THESES OF DOCTORAL (Ph.D.) DISSERTATION

SZENT ISTVÁN UNIVERSITY– KAPOSVÁR CAMPUS
FACULTY OF AGRICULTURAL AND ENVIRONMENTAL SCIENCES
Department of Animal Breeding Technology and Management

Head of Doctorate School:
PROF. DR. ANDRÁS SZABÓ
Doctor of the Hungarian Academy of Sciences

Supervisors:
PROF. DR. PÉTER HORN
Member of the Hungarian Academy of Sciences

PROF. DR. LÁSZLÓ OROSZ
Member of the Hungarian Academy of Sciences

DNA SEQUENCE OF RED DEER CHROMOSOMES, COMPILATION OF CERELA1.0 RED DEER GENOME

Written by:
ÁGNES NÓRA BANA

KAPOSVÁR
2020

1. BACKGROUND

The red deer (*Cervus elaphus hippelaphus*) has been an integral part of human culture since ancient times, with outstanding social, natural, and economic benefits. It is a Holarctic species whose individuals are found in large numbers in the forests and steppes of Eurasia, North America, and certain areas of North Africa. Deer farms are widespread around the world, producing high-quality meat products using modern breeding methods and from which trophies with high CIC scores are derived. The Hungarian fauna stands out from the European populations with record trophies (Szálka, Gamás, Lenti, Pusztakovácsi, Gemenc-Karapancsa, Vajszló, Lábod).

The Deer Farm in Bőszénfa which was established by the University of Kaposvár later renamed as Bőszénfa Game Management Landscape Center between 1991-92 can be converted to the red deer population of Gemenc. The red deer genome program, launched in 1998, is closely linked to the Bőszénfai Deer Farm and the Doctoral School of Animal Science of the University of Kaposvár. The MSc and Ph.D. programs of the Agricultural Biotechnology Center in Gödöllő and the Department of Genetics of the Eötvös Loránd University in Budapest and the SOTE 1st Department of Internal Medicine also played a significant role in the implementation of the project.

The DNA sample on which CerEla1.0 is based was obtained from a 7-year-old red deer with a huge antler (ear number: Crot. N.o. 3016) living in close-to-nature conditions on the Deer Farm in Bőszénfa belongs to Game Management Landscape Center of the University of Kaposvár. After collection of 3x10 ml of blood, total genomic DNA was extracted using a Duplicα Prep Automated DNA / RNA Extraction System (EuroClone S.p.A., Italy) kit.

Purified DNA was sequenced on the Illumina HiSeq2000 platform, resulting in 4 paired-end and 2 mate-pair sequence libraries. These libraries contain a total of more than 2 billion raw read sequences, corresponding to 223 billion base pairs of DNA. This represents an average coverage of 74 times the red deer haploid genome.

From the overlapping paired-end reads, the ALLPATHS-LG assembler program first created so-called contigs, which were contiguous sequences. From the contigs, even larger sequence units were generated, which called scaffolds. In parts where there were no overlapping reads, pairs of the mate-pair read placed the contigs on the same scaffold. A total of 437,412 contigs were generated. They had N50 values of 7.5 Kbp. The number of scaffolds concatenated from the contigs was 34724, and their N50 value was 430 Kbp.

In my work, I used the double reference-guided genome assembly method.

The first reference was a red deer linkage genetic map generated by two New Zealand deer farms using an interspecific back-cross breeding procedure. 7 F1 Père David's deer (Elaphurus *davidianus*) and red deer (*Cervus elaphus*) hybrid deer stags were crossed with 267 red deer hinds, resulting in 351 back-cross offsprings. A genetic map for back-cross progenies identified 714 markers, of which, however, only 621 could be placed as accurate map points. Markers could have been microsatellites, RFLVs, ESTs, proteins, and AFLPs. The genetic map was arranged according to the red deer haploid chromosome number together with the X / Y chromosomes into 34 linkage groups, which footed up 2532 cM long.

The other reference was provided by an online physical genetic map of a closely related species. That was cattle (*Bos taurus*), which had a well-annotated reference genome sequence (Btau_5.0.1).

2. AIMS

My main goal was to create the “Red Deer Reference Genome Sequence (CerEla1.0), which is the first red deer reference genome in the world, and the first complete mammal reference genome, made in Hungary. This project came about as part of the Red Deer/Wonder Deer Genome Program. In the course of my work, I set the following sub-goals.

Match the available red deer genetic map (Slate, et al., 2002) and the *de novo* red deer sequence skeletons (genome assembly / scaffolds).

1. Corresponding to the available red deer genetic map (Slate et al., A Deer (Subfamily Cervinae) Genetic Linkage Map and the Evolution of Ruminant Genomes, 2002) to the *de novo* red deer sequence units (genome assembly/scaffolds).
2. Reconciling the points of the red deer genetic map (DNA markers) and the orthologous sequences of the cattle reference genome.
3. Identifying orthologous sequences of cattle reference genes on red deer scaffolds.
4. Comparing and fit *de novo* red deer scaffolds to the cattle reference genome taking into account evolutionary changes.
5. Searching for genes encoding proteins, repetitive sequences, transfer RNAs, ribosomal RNAs, and microRNAs in the red deer genome.
6. Determining the centromere positions of red deer chromosomes.
7. Uploading the red deer's complete genome to the NCBI server (where gets a unique identifier), so it will be available online and can be downloaded.

3. MATERIALS AND METHODS

As the first step of my work, I collected the DNA sequences of the red deer gene map points by name and identifier from literature data and online Ensembl, UCSC, Uniprot, NCBI, and ENA databases. I created two different blast libraries from all red deer scaffolds and the cattle reference genome with the makeblastdb command. I put the map markers based on their types along with the 10 DeerPlex microsatellites into multifasta files. These files were called queries (search files). The markers (queries) were aligned to the cattle and red deer scaffold blast libraries with the blastn, tblastx, and tblastn commands. The scaffolds caught in this way were called "mapmarker" scaffolds (MMSc). In the following, I identified the orthologous sequences of the red deer genetic map points in the cattle's whole reference genome (comparative gene mapping principle). The points of the red deer genetic map were located in the same order, co-linearly in the bovine genome, ie extensive syntenies and local linkages can be observed.

Due to the similarity and syntenies of the markers of the two species, similarity between their genes can also be assumed, so in the next step, I used MegaBLAST command to align the bovine genes downloaded from the UCSC online database to the red deer scaffold blast library. Scaffolds caught in this way were named reference genes containing scaffolds (RGSc).

I masked low-complexity genomic sequences and repetitive elements of the cattle reference genome with RepeatMasker program (hard-masking) so these did not disturb subsequent aligning. I then split the entire reference genome into chromosomes with the seqretsplit command. The number of red deer and bovine chromosomes is different because chromosomal rearrangements (Robertsonian fusion, Robertsonian translocation, fission, translocation, paracentric inversion, and inversion) have occurred during their evolution.

Due to the chromosomal differences between the two species, the reference bovine chromosomes were transformed according to the red deer genetic map using the seqret program. In some cases, I observed the exchange of two adjacent markers in the course of comparing cattle sequences to the red deer linkage map. I considered inversion to be conceivable when the distance between two neighbor markers was 1 cM or more but below 1 cM distance, I remained the marker ranking corresponding to cattle genome.

BLAST, MUMmer3.0, and BWA (version 0.7.10-r789) programs aligned the markers and MMScs to the bovine genome. The positions of the RGScs were also determined on bovine chromosomes using MegaBLAST, MUMmer, and BWA programs.

Then I used the MegaBLAST, LASTZ_32, MUMmer, and BWA bioinformatics software packages to align non-reference genes (scaffolds containing genes (miRNA, tRNA, rRNA genes) beyond the USCS Refseq genes, i.e. the so-called inter reference genes scaffolds, IRGScs). In this workflow, I determined the localization, order, and orientation of RGScs and IRGScs on the hard masked bovine chromosomes, which had been converted according to red deer linkage groups before. I processed the result files of several programs with homemade Bash scripts, essentially with awk, sed, bedtools merge, and sort pipelines.

I masked the parts of the converted bovine chromosomes, whither MMSc, RGSc and IRGSc scaffolds had been aligned (hard masking). The remaining gaps were filled with scaffolds above 1999 bp so-called gap-filling scaffolds (GFSc) using BWA and MUMmer programs.

The scaffolds placed in the appropriate order and the 2582 individual contigs were concatenated with a Python script, which inlaid 100 bp length of N characters in the gaps between them. I searched for repetitive sequences with RepeatMasker Open-4.0. software. The small and large subunit (SSU, LSU) sequences of mammalian ribosomal RNA came from the SILVA123 reference database.

The precursor format microRNAs were downloaded from the miRBase database, of which 21 mammalian miRNA sequences were used later. I created search sequence files from the rRNAs and separately from the miRNAs, which I aligned to our red deer genome with the blastn command. I searched for the 5S unit of ribosomal RNA using Barrnap 0.6 program. Transfer RNAs were found with tRNAscan-SE-1.3.1 software.

Protein coding genes were identified by the MAKER 2.31.8 gene annotation program by calling different subroutines one after the other (RepeatMasker-open-4.0.5, RepeatRunner, RepeatModeler 1.0.4, RECON, RepeatScout 1.0.5, Exonerate, BLAST, SNAP, AUGUST). MAKER aligned EST, mRNA, and protein sequences from other species (red deer, cattle, sheep, human) to the soft-masking red deer genome using the BLAST algorithm. Ab-inito gene prediction was performed with SNAP and AUGUSTUS. Protein functions and protein-encoding genes were determined using InterProScan software.

For the determination of red deer genetic variants (SNV, INDEL), the reads of the paired-end and mate pair libraries were fitted to the reference genome by BWA -mem program. In the aligned reads, the SAMtools (parameters: mpileup -D -S -E -uf) software localized the genetic variants. The annotations of the variants causing the functional changes of the amino acids were performed by ANNOVAR software.

4. RESULTS

As the first step of the red deer genome program, DNA was isolated from the blood of a seven-year-old red deer stag lived in Bőszénfa. This stag had antlers with 12 kg. The whole genome of the animal was sequenced on the Illumina HiSeq2000 sequencing platform. As a result of this process, 4 paired-end and 2 mate pair sequence libraries were generated with a total of approximately 2.2 x 10^9 reads corresponding to 222.7 Gbp lengths of DNA. The *de novo* assembly was performed by the ALLPATHS-LG program, resulting in 437412 contigs, which were a total of 1.95 Gbp lengths. The contigs were concatenated into larger sequence units, called scaffolds. The number of scaffolds was 34724, and their total length, including gaps, was 3.4 Gbp.

The genetic map of the red deer (Slate et al., A Deer (Subfamily Cervinae) Genetic Linkage Map and the Evolution of Ruminant Genomes, 2002) used as the first reference consisted of 34 linkage groups. The total length of the linkage groups were 2532 cM long, which contained 621 genetic markers. These had an average density of 5.7 cM. I searched sequences of the 621 markers from online genomic databases. Sequences were found for all types (EST, RFLV, STS, protein) except AFLPs. So, I successfully localized the DNA sequence of 365 markers on the genetic map. Scaffolds carrying 365 markers were named mapmarker (MMSc) scaffolds. The MMScs were got in a defined position on the genetic linkage map.

The *B. taurus* reference genome (NCBI Btau_5.0.1) served as the second reference. The 365 *C. elaphus* marker sequences and mapmarker scaffolds (MMSc) were aligned on this genome. Thus, I successfully identified orthologous sequences in the bovine genome similar to 365 red deer MMSc. The orthologous sequences on the chromosomes were located collinearly throughout both species. Also, the order of the genes in the MMSc was the same as that of the bovine genes, i.e., the syntenies also prevailed at the "intra-scaffold and intra-contig" levels.

Due to the collinearity of the red deer and bovine MMSc sequences, it can be assumed that the gene sequences of the two species are also very similar. Because of the putative similarity, I downloaded bovine genes from the UCSC genomic database available online.

Next, I used these search genes as “baits” to fish out “prey” scaffolds that contained red deer sequences orthologous to the bait gene. Such scaffolds were termed “reference gene scaffolds” (RGScs). The more certain the orthologous sequence of bovine genes in the same order, the more certain the position of an RGSc was. As a result of the process, I placed RGScs between the marker points. Following orthologous sequences of bovine genes in the red deer genetic map, I localized a total of 6013 scaffolds.

In the next work phase, I filled the gaps between MMSc-RGSc and RGSc-RGSc with the scaffolds not placed so far. These presumably did not contain the UCSC-provided Refseq protein-encoding genes, but most likely had other protein-encoding genes, such as non-Refseq, rRNA, tRNA, and miRNA genes. I named these searching red deer scaffolds, “inter reference gene scaffolds” (IRGSc). Combining MMScs, RGScs, and IRGScs, I was able to accommodate 13748 scaffolds, however, the localization of scaffolds longer than 15205 2 Kbp remained questionable.

As a solution to the problem, I masked the existing 13748 *Bos taurus* genomic site. Sequences larger than 2 Kbp were aligned to a “masked” reference in this way. Therefore, I found 9845 new scaffolds so-called gap-filling scaffolds (GFSc), which were localized on the no-masked parts of the bovine chromosomes. A total of 23593 scaffolds were aligned to the cattle orthologous regions up to now.

For most of the sequences, more precisely 99.6%, I found a clear correlation between the two species, but 102 scaffolds showed orthology with several distinct chromosomal segments or, conversely, covered identical positions. To solve the problem, I split the scaffolds into separate contigs or merged them into one identical scaffold. The resulting 2582 unique contigs and the new 35 scaffolds were aligned to the reference genome in this form.

As a result of my scientific work, I have created the haploid red deer (*C. elaphus* *hippelaphus*) reference genome (or pseudochromosomes), which is called CerEla1.0. The CerEla1.0 genome consists of 23491 scaffolds plus 35 new scaffolds (MMScs, RGScs, IRGScs, GFScs plus 35 new ones) and 2582 individual contigs, i.e. a total of 26108 sequence elements. The red deer reference genome (CerEla1.0) is 3.4 Gbp lengths.

11444 scaffolds were put into the "unplaced" category because I could not localize them on the new red deer reference chromosomes. The unplaced sequences are a total of 52989442 bp lengths, which is 1.6% of the total genome.

Of great importance were scaffolds that contained multiple genes, i.e., possible linked genetic elements. Comparing the red deer chromosomes with the Btau_5.0.1 bovine genome, it was found that the intra-scaffold genes showed local linkages similar to the bovine orthologous genes. For all multi-gene red deer scaffold (3422) and orthologous bovine chromosomal segments, I observed these extensive syntenies. Thus, it can be said that the order of genes can not only be confirmed at the chromosomal level but also manifested subchromosomal, within the scaffolds. Red deer genes were identified using the bovine, ovine, and human transcriptome and proteome, as well as the *de novo* generated sika deer transcriptome, using the MAKER program. The MAKER with its pipelines has annotated 19368 protein-encoding genes. The sequence of identical genes was predominantly the same in the bovine and red deer genomes. The complete reference genome (CerEla1.0) has been uploaded to NCBI (NCBI ID MKHE00000000.1, GCA_002197005.1), where annotated genes can be displayed using the genome browser. Besides, the position of 589 rRNA coding genes (LSU, SSU) covering a sequence of 98.3 Kbp, representing 0.0029% of the total pseudochromosome length, was determined. 1029 of the 5s rRNA genes (96 Kbp, 0.0028% of CerEla1.0), 2096 of the tRNA genes (128 Kbp, 0.0038% of CerEla1.0) and 264 of the microRNA genes (27.7 Kbp, 0.0008% of CerEla1.0) was localized. The RepeatMasker program recognized 769492957 bp repetitive region, which accounts for 22.73% of the total genome.

The structure of *C. elaphus* chromosomes is quite "primitive" in nature, as almost all are acrocentric (Chromosome A). The majority of red deer chromosomes (19 autosomes and X, Y) can be directly paired with a homeologous bovine chromosome. All the same, 6 acrocentric ancient chromosomes were preserved in cattle, however, during the evolution of red deer, 12 acrocentrics (deer) chromosomes were formed by the fissions of these 6 ancient chromosomes. This means that new centromeres have been formed on the deer evolution line. The Ce5 metacentric (M chromosome) was formed by Robertsonian translocation of two acrocentric chromosomes upon detachment of the Cervinae branch. Based on the karyograms of the Pecora lineages (Bovidae, Cervidae), Robertsonian fusions, and fissions (i.e. when 1 acrocentric chromosome is formed by tandem fusion from 2 acrocentric chromosomes and vice versa) could have occurred very often in the evolution of the two species. The red deer genetic map did not indicate the position of the centromeres.

Based on the points of the linkage groups, it was not possible to determine the position of the red deer centromeres. On the bovine side, the position of centromeres in the *B. taurus* genome was indicated, since, in previous chromosome cytology studies, the location of centromeres and adjacent genes was determined by centromere staining and in situ DNA hybridizations. These genes were already identifiable in the bovine genome sequence, and their orthologs were also included in the red deer genome sequence, CerEla1.0. So I searched orthologous sequences of the centromere near genes of the bovine genome in the homeologous red deer chromosomes. Thus, a comparative analysis of the available orthologous bovine and red deer pseudochromosome sequences and cytogenetic karyograms (band patterns of ranked metaphase chromosomes) determined the possible position of the centromeres (adjusted to the genetic map points) in both the 34 red deer linkage groups and the pseudochromosomes of red deer. In this way, 12 red deer chromosomes (Ce3-22, Ce6-17, Ce8-33, Ce16-29, Ce19-31, Ce26-28) can also be paired with 6 bovine chromosomes (Bt5, Bt6, Bt2, Bt8, Bt1, Bt9) accordingly, the position of the centromeres could be given with certainty. The acrocentric red deer Ce19 is orthologous with the distal half of bovine Bt1, while the also acrocentric red deer Ce31 corresponds with the proximal portion of bovine Bt1. However, the situation is more complex than this, as, in the ancestor of the Ce19 chromosome, a well-defined fracture and translocation (the lower and upper segments swapped) also took place during evolution. The red deer Ce28 and Ce26 acrocentric chromosomes are identical to the two arms of the bovine Bt9 acrocentric chromosome. This is because, in the ancestor of deer, this chromosome is split in two. The centromere of red deer Ce26 has no equivalent in bovine Bt9 sequences, as red deer Ce28 "inherited" the centromere, and based on a comparison of band patterns and genetic map point order, a paracentric inversion had to occur in the red deer. As a result, the large segment of all orthologous markers on the chromosome was completely inverted, while the sequences closest to the centromere remained in place. I discovered two cases where one red deer chromosome is orthologous to two bovine chromosomes. One arm of the metacentric red deer Ce5 corresponds to bovine Bt17 and the other to Bt19 (Robertsonian translocation). The acrocentric red deer Ce15 came into existence by Robertsonian fusion in red deer evolution line, in which the proximal part, near the centromere, is equivalent to bovine Bt28 and the distal portion corresponds with Bt26. Inversions, translocations, chromosomal fusions, and fissions relocate and isolate mappoint markers and create a new environment around the breakpoints of the rearrangements. In terms of red deer-cattle kinship, these "evolution-created" rearrangements led to the formation of new adjoining sequence regions. In the CerEla1.0 genome, I detected 26 deer-bovine rearrangements involving 18 inversions, 2 translocations, and 6 chromosomal fusions and cleavages. In the 6 cases where 1 acrocentric bovine chromosome is identical to two acrocentric red deer chromosomes, I divided the scaffolds and contigs between the two red deer chromosomes according to proportions of recombination distances of the deer genetic linkage groups. I arranged them in the order experienced in cattle. Combined DNA sequences of this type make up 5% (0.166 Gbp) of the CerEla1.0 genome. The inversions resulted in 54 red deer-bovine sequence switching points ("switch points"), which together with their overhanging "flanking" regions were essentially the same as the 462.9 Mbp sequence bounded by neighboring MMScs. This represents 13.5% of the CerEla1.0 genome. In 81.5% of CerEla1.0, red deer genes followed the sequence of bovine orthologous genes in the segments between MMSc, and within MMSc, the sequence of red deer genes prevailed. In 18.5% of the CerEla1.0 sequence (chromosomal fissions/fusions, inversions of flanking pieces), syntenic blocks of red deer and cattle genes were also combined.

In search of the heterozygosity points/genetic variants of the CerEla1.0 genome, 2807458 SNVs and 364689 indels were identified. Additional heterozygous SNVs were annotated based on the MAKER annotation pipeline result files. Thus, a total of 17700 non-synonymous and 14252 synonymous SNVs were found.

* Bt(ordinal number): *Bos taurus*, the ordinal number of cattle chromosome, Ce(ordinal number): *Cervus elaphus*, the ordinal number of red deer chromosome

5. CONCLUSIONS, RECOMMENDATIONS

The topic of the dissertation is extremely important from the point of view of Hungarian genomic research, as it is about the production of Hungary's first, real, internationally recognized complete reference genome sequence. It contributes greatly to the understanding population's genetic profile of red deer, thus providing an opportunity to design new microsatellite and SNP markers. The red deer reference genome can help to understand the evolutionary biology and lineages of this big game, and to improve farm breeding methods. Based on the knowledge of the structure of red deer genes and their promoter sequences, we can explore the genetic background of individuals with capital antlers. The functions and biomedical implications of genes involved in bone-antler metabolism and tumor formation will be mapped. The bioinformatics methods described in the dissertation can serve as a suitable model for establishing a reference genome of another agriculturally or environmentally valuable species.

Bioinformatics work was preceded by sample collection, DNA isolation, and whole genomic DNA sequencing. DNA sampling was performed from the blood of a 7-year-old capital red deer living on the Bőszénfa Deer farm of the University of Kaposvár Game Management Landscape Center, taking into account animal welfare laws. The high-purity DNA resulting from the DNA isolation was sequenced by the Danish company Aros Applied Biotechnology with the Illumina HiSeq2000, resulting in a large amount of bioinformatics raw data 2 billion read sequences suitable for further analysis. From the point of view of later population genetic studies, it would be worthwhile to extend the sampling and the whole genome sequencing to about 150-300 red deer individuals from different populations in Hungary. In terms of the sequencing method, it is advisable to stay with Illumina technology, as their new technologies aim to increase the size of the reads and reduce the cost of the procedure. The Broad Institute ALLPATHS-LG program concatenated the read sequences into contigs and scaffolds. The number of scaffolds was 34724. Relatively many scaffolds below 2000 bp, and few in large size, were created. Of which, those above 8 Kbp proved to be the most useful to generate chromosomes. The N50 of the scaffolds with gaps was 430 Kbp lengths, and N50 value was 265 Kbp for the gap-free scaffolds.

The length of the *C. elaphus* genetic map used as a reference corresponds to 2532 cM, while the total length of the "assembled" CerEla1.0 reference genome corresponds to 3.4 Gbp, i.e. 1 cM to 1.34 Mbp. This value is significantly higher than the commonly used approximate 1cM/1Mbp value (thumb role) or 0.8 Mbp/1 cM established for the bovine genome. The red deer CerEla1.0 reference genome appears to be 25% longer than the bovine Btau_5.0.1 reference genome. To find the reason for the 0.7 Gbp "extra" length of the red deer genome, I compared the orthologous segments of the pseudogenome of CerEla1.0 and Btau_5.0.1. I have found that, with a few exceptions, along with the red deer and bovine pseudogenomes, the red deer segments and additionally the scaffolds are uniformly 1.25 times longer than the bovine segments (except for this is the Ce11 chromosome, where this ratio is 2.2). Comparing the CerEla1.0 pseudogenome, which has a 25% (0.7 Gbp) extra lengths (3.4 Gbp) with the Btau_5.0.1 pseudogenome (2.7 Gbp), it can be seen that ALLPATHS-LG program has inlaid too many gap regions ("NNN") between contigs during the constructing of scaffolds. Assuming that the genomes of *B. taurus* and *C. elaphus* are essentially the same in size, and considering that the total DNA sequences of the contigs in CerEla1.0 are 1.9 Gbp, it can be assumed that instead of a 0.8 Gbp lengths gap region, which is proportionate with the bovine genome, a total of 1.5 Gbp gap was added to the scaffolds. However, their distribution is proportional to physical distances. A ratio of 1.25 was observed for deer-bovine segment lengths, however, this ratio was significantly higher, 2.2 in the Ce11 chromosome. We have not had an explanation for this value yet. In the red deer scaffolds, the segments of the genetic map markers and the order of the genes showed syntenies with the bovine (which is a closely related species) orthologous segments, i.e., the major substantial sequences of the scaffolds proved to be good. We will consider using other ALLPATHS-LG parameters or programs (such as DISCOVAR) when generating a newer red deer genome assembly.

The scaffolds were aligned into chromosomes using Jon Slate's red deer genetic marker map and the bovine reference genome. 621 EST, RFLV, STS, protein genetic markers were identified on the red deer genetic map, but their sequences had to be collected from literature data and bioinformatics databases. The problem was that no sequence could be found anywhere for 229 AFLPs, so these were excluded from further studies, and the sequences of some markers remained unknown in the other categories as well. For these reasons, I identified the DNA sequence of 365 marker points and further matched it to the scaffolds and the bovine reference genome as a search sequence using the BLAST program. The scaffolds found in this way were named mapmarker stands (MMScs). I also used several reference genome versions (NCBI Btau_4.6.1, Btau_5.0.1) for the study, although the latest one at that time would have been sufficient (Btau_5.0.1).

The points of the red deer gene map were located in long sections in the same order, co-linearly in both the bovine genome and the red deer scaffolds, therefore I filled between parts of the marker points up with red deer scaffolds (RGSc) caught with bovine orthologous genes using the comparative gene mapping principle. As a control, the scaffolds containing the genes and the other 15205 scaffolds longer than 2 Kbp were aligned to the bovine reference genome using LASTZ, MUMmer, BWA, and MegaBLAST programs. I called these inter-reference gene scaffolds (IRGSc). During the annotation process of the MAKER program, 19368 protein-encoding genes have been annotated in the red deer genome, which fully corresponds to the number of protein-encoding genes between 19,000 and 21,000 described in other mammalian species and covering 90% of ruminants genes. The order of the red deer genes and their bovine orthologs shows collinearity in the CerEla1.0 and Btau_5.0.1 genomes. We have determined the detailed structure of genes (exon, intron structure, their 5 'and 3' UTR regions), their evolutionary relationships (similar genes in other species), their functions (their role in development within the cell, organs, tissues). We have identified repetitive sequences, LTR elements, genes of RNAs (transfer, small and ribosome RNAs), and the position of all of these genetic elements on chromosomes. We have localized 2.8 million heterozygous point variants (SNVs) and 365000 small deletions and insertions in the red deer genome.

Completely independent of the CerEla1.0 red deer genome, a high-density red deer genetic map was also prepared by an English-New Zealand research team that determined more than 38000 recombination points. These SNPs came from mostly red deer, but a little part of these points from sika deers. This genetic map contains a large number of 38083 SNP marker sequences, which can validate the Hungarian red deer reference genome, CerEla1.0, with a comparative study. As a first step in the comparative work, I searched for and downloaded all English-New Zealand SNP markers with their 150-150 base pair flanking regions. I aligned these markers as search sequences to our red deer genome with the BLASTN command. I selected the best results. After that, I generated tables for each autosome in such a way as to show CerEla1.0 chromosomal localizations of the SNPs in megabase pairs and megabase pairs and cM positions of SNP markers estimated by the English-New Zealand researchers. From the obtained results I created point diagrams with an excel program. Based on the data and figures, the possible breakpoints and chromosome rearrangements were visible. Of the 38083 points, I was able to place 32408 on the chromosome as expected, i.e. 85% of all points. At these points, I experienced a larger inversion of the multi-point part in 33 cases. The most likely reason for the turns was the difference between Slate's and Johnston's genetic marker maps. Slate's linkage groups contain 621 markers in contrast with Johnston's genetic map, which has more than 38000 SNP markers. Respectively, in the case of two inverted Slate's markers in opposite positions, I could not be sure how large flanking parts belonged to the inverted sections. Furthermore, when constructing the red deer reference genome, some adjacent marker points on the Slate map also appeared to be inverted relative to the bovine reference genome, yielding 54 red deer-bovine inversions, for a total sequence of 462.9 Mbp. Slate's gene map was available to create CerEla1.0, so scaffolds were placed in this order. Of the aligned Johnston points, only 604 gave a false value, which is less than 1.9% of the placed points. 1019 SNPs were localized on incorrect chromosomes. The discrepancy of these points may have been caused by the fact that the chromosome number of the close relative bovine species, used to make the red deer pseudochromosome, does not match that of the red deer. In cases where 12 red deer chromosomes were orthologous to 6 bovine chromosomes, it was not possible to know for sure how, beyond the last genetic map markers, how to share between two red deer chromosomes the segment of the orthologous cattle chromosome. These uncertain portions represent a DNA sequence of 0.166 Gbp in length. Overall, however, the results obtained provide convincing evidence to confirm the correctness of the CerEla1.0 red deer genome.

In addition to the high-density SNP genetic map, the DeerPlex, a "parentage control kit" based on an extremely sensitive 10 autosomal tetranucleotide microsatellite locus, was available from the start of the work. During the research, DeerPlex microsatellites were placed successfully in an extensive DNA sequence environment.

CerEla1.0 can be used for SNP and/or microsatellite-based individual identifications, population genetic level studies, lineage, and evolutionary relationship findings. We are currently publishing an article on the development of new XY chromosomal microsatellite markers. These sex chromosomal markers greatly facilitate the exploration of paternal and maternal lineages, allow for objective and unambiguous identification, more conscious and reliable animal breeding, and can be a reliable tool in curbing poaching.

Further biomedical examinations and articles can be based on the analysis of promoter sequences of red deer bone and antler metabolism genes. During their annual antler cycle, red deer stags shed their bony antlers from February to May, after which it takes 100-120 days for new antlers to build. In this case, large amounts of minerals must be provided and delivered to the bone organ, which can weigh up to 14-17 kg. Nutrient calcium intake is low in early spring, so calcium and phosphate are carried from skeletal elements (sternum, ribs, and individual vertebrae) into antler, thereby inducing a decrease in bone density, i.e., physiological osteoporosis occurs in the affected skeletal bones. Later, mineral salts are replenished from the rich vegetation through nutrition, i.e., the bone density restores in the skeleton. According to the preliminary studies of our research group, 8 genes that play an important role in bone development may be expressed a much higher (10-fold to 30-fold) in the bony part of the antler than in skeletal bones. Of these genes, there are more runx2 transcription factor binding sites in the 1 and 5 kb promoter regions of the col1A1 gene in red deer than in humans or cattle because these genes are more active in red deer. In the future, we would like to investigate whether other genes associated with bone development have multiple runx2 or osx transcription factor binding site motifs, or whether some SNP or INDEL differences may cause any phenomena associated with osteogenesis or osteoporosis. In addition to the binding sites of the runx2 and osx transcription factors (master regulators, central regulatory genes for bone development), other conserved bone gene promoter motifs may be of interest, so it would be worthwhile to create multiple sequence alignment. These sequences may be derived from red deer and closely related species or species of interest for bone development. Other medical-biological research in the field of organ development/regeneration, robust tissue growth/tumor biology can be performed on the model of genomic studies of osteoporosis.

6. NEW SCIENTIFIC RESULTS

My doctoral research was organically linked to the red deer *(Cervus elaphus*) genome program, which is the compilation of the world's first recognized red deer reference genome. My tasks were mainly focused on the field of bioinformatics. Within the diversified project, my new scientific findings related to my work are as follows:

1) The double reference-guided alignment used in the scientific study made it possible to align shorter red deer DNA sequences (*de novo* scaffolds) into pseudochromosomes with greater accuracy than in the case of the commonly used single reference alignment. I utilized as a reference a red deer closely related species, the well-defined and annotated reference genome of cattle (*Bos taurus*), as well as Slate's red deer genetic map of 34 linkage groups and 621 marker points. However, I could only use 361 of the marker points for my work.
2) During the bioinformatics work I worked based on a newly introduced algorithm:
   a) I matched the red deer genetic map markers and the *de novo* red deer scaffolds.
   b) I searched for sequences very similar to red deer gene map markers on the bovine reference genome.
   c) I identified the orthologous sequences of the bovine reference genes on the red deer scaffolds.
   d) I aligned *de novo* red deer scaffolds to the bovine reference genome transformed according to red deer's linkage groups.
3) For all red gene-bearing red deer scaffolds, it can be seen that their intra-scaffold genes showed extensive syntenies with bovine orthologous genes. Consequently, the order of the genes was the same not only in the segments bounded by map point markers and at the chromosomal level, but also within the scaffolds.

4) Based on the observed relationships, it can be said that in 81.5% of the red deer genome, the genes follow the sequence of bovine orthologous genes in the segments between MMSc and the sequence of red deer within MMSc. In the remaining 18.5% of CerEla1.0, where chromosomal cleavages, fusions, and inversions occurred, substantial blocks of bovine and red deer genes were combined.
5) My scripts were created for the different work phases:
   a) I interpreted and processed the result files of BWA and MUMmer programs with my bash scripts.
   b) I wrote a bash script to bind localized scaffolds into chromosomes. The program filled up the gap between scaffolds with 100 base pairs of N characters on a sample of NCBI genomes. (Eventually, however, we uploaded chromosomes created by running a Python script created by my colleague to NCBI.)
6) The structure and localization of red deer genes were the first in the world to be determined. Complete annotation of the chromosomally arranged red deer reference genome has been completed, including the protein-coding genes, repetitive sequences, and gene sequences of ribosomal RNA, transfer RNA, and micro RNA.
7) Based on the karyograms of the bovine chromosomes and the centromere positions of the bovine chromosomes, I placed the centromeres of the red deer chromosomes.
8) The bioinformatics study confirmed the putative evolutionary changes in red deer chromosomes: 6 fission, 1 Robertsonian translocation, and 1 Robertsonian fusion compared to bovine chromosomes. It has been shown that fission and a translocation occurred in Ce19, which is equivalent to the distal arm of the Bt1 centromere. Based on the position of chromosomal bands and genetic map markers, a paracentric inversion event probably occurred in Ce28.

9) Overall, with my bioinformatics research, I made a substantial contribution to the compilation of the world's first internationally recognized complete red deer reference genome, which is arranged in chromosomes. The red deer reference genome (CerEla1.0) was uploaded to the NCBI genomics database and browser website. Thus, we made the entire genome sequence available and free online download (MKHE00000000.1) for everyone. As a result, 12 scientific articles have been written in the last 3 years referring to the CerEla1.0 genome.

7. PUBLICATIONS ABOUT THE SUBJECT OF DISSERTATION

Peer-reviewed papers published in English

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

Peer-reviewed papers published in Hungarian

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

Papers published in Hungarian

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

Papers published in conference proceedings

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.

Publication outside the topic of the dissertation: Peer-reviewed papers published in English

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

Scientific lectures

Bana Á. N. (2019) A gímszarvas genom program bővítése (CerEla1.0) egy nagy sűrűségű SNP alapú géntérképpel. ÚNKP beszámoló, Kaposvár, 2019.04.03.

Bana Á. N (2018) A Gímszarvas/Csodaszarvas Genom Program, GENETIKAI MŰHELYEK MAGYARORSZÁGON” XVII. Minikonferencia, előadás. Szeged, 2018.09.21.

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) The Red Deer *Cervus elaphus* Reference Genome CerEla1.0. 9th International Deer Biology Congress, Estes Park, Colorado, USA, August 5–10, 2018. pp. 163-164

Frank K., Barta E., Bana Á. N., Nagy J., Horn P., Orosz L., Stéger V. (2016) A magyarországi gímszarvas (*Cervus elaphus hippelaphus*) teljes mitokondriális genomja. Fiatal Biotechnológusok Országos Konferenciája „FIBOK 2016”, 2016. március 22.

Bana Á. N. (2016) Az első gímszarvas referencia genom szekvencia összerakása és annotálása. MBK napok 2016, Mezőgazdasági Biotechnológiai Kutatóintézet, Gödöllő, 2016. december 15.

Bana Á. N. (2016) A gímszarvas (*Cervus elaphus hippelaphus*) csont és agancs metabolizmus gének promóter szekvenciáinak bioinformatikai vizsgálata. Fiatal Biotechnológusok Országos Konferenciája 2016. Szent István Egyetem, Gödöllő, 2016. március 21.

Bana Á. N. (2015) Nagy teljesítményű genom analízisek bevezetése a dél-dunántúli gímszarvas populáció állományaiban, Kaposvár 2015. június 1.

Bana Á. N. (2013) A DeerPlex-ben felhasznált lókuszok kapcsoltsági vizsgálata, Gímszarvas workshop Bőszénfa, 2013. június 25.