University of Central Lancashire
UCLan

Central Lancashire Online Knowledge (CLoK)

Title	An announcement of a new genome sequence available for Dama dama (fallow deer)
Type	Article
URL	https://clok.uclan.ac.uk/id/eprint/49259/
DOI	https://doi.org/10.1016/j.fsiae.2023.100074
Date	2023
Citation	Barnard, Rebecca, Smith, Judith Alexis, Yuan, Na, Liu, Fan and Hadi, Syed Sibte (2023) An announcement of a new genome sequence available for Dama dama (fallow deer). Forensic Science International: Animals and Environments, 4. ISSN 2666-9374
Creators	Barnard, Rebecca, Smith, Judith Alexis, Yuan, Na, Liu, Fan and Hadi, Syed Sibte

It is advisable to refer to the publisher's version if you intend to cite from the work.
https://doi.org/10.1016/j.fsiae.2023.100074

For information about Research at UCLan please go to http://www.uclan.ac.uk/research/

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the http://clok.uclan.ac.uk/policies/

Forensic Science International: Animals and Environments 4 (2023) 100074

Contents lists available at ScienceDirect

Forensic Science International: Animals and Environments

journal homepage: www.sciencedirect.com/journal/forensic-science-international-animals-and-environments

An announcement of a new genome sequence available for *Dama dama* (fallow deer)

Rebecca K. Barnard [a,*], Judith A. Smith [a], Na Yuan [b], Fan Liu [c], Syed Sibte Hadi [c]

[a] *University of Central Lancashire, UK*
[b] *Beijing Institute of Genomics, UK*
[c] *Naif Arab University for Security Sciences, UK*

ARTICLE INFO

Keywords:
Dama dama
Fallow deer
Whole Genome Sequencing (WGS)
Chromosome
Next Generation Sequencing (NGS)
PacBio
Genome

ABSTRACT

This publication presents the genome for a male individual of *Dama dama* (Fallow deer). The Genome is 3108,385,535 bp in length with the sequence assembled in to 35 Chromosomes including an assembly of the X and Y allosomes. The final assembly also had a BUSCO completeness score of 96.4%. The annotated genome was found to consist of 22,619 genes, which are currently mostly all protein coding genes. The final BUSCO v4.1.4 completeness score for the annotation was 72%. Predicted genes have an average transcript length of 49,000 bp. On average each gene has 10 exons. Overall, the number of identified genes appeared lower than that expected for mammals (typically 20–30 K), however, these results can be expected due to the lack of additional RNA-seq data at this time, therefore not all non-coding genes have yet to be annotated. This project is the first to fully sequence the genome of a male *Dama dama* to reference genome level and provides the groundwork for further research involving this species. The result of this project not only aids to increase knowledge of the *Dama dama* genetic history but also benefits the welfare of the *Dama dama* species across the world, contributing to the global effort to combat wildlife crime.

Taxonomy

As according to Linnaeus in 1758, the Linage for *Dama dama* is as follows: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Artiodactyla; Ruminantia; Pecora; Cervidae; Cervinae; *Dama*; *Dama dama* (NCBI:txid30532) [18].

Background

In wildlife conservation genetics, reference genomes are highly sort after and are considered the gold standard (when undertaken using voucher specimens) from which to conduct a wide range of population, disease related and forensic studies. We are currently in the middle of a sixth mass extinction event [5], it is estimated that during the next few decades over one million animal and plant species are at risk of extinction demonstrating that conservation efforts are now more essential than ever [4]. Only a few of the endangered species listed on the IUCN as threatened have utilised the powerful technology of whole genome sequencing. It is estimated that only 1% of the 13,505 species listed as threatened have a fully sequenced genome available [4]. The shortage of reference genomes available is often due to the lack of understanding surrounding their potential. However, expertise isn't always the primary cause, lack of funds greatly impedes the ability to generate reference genomes, which is not often available to small research laboratories and conservation teams [4,9]. On NCBI there are more than 1800 animal genomes currently available, however only 6% of these are associated to threatened and endangered species listed on the IUCN Red List. However, the Wellcome Sanger Institute Darwin Tree of Life programme aims to change this by sequencing approximately 70,000 species genomes from Britain and Ireland [25]. Developing a reference genome which is capable of determining aspects of a species evolutionary past and biology as well as identifying variants, long repetitive regions, indels and mononucleotide regions (which can create gaps if sequencing resolution is low) is a costly task and often resides within institutions where considerable funding is available. Due to the enormous task the Darwin Tree of Life project has undertaken, it may be years until they sequence the *Dama dama* genome, despite the option to suggest species for sequencing. The *Dama dama* genome sequencing should have been made a priority previously due to interest in their genetic diversity, their role in deer ecology in Europe and their

* Corresponding author.
E-mail address: rbarnard@uclan.ac.uk (R.K. Barnard).

https://doi.org/10.1016/j.fsiae.2023.100074
Received 25 January 2023; Received in revised form 20 September 2023; Accepted 4 October 2023
Available online 5 October 2023
2666-9374/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

increasing involvement in wildlife crime. Therefore, due to the current demand for the *Dama dama* genome, this project sourced internal funding from the University of Central Lancashire to sequence the genome, at Edinburgh Genomics, ensuring significant coverage to achieve a full genome assembly.

Reference genomes are highly important in conservation genetics and play a huge role in conservation, offering insights into species biology and evolution [4]. A reference genome for *Dama dama* would be highly beneficial. During previous studies on the *Dama dama* species, the genome has shown to be highly conserved [16], however with the availability of a reference genome this can be investigated further. Many studies have been conducted into *Dama dama* via the Dama International Fallow Deer Project, however most of these studies have been archaeological based using Isotope analysis of bones to determine age and geographical origin [15,24,2,7]. Little has been done with Genomic DNA and the *Dama dama* genome was not fully sequenced before this work. A reference genome for *Dama dama* makes many research themes possible. Many future studies will be possible, including geographical assessment [24] and accurate species determination, the latter is particularly important in a forensic context [24], for example when determining ingredients in food products [13] or establishing poached trophy origin [7]. Geographical assessment via genetics is not only forensically applicable but will help in answering many questions linked to the historical origin of *Dama dama* [2]. Furthermore, many areas in conservation genetics can be studied utilising the reference genome including: effective population size (Ne), hybridization, population substructure, kinship, evolutionary history, population connectivity, local adaptation, adaptive genetic variation, and inbreeding [9]. The reference genome can be used to compare the European *Dama dama* to its close relative Persian *Dama dama*, which has often caused debate as to whether it is a true sub-species or a separate species in its own right. Understanding the genetics of different geographical populations of *Dama dama* will also give a more informed approach to herd management for use in deer parks across the UK and Europe. The reference genome can be utilised by scientists across the world in many different research themes involving *Dama dama*, as little is currently known about the species of *Dama dama*; publication of the *Dama dama* genome aims to change that.

The sample selected for genome sequencing originated from Richmond Park in London which is part of the Royal Parks organisation. Richmond Park, originally a royal hunting ground, is home to ~315 *Dama dama* [23]. The herd was initially introduced by Charles I in 1637 as a status symbol and for hunting purposes [23]. The deer sampled for this project was part of the wild herd and was sampled during part of the 2021–2022 cull season by Tony Hatton, head keeper at Richmond Park. Due to the large, long established herd of the *Dama dama* in Richmond Park, the specimen used for this project can be considered a true *Dama dama* and the sequence produced as a result is as true to the species as possible. The deer was a young male Prickett with distinctive *Dama dama* markings and unbranched antlers, see Fig. 1. This long-standing herd is well maintained by Richmond park, due to the long line of heredity in the well-established herd of *Dama dama* at Richmond, it made for an ideal sampling location for the genome sequencing.

Genome sequence report

The genome was sequenced from a sample of muscle taken from a male *Dama dama* individual which was collected from Richmond Royal Park in London UK (Latitude 51.443225, longitude −0.27042112). 36x coverage was achieved using PacBio® Long-Read sequencing, using the SMRTBell™ barcoded adapter on the Sequel IIe system. The genome was assembled using HiFiasm (v0.16.1-r375) [6]. QUAST (v5.0.2) was used to generate the assembly validation metrics [10] and BUSCO (v5.3.0) was used on the assembly generated in order to assess their quality in terms of gene completeness based on the mammalian lineage [19]. The final genome assembly has a total length of 3108 Mb in a total of 471 contigs which were assembled into 72 chromosomal scaffolds, see Table 1. The final assembly also had a BUSCO completeness score of 96.4%.

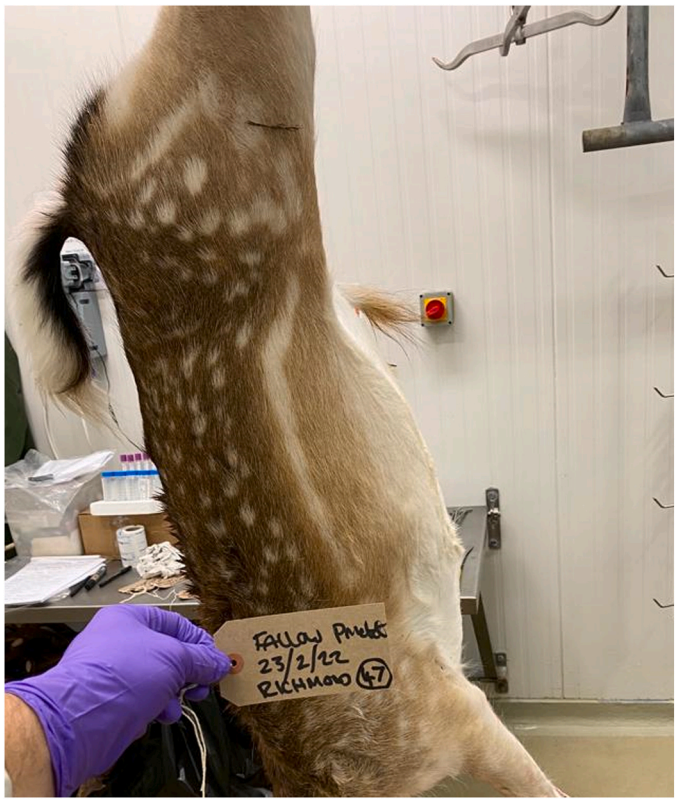

Fig. 1. Image of the male *Dama dama,* from which had a muscle sample taken from it for sequencing.

Genome annotation report

The annotated genome was found to consist of 22,619 genes. Statistically, the average mRNA length was 49.1 kb and 10.7 exons per transcript. The final BUSCO v4.1.4 completeness score for the annotation was C:74.4%[S:72.8%,D:1.6%],F:3.8%,M:21.8%,n:9226.

Genome comparison report

OrthoFinder (v2.2.6) was used to do phylogenetic-orthology inference for comparative genomics (total of eight genomes including *Dama dama*) and mapped to a phylogenetic tree by FigTree (v1.4.3). The seven genomes used to compare to the *Dama dama* genome were: elk *Cervus canadensis*), red deer (*Cervus elaphus*), white-tailed deer (*Odocoileus virginianus texanus*), Reeves' muntjac (*Muntiacus reevesi*), Central European red deer (*Cervus elaphus hippelaphus*), Yarkand deer (*Cervus hanglu yarkandensis*) and muntjac, (*Muntiacus muntjak*). Results of the genome comparison phylogenetic tree can be seen below in Fig. 2.

As expected, the four Cervus genus deer species show to be very closely related, situated on the same node of the tree and inline on the same short branch. As are the two species of muntjac which share a branch albeit on a separate node from the rest of the species. This suggests the inter genus evolutionary distance is low for these species showing a strong relationship between them. The White-tailed deer is the most unrelated, situated on its own longer branch near the root of the tree, this suggests this species has the largest evolutionary distance between it and the other genomes. *Dama dama* is showing to also be a species with low phylogenetic relation to the other species, situated on its own branch. Nonetheless, *Dama dama* does present some relation to

Table 1
Genome data for *Dama dama*, Ldn47.

Project data accession numbers	
Species	*Dama dama*
Specimen	Ldn47
NCBI taxonomy ID	ID_30532
BioProject	PRJNA905851
BioSample ID	SAMN31886425
Isolate information	Male, Muscle Tissue
Raw Data Sequence Read Archive (SRA) database accession number	
PacificBiosciences SEQUEL II	SRR24718831
Genome assembly	
Assembly accession	JASJUW000000000
Span (Mb)	3108
Number of contigs	471
Contig N50 length (Mb)	68
Number of scaffolds	72
Scaffold N50 length (Mb)	68
Longest scaffold (Mb)	132
BUSCO genome score	C:96.4%[S:93.0%,D:3.4%],F:0.9%,M:2.7%,n:9226
Genome annotation	
Number of genes	22,619
Average length of gene (bp)	49,098.65
Average number of exons per transcript	10.7
Average exon size (bp)	172.83
Average intron size (bp)	4871.47
BUSCO annotation score	C:74.4%[S:72.8%,D:1.6%],F:3.8%,M:21.8%,n:9226

C= complete [S= single copy, D=duplicated], F=fragmented, M=missing, n = number of orthologues in comparison.

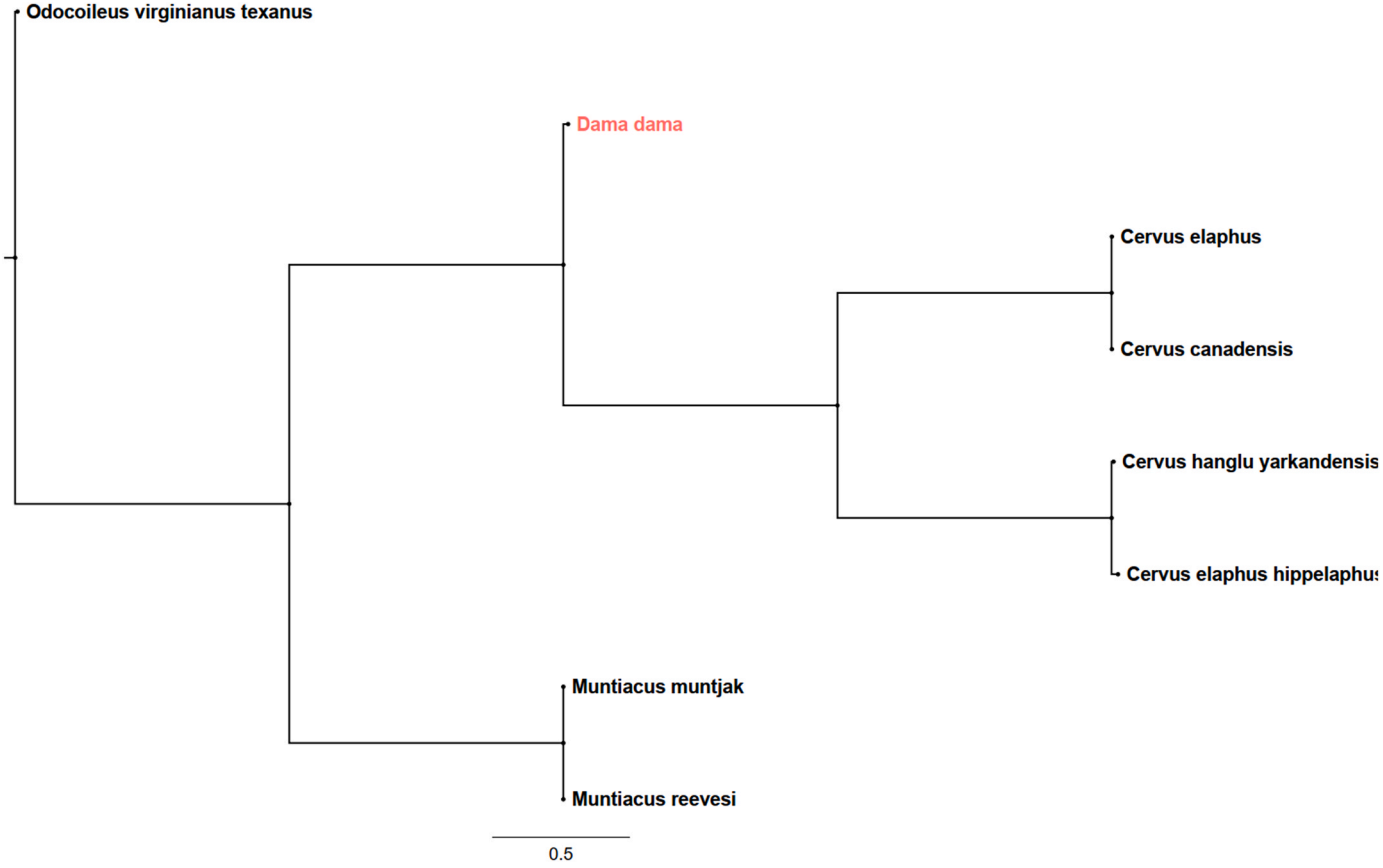

Fig. 2. Phylogenetic tree visualized by using FigTree (v1.4.3) depicting the results of the phylogenetic-orthology inference of the *Dama dama* genome along with seven other species.

the four Cervus species, with the node of the tree branching down to Cervus, which is interesting as *Dama dama* live closely in Britain with *Cervus elaphus*. As depicted by the phylogenetic tree, the evolutionary distance between *Dama dama* is high, but not as high as the white tailed deer. As previous studies suggested [12,16,17], *Dama dama* have low genetic relationship to other species of deer, this comparison study supports their findings, but shows a new evolutionary relationship to the Cervus genus.

Methods

Sample Collection, DNA Extraction, Species ID and Genome Sequencing.

A sample of muscle from a male *Dama dama* individual was collected from Richmond Royal Park in London UK (Latitude 51.443225, longitude −0.27042112) by Tony Hatton, Head Keeper, Richmond Park.

DNA extraction from muscle was performed using the Gentra® Puregene® tissue kit, following the protocol: DNA purification from muscle tissue using the Gentra Puregene tissue kit page 19–21 of the Gentra® Puregene® Handbook (QIAGEN, Germany). However, the following changes were made to the Gentra® Puregene® Tissue Kit protocol, to maximise yield and quality of DNA extracted:

Step 1: Pestle and mortar was used rather than liquid nitrogen.

Step 2: 1.5 ml cell lysis solution and 1.5 ml of tissue lysis solution used rather than just 3 ml cell lysis.

Step 2b: 30 μl Puregene proteinase K and 15 μl of DTT used rather than just 15 μl Puregene proteinase K. Incubate for 4 h.

Step 3: 30 μl RNase A solution used rather than 15 μl.

Step 5: 1.5 ml protein precipitation solution used rather than just 1 ml with an additional 10 min incubation on ice.

Step 6: Centrifugation speed was increased to 7000 x g.

Step 9 and 12: Centrifugation speed was increased to 4000 x g.

Step 13: Air dry overnight.

Step 15: Incubation at 65 °C was done for 30 min rather than 1 h.

Species ID was carried out prior to sending the sample for sequencing. PCR amplification was carried out using the COI universal primer set LCO1490: 5'-ggtcaacaaatcataaagatattgg-3' HC02198: 5'-taaacttcagggtgaccaaaaaatca-3' [8]. PCR amplification was conducted in 15 μl reactions containing 1X Platinum® PCR multiplex master mix (Life Technologies, USA), 25 mM $MgCl_2$, 10 pmol primer, 3 ng DNA, and PCR grade water was used to bring the reaction up to the final volume. The reaction was carried out in a Veriti® thermal cycler (Applied Biosystems™, USA) under the following conditions: initial denaturation for 15 min at 95 °C; 40 cycles of 30 s at 94 °C, 30 s at 50 °C, and 1 min at 72 °C; followed by a final extension for 15 min at 72 °C. Cycle sequencing was performed using the BigDye™ Terminator v3.1 cycle sequencing kit (ThermoFisher Scientific, United States) at half reaction volume using the forward LCO1490 primer only (3.2 pmol). The 3500 genetic analyser (ThermoFisher Scientific, United States) was used for DNA sequencing of the COI gene. The purified and dried sequencing reactions were resuspended in 11 μl of Hi-Di™ Formamide. A 50 cm capillary array was used, the injection time was 8 s, and the run-time was 5000 s. Sequence Analysis v.6 software was used to analyse and interpret the data. The generated sequences were compared to previously uploaded sequences on the NCBI GenBank database via the BLAST search function. Species identifications were derived from this information based upon the percentage match identities to ensure the sample was *Dama dama*.

PacBio® SMRTBell™ sequencing libraries were constructed according to the manufacturer's protocol by Edinburgh Genomics. The BluePippin automated size selector was used to size select the DNA after shearing on The Megaruptor® according to the manufacturer's guidelines by Edinburgh Genomics. Sequencing was carried out by Edinburgh Genomics in their facility at The University of Edinburgh via PacBio® long-read sequencing, using the SMRTBell™ barcoded adapter on the Sequel IIe system.

Genome assembly

The genome was assembled using HiFiasm (v0.16.1-r375) [6]. QUAST (v5.0.2) was used to generate the assembly validation metrics [10] and BUSCO (v5.3.0) was used on the assembly generated in order to assess their quality in terms of gene completeness based on the mammalian lineage [19].

Using the software RagTag [1], the assembled original genome sequence was re-mounted at the chromosome level (based on the results of RagTag, 80 contigs were successfully mounted), the reference genome used for mounting the Chromosomes was the Cervus elaphus genome (33 chr+X) (https://www.ncbi.nlm.nih.gov/genome/?term=-txid9860 [orgn]) and the Cervus canadensis genome (Y chromosome) (https://www.- ncbi.nlm.nih.gov/genome/34916). The MT genome was mounted to the *Dama dama* MT genome already available (https://www.ncbi.nlm.nih.gov/nucleotide/NC_-020700.1).

Genome annotation

A repeat-masked genome was firstly achieved using RepeatMasker [21] and the Repbase library [3]. Gene Mapping was undertaken using the repeat-masked genome, two approaches were combined to predict genes. The first approach relies on ab initio prediction. Two different software packages were used including AUGUSTUS [22] and Glimmer-HMM [14]. The second approach relies on homolog sequence comparison, where the homolog protein sequences were downloaded from two published homolog species including Cervus elaphus (https://www.ncbi.nlm.nih.gov/genome/?term=txid9860[orgn]) and *Dama dama* (https://www.ncbi.nlm.nih.gov/protein/?term=Dama+dama). For genome annotation, these protein sequences were aligned to the assembly of the *Dama dama* genome using the software package, Exonerate [20]. The gene sets predicted by the two approaches were integrated using software package EVM [11] to produce weighted consensus gene models.

Ethical/compliance issues

Appropriate Ethical clearance was obtained from The Animal Welfare and Ethics Review Board at the University of Central Lancashire, for the collection of the sample from the *Dama dama* species and the undertaking of the project.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

NCBI Accession Number: JASJUW000000000.

Acknowledgements

Sequencing and bioinformatics analysis was carried out by Edinburgh Genomics, The University of Edinburgh, which is partly supported with core funding from the Natural Environment Research Council (UKSBS PR18037). Funding for the Genome sequencing was granted by the University of Central Lancashire.

References

[1] M. Alonge, L. Lebeigle, M. Kirsche, S. Aganezov, X. Wang, Z.B. Lippman, M. C. Schatz, S. Soyk, Automated Assembly scaffolding elevates a new tomato system for high-throughput genome editing, BioRxiv Prepr. Serv. Biol. (2021), https://doi.org/10.1101/2021.11.18.469135.

[2] K.H. Baker, H.W. Gray, V. Ramovs, D. Mertzanidou, Ç. Akın Pekşen, C.C. Bilgin, N. Sykes, A.R. Hoelzel, Strong population structure in a species manipulated by humans since the neolithic: The European fallow deer (*Dama dama dama*), Heredity 119 (2017) 16–26, https://doi.org/10.1038/hdy.2017.11.

[3] W. Bao, K.K. Kojima, O. Kohany, Repbase update, a database of repetitive elements in eukaryotic genomes, Mob. DNA (2015) 6, https://doi.org/10.1186/s13100-015-0041-9.

[4] Peel Brandies, Belov Hogg, The value of reference genomes in the conservation of threatened species, Genes 10 (2019) 846, https://doi.org/10.3390/genes10110846.

[5] G. Ceballos, P.R. Ehrlich, A.D. Barnosky, A. García, R.M. Pringle, T.M. Palmer, Accelerated modern human–induced species losses: Entering the sixth mass extinction, Sci. Adv. (2015) 1, https://doi.org/10.1126/sciadv.1400253.
[6] H. Cheng, G.T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de Novo Assembly using phased assembly graphs with hifiasm, Nat. Methods 18 (2021) 170–175, https://doi.org/10.1038/s41592-020-01056-5.
[7] Di Giovanni, G., 2020. Fallow deer DNA marker research [WWW Document]. Australian Deer Association. URL https://www.austdeer.com.au/2020/05/19/1380477/fallow-deer-dna-marker-research (accessed 09.17.23).
[8] O. Folmer, M. Black, W. Hoeh, R. Lutz, R. Vrijenhoek, DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates, Mol. Mar. Biol. Biotechnol. 3 (1994) 294–299.
[9] A.P. Fuentes-Pardo, D.E. Ruzzante, Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations, Mol. Ecol. 26 (2017) 5369–5406, https://doi.org/10.1111/mec.14264.
[10] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, Quast: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, https://doi.org/10.1093/bioinformatics/btt086.
[11] B.J. Haas, S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C. R. Buell, J.R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments, Genome Biol. (2008) 9, https://doi.org/10.1186/gb-2008-9-1-r7.
[12] G.B. Hartl, A. Schleger, M. Slowak, Genetic variability in fallow deer, *Dama dama* L, Anim. Genet. 17 (2009) 335–341, https://doi.org/10.1111/j.1365-2052.1986.tb00726.x.
[13] E. Haynes, E. Jimenez, M.A. Pardo, S.J. Helyar, The future of NGS (next generation sequencing) analysis in Testing Food Authenticity, Food Control 101 (2019) 134–143, https://doi.org/10.1016/j.foodcont.2019.02.010.
[14] W.H. Majoros, M. Pertea, S.L. Salzberg, Tigrscan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders, Bioinformatics 20 (2004) 2878–2879, https://doi.org/10.1093/bioinformatics/bth315.
[15] Miller, H., 2013. Dama international- the fallow deer project by Holly Miller [WWW Document]. Britgeopeople.blogspot.com. URL https://britgeopeople.blogspot.com/2013/04/dama-international-fallow-deer-project.html (accessed 6.17.22).
[16] J.M. Pemberton, R.H. Smith, Lack of biochemical polymorphism in British fallow deer, Heredity 55 (1985) 199–207, https://doi.org/10.1038/hdy.1985.92.
[17] E. Randi, M. Apollonio, Low biochemical variability in European fallow deer (*Dama dama* L.): natural bottlenecks and the effects of domestication, Heredity 61 (1988) 405–410, https://doi.org/10.1038/hdy.1988.131.
[18] C.L. Schoch, S. Ciufo, M. Domrachev, C.L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J.P. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, NCBI taxonomy: A comprehensive update on curation, resources and Tools, Database 2020 (2020), https://doi.org/10.1093/database/baaa062.
[19] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, Busco: Assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, https://doi.org/10.1093/bioinformatics/btv351.
[20] G.S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, BMC Bioinforma. (2005) 6.
[21] Smit, A.F.A., Hubley, R., Green, P., 2015. [WWW Document]. RepeatMasker. URL http://www.repeatmasker.org/ (accessed 09.07.22).
[22] M. Stanke, R. Steinkamp, S. Waack, B. Morgenstern, Augustus: A web server for gene finding in Eukaryotes, Nucleic Acids Res. (2004) 32, https://doi.org/10.1093/nar/gkh379.
[23] The Royal Parks, 2022. Richmond Park Deer: Richmond Park [WWW Document]. Deer in Richmond Park. URL https://www.royalparks.org.uk/parks/richmond-park/things-to-see-and-do/wildlife/deer-in-richmond-park#:~:text=Richmond%20Park%20is%20a%20National,have%20shaped%20the%20landscape%20too. (accessed 06.16.22).
[24] The British Deer Society (BDS), 2016. Fallow deer [WWW Document]. Bds.org.uk. URL https://bds.org.uk/information-advice/issues-with-deer/poaching-and-wildlife-crime/ (accessed 17.09.23).
[25] Wellcome Sanger Institute, 2019. [WWW Document]. Darwin Tree of Life Project. URL https://www.sanger.ac.uk/collaboration/darwin-tree-of-life-project/ (accessed 06.16.22).