Luciani, S. (2014). Positioning the red deer (Cervus elaphus) hunted by the Tyrolean Iceman into a mitochondrial DNA phylogeny. *PloS one*, *9*(7), e100136. https://doi.org/10.1371/journal.pone.0100136

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, *16*(6), 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2

Rubini, M., & Fontana, F. (1988). Standard G-banded karyotype, constitutive heterochromatin and nucleolus organizer regions in the roe deer (Capreolus capreolus L.). *Genetica*, *77*(2), 143-148. doi:10.1007/BF00057765

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, *32*(Database issue), D91–D94. https://doi.org/10.1093/nar/gkh012

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, *94*(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic acids research*, *33*(Web Server issue), W686–W689. https://doi.org/10.1093/nar/gki366

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., & Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome research*, *13*(1), 103–107. https://doi.org/10.1101/gr.809403

Seabury, C. M., Bhattarai, E. K., Taylor, J. F., Viswanathan, G. G., Cooper, S. M., Davis, D. S., Dowd, S. E., Lockwood, M. L., & Seabury, P. M. (2011). Genome-wide polymorphism and comparative analyses in the white-tailed deer (Odocoileus virginianus): a model for conservation genomics. *PloS one*, *6*(1), e15811. https://doi.org/10.1371/journal.pone.0015811

Slate, J., Van Stijn, T. C., Anderson, R. M., McEwan, K. M., Maqbool, N. J., Mathias, H. C., Bixley, M. J., Stevens, D. R., Molenaar, A. J., Beever,

J. E., Galloway, S. M., & Tate, M. L. (2002). A deer (subfamily Cervinae) genetic linkage map and the evolution of ruminant genomes. *Genetics*, *160*(4), 1587–1597.

Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, *6*, 31. https://doi.org/10.1186/1471-2105-6-31

Smit, A. F. A., Hubley, R., & Green P. (2013-2015). RepeatMasker Open-4.0. (http://www.repeatmasker.org/)

Sokolov, V. E., Orlow, V. N., Chudinovskaja G. A., & Danilkin A. A. (1978). Differences in chromosomes between two subspecies Capreolus capreolus capreolus L. and C.c. pygargus. *Pall. Zool. Zh*., *57*, 1109–1112.

Sonkoly, K., Bleier, N., Heltai, M., Katona, K., Szemethy, L., Szabó, L., Beregi, A., & Csányi, S. (2013). Big game meat production in Hungary: A special product of a niche market. *Hungarian Agricultural Research*, *22*(2): 12–16.

Stanke, M., Tzvetkova, A., & Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome biology*, *7 Suppl 1*(Suppl 1), S11.1–S11.8. https://doi.org/10.1186/gb-2006-7-s1-s11

Stéger, V., Molnár, A., Borsy, A., Gyurján, I., Szabolcsi, Z., Dancs, G., Molnár, J., Papp, P., Nagy, J., Puskás, L., Barta, E., Zomborszky, Z., Horn, P., Podani, J., Semsey, S., Lakatos, P., & Orosz, L. (2010). Antler development and coupled osteoporosis in the skeleton of red deer Cervus elaphus: expression dynamics for regulatory and effector genes. *Molecular genetics and genomics: MGG*, *284*(4), 273–287. https://doi.org/10.1007/s00438-010-0565-0

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology, 14*(1), 43-59. doi:10.1002/jez.1400140104

Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T., & Tateno, Y. (2008). DDBJ with new system and face. *Nucleic acids research*, *36*(Database issue), D22–D24. https://doi.org/10.1093/nar/gkm889

Szabolcsi, Z., Egyed, B., Zenke, P., Borsy, A., Pádár, Z., Zöldág, L., . . . Orosz, L. (2008). Genetic identification of red deer using autosomal STR markers. *Forensic Science International: Genetics Supplement Series, 1*(1), 623-624. doi: 10.1016/j.fsigss.2007.10.003

Szabolcsi, Z., Egyed, B., Zenke, P., Padar, Z., Borsy, A., Steger, V., Pasztor, E., Csanyi, S., Buzas, Z., & Orosz, L. (2014). Constructing STR multiplexes for individual identification of Hungarian red deer. *Journal of forensic sciences*, *59*(4), 1090–1099. https://doi.org/10.1111/1556-4029.12403

Swindell, S. R., & Plasterer, T. N. (1997). Seqman. *Sequence Data Analysis Guidebook,* 75-89. doi:10.1385/0-89603-358-9:75

Tate, M. L., Mathias, H. C., Fennessy, P. F., Dodds, K. G., Penty, J. M., & Hill, D. F. (1995). A new gene mapping resource: interspecies hybrids between Père David's deer (Elaphurus davidianus) and red deer (Cervus elaphus). *Genetics*, *139*(3), 1383–1391.

Todd, N. B. (1975). Chromosomal Mechanisms in the Evolution of Artiodactyls. *Paleobiology, 1*(2), 175-188. doi:10.1017/S0094837300002360

Todd, N.B. (2000). Kinetochore reproduction underlies kariotypic fission theory: Possible legacy of symbiogenesis in mammalian chromosome evolution. *Symbiosis, 29:*319-327.

Tsipouri, V., Schueler, M. G., Hu, S., NISC Comparative Sequencing Program, Dutra, A., Pak, E., Riethman, H., & Green, E. D. (2008). Comparative sequence analyses reveal sites of ancestral chromosomal fusions in the Indian muntjac genome. *Genome biology*, *9*(10), R155. https://doi.org/10.1186/gb-2008-9-10-r155

UniProt Consortium (2007). The Universal Protein Resource (UniProt). *Nucleic acids research*, *35*(Database issue), D193–D197. https://doi.org/10.1093/nar/gkl929

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), 180. https://doi.org/10.1186/1471-2164-15-180

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Y. X., Ma, S. L., & Li, C. Y. (1993). The taxonomy, distribution and status of forest musk deer in China, in Ohtaishi N, Sheng HL (eds). *Deer of China: Biology and Management*, pp 22–30 (Elsevier Science, Tokyo).

Warren, W. C., Hillier, L. W., Marshall Graves, J. A., Birney, E., Ponting, C. P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A. T., Yang, S. P., Heger, A., Locke, D. P., Miethke, P., Waters, P. D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., … Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, *453*(7192), 175–183. https://doi.org/10.1038/nature06936

Yao, B., Zhao, Y., Zhang, H., Zhang, M., Liu, M., Liu, H., & Li, J. (2012). Sequencing and de novo analysis of the Chinese Sika deer antler-tip transcriptome during the ossification stage using Illumina RNA-Seq technology. *Biotechnology letters*, *34*(5), 813–822. https://doi.org/10.1007/s10529-011-0841-z

Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V., & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic acids research*, *26*(17), 3986–3990. https://doi.org/10.1093/nar/26.17.3986

Zsolnai, A., Lehoczky, I., Gyurmán, A., Nagy, J., Sugár, L., Anton, I., . . . Magyary, I. (2009). Development of eight-plex microsatellite PCR for parentage control in deer. *Archives Animal Breeding, 52*(2), 143-149. doi: 10.5194/aab-52-143-2009

Felhasznált weboldalak jegyzéke

https://annovar.openbioinformatics.org/en/latest/user-guide/startup/.
https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate.
http://elte.prompt.hu/sites/default/files/tananyagok/ABiokemiaEsMolekulariBiologiaAlapjai/ch19s05.html.
https://en.wikipedia.org/wiki/DNA_sequencing.
https://en.wikipedia.org/wiki/Whole_genome_sequencing.
https://github.com/lh3/samtools/blob/master/bcftools/vcfutils.pl.
https://github.com/tseemann/barrnap, v06., Seemann T. barrnap.
http://gregoryzynda.com/.
https://mek.oszk.hu/03400/03408/html/200.html.- Brehm A. (1863-69) Thierleben (Az állatok világa) alapján Digitális kiadás: Arcanum Adatbázis Kft. 2000
https://github.com/trinityrnaseq/trinityrnaseq.
https://www.eurofinsgenomics.co.in/en/eurofins-genomics/product-faqs/next-generation-sequencing/general-technical-questions/what-is-the-principal-of-the-illumina-sequencing-technology.aspX.
https://www.ncbi.nlm.nih.gov/genome/?term=txid9913[orgn].
https://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011_0025_bio_4/ch35.html.
https://www.tankonyvtar.hu/hu/tartalom/tamop425/2011_0001_521_Vadaszati-allattan/ch04s22.html.

15 A DISSZERTÁCIÓ TÉMAKÖRÉBŐL MEGJELENT PUBLIKÁCIÓK

Lektorált folyóiratban megjelent idegen nyelvű publikáció

Bana, N. Á., Nyiri, A., Nagy, J., Frank, K., Nagy, T., Stéger, V., Schiller, M., Lakatos, P., Sugár, L., Horn, P., Barta, E., & Orosz, L. (2018). The red deer Cervus elaphus genome CerEla1.0: sequencing, annotating, genes, and chromosomes. *Molecular genetics and genomics: MGG*, *293*(3), 665–684. https://doi.org/10.1007/s00438-017-1412-3

Frank, K., Barta, E., Bana, N. Á., Nagy, J., Horn, P., Orosz, L., & Stéger, V. (2016). Complete mitochondrial genome sequence of a Hungarian red deer (Cervus elaphus hippelaphus) from high-throughput sequencing data and its phylogenetic position within the family Cervidae. *Acta Biologica Hungarica*, *67*(2), 133–147. https://doi.org/10.1556/018.67.2016.2.2

Lektorált folyóiratban megjelent magyar nyelvű publikáció

Bana, Á. N. (2020). A genom összerakás elmélete és alkalmazása a gímszarvas genom projektben. *Acta Agraria Kaposváriensis*, 24(1), 14- 34. https://doi.org/10.31914/aak.2370

Nem lektorált folyóiratban megjelent magyar nyelvű publikáció

Orosz, L., & Bana, N. Á. (2019). Mire jó a szarvasgenom? *Élet és Tudomány, 74*(16), 498-500.

Konferenciakiadványban teljes terjedelemben idegen nyelven megjelent

Bana, Á. N., Nyiri, A., Nagy, J., Frank, K., Nagy, T., Stéger, V., Schiller, M., Lakatos P., Sugár L., Horn P., Barta E., & Orosz L. (2018, August 5-10). *The Red Deer Cervus elaphus Reference Genome CerEla1.0*. 9th International Deer Biology Congress, Estes Park, Colorado, USA, https://static.sched.com/hosted_files/idbc2018/3e/459317.pdf

Frank K., Barta, E., Bana, N.Á., Nagy, J., Horn, P., & Orosz, L., & Stéger, V., (2016, március 21-22). *Complete mitochondrial genome of the hungarian red deer* (*Cervus elaphus hippelaphus*). Fiatal Biotechnológusok Országos Konferenciája 2016, Szent István Egyetem, Gödöllő, oldalszám: 69. ISBN 978-963-269-536-5.

16 A DISSZERTÁCIÓ TÉMAKÖRÉN KÍVÜL PUBLIKÁCIÓK

Lektorált folyóiratban megjelent idegen nyelvű publikáció

Frank, K., Bana, N. Á, Bleier, N., Sugár, L., Nagy, J., Wilhelm, J., & Stéger, V. (2020). Mining the red deer genome (CerEla1.0) to develop X-and Y-chromosome-linked STR markers. *Plos One, 15*(11). doi:10.1371/journal.pone.0242506

Konferenciakiadványban teljes terjedelemben idegen nyelven megjelent

Bana, N.Á., Frank K., Nyiri, A., Barta, E., Nagy, J., Horn, P., Stéger, V., & Orosz, L. (2016. március 21-22). *Bioinformatic analysis of promoter sequences of the red deer (Cervus elaphus hippelaphus) bone and antler metabolism genes.* Fiatal Biotechnológusok Országos Konferenciája 2016, Szent István Egyetem, Gödöllő, ISBN 978-963-269-536-5.

Frank, K., Stéger, V., Nagy, T., Bana Á.N., Nagy, J., Szabolcsi, Z., Wilhelm, J., Kálmán, Zs., Barta, E., Horn, P., & Orosz, L. (2015, március 27-29). *Microsatellite markers developed for red deer using genome sequence data*. Hungarian Molecular Life Sciences 2015, Hotel Eger-Park, Eger, oldalszám: 190. ISBN 978-615-5270-15-4.

Németh, A., Frank, K., Bana, Á. N., Molnár, J., Tóth, G., Nagy, T., Barta, E., Marincs, F., Bodó, Sz., & Stéger, V. (2014, május 23-24). *Development of Mangalica breed-specific DNA marker*. 20th Youth Scientific Forum, Pannon Egyetem, Keszthely, pp. 57-67. ISBN 978-963-9639-57-7.

Konferenciakiadványban teljes terjedelemben magyar nyelven megjelent

Frank, K., Stéger, V., Bana Á.N., Nagy, T., Nagy, J., Wilhelm, J., Kálmán, Zs., Barta, E., Horn, P., Orosz, L. (2015, május 21). *Gímszarvas mikroszatellita marker fejlesztés újgenerációs szekvenálási adatok segítségével*. XXI. Ifjúsági Tudományos Fórum, Pannon Egyetem, Keszthely, ISBN 978-963-9639-78-2.

17 RÖVID SZAKMAI ÉLETRAJZ

1982. november 3.-án születtem Debrecenben. A gyermekkoromat Tiszacsegén töltöttem. Középiskolai tanulmányaimat a debreceni Tóth Árpád Gimnáziumban biológia tagozatos hallgatóként végeztem, ahol 2001-ben sikeres érettségi vizsgát tettem.

Az érettségi megszerzése után a Debreceni Egyetem Természettudományi Karának nappali tagozatán biológus szakon, ökológus szakirányon kezdtem felsőfokú tanulmányaimat, és 2007-ben biológus MSc oklevelet szereztem.

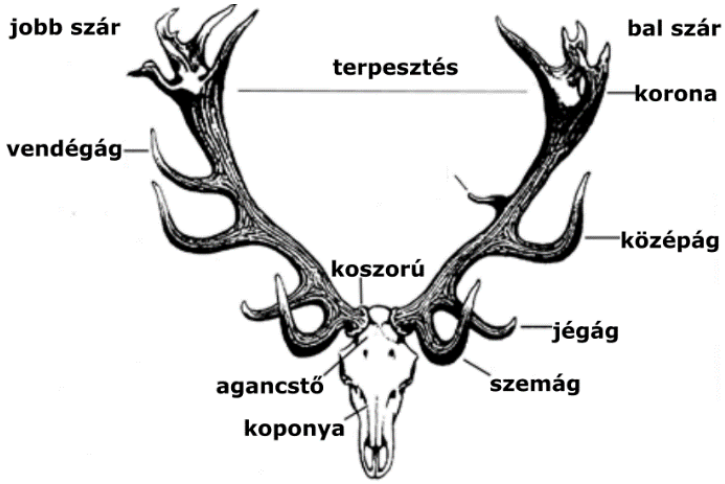
Közben 2006-ban sikeresen felvételiztem a Debreceni Egyetem Természettudományi Karára molekuláris biológus szakra genetikus szakirányra. Ezen a szakon 2011-ben szereztem meg a molekuláris biológus mesterszakos diplomámat.

Az MSc. egyetemi tanulmányaim befejezése után 2012-ben felvételt nyertem a Kaposvári Egyetem Állattenyésztési-tudományok Doktori Iskolájába, amelynek nappali tagozatos hallgatója voltam. Doktoranduszként, és tudományos segédmunkatársként a kutatómunkámat Gödöllőn a Nemzeti Agrárkutatási és Innovációs Központ Mezőgazdasági Biotechnológiai Kutatóintézetében végeztem egészen 2018 márciusáig. 2018 és 2019 között a budapesti Eötvös Lóránd Tudományegyetem Genetikai Tanszékén dolgoztam, mint oktató és bioinformatikus. Témavezetőként egy külföldi MSc diák diplomamunkáját irányítottam.

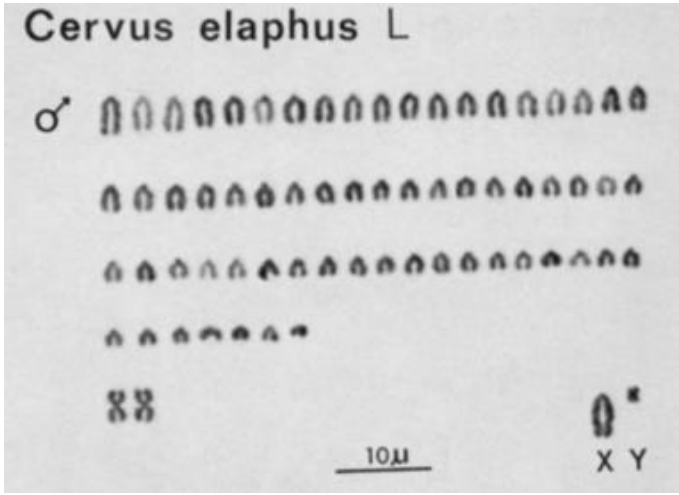
Jelenleg az másfél éves kisfiammal vagyok otthon.

18 MELLÉKLETEK

1. melléklet. A gímszarvas agancsa (forrás: Nimród, 1996, https://www.tankonyvtar.hu/hu/tartalom/tamop425/2011_0001_521_Vadaszati-allattan/ch04s22.html).

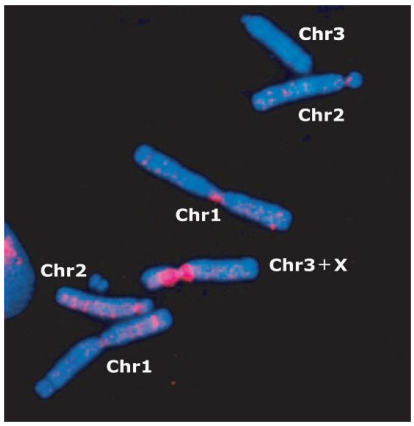


2. melléklet. *Cervus elaphus* szarvasbika kariotípusa a mitózis metafázisában. (2n =68, N.F.=70), (Gustavsson és Sundt 1968).



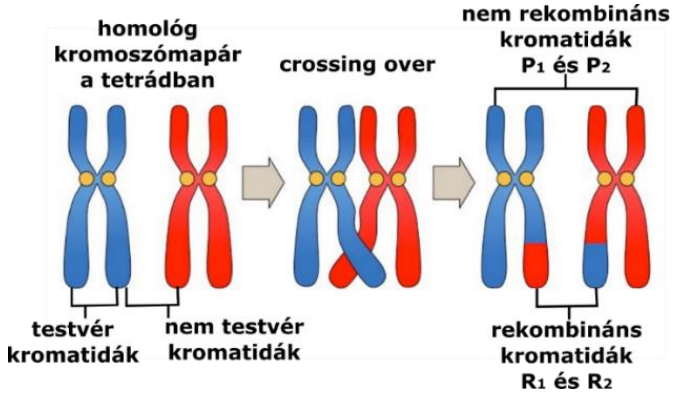
Magyarázat: Valamennyi végállású kromoszómát telocentrikus (T) kromoszómának neveztek egykor, de a finomabb felbontású technológiákkal észlelhető a p kar is, ezért átnevezték ezeket akrocentrikusnak (A).

3. melléklet. Indiai muntyákszarvas kromoszómái in situ hibridizálása.

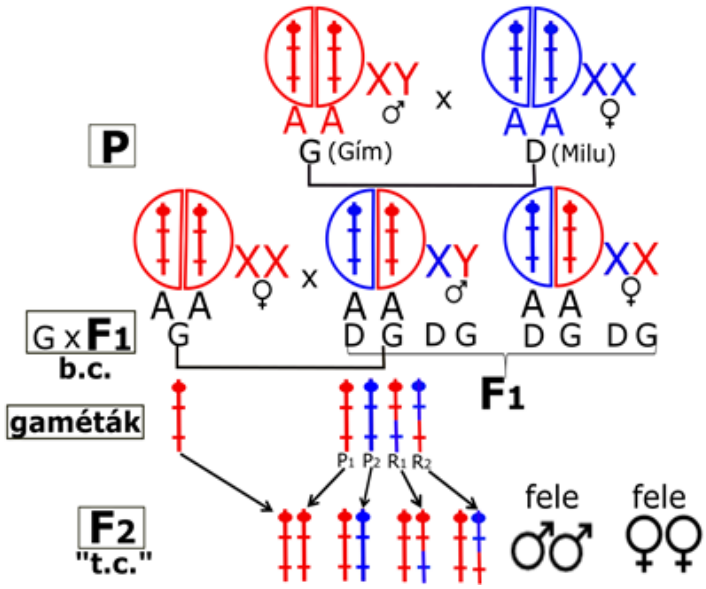


Magyarázat: A hibridizálás BAC klónokkal, amelyeket fluorescensen jelöltek (FISH technológia). A kromoszómák DAPI kékkel vannak megfestve, narancs festék különböző spektrumai jelzik a hibridizációs (transzlokációs) helyeket Tsipouri és mtsai., 2008).

4. melléklet. Crossing over (forrás: https://sites.google.com/a/wyckoffschools.org/cells-and-heredity/home/chapter-5-1/crossing-over).

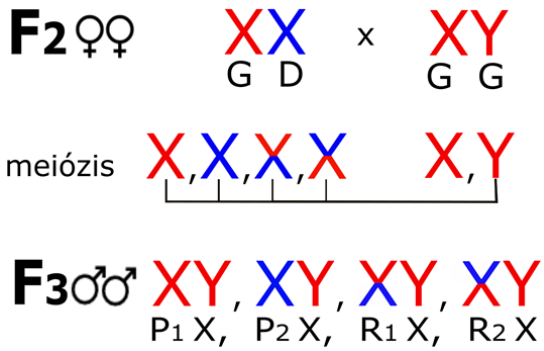


5. melléklet. Interspecifikus back-cross keresztezések.



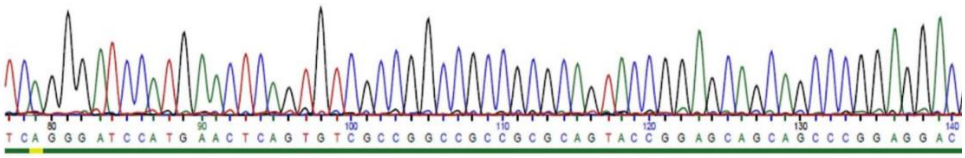
Magyarázat: A: autoszóma, G: gímszarvas, D: Dávid szarvas, P: parentális nemzedék, F_1: 1. utódnemzedék, b.c.: backcross keresztezés, F_2 utódnemzedék, t.c.: teszt-cross keresztezés. X: X kromoszóma, Y: Y kromoszóma.

6. melléklet. Az X kromoszóma géntérképezése, criss-cross szabály alapján.



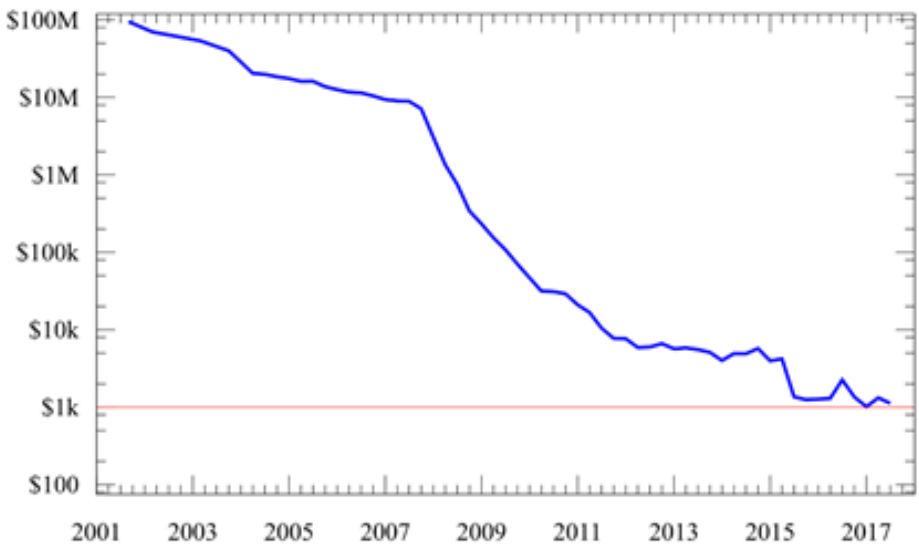
Magyarázat: G: gímszarvas, D: Dávid szarvas, P: parentális nemzedék, F_2: 2. utódnemzedék, F_3: 3. utódnemzedék, X: X kromoszóma, Y: Y kromoszóma.

7. melléklet. Szekvenáló kromatogram (szekvenogram).



Magyarázat: Azt mutatja meg, hogy adott időpillanatban mely fluorofór haladt át a detektor előtt. Ezáltal leolvashatóvá válik a templát DNS bázissorrendje. Négy szín reprezentálja a különböző nukleotidokat: piros = T, zöld = A, sárga = G, kék = C (http://elte.prompt.hu/sites/default/files/tananyagok/ABiokemiaEsMolekularisBiologiaAlapjai/ch19s05.html).

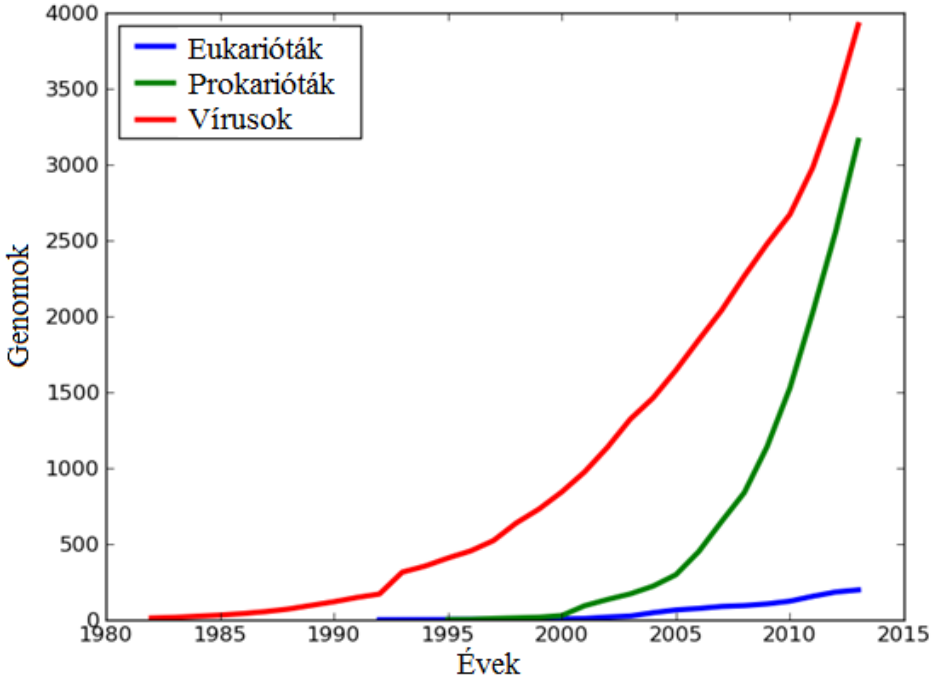
8. melléklet. A teljes humán genom totális szekvenálási költségeinek évenkénti alakulása NHGRI adatok alapján (https://en.wikipedia.org/wiki/Whole_genome_sequencing).



9. melléklet. Nagy áteresztőképességű módszerek csoportosítása (https://en.wikipedia.org/wiki/DNA_sequencing).

Módszer	PCR	Szekvenálás	Read hossz	Pontosság	Read/futás	Futási idő	1 Mbp költsége ($)
Piroszekvenálás (454)	Emulziós	Szintézissel	700 bp	99,90%	1 millió	24 óra	$10
Illumina	Híd-amplifikáció	Szintézissel	MiniSeq:75-300 bp NextSeq:75-300 bp MiSeq: 50-600 bp HiSeq2500: 50-500 bp HiSeq3/4000: 50-300 bp HiSeq X: 300 bp	99,90%	MiniSeq:1-25 millió MiSeq:1-25 millió NextSeq:130 millió HiSeq2500:300 millió-2 milliárd HiSeq3/4000:2,5 milliárd HiSeqX:3 milliárd	1-11 nap	$0,05-$0,15
SOLiD	Emulziós	Ligálással	50+35 bp vagy 50+50 bp	99,90%	1,2 to 1,4 milliárd	1-2 hét	$0,13
Pacific Biosciences	Egymolekulás kamrában	Szintézissel	30000 bp	87%	10-20 Gbp	30 perc -20 óra	$0,05-$0,08
DNS nanolabda	Rekombináns	Szintézissel	BGISEQ-50:35-50 bp MGISEQ-200:50-200 bp BGISEQ-500:50-300 bp MGISEQ-2000:50-300 bp	99,90%	BGISEQ-50:160 Mbp MGISEQ-200:300 Mbp BGISEQ-500:1300 Mbp/cella MGISEQ-2000:375 Mbp/cella	1-9 nap	$0,035-$0,12
Ion Torrent	Emulziós	Félvezető	600 bp fölött	99,60%	80 millió fölött	2 óra	$1
Nanoporus	Nincs	Elektromos-potenciál	500 kbp fölött	92-97% single read	Read hosszúságtól függ	1 perc-48 óra	$500-999 cellánként

10. melléklet. Az elmúlt 3 évtizedben szekvenált és NCBI-ba feltöltött teljes genom szekvenciák száma (Greg Zynda-tól átvéve, http://gregoryzynda.com/).



11/a. melléklet. A legjelentősebb házi- és haszonállatok legújabb, kromoszómákba rendezett teljes referencia genom szekvencia adatai (forrás: NCBI)

Név	Latin név	Fajta	Nem	Kromoszóma szám (n)	Szekvenálási technika	Legfrissebb verzió	Genom lefedettség	Genom hossz (Gbp)	Fehérje kódoló gének
Házi macska	*Felis catus*	Abesszin	♀	18+X+MT	PacBio; 454 Titanium; Illumina; Sanger	Felis_catus_9.0 (2017)	72,0x	2,52	19748
Kutya	*Canis lupus familiaris*	Boxer	♀	38+X+MT	Sanger	CanFam3.1 (2011)	7,0x	2,41	20039
Szarvasmarha	*Bos taurus*	Hereford	♀	29+X+MT	PacBio; Illumina NextSeq 500/HiSeq/Gall	ARS-UCD1.2 (2018)	80,0x	2,72	21039
Juh	*Ovis aries*	Rambouillet	♀	26+X+MT	HiSeq X Ten; PacBio RS II	Oar_rambouillet_v1.0 (2017)	126,0x	2,87	21160
Ló	*Equus caballus*	Telivér	♀	31+X+MT	Sanger; Illumina HiSeq; PacBio	EquCab3.0 (2018)	88,0x	2,51	21129

11/b. melléklet folytatása. A legjelentősebb házi- és haszonállatok legújabb, kromoszómákba rendezett teljes referencia genom szekvencia adatai (forrás: NCBI)

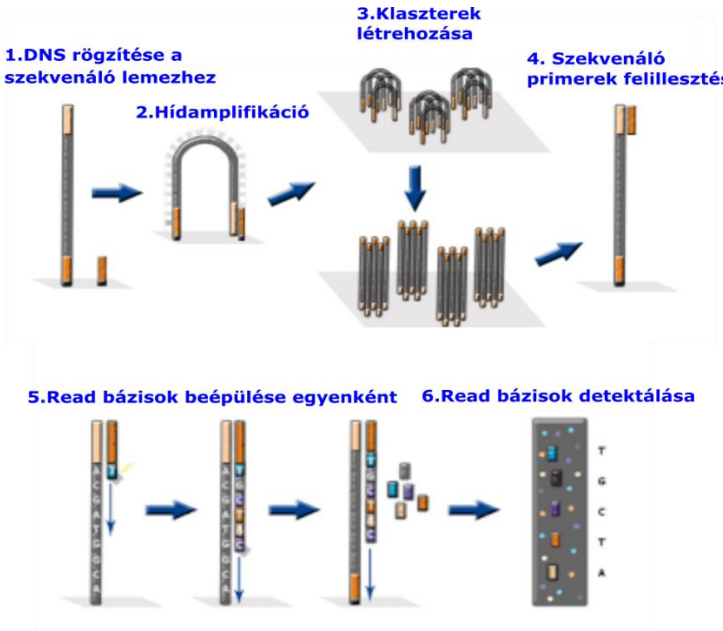
Név	**Latin név**	**Fajta**	**Nem**	**Kromoszóma szám (n)**	**Szekvenálási technika**	**Legfrissebb verzió**	**Genom lefedettség**	**Genom hossz (Gbp)**	**Fehérje kódoló gének**
Sertés	*Sus scrofa*	Duroc	♀	18+X/Y+MT	PacBio	Sscrofa11.1 (2017)	65,0x	2,50	2079
Nyúl	*Oryctolagus cuniculus*	Thorbecke	♀	21+X+MT	ABI	OryCun2.0 (2009)	7,48x	2,74	20547
Baromfi	*Gallus gallus*	Bankivatyúk (UCD001)	♀	33+W/Z+MT	Pacific Biosciences RSII	GRCg6a (2018)	82,0x	1,07	17477
Ponty	*Cyprinus carpio*	-	-	50+MT	-	common carp genome (2014)	-	1,71	49579
Méh	*Apis mellifera*	(DH4)	♂	16+MT	PacBio; 10X Chromium; Bionano	Amel_HAv3.1 (2018)	192,0x	0,23	9935

12. melléklet. BLAST algoritmus típusok.

Kereső-szekvencia query	**Adatbázis subject**	**BLAST program**
fehérje	fehérje	***blastp***
nukleotid	nukleotid	***blastn***
6 keretben lefordított nukleotid	fehérje	***blastx***
fehérje	6 keretben lefordított nukleotid	***tblastn***
6 keretben lefordított nukleotid	6 keretben lefordított nukleotid	***tblastx***

13. melléklet. Illumina szekvenálás menete.

https://www.eurofinsgenomics.co.in/en/eurofins-genomics/product-faqs/next-generation-sequencing/general-technical-questions/what-is-the-principal-of-the-illumina-sequencing-technology.aspX



14. melléklet. Az ALLPATHS-LG program egyik kimeneti fájlja, a contig-ok és a scaffold-ok statisztikai tulajdonságai.

ALLPATHS-LG ASSEMBLY PERFSTAT:	
contig minimum size for reporting [ap_report_min_contig]	1000
number of contigs [n_contigs]	437412
number of contigs per Mb [contigs_per_Mb]	128,3
number of scaffolds [n_scaffolds]	34724
total contig length [contig_length]	1952556261
total scaffold length, with gaps [scaff_length_gap]	3409397972
N50 contig size in kb [N50_contig]	7,5
N50 scaffold size in kb [N50_scaffold]	265
N50 scaffold size in kb, with gaps [N50_scaff_gap]	430
number of scaffolds per Mb [scaff_per_Mb]	10,18
median size of gaps in scaffolds [median_gap]	3560
median dev of gaps in scaffolds [median_gap_dev]	2113
% of bases in captured gaps [frac_captured_gaps]	43,56
% of bases in negative gaps (after 5 devs) [frac_negative_gaps]	1,01
%% of ambiguous bases [amb_base_frac]	122,82
ambiguities per 10,000 bases [ambiguity_frac]	14,94

15. melléklet. UCSC genom böngészőben a Table browse alkalmazásban a szarvasmarha referencia gének genomi koordinátáinak lekérése.

Select Fields from bosTau7.refGene

☐	bin	
☐	name	Name of gene (usually transcript_id from GTF)
☑	chrom	Reference sequence chromosome or scaffold
☑	strand	+ or - for strand
☑	txStart	Transcription start position (or end position for minus strand item)
☑	txEnd	Transcription end position (or start position for minus strand item)
☐	cdsStart	Coding region start (or end position for minus strand item)
☐	cdsEnd	Coding region end (or start position for minus strand item)
☐	exonCount	Number of exons
☐	exonStarts	Exon start positions (or end positions for minus strand item)
☐	exonEnds	Exon end positions (or start positions for minus strand item)
☐	score	score
☑	name2	Alternate name (e.g. gene_id from GTF)
☐	cdsStartStat	enum('none','unk','incmpl','cmpl')
☐	cdsEndStat	enum('none','unk','incmpl','cmpl')
☐	exonFrames	Exon frame {0,1,2}, or -1 if no frame for exon

get output | cancel | check all | clear all

16. melléklet. UCSC genom böngésző a Table browse alkalmazásból lekért szarvasmarha referencia gének genomi koordinátái (részlet).

#chrom	strand	txStart	txEnd	name2
chr15	-	451151	521967	MRE11
chr15	+	537280	540085	ANKRD49
chr15	-	580547	607408	AASDHPPT
chr15	+	784247	794656	MSANTD4
chr15	-	907019	1259272	GRIA4
chr15	+	2088749	2108214	CASP4
chr15	+	3216479	3461774	PDGFD
chr15	-	3315564	3317635	DDI1
chr15	+	4471133	4499828	DCUN1D5
chr15	+	4611282	4621971	MMP13
chr15	+	4699014	4709377	MMP12
chr15	+	4719640	4726226	MMP3
chr15	+	4752099	4784225	MMP1
chr15	+	4878392	4892182	MMP27
chr15	+	5052036	5109875	MMP20
chr15	+	5163173	5173344	MMP7
chr15	-	5391794	5422552	BIRC3
chr15	-	5679737	5709020	CFAP300
chr15	+	5857874	6018302	TRPC6
chr15	+	6401922	6518238	PGR
chr15	-	7227510	7806604	CNTN5

17. melléklet. Szarvasmarha referencia genom átalakítása.

Bos taurus				***Cervus elaphus***			
kr	**kezdet**	**vég**	**irány**	**evolúciós folyamat**	**LG**	**marker kezdet**	**marker vég**
1	59135672	105214781	-	hasadás, transzlokáció	19	INRA11	OaRMAF109
	105214782	158972876	-	hasadás, transzlokáció		CSSM19	C276
	1	59135671	+	hasadás	31	BM6438	RM95
2	1	93808892	+	hasadás	33	INRA40	INHBB
	93808893	137479425	-	hasadás	8	ALPL	TGLA226
5	1	70564518	+	hasadás	3	ILSTS42	OaMAF23
	70564519	121584146	-	hasadás	22	TEXAN15	ACO2
6	63599871	119927886	-	hasadás	6	PDGRFRA	PDE6B
	1	63599870	+	hasadás	17	ILSTS93	SPP1
8	1	64530926	+	hasadás	29	UWCA47	CM100
	64530927	114055384	-	hasadás	16	LPL	ORM1
9	1	64019783	+	hasadás	28	ETH225	OaRCP021
	64019784	106206484	+	hasadás	26	CM102	SOD2
10	1	104654538	-	fordulás	12	BM3033	BMS2614
11	1	107582167	-	fordulás	11	SLC8A1	ASS
17	1	75367154	-	Robertsoniális transzlokáció	5	FGG	TGLA322
19	1	64325510	+			CM332	P4HB
28	1	46481278	+	Robertsoniális fúzió	15	RT5	IRBP
26	1	51889111	+			ABS12	CYP17
29	1	51869468	-	fordulás	2	TGLA86	KRN1

18. melléklet. MMSc-k RGSc-k, IRGSc-k és GFSc-k illesztése Btau_5.0.1 genomra BWA programmal szkript.

```
#!/bin/sh

# szkript: bwa_cesc_vs_bt501_parameter.sh: Szarvas_scaffoldokat illeszt \
# az ortológ gímszarvas géntérkép szerint átalakitott szarvasmarha hard \
# maszkolt kromoszómákra (1-33, x, y)
# bed fájlokat hoz létre

#parameters: -t INT     Number of threads
# -a    Output all found alignments for single-end or unpaired paired-end \
# reads. These alignments will be flagged as secondary alignments.
#-O INT Gap open penalty
#-L INT Clipping penalty
# -E Gap extension penalty. A gap of length k costs O + k*E (i.e. -O is for
# opening a zero-length gap).

#futtatás:
#nohup ./data/szarvas/BN/Nora_mapp/Nora_scaffolds/out/\
#bwa_cesc_vs_bt501_parameter.sh > logs/bwa_parameter.log 2\
# > logs/bwa_parameter.err &

#konyvtarak utvonalakkal:
dir1="/data/szarvas/BN/Nora_mapp/Nora_scaffolds/out/"
dir2="/data/szarvas/BN/Nora_mapp/BosTau_501_db_bwa_index/chr_index"
dir3="/data/szarvas/BN/Nora_mapp/Nora_scaffolds/out/query"

cd $dir1 #belép a dir1 könyvtárba

# szarvasmarha chr-k,DE gímszarvas géntérkép szerint széttört/illesztett\
# és ortológ gím chr számmal
for i in `seq 1 33` x y ; do
echo "processing chr$i at `date`"

# indexelt szarvasmarha chr-re illeszti a gímszarvas scaffoldokat
bwa mem -t 6 -a -O 20,20 -L 100,100 $dir2/chr${i}.fasta $dir3/chr${i}.fa \
> chr${i}.sam

# sam fájlok sorba rendezése
samtools view -bS chr${i}.sam | samtools sort - chr${i}_sorted

# bam fájlok indexelése,
samtools index chr${i}_sorted.bam chr${i}_sorted.bai

# 12 oszlopos bed fájlok létrehozása a bam fájlokból
bedtools bamtobed -i chr${i}_sorted.bam -bed12 >chr${i}.bed

done
exit
```

Magyarázat: A színkód a bash program nyelv szerinti. A #-el jelölt, zöld színű komment szekciók nem futnak le, viszont magyarázzák az egyes kód részeket.

19. melléklet. MMSc-k RGSc-k, IRGSc-k és GFSc-k illesztése Btau_5.0.1 genomra MUMmer programmal szkript.

```
#!/bin/sh

# szkript: run_mummer_m_bt501_vs_ce_scaff.sh: Szarvas_scaffoldokat illeszt\
# az ortolog gímszarvas géntérkép szerint átalakitott szarvasmarha\
# hard maszkolt kromoszómákra (1-33, x, y)
# coords fájlokat hoz létre

#Parameters:
#--mum  Use anchor matches that are unique in both the reference and query
#-c int Minimum cluster length (default 65)
#--coords   Automatically generate the <prefix>.coords file using the \
#'show-coords' program with the -r option

#futtatás:
#nohup ./run_mummer_m_bt501_vs_ce_scaff.sh \
#> logs/mummer_bt501_vs_ce_scaff.log 2> \
#logs/mummer_bt501_vs_ce_scaff.err &

#Munka könyvtár útvonallal
base='/home/norah007/mummer_bt_ce'

cd $base

# szarvasmarha chr-k,DE gímszarvas géntérkép szerint széttört/illesztett és\
# ortológ gím chr számmal
for i in `1 33` x y ; do

echo "processing chr$i at `date`"

# Illesztés, coords fájlok létrehozása
nucmer -p chr${i} --mum --coords -c 100 ${base}/subject/chr${i}.fasta \
${base}/query/chr${i}.fa

#könyvtár létrehozása a pdf fájloknak
mkdir pdf_chr${i}

#pdf készül az illesztésről
mapview --format pdf --mag 2.0 -I -Ir chr${i}.coords

#pdf fájlok áthelyezése a könyvtárba
mv *.pdf pdf_chr${i}

done
exit
```

20/a. melléklet. BWA illesztéseket feldolgozó szkript.

```
#!/bin/bash

# futtatás ./bwa_bed_feldolg.sh
# soronként szedi ki a kromoszóma neveket(chr1-tol chry-ig) a bwa_list.txt\
# fájlbol, es behelyetesíti az $line változóba

 i=0

while read line

do

echo $line

#a bwa illesztés eredményeként kapott 12 oszlopbol 6, tabulátorral\
# elválasztott oszlopot készít : 6 oszlop $1=chr(szám)  $2=start_pozició bp\
# $3=stop_pozició bp    $4=scaffold_azonositó   $5=.    $6=orientácio(+/-)
awk '{print $1"\t"$2"\t"$3"\t"$4"\t""."\t"$6}' $line.bed |sed '/^$/d'\
 > $line_.bed

#$_line bed fájlok átalakítása, mínusz (reverz) helyzetű scaffoldok \
# jelölése, 100 millió bp-os régióban összevonja az azonos scaffoldokat
awk '{print $4"\t"$2"\t"$3"\t""."\t"$6}' $line_.bed \
|awk '{if ($5=="-") {print $1"_minus""\t"$2"\t"$3"\t"$4"\t"$5} \
else {print $1"\t"$2"\t"$3"\t"$4"\t"$5}}' \
| bedtools merge -i stdin -d 100000000 > $line.bwa_sorted.bed

rm -rf $line_.bed

i=$(($i+1));

done <bwa_list.txt
```

20/b. melléklet. MUMmer illesztéseket feldolgozó szkript.

```
#!/bin/bash

# futtatás ./mummer_coords_feldolg.sh
# soronként szedi ki a kromoszóma neveket(chr1-tol chry-ig) a\
# coords_list.txt fájlbol, és behelyetesíti  az $line változóba

 i=0

while read line

do

echo $line

#a mummer nucmer illesztés eredményeként kapott coords fájlokbol 6,\
# tabulátorral elválasztott oszlopot készít 6 oszlop $1=chr(szám)\
# $2=start_pozició bp   $3=stop_pozició bp  $4=scaffold_azonositó   $5=. \
# $6=orientácio(+/-)
grep -vw "Q"  $line.coords | grep -vw "home" | grep -vw "NUCMER" \
 | grep -vw "=" | grep -vw "[S1]"|sed '/^$/d' \
| awk '{print $12"\t"$1"\t"$2"\t"$13"\t"."\t"$4"\t"$5}' \
| awk '{if ($6 < $7) { print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t""+"} \
else {print $1"\t"$2"\t"$3"\t"$4"\t"$5"\t""-"}}' > $line.bed

#$line.bed fájlok átalakítása, mínusz(reverz) helyzetű scaffoldok \
# jelölése, 100 millió bp-os régióban összevonja az azonos scaffoldokat
awk '{print $4"\t"$2"\t"$3"\t"."\t"$6}' $line.bed \
|awk '{if ($5=="-") {print $1"_minus""\t"$2"\t"$3"\t"$4"\t"$5} \
else {print $1"\t"$2"\t"$3"\t"$4"\t"$5}}' \
| bedtools merge -i stdin -d 100000000 > $line.mum_sorted.bed

rm -rf $line.bed

i=$(($i+1));

done <coords_list.txt
```

20/c. melléklet. BWA és MUMmer illesztéseket kezelő szkriptek eredményfájljait feldolgozó szkript.

```
#!/bin/bash

# futtatás ./bwa_mummer_ossz_feldolg.sh
# soronként szedi ki a kromoszóma neveket(chr1-tol chry-ig) a bed_list.txt\
#  fájlbol, es behelyetesíti az $line változóba

 i=0

 while read line

 do

 echo $line

# Kromoszómánkent fűzi ossze a bwa_sorted.bed és a mum_sorted.bed fájlokat.\
# Az azonos azonosítójú scaffold illeszkedő HSP-kből a leghosszabbat\
# választja ki, és a szarvasmarhára illeszkedő rész közepérteket veszi\
# bp-ban megadva. E szerint rendezi sorba a scaffoldokat \
# az orientaciót(+/-) is figyelembe véve. Végül a scaffoldokat\
# kromoszómába fűző szkriptnek készít megfelelő formátumú bemeneti fájlt.
cat $line.bwa_sorted.bed $line.mum_sorted.bed \
|awk '{print $2"\t"$3"\t"$3-$2"\t"$1}' |sed 's/_minus/\tminus/g' \
|sort -k4,4 -k3,3nr |sort -u -k4,4 --merge \
|awk -v OFS="\t" '{if ($5=="minus"){print $1,$2,$3,$4,"-",($2-$1)%2+$1}\
 else {print $1,$2,$3,$4,"+",($2-$1)%2+$1}}' \
|sort -nk6,6| awk -v chr=$line -v OFS="|" '{print $4,chr,$5}' >$line.txt

 i=$(($i+1));

 done <bed_list.txt
```

21/a. melléklet. Scaffold összefűző szkript.

```
#!/bin/bash

file="chr1.txt"
wdir="/molbio/projects/MBK/szarvas/chromosomes/seqs2"
scaf_dir="/molbio/blastdb/MBK/szarvas"

declare -a myarray

i=0

while IFS=$'\n' read -r line_data; do
    myarray[i]="${line_data}"
    ((++i))

done < $file

for gen in ${myarray[@]}
        do

IFS='|' read -a array <<< "${gen}"
        echo ${array[0]}

        mkdir -p  ${array[1]}

       if [ "${array[2]}" = "+" ]; then
                echo "+"

                fastacmd -d ${scaf_dir}/Ce_ALLPATHS-scaffolds -s ${array[0]}\
                |sed 's/ No definition line found//g' | sed 's/lcl|//g' \
                > ${wdir}/${array[1]}/${array[0]}.fasta
        else
                echo "-"

                fastacmd -d ${scaf_dir}/Ce_ALLPATHS-scaffolds -s ${array[0]} \
                -S2 |sed 's/ No definition line found//g' | sed 's/lcl|//g'\
                > ${wdir}/${array[1]}/${array[0]}.fasta
        fi

        cd ${wdir}/${array[1]}

        cat ${array[0]}.fasta >> chr1.dfa

        cd ..

done
```

Magyarázat: A bash szkript chr1.txt-ben leírt sorrendben és orientációban veszi ki a scaffoldokat az Ce-ALLPATHS-scaffolds BLAST könyvtárból és fűzi össze őket az 1. gímszarvas kromoszómába.

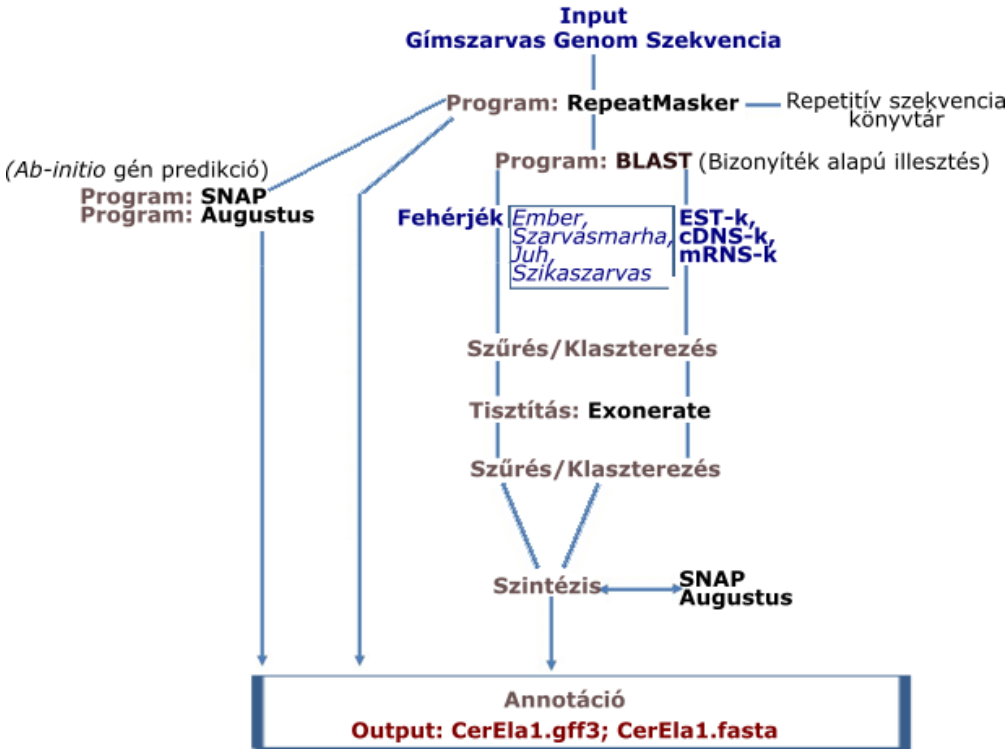
21/b. melléklet. Kromoszóma és genom készítés.

```
# kitörlöm az összes scaffold fastáját
rm -rf *.fasta

# a fasta > fejléceket kicserélem 100 N-nel, majd kézzel kiszedem az 1. 100 N\
#helyette >chr1-t kell beírni
sed -i "s/^>.*$/`for i in {1..100}; do echo -n N; done`/g" chr1.dfa

# átformázom a chr1.dfa-t chr1.fasta fájlla a seqret paranccsal
seqret chr1.dfa chr1.fasta

# megismétlem a fenti lépéseket chr1-tol chry-ig

# összefűzöm az összes gímszarvas kromoszómát egy Ce_pseudogenome.fasta fájlá
cat chr1.fasta chr2.fasta chr3.fasta chr4.fasta chr5.fasta chr6.fasta chr7.fasta\
 chr8.fasta chr9.fasta  chr10.fasta chr11.fasta chr12.fasta chr13.fasta\
 chr14.fasta chr15.fasta chr16.fasta chr17.fasta chr18.fasta chr19.fasta\
 chr20.fasta chr21.fasta chr22.fasta chr23.fasta chr24.fasta chr25.fasta\
 chr26.fasta chr27.fasta chr28.fasta chr29.fasta chr30.fasta chr31.fasta\
 chr32.fasta chr33.fasta chrX.fasta chrY.fasta> Ce_pseudogenome.fasta
```

22. melléklet. A gímszarvas transzfer, mikro és riboszómális RNS és repetitív genomi szekvenciák annotálása

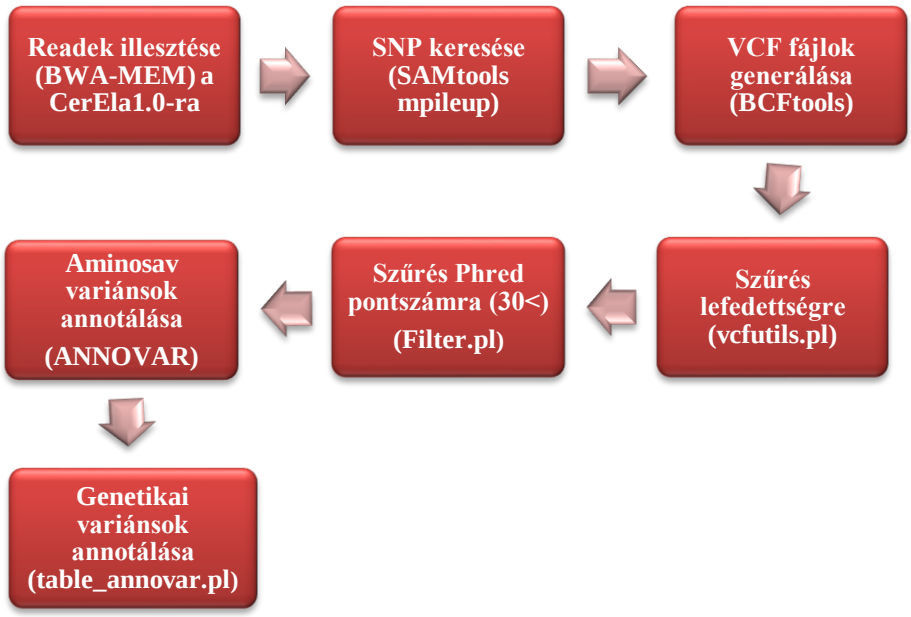
Repetitív szekvencia és t-, mi-, rRNS annotáció		
DNS szekvencia	**Program**	**Adatbázis**
repetitív szekvenciák	RepeatMasker	belső: artiodactyl
transzfer RNS	tRNAscan-SE	belső: eukaryotic
mikro RNS	BLASTN	miRBase: mammals
riboszómális RNS (LSU, SSU)	BLASTN	SILVA123 arkdb
	RepeatMasker	belső: artiodactyl
riboszómális RNS (5s)	Barrnap	belső: eukaryotic
	RepeatMasker	belső: artiodactyl

23. melléklet. A gímszarvas fehérje kódoló gének annotációja MAKER programmal (Cantarel és mtsai., 2008, 2. ábrája alapján).

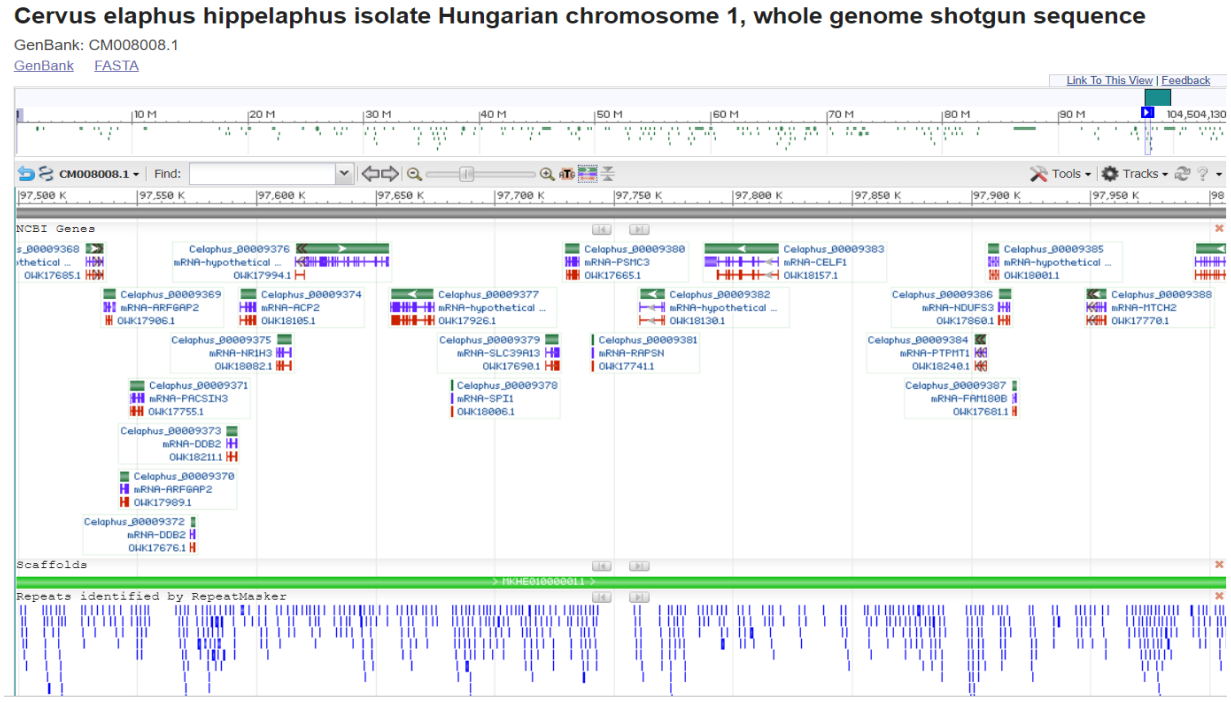


Magyarázat: Színkód: Kékkel jelöltem a bemeneti szekvenciákat, Szürkével a munkafázisokat, feketével a számítógépes programok nevét és pirossal a kimeneti fájlokat.

24. melléklet. Gímszarvas heterozigocitás (SNV, INDEL) vizsgálat folyamatábrája.



25. melléklet. A gímszarvas 1 kromoszóma 500 Kbp-os részlete az NCBI genom böngészőben.



Magyarázat: A gímszarvas 1 kromoszóma 500 Kbp-os részlete az NCBI genom böngészőben. Felül a kromoszóma szám egyenessel, középen a gímszarvas gének zöld téglalappal jelölve, nyilak mutatják az irányultságukat, alattuk kék jelzéssel mutatja, hogy melyik szarvasmarha mRNS-sel ortológok, pirossal a CDS szekvenciájuk, a vertikális boxok az exonok, a horizontálisak az intronokat jelölik. A scaffold alul zöld színű csík, legalul a repetitív szekvenciák vannak.

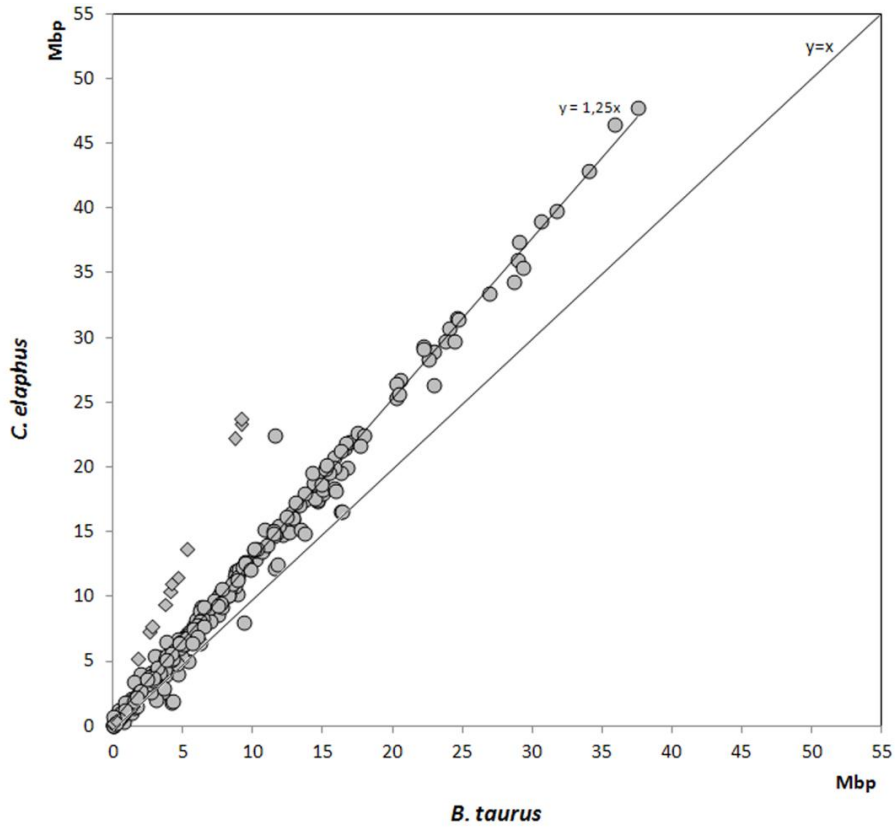
26. melléklet. A RepeatMasker program táblázatos formátumú eredmény fájlja.

file name: CerEla1.0.fasta
sequences: 35
total length: 3385636737 bp (1928192114 bp excl N/X-runs)
GC level: 41,51 %
bases masked: 769492957 bp (22,73 %)

		number of elements*	length occupied		percentage of sequence	
SINEs:		1443280	209069252	bp	6,18	%
	Alu/B1	0	0	bp	0	%
	MIRs	348401	49923175	bp	1,47	%
LINEs:		1039158	393951179	bp	11,64	%
	LINE1	520359	231587901	bp	6,84	%
	LINE2	227931	56065241	bp	1,66	%
	L3/CR1	30317	6214363	bp	0,18	%
	RTE	259649	99947899	bp	2,95	%
LTR elements:		300340	96940119	bp	2,86	%
	ERVL	66416	24866835	bp	0,73	%
	ERVL-MaLRs	107869	34041529	bp	1,01	%
	ERV_classI	77337	31321116	bp	0,93	%
	ERV_classII	33875	3264732	bp	0,1	%
DNA	elements:	250418	49072131	bp	1,45	%
	hAT-Charlie	141928	26225569	bp	0,77	%
	TcMar-Tigger	38629	10057487	bp	0,3	%
Unclassified:		4585	769585	bp	0,02	%
Total interspersed repeats:			749802266	bp	22,15	%
Small RNA:		216401	35304753	bp	1,04	%
Satellites:		2624	699977	bp	0,02	%
Simple repeats:		377096	15587604	bp	0,46	%
Low complexity:		62511	2988665	bp	0,09	%

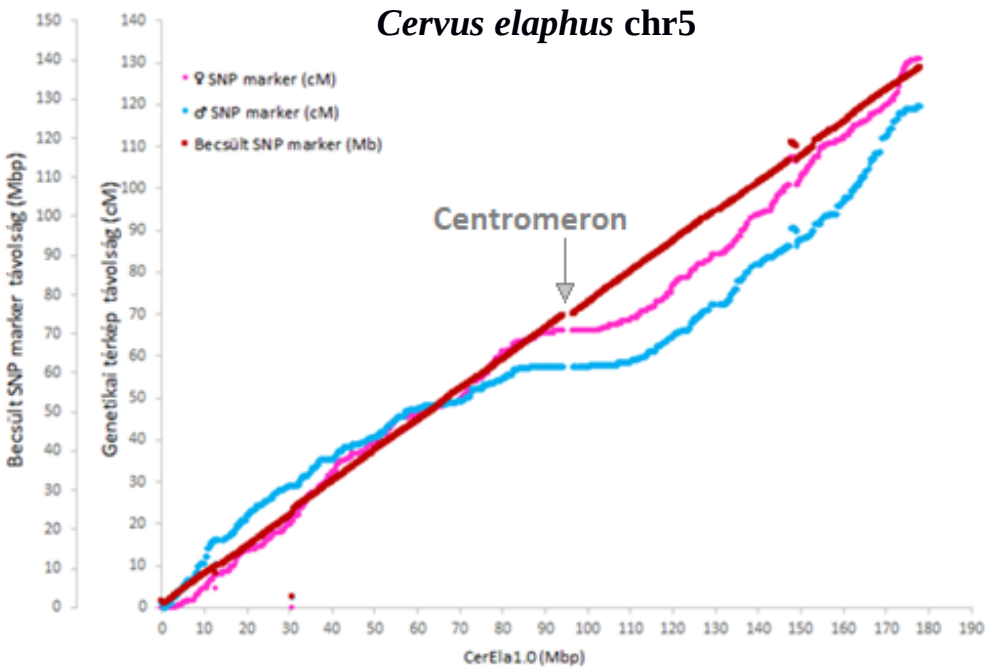
Magyarázat: Megadja a referencia genomban található repetitív elemek fajtáját, pontos hosszúságát és százalékos arányát. Retrotranszpozonok: SINE (short interspersed nuclear elements, rövid megszakított nukleáris elemek), LINE (long interspersed nuclear elements, hosszú megszakított nukleáris elemek), LTR (long terminal repeat, hosszú ismétlődő szekvencia).

27. melléklet A gímszarvas és a szarvasmarha genomális szegmensek összevetése



Magyarázat: A CerEla1.0 térképpontok által meghatározott genomi szegmenseit összehasonlítottam a Btau_5.0.1. ortológ szegmenseivel. Megjegyzés: (i) A körök jelzik azokat a távolságokat, ahol térképpont távolságok 1,25-szer hosszabbak a CerEla1.0-ban, mint a Btau_5.0.1-ban. A körök teszik ki a jelzések nagy részét, vagyis ez a hosszkülönbség általánosan érvényes a két genomra nézve. (ii) Négyzetekkel jelöltem a gímszarvas és szarvasmarha 11 kromoszóma ortológ szegmensei közötti hosszúsági eltérést, amely 2,2-szer nagyobb értéket mutatott a Ce11-ben, mint a Bt11-ben.

28. melléklet. Az 5. gímszarvas kromoszómára felillesztett Johnston SNP markerek.



Magyarázat: X tengely: Az angol-újzélandi SNP géntérképi marker pontok lokalizációja a CerEla1.0-ban az 5. kromoszómán Mbp-ban megadva. y1 tengely, Az angol-újzélandi SNP markerek géntérképe cM-ban meghatározott géntérképi távolság skálán. Rózsaszín körök az ünőkre, a kék gyémánt a gímszarvas bikákra vonatkozik. y2 tengely: A angol-újzélandi SNP markerek Mbp-ra átszámított távolság értékei (vörösesbarna négyzetek).