The Red Deer *Cervus elaphus* Reference Genome CerEla1.0

Nóra Á. BANA\(^1,2\), Anna NYIRI\(^1\), János NAGY\(^3\), Krisztián FRANK\(^1,2\), Tibor NAGY\(^1\), Viktor STÉGER\(^1\), Mátéás SCHILLER\(^1\), Péter LAKATOS\(^4\), László SUGÁR\(^5\), Péter HORN\(^2\), Endre BARTA\(^1,6\) and László OROSZ\(^1,7\)

\(^1\)Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, 2100, Hungary
\(^2\)Department of Animal Breeding Technology and Management, Kaposvár University, Kaposvár, 7400, Hungary
\(^3\)Game Management Center, Kaposvár University, Kaposvár, 7400, Hungary
\(^4\)1st Department of Internal Medicine, Semmelweis University, Budapest, 1083, Hungary
\(^5\)Department of Game Biology and Ethology, Kaposvár University, Kaposvár, 7400, Hungary
\(^6\)Department of Biochemistry and Molecular Biology, University of Debrecen, Debrecen, 4032, Hungary
\(^7\)Department of Genetics, Eötvös Loránd University, Budapest, 1117, Hungary

*Corresponding author: orosz.laszlo@abc.nai.k.hu*

The first reference genome assembly CerEla1.0 (NCBI, MKHE00000000) for red deer (*Cervus elaphus*) has been presented [1]. The CerEla1.0 could prove useful in many cervid Genome Wide Association Studies (GWAS). Red deer are an emblematic member of the natural megafauna of the Northern Hemisphere and have been present in human culture since the Neolithic. Humans introduced and spread the species to many places, including the Southern Hemisphere. Red deer, the mythological Wonder Deer, also the Royal Game of the Middle Ages in Europe, is respected and revered in many cultures. Today red deer are not only one of the most desired game, but also farmed for venison, velvet antler products and tonic. Red deer are becoming increasingly recognized as an animal model for bone, osteoporosis and regeneration research, as well as for population and evolutionary studies.

Red deer stag DNA was sequenced by Illumina technology at 74x coverage. The ALLPATHS-LG assembly of the reads resulted in 34.7x10\(^3\) scaffolds. For building the red deer pseudochromosomes a pre-established genetic (recombination) map was used as the main anchor point [2]. A nearly complete co-linearity appeared between the array of the deer map points/map marker scaffold sequences and the order and orientation of the orthologous sequences in the syntenic bovine regions. Syntenies were also conserved at the in-scaffold level. The final CerEla1.0 assembly contains 26108 scaffolds and contigs and spans 3.4 Gbp including the NNN-s inserted between contigs during the scaffolding and between scaffolds. In nearly all genomic segments the cM distances corresponded uniformly to 1.34 Mbp, due to the many "NNN-s" inserted by the ALLPATHS-LG, 1.25-fold uniformly more than in the bovine homologous regions. In the resulting red deer pseudochromosome sequences, 2.8 million heterozygous sites/SNPs, 365 thousand indels and 19368 protein coding genes were identified along with positions for centromerons. This de novo assembly demonstrates the use of an approach of dual references, i.e. when a target genome (red deer) has already a pre-established genetic map, and is combined with the well-established whole genome sequence of a closely related species (cattle).

The reference genome CerEla1.0 of *Cervus elaphus hippelaphus* and its annotation, in accordance with new data from other programmes, is under continuous monitoring and updating. If the sequence data of a SNP-based map markers will be available the updating of CerEla1.0 will be possible using the approach described in this work. A large number of SNP/heterozygotic sites were identified (2.8x10\(^6\) SNVs, 3.6x10\(^5\) indels) and aligned to the deer pseudochromosomes. 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 as well as in a wider neighborhood of ruminants and Pecora. CerElal.0 also provides a source for chromosome-specific microsatellite sets, which may shed light on inbreeding/outbreeding, help in the identification of gene introgressions, of descents for autosomal, maternal and paternal lineages, forensic identification, or defining allelic compositions behind of phenotypes important, for example, in game management. The exploration of the genetic component of record antlers becomes possible by Genome Wide Association Studies. The applications and use in several fields of medical research (e.g. bone and osteoporosis research, organ development and regeneration, robust tissue proliferation/tumor biology) are also feasible. Data availability: the raw reads have been deposited into the SRA database (SRR4013902). The reference genome sequence has been submitted to the NCBI database and can be accessed at the accession number MKHE00000000. The gene annotation and the variation tracks are available for browsing and downloading from the JBrowse web page http://embo.abc.hu/wonderdeer/JBrowse.

References