



UNIVERSITAT DE
BARCELONA

**Novel bioinformatic methods
for the metagenomic analysis of non-invasive
samples from the endangered species
Galemys pyrenaicus: monitoring population
genomics, ecology, and health**

Román Sapino Funes

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

**NOVEL BIOINFORMATIC METHODS
FOR THE METAGENOMIC ANALYSIS
OF NON-INVASIVE SAMPLES
FROM THE ENDANGERED SPECIES
GALEMYS PYRENAICUS:
MONITORING POPULATION
GENOMICS, ECOLOGY, AND HEALTH**

ROMÁN SAPINO FUNES

Doctoral thesis | January 2025



UNIVERSITAT DE
BARCELONA



UNIVERSITAT DE BARCELONA

Tesi Doctoral

Novel bioinformatic methods for the metagenomic analysis of non-invasive samples from the endangered species *Galemys pyrenaicus*: monitoring population genomics, ecology, and health

Memòria presentada per:

Román Sapino Funes

Per optar al grau de

Doctor per la Universitat de Barcelona

Programa de Doctorat "Genètica"

Treball realitzat a l'Institut de Biologia Evolutiva (CSIC-UPF)

Director: José Castresana Villamor

Tutor: Julio Rozas Liras

Doctorand: Román Sapino Funes

Barcelona, gener del 2025

**NOVEL BIOINFORMATIC METHODS FOR THE METAGENOMIC
ANALYSIS OF NON-INVASIVE SAMPLES FROM THE
ENDANGERED SPECIES *GALEMYS PYRENAICUS*:
MONITORING POPULATION GENOMICS, ECOLOGY, AND HEALTH**

This work was supported by the grant PRE2018-085160, funded by MICIU/AEI /10.13039/501100011033, as well as the research projects PID2020-113586GB-I00, funded by MICIU/AEI /10.13039/501100011033, CGL2017-84799-P, funded by MICIU/AEI /10.13039/501100011033 and by FEDER 'Una manera de hacer Europa,' and TED2021-130149B-I00, funded by the European Union NextGenerationEU/PRTR.



ACKNOWLEDGEMENTS

La finalización de esta tesis no es un fin, sino un inicio que, a su vez, es una continuación del camino que comenzó con mis lecturas de enciclopedias de ciencias naturales antes de ir a dormir. Este tramo que llamamos tesis doctoral ha sido una etapa de enorme crecimiento profesional y científico, además de una fuente de inspiración para continuar. Al mismo tiempo, ha representado un profundo crecimiento personal, un proceso de maduración al otro lado del océano. Pero lo mejor de todo es que este recorrido no lo hice solo, sino junto a muchas personas que me acompañaron a lo largo de estos años.

En primer lugar, quiero expresar mi más sincero agradecimiento a mi director de tesis, el Dr. José Castresana, por brindarme la oportunidad de desarrollar esta tesis en su grupo, así como por las incontables horas de reuniones que hicieron posible el avance de ésta y que contribuyeron enormemente a mi formación. Agradezco también a Julio Rozas por su papel como tutor y su ayuda con las gestiones universitarias.

Además, quiero agradecer especialmente a Alfonso y Lidia por haberme acogido en el grupo, por su ayuda y compañía durante los primeros años de la tesis. También a mis compañeras y compañeros del IBE, y en particular a quienes formaron y forman parte del Programa de Biodiversidad, por los años compartidos, que hicieron las horas de trabajo más divertidas. Asimismo, quiero reconocer a los miembros de la administración del IBE por su trabajo y disponibilidad constante.

Por último, quiero dedicar esta tesis a mi familia, por su apoyo incondicional de ahora y siempre. Y a las personas que conocí, ya sea por primera o por segunda vez, en esta etapa de mi vida en Barcelona, quienes también se convirtieron en familia.

Graphic design

Ferran Mitjans Sanz

Cover & Inner cover image

The Iberian desman (*Galemys pyrenaicus*) in the Elama Stream. Photo by Jorge González–Esteban, from Esnaola, González-Esteban, et al. (2018).

Title page illustration

Figure "Défécation" from *Le Desman des Pyrénées: Un mammifère inconnu à découvrir* by Bernard Richard, 1986. Part of the Science et Découvertes collection, published by Le Rocher.

**NOVEL BIOINFORMATIC METHODS
FOR THE METAGENOMIC ANALYSIS
OF NON-INVASIVE SAMPLES
FROM THE ENDANGERED SPECIES
GALEMYS PYRENAICUS:
MONITORING POPULATION
GENOMICS, ECOLOGY, AND HEALTH**



ROMÁN SAPINO FUNES

Doctoral thesis | January 2025

INDEX OF CONTENTS

I. ABSTRACT	13
RESUM	15
RESUMEN	17
II. INTRODUCTION	19
1. GENOMICS AND ITS APPLICATION IN CONSERVATION BIOLOGY	20
1.1. Reference genomes	21
1.2. Whole-genome re-sequencing	21
1.3. Reduced representation sequencing	22
1.4. Metabarcoding	22
1.5. Metagenomics	23
2. NON-INVASIVE SAMPLES IN GENETIC AND GENOMIC MONITORING OF WILDLIFE	24
2.1. Feces	25
2.2.1. Metagenomics analyses from feces	28
3. METHODS AND DATABASES IN METAGENOMIC ANALYSIS	32
3.1. Host DNA filtering	32
3.2. Assessing taxonomic diversity	32
3.2.1. Reference-based taxonomic profiling	32
3.2.2. Assembly-based methods	35
3.2.3. Methods for species-level pathogen identification	36
3.3. Databases resources	37
4. THE IBERIAN DESMAN (<i>Galemys pyrenaicus</i>)	40
4.1. Morphology adapted to aquatic habitats	40
4.2. Sexual characters, sex identification and reproduction	41
4.3. Distribution	42
4.4. Evolutionary history and phylogeographic structure	42
4.5. Current conservation status	45
4.6. Habitat and trophic ecology	46
4.6.1. Genetic approaches in the study of the diet of the Iberian desman	48
4.7. Parasites and pathogens	50
III. OBJECTIVES	55
IV. METHODS	59
1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES	59
1.1. Fecal samples collection and DNA extraction	59
1.2. Shotgun-metagenomics library construction and sequencing	60
1.3. Sequencing data processing: quality control and separation of endogenous and exogenous reads	60
1.4. Genomics analysis from endogenous data	63
1.4.1. Population structure analysis	63
1.4.2. Genetic sex determination	64
1.4.3. Mitogenome phylogenetic reconstruction	64
1.5. Metagenomics analysis from exogenous data: overall DNA composition, diet and parasites	64
1.5.1. Overall DNA composition in the feces	64
1.5.2. Databases for diet and parasites identification	65
1.5.3. Taxonomic classification of consensus sequences from diet and parasites	66

NOVEL VASIVE MONITORING	BIOINFORMATIC SAMPLES	METHODS FROM THE POPULATION	FOR THE ENDANGERED GENOMICS,	METAGENOMIC SPECIES ECOLOGY,	ANALYSIS OF GALEMYS AND	OF NON-IN-PYRENAICUS: HEALTH
2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS	72					
2.1. Reference bacterial genomes	72					
2.2. Alignment of sequences to reference genomes of <i>Yersinia</i> species	72					
2.3. Alignments to reference genomes of pathogenic bacterial species	73					
V. RESULTS	75					
1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES	75					
1.1. Performance of metagenomic libraries and illumina sequencing from fecal samples	75					
1.2. Total DNA composition in fecal samples	75					
1.3. Sex determination	77					
1.4. Population structure analysis	78					
1.5. Phylogenetic Tree of the mitochondrial genomes	79					
1.6. Taxonomic classification of diet and parasites	80					
1.6.1. Classification of diet components	81					
1.6.2. Identification of parasites	84					
1.6.3. Identification of consensus sequences at the species level	86					
2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS	89					
2.1. Species-level identification of <i>Yersinia</i> species	89					
2.2. Verification of the identification criterion by comparison with other mapping tools	91					
2.3. Application to the detection of pathogenic bacterial species	92					
2.4. Overall health assessment of the Iberian desman population	95					
VI. DISCUSSION	97					
1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES	98					
1.1. DNA composition of fresh fecal samples	98					
1.2. Use of endogenous DNA in population genomic studies	99					
1.3. New method for the classification of diet elements and parasites using metagenomic data	100					
1.4. Taxonomic identification parameters	101					
1.5. Assessment of the diet using metagenomic data	102					
1.6. Identification of parasites using metagenomic data	105					
2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS	106					
2.1. Validation of a mapping methodology for the detection of pathogens from fecal samples	106					
2.2. Implications for the conservation of the Iberian desman	108					
2.3. Method challenges and prospects	110					
3. NOVEL APPROACHES FOR WILDLIFE CONSERVATION: FECAL METAGENOMICS FOR MONITORING POPULATION GENOMICS, ECOLOGY AND HEALTH	111					
3.1. Novel metagenomic methods developed in this thesis	111					
3.2. From the Iberian desman to broader conservation: applications of the developed methods	112					
VII CONCLUSIONS	117					
VIII REFERENCES	121					
IX ANNEX	155					

I. ABSTRACT

The use of genomic methods has become an important tool in conservation programs for endangered species. These methods have been used mainly to study population genomics, but they have the potential to provide additional information, especially about the trophic ecology and health of individuals and species. One promising way to expand conservation genomics methods is to analyze fecal samples using metagenomic techniques. Fecal samples can be collected with minimal disturbance to the animals, and they can provide important insights into diet, gut microbiome, and pathogens. Combining these samples with modern bioinformatic tools could help to learn critical aspects for the conservation of endangered species. However, metagenomic analysis in non-model species is a major challenge due to the lack of reference genome sequences or data limited to distantly related species. Fresh fecal samples of the Iberian desman (*Galemys pyrenaicus*) were used to construct metagenomic libraries, from which it was possible to analyze the overall composition of the DNA present in the feces. Using endogenous nuclear DNA, population structure was analyzed through Principal Components Analysis (PCA), while endogenous mitochondrial genomes were also reconstructed. For the analysis of exogenous DNA of eukaryotic origin, a novel bioinformatic method was developed that uses assembled consensus sequences for mapping instead of individual reads, which allowed improving classifications using various reference databases. This method facilitated the characterization of diet composition and the identification of parasites. Additionally, it enabled the estimation of relative abundances through depth of coverage. For the exogenous DNA of bacterial origin, a specific method was developed to detect potentially pathogenic bacteria at the species level from metagenomic sequences by mapping unassembled reads to complete reference genomes. The approach used the breadth of coverage rather than the number of mapped reads for species identification, thereby minimizing false positives due to conserved or repetitive genomic regions. The analysis revealed the presence of 19 potentially pathogenic bacterial species, with prevalences ranging from a single individual to 30% of the samples. Desmans with elevated or altered pathogen loads were detected, suggesting differences in individual health status or variations in exposure to environmental factors. Application of this method across populations and over time for endangered species may provide essential health and epidemiological information to improve conservation strategies.

Keywords: Metagenomics, Bioinformatics, Non-invasive samples, *Galemys pyrenaicus*

Nous mètodes bioinformàtics per a l'anàlisi metagenòmica de mostres no invasives de l'espècie en perill d'extinció *Galemys pyrenaicus*: monitorització de la genòmica de poblacions, ecologia i salut



RESUM

L'ús de mètodes genòmics s'ha convertit en una eina important en programes de conservació d'espècies amenaçades. Aquests mètodes s'han utilitzat principalment per estudiar la genòmica de poblacions, però tenen el potencial de proporcionar informació addicional, especialment sobre l'ecologia tròfica i la salut dels individus i les espècies. Una manera prometedora d'expandir els mètodes de genòmica de conservació és analitzar mostres fecals fent servir tècniques metagenòmiques. Les mostres fecals es poden recollir amb una mínima pertorbació per als animals, i poden proporcionar informació important sobre la dieta, el microbioma intestinal i els patògens. La combinació d'aquestes mostres amb eines bioinformàtiques modernes podria ajudar a aprendre aspectes crítics per a la conservació d'espècies amenaçades. No obstant això, l'anàlisi metagenòmica en espècies no model és un repte important a causa de la manca de seqüències genòmiques de referència o de dades limitades a espècies distants. Es van utilitzar mostres fecals fresques de l'almesquera (*Galemys pyrenaicus*) per construir biblioteques metagenòmiques, a partir de les quals es va poder analitzar la composició general de l'ADN present en les femtes. Fent servir ADN nuclear endogen, l'estructura poblacional es va analitzar a través de l'Anàlisi de Components Principals (PCA), mentre que els genomes mitocondrials endògens també es van reconstruir. Per a l'anàlisi de l'ADN exogen d'origen eucariota, es va desenvolupar un nou mètode bioinformàtic que fa servir seqüències de consens assemblades per al mapejat en lloc de lectures individuals, cosa que va permetre millorar les classificacions fent servir diverses bases de dades de referència. Aquest mètode va facilitar la caracterització de la composició de la dieta i la identificació de paràsits. A més, va permetre l'estimació d'abundàncies relatives a través de la profunditat de cobertura. Per a l'ADN exogen d'origen bacterià, es va desenvolupar un mètode específic per detectar bacteris potencialment patògens a escala d'espècie a partir de seqüències metagenòmiques, mitjançant el mapatge de lectures no assemblades per completar genomes de referència. L'enfocament feia servir l'amplitud de la cobertura en lloc del nombre de lectures mapades per a la identificació d'espècies, minimitzant així falsos positius a causa de regions genòmiques conservades o repetitives. L'anàlisi va revelar la presència de 19 espècies bacterianes potencialment patògenes, amb prevalences que van des d'un sol individu fins al 30% de les mostres. Es van detectar almesqueres amb càrregues de patògens elevades o alterades, suggerint diferències en l'estat de salut individual o variacions en l'exposició a factors ambientals. L'aplicació d'aquest mètode a través de les poblacions i amb el temps per a les espècies en perill d'extinció pot proporcionar informació essencial de salut i epidemiologia per millorar les estratègies de conservació.

Paraules clau: Metagenòmica, Bioinformàtica, Mostres no invasives, *Galemys pyrenaicus*

Nuevos métodos bioinformáticos para el análisis metagenómico de muestras no invasivas de la especie en peligro de extinción *Galemys pyrenaicus*: monitorización de la genómica poblacional, la ecología y la salud



RESUMEN

El uso de métodos genómicos se ha convertido en una herramienta importante en los programas de conservación de especies en peligro de extinción. Estos métodos se han utilizado principalmente para estudiar la genómica de poblaciones, pero tienen el potencial de proporcionar información adicional, especialmente sobre la ecología trófica y la salud de individuos y especies. Una forma prometedora de ampliar los métodos de genómica para la conservación es analizar muestras fecales utilizando técnicas metagenómicas. Las muestras fecales pueden recolectarse con un mínimo de perturbación para los animales y pueden aportar información clave sobre la dieta, el microbioma intestinal y los patógenos. Combinar estas muestras con herramientas bioinformáticas modernas podría ayudar a obtener datos críticos para la conservación de especies en peligro. Sin embargo, el análisis metagenómico en especies no modelo presenta un desafío importante debido a la falta de secuencias de genomas de referencia o a datos limitados a especies lejanamente relacionadas. Se utilizaron muestras fecales frescas del desmán ibérico (*Galemys pyrenaicus*) para construir bibliotecas metagenómicas, a partir de las cuales fue posible analizar la composición general del ADN presente en las heces. Mediante ADN nuclear endógeno, se analizó la estructura poblacional a través de un Análisis de Componentes Principales (PCA), mientras que también se reconstruyeron genomas mitocondriales endógenos. Para el análisis de ADN exógeno de origen eucariota, se desarrolló un nuevo método bioinformático que utiliza secuencias consenso ensambladas en lugar de lecturas individuales para el mapeo, lo que permitió mejorar las clasificaciones utilizando varias bases de datos de referencia. Este método facilitó la caracterización de la composición de la dieta y la identificación de parásitos, además de permitir la estimación de abundancias relativas mediante la profundidad de cobertura. Para el ADN exógeno de origen bacteriano, se desarrolló un método específico para detectar bacterias potencialmente patógenas a nivel de especie a partir de secuencias metagenómicas, mediante el mapeo de lecturas no ensambladas a genomas de referencia completos. Este enfoque utilizó la amplitud de cobertura en lugar del número de lecturas mapeadas para la identificación de especies, minimizando así los falsos positivos causados por regiones genómicas conservadas o repetitivas. El análisis reveló la presencia de 19 especies bacterianas potencialmente patógenas, con prevalencias que iban desde un solo individuo hasta el 30 % de las muestras. Se detectaron desmanes con cargas elevadas o alteradas de patógenos, lo que sugiere diferencias en el estado de salud individual o variaciones en la exposición a factores ambientales. La aplicación de este método en diferentes poblaciones y a lo largo del tiempo para especies en peligro puede proporcionar información esencial sobre salud y epidemiología para mejorar las estrategias de conservación.

Palabras clave: Metagenómica, Bioinformática, Muestras no invasivas, *Galemys pyrenaicus*

II. INTRODUCTION

Biodiversity is increasingly threatened due to a variety of natural and anthropogenic activities, such as habitat fragmentation and loss, overexploitation, environmental pollution, and climate change (Kumar et al., 2022). Therefore, there is an urgent need to implement effective strategies to conserve biodiversity, which is essential for maintaining healthy ecosystems and meeting society's needs for sustainable development (Díaz et al., 2015).

Numerous programs have been initiated worldwide for the conservation, protection, and management of natural resources and wildlife. This involves biodiversity monitoring, managing small populations of endangered species, restoring and enhancing the genetic diversity of target species or populations, and supporting species adaptation to a changing environment (Theissinger et al., 2023). In recent years, molecular and genomic techniques have introduced a new set of tools for biodiversity conservation plans. These techniques reveal genetic diversity and genetic variations at different levels—from populations to species and ecosystems—, facilitate the reconstruction of phylogenetic relationships to delineate species boundaries and enable the identification of ecological interactions between organisms, among other critical aspects in conservation (Bernatchez et al., 2024; Breed et al., 2019; Khan et al., 2016; Kumar et al., 2022).



Habitat of the Iberian desman. The Elama Stream, Basque Country, Photo by Amalur Eñola, from Eñola, González-Esteban, et al. (2018).

1. GENOMICS AND ITS APPLICATION IN CONSERVATION BIOLOGY

Over the past decade, advances in the field of genomics and high-throughput sequencing technologies have been applied to the study of a wide variety of organisms, including wildlife species. As a result, genomic approaches have become an important part of the toolkit for understanding the fundamentals of wildlife biology and for helping to determine management and conservation actions for wild populations and their habitats (Bernatchez et al., 2024; Hohenlohe et al., 2021; Nielsen et al., 2023; Porter & Hajibabaei, 2018; Supple & Shapiro, 2018; Theissinger et al., 2023).

Genomic applications can contribute significantly on two main fronts. On the one hand, they are widely used in population genomics studies, where they allow for the identification of individuals, conservation units, and species, describing kinship relationships, population structure, and connectivity, as well as detecting hybridization and evaluating population size changes (Formenti et al., 2022; Supple & Shapiro, 2018; Theissinger et al., 2023). Additionally, given the close relationship between genetic diversity and the fitness of individuals, as well as the adaptive potential of populations, these applications serve as a foundation for planning conservation actions aimed at maximizing population persistence in changing environments (Bertola et al., 2023). Currently, genomic approaches commonly used in population genomics research focus on identifying variable positions (single nucleotide polymorphisms, SNPs) distributed across the entire genome, based on genomic data generated through reduced-representation sequencing (RRS) or whole-genome re-sequencing (WGR) (Bertola et al., 2023). The increase in the number of markers used, compared to traditional methods, significantly enhances the statistical power and resolution of analyses. Genomic analyses also allow researchers to address new questions, such as the detection of genes associated with important traits or fitness (Hohenlohe et al., 2021).

On the other hand, genomics has become an important tool in biodiversity monitoring studies due to its potential to process large amounts of samples collected across time and space (Porter & Hajibabaei, 2018). This allows, for example, the identification of an ecosystem's taxonomic diversity, the mapping of ecological networks, the detection of invasive species, or the assessment of species turnover during ecosystem restoration (Theissinger et al., 2023). While methods such as DNA barcoding and metabarcoding are widely used in biomonitoring (Bernatchez et al., 2024; Ruppert et al., 2019), there is a growing interest in employing genomic tools in this context. In this regard, the application of metagenomic allows researchers to study which organisms are present and what the genomic potential of a community is, based on the sequencing of total DNA extracted from a complex sample, such as environmental samples (Porter & Hajibabaei, 2018). However, this method is still primarily used in studies of microbial communities, mainly due to the availability of reference genomes (Theissinger et al., 2023). Nonetheless, some studies have begun to apply this method to the monitoring of eukaryotic organisms (Curto et al., 2024).

The successful implementation of these methodologies depends on factors such as the quantity and quality of the available biological material, proficiency in laboratory and bioinformatics techniques, feasibility and costs, and the quality and quantity of reference genomes available in databases (Theissinger et al., 2023).

1.1 Reference genomes

Reference genomes, i.e., high-quality annotated assemblies, significantly enhance genomic studies by providing an accurate representation of species' genome structure and organization. By using these genomes as a guide, it is possible to identify genetic variants, such as SNPs and structural variants (Formenti et al., 2022). Additionally, reference genomes play a critical role in metagenomic studies, facilitating the taxonomic and functional classification of DNA sequences in a sample (Formenti et al., 2022). With advances in sequencing technologies, the number of chromosome-level reference genomes has increased in recent years, covering a wide range of species and establishing themselves as a key tool in conservation genomics (Corominas et al., 2024; Lewin et al., 2018; Mc Cartney et al., 2024).

1.2 Whole-genome re-sequencing

The goal of whole-genome re-sequencing (WGR) is to compare genomic variability among individuals or populations. This approach requires the availability of a reference genome for mapping sequences and identifying variants. The genomic DNA of each sample is fragmented to create a library, which is then sequenced using high-throughput sequencing technologies. The resulting reads, obtained at a given coverage, are aligned to the reference genome based on sequence similarity (Fuentes-Pardo & Ruzzante, 2017).

The main limitation of whole-genome re-sequencing is the lack of a reference genome for the species under study. While a closely related species' genome can be used to align reads or generate a preliminary assembly, this approach requires caution, as differences in genomic organization—such as variations in copy number or structural variants—may exist even among closely related species (Hohenlohe et al., 2021). Furthermore, using a reference genome from another species restricts read mapping to conserved regions between taxa, potentially reducing the ability to detect informative variation unique to the species of interest. Genomic differences between taxa can also affect the accuracy of both read mapping and SNP identification (Fuentes-Pardo & Ruzzante, 2017).

While high depth of coverage (>30X) per individual is considered necessary for accurate genotype calling (Fuentes-Pardo & Ruzzante, 2017), approaches using medium (5-20X) or low coverage (<5X) have proven to be cost-effective alternatives (Lou et al., 2021). These approaches enable whole-genome information to be obtained from a larger number of individuals with the same sequencing effort, providing estimates of population parameters (Lou et al., 2021). At these coverage levels, genotype calling at the individual level is generally avoided. Instead, methods based on the calculation of genotype likelihoods per site, using sequencing data from all individuals within probabilistic frameworks, are employed, such as ANGSD (Korneliussen et al., 2014). These methods allow most population-level questions to be addressed without requiring individual genotypes (Lou et al., 2021). Using this

approach, Martchenko & Shafer (2023) achieved comparable population estimates between WGR data from 35 samples at 9X and RADseq data from 254 individuals, applied to the study of the North American mountain goat. Furthermore, it has been employed in the study of historical and non-invasive samples with medium coverage data ranging from approximately 3-12X (Alvarez-Estape et al., 2023; Sánchez-Barreiro et al., 2021; Van Der Valk et al., 2019).

1.3. Reduced representation sequencing

In reduced representation sequencing approaches, a small fraction of the genome is reproducibly sampled in each specimen. This can be achieved using restriction endonucleases and size selection, as in restriction site-associated DNA sequencing (RADseq) (Davey et al., 2011) and related methods (Andrews et al., 2016). Among these, a particularly popular approach is double-digest RAD sequencing (ddRAD) (Peterson et al., 2012), in which DNA is simultaneously digested using two restriction enzymes and ligated to specific adapters designed for the ends generated by each enzyme, which contain a different barcode for each sample. The adapter-ligated restriction fragments from all samples are pooled and subjected to size selection. Subsequently, the same adapters provide binding sites for PCR primers, enabling amplification of the fragments while also incorporating platform-specific adapters required for sequencing. The combined use of double digestion and size selection offers greater consistency and reproducibility, allowing the recovery of the same genomic regions across individuals and improving SNP identification (Peterson et al., 2012).

RADseq approaches provide data on thousands of SNPs distributed across the genome, which are sufficient for population genomics analyses (Andrews et al., 2016; Fuentes-Pardo & Ruzzante, 2017). Since reduced representation sequencing methods do not require prior genomic information for the studied species to call SNPs (reads can be assembled *de novo*), they have enabled the widespread use of these methods in population genomics studies of non-model organisms. However, the availability of reference genomes improves the inferences obtained from reduced representation sequencing data, as it provides genomic coordinates for most SNPs, facilitating the identification of linked loci, which is critical for many population genetic inferences (Theissinger et al., 2023). The low cost of RADseq approaches has allowed their application to large sample sizes, enabling the estimation of genetic diversity, inbreeding, effective population size, population structure, gene flow, phylogeographic patterns, and phylogenetic relationship (Balmori-de la Puente et al., 2023; Burriel-Carranza et al., 2024; Escoda et al., 2017, 2019).

1.4. Metabarcoding

In DNA barcoding, universal or taxon-specific DNA loci are analyzed using PCR primers that bind to conserved regions, followed by sequencing of the fragment and its

classification against a reference database (Antil et al., 2023; Kress et al., 2015). DNA metabarcoding, in turn, combines the principles of PCR-based barcoding with next-generation sequencing (NGS) techniques, enabling the analysis of complex samples containing mixtures of organisms (Deiner et al., 2017). This technique is essential in the study of environmental DNA (eDNA), allowing for biodiversity assessment in environmental samples such as water, sediments, air, or those emitted by living organisms, including secretions and feces (Ruppert et al., 2019). Using these data, the presence of species can be determined, such as detecting invasive invertebrate species from water samples (Klymus et al., 2017), identifying terrestrial vertebrate communities through airborne eDNA (Lynggaard et al., 2024), or determining diets from fecal samples (Hawllitschek et al., 2018). However, the accuracy of genetic and taxonomic diversity characterization using DNA barcoding and metabarcoding can be limited by several factors, such as the inability to differentiate closely related taxa that possess identical marker sequences, intraspecific variations across geographic locations, and issues related to PCR amplification, such as the failure to amplify markers when using universal primers (Johri et al., 2019; Theissinger et al., 2023).

1.5. Metagenomics

The shotgun metagenomics technique (or simply metagenomics) avoids the development of specific markers for multiple species and the biases associated with their selection and amplification. This approach involves randomly fragmenting the genomic material extracted from a complex sample and sequencing the resulting fragments using high-throughput sequencing technologies. This process generates a sampling of sequences across the entire genome enabling an accurate estimation of both the taxonomic and functional diversity present in the sample using reference databases (Johri et al., 2019; Porter & Hajibabaei, 2018).

Traditionally, metagenomics has been employed to explore the composition of microbial communities in diverse environments. Since its inception, shotgun metagenomic sequencing has demonstrated its superiority over 16S ribosomal RNA (rRNA) gene sequencing for analyzing microbial communities. Early applications, such as the analysis of seawater from the Sargasso Sea (Venter et al., 2004) and human feces (Qin et al., 2010), showcased its capacity to provide deeper insights into microbial diversity. Furthermore, it has been proposed as a valuable tool for identifying non-culturable organisms that are challenging to detect using traditional PCR-based methods (Handelsman, 2004).

All animals and plants contain an enormous diversity of microorganisms, living inside or outside them, forming the microbiome. The microbiome plays a crucial role in the host health (Huttenhower et al., 2012), contributing to nutrient acquisition, reproduction, development and functioning of the immune system, responses to pollutants, and environmental degradation, among other factors (Johri et al., 2019). Environmental disturbances, changes in diet, or host health can alter the microbiome composition and destabilize it, potentially leading to specific disease phenotypes in the host (Janssens et al., 2018). Since shotgun metagenomics sequences all the

DNA present, this method allows the recovery of both microbial DNA and host DNA. Therefore, it provides a non-invasive alternative to access host and microbial genetic material simultaneously (Johri et al., 2019).

Additionally, metagenomics offers the possibility of reconstructing metagenome-associated genomes. However, its use requires specific conditions such as adequate sequencing depth, low taxonomic complexity in the sample, and small genome size of the organisms (Papudeshi et al., 2017). This approach allows the description of novel microbes, including those not cultivable (Parks et al., 2017).

A challenge in metagenomic studies is that, by analyzing entire genomic regions instead of only phylogenetic markers, the number of taxa that can be confidently identified may be limited. Additionally, taxonomic assignment of certain genome regions can be biased toward organisms with available complete genomes in databases, leading to potential false positives (Porter & Hajibabaei, 2018). For this reason, this technique is primarily applied to the study of microbial communities, which benefit from a broader availability of reference genomes, particularly for known bacteria, archaea, and fungi (Theissinger et al., 2023). In contrast, many genomic sequences from eukaryotes are not yet sequenced (Corominas et al., 2024; Lewin et al., 2018; Mc Cartney et al., 2024). As the number of annotated genomes in public databases increases, it will enable a more comprehensive characterization of communities across broader taxonomic groups, facilitating the description and monitoring of biodiversity from entire ecosystems.

2. NON-INVASIVE SAMPLES IN GENETIC AND GENOMIC MONITORING OF WILDLIFE

The development of genomic methods, together with advances in massive sequencing technologies, has proven to be a powerful tool for monitoring and studying wild species, enabling key questions about ecology, evolution, and conservation to be addressed. However, one of the main obstacles to its application lies in obtaining samples for DNA extraction. This challenge is particularly critical in the case of rare, elusive, or endangered animal species due to logistical, legal, and ethical restrictions that limit direct access to individuals. In this context, the use of non-invasive samples (e.g., feces, hair, and feathers), environmental samples (e.g., water, soil, and air), and archival samples (e.g., museum specimens and subfossils) has emerged as a viable alternative (Andrews et al., 2018).

These samples allow genomic information to be obtained from species without the need to capture, handle, or even directly observe them, reducing the impact on the studied species and facilitating the collection of hard-to-access data (Beja-Pereira et al., 2009; Taberlet & Luikart, 1999). Depending on the degree of contact involved, non-invasive samples can be classified as strictly non-invasive (e.g., naturally shed feathers) or minimally invasive (e.g., buccal swabs or feces taken from captured specimens) (Carroll et al., 2018).

The first studies utilizing non-invasive samples in wildlife research were conducted in 1992, when mitochondrial DNA was successfully obtained and amplified from brown bear (*Ursus arctos*) hairs (Taberlet & Bouvet, 1992) and feces (Höss et al., 1992), as well as nuclear DNA from chimpanzee (*Pan troglodytes*) hairs (Morin & Woodruff, 1992). Höss et al. (1992) also demonstrated for the first time the possibility of detecting dietary items in feces by sequencing a fragment of the chloroplast *rbcL* gene.

Although the use of this type of sample presents significant challenges, mainly due to the low concentration and quality of DNA, as well as the presence of contaminating exogenous DNA (when it is not the target of the study), laboratory techniques and bioinformatic tools have been developed and optimized to adapt to these types of samples (Beja-Pereira et al., 2009; Carroll et al., 2018). In this way, its application has expanded to various approaches, methodologies, and additionally sample types, including eggshells (Egloff et al., 2009), skin swabs (Monteiro et al., 2014) and regurgitations, such as owl pellets (Balmori-de la Puente et al., 2019), among others. Environmental samples as well as ancient materials from biological archives, including fossils or museum specimens, have also been incorporated (Leonard, 2008; Willerslev et al., 2009).

2.1. Feces

The sources of DNA in organisms can include tissues and liquid or solid waste products. The quantity and quality of extracted DNA are influenced by the type of material. Tissues with higher cell density (e.g., muscle, liver, blood) generally provide DNA of superior quality and quantity. In contrast, those with lower cell density (e.g., bone, hair, nails, scales, feathers) or waste products such as feces, which contain shed intestinal epithelial cells, tend to yield lower amounts and are often accompanied by some contamination or degradation (Bengtsson et al., 2012).

Feces are the most common type of non-invasive sample, as they can be easily collected during field surveys in large quantities and without the need to capture or observe individuals. For example, it has been shown that feces allow for reliable population size and density estimates through individual identification using microsatellite data, providing results that are comparable and more cost-effective than those obtained with traditional methods, such as camera traps for tigers (*Panthera tigris*) (Mondol et al., 2009) or helicopter-based observations for brown bears (*Ursus arctos*) (Solberg et al., 2006). Additionally, feces are considered a complex mixture of DNA from multiple organisms, partially digested in the host animal's intestines and sometimes degraded due to exposure to the external environment (Taberlet et al., 2018). As such, they have the potential to provide information about both the host (endogenous DNA) and, through exogenous DNA, its diet, parasites, and microbiome.

Fecal samples present significant methodological challenges, such as low DNA quality and quantity, which lead to the retrieval of short fragments. This can result in genotyping errors, including allelic dropout (the loss of one allele at a diploid locus, resulting in a false homozygote), null alleles (undetected loci, leading to either



Figure 1
Iberian desman feces
(Photo by Ángel Fernández-González)

false homozygotes or missing data), and false alleles (variations that are not true alleles) (Beja-Pereira et al., 2009; Carroll et al., 2018). Different DNA success rates

and associated errors have been observed depending on the storage method (Piggott & Taylor, 2003), collection season, and fecal sample age (Piggott, 2004).

However, this type of sample has enabled the development of a wide variety of studies across different taxa, primarily based on mitochondrial data and the genotyping of microsatellite loci. For example, partial mitochondrial sequences of the cytochrome *b* gene and the D-loop region have been used to determine the presence and distribution of elusive species, such as the Iberian lynx (*Lynx pardinus*) (Palomares et al., 2002). The same markers were employed in phylogeographic analyses to elucidate the evolutionary history of the Iberian desman (*Galemys pyrenaicus*) using feces (Figure 1) (Igea et al., 2013), as shown later, and the Cabrera's water shrew (*Neomys anomalus*), to assess the spatial distribution of its genetic diversity (Querejeta & Castresana, 2018). Additionally, the analysis of microsatellite loci allowed the individual identification of desert bighorn sheep (*Ovis canadensis*), as well as the estimation of nuclear genetic diversity and gene flow, to evaluate the impact of human barriers on the conservation of this species (Epps et al., 2005). In addition, metabarcoding has been widely used for studying exogenous DNA present in feces. This method has been applied to describe the diet of a variety of taxa, including herbivores, carnivores, insectivores, and omnivores, as well as to characterize the microbiome, primarily through comparative studies

of microbial diversity across different conditions or habitats, focusing on its effect on microbiome composition (Andrews et al., 2018).

NGS methods have enabled the construction of genomic libraries that attach adapter sequences to the ends of DNA templates, providing binding sites for PCR and sequencing primers. This facilitates the amplification and sequencing of millions of sequences in a single reaction, maximizing the use of available DNA (Bengtsson et al., 2012). This approach allows for the acquisition of genome-scale information and helps overcome two key limitations of non-invasive samples, as it enables the analysis of small quantities of fragmented DNA. However, when the target is endogenous DNA, an additional challenge in fecal samples is the low availability of host DNA, as most of the sequenced DNA may originate from exogenous sources. For example, qPCR analyses showed that the proportion of endogenous DNA in feces from western chimpanzees (*Pan troglodytes verus*) averaged 1.8% (range: 0.5–5.2%) (Perry et al., 2010) and 2.04% (range: 0.19–8.37%) in yellow baboons (*Papio cynocephalus*) (Snyder-Mackler et al., 2016).

Two strategies used to address these challenges with endogenous DNA are SNP arrays and target enrichment methods. SNP array platforms, such as the Fluidigm SNPtype assay, enable the genotyping of a set of known and validated SNPs using fluorescence measurements, with a low error rate, and have been employed in several wildlife monitoring studies, such as in grey wolves (*Canis lupus*) (Kraus et al., 2015), brown bears (*Ursus arctos*) (Norman & Spong, 2015), and river otters (*Lontra canadensis*) (Stetz et al., 2016). On the other hand, target enrichment methods aim to increase the proportion of target DNA before sequencing using DNA or RNA probes. For example, Perry et al. (2010) designed a DNA capture protocol with probes targeting approximately 1.5 Mb of the nuclear genome, including regions of the X chromosome and the complete mitochondrial genome of western chimpanzees (*Pan troglodytes verus*), achieving sufficient read coverage for SNP identification and the reconstruction of complete mitochondrial genomes. Similarly, (Snyder-Mackler et al. (2016) developed a protocol based on RNA probes to capture host-specific DNA at a genomic scale in wild baboons (*P. cynocephalus* and *P. papio*), successfully increasing the proportion of endogenous DNA 40-fold.

Another strategy that avoids the development of probes (a costly, labor-intensive, and time-consuming process) was developed by (Chiou & Bergey, 2018) under the name *FaecalSeq*. This method utilizes differences in CpG methylation density between vertebrate and bacterial genomes, employing a protein that selectively binds to methylated DNA (host DNA) to isolate it. *FaecalSeq* was validated with fecal samples from wild baboons (genus *Papio*), where it was demonstrated an increase in the mean proportion of endogenous DNA from 0.3% (range: <0.01–3.1%, measured by qPCR) to 28.8% (range: 1.5–73.6%, measured as the number of reads mapped to the baboons genome). This method has the advantages of being cost-effective and not requiring prior knowledge of the host genome. However, a potential disadvantage is the co-enrichment of non-target DNA, such as dietary DNA, although this aspect could also be exploited for specific dietary studies.

A methylation-based DNA enrichment method based on *FaecalSeq* was used to obtain genome-wide SNP data from ddRAD applied to fecal samples of the jungle cat (*Felis chaus*) (Tyagi et al., 2022), as well as the gaur (*Bos gaurus*) and sambar (*Rusa unicolor*) (Tyagi et al., 2024), to evaluate anthropogenic pressures on population connectivity, providing valuable insights aimed at improving biodiversity management.

In contrast, the difference in CpG methylation density has also been used to remove host DNA when the goal is to study the associated microbial community (Feehery et al., 2013). Although several methods have been developed to enrich samples with limited microbial content, these procedures can introduce biases into sequencing data (Quince et al., 2017). Given that the majority of DNA present in feces is considered to be of bacterial origin (Qin et al., 2010), the most commonly used approach in shotgun DNA sequencing with fecal samples is to filter endogenous reads after sequencing through bioinformatic analyses by mapping them to the host genome (Lu et al., 2022).

2.2.1. Metagenomics analyses from feces

Metagenomic shotgun sequencing offers significant advantages over targeted methods (such as PCR-based metabarcoding), as it avoids PCR biases and the lack of resolution associated with the use of a limited number of loci. By sequencing all the DNA present in feces, it enables the identification of a wide variety of taxa in a single sequencing run, including both enriched and unenriched DNA, while also enabling sequence assembly (Andrews et al., 2018). This approach has the potential to facilitate the exploration of questions related to host population genetics, ecology, and health by analyzing DNA from diverse sources (Figure 2).

The primary use of metagenomics in fecal samples so far has been to analyze the composition of the gut microbiome, taking advantage of the abundance of microbial DNA in feces, which allows for the taxonomic and functional analysis of microbial communities (Akoijam & Joshi, 2022). This approach has been used to investigate various aspects of host-microbiome interactions and to evaluate host health status and responses to environmental stressors (Wei et al., 2019). Similar to the applications of metabarcoding, numerous studies have focused on differences between the microbiomes of animals under different conditions and their implications for host health (Williams et al., 2018). For example, Gibson et al. (2019), through metagenomic analysis of feces, detected differences in the taxonomic and functional composition of the microbiomes of wild and captive black rhinoceros, attributing these differences to variations in diet and providing recommendations to improve the welfare of captive animals.

On the other hand, metagenomics has been barely explored for studying other sources of DNA present in the fecal samples, for instance, in diet

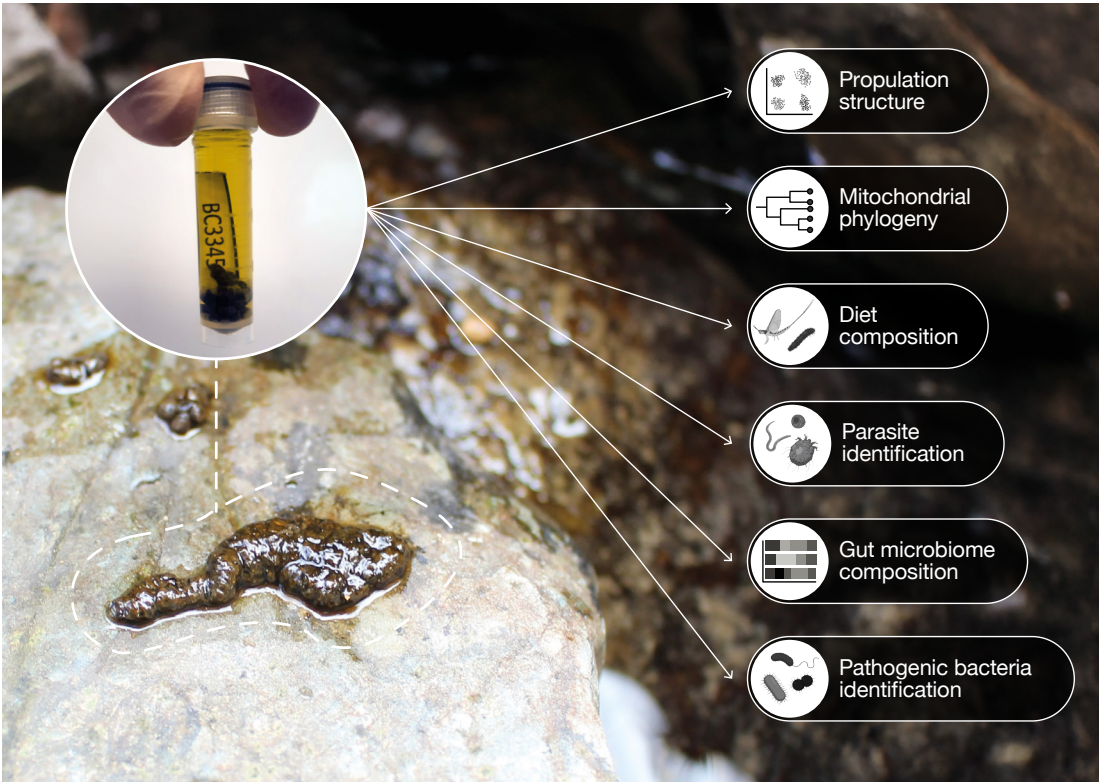


Figure 2
Diagram of the analyses that can be performed from shotgun metagenomics data obtained from fresh fecal samples (Photo: fecal sample of *G. pyrenaicus*, by José Castresana).

characterization. This is mainly because a great effort is required in both sequencing and bioinformatic analysis, as the standardization of pipelines for

result processing is still underdeveloped (Alberdi et al., 2019). Srivathsan et al. (2015) conducted a pioneering study using both metabarcoding and metagenomics to evaluate the diet of captive red-shanked doucs langurs (*Pygathrix nemaeus*) from their feces, comparing the results with known diet data. Metagenomics, through shotgun sequencing with multiple and longer barcodes, achieved greater taxonomic resolution but failed to detect low-abundance diet items. The authors suggested that increased sequencing effort or a more comprehensive reference database with longer sequences could enhance resolution. For instance, they demonstrated that, in some cases, metagenomic reads were uniformly distributed across the chloroplast genome, which could be improved if more complete chloroplast genomes were available as references. Following this work, Srivathsan et al. (2016) applied the same methodologies to fecal samples of banded leaf monkeys (*Presbytis femoralis*) collected in the field, comparing the results with observational data. They found a good agreement between both methods and field observations, while once again demonstrating the higher taxonomic resolution achieved with metagenomics. This time, they successfully identified rare dietary items by increasing sequencing depth. In addition,

metagenomics also enabled the reconstruction of complete mitochondrial genomes of the host and the identification of some intestinal parasites.

In another example, (Chua et al., 2021) compared the identifications obtained through metabarcoding and metagenomics using fecal samples from western capercaillie (*Tetrao urogallus*), employing chloroplast and nuclear markers for metagenomic identifications and a chloroplast marker for metabarcoding. In this case, the methods did not show significant differences in the number of taxa identified or in species-level resolution. However, they found certain discrepancies between the detected items, mainly at more specific taxonomic levels (genus, species), although in both cases the results aligned with expected diet items. They highlighted, once again, the need for more comprehensive reference databases, especially those representing local species. While both methods proved effective for describing dietary profiles, the greater potential of metagenomics lies in its ability to simultaneously provide additional information about the host, as they were also able to assemble the mitochondrial genomes of the capercaillie.

Some studies have used the potential of recovering mitochondrial genomes from fecal samples. Ang et al. (2020) reconstructed the phylogeny of the genus *Presbytis* (Asian colobines) using mitogenomes obtained from fecal samples to delimit new species, which could have implications for updating their conservation status, according to the authors. Similarly, Baeza (2022) demonstrated the ability to *de novo* assemble the mitogenome of *Panthera tigris altaica* (Amur tiger) from fecal samples, and through phylogenetic analyses confirmed its capacity to distinguish between closely related species within the genus *Panthera*, representing a useful tool for monitoring.

Recent studies on this topic have focused on analyzing the global DNA content in fecal samples and testing the capacity of recovered endogenous nuclear DNA to obtain genome-wide data for population-level analyses. Taylor et al. (2022) carefully selected fecal samples from caribou (*Rangifer tarandus*) for sequencing, prioritizing those with the highest DNA quantity and quality. DNA was extracted from the mucosal layer of the fecal pellets to maximize the likelihood of retaining intestinal epithelial cells. These high-quality samples allowed the recovery of a significant proportion of nuclear endogenous DNA, with a mean of 30.87% (range 5.81%–59.71%) of the reads mapped to the reference genome, which represented values significantly higher than those previously reported. These data enabled SNP calling and population genomic analyses, including Principal Components Analysis (PCA), in combination with tissue samples. Additionally, they performed taxonomic classification of the reads. Fecal samples contained reads from green plants, SAR (Stramenopiles, Alveolata, and Rhizaria), archaea, viruses, and other eukaryotes, with a substantial proportion of unclassified reads, although the exact number was not reported.

On the other hand, de Flamingh et al. (2023) used shotgun libraries to analyze the impact of the time elapsed between defecation and sample collection (0h, 24h, 48h, 72h) on the overall composition of feces from captive African savanna elephants (*Loxodonta africana*), sampling the outside of the same dung bolus. It was observed that, over time, total DNA decreased, along with the proportion of endogenous DNA (both nuclear and mitochondrial) and other taxonomic groups, except for bacteria, which increased. Additionally, the number of unclassified reads also increased with time. They also quantified the proportion of nuclear endogenous DNA in fresh feces from free-ranging elephants by mapping the reads to the reference genome of the species, recording a mean of 12.41% (range 5.54% to 21.65%) and generating low-coverage genomic data sufficient to perform PCA analysis, in combination with other samples from the same geographic location. They were also able to reconstruct the mitochondrial genomes of *L. africana*. Furthermore, they performed a metagenomic classification of the reads, identifying taxa that included *Loxodonta*, green plants, fungi, arthropods, bacteria, archaea, and viruses, with an average of 74% unclassified reads.

To the best of our knowledge, these studies represent the metagenomic approaches developed with fecal samples from wild animals to date, where microbiome analysis was not the primary objective. In all these studies, it became clear that there is a large number of unidentified reads and therefore large room for improving metagenomic pipelines and analyses. In this context, the read classification methods developed in the first part of this thesis for the identification of diet elements and parasites represent an innovative advance to increase the quantity and quality of assignments. Furthermore, the combined development of population analyses using nuclear and mitochondrial endogenous DNA allows to maximize the utility of fecal samples in conservation genomics studies.

3. METHODS AND DATABASES IN METAGENOMIC ANALYSIS

Metagenomic studies present several challenges, primarily due to the complexity and volume of data generated, which complicates computational analyses. For instance, read classification, that is, determining the genome from which a read originates, can be difficult. Additionally, due to the diversity of genomes present and their relative abundance, not all genomes may be fully and equally represented in the sequence dataset (Breitwieser et al., 2019). Therefore, the vast amount and variety of genomic information sampled require substantial data volumes to generate meaningful results, resulting in high computational demands. Fortunately, significant advancements in computational software have been made to address these challenges, improving the efficiency and accuracy of metagenomic analyses, primarily developed for the characterization of microbial communities (Sharpton, 2014).

3.1. Host DNA filtering

In addition to sequence quality controls, one of the first steps in metagenomic analysis is the filtering of host DNA. If the host reference genome is available, a commonly used method is the mapping of reads with short-read alignment programs such as Bowtie2 (Langmead & Salzberg, 2012) or BWA-MEM (H. Li, 2013).

3.2. Assessing taxonomic diversity

One of the main approaches to characterizing microbial communities is analyzing their diversity by determining which organisms are present (taxonomic classification) and in what abundance (usually related to the number of assigned reads), and comparing it with other communities. To achieve this, numerous computational approaches have been developed for the taxonomic classification of metagenomic read pools, which can be grouped into two main categories: reference-based methods and assembly-based methods (Figure 3) (Breitwieser et al., 2019; Pinto & Bhatt, 2024). These methods are not mutually exclusive and can often complement each other (Blanco-Míguez et al., 2023). Moreover, although they were initially developed to work with short-reads, their application with long-reads has been progressively integrated.

3.2.1. Reference-based taxonomic profiling

Reference-based methods rely on matching reads to closely related sequences in reference databases to identify the corresponding taxon for each read. These databases can include both taxonomically informative genetic markers and complete genomes. Initially, the most commonly used method was BLAST (Altschul et al., 1990), which provides high sensitivity and accuracy in taxonomic assignments (Breitwieser et al., 2019). This approach was used, for instance, by Srivathsan et al. (2015, 2016) with

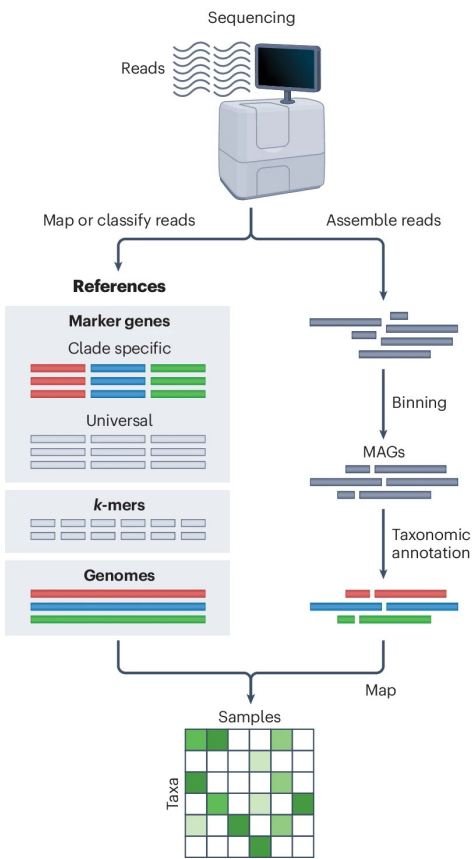


Figure 3. Profiling taxonomic composition using shotgun metagenomic sequencing. Two primary approaches are shown: The first is a reference-based method (left), where reads are mapped to marker sequences (clade-specific or universal) or whole genome sequences, or classified using k-mer-based algorithms. The second approach involves assembly-based methods (right), where reads are assembled into contigs, binned into metagenome-assembled genomes, and assigned taxonomy. To quantify abundance, reads are mapped back to the assembled genomes. Both approaches ultimately generate a table of taxa abundance per sample as the final output (from Pinto & Bhatt (2024)).

databases of universal plant chloroplast barcodes (rbcL, matK, trnL-F, and trnH-psbA) to determine the diet of colobines. Similarly, Chua et al. (2021) used three chloroplast markers (rbcL, matK, and trnL-F) and two nuclear markers (ITS1 and ITS2), commonly applied in plant identification, to describe the Capercaillie diet. This method is feasible with limited sequencing data. However, with the expansion of databases and the growth of sequencing data for the samples, BLAST (Altschul et al., 1990) alignments became computationally inefficient. This led to the development of other metagenomic classifiers, which, while less sensitive than BLAST (Altschul et al., 1990), address these challenges (Ye et al., 2019).

Tools designed to enhance speed and computational efficiency in sequence classification typically employ three main approaches: the use of marker gene databases, the application of k-mer-based methods, and the translation of DNA sequences into proteins. For example, mOTUs (Sunagawa et al., 2013) aligns reads against a reduced database of phylogenetically informative universal markers using the HMMER (Eddy, 2011) aligner. MetaPhlAn (Segata et al., 2012), on the other hand, employs a carefully curated database containing thousands of unique clade-specific marker genes with high discriminatory power, which are aligned using Bowtie2 (Langmead & Salzberg, 2012). These methods enable faster alignments by reducing the

size of the databases (compared to whole genomes) and provide accurate assignments that offer a detailed view of microbial composition, even allowing strain-level assignments (Quince et al., 2017). However, because they rely on only a fraction of the reference genomes, these methods are less sensitive for detecting low-abundance microorganisms compared to approaches that utilize complete reference genomes in the database (Pinto & Bhatt, 2024). Additionally, these methods are highly database-dependent and are primarily used in environments where a significant portion of microbial diversity is already well-characterized and represented in databases (Quince et al., 2017).

K-mer-based methods enable the use of complete reference genomes by replacing the alignment of reads with the search for matches between k-mers, increasing speed and reducing computational workload (Breitwieser et al., 2019). The first program developed with this approach was Kraken (Wood & Salzberg, 2014), which remains one of the most widely used along with its newer version, Kraken2 (Wood et al., 2019). This program stores the k-mers from each reference sequence along with their species-level taxonomic identifier. These k-mers are used to search for matches with k-mers from metagenomic reads. When a match is found, the originating read is assigned to the corresponding taxon using the lowest common ancestor (LCA) criterion (Huson et al., 2007). That is, if a k-mer matches more than one taxon, it is assigned to the lowest common level. In this way, Kraken classifies reads into a taxonomic tree, enabling an analysis of the global diversity in the sample. When the goal is to analyze the composition at a given level, such as species or genus, Kraken is used in conjunction with the Bracken (Lu et al., 2017) extension. This tool estimates species-level abundances based on the number of reads assigned to each species in the sample, enabling studies of alpha-diversity and beta-diversity (Lu et al., 2022).

Kraken2 (Wood et al., 2019) was the method used by Taylor et al. (2022) and (de Flamingh et al., 2023) to describe the global composition of caribou and elephant feces, respectively, using a database available in the program's repository. This database includes RefSeq sequences of archaea, bacteria, plasmids, viruses, humans, fungi, plants, and protozoa, as well as the NCBI non-redundant nucleotide database and UniVec_Core (NCBI database of vector, adapter, linker, and primer sequences). They supplemented this database with the corresponding host reference genomes.

Lastly, because protein sequences are more conserved than genomic sequences, their comparison allows for the identification of similarities with more distant taxa, improving sensitivity in cases where databases cover only a fraction of the diversity present in the samples. Additionally, by focusing exclusively on coding regions, the size of the databases is reduced, as well as the redundancy of genomic sequences, which often lead to multiple equally

valid alignments for the same read (Bagci, 2022). DIAMOND (Buchfink et al., 2014) is one of the most widely used programs employing this approach, aligning six-frame translations of reads against protein databases (e.g., the NCBI-nr database), similarly to BLASTX (Altschul et al., 1997), but with greater speed. Taxonomic assignment is performed by selecting the matches with the highest scores and using the LCA approach implemented in MEGAN (Huson et al., 2007). Furthermore, the DIAMOND+MEGAN combination enables functional classification by assigning each read to a functional category using classification systems such as eggNOG (Powell et al., 2012), among others (Bağcı et al., 2021). Moreover, DIAMOND+MEGAN allows working with both short reads and long reads (Bağcı et al., 2021).

3.2.2. Assembly-based methods

The goal of these methods is to use reads to generate genomic assemblies for each microorganism present in the sample (Breitwieser et al., 2019; Pinto & Bhatt, 2024). This typically involves three steps: assembling the reads to generate contigs, grouping contigs of the same origin (often called "binning"), and evaluating quality. *De novo* metagenomic assembly is similar to single-genome assembly, as it also uses the de Bruijn graph approach. However, metagenomic assembly presents specific challenges, such as variations in genome coverage depth depending on the different abundances in the sample, and the presence of closely related genomes, which can be exacerbated by the complexity of the sample (Quince et al., 2017).

Specific metagenome assemblers were developed to address these challenges, such as metaSPAdes (Nurk et al., 2017), MEGAHIT (D. Li et al., 2015), and MetaVelvet (Namiki et al., 2012). Since metagenomic assemblies are highly fragmented, they produce thousands of contigs that are then grouped to form metagenome-assembled genomes. Binning can be performed through taxonomic classification, using some of the methods described earlier, or based on features such as tetranucleotide frequencies or depth coverage profiles across multiple samples, using tools like CONCOCT (Alneberg et al., 2014). Finally, one of the most widely used tools for quality control is CheckM (Parks et al., 2015), which assesses the completeness of the assembled genome and contamination with foreign DNA. Although these assemblers primarily use Illumina sequencing reads due to their lower error rate, some metagenomic assemblers, such as metaFlye (Kolmogorov et al., 2020) and hifiasm-meta (Feng et al., 2022), have been developed to work with long-reads (Pacific Biosciences or Oxford Nanopore), showing good results thanks to the generation of more contiguous assemblies.

Metagenomic assemblies have mainly contributed to the recovery of microbial genomes from uncultured or non-isolated organisms and from previously unexplored ecological niches, enabling the discovery of new organisms (Pinto & Bhatt, 2024).

3.2.3. Methods for species-level pathogen identification

The approaches presented previously are mainly used for microbial profiling, aiming to perform analyses related to relative abundances. However, it is important to consider their accuracy and limitations when it comes to identifications at lower levels, such as species or strain. This is especially critical when the goal is the identification of pathogenic species, which require stricter criteria.

LCA-based methods classify the largest possible number of reads by assigning the common taxon for ambiguous reads. In this way, they achieve high accuracy at higher taxonomic levels, such as phylum and family (Francis et al., 2013). However, reads that are assigned at the species level have a high potential for false positives, mainly due to misclassifications in low-complexity regions or highly conserved regions. This issue is particularly pronounced in k-mer-based methods, as they do not provide detailed information about the exact matching site (Pinto & Bhatt, 2024).

Marker-based methods, such as MetaPhlAn (Segata et al., 2012) and its strain-level adaptation, StrainPhlAn (Truong et al., 2017), achieve higher resolution at the species and strain levels. However, they have the drawback of requiring greater sequencing depth (10X or more) to detect low-abundance organisms compared to whole-genome mapping methods (Yan et al., 2020). Additionally, the databases used by these methods do not adequately cover poorly or unexplored environments (Blanco-Míguez et al., 2023). On the other hand, assembly-based methods also enable precise species- and strain-level identification but require even higher sequencing depth (often on the order of 50-100X of the targeted genome) and significantly more computational effort (Pinto & Bhatt, 2024; Schatz et al., 2010).

Pipelines based on aligning reads to complete reference genomes circumvent many of these problems. For instance, to validate the classification performed by Kraken2 (Wood et al., 2019), the classified reads are mapped to a specific species' reference genome using Bowtie2 (Langmead & Salzberg, 2012), assessing alignment quality based on coverage distribution across the entire genome (Lu et al., 2022). Additionally, other pipelines have been specifically designed to detect pathogenic species in clinical applications. These include PathoScope (Francis et al., 2013; Hong et al., 2014) and SURPI (Gu et al., 2021; Naccache et al., 2014), which use read mapping to reference genome databases as their main engine, employing the Bowtie2 (Langmead & Salzberg, 2012) and SNAP (Zaharia et al., 2011) aligners, respectively. Additionally, some mapping-based pipelines have been developed for long reads, such as MetaMaps (Dilthey et al., 2019) and NanoMap (Hall et al., 2020). Mapping methods are optimized for speed in whole genome analysis, which can only be achieved by mapping reads to the genomes of the same or very closely related species. These methods also calculate a mapping quality (MAPQ)

score (Langmead, 2017; H. Li et al., 2008), which reflects the probability that a read is aligned to the correct position in the reference genome, taking into account the difference between the highest-scoring alignment position and all other possible positions. For example, a read originating from a repetitive element will align equally well to every location where the element is present in the genome; therefore, its MAPQ score will be low (Langmead, 2017). Consequently, MAPQ serves to select only informative reads and also acts as a proxy for the confidence that the read truly originates from the genome of the species used for mapping. Due to these characteristics, and although mapping methods were not initially developed for species identification, they have proven to be very useful for this purpose due to their efficiency and accuracy in aligning reads to specific reference genomes.

Despite significant progress in methods targeting exact species classification of pathogens and parasites in human health applications, there remains a notable gap in applying these techniques to non-model species, mainly due to the need of complete genomes of the exact or closely related species in databases. However, the increasing number of complete genomes now supports their broader use (O'Leary et al., 2016). Developing methods to obtain accurate genomic information about pathogens from the feces of endangered species is imperative for enhancing pathogen surveillance and help protect vulnerable populations (ENETWILD-consortium et al., 2023).

In this thesis, a new method for identifying pathogenic bacterial species from metagenomic fecal data of a non-model species is developed and tested. It is based on the mapping of unassembled reads to bacterial reference genomes, and the selection of the best alignment quality criteria for species identification. This method can be employed for monitoring pathogenic microbes in endangered wild species.

3.3. Databases resources

Almost all the described methods rely, to a greater or lesser extent, on databases of sequences or complete genomes taxonomically identified. The accuracy and representation of these databases critically influence metagenomic analyses (Wright et al., 2023; Xu et al., 2023), and therefore, the availability of reference databases with the widest possible biodiversity coverage is highly necessary.

The most commonly used databases are those provided by the NCBI (National Center for Biotechnology Information of the USA) (Sayers et al., 2021), including GenBank (Sayers et al., 2019), the nucleic acid sequence database. GenBank includes sequences submitted by authors and data exchanged with the EMBL-EBI European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ), as part of the International Nucleotide Sequence Database Collaboration (INSDC). This ensures that the information is uniform and available worldwide. GenBank is accessible

through the NCBI nucleotide database, along with sequences from RefSeq (Goldfarb et al., 2024), TPA (Third Party Annotation), and PDB (Protein Data Bank), linking each sequence to related information such as taxonomy, genomes, protein sequences and structures, and literature from PubMed (Sayers et al., 2019).

RefSeq contains curated non-redundant sequences of organelles, plasmids, targeted loci, and annotated genomes of eukaryotes, viruses and, to a greater extent, prokaryotes. These genomes are selected from the INSDC database and then curated and annotated using a specific pipeline (Goldfarb et al., 2024). For eukaryotes and bacteria, the NCBI selects one RefSeq genome per species and designates it as the "reference genome". These sequences, along with other whole genomes from GenBank, are available in NCBI Datasets (previously known as NCBI Genome and NCBI Assembly) (Sayers et al., 2024). Additionally, the NCBI Pathogen Detection Project integrates sequence data of pathogenic bacteria (Sayers et al., 2021).

To complement these resources, NCBI Taxonomy (Schoch et al., 2020a) includes organism names and taxonomic classifications for every sequence in its databases, which are shared with the INSDC. NCBI Taxonomy consists of a hierarchical system covering all domains of life, assigning each node a name, according to nomenclature codes, and a unique taxonomic identifier (TaxId). This enables access to all sequences of a specific taxon across all NCBI resources. Currently, NCBI Taxonomy can be accessed via NCBI Datasets Taxonomy (Cox et al., 2024).

Other available databases mainly contain high-quality DNA sequences of taxonomically classified genes. These standardized DNA loci, commonly known as DNA barcodes or DNA markers, have sufficient variability to theoretically provide species-level resolution (Porter & Hajibabaei, 2020). Since the species-resolving power and taxonomic coverage vary depending on the organisms under study, different marker genes are used for different groups (Creer et al., 2016).

The SILVA (Quast et al., 2013), Greengenes (McDonald et al., 2012), and RDP (Ribosomal Database Project) (Wang & Cole, 2024) databases provide 16S rRNA gene sequences (small subunit rRNA gene (SSU)), which are commonly used for studying prokaryotic diversity (Balvočiute & Huson, 2017). In Bacteria, this gene contains nine hypervariable regions (V1-V9), of which the V2 and V3 regions offer the highest resolution for identifying lower-level taxa, such as species and genus (Bukin et al., 2019).

SILVA also provides sequences of the 23S rRNA gene (large subunit, LSU) for Bacteria and Archaea, and the nuclear 18S and 28S genes (SSU and LSU, respectively) for Eukarya (Quast et al., 2013). The 18S rRNA gene consists of nine highly variable regions (V1-V9) interspersed with conserved regions, among which the V2, V4, and V9 regions have been shown to be the most variable and suitable for biodiversity analyses in eukaryotes (Hadziavdic et al., 2014; Wu et al., 2015). However, its use has been primarily for the identification of protists, with varying degrees of resolution depending on the region used (Antil et al., 2023), as well as in nematodes and unicellular fungi (Tedersoo et al., 2024).

On the other hand, the 28S rRNA gene has a higher evolutionary rate than the 18S gene, providing greater discriminatory power for closely related taxa (Antil et al., 2023). This gene contains 12 variable regions (D1-D12), among which the D1-D2 and D1-D5 regions have demonstrated resolution power between closely related species in protists (Antil et al., 2023; Wylezich et al., 2010). The D1-D2 region has also been proposed as a marker region for species-level identification in metazoans, based on its evaluation in closely related fish species (Sonnenberg et al., 2007). Additionally, the D3 region showed interspecific variation among members of Acari (Lehmitz & Decker, 2017; McClain et al., 2001).

Another widely used database is BOLD (Barcode of Life Data System) (Ratnasingham & Hebert, 2007), which mainly contains fragments of the mitochondrial cytochrome c oxidase I (COI) gene from metazoans. It additionally contains barcodes from other genes, such as the nuclear ITS (Internal Transcribed Spacer) gene for fungi and plants, and the chloroplastic markers matK (maturase K) and rbcL (large subunit of ribulose 1,5-bisphosphate carboxylase) also for plants, among others. The COI gene have been proposed as the core for species-level identification in animals (Hebert et al., 2003), and are well represented in both BOLD and NCBI. However, these two databases use different classification systems and taxonomic identifiers (TaxID), which makes it difficult to generate a unified database from these sources (Meglécz, 2023).

Because the mitochondrial genome evolves at higher rates than the nuclear genome (Antil et al., 2023), mitochondrial sequences are more informative for differentiating between closely related species. For this reason, in addition to COI, other genes such as ND5 (the mitochondrial gene encoding NADH:Ubiquinone Oxidoreductase Core Subunit 5) and COII have been proposed as barcodes, among others (Antil et al., 2023). Likewise, the use of complete mitochondrial genomes obtained through mitochondrial metagenomics has also been suggested (Crampton-Platt et al., 2016).

Databases are a fundamental tool for metagenomic studies, and issues in reference sequences can lead to errors in taxonomic assignments. Some common problems found in reference sequences include incorrect taxonomic labelling, the presence of contaminations in the sequence (e.g., inclusion of vector or viral sequences or chimeric sequences), low-complexity sequences, fragmented or incomplete sequences, and uneven taxonomic representation, with some groups being underrepresented and others overrepresented, that make database searches problematic (Chorlton, 2024). This highlights the need for better curation of databases, as well as for more reference sequences and high-quality complete genomes across the entire tree of life (Corominas et al., 2024; Lewin et al., 2018; Mc Cartney et al., 2024).

4. THE IBERIAN DESMAN (*Galemys pyrenaicus*)

The Iberian desman or Pyrenean desman (*Galemys pyrenaicus*) (E. Geoffroy St. Hilaire, 1811) (Figure 4) is a small semiaquatic insectivore mammal endemic to the Iberian Peninsula and the Pyrenees, belonging to the order Eulipotyphla, family Talpidae, and subfamily Desmaninae. This subfamily, characterized by its aquatic adaptations, currently contains only two species from monotypic genera: *G. pyrenaicus* and the Russian desman (*Desmana moschata*), which is distributed in an area north of the Black and Caspian seas, in Eastern Europe (Palmeirim & Hoffmann, 1993).



Figure 4. The Iberian desman (Photo by Joxerra Aihartza, from Esnaola et al., 2018).

4.1. Morphology adapted to aquatic habitats

G. pyrenaicus is a very small animal, with an average head and body length of 120.5 mm (108–135 mm) and a tail length of 146.9 mm (123–156 mm), giving a total average length of 267.4 mm. Its average body weight is 57.75 g, with a range of 35.5 g to 80 g (Palmeirim & Hoffmann, 1993). It possesses unique characteristics due to its specialized adaptation to aquatic habitats. It has a conical head with a long, flexible snout flanked by vibrissae, which are used to capture prey on the riverbed. A tracheal adaptation separates the digestive tract from the respiratory tract, allowing it to ingest prey underwater. Its eyes are small, and it lacks visible ear pinnae, further highlighting its adaptation to underwater life (Nores, 2017; Palmeirim & Hoffmann, 1993). Its limbs are highly specialized for swimming. The forelimbs are small with short interdigital membranes, while the hindlimbs are much larger and longer,

equipped with wide interdigital membranes. This allows the animal to move swiftly through the water, aided by its long, thick tail. The tail is covered with sparse, stiff hair that reveals scaly skin underneath, providing additional propulsion in aquatic environments (Palmeirim & Hoffmann, 1993). As for its fur, the guard hairs, roughly twice as long as the underhairs, are greatly enlarged distally, a characteristic shared with other aquatic mammals. This unique fur structure effectively insulates the animal from water, ensuring it can maintain body temperature while swimming and diving in its aquatic habitat. (Nores, 2017; Palmeirim & Hoffmann, 1993).

4.2. Sexual characters, sex identification and reproduction

The Iberian desman does not present evident sexual dimorphism, which is common among talpids (Nores, 2017; Palmeirim & Hoffmann, 1993). It is difficult to determine the sex of individuals because both sexes possess similar peniform external genitalia. Males have intra-abdominal testes, and females exhibit masculinized external genitalia, with the urinary papilla having an internal structure similar to the penis of young males (Peyre, 1955). Additionally, females have an open vagina only for a brief period during the breeding season, which facilitates sex identification at that time, when a change in vaginal pigmentation also occurs (Richard, 1986). Another characteristic used for sex identification is the rigidity of the pelvic arch, which can be assessed by palpation. In females, this arch starts as cartilaginous, with a slight central groove that widens with age, eventually becoming a connective ligament in adult females due to a process of osteolysis. In contrast, in males, the cartilaginous arch fully ossifies, becoming a hard structure in adults (Peyre, 1957). However, this difference is not helpful for distinguishing immature specimens. Additionally, González-Esteban et al. (2003) reported the possibility of identifying sex by eversion and observation of the external genitalia after applying slight pressure on the urinary papilla.

More recently, molecular methods for sex determination of the Iberian desman have been established. Vidal et al. (2010) developed a method based on PCR amplification of two regions belonging to intron 8 of the DEAD Box X isoform (DBX) and DEAD Box Y isoform (DBY) genes, differentiating them by size using capillary electrophoresis. This method was used for sex determination based on DNA obtained mainly from tissue and hair samples. On the other hand, Querejeta et al. (2016) developed a sex determination method using ddRAD-seq sequences from tissue samples. Initially, the sex of some individuals was determined by qPCR, amplifying a fragment linked to the Y chromosome. Subsequently, from the ddRAD sequences, they identified 49 Y-linked loci in the males sexed by qPCR, and these loci were used as a reference for aligning the ddRAD sequences through BLAST searches in the rest of the individuals. Very stringent alignment criteria (evalue = 1e-74) were used to avoid detecting duplicates on the X chromosome.

Regarding reproduction, the mating period lasts from January to May. Spermatogenesis in males begins in November and continues until May, while estrus in females extends from January to May, with three reproductive peaks in February, March, and

May. The gestation period is approximately 30 days, which allows females to have postpartum estrus, meaning that lactating females can become pregnant again. This contributes to the possibility of having several litters per season. The number of embryos per female ranges from 1 to 5, with an average of 4 (Peyre, 1962). The offspring are born between April and August (Richard, 1962) and reach maturity in their second year (Peyre, 1962).

4.3. Distribution

The Iberian desman is endemic to the Iberian Peninsula, with a distribution that includes France, Andorra, northern and central Spain, and northern Portugal. Its populations are distributed in a fragmented and discontinuous manner across mountainous systems, from the Pyrenees, through the Cantabrian Arc, Galicia, northern Portugal, the Central System, and the Iberian System (Figure 5) (Quaglietta et al., 2024).

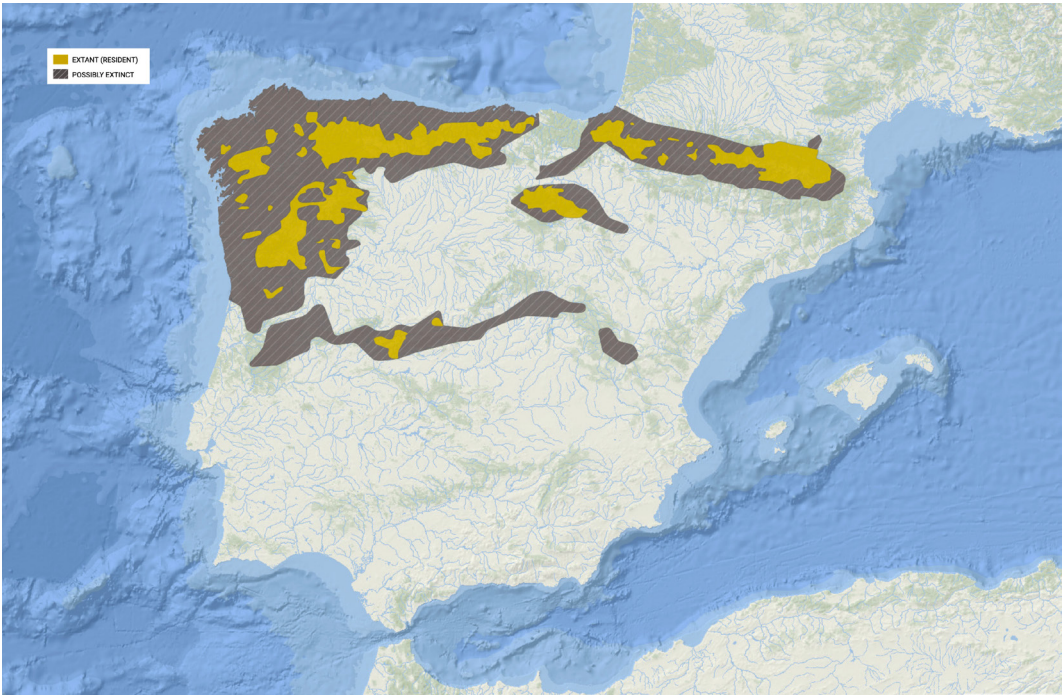


Figure 5. Distribution of the Iberian desman according to the IUCN red list (from Quaglietta et al. (2024)).

4.4. Evolutionary history and phylogeographic structure

The current distribution of *G. pyrenaicus* is linked to its evolutionary history and the marked phylogeographic structure of its populations, which reflect a postglacial expansion scenario following isolation in glacial refugia (Figure 6). This was demonstrated through the analysis of mitochondrial genes (cytochrome *b* and D-loop), using non-invasive samples (feces, tissues from

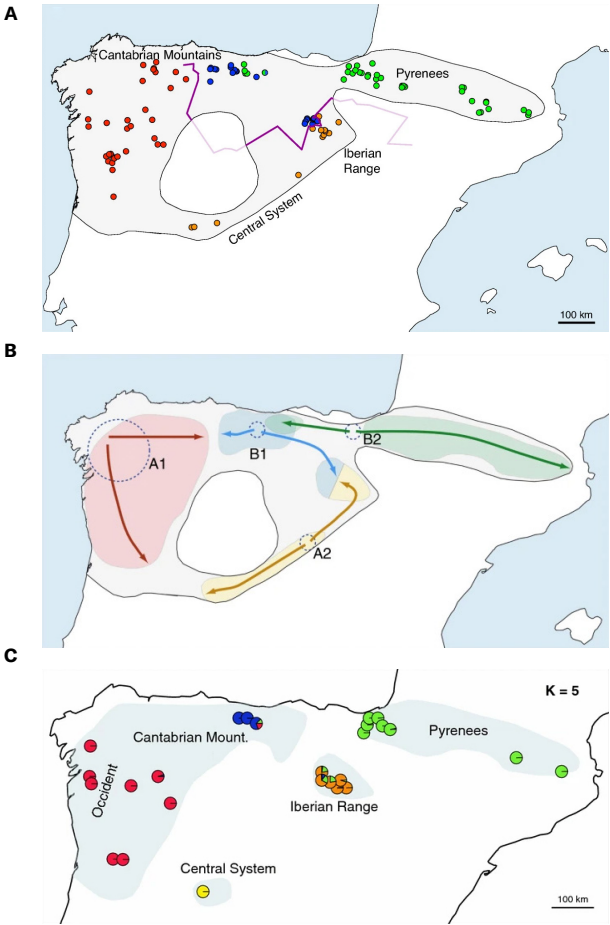


Figure 6. Phylogeography of the Iberian desman. a) Distribution of the four identified mitochondrial lineages. b) Historical distribution of the species (in gray), current distribution of the mitochondrial lineages (in colors), and hypothetical locations of glacial refugia (dotted circles) (from Igea et al. (2013)). c) Distribution of the five nuclear clusters, with admixture proportions for each individual determined through a structure analysis with K=5. (from Querejeta et al. (2016)).

biological collections, and bones from museum specimens) by Igea et al. (2013). The mitochondrial gene phylogeny identified two main clades, A and B, each further subdivided into two subclades: A1, A2, and B1, B2 (Figure 6a). This resulted in four mitochondrial lineages, whose distribution is linked to their geographic location and exhibits a parapatric distribution pattern. Two contact zones between lineages were identified: one between A1 and B1, and another between A2 and B1, though no evidence of genetic mixing was observed between them, at least at the mitochondrial level, which reflects exclusively maternal inheritance (Igea et al., 2013).

A further study, based on nuclear SNP data obtained through ddRAD, identified five main clusters or evolutionary significant units with low levels of genetic admixture between them, which largely corresponded to the major mountain systems: Pyrenees, Cantabrian Mountains, Occident, Central System, and Iberian Range (Figure 6c) (Querejeta et al., 2016). While the results largely aligned with the mitochondrial clades described previously, some discrepancies were observed. Notably, in the Iberian Range, the two mitochondrial clades detected by Igea et al. (2013) were grouped into a single genomic nuclear cluster. Another discrepancy was found in the Central System, where haplotypes from this area grouped with those of the Iberian Range (forming part of clade A2), whereas nuclear phylogeography defines a clearly distinct cluster for this region. These results underscore the importance of conducting genomic analyses to achieve a more comprehensive understanding of species' evolutionary history.

In terms of genetic diversity, Igea et al. (2013) reported a globally low mitochondrial genetic diversity, with a nucleotide diversity (π) value of 0.0073, which is four times lower than the average reported for mammals (Nabholz et al., 2008). Additionally, significant differences were observed among lineages, with

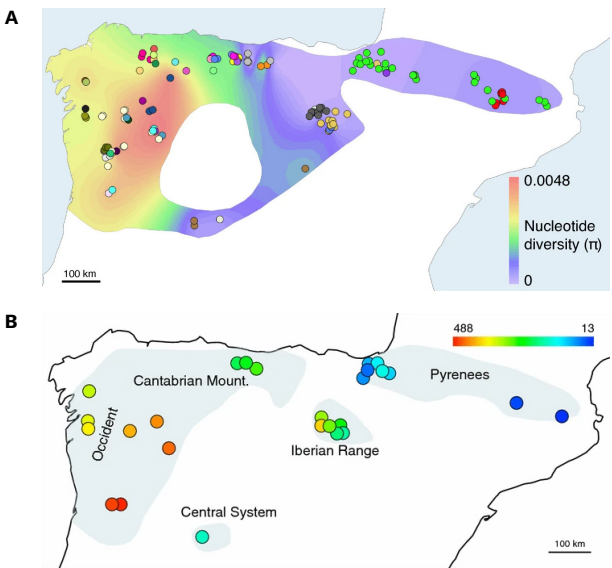


Figure 7. Genetic diversity of the Iberian desman. a) Map of mitochondrial nucleotide diversity (π) (from Igea et al. (2013)). b) Nuclear heterozygosity rates (from Querejeta et al. (2016)).

0.000488) and a progressive decline towards the Pyrenees, where an extremely low value of 0.000013 was recorded (Figure 7b). These values yield an average of 0.000246, corresponding to 246 heterozygous positions per million bases, which represents one of the lowest heterozygosity rates reported in mammals (Querejeta et al., 2016).

These data were recently assessed through the sequencing (121X) and re-sequencing (10.3X–33.3X) of the complete genome of *G. pyrenaicus* using shotgun libraries (Escoda & Castresana, 2021). This study confirmed the evolutionary significant units previously delineated. Additionally, a wide range of heterozygosity rates was detected among different populations, with a maximum of 459 SNPs/Mb in a specimen from the western distribution and a minimum of 12 SNPs/Mb in an individual from the Eastern Pyrenees, which represents the lowest heterozygosity rate recorded in mammals as of the date of the study. The individual from the Central System also exhibited one of the lowest recorded values, with 116 SNPs/Mb. The differences in heterozygosity rates among populations reflect distinct evolutionary histories, as observed in PSMC plots, within a general trend of population decrease (Figure 8). These reveal significant fluctuations in effective population size and divergent demographic trajectories among populations, characterized by bottlenecks and population expansions. These dynamics partly explain the reduction in genetic diversity observed (Escoda & Castresana, 2021). Finally, most individuals showed a high proportion of long ROHs, indicative of recent inbreeding (Escoda & Castresana, 2021). This phenomenon is linked to the Iberian desman’s limited dispersal ability and the presence of natural and artificial barriers that impede population connectivity, leading to isolation and elevated inbreeding levels (Escoda et al., 2017, 2019).

the highest value in lineage A1 (π : 0.036) and a gradual decrease towards lineage B2, which showed a value an order of magnitude lower (π : 0.0004) (Figure 7a). The area with the highest contemporary genetic diversity coincided with the main inferred glacial refuge (Figure 6b), suggesting that the genetic structure of the Iberian desman is strongly influenced by the history of Pleistocene glaciations. A similar pattern was found by Querejeta et al. (2016), who studied heterozygosity using nuclear genomic data, reporting higher values for individuals from the western region (maximum:

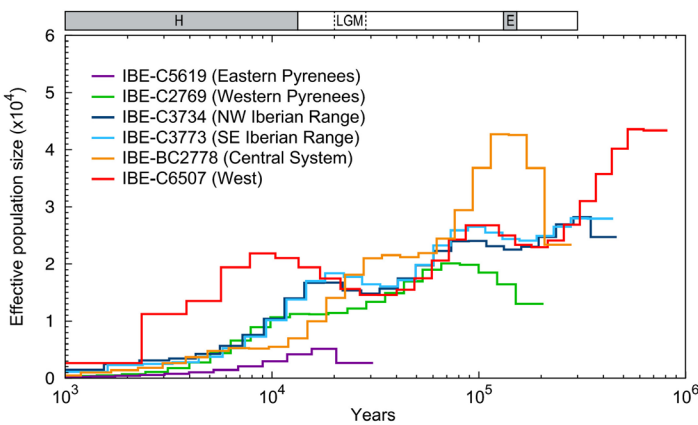


Figure 8. Historical effective population size of the Iberian desman individuals inferred by PSMC. The last two interglacial periods, Holocene (H) and Eemian (E), are indicated with grey boxes and the last glacial maximum (LGM) with dashed lines (from Escoda & Castresana (2021)).

4.5. Current Conservation Status

The main threats to this species are habitat degradation and fragmentation, including the modification of natural river courses and banks, the construction of dams and reservoirs, the deterioration of the benthos, and water pollution (Nores, 2017). These factors have caused a widespread and sustained decline in its populations in recent years. In Spain, its occurrence is estimated to have decreased by 87% over the last 30 years. In Andorra, a 70% reduction has been recorded compared to 24 years ago. Additionally, its distribution has shrunk by 60% in the past two decades in northern Portugal and the French Pyrenees (Quaglietta et al., 2024). The most dramatic decline has occurred in the population of the Central System, where a 97% reduction was recorded in 2014, evidencing a process that began around the 1960s. This significant regression is mainly attributed to the fragmentation of subpopulations, the deterioration of water quality, and the introduction of the invasive American mink (*Neovison vison*), which preys on the species (Gisbert & García-Perea, 2014).

Various conservation programs have been implemented across the species' entire distribution range. Of particular relevance were the six LIFE programs developed: LIFE Conhecimento e Gestão do Património Natural (Portugal, 1994–1997), LIFE+ MARGALULLA - LIFE 09 NAT/ES/000514 (Galicia, 2010–2016), LIFE DESMANIA - LIFE11 NAT/ES/000691 (Castilla y León + Extremadura, 2012–2018), LIFE IREKIBAI - LIFE 14 NAT/ES/000186 (Navarra, 2015–2020), LIFE MEDWETRIVERS - LIFE11 NAT/ES/000699 (Castilla y León, 2012–2018), and LIFE+ Desman - LIFE 13 NAT/FR/000092 (France, 2014–2020), which are a European Union funding instrument for environmental support, nature conservation, and climate action projects. Currently, the Iberian desman is strictly protected under the Bern Convention and the EU Habitats and Species Directive, and it benefits from national protection laws in all four countries where it is distributed. In Spain and France, it is classified as Vulnerable, except for the population in the Spanish Central System, which is categorized as Endangered. In Portugal, it is also classified as Endangered, while in Andorra its status was recently updated to Critically Endangered (Quaglietta et al., 2024). Among the actions proposed in Spain to prevent the extinction of the desman, in addition to habitat improvement, translocations (population reinforcement and reintroduction) have been proposed

alongside captive breeding programs (Gisbert et al., 2024), for which genomic tools are essential to understand population genetic characteristics, kinship relationships, as well as aspects of their diet and the pathogens affecting them. Additionally, the species has been proposed to be classified as Endangered at the national level in Spain and as Critically Endangered for the populations in the Central System and the Iberian System (Gisbert et al., 2024).

4.6. Habitat and trophic ecology

The Iberian desman inhabits mountain rivers and streams with cold, clean, and well-oxygenated waters, characterized by a constant flow throughout the year (Nores, 2017; Peyre, 1956) (Figure 9). For this reason, it is mainly found in areas with significant snow cover in the headwaters of rivers, which guarantees a regular flow. It is absent in basins with seasonal droughts and pluvial regimes, which explains its preference for oceanic climates over Mediterranean ones (Aymerich & Gosálvez, 2002). Its presence is not so much determined by altitude (it can be found from sea level to altitudes above 2,500 meters) as by the slope of the rivers, their depth (shallow to moderate), and the high-water flow (Charbonnel et al., 2015; Esnaola, González-Esteban, et al., 2018). A key element of its habitat is the availability of suitable shelters, such as crevices and holes among stones and roots along stream banks. Additionally, it selects rivers with a high presence of rocks and a scarcity of fine sediment (Nores, 2017). Due to its wide altitudinal range, the environments inhabited by this species include a great diversity of riparian and surrounding vegetation (Palmeirim & Hoffmann, 1993).



Figure 9. Iberian desman habitat from the Arga river in Quinto Real, Navarra (Photo by José Castresana).

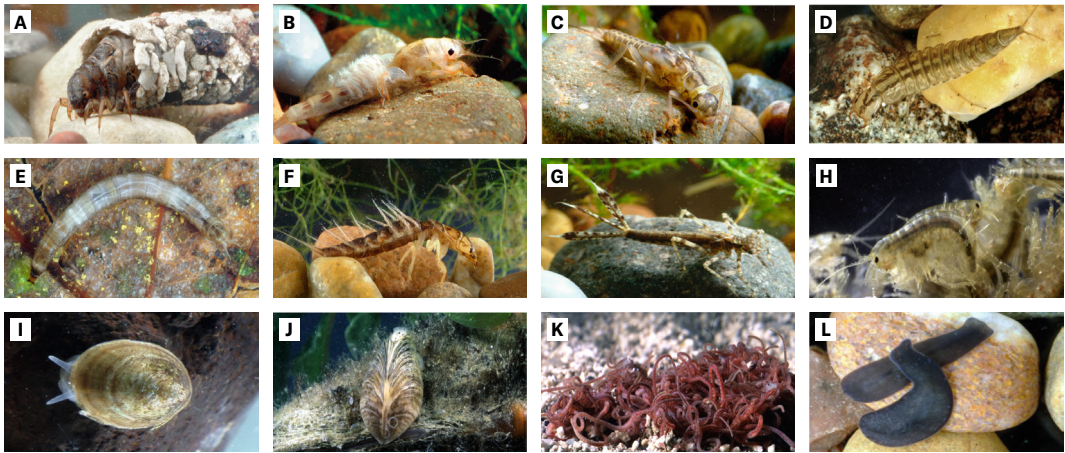


Figure 10. Examples of some components of the diet of the Iberian desman and other aquatic invertebrates commonly found in freshwater ecosystems of Europe: a) Trichoptera (*Potamophylax nigricornis* larva), b) Ephemeroptera (*Ephemera danica* larva), c) Plecoptera (*Isoperla* sp. larva), d) Coleoptera (*Dytiscidae* larva), e) Diptera (*Pediciidae* larva), f) Megaloptera (*Sialis* sp. larva), g) Odonata (*Platycnemidae* larva), h) Crustacea (*Gammarus* sp.), i) Gastropoda (*Ancylus fluviatilis*), j) Bivalvia (*Dreissena polymorpha*), k) Oligochaeta (*Tubifex* sp.), l) Planaria (*Polycelis nigra*) (from Kriska (2023)).

According to (Peyre, 1956), another key requirement for the presence of *G. pyrenaicus* is the abundance of insect larvae and small crustaceans, which make up its main food source (Figure 10). In an analysis of stomach contents from individuals captured in the eastern

Pyrenees, Peyre observed primarily benthic macroinvertebrate larvae from the orders Trichoptera and Plecoptera, crustaceans from the genus *Gammarus*, and terrestrial insects, which the animal likely captured on the banks of streams. On the other hand, Castién & Gosálbez (1995) conducted a detailed study of the digestive tracts of 46 specimens from the western end of the Pyrenees and compared the results with the availability of prey in the aquatic environment throughout the year. These authors concluded that the majority of the desman's diet was composed of aquatic prey. In terms of the number of prey items found, 49.6% were Ephemeroptera larvae, 23.1% were Trichoptera larvae, 13.3% were Diptera larvae, and 9% were Plecoptera larvae. The rest of the diet included less abundant prey such as Oligochaeta, Gastropoda, Chilopoda, Opiliones, larvae and adults of Coleoptera, adult Diptera, Amphipoda, and trout (*Salmo trutta*) eggs.

Regarding seasonality, Ephemeroptera and Trichoptera larvae represented the largest proportion of the diet throughout the year, although Plecoptera larvae were significant in the spring and Diptera larvae in the fall. As for the abundance of dietary elements in the river, it was observed that Ephemeroptera and Trichoptera larvae were positively selected. Trichoptera larvae are relatively large and immobile, while Ephemeroptera larvae are smaller but very abundant. On the other hand, a negative selection of Amphipoda (all from the family Gammaridae) was observed, which contrasted with Peyre (1956)'s findings. This difference was likely due to the greater escape ability of these crustaceans and their high chitin content relative to their available biomass. This result was consistent with the desman's need to optimize its energy intake. Due to the small size of the Iberian desman, the energy required to

maintain homeothermy in an aquatic environment is high relative to its body weight. Thus, the need to obtain the highest biomass in proportion to the time spent foraging seems to be a determining factor in its prey selection (Castián & Gosálbez, 1995).

Other multiple studies have been conducted based on the morphological analysis of invasive samples (digestive tracts) in different populations, describing the desman as a generalist predator of rheophilic species, with a preference for Trichoptera, Ephemeroptera, and Plecoptera, but with a substantial consumption of other insect orders (Diptera, Coleoptera), as well as annelids, crustaceans, and gastropods (Castián & Gosálbez, 1999; Santamarina, 1993). Recently, López-de Sancha et al., (2022) conducted an analysis based on the morphological identification of prey from 28 fecal samples of *G. pyrenaicus* from a population in the western Pyrenees. They compared its diet with that of two other semi-aquatic insectivorous vertebrates that share habitat and food resources: the white-throated dipper (*Cinclus cinclus*) and the Eurasian water shrew (*Neomys fodiens*). The study revealed clear trophic overlap and interspecific competition between the two mammals, with lesser competition with *C. cinclus*. The main components of the desman's diet were Ephemeroptera, Trichoptera, Diptera, Plecoptera, and to a lesser extent, Coleoptera. *G. pyrenaicus* showed a tendency to feed on prey in aquatic life stages, with low mobility and energetically profitable. Its morphological and physiological adaptations explain its prey selection. Additionally, the desman rejected families with low energy yield and avoided terrestrial prey, although the latter constituted a small percentage of the diet. Also, non-preferred but abundant prey, such as Oligochaeta worms, were actively consumed. The authors emphasize the difficulty of identifying soft-bodied prey in morphological studies and recommend the development of genetic studies to complement them, improving accuracy.

4.6.1. Genetic approaches in the study of the diet of the Iberian desman

In recent years, molecular analyses based on metabarcoding of fecal DNA fragments and next-generation sequencing have been developed to study the diet of the Iberian desman. These methods allow for both the identification of the species to which the feces belong and the components of its diet, preventing errors in fecal sampling, reducing the time required for morphological analyses, and increasing the sensitivity of the study. Gillet et al. (2015) and Biffi et al. (2017) applied this approach to the Iberian desman in the Pyrenees. Both studies identified Trichoptera, Plecoptera, Ephemeroptera, and Diptera as the main dietary groups. Gillet et al. (2015), working with 16 samples from different localities in the French Pyrenees, used COI gene sequencing and compared the results with the BOLD database, managing to identify several species within these orders, as well as Coleoptera and Amphipoda in lesser abundance. Biffi et al. (2017), on the other hand, conducted an extensive sampling of 385 samples covering three hydrographic regions of the French Pyrenees. In this study, in addition to the main groups mentioned, they identified a greater diversity of prey, including additional insect orders such as Coleoptera, Orthoptera, Neuroptera,

Hemiptera, Lepidoptera, Odonata, and Hymenoptera, as well as representatives of Malacostraca (Isopoda, Decapoda, Amphipoda), Diplopoda, Lissamphibia, Arachnida, Clitellata (Oligochaeta), and Gastropoda, in many cases identifying them down to the genus level. In particular, Amphipoda (*Gammarus*) had a significant representation in one of the regions studied. According to the authors (Biffi et al., 2017), the high number of identified preys highlights the efficiency of molecular analyses in detecting taxa that are difficult to identify using morphological methods and underscores the adaptability of the Iberian desman to different trophic environments. Although it shows a preference for certain groups, the variety of less common prey suggests a generalist diet, possibly influenced by aquatic stressors or seasonal effects, which increase the availability of both aquatic and terrestrial prey, with the latter being significantly represented. The results suggest that the desman's diet adjusts on a local scale according to resource availability and accessibility, reflecting a flexible and opportunistic foraging strategy.

Esnaola et al. (2018) focused on optimizing the methodology for analyzing the diet of the desman. To achieve this, they analyzed 24 fecal samples from populations in the northernmost part of the western Pyrenees (Basque Country), using five pairs of primers to amplify different regions of the COI gene and compared the sequences with the GenBank nucleotide database. They identified the main prey groups mentioned earlier, with the addition of Salmonidae, Nematoda, Psocoptera, and Mollusca. They found that the different primer sets accurately identified the most common taxa, but varied in detecting less abundant groups. This demonstrated that the choice of molecular markers, and even primers targeting different regions of the same gene, can significantly influence the characterization of diet and the interpretation of trophic ecology, especially in generalist predators. Therefore, the authors emphasize the importance of combining different sets of primers to detect the widest possible range of potential prey and avoid information loss.

In an effort to improve sequence identification in metabarcoding studies, Hawlitschek et al. (2018) developed a protocol based on nested PCR amplification of a fragment of the COI gene, which enhances specificity. The sequences were classified both by constructing a phylogenetic tree with known sequences of potential aquatic prey of *G. pyrenaicus*, as well as using BLAST with the BOLD and GenBank databases. Additionally, they analyzed the impact of the condition of fecal samples (fresh vs. old) on DNA degradation and its effect on diet evaluation. For the analysis, they used 19 fresh samples, collected just after defecation, and 22 old samples, collected from riverbanks and stones in various regions of the Iberian Peninsula, primarily from Zamora. The results from BLAST and the tree-based approach were more similar when the sequences were genetically closer to the reference sequences. The tree-based approach was more accurate in identifying divergent sequences, although its effectiveness was limited by the need to use a reference tree. In contrast, BLAST benefited from a larger database, which provided an

advantage given the wide range of potential prey species for the desman. On the other hand, fresh samples generated more useful reads and more identifications than old samples, although the latter still produced enough reads to identify a significant number of prey items. The overall taxonomic composition of the diet was consistent with previous studies in terms of detecting freshwater invertebrates, including the aforementioned insect orders, as well as Crustacea and Annelida. Additionally, unexpected organisms such as Acari, Hemiptera, Hymenoptera, and Nematoda were identified. A variation in composition was observed between the samples, which could indicate different feeding strategies in different areas or at different times of the year (Hawlotschek et al., 2018).

Fecal sample metabarcoding offered several advantages over traditional morphological methods, primarily because it is a non-invasive approach that allowed the analysis of complex samples containing DNA from multiple species without the need for prior isolation. However, it also presented certain limitations. As it is a PCR-based method, it could fail to amplify certain species in the diet, especially when only a single primer pair is used, in addition to the potential bias caused by the differential affinities of primers to the alignment regions. This issue is further exacerbated by the fact that the quality and quantity of DNA in feces are often low due to digestive processes prior to, and decomposition after, defecation (Esnaola et al. 2018, Hawlotschek et al., 2018). On the other hand, for feces collected from the river, it is important to consider potential contaminations. These include environmental DNA or microorganisms, especially relevant in humid habitats, as well as insects involved in the decomposition of feces (Hawlotschek et al., 2018). This is particularly significant for insectivorous species, where it can be challenging to distinguish between prey and insects contributing to decomposition (López-de Sancha et al., 2022). Finally, there was the need for a complete and reliable database. Although freshwater invertebrates are relatively well-represented, not all taxonomic groups and geographical regions are equally covered (Hawlotschek et al., 2018). Nevertheless, the different studies agreed on the identification of the main components of the desman's diet and demonstrated greater sensitivity in detecting rare or unexpected elements. This showed that, although the desman has a clear preference for aquatic invertebrates, it exhibits opportunistic behavior by resorting to other prey depending on their availability and accessibility.

4.7. Parasites and pathogens

The importance of identifying the parasites and pathogens affecting the Iberian desman lies in its status as an endangered species. Infectious diseases represent a significant risk factor for the survival of wildlife species, especially those with small, fragmented populations and low genetic variability, such as the desman. Under these conditions, susceptibility to diseases increases, making infections an additional

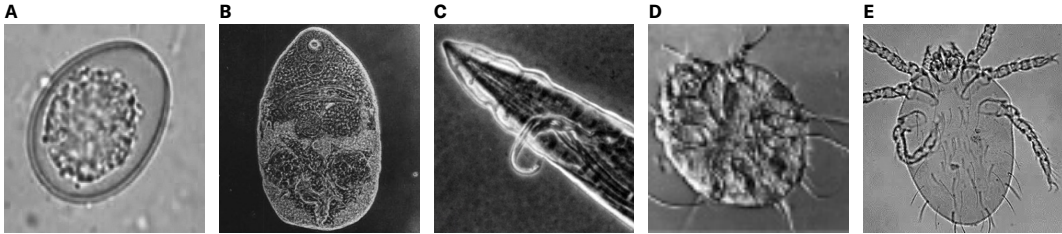


Figure 11. Examples of parasites detected in the Iberian desman, endoparasites: a) *Eimeria* spp. oocysts (from Bangoura & Dauschies (2018)), b) Trematode *Maritrema pyrenaica* (from Jourdan (1979)), c) Nematode *Paracuarria hispanica* (from Alvarez et al. (1994)), d) Sarcophagid mite (from Mehlhorn (2016)), e) Trombidiform mite (from Peng et al. (2016)).

factor that raises the risk of extinction by reducing individual fitness and survival or causing population declines, particularly when combined with other environmental stressors (Pedersen 2007, Kolodziej 2019).

Among small mammals, studies of parasites and pathogens have primarily focused on rodents and shrews, the latter belonging to the order Eulipotyphla (Barandika et al., 2007; Hammerl et al., 2017; Haring et al., 2023). These small mammals have been shown to be significant reservoirs of zoonotic diseases, which can be transmitted directly or indirectly, often via arthropod vectors, to domestic animals and eventually to humans (Tian et al., 2023). In the context of significant anthropogenic land-use changes, including agricultural and urban frontier expansion, where interactions between wildlife, livestock, and humans are increasingly frequent, this underscores the importance of monitoring parasites and infectious diseases in wildlife, not only for their conservation but also for their critical relevance in public health (Cantas & Suer, 2014; White & Razgour, 2020).

The Iberian desman is susceptible to a variety of parasites, both internal (endoparasites) and external (ectoparasites) (Nores, 2017) (Figure 11). The first reports of endoparasites affecting the Iberian desman were digeneans (a subclass of Trematoda): *Omphalometra flexuosa* var *peyrei*, *Maritrema pyrenaica* and *Mathovius galemydis*. *O. flexuosa* var *peyrei* was found in populations from the French Pyrenees and the mountains of the Central System and was also detected in crustaceans (*Gammarus pulex*) (Timon-David, 1960; Vaucher, 1975). Similarly, *M. pyrenaica* was found in populations from the French Pyrenees, where it affects *G. pyrenaicus* and *N. fodiens*, involving two intermediate hosts: the mollusk *Bythinella reyniesii* and the crustacean *G. pulex* (Deblock & Combes, 1965; Jourdan, 1979). Finally, *M. galemydis* was reported in different populations from Spain (Mas Coma et al., 1985).

Additionally, Ribas et al. (2009) identified two digenean parasites with aquatic life cycles: *Pseudocephalotrema pyrenaica* and *Skrjabinophyetus neomydis*. They were found in aquatic invertebrates, such as the mollusc *Bythinella rufescens* and insect larvae (ephemeropterans and trichopterans), which are an important part of the desman's diet. These invertebrates were collected in the Pyrenees, where

the desman inhabits. It is believed that the desman is a potential host for these parasites because it shares habitat and diet with other semi-aquatic mammals, such as water shrews (*N. fodiens*), where these parasites has already been reported (Ribas et al., 2005).

Studies by Alvarez Mascato, (1990) and Alvarez et al. (1994) on the parasitology of the Iberian desman revealed the presence of several nematodes affecting this species. In Alvarez Mascato (1990), three nematodes were identified in desman populations from Galicia: *Parastrongyloides winchesi*, *Paracuaria* sp., and *Eucoleus* sp. In Alvarez et al. (1994), a new species, *Paracuaria hispanica*, was described, which was found in the stomach of 45% (9 out of 20) of the individuals studied in populations from central and northern Spain. The authors define the desman as the definitive host of this nematode species and suggest that aquatic insect larvae and crustaceans that form part of its diet could act as intermediate hosts.

Regarding ectoparasites, only three Acari species have been reported so far: *Eadiea desmanae* (Lukoschus, 1969), *Lynxacarus dubinini*, and *Orycterovenus galemys* (Cordero del Campillo et al., 1994).

Recently, Ripa et al. (2023) conducted a study using molecular methods based on fecal samples to detect pathogens in a population of the Central System of the Iberian desman. The study focused on parasites previously described in this species, such as the trematode *Omphalometra flexuosa*, as well as other pathogens commonly affecting small mammals. The detection of these pathogens was carried out through real-time PCR (qPCR), mainly using commercial kits to identify nematodes from the family Ascarididae (covering only the genera *Baylisascaris*, *Parascaris*, *Ascaris*, *Toxascaris*, and *Toxocara*), coccidia such as *Cryptosporidium* spp. and *Eimeria* spp., and bacteria such as *Salmonella enterica*, *Clostridium perfringens*, *Staphylococcus* spp., *Leptospira* spp., and *Listeria* spp.. For the identification of *O. flexuosa*, they developed a real-time PCR assay targeting a fragment of the 28S rDNA gene. The highest prevalence detected was for *Eimeria* spp., found in 92.3% of the samples (191 out of 207), followed by *Staphylococcus* spp. (14.5%, 30 samples), *Omphalometra* spp. (8.2%, 17 samples), *Salmonella* spp. (4.3%, 9 samples), *Leptospira* spp. (2.4%, 5 samples), *Cryptosporidium* spp. (1.15%, 3 samples), and *Listeria* spp., detected in only one sample (0.5%). All samples tested negative for nematodes from the Ascarididae family and for *Clostridium perfringens*. Excluding *Eimeria* spp., which was present in nearly all samples, 22.2% of the samples showed at least one infection from the other parasites or pathogens, 3.9% had two infections, and 0.5% had triple infections. Regarding the temporal and spatial analysis, no significant differences were found between the different years, but differences were observed in the presence of *Eimeria* spp., *Salmonella* spp., and *Staphylococcus* spp. between sampling sites, suggesting a concentration of these infections in specific areas.

This work represented a significant advance in understanding the parasites and pathogens that affect *G. pyrenaicus* through the use of molecular methods and non-invasive samples. However, the targets studied were highly limited by the availability

of commercial kits. Additionally, due to the method's directed specificity, it was only possible to identify the taxa previously selected for the assay, without detecting closely related organisms. An example of this is the case of nematodes from the Ascarididae family, where the test used only covered five specific genera, without including the nematodes previously described for the desman, so their presence cannot be ruled out. The issue becomes more critical when even members of the same species cannot be detected, as in the case of *Clostridium perfringens*, whose assay was specifically designed to detect the beta toxin; therefore, only the presence of this particular strain could have been ruled out.

Overall, there is a clear need to expand genetic studies to gain more knowledge about the health of *G. pyrenaicus*. Specifically, new genomic methods are needed for the detection of parasites and pathogens that allow for broader screening, both for previously reported taxa and for the identification of new parasites and pathogens (ENETWILD-consortium et al., 2023). The conservation of biodiversity depends on a deep understanding of the interactions between species and their ecosystems. Among these factors, parasites and pathogens have a significant impact on the populations and communities of their hosts. Therefore, studying the relationships between the parasites and pathogens affecting the Iberian desman is essential for its conservation.

III. OBJECTIVES

The development and application of genomic methods has proven to be a powerful tool for conservation programs of endangered species. These methods have primarily focused on population studies, leaving significant potential for their application in generating ecological and health information about species. A substantial contribution can be achieved through the use of fecal samples in combination with metagenomic approaches to maximize the extraction of information from such samples. In this context, the main objective of this thesis is to develop novel pipelines and optimize bioinformatic methods to extract metagenomic information from feces. As a case study, the endangered semi-aquatic mammal, *Galemys pyrenaicus*, was used in order to explore aspects of its population genomics, trophic ecology, and the presence of potential parasites and pathogens that may impact its health.

To achieve this general goal, the following specific objectives were proposed:

1.

Obtain genomic information from shotgun metagenomic libraries based on fresh fecal samples. Use samples of the Iberian desman (*Galemys pyrenaicus*) from the Central System population, considered the most threatened population of this species, as a case study. Focus on both endogenous and exogenous DNA to provide a foundation for subsequent analyses of the species' biology.

2.

Determine the global composition of DNA present in the feces of *G. pyrenaicus*, including endogenous and exogenous DNA classified into high-rank taxonomic groups (e.g., Bacteria, Archaea, Eukaryotes, and Viruses), using bioinformatic tools for taxonomic classification and DNA sequence databases.



The Iberian desman (*Galemys pyrenaicus*). Photo by Jorge González-Esteban, from Esmola et al. (2021).

3.

Evaluate the utility of endogenous DNA for determining the sex of individuals and conducting population genomic analyses.

4.

Develop a novel bioinformatic method for the characterization of exogenous eukaryotic DNA present in feces, through taxonomic classification adapted to the characteristics of metagenomic libraries and available reference sequences.

5.

Build a curated database of mitochondrial DNA sequences to enable precise taxonomic classification of the reads obtained from feces, facilitating diet and parasitology analysis.

6.

Use other available sequence databases for higher-level taxonomic classification through homology searches in the context of diet and parasitology analysis.

7.

Analyze the trophic ecology of *G. pyrenaicus* using exogenous DNA sequences taxonomically classified into elements of its diet.

8.

Detect potential parasites that may impact the health of *G. pyrenaicus* by determining exogenous DNA sequences of parasitic taxa.

9.

Develop a novel bioinformatic method for the identification at the species level of pathogenic bacterial in the feces of *G. pyrenaicus*, using complete reference genomes, and adapting the method to the unique challenges associated to non-model species.

The development of specific objectives 1 to 8 is presented in the subsections titled "1. DNA composition, endogenous DNA, diet and parasites from Iberian desman feces", within the Methods, Results, and Discussion sections of the thesis. Specific objective 9 is addressed in the subsections titled "2. Species-level identification of bacterial pathogens", also within those same sections. Finally, an integrative discussion is provided, in which the newly developed methods, their main differences, the challenges addressed, and their potential applications in future studies are analyzed in relation to the general objective of the thesis.

IV. METHODS

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES

The following methods correspond to specific objectives 1 to 8.

1.1. Fecal samples collection and DNA extraction

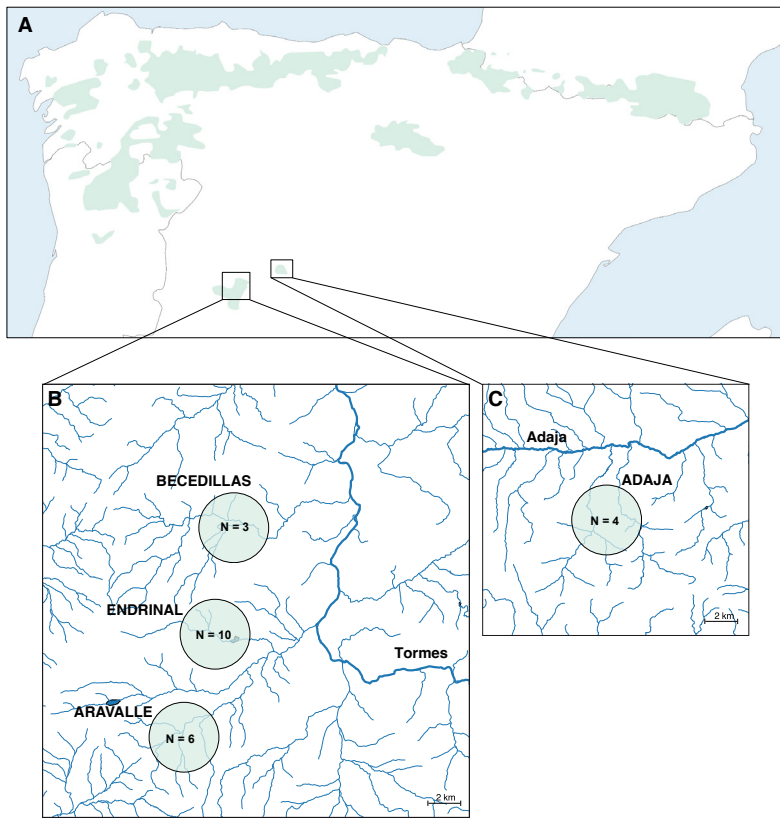
The analyzed samples originate from the population of the Iberian desman (*Galemys pyrenaicus*) located in the Central System, in the center of the Iberian Peninsula, which has been proposed to be classified as Critically Endangered (Gisbert et al., 2024). A total of 23 fecal samples (Table S1) were obtained during the summer months of 2018 and 2019 from four different subpopulations: Aravalle, Becedillas, Endrinal, and Adaja (Figure 12), which were defined based on the important hydrological units in the area. These samples were collected during the capture of individuals immediately after deposition, and placed in tubes with ethanol. This approach preserves DNA integrity and minimizes the risk of environmental contamination that can occur with fecal samples (Hawlitshcek et al., 2018; Oliveros et al., 2023). The work of capturing individuals was part of a conservation program independent of this study promoted by the Ministry of Environment through the Biodiversity Foundation, the Duero River Basin Authority, and the Autonomous Government of Castilla y León through the Patrimonio Natural Foundation, in Spain.

DNA was extracted using the QIAamp DNA Mini Kit (QIAGEN), following manufacturer's instructions, and quantified using a Qubit fluorometer (Thermo Fisher Scientific) with the Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific).



Iberian desman feces. Photo by José Castresana.

Figure 12. Geographic sample distribution. (a) Map of the northern part of the Iberian Peninsula showing the current distribution of the Iberian desman (*Galemys pyrenaicus*) (Quaglietta et al. (2024)) and the origin of the studied subpopulations, which come from the Central System. The number of samples (N) and the location of each studied subpopulation are indicated: (b) Aravalle, Becedillas, and Edrinal; and (c) Adaja, according to the sub-basin where they are located and the main rivers to which they belong: Tormes and Adaja.



1.2. Shotgun-metagenomics library construction and sequencing

Shotgun metagenomic libraries were constructed using the NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs). Extracted DNA (26 µL per sample) was enzymatically fragmented at 37°C for 15 minutes. Specific Illumina adapters were ligated, and fragments of 150-250 bp insert size were selected using NEBNext Sample Purification Beads (New England Biolabs). Each sample was then indexed and amplified through 12 cycles of PCR, cleaned, and quantified using the Qubit fluorometer. Fragment size distribution (270 – 370 bp) was assessed by E-gel EX 2% agarose gel electrophoresis (Invitrogen). Finally, equimolar amounts of each library were pooled and sequenced on an Illumina platform.

1.3. Sequencing data processing: quality control and separation of endogenous and exogenous reads

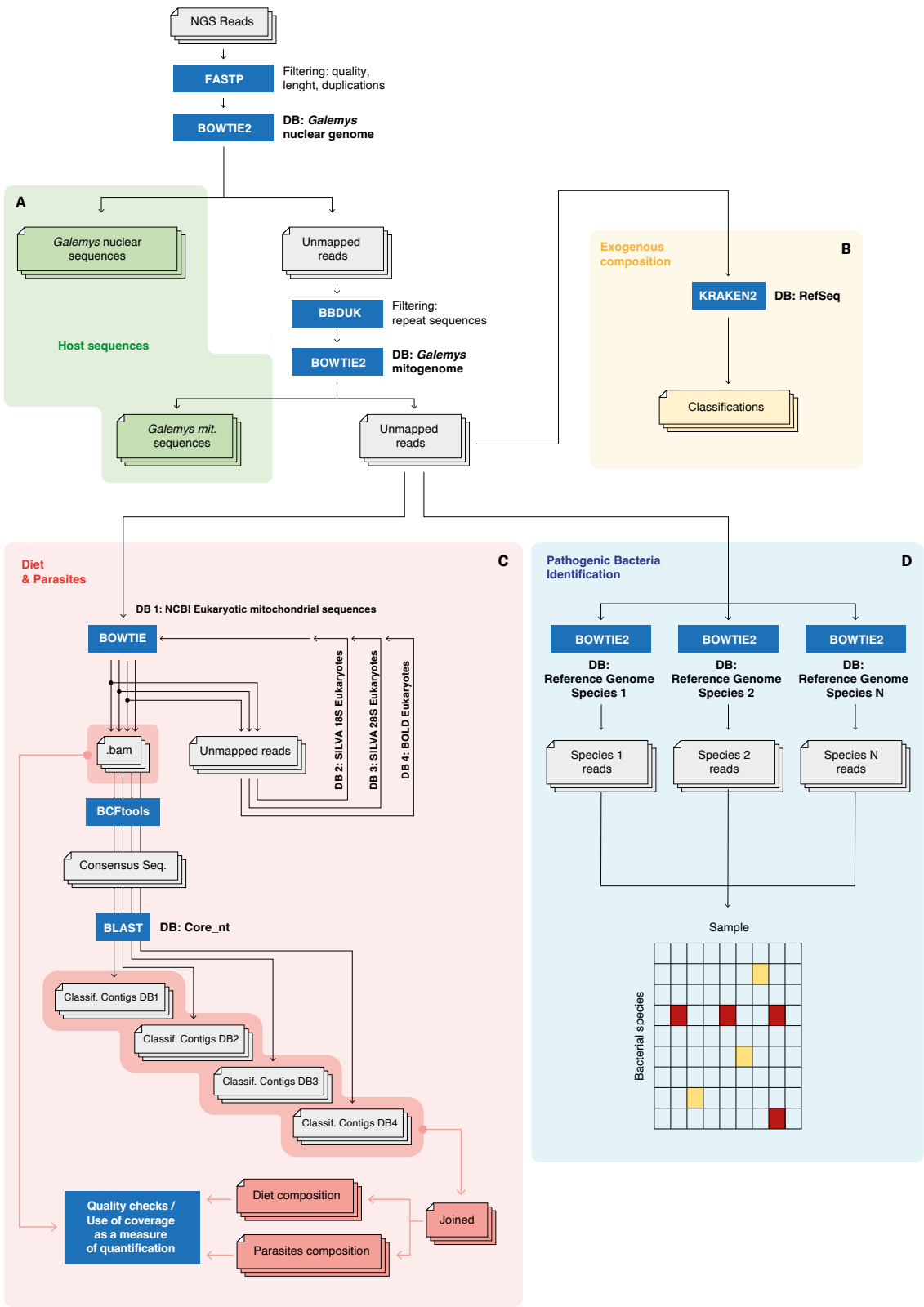
A bioinformatics pipeline was developed to process sequences, incorporating tools for quality filtering of reads, mapping, assembly, and taxonomic classification, using various reference sequence and genome databases (Figure 13). The pipeline began by removing low-quality, duplicate, and reads shorter than 100 bp for each sample using FASTP version 0.23.2. (Chen, 2023).

The resulting reads were aligned in pairs with the nuclear reference genome of *G. pyrenaicus* (Escoda & Castresana, 2021) using Bowtie2 version 2.5.0 with the --end-to-end and --sensitive options (Langmead & Salzberg, 2012). In Bowtie2, the "end-to-end" alignment mode and the preset "sensitive" configuration are combined to optimize alignment accuracy and sensitivity. The "end-to-end" mode requires that the entire read aligns from end to end without any trimming (or "soft clipping") at the ends, minimizing the risk of partial or spurious alignments and increasing precision by ensuring complete matches (Langmead & Salzberg, 2012). The "sensitive" configuration adjusts parameters to enhance alignment sensitivity by using a seed length of 22 nucleotides (-L 22) without mismatches in the initial seed (-N 0), ensuring exact matches at the start of the alignment. It also allows up to 15 consecutive failed seed extensions (-D 15) before testing another alignment site or discarding the read, improving the detection of alignments in sequences with slight variations and increasing the likelihood of finding the best possible match. Additionally, output options were included to manage both aligned and unaligned reads. Aligned reads were directly saved in SAM format using the -S option together with the --no-unal option to exclude unaligned reads. The --un-conc-gz option was used to save unaligned read pairs in a concordant manner (i.e., those that do not align with the expected orientation and distance between mates), which constitute the input for the next step. The SAM files were then transformed into BAM format using the SAMtools package v1.9 (H. Li et al., 2009). During this process, the paired reads were sorted by name, mate information was fixed, and PCR duplicates were marked to avoid them in subsequent analyses. Finally, the BAM files were sorted by genomic position and indexed.

The set of unmapped reads was then processed with BBDUK version 39.01 (<https://jgi.doe.gov/data-and-tools/software-tools/bbtools/>) to remove sequences containing repetitive motifs. An entropy filter was applied; this parameter measures the complexity of each read and ranges from 0 to 1. A sequence composed entirely of a single base would have an entropy of zero, while a completely random sequence would have an entropy of 1. Specifically, reads with an average entropy below 0.9 were filtered out. In this way, the largest possible number of sequences was mapped to the nuclear genome of *Galemys*, including low-complexity sequences that are considered legitimate parts of this genome, such as microsatellites or other repetitive sequences that may be part of the species' nuclear genome. This prevented their removal at an early stage of processing, allowing for a more accurate quantification of the proportion of this source in fecal samples. However, these sequences were discarded in subsequent analyses.

Following the pipeline, the reads not removed by BBDUK were aligned to the reference *G. pyrenaicus* mitochondrial genome (Escoda & Castresana, 2021), also using Bowtie2 with the same configuration as before. The unmapped reads constituted the exogenous set of each fecal sample. This process thus generated three sets of reads: endogenous nuclear, endogenous mitochondrial, and exogenous.

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.3. SEQUENCING DATA PROCESSING: QUALITY CONTROL AND SEPARATION OF ENDOGENOUS AND EXOGENOUS READS



1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES

← Figure 13. Pipeline for metagenomic content analysis of fecal samples

The developed pipeline consists of several steps: first, a quality filtering of raw reads using Fastp, followed by the alignment of sequences to the *G. pyrenaicus* reference genome with Bowtie2 to extract nuclear reads. The unmapped reads are filtered to remove sequences with repetitive motifs using BBduk and subsequently aligned to the *G. pyrenaicus* mitochondrial reference genome to extract mitochondrial-origin reads. This process results in two sets of endogenous reads: nuclear and mitochondrial (a), as well as a set of exogenous reads. From this point, further analyses are conducted: (b) the global composition of the exogenous content is characterized using Kraken2 with the RefSeq database. (c) The diet and parasite composition are identified using a novel method. This method involves the alignment of endogenous reads with Bowtie2 against four databases: (DB1) a curated mitochondrial sequence database from NCBI containing sequences over 500 bp from eukaryotes (excluding plants), (DB2) the Silva 18S

rDNA database for eukaryotes, (DB3) the Silva 28S rDNA database for eukaryotes, and (DB4) the BOLD eukaryotic database. Mapped reads are separated from unmapped ones after alignment with each database, following the mentioned order. The unmapped reads are aligned to the next database sequentially. Mapped reads are then assembled into consensus sequences longer than 500 bp using BCFtools, which are subsequently classified taxonomically using BLAST against the Core Nucleotide database as a reference. Finally, the classifications obtained from each database are combined to describe the diet and parasites. A series of quality checks is performed to ensure the accuracy and reliability of the taxonomic assignments. (d) The identification of pathogenic bacteria using a new method that involves aligning the exogenous sequence set of each sample to individual bacterial genome databases with Bowtie2. Presence is defined using breadth of coverage as a threshold parameter.

1.4. Genomics analysis from endogenous data

1.4.1. Population structure analysis

To evaluate the potential of DNA obtained from fecal samples for genomic analyses, the ANGSD version 0.941 software was used (Sand Korneliussen et al., 2014). This software estimates single nucleotide polymorphisms (SNPs) from BAM files based on genotype likelihoods calculated from aligned reads and associated mapping and sequencing quality scores, integrating information from all samples and accounting for the statistical uncertainties inherent in low-coverage data. The following options we used: -GL 1 (genotype likelihood model), doMajorMinor 1 (to infer major and minor alleles per site from the genotype likelihoods), -doMaf 1 (to estimate allele frequencies per site with a fixed major and minor alleles), -SNP_pval 1e-6 (to keep variant sites with a p-value less than 1e-6, indicating significant variants), and -doCounts 1 (to calculate the counts of each of the four possible bases per site). To focus on well-assembled regions of the reference genome, we restricted the variant site search to autosomic scaffolds greater than 40,000 bp using -rf. Quality filters were applied with -minMapQ 20 (to set the minimum mapping quality), -remove_bads 1 (to exclude not primary, failure and duplicate reads (default)) and -only_proper_pairs 1 (to include only pairs of read with both mates mapped correctly (default)). The output was specified as a covariance matrix with -docov 1. Using the same software, a principal component analysis (PCA) was performed using the covariance matrix to verify that the samples clustered according to their subpopulation of origin.

1.4.2. Genetic sex determination

Sex determination for each individual was estimated based on the ratio of the mean depth of coverage values for the Y chromosome (Y) and the entire nuclear genome (NG), an approach similar to that used in previous studies (Alvarez-Estape et al., 2023; Cabrera et al., 2022), for which a Y:NG ratio ~ 0.5 is expected in males and a Y:NG ratio ~ 0 is expected in females. To obtain these values, sequences aligned to the nuclear genome from each sample were analyzed from the BAM files. SAMtools view was used to retain properly aligned read pairs (-f 2) and to filter out those with supplementary alignments (-F 2048). Additionally, a MAPQ filter of 20 (-q 20) was applied. Since the MAPQ value reflects the difference between the best alignment score and the second-best alignment score, applying this filter helps to remove reads from the X chromosome that might be incorrectly aligned to Y chromosome sequences.

1.4.3. Mitogenome phylogenetic reconstruction

The mitochondrial genome of each sample was extracted from the BAM files corresponding to the reference mitochondrial genome using the BCFtools package v1.17 (Danecek et al., 2021) with the "mpileup" tool. Subsequently, variant calling was performed with the "bcftools call" command using the -c option. Consensus sequences in FASTQ format were generated using the vcf2fq option of the VCFtools package (Danecek et al., 2021)

The FASTQ sequences were converted to FASTA format and aligned with the MAFFT program (v7.464) using the options "maxiterate 1000" and "localpair". A maximum-likelihood phylogenetic tree was then constructed from the alignment with the RAXML program (v8.2.12) using a GTR model of nucleotide substitution and a gamma distribution of evolutionary rates (Stamatakis, 2014).

1.5. Metagenomics analysis from exogenous data: overall DNA composition, diet and parasites

Once the endogenous reads were separated, the pipeline development focused on the taxonomic classification of the exogenous read set, with the aim of analyzing, on the one hand, the overall composition of DNA present in the feces (Figure 13b) and, on the other, identifying elements that make up the desman's diet and potential parasites (Figure 13c).

1.5.1. Overall DNA composition in the feces

To assess the overall DNA composition in the feces, taxonomic classification of the exogenous DNA set from each sample was performed using Kraken 2 (Wood et al., 2019) with the RefSeq database (O'Leary et al., 2016) as a reference

(Figure 13b). For this, the pre-built "kraken2_RefSeqV208_nt" indexes created by Wright et al. (2023) were used. The RefSeq nucleotide database contains well-annotated, non-redundant DNA sequences from NCBI. The version used (Release 208, September 2021) includes 113,002 species represented across all domains of life: 1,360 of Archaea, 66,539 of Bacteria, 11,557 of Viruses, and 27,937 of Eukaryota. The reads were processed in a paired-end manner (--paired), meaning that mates were classified together, and the minimum number of hit groups required to make a classification call was set to 3 (--minimum-hit-groups) as recommended by the Kraken 2 authors, to improve accuracy (Lu et al., 2022). The number of reads assigned to each domain—Archaea, Bacteria, Viruses, and Eukaryota—was recorded, as well as to the lineages within Eukaryota and the lineages within Opisthokonta, according to NCBI taxonomy (Schoch et al., 2020b).

1.5.2. Databases for diet and parasites identification

To identify dietary elements and potential parasites from exogenous DNA present in feces, a new procedure was developed (Figure 13c, 14) that leverages the efficiency of Bowtie2 to align reads to reference sequences with precision and sensitivity, along with the BLAST suite's capability to identify similarities in large databases. This process is further enhanced by using various databases to improve classification accuracy.

First, a new database of mitochondrial sequences from eukaryotes was constructed from the NCBI Nucleotide database (Downloaded on 04-May-2022) (Cooper et al., 2010). Using the Entrez Direct software from E-utilities (Kans, 2013), we selected mitochondrial sequences between 500 and 30,000 base pairs (bp), excluding ribosomal RNA (rRNA) and transfer RNA (tRNA) sequences, and all sequences from organisms in the Viridiplantae domain. Additionally, pseudogenes and unverified sequences, which could lead to incorrect classifications, were also excluded. We called Mit11 to this curated mitochondrial database. To avoid redundancies and ensure that reads from the same source align to a single reference, duplicate sequences were removed using the SeqKit rmdup v2.3.0 program (Shen et al., 2016). This resulted in a database of 5,036,773 sequences, including specific genes, such as the COI gene, widely used in taxonomic identification (Antil et al., 2023; Porter & Hajibabaei, 2020), as well as other mitochondrial fragments and genes and complete mitochondrial genomes from a wide variety of taxa.

To take advantage of the nuclear genomic information present in the samples, we used additional databases. For this, we incorporated the SILVA 18S (SSU: Small Subunit) and SILVA 28S (LSU: Large Subunit) databases (Release 138.1 27-August-2020) (Quast et al., 2013), which contain curated nuclear rDNA sequences from the small subunit (18S) and large subunit (28S), respectively. These sequences are widely used for species identification because they

combine conserved regions with areas of high variability, allowing for discrimination among different taxa. From these databases, only sequences from the Eukarya domain were selected, providing 646,567 and 742,677 sequences for 18S and 28S, respectively.

Finally, to achieve a broader representation of taxa, we also included the BOLD database (Release 07-Apr-2023) (Ratnasingham & Hebert, 2007), which primarily contains barcodes from the mitochondrial COI gene for animals, as well as barcodes from other genes such as ITS/ITS2, matK, and rbcL for fungi and plants. This database is widely used for species identification and is not directly linked to NCBI, which allows it to provide additional sequences that are not present in the mitochondrial or SILVA databases. In total, BOLD contributes 9,338,938 eukaryotic sequences.

These databases were used in the initial phases of the pipeline for mapping reads with Bowtie2, while for taxonomic classification with BLAST, an additional database, the NCBI Core nucleotide BLAST database (BLAST Core_nt, version 1.1 – 21-October-2024; 109,874,994 sequences) (Sayers et al., 2025), was used. This database is an optimized version of the standard NCBI nucleotide database, designed to be non-redundant and to exclude preliminary or fragmentary sequences. It provides a more efficient representation by merging identical sequences of highly represented organisms into a single entry and excludes most eukaryotic chromosome sequences, retaining only transcript and gene-related sequences. By classifying the consensus sequences with this unique and more up-to-date database, it is possible to unify taxonomy, correct potential misannotations from the other databases, and ensure comparability of e-value scores for the detected alignments, which would not be accurate when using different databases.

1.5.3. Taxonomic classification of consensus sequences from diet and parasites

The assessment of the diet and parasites was performed using the set of exogenous sequences from each sample (Figure 13c). Using each database consecutively –Mit11, Silva 18S Eukaryotes, Silva 28S Eukaryotes, BOLD Eukaryotes–, the procedure was as follows (Figure 14a): First, the reads were mapped against our custom NCBI Mitochondrial database, Mit11, using the Bowtie2 in "unpaired" mode with the --end-to-end and --very-sensitive options. The "end-to-end" alignment mode requires that the read aligns entirely, using every character in the read, as explained before. The "very-sensitive" setting is designed to maximize alignment sensitivity by allowing more attempts and adjusting parameters to tolerate certain sequence variations. Specifically, this configuration uses a seed length of 20 nucleotides (-L 20) with no mismatches allowed in the seed (-N 0) and increases the number of consecutive attempts to extend a seed to 20 (-D 20) before giving up on finding an alignment at that location, which also allows

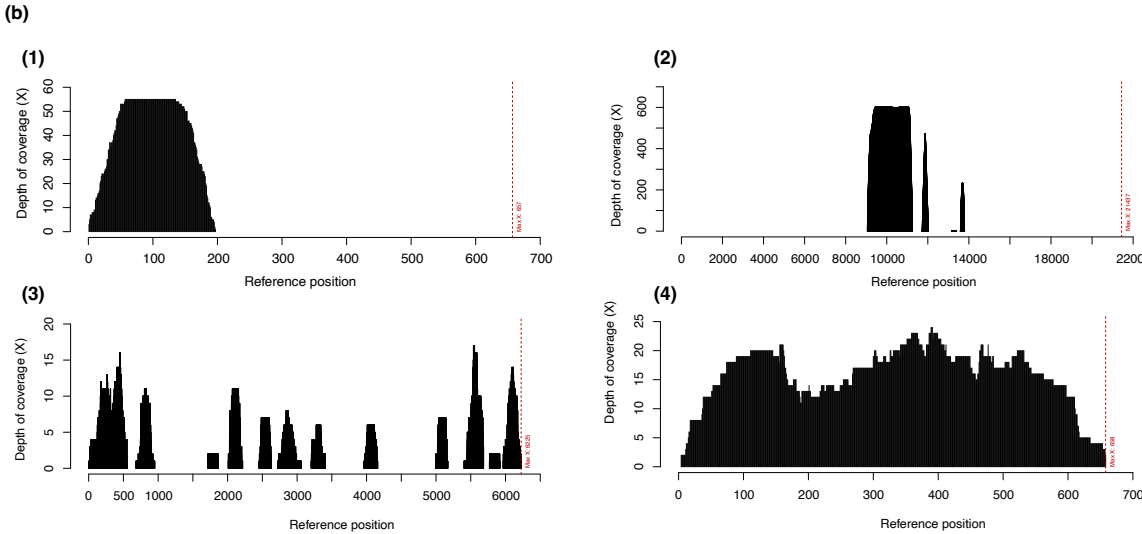
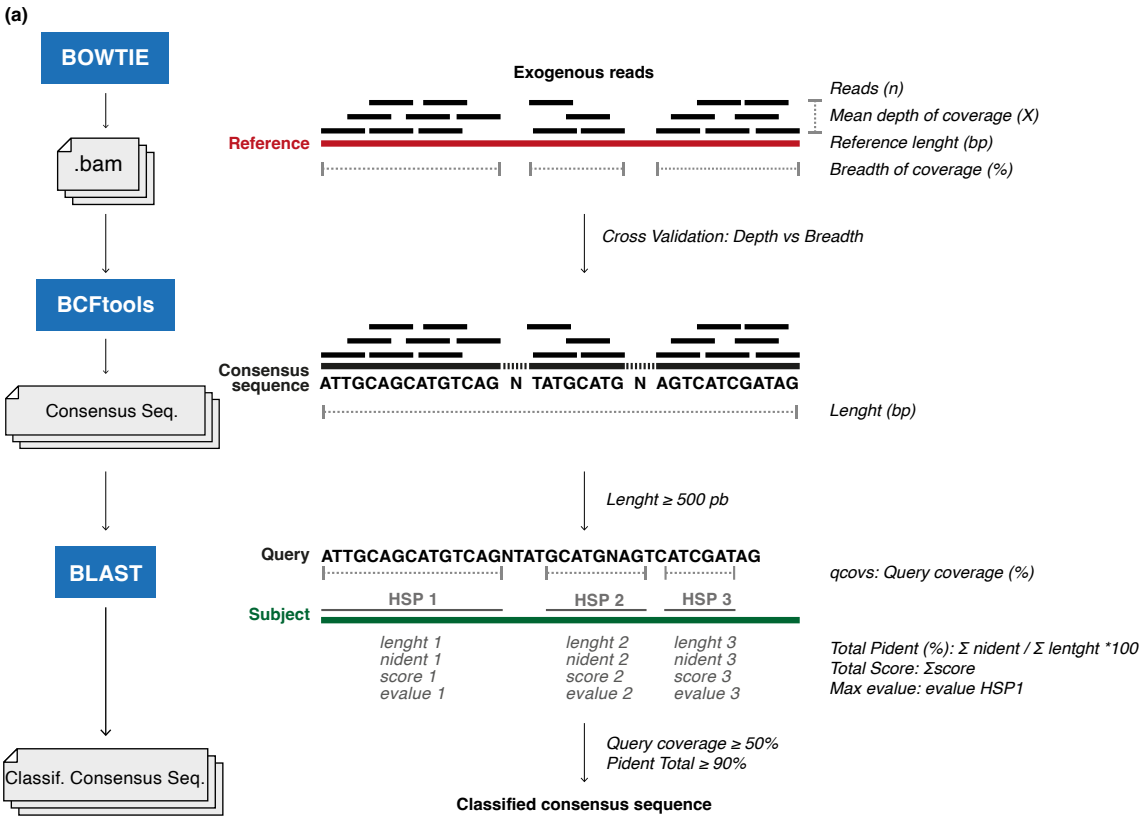
for finding better alignments at each position. This approach enables the detection of more alignments by accommodating small differences from the reference, which is particularly useful when the reads from the samples are divergent from the reference sequences, possibly due to sequences of the organisms being absent in the databases. Additionally, the unpaired alignments mode was chosen in this case because the databases used contain many partial or fragmented genetic sequences, which can cause only one of the mates in a pair to align correctly. The --un-gz option was used to record unpaired reads that did not align, which was used as the input for alignment with the next database consecutively, using the same parameters. This sequential procedure allows for progressive filtering of aligned reads, preventing reads from aligning more than once in cases where the databases are redundant. The -S and --no-unal options were used to record aligned reads in SAM files, which were subsequently converted to BAM format using SAMtools. During this process, reads were sorted by genomic position, PCR duplicates were marked, and the BAM files were indexed.

Instead of classifying the individual reads according to the matched database sequences, a step was included to generate consensus sequences from contiguously mapped reads in order to generate longer sequences for a subsequent database search with BLAST (Figures 13c and 14a). For this, the BCFtools package v1.17 (Danecek et al., 2021) was used to obtain consensus sequences from the generated BAM files, using the mpileup command with the --ns DUP option to exclude reads marked as duplicates. Subsequently, the call command with the -c option was used to perform variant calling based on genotype likelihoods, using a consensus model to assign genotypes for each position. These variant calls were subsequently used to generate consensus sequences, which were recorded in FASTQ format with the help of the vcfutils.pl vcf2fq program. The FASTQ files were then processed with Seqtk v1.3-r106 (H. Li, 2018) to transform them into FASTA format. During this process, low-quality or undefined bases were replaced with "N" and consecutive "N" sequences were reduced to a single "N." Additionally, the consensus sequences were filtered to retain only those with a length of at least 500 bp. For these consensus sequences, the number of reads that constituted each sequence was recorded. Depth of coverage (X) was calculated as the number of reads constituting the consensus sequence, multiplied by the average length of the reads in the sample, and divided by the length of the reference sequence where they were mapped. Breadth of coverage (%) was calculated as the percentage of the reference sequence covered by the generated consensus sequence.

The next step in the pipeline was the taxonomic classification of the retained consensus sequences. This was performed using the BLASTN v2.13.0+ program from the BLAST+ suite (Camacho et al., 2009) with the "megablast" task option. This option configures BLASTN to prioritize finding highly similar

IV. METHODS

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.5. METAGENOMICS ANALYSIS FROM EXOGENOUS DATA: OVERALL DNA COMPOSITION, DIET & PARASITES
1.5.3. TAXONOMIC CLASSIFICATION OF CONSENSUS SEQUENCES FROM DIET AND PARASITES



IV. METHODS

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.5. METAGENOMICS ANALYSIS FROM EXOGENOUS DATA: OVERALL DNA COMPOSITION, DIET & PARASITES
1.5.3. TAXONOMIC CLASSIFICATION OF CONSENSUS SEQUENCES FROM DIET AND PARASITES

sequences, making it ideal for comparisons within the same species or among closely related species. As mentioned previously, the database used in this step was the BLAST Core_nt.

For each consensus sequence evaluated (considered a query in BLAST), the top 5 hits (subjects from the database) were recorded based on their e-value (obtained with the option -max_target_seqs 5). Usually, the e-values were very similar and frequently all equal to 0. Therefore, for each of the 5 subjects, the Total Score, not provided in the BLAST output, was calculated as the sum of the scores for each High-scoring Segment Pair (HSP) that it contains (-outfmt score). Then, the subject with the highest Total Score was selected for taxonomic classification of the query. This parameter reflects both the similarity between the query and the subject and the alignment length, as a more extensive alignment with a good proportion of matches generally results with a higher Total Score. For the selected subject, the total percentage of identical matches (Total Pident) was calculated as the sum of the number of identical matches (-outfmt nident) in each HSP, divided by the sum of the alignment lengths of the HSPs (-outfmt length), multiplied by 100. Additionally, the query coverage (percentage of the query length involved in the alignment) was recorded (-outfmt qcovs), along with the e-value of the best HSP (-outfmt eval), and information about the selected subject, including Subject Seq-id, Subject Title, and Subject Taxonomy ID (-outfmt sseqid stitle staxid).

← Figure 14. Method developed for the taxonomic classification of diet and parasites

(a) The figure shows the alignment of exogenous reads to a reference sequence using Bowtie2, where the number of reads (Reads (n)), the depth of coverage (X), the breadth of coverage (%), and the reference length (bp) are recorded. These parameters are used for cross-validation of depth and breadth. Subsequently, the construction of the consensus sequence is shown, where regions of the reference sequence not covered by reads are replaced with an "N." Only consensus sequences longer than 500 bp proceed to the next step, which involves classification via BLAST, where each consensus sequence serves as a query. In this step, a similarity search is conducted against a reference sequence (Subject), and parameters for each alignment are recorded to assess the quality of classifications. Only classifications where more than 50% of the query's length participates in the alignment (Query coverage) and the overall percentage identity (Total Pident) exceeds 90% are retained. The figure also illustrates HSPs (High-Scoring Pairs), which are locally aligned segments of high similarity identified by BLAST. For each HSP, its length, the number of identical positions in the alignment between the query and the subject (nident), the alignment score (score), and the statistical significance (eval) are recorded. The overall percentage identity (Total Pident) is calculated as the sum

of identical positions across all HSPs divided by the sum of their lengths, multiplied by 100. The total alignment score (Total Score) is calculated as the sum of the scores of all HSPs, and the maximum e-value (Max eval) is equal to the e-value of HSP1, which is always the longest HSP. (b) This method allows for the evaluation of consensus sequence quality through the analysis of depth of coverage across the reference sequence. The figure presents four examples of consensus sequences from the same sample: (1) a low-quality alignment where all reads mapped to a region of less than 200 bp; this consensus sequence was filtered out for having a length below 500 bp; (2) another low-quality alignment that, although exceeding 500 bp in length, was discarded during depth:breadth cross-validation due to excessively high depth in relation to low breadth. The figure also includes two high-quality consensus sequences: (3) a sequence with homogeneous depth across the reference, though it does not achieve 100% breadth; and (4) a sequence with uniformly distributed depth of coverage along the reference, where the reference sequence is fully sequenced. These examples demonstrate the criteria used to validate and ensure the quality of the consensus sequences. (Max X (x-axis): reference sequence length).

Subsequently, based on each TaxID, the complete taxonomic lineage of each subject was recorded using the nodes.dmp and names.dmp files provided by NCBI. The nodes.dmp file enables the reconstruction of the hierarchical lineage of each TaxID, identifying the TaxIDs corresponding to each level. Meanwhile, names.dmp translates these TaxIDs into scientific names, thus generating a complete taxonomic description that was linked to the corresponding query.

Finally, the results obtained from the Bowtie2 alignments with each initial database – Mit11, Silva 18S Eukaryotes, Silva 28S Eukaryotes, and BOLD Eukaryotes – and classified with BLAST, were categorized into diet and parasites and combined for quality assessment. In the initial curation step, problematic sequences that could introduce errors in the analyses were removed. To achieve this, consensus sequences assigned to subjects labeled by NCBI as PREDICTED (annotated using a gene prediction method) or TPA (Third Party Annotation) were filtered out. Additionally, sequences assigned to subjects with any of the following taxonomic categories were excluded: unclassified, environmental, artificial, Bacteria, or without taxonomy. Subsequently, a filter was applied to retain only high-quality consensus sequences based on the parameters generated from Bowtie2 results (Figure 14b). Thus, consensus sequences with depth of coverage values of 50X or higher were discarded if they did not reach a breadth of coverage of at least 70%. This was done because a high depth of coverage should lead to maximum coverage if the reads were mapped to the correct species or a close one. However, if this coverage does not exceed a high value, like a 70% of the reference sequence, it suggests that the sample reads and the reference sequence come from evolutionarily distant organisms, where only highly conserved, non-informative regions at lower taxonomic levels, are being mapped. These results do not lead to a suitable taxonomic identification and therefore were removed.

Then, filters based on BLAST results were applied, retaining consensus sequences assigned to subjects with a Total Pident of more than 90% (percentage of identical matches across all HSPs for the subject) and with at least 50% query coverage, meaning that at least 50% of the consensus sequence length participated in the alignment. At this point, some problematic subjects were visually inspected, and all consensus sequences aligning to these were filtered out. For example, low-complexity reference sequences or those containing repetitive motifs were identified. This was evidenced when consensus sequences from different samples were assigned to the same reference sequence, and analysis of their depth and breadth values revealed significant variation in depth (reflecting the number of reads used to construct the consensus sequence) but consistently low breadth values. This indicated that these consensus sequences were constructed based on the alignment of reads to the same small regions of the reference sequence.

We also attempted to estimate the relative abundance of each element in the samples. Since organisms that are better represented in the databases could generate a higher number of mapped sequences, it is necessary to take the lengths of the reference sequences into account. Thus, for each element identified in the diet and parasites, the total depth of coverage (X) was calculated by summing the number of reads that contributed to the generation of the consensus sequences assigned to a given taxon, multiplying it by the average length of the exogenous reads in the sample, and dividing it by the sum of the lengths of the reference sequences where these reads aligned. Additionally, to estimate the overall diet or parasite composition across all samples, the weighted mean of the total depth of coverage (X) was calculated, using the number of exogenous reads in each sample as the weighting factor. Finally, the prevalence of each element was determined as the number of samples in which it was present.

As sequences from most species are not present in databases, most classifications were considered adequate at the class or order level. However, the Bowtie2 and BLAST results indicated the existence of some hits at the species level. To determine consensus sequences likely classified at the species level we only considered mitochondrial sequences, which generally have greater variability along their entire length, facilitating differentiation between species. In addition, we evaluated two aspects: the quality of the consensus sequence generated from the Bowtie2 alignments, as estimated by the breadth of coverage, and the quality of the alignment performed by BLAST, as estimated by the Total Pident and the query coverage. Specifically, the following thresholds were set for considering a species hit: a breadth of coverage of at least 50%, a minimum total identity of 94% with the subject, and an alignment involving at least 90% of the query length.

2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS

The following methods correspond to specific objective 9.

2.1. Reference bacterial genomes

The genus *Yersinia* was detected in initial analyses of desman feces samples and was chosen as a case study to evaluate different methods for species identification. We specifically selected for testing a group of closely related species, which can present significant identification problems. The genus contains 26 species listed in the NCBI taxonomy, four of which are pathogenic (Table S2), so accurate identification to species level is essential in this genus, highlighting the need for robust methods for this task. Reference genomes for the bacterial species were obtained from the NCBI Genome Database (<https://www.ncbi.nlm.nih.gov/datasets/>).

For a broader analysis of other pathogenic bacteria, the reference genomes of species found in the Pathogen Detection Project of NCBI (<https://www.ncbi.nlm.nih.gov/pathogens/>) and the Virulence Factor Database (VFDB, <http://www.mgc.ac.cn/VFs/>) (Liu et al., 2022) were downloaded, in addition to other bacterial species found in the literature (Ali & Alsayeqh, 2022; Barandika et al., 2007; Cantas & Suer, 2014; Pedersen et al., 2007; Sabour et al., 2022; Suminda et al., 2022; White & Razgour, 2020) and online sources (e.g. <https://ewda.org/diagnosis-cards/>, animal health section of [https://www.mapa.gob.es, etc.\)](https://www.mapa.gob.es, etc.)), resulting in 137 species that included the four pathogenic *Yersinia* species (Table S3). *Cutibacterium acnes* was included in the initial analyses and was found in most samples, but it has been reported that this species is a likely contaminant of kits and reagents (Gu et al., 2019) and was excluded from the final analyses, which were based on 136 species.

2.2. Alignment of sequences to reference genomes of *Yersinia* species

The set of 26 *Yersinia* species was used to test the performance of different methods and parameters in the developed pipeline. First, the set of exogenous reads was aligned separately to each of the respective reference genomes using Bowtie2 (Langmead & Salzberg, 2012) in the "end-to-end" alignment mode and the --sensitive option. In addition, the options --no-discordant and --no-mixed were used to capture only read pairs where both reads align and have the expected orientation and distance between them. Subsequently, alignments were converted to BAM format, PCR duplicates removed, and reads with MAPQ values lower than 20 filtered using SAMtools v1.9 (H. Li et al., 2009). General mapping quality was visualized using the coverage across the genome graph constructed with Qualimap 2.2.2. (Okonechnikov et al., 2016) which represents the mean depth of coverage in 4,000 windows of the genome.

Using SAMtools, genome alignment statistics were calculated for each sample and bacterial species, including the number of mapped reads, the mean MAPQ of the mapped reads, the number of positions in the reference genome sequenced with at least one read, the breadth of coverage (%), and the depth of coverage (X) . Different thresholds based on percent coverage (breadth of coverage) were set to determine the presence of each species in the samples. We also calculated the percent coverage when filtering out positions with extreme values of depth of coverage, defined as those sites with a depth of coverage greater than the sample mean plus 3 times the standard deviation, calculated in logarithmic space.

Bowtie2 can be used with single-species or multi-species databases. A Bowtie2 database was first constructed for each species for the main analyses. In addition, a combined database containing the set of all species was constructed to assess the effect of database configuration on analysis performance. For samples with multiple positives of species in the same genus or closely related species, the numbers of unique and shared reads for each positive species were compared using Venn diagrams.

The set of exogenous sequences was also aligned against the *Yersinia* reference genomes using other methods and conditions. Thus, we used Bowtie2 in "local" alignment mode, as well as another widely used DNA mapping program, BWA 0.7.17 (H. Li & Durbin, 2009). For the latter, both the BWA-mem and BWA-aln algorithms were evaluated in their default configurations. The MAPQ scales are different in Bowtie2 and BWA, with a maximum of 42 for Bowtie2 and 60 for BWA. As mentioned above, a MAPQ cut-off of 20 was used for Bowtie2. The distributions of MAPQ values for both methods in the same set of samples showed that the equivalent threshold for BWA was ~30, so this value was used to filter mapped reads with this program.

2.3. Alignments to reference genomes of pathogenic bacterial species

The Bowtie2 "end-to-end" mode, which provided consistent results for the *Yersinia* species, was used to align the exogenous reads from each sample to the 136 pathogenic bacterial species using both single-species and combined databases. Identification was based on the percent coverage thresholds as described above. Principal Component Analyses (PCA) was performed on the percent coverage values to compare the overall health status of the Iberian desmans using the single-species databases (similar results were obtained using the combined database).

V. RESULTS

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES

The following results correspond to specific objectives 1 to 8.

1.1 Performance of metagenomic libraries and illumina sequencing from fecal samples

The total DNA concentrations extracted from the fecal samples were generally quite low; only three samples exceeded 1 ng/μL (Table S4). The minimum concentration was 0.14 ng/μL, with a mean of 0.6 ng/μL and a coefficient of variation (CV) of 128%. Despite these low concentrations, the quantities were sufficient to construct the libraries. The sequencing yielded a total of 1,867,181,444 Illumina reads in the metagenomic library pool, with an average of 81,181,802 reads per sample. Additionally, a notable decrease in variability among samples was observed, reflected in a CV of 16.65% in the number of reads per sample, made possible by the equimolar representation of the samples in the library pool. After applying quality filters, an average of 81% of the reads in each sample was retained, resulting in a mean of 66,059,624 reads per sample and a CV of 17.16%. Sample BC3015 had the lowest number of reads with 42,603,532, while sample BC2408 had the highest number, reaching 84,184,200 reads.

1.2 Total DNA composition in fecal samples

The number of reads successfully aligned to the *G. pyrenaicus* nuclear genome varied significantly across samples (Figure 15, Table S5). The lowest number of mapped

reads was 47,582 in sample BC2042, accounting for only 0.07% of the total filtered reads for that sample, while the highest mapped reads count was 38,876,690 in sample BC2865, representing 54.67% of its total filtered reads (Table 1). This sample achieved the highest depth of coverage, at 3.09X, and the highest breadth of coverage, at 79.04% (Table S5). The weighted average of aligned reads was 6,079,810.78, corresponding to 9.32% of the initial reads. The weighted average values for depth of coverage and breadth of coverage were 0.48X and 19.34%, respectively.

Regarding sequences mapped to the G. pyrenaicus mitogenome, their contribution to the overall fecal DNA content was very low, with a weighted average percentage relative to the total reads per sample of only 0.03%. The lowest count of mitochondrial reads was in sample BC2042, with just 2,308 reads, while the highest was in sample BC3345, with 125,918 reads (Table 1). Despite these low values, complete mitogenome reconstruction was achieved in all samples, with weighted averages of breadth of coverage at 99.4% (95.37–100) and depth of coverage at 154.25X (19.35–991.4) (Table S5). Notably, the quantities of endogenous sequences from nuclear and mitochondrial genomes showed no correlation (Figure S1).

Metagenomic analysis of exogenous sequences (Figure 15, Table 1, Table S6) revealed

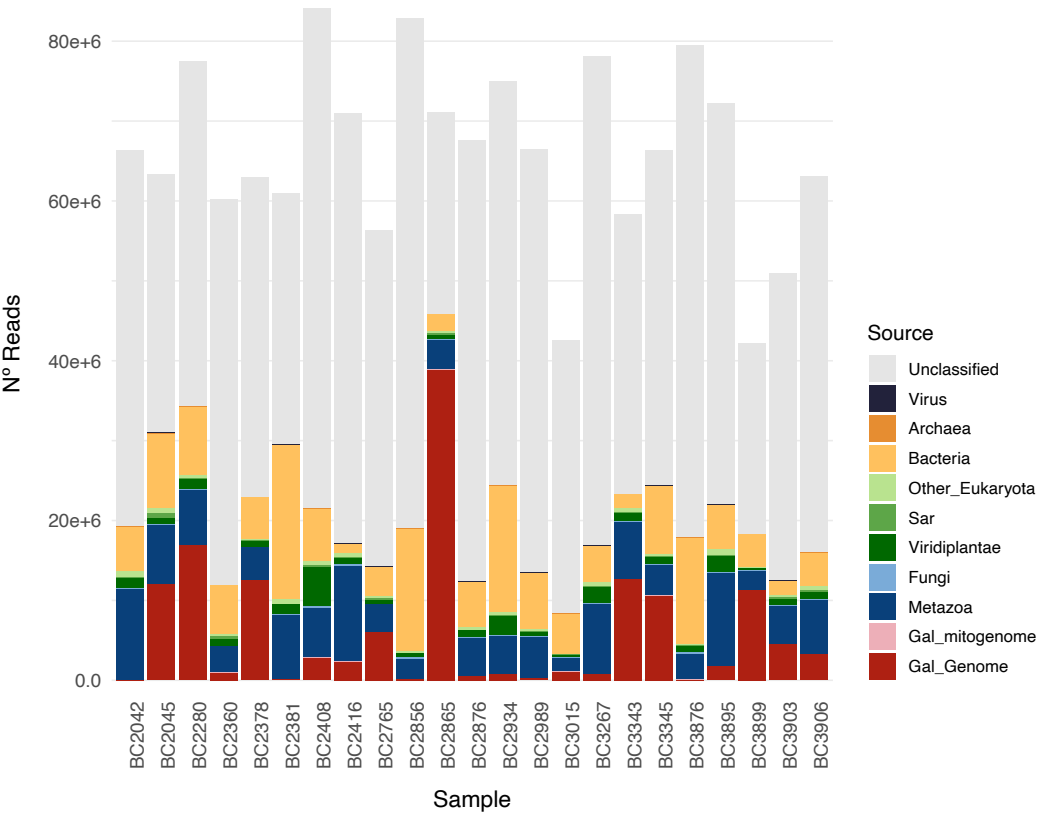


Figure 15. Global fecal DNA composition. Number of reads assigned to each source category in each sample.

a high proportion of reads of metazoan origin. The total number of reads assigned to this taxonomic group was 133,019,468, with a weighted mean of 5,926,054.93 per sample, accounting for 8.75% of the total fecal DNA composition, with a maximum proportion of 17.15% in sample BC2042 and a minimum of 2.96% in sample BC2856. Other eukaryotic groups showed much lower representation: Viridiplantae had a mean proportion of 1.71%, SAR (Stramenopiles, Alveolata, and Rhizaria) of 0.24% and Fungi of 0.20%. Reads from other eukaryotic and opisthokont groups fell between 0 and 0.01%, so these were grouped as "Other Eukaryota".

Among other domains, Bacteria represented the largest proportion, with 10.53% (159,943,278 total reads), with a minimum of 1.61% in sample BC2416 and a maximum of 31.51% in sample BC2381. In contrast, the contribution of Archaea was minimal, comprising only 0.04% of the total. Additionally, reads classified as viruses accounted for 0.09% of the total DNA composition. Lastly, a significant portion of reads remained unclassified, representing 68.51% of all reads, with a maximum of 81.57% in sample BC2876 and a minimum of 35.46% in sample BC2865 (Figure 15, Table 1, Table S6).

Table 1. Contribution of different sources to the overall DNA composition present in fecal samples. For each source, the table shows the total number of reads assigned across all samples, the weighted average number of reads based on the total filtered reads in each sample, the minimum and maximum read counts, the weighted average percentage, and the minimum and maximum percentages for each source.

Source	Total filtered reads (n)	W. Average Reads (n)	Min reads (n)	Max reads (n)	W. Average Reads (%)	Min reads (%)	Max reads %
Nuclear Genome	141,634,060	6,079,810.78	47,582	38,876,690	9.32	0.072	54.67
Mitogenome	489,528	21,319.96	2,308	125,918	0.03	0.003	0.19
Metazoa	133,019,468	5,926,054.93	1,690,652	11,997,110	8.75	2.962	17.15
Fungi	2,967,930	135,497.44	58,734	237,142	0.20	0.095	0.29
Viridiplantae	25,980,708	1,216,066.11	249,776	4,904,016	1.71	0.578	5.83
Sar	3,574,334	155,459.01	29,978	615,448	0.24	0.051	0.97
Other Eukaryota	8,897,684	397,125.34	115,858	767,206	0.59	0.230	1.16
Bacteria	159,943,278	7,241,293.43	1,143,180	19,229,912	10.53	1.610	31.51
Archaea	586,306	26,650.20	9,988	56,294	0.04	0.014	0.07
Virus	1,302,804	57,691.71	17,322	140,172	0.09	0.024	0.23
Unclassified	1,040,975,242	46,662,243.02	23,881,254	63,810,482	68.51	35.456	81.57

1.3. Sex determination

For sex determination we first applied quality filters to the reads mapped to the G. pyrenaicus nuclear genome: pairs of concordantly aligned reads, without supplementary alignments, and with MAPQ >20. A high proportion of reads was retained (Table S7), with a weighted average of 99.14%. This indicates that, overall, most of the reads aligned to the G. pyrenaicus nuclear genome were of high quality.

Based on these filtered data, depth of coverage ratios (Y:NG) were 0 in 8 samples, due to the absence of reads aligned to the Y chromosome, classifying them as female. In the remaining 15 samples, Y:NG values ranged between 0.54 and 0.80, classifying them as male (Figure 16). The maximum Y:NG ratio (0.80) was observed in sample BC3876, based on the alignment of only 12 reads to the Y chromosome. This sample also had a low number of reads aligned to the nuclear genome (135,168 high-quality reads), but it resulted in a sufficient depth ratio for classification as male, demonstrating the method's sensitivity even in samples with a low read count. In contrast, sample BC2865, classified as female, showed the highest number of reads among all samples, with 38,632,796 high-quality reads aligned to the nuclear genome, without any aligning to the Y chromosome. This result highlights the accuracy of the sex determination method based on Bowtie2 alignments and the applied quality filters.

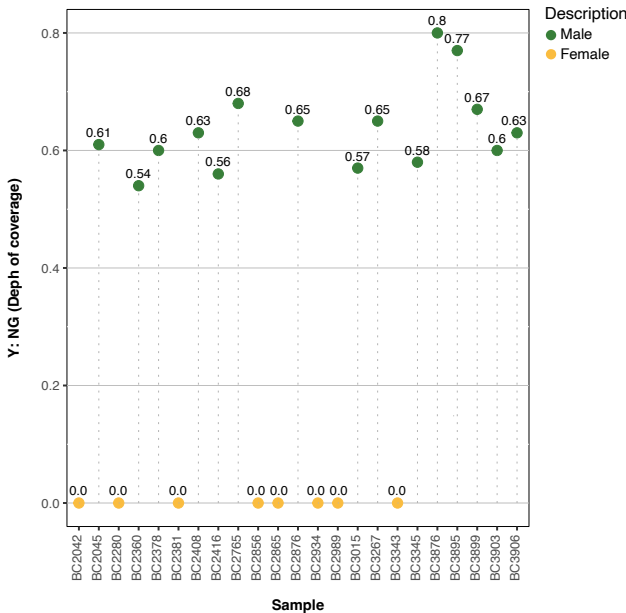


Figure 16. Sex Determination from coverage. Ratio between the depth of coverage for alignments to the Y chromosome (Y) and the depth of coverage for alignments to the nuclear genome of G. pyrenaicus (NG).

1.4. Population structure analysis

The PCA based on genotype probabilities showed a clear separation between the subpopulations (Figure 17). PC1 separated the subpopulations according to the rivers where they are distributed: Aravalle-Endrinal, Becedillas and Adaja, with Adaja being the most geographically distant. Meanwhile, PC2 distinguished the subpopulations from the sub-basins of the Aravalle river: Aravalle and Endrinal. In general, individuals grouped according to their subpopulation of origin and consistently with their geographic location, indicating the existence of low connectivity and structure among the subpopulations. However, the BC2865 sample from Endrinal showed greater affinity with the Aravalle group than with the Endrinal group, which could be due to the dispersal of this individual from Aravalle to Endrinal, indicating that these populations still maintain some level of connectivity.

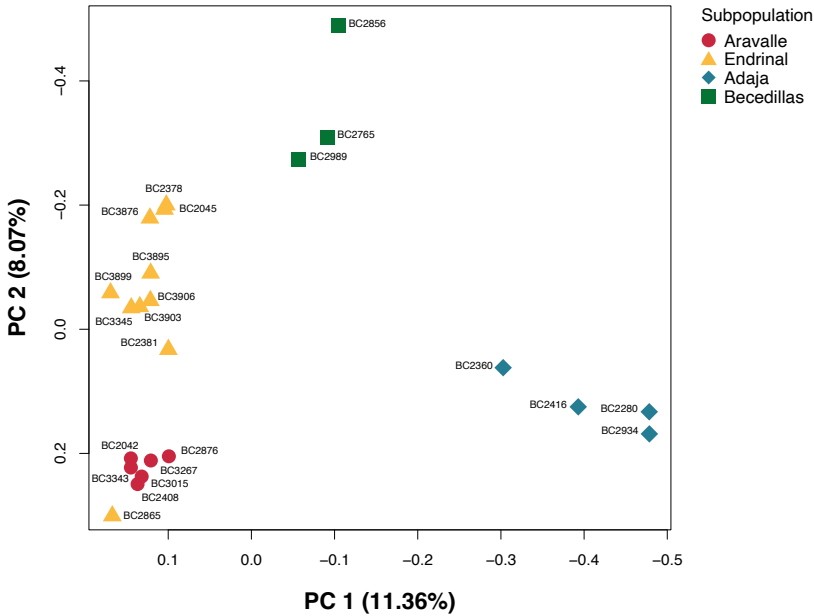


Figure 17. Principal Component Analysis (PCA) illustrating the genomic variation among the analyzed samples based on genotype probabilities. The plot displays the first two principal components (PC1 and PC2), along with the percentage of total genomic variation explained by each component (%).

1.5. Phylogenetic Tree of the mitochondrial genomes

The reconstructed consensus sequences of the mitochondrial genomes corresponded to five different haplotypes identified through the phylogenetic analysis (Figure 18). The tree did not show a genealogy consistent with the origin of the individuals, with the Adaja population being the only one to present a unique haplotype.

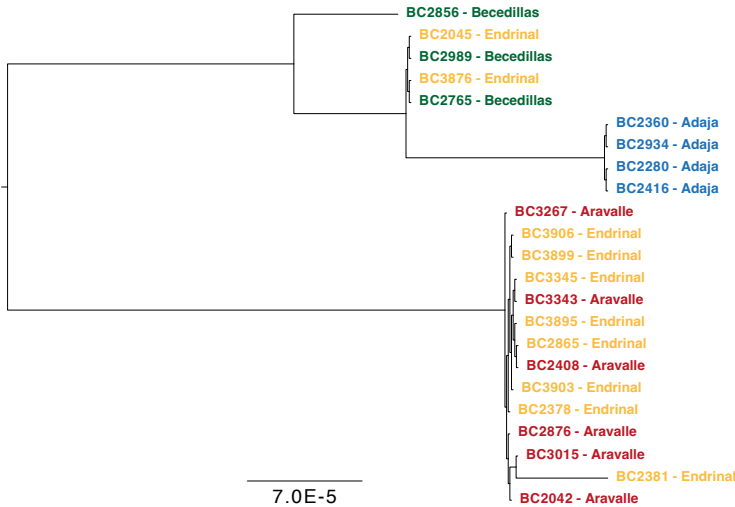


Figure 18. Mitochondrial Phylogenetic Tree. The subpopulation of each sample is indicated. The scale is in substitutions per site.

1.6. Taxonomic classification of diet and parasites

The newly developed pipeline (Figure 13c, 14) enabled the assembly of 13,016 consensus sequences through the sequential mapping of the exogenous reads to four eukaryotic databases (Table 2). Specifically, 2,797 consensus sequences were obtained from the Mit11 database, 5,924 sequences from Silva 18S, 3,183 sequences from Silva 28S, and 1,112 sequences from BOLD.

Table 2. Databases and processing of generated consensus sequences. The top section provides details on each database used to generate consensus sequences: Mit11 (NCBI Mitochondrial Eukaryotes, excluding plants), Silva 18S Eukaryota, Silva 28S Eukaryota, and BOLD Eukaryota. For each database, the table shows the total number of reference sequences it contains, the number of consensus sequences generated from aligning exogenous reads, the eukaryotic consensus sequences obtained, those retained after Bowtie2 and BLAST filtering (including visual curation), the number classified as diet or parasite components (% classified relative to the total), and the total reads forming the classified consensus sequences. The bottom section summarizes the totals for each category (combined across databases) and the percentage of retained consensus sequences at each processing step.

Database	Reference seq.	Consensus seq.	Eukaryota seq.	Retained (Bowtie2)	Retained (BLAST)	Classified seq.	Classified reads
Mit11	5,036,773	2,797	2,322	2,302	2,240	1,645 (58.8%)	345,416
Silva 18S Eukaryota	646,567	5,924	3,778	3,746	3,653	2,683 (45.3%)	128,858
Silva 28U Eukaryota	742,677	3,183	1,876	1,850	1,795	1,024 (32.2%)	330,820
Bold Eukaryota	9,338,938	1,112	1,060	1,060	1,056	18 (1.6%)	6,139
Total	15,764,955	13,016	9,036	8,958	8,744	5,370	811,233
% Consensus seq.	-	100	69.4	68.8	67.2	41.3	-

In a first quality filtering step, we automatically eliminated a set of 3,980 consensus sequences corresponding to sequences with the following characteristics: taxonomy labeled as "unclassified" (57.94%) or "environmental" (7.39%), sequences assigned by BLAST to subjects labeled as "PREDICTED" (29.95%), and sequences classified as Bacteria (3.34%), which were not object of this part of the study. The retained set of 9,036 sequences constituted the group of classified eukaryotic sequences.

The second filtering step involved two main assessments: evaluating the quality of the consensus sequences generated from Bowtie2 alignments and assessing the quality of the alignments performed by BLAST. This resulted in removing 78 consensus sequences with more than 50X depth of coverage but that did not reach more than 70% breadth of coverage with the reads mapped by Bowtie2, considered to correspond to very distant reference sequences relative to those present in the samples. Additionally, 213 sequences were filtered out due to a Total Pident lower than 90% –meaning that less than 90% of the positions in the query (consensus sequence) were identical to those in the subject (reference sequence) assigned by BLAST– or due to a query coverage –query length involved in the alignment– of less than 50%.

At this point, 47 consensus sequences were also removed as they aligned to 9 subjects visually identified as problematic, as it was impossible to precisely

determine their origin. For instance, consensus sequences aligned to subject EU815285.1, annotated as belonging to the Megaloptera *Sialis californica*, also aligned to sequences from the nematode *Mermithidae* sp., a parasite of insects (Dorris et al., 1999). Similarly, consensus sequences aligned to subject MW384893.1, annotated as the Hemiptera *Haplaxius skarphion*, also matched sequences from α -proteobacteria, such as *Candidatus Tisiphia* and *Rickettsia*, both of which are insect endosymbionts (Davison et al., 2022). This observation, together with previous automatic filtering steps, highlights the issues that can be present in sequence databases and underscores the necessity of visually inspecting and curating these databases to avoid artefactual results.

The final set of consensus sequences after quality filtering included 8,744 sequences, representing 67.2% of the assembled consensus sequences, with the following characteristics: an average of 132.6 reads per sequence, a length range of 500-16,329 bp (mean 992.6 bp), a Total Pident > 90% (mean 98.98%), a query coverage > 50% (mean 98%), and e-values ranging from 0 to 4.75e-57 (Table S8, Figure S2). From this group, we identified groups of metazoans and SAR that could belong to diet and parasite components, resulting in 5,370 consensus sequences being considered. The remaining consensus sequences that passed the quality filters but were not classified as part of the diet or parasites (3,374) primarily corresponded to Viridiplantae, Fungi, Rotifera and Rhodophyta and were outside the scope of this study.

For the description of diet and parasites, taxonomic units at various levels were used to ensure representative and identifiable names according to the literature of diet and parasites. These taxonomic units were then grouped into broader categories to facilitate analysis (also at different taxonomic ranks). Taxonomic units were further differentiated by certain ecological aspects, such as habitat type (aquatic, aquatic-terrestrial, terrestrial) for diet, and parasitism type (endoparasite, ectoparasite) for the detected parasites (Tables 3 and 4).

1.6.1. Classification of diet components

The proportion of different diet elements detected in each sample, based on total depth of coverage values, is shown in Figure 19a (Table S9), along with the overall composition in terms of abundance (calculated using weighted average depth of coverage values across all samples) (Figure 19b) and prevalence (number of samples in which each element was detected) (Figure 19c). Some global descriptive statistics are presented in Table 3.

Taking the defined taxonomic units, the diet consisted of 11 insect orders, with Trichoptera and Ephemeroptera showing the highest abundance at 33.5% and 19.13%, respectively. These two orders also exhibited the highest prevalence, being present in 100% (23 samples) and 95.7% (22 samples) of cases, respectively. Other relatively abundant orders included Lepidoptera (4.54%; 6 samples), Diptera (4.31%; 18

V. RESULTS
1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.6. TAXONOMIC CLASSIFICATION OF DIET AND PARASITES
1.6.1. CLASSIFICATION OF DIET COMPONENTS

samples), and Plecoptera (4.13%; 13 samples). Although Lepidoptera had a slightly higher abundance, its prevalence was lower, indicating that it was concentrated in a few samples. The remaining insect orders were detected at lower abundance levels: Hemiptera (1.17%), Coleoptera (0.9%), Hymenoptera (0.89%), Odonata (0.51%), Orthoptera (0.23%), and Megaloptera (0.18%). Generally, these orders exhibited low prevalence, with the exception of Coleoptera, which was present in 14 out of the 23 samples (60.9%), indicating that it was a constant element although always at low proportion.

Table 3. Overall composition of the diet. This table shows the major groups of detected dietary elements, the individual elements, their common names, their ecology based on habitat type (where Aquatic-Ter is Aquatic-Terrestrial), the total number of reads comprising the consensus sequences assigned to these taxonomic units, their weighted average depth of coverage (average depth) values (X) and its global percentage (%), as well as their prevalence in terms of number of samples (n) and prevalence percentage (%).

Group (taxonomic rank)	Diet Element (taxonomic rank)	Common name	Habitat type	Total Reads (n)	Average Depth		Prevalence	
					(X)	(%)	(n)	(%)
Insecta (class)	Trichoptera (order)	Caddisflies	Aquatic	296,312	6.59	33.5	23	100.0
Insecta (class)	Ephemeroptera (order)	Mayflies	Aquatic	111,578	3.77	19.1	22	95.7
Insecta (class)	Plecoptera (order)	Stoneflies	Aquatic	12,099	0.81	4.1	13	56.5
Insecta (class)	Odonata (order)	Dragonflies & damselflies	Aquatic	1,079	0.10	0.5	4	17.4
Insecta (class)	Megaloptera (order)	Alderflies	Aquatic	348	0.04	0.2	2	8.7
Insecta (class)	Diptera (order)	Flies	Aquatic-Ter.	9,532	0.85	4.3	18	78.3
Insecta (class)	Hemiptera (order)	Bugs	Aquatic-Ter.	2,359	0.23	1.2	7	30.4
Insecta (class)	Coleoptera (order)	Beetles	Aquatic-Ter.	2,574	0.18	0.9	14	60.9
Insecta (class)	Lepidoptera (order)	Butterflies & moths	Terrestrial	618	0.89	4.5	6	26.1
Insecta (class)	Hymenoptera (order)	Wasps, ants & bees	Terrestrial	388	0.17	0.9	5	21.7
Insecta (class)	Orthoptera (order)	Grasshoppers & crickets	Terrestrial	1,260	0.04	0.2	5	21.7
Platyhelminthes (phylum)	Tricladida (order)	Planarians	Aquatic	13,161	5.01	25.4	18	78.3
Nematomorpha (phylum)	Gordioda (class)	Horsehair worms	Aquatic	493	0.29	1.5	7	30.4
Mollusca (phylum)	Bivalvia (class)	Bivalves	Aquatic	11	0.02	0.1	1	4.3
Mollusca (phylum)	Gastropoda (class)	Slugs & snails	Aquatic-Ter.	17	0.05	0.3	2	8.7
Annelida (phylum)	Oligochaeta (subclass)	Earthworm	Aquatic-Ter.	17	0.01	0.0	1	4.3
Crustacea (subphylum)	Isopoda (order)	Woodlice	Terrestrial	47	0.12	0.6	1	4.3
Arachnida (class)	Araneae (order)	Spiders	Terrestrial	208	0.13	0.6	3	13.0
Arachnida (class)	Opiliones (order)	Harvestmen	Terrestrial	92	0.08	0.4	1	4.3
Actinopterygii (superclass)	Salmoniformes (order)	Salmons & trouts	Aquatic	279	0.24	1.2	6	26.1
Amphibia (class)	Caudata (order)	Salamanders	Aquatic	1,760	0.06	0.3	1	4.3

V. RESULTS
1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.6. TAXONOMIC CLASSIFICATION OF DIET AND PARASITES
1.6.1. CLASSIFICATION OF DIET COMPONENTS

Other invertebrate groups detected included the planarians (Tricladida), with a notable high abundance of 25.43%, placing it second overall, and a high prevalence, appearing in 18 samples (78.3%). The other groups were detected at considerably lower abundance levels, with the Nematomorpha Gordioda standing out, present in 7 samples with an abundance of 1.48%. The remaining groups had lower abundances and were found in 1 to 3

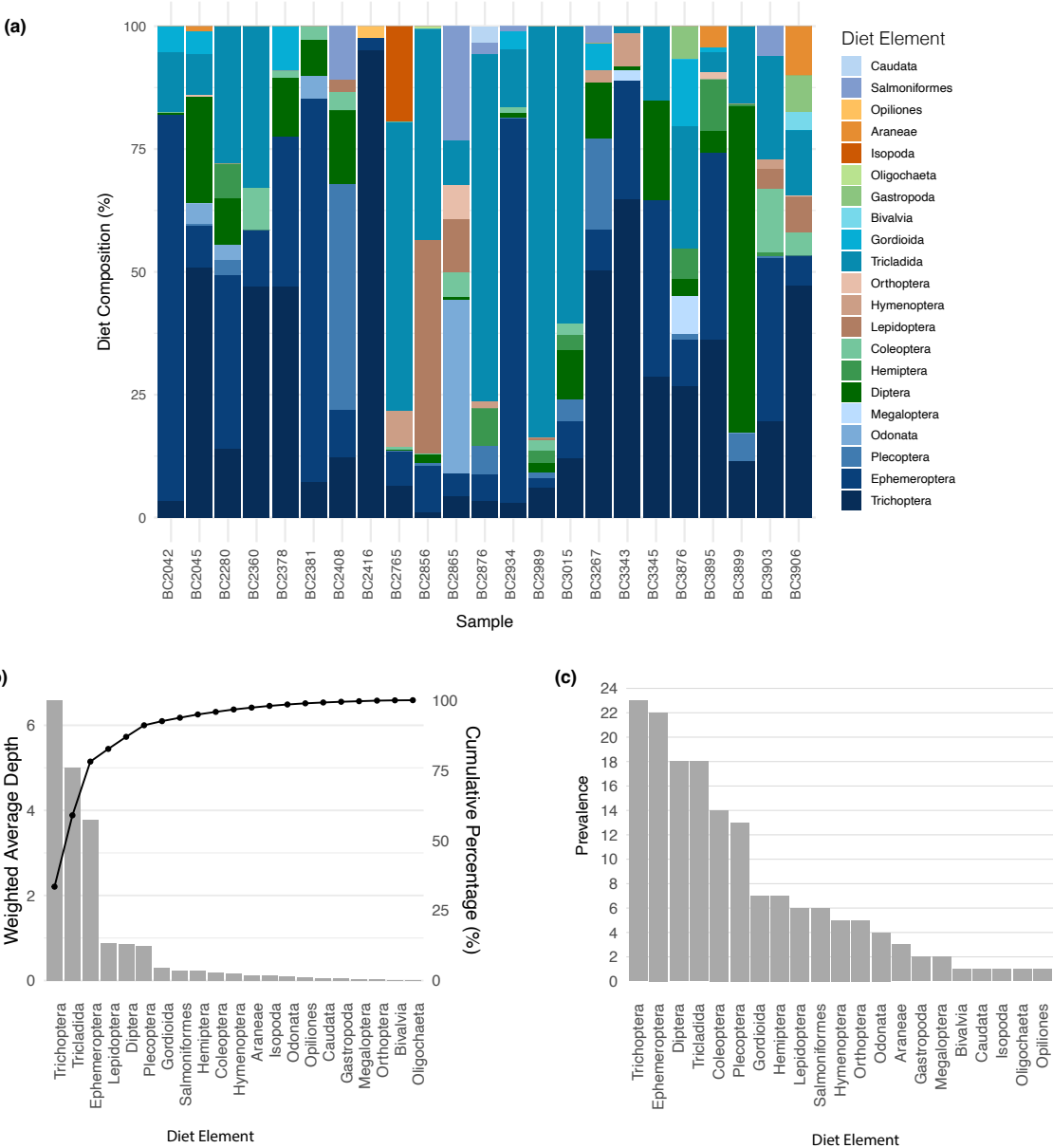


Figure 19. Diet Composition. (a) Proportion of each diet element in each sample based on its total depth of coverage value. (b) Comparison of the abundance of each diet element, represented by the weighted mean depth of coverage for each element across all samples (left axis), and its cumulative percentage (right axis). (c) Prevalence of each diet element, represented in decreasing order.

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.6. TAXONOMIC CLASSIFICATION OF DIET AND PARASITES

samples each. These included the arachnids Araneae (0.64%) and Opiliones (0.42%), the crustacean Isopoda (0.61%), the mollusks Gastropoda (0.28%) and Bivalvia (0.10%), and the annelid Oligochaeta (0.05%). Additionally, two vertebrate groups were detected. Salmoniformes was present in 6 samples and Caudata was found in one sample. Regarding the habitat type of detected prey, 86% were aquatic, 7.3% were terrestrial, and 6.7% were groups with both aquatic and terrestrial representatives.

1.6.2. Identification of parasites

Regarding parasites, the taxonomic classification of the consensus sequences generated revealed the presence of several groups of endoparasites and ectoparasites (Figure 20, Table S10). Among the endoparasites, apicomplexans from the subclass Coccidia were the most abundant. This parasite group was detected in all samples (23/23, 100%) (Table 4). To a lesser extent, metazoan endoparasites were identified. The flatworm of the class Cestoda was present only in one sample. Similarly, nematodes from the orders Mermithida and Rhabditida were detected, with lower global abundances of 1.58% and 0.74%, respectively, although they showed notable proportions at the individual sample level. The nematode Mermithida was detected in two samples (prevalence of 8.7%) and the Nematode Rhabditida was present only in one sample. Regarding ectoparasites, these were represented by mite orders (Acari). Trombidiformes was in four samples. Mesostigmata was detected in only one sample and Sarcoptiformes in three.

Table 4. Taxonomic units identified as parasites. This table presents the major groups of identified parasites, the corresponding taxonomic units they contain, their common names, their ecology based on the type of parasitism, the total number of reads comprising the consensus sequences assigned to these taxonomic units, their weighted average depth of coverage (Average Depth) values (X) and its global percentage (%), as well as their prevalence in terms of the number of samples (n) and prevalence percentage (%).

Group (taxonomic rank)	Parasite (taxonomic rank)	Common name	Parasitism type	Total Reads (n)	Average Depth		Prevalence	
					(X)	(%)	(n)	(%)
Apicomplexa (phylum)	Coccidia (subclass)	Coccidia	Endoparasites	349017	4.51	74.49	23	100.0
Platyhelminthes (phylum)	Cestoda (class)	Tapeworms	Endoparasites	1908	0.48	7.95	1	4.3
Nematoda (phylum)	Mermithida (order)	Roundworms	Endoparasites	103	0.10	1.58	2	8.7
Nematoda (phylum)	Rhabditida (order)	Roundworms	Endoparasites	19	0.04	0.74	1	4.3
Acari (subclass)	Trombidiformes (order)	Mites	Ectoparasites	643	0.57	9.48	4	17.4
Acari (subclass)	Mesostigmata (order)	Mites	Ectoparasites	642	0.19	3.12	1	4.3
Acari (subclass)	Sarcoptiformes (order)	Mites	Ectoparasites	225	0.16	2.64	3	13.0

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES
1.6. TAXONOMIC CLASSIFICATION OF DIET AND PARASITES

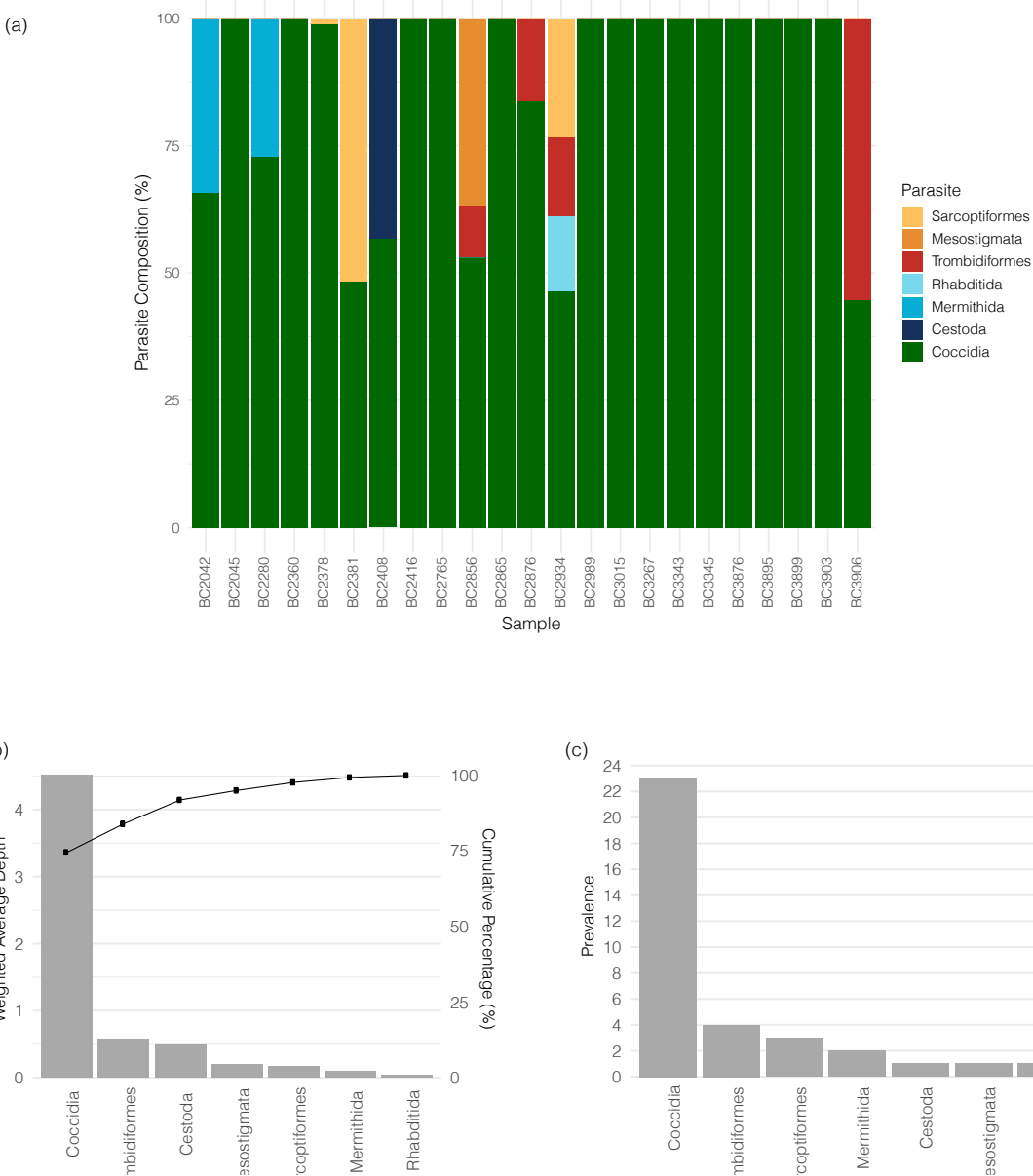


Figure 20. Parasites detected. (a) Proportion of each parasite in each sample based on its total depth of coverage value. (b) Comparison of the abundance of each parasite, represented by the weighted mean depth of coverage for each element across all samples (left axis), and its cumulative percentage (right axis). (c) Prevalence of each parasite, represented in decreasing order.

1.6.3. Identification of consensus sequences at the species level

To attempt to classify consensus sequences at lower taxonomic levels, eventually, at the species level, we applied stricter filter thresholds for breadth of coverage (50%), Total Pident (94%) and query coverage (90%). We only applied them to Metazoans and only to mitochondrial sequences, which are more suitable for species-level identification (Antil et al., 2023; Elyasigorji et al., 2023). These thresholds were exceeded by a total of 181 consensus sequences, which were assigned to 37 species. Of these, 34 corresponded to species of the phylum Insecta, distributed among the following orders: 11 species of Trichoptera, 6 of Coleoptera, 4 of Diptera, 4 of Ephemeroptera, 3 of Hemiptera, 2 of Hymenoptera, 1 of Megaloptera, 1 of Odonata, 1 of Orthoptera and 1 of Plecoptera. Additionally, 1 species of Isopoda, 1 of Tricladida, and 1 of Caudata were also likely identified at the species level. (Table 5). To confirm if these classifications were correct at the species level, we first considered if the resulting species were present in the Iberian Peninsula and, second, we analyzed the depth of coverage plots.

The presence of the species in the Iberian Peninsula was verified using the Iberian fauna database Iberfauna (http://iberfauna.mncn.csic.es/). When the species was not found in this database, its presence at the genus level was investigated. In only three cases, dietary elements were not described in the Iberian Peninsula at the species level but were identified at the genus level. In these cases, it was considered that the reads originated from a sister species or a closely related species.

A representative example of depth of coverage plot of a likely species-level identification is presented by the Tricladida species Polycelis felina, which had not previously been described as part of the Iberian desman's diet and this study found in notable abundance and prevalence. The consensus

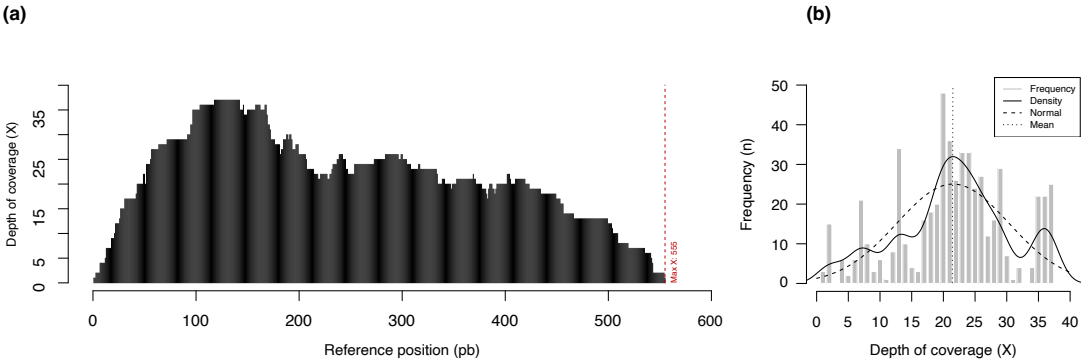


Figure 21. Depth coverage across reference. Consensus sequence assigned to the triclad Polycelis felina: (a) depth of coverage across the reference for a partial sequence of the COI gene, where Max X (x-axis) indicates the length of the reference sequence; (b) the distribution of depth of coverage values, showing the density curve estimated, the fitted normal curve, and the indicated mean value.

Table 5 . Identification of consensus sequences at the species level. For each taxonomic unit, its family and species are shown, along with the origin of the classified consensus sequences, number of consensus sequences classified for that species (Freq.), and statistics of the longest sequence: length (bp), depth and breadth of coverage, total percent identity (Pident), and query coverage (Query cov.). Taxa described in the Iberian Peninsula are marked with an asterisk (*). Mit: Complete or partial mitochondrial genomes. Fragments of mitochondrial genes: COI, Cytb, 16S+tRNA+NAD1, rrnL+trnL+NAD1, ATP6.

Taxonomic Unit	Family	Species	Origin	Freq. (n)	Longest Consensus Sequence				
					Length (bp)	Depth (X)	Breadth (%)	Pident (%)	Query cov. (%)
Trichoptera	Hydropsychidae	Hydropsyche ambigua*	COI	5	607	3.14	93.94	100	99
	Hydropsychidae	Hydropsyche fulvipes*	Mit	1	8516	8.89	56.07	98	99
	Hydropsychidae	Hydropsyche siltalai*	COI	18	622	6.95	96.43	100	99
	Hydropsychidae	Hydropsyche tenuis*	COI	2	553	12.66	83.89	99	99
	Limnephilidae	Allogamus laureatus*	COI	8	658	24.95	99.85	99	99
	Limnephilidae	Potamophylax cingulatus*	COI, Mit	11	12396	135.63	97.39	98	99
	Limnephilidae	Potamophylax latipennis*	Mit, COI	5	16329	339.95	98.87	99	99
	Odontoceridae	Odontocerum albicorne*	Mit, COI	8	14671	191.33	96.77	97	99
	Philopotamidae	Philopotamus montanus*	Cytb, COI	34	699	1717.29	99.57	97	100
	Rhyacophilidae	Rhyacophila oblitterata*	COI	2	619	5.05	93.92	100	99
Coleoptera	Rhyacophilidae	Rhyacophila* nubila	Mit	2	3303	19.48	78.99	97	99
	Dytiscidae	Agabus didymus*	COI	2	646	5.43	98.18	100	100
	Dytiscidae	Hydroporus dobrogeanus*	Mit	1	2983	2.48	92.02	98	99
	Dytiscidae	Hydroporus pubescens*	16S+tRNA+NAD1	1	614	1.14	76.98	100	99
	Dytiscidae	Platambus* princeps	rrnL+trnL+NAD1	1	516	1.48	63.97	99	99
	Elateridae	Dalopius marginatus*	COI	2	555	1.11	83.74	99	99
	Hydraenidae	Limnebius truncatellus*	COI	12	790	4.68	95.4	100	99
Diptera	Chironomidae	Heterotrissocladius marcidus*	COI	1	564	5.58	85.41	99	99
	Chironomidae	Tvetenia calvescens*	COI	8	658	18.64	100	100	100
	Mycetophilidae	Mycetophila marginata*	COI	1	508	1.81	77.05	99	99
	Pediciidae	Pedicia* albivitta	Mit	1	585	9.81	78.47	98	99
Ephemeroptera	Baetidae	Baetis rhodani*	COI	3	552	1.89	80.03	98	99
	Ephemerellidae	Serratella ignita*	Mit	2	12131	4.47	81.93	98	99
	Ephemeridae	Ephemerella danica*	COI, Mit	2	14750	168.11	95.04	97	99
	Leptophlebiidae	Callibaetis humilis*	COI	5	658	271.69	100	100	100
Hemiptera	Aphididae	Rhopalosiphum padi*	COI, ATP6, Cytb, Mit	10	10405	6.68	68.31	98	98
	Cicadellidae	Rhytistylus proceps*	COI	1	658	8.03	100	100	100
	Nepidae	Nepa cinerea*	COI	2	585	2.23	88.75	100	99
Hymenoptera	Formicidae	Myrmica sabuleti*	COI	1	510	1.11	63.9	100	99
	Formicidae	Myrmica scabrinodis*	Mit	2	9205	1.32	59.93	98	95
Megaloptera	Sialidae	Sialis fuliginosa*	Mit	1	7606	3.5	76.15	98	98
Odonata	Cordulegastridae	Cordulegaster boltonii*	COI	1	584	5.14	94.17	95	99
Orthoptera	Acrididae	Omocestus panteli*	COI	1	529	3.34	87.27	100	99
Plecoptera	Perlodidae	Isoperla grammatica*	COI, Mit	18	14425	55.74	97.52	98	99
Isopoda	Porcellionidae	Porcellio scaber*	COI	1	536	3.56	81.06	99	99
Tricladida	Planariidae	Polycelis felina*	COI	1	555	26.01	100	95	100
Caudata	Salamandridae	Salamandra salamandra*	Mit	4	11604	2.54	70.91	99	99

sequences assigned to this species originated from a partial sequence of the mitochondrial COI gene (555 bp) (Table 5). Through the analysis of depth of coverage across the reference, Figure 21 illustrates the alignments of the reads that formed the consensus sequence identified as *P. felina* in sample BC2989. The reads fully covered the reference sequence uniformly along its entire length, with a typical decrease at the ends, an intrinsic characteristic of the methodology used (Figure 21a). This resulted in a distribution with a tendency toward normality in the depth of coverage values (Figure 21b).

Another example is provided by the identification of *Salamandra salamandra* (Caudata) in sample BC2876, another element not previously identified in the diet of the Iberian desman. In this case, the longest sequence was constructed from 279 reads aligned to the mitochondrial reference genome of 16,331 bp, covering 70.91% of it (Table 5). Although the mapped reads do not completely cover the reference sequence, the alignment profile analysis shows that they are homogeneously distributed (Figure S3).

2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS

The following results correspond to specific objective 9.

2.1. Species-level identification of *Yersinia* species

To evaluate the mapping approach for identifying closely related bacterial species, we first used the Bowtie2 aligner (end-to-end mode) to map reads from the exogenous set of each sample against the genomes of 26 *Yersinia* species (Table S11), with a Bowtie2 database constructed for each genome (Figure 13d).

Traditional species identification often relies on the number of mapped reads. However, coverage plots across the genome revealed that certain regions, either because they are highly conserved among different species in the sample or because they are repetitive, tend to accumulate a large number of reads, while the rest of the genome maps any read. This is illustrated in Figure 22, which compares the mappings of two samples, BC3903 and BC3345, to the *Yersinia aldovae* genome. Sample BC3903, with 12,378 reads, showed a mostly even distribution of reads across the genome, as expected for the mapping of reads in the genome of the same species. In contrast, BC3345, with a similar number of reads (13,790), showed extremely high coverage in certain regions (up to ~250X), while most regions of the genome were not covered by reads, suggesting that *Y. aldovae* is not present in this sample. Similar contrasting examples are shown in Figure S4. Therefore, direct application of the number of mapped reads to species identification may lead to false positives.

Given the large number of unspecific reads mapped, we found that the breadth of coverage was a more reliable parameter for species identification than the total number of reads. For example, in the previous example, sample BC3903 had 19.03% of its genome covered by at least one read, strongly suggesting the presence of *Y. aldovae* or a close relative, whereas BC3345 had only 1.5% covered (Figure 22 and Table S11), despite having a similar number of mapped reads. However, setting a threshold for the percent coverage for positive identification is not straightforward. According to initial analyses of *Yersinia* and other species, we explored the performance of two different thresholds for species identification: a first threshold at ≥3% coverage and a second at ≥2% coverage (corresponding to ~139 kb and 92 kb of an average *Yersinia* genome covered, respectively).

Using the stricter 3% threshold, we identified 13 instances of 5 distinct *Yersinia* species (Figure 23a). *Y. intermedia* had the highest prevalence, being detected in 9 samples, while *Y. aldovae*, *Y. entomophaga*, *Y. nurmii*, and *Y. ruckeri* were found in only one sample each. Percent coverages of these 13 identifications ranged from 3.35 to 45.38% of the genome, which perfectly correlated with number of mapped reads (Table S11). The mean MAPQ values for the mapped reads of the positive identifications were close to the maximum value (ranging from 40.66 to 41.89), meaning that most reads were perfectly aligned in the correct position in the reference genome used. Lowering the threshold

V. RESULTS
2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS
2.1. SPECIES-LEVEL IDENTIFICATION OF YERSINIA SPECIES

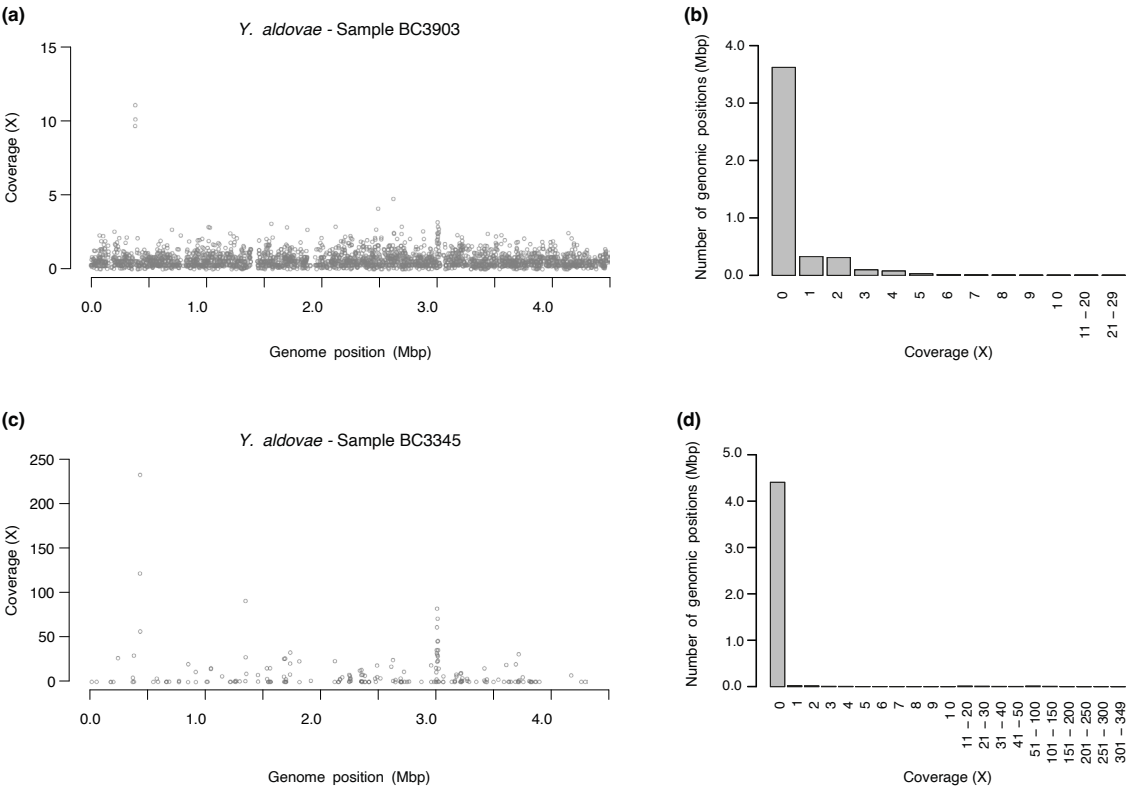


Figure 22. Comparison of depth of coverage across the reference genome of Yersinia aldovae and the corresponding depth of coverage barplots for two samples of Iberian desman. (a) coverage across the reference genome in 4,000 windows and (b) coverage barplot for the BC3903 sample, in which 12,378 reads were assigned, resulting in a mean depth coverage of 0.41X and a breadth of coverage of 19.03%. (c) coverage across the reference genome and (d) coverage barplot for the BC3345 sample, in which 13,790 reads were assigned, resulting in a mean depth coverage of 0.46X and a breadth of coverage of 1.5%. In (a) and (c), windows with zero coverage are not plotted. Note the different coverage scales in (a) and (c).

to 2% genome coverage resulted in three additional identifications (Figure 23a). The previously found species Y. intermedia was detected in one additional sample, and two new species, Y. alsatica and Y. enterocolitica, were detected in only one sample each. Similar identification results were obtained when using a combined Bowtie2 database with all Yersinia species (Figure S5a). However, three of the detections were missing, likely indicating fewer unspecific mapped reads with combined databases. We also verified that the percent coverage was very similar when positions with extreme depth of coverage were filtered out, resulting in the same identifications.

In sample BC3876, where four Yersinia species were identified, we used a Venn diagram to compare mapped reads and assess whether all four species were actually present in the sample. We found that each species had a significant number of unique reads, with the only exception of Y. entomophaga and Y. nurmii, which shared a substantial number of reads (1,660, representing 18% of the reads mapped to Y. nurmii; Figure 23b). The analysis with the combined database including all Yersinia species corroborated this finding, as Y. nurmii was not positive in BC3876

V. RESULTS
2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS
2.1. SPECIES-LEVEL IDENTIFICATION OF YERSINIA SPECIES

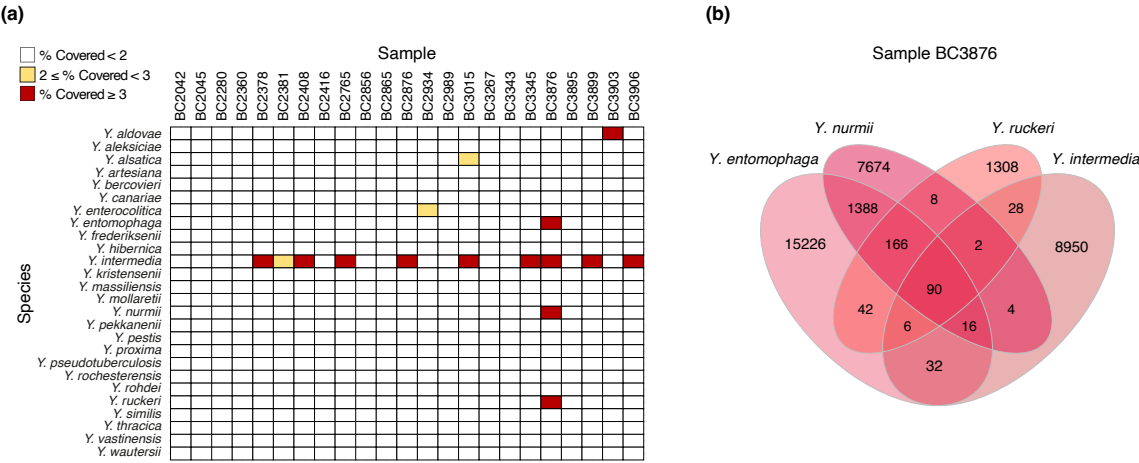


Figure 23. Identification of Yersinia species in 23 samples of Iberian desman and evaluation of read sharing of multiple species found in a single sample. (a) Species identification according to the percentage of the genome covered by mapped reads. (b) Venn diagram comparing reads assigned to four Yersinia species in sample BC3876, showing shared reads (intersections) and unique reads (non-intersecting areas) for each species.

(Figure S5a). However, all other three species were consistently detected in the same sample in both analyses, which, together with the low number of shared reads among the three species, suggests that all three are likely present in the sample.

2.2. Verification of the identification criterion by comparison with other mapping tools

To verify the mappings obtained with the default Bowtie2 end-to-end algorithm, we compared them with those generated using Bowtie2 in local mode, BWA-mem, and BWA-aln. For all samples, BWA-mem reached high mapping rates (e.g., an average of 270,298 reads aligned to Y. intermedia, compared to 6,511, 6,459, and 7,540 reads for Bowtie2 end-to-end, Bowtie2 local, and BWA-aln, respectively) and also higher breadth of coverage values. This was achieved mostly by reducing both read length and insert size, indicating that this mapping algorithm, in its default configuration, exhibits less stringent behavior and has no discrimination power at the species level. As for the other three algorithms, the number of reads mapped to each sample and the breadth of coverage values were similar (Figure 24 shows the results for Y. intermedia). BWA-aln achieved a higher number of mapped reads, although a consistent pattern was maintained for the three methods across all samples. Consequently, the determination of the Yersinia species was similar for these three methods, with both Bowtie2 algorithms giving identical identifications and BWA-aln giving some additional positives (Figures S5b and S5c). However, the results would also be equivalent for the latter method if the breadth of coverage thresholds used for positive identification were raised slightly. Therefore, the three stringent mapping algorithms showed similar performance in species identification.

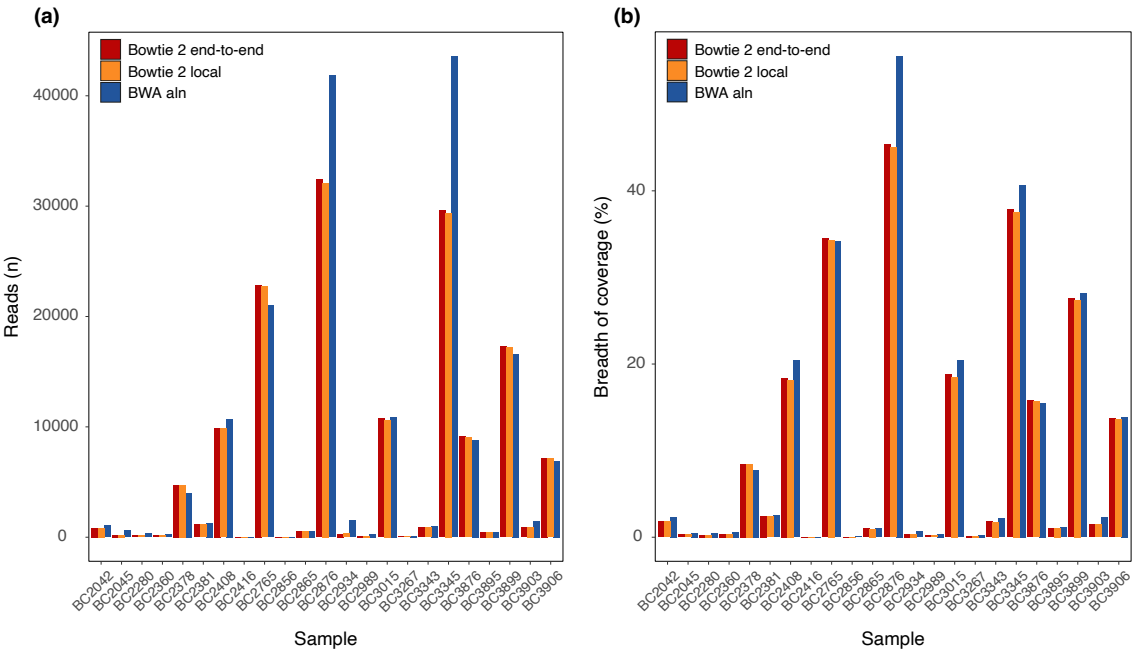


Figure 24. Comparison of different mapping tools (Bowtie 2 end-to-end, Bowtie 2 local, and BWA-aln) using the *Yersinia intermedia* genome as an example. (a) Number of reads aligned to each sample by each mapping tool. (b) Percent of the genome covered achieved by each mapping tool.

2.3. Application to the detection of pathogenic bacterial species

Due to its consistent performance, the Bowtie2 end-to-end algorithm was used in conjunction with the single-species databases for the pathogenic bacteria identification. Using the more stringent 3% genome coverage threshold, we identified 57 instances of pathogenic species belonging to 19 different species (Figure 25). Table 6 details these 19 bacterial species, their associated pathologies, and other relevant wildlife species in which these pathogens were found. The most frequently detected species were *Acinetobacter baumannii* and *Fusobacterium necrophorum*, each found in 7 samples. Samples BC2381 and BC2934 had the highest number of pathogenic species, with 14 and 8 species, respectively. Notably, all but 7 samples were found to contain at least one pathogen. Percent coverages started at 3.29% and were as high as 91.74% for *Shigella sonnei* in one sample. The mean MAPQ values for the positive identifications were close to the maximum (ranging from 40.95 to 41.98). When the threshold was lowered to 2% coverage, 19 additional identifications were obtained. However, many of these corresponded to species detected in only one sample (Figure 25). A similar identification pattern was observed when all genomes were analyzed using a combined Bowtie2 database, although the total number of identifications was reduced to 50 with the 3% threshold, and only 7 additional identifications were obtained with the 2% threshold (Figure S6).

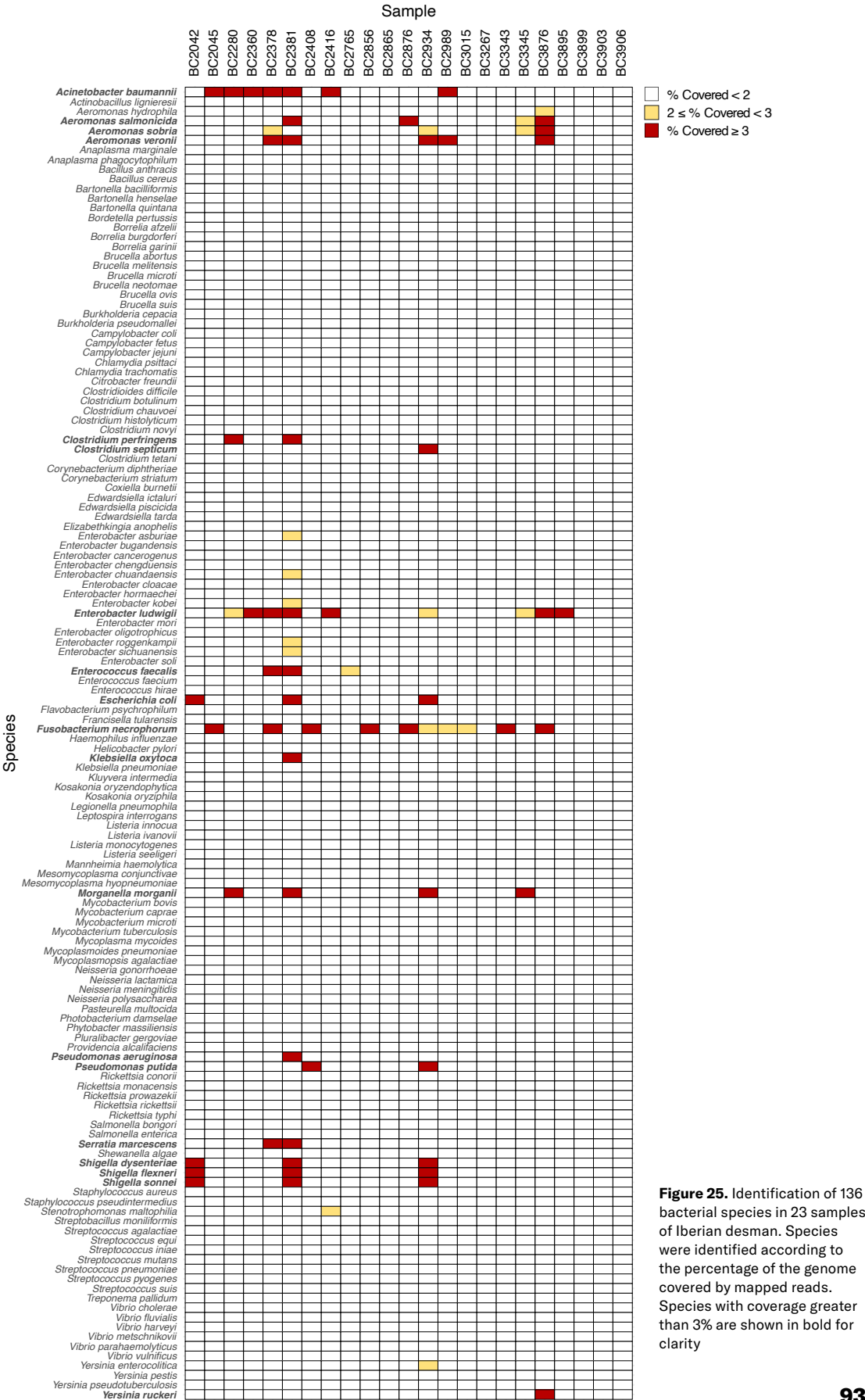


Figure 25. Identification of 136 bacterial species in 23 samples of Iberian desman. Species were identified according to the percentage of the genome covered by mapped reads. Species with coverage greater than 3% are shown in bold for clarity

2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS
2.3. APPLICATION TO THE DETECTION OF PATHOGENIC BACTERIAL SPECIES

Table 6. Bacterial species found in Iberian desman feces with the threshold of 3% of the genome covered by mapped reads.
For each species, its prevalence and a brief description of associated pathologies as found in different sources are given. In addition, examples of other relevant wildlife species in which the bacterial species has been found are given.

Bacterial species	Prevalence	Associated pathologies	Examples of wildlife species where the bacterial species was found
<i>Acinetobacter baumannii</i>	7	Opportunistic infections, pneumonia, bloodstream infections, meningitis	European mink (Cano-Terriza et al., 2017)
<i>Aeromonas salmonicida</i>	3	Furunculosis in fish, particularly salmonids	Salmonids (Austin & Austin, 2016); Lamprey (Diamanka et al., 2014)
<i>Aeromonas sobria</i>	1	Wound infections and gastrointestinal infections in fish	Fish (Austin & Austin, 2016)
<i>Aeromonas veronii</i>	5	Wound infections, diarrhea, and sepsis in fish	Fish (Austin & Austin, 2016); Pond turtle (Guz et al., 2021)
<i>Clostridium perfringens</i>	2	Food poisoning, gas gangrene, enteritis necroticans.	Eurasian otter (Rohner et al., 2022)
<i>Clostridium septicum</i>	1	Gas gangrene	
<i>Enterobacter ludwigii</i>	6	In humans, catheter associated bloodstream infections	
<i>Enterococcus faecalis</i>	2	In humans, urinary tract infections, bacteremia, peritonitis, endocarditis	Beaver (Laukova et al., 2015)
<i>Escherichia coli</i>	3	In humans, diarrhea, urinary tract infections, sepsis, meningitis	Eurasian otter, Pond slider, American mink (Vulfson et al., 2001; Rohner et al., 2022; Mengistu et al., 2022)
<i>Fusobacterium necrophorum</i>	7	Hoof disease in horses, lameness in sheep, liver abscesses in cattle	
<i>Klebsiella oxytoca</i>	1	In humans, opportunistic infections	Pond slider (Mengistu et al., 2022)
<i>Morganella morganii</i>	4	In humans, post-operative wound and urinary tract infections	Pond slider, American mink (Mengistu et al., 2022)
<i>Pseudomonas aeruginosa</i>	1	Opportunistic infections	Pond slider (Mengistu et al., 2022); American mink (Bai et al., 2023)
<i>Pseudomonas putida</i>	2	Opportunistic infections, ulceration in fish	Fish (Ardura et al., 2013)
<i>Serratia marcescens</i>	2	Opportunistic infections	
<i>Shigella dysenteriae</i>	3	Shigellosis (Dysentery)	
<i>Shigella flexneri</i>	3	Shigellosis (Dysentery)	
<i>Shigella sonnei</i>	3	Shigellosis (Dysentery)	
<i>Yersinia ruckeri</i>	1	Enteric redmouth disease in salmonids	Salmonids (Austin & Austin, 2016); Fish, Muskrat (Pajdak-Czaus et al., 2019)

In four cases where species of the same genus or closely related species were detected in a single sample, we used Venn diagrams to check species identification accuracy. For instance, three *Aeromonas* species were detected in sample BC3876 with each having a significant number of unique reads, confirming the likely presence of the three species in this sample (Figure S7). Samples where only one or two of the three species were assigned support this result. Conversely, *Escherichia coli* and the three species of the genus *Shigella*, which are four very closely related species (Chattaway et al., 2017), coincided in three samples, but the Venn diagram shows that these species share most reads (Figure S7). When all genomes were analyzed in a combined database, only one sample was positive for these four species whereas the other two samples were negative for the four bacteria (Figure S6). Thus, in

2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS

analyses of very closely related species, read sharing due to unspecific mappings is found when using separate databases, whereas loss of read mappings due to ambiguous mapping and low MAPQ values occur in combined databases.

2.4. Overall health assessment of the Iberian desman population

A PCA based on the breadth of coverage showed that most of the Iberian desman samples clustered together, with several outliers clearly defined (Figure 26). In particular, samples BC2381 and BC2934, previously mentioned as having the highest number of pathogenic bacterial species, are appreciably separated from the main group. Other outliers in the plot (BC3876, BC2045 and BC2378) show varying numbers of assigned pathogenic species, although they generally had a high number of pathogens with a high breadth of coverage. Four of the desmans with atypical values were from the Endrinal, while one belonged to the Adaja hydrological unit (Table S1).

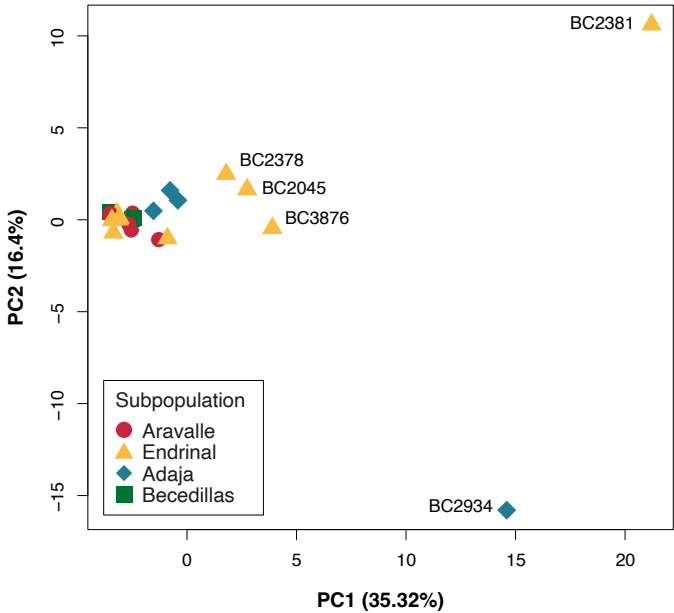


Figure 26. Principal Component Analysis based on breadth of coverage of each pathogenic bacterial species in the 23 Iberian desman samples. The sample names of outliers are indicated. The percentage of variance explained by each principal component (PC1 and PC2) is shown on the corresponding axis.

VI. DISCUSSION

Fresh fecal samples from *Galemys pyrenaicus* proved to be a valuable source of DNA, which, in combination with shotgun metagenomics sequencing, enabled the acquisition of both endogenous and exogenous genomic information. This approach facilitated the analysis of population characteristics of the species and key aspects of its ecological interactions. These findings highlight the potential of such minimally invasive samples for studying endangered species, providing crucial information for their management.

The focus of this work was on developing a bioinformatic pipeline that enabled the separation of the initial metagenomic sequences into different sources and the analysis of each set. It was demonstrated that endogenous DNA was sufficient for reconstructing complete mitochondrial genomes and obtaining low-coverage nuclear genome sequences, which facilitated genomic analyses. Additionally, a new methodology was developed for the taxonomic classification of exogenous metagenomic sequences, using various reference databases to describe the overall fecal content, as well as to identify diet components and parasites that may affect the health of this species. This pipeline made use of existing database information, consisting mainly of partial sequences of commonly used nuclear and mitochondrial genes for taxonomic species classification. For the identification of pathogenic bacteria in the feces of the Iberian desman, the bioinformatic approach focused on precise species-level classification, using a methodology adapted to the available complete reference genomes for these organisms.

1. DNA COMPOSITION, ENDOGENOUS DNA, DIET AND PARASITES FROM IBERIAN DESMAN FECES

The following discussion corresponds to specific objectives 1 to 8.

1.1. DNA composition of fresh fecal samples

Metagenomic analysis of the DNA content present in the fresh feces of *G. pyrenaicus* allowed us to assess its potential as a valuable DNA source for genomic studies relevant to the conservation of this species. Using fresh feces overcame two of the most critical challenges associated with non-invasive samples. First, DNA of sufficient quantity and quality was obtained. Furthermore, this approach minimized contamination with environmental DNA, reducing artifacts in the results and ensuring more reliable data for genomic and metagenomic analyses (Hawlitschek et al., 2018; Oliveros et al., 2023).

The use of fresh feces avoided the need for enrichment steps of endogenous DNA, achieving amounts similar to those reported in recent studies using metagenomic approaches on the feces of wild animals. The average proportion of nuclear endogenous DNA obtained in our study was 9.32% (Table 1), comparable to the values reported by de Flamingh et al. (2023) for thirteen samples collected from fresh elephant (*L. africana*) feces, which achieved an average of 12.41%. Additionally, we achieved similar averages for depth and breadth of coverage, with values of 0.49X and 19.81% (Table S5), respectively, compared to the 0.36X and 21.11% reported by de Flamingh et al. (2023). Notably, while the maximum breadth of coverage recorded by de Flamingh et al. (2023) was 35.98%, our results achieved nearly 80% coverage of the nuclear genome of *G. pyrenaicus* for one sample, with several samples exceeding 20% (Table S5).

On the other hand, Taylor et al. (2022) achieved a higher mean, with 30.87% endogenous DNA obtained from caribou (*Rangifer tarandus*) feces. This result may have been due to the careful selection of samples used, as they processed only four samples, and the range of values obtained varied significantly, from 5.81% to 59.71%. This aligns with the considerable variation observed in our results (Figure 15), where six samples achieved nuclear DNA amounts of *G. pyrenaicus* close to or above 20% (Table S5). In any case, all values obtained by metagenomics are much higher than those calculated in the first attempts to shotgun sequence fecal DNA, where only 1% of the reads were found to be endogenous (Perry et al. 2010).

The proportion of mitochondrial endogenous DNA we obtained was also similar to that reported by de Flamingh et al. (2023), representing a substantially smaller quantity compared to nuclear DNA. However, in our case, we achieved a higher recovery of mitochondrial reads, with an average of 0.03% (Table 1) and mean depth and breadth of coverage of 154.25X and 99.40% (Table S5), respectively, compared to the average mitochondrial reads of 0.00006% and mean depth and breadth of coverage of 79.4X and 99.90%, respectively, reported by de Flamingh et al. (2023).

At the same time, the use of fresh samples prevented the proliferation of bacterial DNA and unclassified sequences, as documented by de Flamingh et al. (2023). However, the number of unclassified sequences represented a significant fraction (68.51%) of the total DNA in our study (Table 1), although consistent with the studies cited previously (de Flamingh et al., 2023; Taylor et al., 2022). A considerable portion of these reads may be related to the degradation state of the DNA present in the feces, as it has been demonstrated that the digestive process prior to defecation affects its quality (Paula et al., 2015). Moreover, the fraction of unclassified reads may be related to the representation bias in the database. The RefSeq database used in this study (V208) contained only 4% of invertebrate species relative to the total species represented, while bacteria accounted for 59%. This disparity remains evident in the latest version of RefSeq (V227) (<ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/archive/>). Although the number of invertebrate species nearly doubled compared to V208 (from 4714 to 8381), they now represent only 5% of the total, in contrast to 57% for bacteria. Since invertebrates constitute the main diet components of *G. pyrenaicus*, improved representation of this group in databases is expected to contribute to a higher number of classified reads. The influence of database completeness on metagenomic results has already been well documented (Wright et al., 2023), as well as the need for more reference genomes from across the entire tree of life (Theissinger et al., 2023).

Nevertheless, a considerable proportion of reads of metazoan origin (8.75% of the total number of reads) has been classified, as well as those of bacterial origin (10.53%), which together with endogenous nuclear DNA constitute the three main sources of DNA present in feces (Table 1). These results support the use of shotgun metagenomics without enrichment steps to enable both analyses based on endogenous DNA and those based on exogenous DNA, thereby maximizing the information that can be obtained with this method.

1.2. Use of endogenous DNA in population genomic studies

In our work, endogenous DNA at the whole-genome scale, combined with bioinformatic methods optimized for low-coverage data (Korneliussen et al., 2014), demonstrated great potential for conducting population structure analyses of the Iberian desman and evaluating connectivity between subpopulations. Through PCA analysis (Figure 17), population structure was assessed, identifying four groups based on the sub-basins from which the individuals originated, with genetic affinities between subpopulations consistent with their geographical distance. While this method generally provides information on ancient migration patterns, a recent dispersal event was also detected from Aravalle to Endrinal, corresponding to individual BC2865, identified as a female based on sex determination using depth of coverage values (Table S7, Figure 16). This pattern of structuring tied to river of origin was also observed in *G. pyrenaicus* by Escoda et al. (2017) using ddRAD data, where low levels of contemporary inter-

river dispersal compared to intra-river dispersal were also observed based on relatedness networks, showing that desmans rarely move between different river basins. This was associated with the philopatric behavior of both males and females and the high levels of inbreeding characteristic of the species (Escoda et al., 2017). However, philopatric behavior does not exclude the possibility of some long-distance exploratory movements in desmans, possibly associated with changes in their habitat (Gillet et al., 2016).

Furthermore, phylogenetic analyses were conducted on the reconstructed complete mitochondrial genomes (Figure 18). As expected for the mitochondrial genome, the tree did not show sufficient resolution to address the fine-scale geographic patterns examined in this study, but it could be very useful for future studies of the entire species range. Taken together, these approaches collectively highlight the versatility and applicability of metagenomic methodologies in genomic studies relevant to conservation efforts. Future research should explore whether analyses of admixture, gene flow, and kinship relationships, among others, using genotype probabilities, are feasible with this type of data.

1.3. New method for the classification of diet elements and parasites using metagenomic data

The method developed (Figures 13 and 14) for the identification of dietary elements and parasites aimed to maximize the information present in feces while ensuring accuracy in taxonomic assignments. Several features of the method accounted for its effectiveness in obtaining the results.

The use of Bowtie2 enabled high-quality alignments of the primary reads and the generation of a large number of consensus sequences larger than 500 bp, allowing for more precise identifications compared to those obtained through the direct classification of the reads or of shorter sequences used in other studies, for example, in dietary studies of the Iberian desman through metabarcoding (Biffi et al., 2017; Gillet et al., 2015; Hawlitschek et al., 2018). Additionally, the generation of consensus sequences allowed us to use of depth of coverage as an abundance parameter, in relation to the length of the reference sequence where the reads map, and to account for representation bias in the databases, given that a more represented taxon could achieve a higher number of mapped reads. The calculation performed here for estimating abundance also provided a more accurate value compared to the direct counting of reads, which is typically used (Chua et al., 2021; Srivathsan et al., 2015, 2016).

The reference databases used take advantage of two key features. First, the higher number of copies of mitochondrial genomes and ribosomal genes, which ensures a good fraction of these DNA fractions in the sequencing data (Andrews et al., 2018). Second, they are well represented in the databases, especially because they are used as barcodes or phylogenetic markers due to their discriminatory power among

taxa. To maximize the effectiveness of the analysis, four different databases were used for this purpose.

The mitochondrial sequence database was constructed with different mitochondrial fragments and genomes used in taxonomic identification, with the aim of diversifying the markers to be used and not centering classifications on a single locus, such as COI. Mitochondrial sequences are generally used for metazoan identification (Antil et al., 2023). The SILVA database (Quast et al., 2013) contributed ribosomal markers of nuclear origin, and was specifically incorporated because it has many markers commonly used in identification of nematodes, flatworms, and protists (Antil et al., 2023; Kounosu et al., 2019). Additionally, BOLD (Ratnasingham & Hebert, 2007) was included as a complement, providing sequences absent in the other databases, including both mitochondrial and ribosomal markers. These four databases served as references for sequence mapping, but taxonomic classification was performed using the latest version (as of October 2024) of the BLAST Core_nt database from NCBI (Sayers et al., 2025) to unify taxonomy, standardize e-value comparisons, and serve as a double-check to verify the quality of consensus sequences. For example, the use of this database allowed the detection of errors originating from chimeric reference sequences formed by eukaryotic consensus sequences contaminated with bacterial sequences. These errors could be identified because the BLAST Core_nt database contains reference sequences from all three domains, including Bacteria, which were excluded from the primary databases to optimize computational efficiency.

Thus, the novel pipeline took advantages of the combined use of Bowtie2 (Langmead & Salzberg, 2012) and BLAST (Altschul et al., 1990). It allowed for the rapid and precise alignment of the large volume of reads with Bowtie2 and, through the key step of generating consensus sequences, overcame one of BLAST's main limitations by reducing its processing time (Breitwieser et al., 2019), making it possible to leverage the sensitivity and accuracy of this powerful sequence search tool.

1.4. Taxonomic identification parameters

A commonly used criterion for species-level identification, particularly in metabarcoding studies, is 98% identity between reads and reference sequences (Biffi et al., 2017; Esnaola, Arrizabalaga-Escudero, et al., 2018; Gillet et al., 2015). If the sequence does not achieve this value, it is considered as not classified. Although this criterion is mainly applied to the COI gene, its use has been extended to other markers. However, this fixed threshold can result in low assignment rates depending on the representation of the database used. For this reason, incorporating higher taxonomic levels helps increase classified sequences and provides more representative results of the diversity in the samples.

Assignments to higher taxonomic levels are primarily based on relaxing the required percentage of identity. For example, for insect orders, 93% identity is considered

sufficient to make reliable assignments with COI gene fragments of 157 bp and 16S rDNA fragments of 106 bp (Alberdi et al., 2018). In our case, the use of longer queries (> 500 bp) increased the reliability of alignments (Alberdi et al., 2019), allowing the identity threshold to be relaxed to 90%. Nonetheless, the vast majority of identifications recorded 100% identity (Figure S2).

This relaxation of the percent identity threshold was coupled with the incorporation of query coverage as additional quality parameter. With a threshold of 50% query coverage, it was required that at least 250 bp participate in the alignment, although in the vast majority of the assignments 99% of the query sequence was involved (Figure S2). In contrast, other studies have used thresholds of only 50 to 100 bp of overlap, employing various markers such as COI, 18S, and 28S rDNA, chloroplast markers (rbcL, matK, trnL-F, and trnH-psbA), and nuclear markers ITS1 and ITS2 (Chua et al., 2021; Srivathsan et al., 2015, 2016)

These parameters allowed for the identification and description of diet and parasites at various higher taxonomic levels. To assess the possibility of making species-level assignments, stricter criteria were applied. In addition, we used exclusively mitochondrial reference sequences due to their higher evolution rates and therefore better discrimination between closely related species (Antil et al., 2023; Elyasigorji et al., 2023; Gissi et al., 2008). As before, a 50% breadth of coverage was required to ensure the inclusion of variable regions, particularly when reference sequences came from complete or partial genomes. Additionally, the identity threshold was increased to 94%, and at least 90% of the consensus sequence had to participate in the classification, which in many cases led to the recovery of nearly complete mitochondrial genomes (Table 5).

Although species-level identifications have been performed using the ribosomal genes 18S and 28S in metabarcoding approaches, primer pairs targeted the regions of highest diversity (Bradley et al., 2016; Kang et al., 2024; Kounosu et al., 2019; Wylezich et al., 2010). Since most sequences in the SILVA database contain both conserved and variable regions, it was not possible to use it for species-level identification.

1.5. Assessment of the diet using metagenomic data

In general, the dietary components of the Iberian desman detected in the 23 samples analyzed from the populations of the Central System were consistent with those previously identified in other populations, using both morphological (Casti n & Gos lbez, 1995, 1999; L pez-de Sancha et al., 2022; Santamarina, 1993) and genetic methods (Biffi et al., 2017; Esnaola, Arrizabalaga-Escudero, et al., 2018; Gillet et al., 2015; Hawlitschek et al., 2018). Most of the detected groups belonged to aquatic organisms, although some exclusively terrestrial groups were found with lower prevalence (Table 3). Additionally, some fewer common groups in the desman's diet were identified.

To analyze the diet composition in quantitative terms and in an attempt to detect prey preferences, the combined analysis of abundance and prevalence can provide a good approximation (Figure 19), although not in definitive terms. On the one hand, prevalence may serve as a more direct estimator of preference, while abundance, relative to the number of reads, reflects the biomass contribution of each prey. However, factors such as digestion rates influenced by body structural differences in the diet elements could affect the relationship between biomass and the number of reads (Deagle et al., 2010; Srivathsan et al., 2015). For example, Diptera, Coleoptera, and Plecoptera showed higher prevalences but low abundance. These results suggest trends in the feeding behavior of the desman; however, data such as the abundance of each prey in the ecosystem or the number of items present and their size could complement this analysis to draw more accurate conclusions on the diet preferences of the Iberian desman (Esnaola et al., 2021; L pez-de Sancha et al., 2022).

The main groups of insects detected here were Trichoptera, Ephemeroptera, Diptera, Coleoptera, and Plecoptera, with lower prevalence of Hemiptera, Odonata, and Megaloptera (Table 3). Except for Megaloptera, all groups are commonly described as regular components of the Iberian desman's diet (Biffi et al., 2017; Casti n & Gos lbez, 1995; Gillet et al., 2015; Hawlitschek et al., 2018; L pez-de Sancha et al., 2022; Santamarina, 1993). Additionally, other less frequent groups, previously described in the literature were identified, such as Salmoniformes, recorded as *Salmo trutta* eggs by Casti n & Gos lbez (1995). Gastropoda and Oligochaeta were reported in similar proportions (Biffi et al., 2017; Casti n & Gos lbez, 1995), while new items detected in this work included Bivalvia and Caudata (Table 3). All four groups were found in only 1 or 2 samples, representing very isolated cases. The latter was identified at the species level as *Salamandra salamandra* through mitochondrial sequences, the longest of which corresponded to the complete mitochondrial genome, with 71% breadth of coverage (Figure S3). This generated a consensus sequence of 11,604 bp that contributed 99% to the BLAST alignment, achieving 99% identity.

Two particular cases were the detection of the nematomorph Gordioida, with medium prevalence (in 7 individuals), and more notably, the planarian Tricladida, which was detected with high prevalence (in 18 individuals) and also high abundance, representing the second most abundant group (Table 3). Neither of these groups had been previously reported in the literature. In morphological studies this is likely due to the difficulty of identifying soft-bodied prey in feces after digestion (L pez Sancha, 2022). Additionally, Tricladida is a group of Platyhelminths, known to have very high mitogenomic evolutionary rates (Jakovli  et al., 2023), making it difficult their amplification using universal primers in genetic studies of diet. The difficulty of identifying Platyhelminthes and Nematomorphs using eDNA metabarcoding in riverine environments was reported by (Poyntz-Wright et al., 2024), possibly due to PCR failures and limitations in the markers used.

Regarding planarians, it is noteworthy that the species *Polycelis felina* was identified through the generation of a consensus sequence based on 100% breadth of coverage of a COI gene sequence, producing a 555 bp sequence

that participated fully (100%) in the BLAST alignment, achieving 94% identity (Table 5), strongly suggesting the presence of this species in the diet of the desman. *Polycelis felina* occurs in the Iberian Peninsula and inhabits cold and fast flowing mountain streams (Roca et al., 1992), which overlap with the desman's habitat, further supporting the likelihood of its inclusion in the diet of the desmans analyzed in this work. However, as the samples of this study are from a single population and were all collected in summer, it remains to be seen in future studies whether planarians are important dietary components in other populations and at other times of the year.

Although Gordioida have a free-living adult phase, potentially making them an active prey for the desman, they are also known as parasites of arthropods during their larval stage (Poinar, 2008), so it cannot be ruled out that the detected sequences originate from the parasitic forms of preyed insects.

Finally, as previously reported, terrestrial groups were detected as part of the desman's diet, although to a lesser extent than aquatic groups. They were mainly insects such as Lepidoptera (Biffi et al., 2017; Esnaola et al., 2021), Hymenoptera and Orthoptera (Biffi et al., 2017), and, with lower prevalence, arachnids like Araneae and Opiliones, as well as the crustacean Isopoda, also reported by Biffi et al. (2017) (Table 4). The latter was identified at the species level, Porcellio scaber, through 99% overlap of a 536 bp COI consensus sequence, with 99% identity (Table 5). Unlike other studies (Biffi et al., 2017; Esnaola et al., 2021; Gillet et al., 2015; Peyre, 1956; Santamarina, 1993), no amphipod crustaceans were found, which have been described as negatively selected by the desman using morphological analysis of the diet (Castián & Gosálbez, 1995).

In general, the results were comparable to those previously described through both morphological and genetic methods, for both common dietary components and particular cases. However, the newly identified groups demonstrate that the application of shotgun sequencing combined with the developed bioinformatics method is highly sensitive, allowing the challenges of previously applied techniques to be overcome. The use of non-invasive samples enabled the identification of soft-bodied organisms, like planarians and Gordioida, difficult to find in morphological analyses and PCR-based methods. This resulted in the detection of a high diversity of groups. For example, the average number of orders per sample detected by Biffi et al. (2017) was 3.7, while in our study, an average of 7 orders/classes was detected.

Regarding the species level identifications, it is important to highlight that almost all correspond to species described in the Iberian Peninsula. In addition, all species were part of the diet and no parasite was identified at the species level, probably due to the lack of reference sequences. This result underscores the importance of having reference sequences from the geographical area of study and the need to expand them to include other groups in order to achieve better species level identifications.

1.6. Identification of parasites using metagenomic data

Regarding parasite identification, this is the first study to conduct an in-depth analysis using shotgun metagenomics in feces for their detection in *G. pyrenaicus*. The most frequently detected order of parasites, present in all samples and with high abundance, was Coccidia (Figure 20). This group had the highest number of reads assigned (Table 4) This result aligns with that reported by Ripa et al. (2023), where the genus of coccidia *Eimeria spp.* was detected with a prevalence of 92.3% in another Iberian desman population using commercial qPCR kits.

Genera identified in our analysis included *Eimeria* as well as *Isospora*. Both belong to the family Eimeriidae, known for completing their life cycle in the intestinal tract of the host (Berná et al., 2021). They have been identified in the feces and intestinal lumen of animals, which could explain the abundance of Coccidia reads (Mehlhorn, 2016). This identification is particularly important given that these parasites cause coccidiosis, one of the most significant diseases in farm animals (Mehlhorn, 2016). Eimeriidae have been identified in insectivores of the same family as the Iberian desman, such as the European mole (*Talpa europaea*), where four species of *Eimeria* and four of *Isospora* were described (Duszynski & Upton, 2000), with multiple infections by intestinal eimeriids consistently detected (Duszynski, 2021) Species level identification is not easy at the moment with these species. For example, the genus *Eimeria* contains more than 800 species, and its mitogenome has been shown to be highly conserved among species. Thus, a variation of only 0–0.5% has been observed between species infecting turkeys (Ogedengbe et al., 2014). Additionally, as a highly host-specific parasite, it is likely that the lack of reference sequences for the species affecting the Iberian desman results in the distribution of reads across closely related sequences, preventing sufficient discrimination for confident species-level assignment. Undoubtedly, having the reference genomes of the species that infect the Iberian desman would be essential for their detection by metagenomic sequences. Alternatively, other methods, such as those based on assembly could help describe the species affecting the Iberian desman.

Three helminth parasites were detected: two nematodes, Rhabditia and Mermithida, and one platyhelminth of the class Cestoda (Table 4). The order Rhabditia includes *Parastrongyloides winchesi*, a species previously described in the Iberian desman (Alvarez Mascato, 1990). The order Mermithida is primarily known as parasites of arthropods (Dorris et al., 1999); consequently, the recovered DNA could also originate from elements of the desman's diet. This is an abundant order in freshwater environments, and some of its members have life cycles with free-living stages, which can be found in freshwater bodies or sediments (Abebe et al., 2008). Therefore, it cannot be ruled out that they may also be ingested during foraging in riverine benthos. Cestodes have not been previously reported in *G. pyrenaicus*, although they are common parasites in small mammals, such as insectivores of the genus *Sorex*, which can serve as intermediate or definitive hosts (Morand et al., 2006). In rats and mice, cestodes cause bloody, slimy diarrhea and may lead to intestinal blockage (Mehlhorn, 2016).

Regarding ectoparasites, three orders of mites were detected (Table 4). Two of them include species previously described in the Iberian desman: Trombidiformes, which includes *Eadiea desmanae* (Lukoschus, 1969), and Sarcoptiformes, which includes *Lynxacarus dubinini* and *Orycteroxenus galemys* (Cordero del Campillo et al., 1994). Additionally, the order Mesostigmata was identified, which, although no members have been previously described as parasites of the Iberian desman, is widely known as a parasite of small mammals, such as rodents and insectivores (Karbowski et al., 2023; Morand et al., 2006).

Although some studies on the parasitofauna of the Iberian desman have been conducted, most have focused on intestinal helminths, and few have used non-invasive samples and molecular tools. Compared to other insectivores in the same order, such as Soricinae shrews, recognized as zoonotic reservoirs and/or amplifiers of arthropod-borne pathogens (Karbowski et al., 2023; Tian et al., 2023), little is known about the parasites affecting the Iberian desman, their implications for the species' health, or their role as intermediaries in pathogen transmission. In this context, this study represents a first approach using genomic and bioinformatics tools, providing comprehensive and exploratory information and opening the door to future, more in-depth analyses. Although it was not possible to achieve detailed identification at lower taxonomic levels, such as genus or species, due to the limited availability of reference sequences, ongoing efforts to improve the representation of all taxa in databases are expected to enable greater specificity with this method (Corominas et al., 2024; Lewin et al., 2018; Mc Cartney et al., 2024).

2. SPECIES-LEVEL IDENTIFICATION OF BACTERIAL PATHOGENS

The following discussion corresponds to specific objective 9.

2.1. Validation of a mapping methodology for the detection of pathogens from fecal samples

In this part of the thesis, we developed and validated a method to identify bacterial pathogens from fecal samples based on the mapping of unassembled reads obtained by metagenomic sequencing to reference genomes. This approach differs fundamentally from methods designed to profile the general composition of the microbiome, such as Kraken (Wood & Salzberg, 2014) or DIAMOND (Buchfink et al., 2014). Instead, it aligns more closely with pathogen detection methods used in human disease studies, where the primary objective is species-level identification of target pathogens, such as PathoScope (Francis et al., 2013; Hong et al., 2014) or SURPI (Gu et al., 2021; Naccache et al., 2014). To our knowledge, no studies have systematically applied these species-level pathogen detection methods to endangered wildlife species.

The effectiveness of our pipeline for wildlife health monitoring was demonstrated by its ability to identify pathogenic bacterial species in fecal samples from the Iberian desman. Several factors affecting the identifications were analyzed to better understand their impact, optimize the method, and guide their application in different scenarios. These factors included the choice of identification parameters, threshold settings, database configuration, and mapping tool.

To improve the accuracy of species identification, we used the breadth of genome coverage rather than the more commonly used number or proportion of mapped reads (Francis et al., 2013; Gu et al., 2021; Hong et al., 2014; Naccache et al., 2014), thus avoiding false positives caused by conserved or repetitive genomic regions, where reads from different species may map. Unspecific mappings can be particularly significant when the genome of the exact bacterial species present in the sample is not in the database and rather a closely related species is used for mapping. Therefore, using breadth of coverage for identification is especially important in studies of non-model and wildlife species, where the representation of pathogens in databases is often incomplete.

We analyzed two different operational thresholds for identifications using breadth of coverage, corresponding to 2% and 3% coverage, to evaluate their sensitivity and specificity. In general, the more stringent 3% threshold identified species that were present in multiple samples, each with different genome coverage values that correlated with the number of mapped reads and, by extension, with species abundance (e.g., *Yersinia intermedia*, Table S11). This scaling of genome coverage with the number of reads strongly supports correct mapping to the reference genome and therefore that the species is truly present in the population. Conversely, species identified with breadth of coverage values between 2 and 3% (or at lower levels) were typically detected in only one individual (Figures 23a and 25). Single-sample detections lacked the supporting evidence provided by the scaling of breadth of coverage with the number of reads and, while their presence cannot be entirely ruled out, it is more difficult to dismiss for them the possibility of mappings to conserved or repetitive genomic regions of the genome of a distantly related species. Therefore, a 3% threshold provided more confident identifications in this particular study, given the sequencing effort applied. However, it is important to note that the most appropriate threshold may vary depending on the sequencing depth and specific conditions of individual studies. For instance, in cases involving highly pathogenic species or potential outbreaks, all samples, even those with low numbers of mapped reads, should be considered for further investigation to ensure comprehensive monitoring and risk assessment.

The type of database to be used with the mapping algorithm is another critical factor and its choice depends on the objectives of the study. For example, for targeted monitoring of highly pathogenic species, it would be essential to first perform an analysis of unspecific mappings (Figures 22 and S4) and shared read levels (Figures 23b and S7) for the pathogen of interest in relation to closely related species using different databases. Our results showed that unspecific mappings from sequences mapping

to phylogenetically closely related species in the database, particularly in conserved regions, may be higher when using a single Bowtie2 database for each genome analyzed compared to the combined database. Furthermore, unspecific mappings may lead to more shared reads when using single-species databases. Despite this, single-species databases may be preferable for initial surveys, as using a combined Bowtie2 database for all genomes may result in the loss of potential positives due to ambiguous mappings with low mapping quality values, as observed for the *Shigella* and *Escherichia* species (Figures 25 and S6). In essence, single-species databases typically provide greater sensitivity because they focus on detecting specific pathogens without competition from sequences of other species, whereas combined databases improve specificity by offering a broader reference to discriminate between similar species, although they may reduce sensitivity due to ambiguous or low-confidence mappings to multiple species. Therefore, preliminary analyses should be conducted to determine the most suitable database and detection criteria for each pathogen under study to ensure that the method provides robust identifications.

The mapping tools Bowtie2 (in two different modes: local and end-to-end) and BWA (in aln mode) produced similar results in terms of number of mapped reads and breadth of coverage achieved (Figure 24). This was not the case with the less stringent BWA program in "mem" mode, which uses a less stringent algorithm. The fact that Bowtie2 and BWA in their optimal configurations consistently produced similar mapping results and thus similar species identifications, despite their algorithmic differences, reinforces confidence in their ability to provide accurate taxonomic classifications.

The analysis of the genus *Yersinia* illustrates the efficiency of our approach. Of the 26 species tested, seven different species were detected in the Iberian desman samples when using the more comprehensive 2% threshold. Among these, two species (*Y. enterocolitica* and *Y. ruckeri*) are potentially pathogenic, with each found in a single sample (Figures 23 and 25). These results highlight the importance of robust species-level identifications and demonstrate the power of the method proposed here to discriminate between species within the same genus. However, challenges remain when resolving exact species in groups of very closely related taxa, such as *Shigella* and *Escherichia* species, where shared genomic regions can lead to ambiguous mappings. In these cases, assembly-based approaches combined with high-depth sequencing may be necessary to achieve accurate species-level resolution (Blanco-Míguez et al., 2023; Nurk et al., 2017).

2.2. Implications for the conservation of the Iberian desman

Understanding pathogen burden and diversity in endangered species is critical to inform management strategies, particularly in the context of conservation efforts such as captive breeding and translocation programs (Gaywood et al., 2022). In our study, we identified 19 species of pathogenic bacteria in 23 Iberian desman fecal samples using a strict coverage threshold, with varying levels of genome

coverage and prevalence. The sample size was very limited, which prevents any generalization of the results. Undoubtedly, a larger sample size, covering a wider geographical area and different time points, would provide a more comprehensive understanding of pathogen prevalence and diversity in this mammal. Although the main focus of the work was methodological, with the aim of developing a strategy that can be used for health monitoring, this pilot analysis also provided essential baseline data on the presence and diversity of pathogens in the Iberian desman, which could help shed light on the causes of the rapid decline of some of its populations. Remarkably, as shown in Table 6, many of the identified pathogenic species are found in aquatic environments and are known to infect fish (Ardura et al., 2013; Austin & Austin, 2016; Diamanka et al., 2014; Pajdak-Czaus et al., 2019) and other aquatic or semi-aquatic species, including European mink, American mink, Eurasian otter, beaver, muskrat, pond slider, and pond turtle (Bai et al., 2023; Cano-Terriza et al., 2017; Guz et al., 2021; Lauková et al., 2015; Mengistu et al., 2022; Pajdak-Czaus et al., 2019; Rohner et al., 2022; Vulfson et al., 2001). In some cases, these infections have been associated with disease or mortality in these animals, underscoring the potential severity of some of these pathogens. The sharing of pathogens among aquatic species may be explained by the fact that the aquatic environment can facilitate the transmission of infections, spreading pathogens over a wide area and potentially increasing exposure to multiple microbial threats (Cabral, 2010). Thus, these results highlight the complex and potentially pathogenic microbial landscape to which the Iberian desman is exposed.

The PCA (Figure 26) revealed substantial variability in pathogen profiles among Iberian desmans, suggesting differences in pathogenic bacterial load or diversity of species present, which may be highly relevant for assessing the health status of the analyzed individuals. In particular, five desmans emerged as clear outliers in the PCA plot, indicating altered pathogen loads. It remains unclear whether these desmans contained these levels of pathogens by chance, due to a compromised immune system potentially related to inbreeding, or as a result of specific environmental conditions influencing pathogen loads in certain populations. The geographic distribution of the outlier desmans may provide a clue, as four of them belonged to the same hydrological unit, Endrinal, so that 40% of the desmans in this subpopulation would have an elevated or altered pathogen load. The Endrinal river supports a significant livestock density, which could be behind the unique pathogen load found in the desmans of this river. However, additional sampling and detailed environmental analyses in this and other hydrological units is needed to better understand the origin of these pathogens and the factors driving the altered pathogen loads in some desmans and populations.

Future studies of Iberian desman pathogens should include different populations across the species' range and throughout the year to understand population health trends and to prioritize conservation efforts to the most vulnerable populations. Integrating the analysis of genetic factors, such as inbreeding and mutational load, will be crucial for understanding their influence on pathogen load and susceptibility. Furthermore, epidemiological studies should be a priority to understand how these

pathogens are transmitted, for example through water or through contact with other wildlife species, to identify potential transmission hotspots, such as excessive livestock densities in mountainous areas or wastewater discharge points near human settlements, and to assess their potential impact on the health and viability of Iberian desman populations.

2.3. Method challenges and prospects

This study represents a significant advance in pathogen detection methods for endangered species by providing a novel, effective, and non-invasive metagenomic approach for monitoring microbial diseases. To demonstrate the effectiveness of the method, we used a dataset of 136 bacterial reference genomes for mapping. We focused on bacteria because of their adequate representation of complete genomes in databases, as well as their reduced genome sizes, which allowed us to test a wide variety of parameters and analysis conditions with reasonable computational effort. Viral and eukaryotic pathogens, on the other hand, present unique challenges for metagenomic detection, which were beyond the scope of this study. Future research should address these challenges and optimize tools for detecting such pathogens, as they can also have a significant impact on the overall health and viability of wildlife species (Nunn & Altizer, 2005; Pedersen et al., 2007).

The dependence on reference genome databases is a major limitation of mapping methods, particularly for less studied species such as the Iberian desman. Our results showed that many pathogenic bacterial species present in genome databases (19 out of 136 analyzed) can be found in the Iberian desman. Most of these bacteria also infect humans, suggesting a wide host range for these pathogens (Shaw et al., 2020). However, since bacterial pathogens specific to the Iberian desman have not yet been sequenced, only closely related species present in databases can be detected using this bioinformatics method. Eukaryotic parasite species, especially those with complex life cycles, tend to be more host-specific (Poulin, 2006), so it remains to be seen whether genomes currently present in genome databases, usually isolated from humans, will be useful for detecting eukaryotic pathogens in the Iberian desman or other endangered mammals. In the long term, the generation of complete reference genomes for a wide range of relevant pathogenic and parasitic species will be key to advancing the detection and assessment of pathogens in wildlife, especially in critically endangered species.

Accurate identification of pathogens is not only critical in conservation biology, but also an integral part of the One Health approach, which emphasizes the important interrelationships that exist between human, animal, and environmental health (Destoumieux-Garzón et al., 2018; White & Razgour, 2020). By leveraging advances in high-throughput sequencing and metagenomic technologies, this study provides a reliable and non-invasive method to monitor pathogens that

may pose risks, not only to wildlife and endangered species, but also to humans. This underscores the importance of further research in this area, particularly to improve genomic tools and the representation of reference genomes of wildlife pathogens and parasites in databases, thereby contributing to the broader goal of protecting endangered species and maintaining ecosystem health.

3. NOVEL APPROACHES FOR WILDLIFE CONSERVATION: FECAL METAGENOMICS FOR MONITORING POPULATION GENOMICS, ECOLOGY AND HEALTH

The following discussion addresses the general objective.

3.1 Novel metagenomic methods developed in this thesis

In this thesis, the Iberian desman was used as a model species to develop metagenomic techniques that maximize the use of exogenous information contained in fecal samples. To begin with, the global DNA composition in desman feces was explored using the Kraken2 program (Wood et al., 2019) with the RefSeq database (O’Leary et al., 2016), which enables classification of sequence reads at high taxonomic levels across all domains of life, revealing the complex and information-rich nature of fecal DNA. However, to achieve finer taxonomic resolution, new approaches had to be developed, considering the specific characteristics of the data, such as low sequencing depth and the limited availability of reference sequences for species detection. These efforts resulted in the two methodological frameworks presented in the thesis, which apply different taxonomic identification approaches tailored to distinct types of databases, specifically phylogenetic markers and mitochondrial genomes on one hand, and whole genomes on the other.

To identify eukaryotic sequences derived from dietary components and parasites, mitochondrial genomes and phylogenetic marker sequences were selected. These sequences, selected to be of any length greater than 500 bp, have the advantage of being broadly representation in reference databases. Several databases (our own Mit11, Silva SSU & LSU (Quast et al., 2013) and BOLD (Ratnasingham & Hebert, 2007)) were used in a complementary manner to maximize taxonomic coverage. A method was developed to generate consensus sequences from Bowtie2 (Langmead & Salzberg, 2012) mappings and to perform their taxonomic identification using BLAST (Altschul et al., 1990). This approach used a low-stringency Bowtie2 configuration ("very-sensitive" option and unpaired alignment mode) to recover as many sequences as possible, delegating the taxonomic identification to BLAST. This was made feasible by constructing consensus sequences for BLAST analysis instead of using individual reads, which reduced computational workload, improved the reliability of taxonomic assignments, and enabled abundance

estimation. The intrinsic characteristics and resolution power of each marker were also considered, using only mitochondrial markers for lower-level taxonomic assignments (i.e. genus and species). The dietary analysis revealed not only the expected aquatic insect taxa but also novel components that had not been previously detected using other methods. These results significantly expand the known dietary repertoire and foraging behavior of *G. pyrenaicus*, highlighting the power of metagenomic approaches in ecology. Regarding parasitology, this study presents the first comprehensive analysis using metagenomic data from fecal samples to detect parasites in an endangered species, such as the Iberian desman. Although species-level assignments were constrained by the limited availability of reference sequences, the study provides valuable baseline information and opens new avenues for future research into the parasite burden and health status of endangered species.

Regarding the identification of pathogenic bacteria, the method implemented was based on the use of complete reference genomes, as they are more broadly available for this taxonomic group. In this case, the method was based on the direct mapping of metagenomic reads, and species identification was derived from the analysis of alignment coverage and quality. To achieve this, a more stringent Bowtie2 configuration was used (including the "sensitive" option, paired-end alignment mode with the --no-discordant and --no-mixed flags), along with a MAPQ threshold of ≥ 20 . The performance of this approach was evaluated through comparisons with other alignment algorithms, the use of different database configurations, and the analysis of alignment-derived metrics. We used thresholds for breadth of genome coverage that balances sensitivity and specificity, but with the primary goal of avoiding false positives. After carrying out all these controls, the method successfully identified potentially pathogenic bacterial species in the desman's fecal samples, including several species known to infect other aquatic mammals and fish, highlighting the microbial risks to which the desman is exposed in its aquatic environment. The developed pipeline proved effective for screening bacterial pathogenic species and could be applied to evaluate pathogen load at broader spatial scales, across populations, and in relation to environmental factors such as water quality, or even individual immune status associated with inbreeding.

These methods represent novel advances in the identification of eukaryotes and bacteria from metagenomic fecal sequences of a wildlife species, with parameters adapted to the characteristics of the data and the specific objectives of this study. They provide a starting point for future improvements and new approaches needed for metagenomic applications in non-model species, an area that remains largely unexplored to date.

3.2 From the Iberian desman to broader conservation: applications of the developed methods

The combined use of fresh fecal samples and shotgun metagenomic sequencing, as applied in this thesis, has demonstrated great potential for the conservation of endangered species. This approach enables the exploration of both the genomic characteristics of the target species, its health and its ecological interactions with the environment, information that is essential for in situ management as well as for planning future conservation strategies, including captive breeding and translocation programs.

Analyses based on endogenous DNA open the possibility of expanding population genomic studies using non-invasive fecal samples. Although these samples must be very fresh and collected shortly after defecation to ensure sufficient DNA quality and quantity, species-specific protocols can be developed to efficiently collect these samples, which is often more feasible than capturing individuals. This would allow, for instance, relatedness analyses to study connectivity between subpopulations, as has been done with the Iberian desman (Escoda et al., 2017, 2019), or to assess the recruitment of new individuals into populations through pedigree analysis.

These analyses can be complemented by examining pathogen and parasite loads at both individual and population levels using fecal samples. Biological factors such as sex or inbreeding and their impact on immune status can be explored in this context. Differences in parasite prevalence or intensity between sexes have been observed in several species, and may be attributed to behavioral or physiological differences (Kołodziej-Sobocińska, 2019). Male-biased parasitism has been linked to the immunosuppressive effects of testosterone and to behavioral patterns during the breeding season that increase exposure and susceptibility to infection, as observed in Pyrenean chamois (Klein, 2004; Martínez-Guijosa et al., 2015; Oliver-Guimerá et al., 2017). Conversely, female-biased parasitism has been reported in European bison, possibly resulting from the formation of larger female groups that facilitate parasite transmission and from immunosuppression associated with pregnancy (Kołodziej-Sobocińska, Demiaszkiewicz, et al., 2016; Kołodziej-Sobocińska, Pyziel, et al., 2016). On the other hand, inbreeding has been associated with reduced fitness and lower immunocompetence, leading to higher pathogen loads and increased susceptibility to infections in wild species (Ewen et al., 2012). For instance, lower genetic diversity in immune-related genomic regions may impair the recognition of pathogens (Smith et al., 2009). In the Iberian desman, differences in genetic diversity at the Major Histocompatibility Complex class I (MHC-I) genes have been observed among individuals (Escoda & Castresana, 2021). This is particularly relevant, as inbred individuals may act as sources of pathogen introduction into the population or serve as reservoirs (Smith et al., 2009). Therefore, exploring the relationship between biological factors or host genetics and pathogen or parasite load using fecal metagenomics could provide valuable insights into how these factors influence disease susceptibility in fragmented and endangered populations.

Due to the risks that infectious and parasitic diseases pose to the conservation of natural populations, there is a growing need among environmental authorities to monitor the health status of wildlife species in situ (ENETWILD-consortium et al., 2023). Numerous well-documented cases have shown how infectious agents have affected threatened populations (Pedersen et al., 2007), such as the feline leukemia virus in the Iberian lynx (Nájera et al., 2024), the Aleutian disease virus in the European mink (Sánchez-Migallón Guzmán et al., 2008), the sarcoptic mange in chamois and mountain goats (Unterköfler et al., 2023), and the chytrid fungus (*Batrachochytrium dendrobatidis*) in various amphibian species (Murray et al., 2009). Moreover, zoonotic pathogens such as *Mycobacterium bovis* pose a risk to the health not only of wildlife but also of livestock and humans (Ali & Alsayeqh, 2022; Justus et al., 2024), highlighting the need for integrated surveillance approaches to understand the role of wildlife species in their transmission (ENETWILD-consortium et al., 2023). Developing metagenomic methods capable of supporting large-scale analyses in the future could enable early detection of disease outbreaks and facilitate the implementation of timely management and conservation measures.

The risk associated with infectious diseases is also well documented in the context of conservation actions involving captive breeding and translocation of individuals (Daszak et al., 2000; Ewen et al., 2012; Sainsbury & Vaughan-Higgins, 2012). The main concern is the potential transfer of pathogens to wild populations that have not been previously exposed, which could result in negative impacts on the entire ecosystem. For example, infection by the parasitic nematode *Ashworthius sidemi* was detected in European bison in a reserve in the Czech Republic following the translocation of individuals from Poland. In the same region, infections were also detected in red deer, suggesting possible local transmission between species (Vadlejch et al., 2017). In addition, reduced exposure to natural antigens and the loss of immunogenetic variation during captive breeding can negatively affect the survival of individuals after reintroduction (Ewen et al., 2012; Kołodziej-Sobocińska, 2019). Therefore, monitoring of infectious diseases should continue in reintroduced individuals as well as in their offspring, to assess survival risks and anticipate potential disease outbreaks that could threaten the entire reintroduced population (Nájera et al., 2021).

Another key aspect in developing effective conservation programs is understanding the trophic ecology of species. Information on diet composition under natural conditions can help to better understand foraging behavior in relation to the composition and availability of food resources, as well as habitat characteristics, and support in situ conservation measures (Esnaola et al. 2021). Understanding the composition of the wild diet is also considered fundamental for ex situ conservation programs, as replicating the natural diet can likely facilitate both the acclimatization of wild-caught individuals to captivity and the successful establishment of captive-bred animals in the wild (Ewen et al., 2012). Moreover, understanding the composition of the wild diet is essential for formulating appropriate captive diets, as suboptimal nutrition can lead to disease, reduced

reproductive success, and changes in the microbiome, among other effects (Crates et al., 2023; Hendier et al., 2021; Jayson et al., 2018; Quiroga-González et al., 2021; Zhang et al., 2017). In addition, monitoring the diet of released individuals in reintroduction programs is crucial to evaluate their adaptation and successful establishment in the wild, especially for those born in captivity (Pödra et al., 2013). Traditional genetic methods, for instance, were unable to detect planarians as part of the Iberian desman's diet, despite their high abundance and prevalence in the population studied in this work. Although further research is needed to determine whether planarians are an important trophic component and are actively consumed in different populations and time periods, this finding highlights the potential of fecal metagenomics to accurately monitor the diet of species, with a view to informing future in situ and ex situ conservation programs..

Overall, the integrative approach developed in this thesis advances the genomic toolkit for the Iberian desman and demonstrates the feasibility and value of non-invasive metagenomic methods for wildlife monitoring and conservation management. Future research should aim to increase sample sizes across regions and seasons and explore additional infectious agents such as viral or eukaryotic pathogens. Such efforts will further strengthen the role of genomics as a highly useful tool in wildlife conservation and management.

VII. CONCLUSIONS

1. The DNA present in fresh fecal samples of the Iberian desman (*Galemys pyrenaicus*) of the Central System population enabled the successful construction of shotgun metagenomic libraries and the analyses of both endogenous and exogenous DNA, demonstrating the potential of fecal samples combined with advanced genomic and bioinformatic tools for the non-invasive genomic monitoring of wild species.
2. Endogenous nuclear DNA was obtained, representing an average of 9% of the total DNA, which allowed for population genomic analyses and the determination of the individuals' sex. Although mitochondrial DNA accounted for a smaller proportion (0.03% on average), it was sufficient to reconstruct complete mitochondrial genomes in all samples.
3. Principal Component Analysis of nuclear DNA allowed for the evaluation of the structure of subpopulations within the Central System population of the Iberian desman, revealing a strong correspondence with their geographical location and identifying connectivity patterns among subpopulations.
4. The recovered mitochondrial genomes enabled phylogenetic analyses of the *G. pyrenaicus* haplotypes, which could be useful at high geographical scales.
5. The exogeneous DNA of the feces was classified into the following groups: Metazoa, Fungi, Viridiplantae, SAR, Other Eukaryota, Bacteria, Archaea, and Virus. The two main sources of exogeneous DNA were bacterial (10%) and metazoan (9%). This highlights the utility of fecal samples for analyzing the trophic ecology, parasitology, and microbiology of the Iberian desman.

- 6. A novel bioinformatic method was developed to use assembled consensus sequences over database sequences instead of individual reads for analyzing exogenous eukaryotic DNA, which allowed precise taxonomic classification using various reference databases. Additionally, it enabled the estimation of relative abundances using depth of coverage values and the verification of assignment quality through depth of coverage across the reference analysis.
- 7. The characterization of diet composition revealed primarily aquatic organisms and groups previously described for *G. pyrenaicus*, such as Trichoptera, Ephemeroptera, Diptera, Coleoptera, and Plecoptera insects. Other elements previously reported in smaller quantities or isolated cases were also identified, including Hemiptera, Odonata, Gastropoda, Oligochaeta, and Salmoniformes. Terrestrial elements reported in lower quantities included Lepidoptera, Hymenoptera, Orthoptera, Araneae, Opiliones, and Isopoda. Additionally, taxa not previously described were detected, such as Bivalvia, Caudata, Gordioida and, notably, Tricladida, the latter showing very high prevalence and abundance in the desmans analyzed in this work. This demonstrates the high sensitivity of this technique compared to previous approaches based on metabarcoding and morphological analysis.
- 8. The parasitological analysis identified both endoparasites and ectoparasites in *G. pyrenaicus*. Coccidia were the most abundant and prevalent, while cases of nematodes, cestodes, and mites were also detected. This study represents the first parasitological analysis of the Iberian desman conducted using metagenomic techniques.
- 9. The species-level analysis of diet demonstrated the efficacy of the developed method in identifying specific organisms but also highlighted its dependency on complete and representative sequence databases for the geographical study area, underscoring the need for increased efforts to sequence reference genomes across the tree of life.
- 10. A novel method was developed for identifying pathogenic bacteria at the species level using reference genomes. The method employed sequence mapping tools and rigorous quality thresholds based on depth of coverage rather than the number of reads for determining species. This represents a significant advance in the study of pathogenic bacteria in wild animals, particularly for unexplored species where detailed genomic information of bacterial pathogens may be lacking.
- 11. Nineteen pathogenic bacterial species were identified in the feces of *G. pyrenaicus*. Many of them have been previously found to infect fish and aquatic mammals, highlighting the role of the aquatic environment in the transmission of diseases.
- 12. Five desmans exhibited a distinct pathogenic load, suggesting a potentially higher vulnerability in them. These results may be highly relevant for conservation programs such as captive breeding and translocations.



Photo credit: © David Pérez (DPC), Wikimedia Commons, Licencia cc-by-sa-4.0.

VIII.

REFERENCES

- Abebe, E., Decraemer, W., & De Ley, P. (2008). Global diversity of nematodes (Nematoda) in freshwater. *Hydrobiologia*, 595(1), 67–78.
<https://doi.org/10.1007/s10750-007-9005-5>
- Akoijam, N., & Joshi, S. R. (2022). Conservation Metagenomics: Understanding Microbiomes for Biodiversity Sustenance and Conservation. In A. Kumar, B. Choudhury, S. Dayanandan, & M. L. Khan (Eds.), *Molecular Genetics and Genomics Tools in Biodiversity Conservation*. Springer, Singapore.
https://doi.org/https://doi.org/10.1007/978-981-16-6005-4_3
- Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., & Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources* (Vol. 19, Issue 2, pp. 327–348).
<https://doi.org/10.1111/1755-0998.12960>
- Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, 9(1), 134–147.
<https://doi.org/10.1111/2041-210X.12849>
- Ali, S., & Alsayeqh, A. F. (2022). Review of major meat-borne zoonotic bacterial pathogens. *Frontiers in Public Health*, 10.
<https://doi.org/https://doi.org/10.3389/fpubh.2022.1045599>

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11), 1144–1146. <https://doi.org/10.1038/nmeth.3103>

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* (Vol. 25, Issue 17). Oxford University Press. <https://academic.oup.com/nar/article/25/17/3389/1061651>

Altschul, Stephen. 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. [https://doi.org/doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/doi.org/10.1016/S0022-2836(05)80360-2)

Alvarez, F., Gijón-Botella, H., Quinteiro, P., Rey, J., López-Román, F., & Sanmartín, M. L. (1994). *Paracuaría hispanica* n. sp. (Nematoda: Acuariidae), a stomach parasite of the Iberian desman *Galemys pyrenaicus* Geoffr. (Insectivora: Talpidae), with a redefinition of the genus *Paracuaría* Rao, 1951. *Systematic Parasitology*, 29, 105–112.

Alvarez Mascato, M. F. (1990). Contribución al estudio de la helmintofauna de mamíferos en la Comunidad Autónoma Gallega. Universidad de Santiago de Compostela.

Alvarez-Estape, M., Pawar, H., Fontseré, C., Trujillo, A. E., Gunson, J. L., Bergl, R. A., Bermejo, M., Linder, J. M., McFarland, K., Oates, J. F., Sunderland-Groves, J. L., Orkin, J., Higham, J. P., Viaud-Martinez, K. A., Lizano, E., & Marques-Bonet, T. (2023). Past Connectivity but Recent Inbreeding in Cross River Gorillas Determined Using Whole Genomes from Single Hairs. *Genes*, 14(3). <https://doi.org/10.3390/genes14030743>

Andrews, K. R., De Barba, M., Russello, M. A., & Waits, L. P. (2018). Advances in Using Non-invasive, Archival, and Environmental Samples for Population Genomic Studies. In P. A. Hohenlohe & O. P. Rajora (Eds.), *Population Genomics: Wildlife*. (pp. 63–99). Springer, Cham. https://doi.org/10.1007/13836_2018_45

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* (Vol. 17, Issue 2, pp. 81–92). Nature Publishing Group. <https://doi.org/10.1038/nrg.2015.28>

Ang, A., Roesma, D. I., Nijman, V., Meier, R., Srivathsan, A., & Rizaldi. (2020). Faecal DNA to the rescue: Shotgun sequencing of non-invasive samples reveals two subspecies of Southeast Asian primates to be Critically Endangered species. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-66007-8>

Antil, S., Abraham, J. S., Sripoorna, S., Maurya, S., Dagar, J., Makhija, S., Bhagat, P., Gupta, R., Sood, U., Lal, R., & Toteja, R. (2023). DNA barcoding, an effective tool for species identification: a review. *Molecular Biology Reports* (Vol. 50, Issue 1, pp. 761–775). Springer Science and Business Media B.V. <https://doi.org/10.1007/s11033-022-08015-7>

Ardura, A., Linde, A. R., & Garcia-Vazquez, E. (2013). Genetic detection of *Pseudomonas* spp. in commercial Amazonian fish. *International Journal of Environmental Research and Public Health*, 10(9), 3954–3966. <https://doi.org/10.3390/ijerph10093954>

Austin, B., & Austin, D. A. (2016). Bacterial Fish Pathogens: Disease of Farmed and Wild Fish (Sixth edition). *Switzerland: Springer*. <https://doi.org/https://doi.org/10.1007/978-3-319-32674-0>

Aymerich, P., & Gosálvez, J. (2002). Factors de distribució de *Galemys pyrenaicus* (Insectivora, Talpidae) a Catalunya. *Orsis*, 17, 21–35.

Baeza, A. J. (2022). Mitochondrial genomes assembled from non-invasive eDNA metagenomic scat samples in the endangered Amur tiger *Panthera tigris altaica*. *PeerJ*, 10. <https://doi.org/10.7717/peerj.14428>

Bağcı, C., Patz, S., & Huson, D. H. (2021). DIAMOND+MEGAN: Fast and Easy Taxonomic and Functional Analysis of Short and Long Microbiome Sequences. *Current Protocols*, 1(3). <https://doi.org/10.1002/cpz1.59>

Bai, J., Wang, X., Zhang, Z., Lian, P., & Qiao, J. (2023). Overview of Mink Immunity and Resistance to *Pseudomonas aeruginosa*. In *Veterinary Medicine International* (Vol. 2023). Hindawi Limited. <https://doi.org/10.1155/2023/6158844>

Balmori-de la Puente, A., Nores, C., Román, J., Fernández-González, A., Aymerich, P., Gosálbez, J., Escoda, L., & Castresana, J. (2019). Size increase without genetic divergence in the Eurasian water shrew *Neomys fodiens*. *Scientific Reports*, 9(1). <https://doi.org/10.1038/s41598-019-53891-y>

Balmori-de la Puente, A., Escoda, L., Fernández-González, Á., Menéndez-Pérez, D., González-Esteban, J., & Castresana, J. (2023). Evaluating the use of non-invasive hair sampling and ddRAD to characterize populations of endangered species: Application to a peripheral population of the European mink. *Ecology and Evolution*, 13(9). <https://doi.org/10.1002/ece3.10530>

Balvočiute, M., & Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare?. *BMC Genomics*, 18. <https://doi.org/10.1186/s12864-017-3501-4>

Bangoura, B., & Dauschies, A. (2018). *Eimeria*. In M. Florin-Christensen & L. Schnittger (Eds.), *Parasitic Protozoa of Farm Animals and Pets* (pp. 55–101). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-70132-5_3

Barandika, J. F., Hurtado, A., García-Esteban, C., Gil, H., Escudero, R., Barral, M., Jado, I., Juste, R. A., Anda, P., & García-Pérez, A. L. (2007). Tick-borne zoonotic bacteria in wild and domestic small mammals in Northern Spain. *Applied and Environmental Microbiology*, 73(19), 6166–6171. <https://doi.org/10.1128/AEM.00590-07>

Beja-Pereira, A., Oliveira, R., Alves, P. C., Schwartz, M. K., & Luikart, G. (2009). Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular Ecology Resources* (Vol. 9, Issue 5, pp. 1279–1301). <https://doi.org/10.1111/j.1755-0998.2009.02699.x>

Bengtsson, C. F., Olsen, M. E., Brandt, L. Ø., Bertelsen, M. F., Willerslev, E., Tobin, D. J., Wilson, A. S., & Gilbert, M. T. P. (2012). DNA from keratinous tissue. Part I: Hair and nail. *Annals of Anatomy*, 194(1), 17–25. <https://doi.org/10.1016/j.aanat.2011.03.013>

Berná, L., Rego, N., & Francia, M. E. (2021). The Elusive Mitochondrial Genomes of Apicomplexa: Where Are We Now? In *Frontiers in Microbiology* (Vol. 12). Frontiers Media S.A. <https://doi.org/10.3389/fmicb.2021.751775>

Bernatchez, L., Ferchaud, A. L., Berger, C. S., Venney, C. J., & Xuereb, A. (2024). Genomics for monitoring and understanding species responses to global climate change. In *Nature Reviews Genetics* (Vol. 25, Issue 3, pp. 165–183). Nature Research. <https://doi.org/10.1038/s41576-023-00657-y>

Bertola, L. D., Brüniche-Olsen, A., Kershaw, F., Russo, I. M., MacDonald, A. J., Sunnucks, P., Bruford, M. W., Cadena, C. D., Ewart, K. M., de Bruyn, M., Eldridge, M. D. B., Frankham, R., Guayasamin, J. M., Grueber, C. E., Hoareau, T. B., Hoban, S., Hohenlohe, P. A., Hunter, M. E., Kotze, A., ... Segelbacher, G. (2023). A pragmatic approach for integrating molecular tools into biodiversity conservation. *Conservation Science and Practice*. <https://doi.org/10.1111/csp2.13053>

Biffi, M., Gillet, F., Laffaille, P., Colas, F., Aulagnier, S., Blanc, F., Galan, M., Tiouchichine, M. L., Némot, M., Buisson, L., & Michaux, J. R. (2017). Novel insights into the diet of the Iberian desman (*Galemys pyrenaicus*) using next-generation sequencing molecular analyses. *Journal of Mammalogy*, 98(5), 1497–1507. <https://doi.org/10.1093/jmammal/gyx070>

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., ... Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, 41(11), 1633–1644. <https://doi.org/10.1038/s41587-023-01688-w>

Bradley, I. M., Pinto, A. J., & Guest, J. S. (2016). Design and evaluation of illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Applied and Environmental Microbiology*, 82(19), 5878–5891. <https://doi.org/10.1128/AEM.01630-16>

Breed, M. F., Harrison, P. A., Blyth, C., Byrne, M., Gaget, V., Gellie, N. J. C., Groom, S. V. C., Hodgson, R., Mills, J. G., Prowse, T. A. A., Steane, D. A., & Mohr, J. J. (2019). The potential of genomics for restoring ecosystems and biodiversity. In *Nature Reviews Genetics* (Vol. 20, Issue 10, pp. 615–628). Nature Publishing Group. <https://doi.org/10.1038/s41576-019-0152-0>

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20(4), 1125–1139. <https://doi.org/10.1093/bib/bbx120>

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. In *Nature Methods* (Vol. 12, Issue 1, pp. 59–60). Nature Publishing Group. <https://doi.org/10.1038/nmeth.3176>

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16s rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, 6. <https://doi.org/10.1038/sdata.2019.7>

Burriel-Carranza, B., Tejero-Cicuéndez, H., Carné, A., Mochales-Riaño, G., Talavera, A., Al Saadi, S., Els, J., Šmíd, J., Tamar, K., Tarroso, P., & Carranza, S. (2024). Integrating genomics and biogeography to unravel the origin of a mountain biota: The case of a reptile endemism hotspot in Arabia. *Systematic Biology*. <https://doi.org/https://doi.org/10.1093/sysbio/syae032>

Cabral, J. P. S. (2010). Water microbiology. Bacterial pathogens and water. *International Journal of Environmental Research and Public Health* (Vol. 7, Issue 10, pp. 3657–3703). <https://doi.org/10.3390/ijerph7103657>

Cabrera, A. A., Rey-Iglesia, A., Louis, M., Skovrind, M., Westbury, M. V., & Lorenzen, E. D. (2022). How low can you go? Introducing SeXY: sex identification from low-quantity sequencing data despite lacking assembled sex chromosomes. *Ecology and Evolution*, 12(8).
<https://doi.org/10.1002/ece3.9185>

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10.
<https://doi.org/10.1186/1471-2105-10-421>

Cano-Terriza, D., Guerra, R., Mozos, E., Rodríguez-Sánchez, B., Borge, C., & García-Bocanegra, I. (2017). Fatal *acinetobacter baumannii* infection in the critically endangered european mink (*mustela lutreola*). *Journal of Zoo and Wildlife Medicine*, 48(1), 220–223.
<https://doi.org/10.1638/2016-0082.1>

Cantas, L., & Suer, K. (2014). Review: The important bacterial zoonoses in "One Health" concept. *Frontiers in Public Health* (Vol. 2, Issue OCT). Frontiers Media S. A.
<https://doi.org/10.3389/fpubh.2014.00144>

Carroll, E. L., Bruford, M. W., DeWoody, J. A., Leroy, G., Strand, A., Waits, L., & Wang, J. (2018). Genetic and genomic monitoring with minimally invasive sampling methods. *Evolutionary Applications*, 11(7), 1094–1119.
<https://doi.org/10.1111/eva.12600>

Castién, E., & Gosálbez, J. (1995). Diet of *Galemys pyrenaicus* (Geoffroy, 1811) in the North of the Iberian Peninsula. *Netherlands Journal of Zoology*, 45, 422–430.

Castién, E., & Gosálbez, J. (1999). Habitat and food preferences in a guild of insectivorous mammals in the Western Pyrenees. *Acta Theriologica*, 44, 1–13.

Charbonnel, A., Buisson, L., Biffi, M., D'Amico, F., Besnard, A., Aulagnier, S., Blanc, F., Gillet, F., Lacaze, V., Michaux, J. R., Némóz, M., Pagé, C., Sanchez-Perez, J. M., Sauvage, S., & Laffaille, P. (2015). Integrating hydrological features and genetically validated occurrence data in occupancy modelling of an endemic and endangered semi-aquatic mammal, *Galemys pyrenaicus*, in a Pyrenean catchment. *Biological Conservation*, 184, 182–192.
<https://doi.org/10.1016/j.biocon.2015.01.019>

Chattaway, M. A., Schaefer, U., Tewolde, R., Dallman, T. J., & Jenkins, C. (2017). Identification of *Escherichia coli* and *shigella* species from whole-genome sequences. *Journal of Clinical Microbiology*, 55(2), 616–623.
<https://doi.org/10.1128/JCM.01790-16>

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *IMeta*, 2(2).
<https://doi.org/10.1002/imt2.107>

Chiou, K. L., & Bergey, C. M. (2018). Methylation-based enrichment facilitates low-cost, noninvasive genomic scale sequencing of populations from feces. *Scientific Reports*, 8(1).
<https://doi.org/10.1038/s41598-018-20427-9>

Chorlton, S. D. (2024). Ten common issues with reference sequence databases and how to mitigate them. In *Frontiers in Bioinformatics* (Vol. 4). Frontiers Media SA.
<https://doi.org/10.3389/fbinf.2024.1278228>

Chua, P. Y. S., Crampton-Platt, A., Lammers, Y., Alsos, I. G., Boessenkool, S., & Bohmann, K. (2021). Metagenomics: A viable tool for reconstructing herbivore diet. *Molecular Ecology Resources*, 21(7), 2249–2263.
<https://doi.org/10.1111/1755-0998.13425>

Cooper, P., Landrum, M., Mizrachi, I., & et al. (2010). Entrez Sequences Quick Start. In *Entrez Sequences Help* (Updated 2024 Jul 5). National Center for Biotechnology Information (US).
<https://www.ncbi.nlm.nih.gov/books/NBK44863/>

Cordero del Campillo, M., Castañón Ordóñez, L., & Reguera Feo, A. (1994). Índice - catálogo de zooparásitos ibéricos. (2nd ed.). Secretariado de publicaciones, Universidad de León.

Corominas, M., Marquès-Bonet, T., Arnedo, M. A., Bayés, M., Belmonte, J., Escrivà, H., Fernández, R., Gabaldón, T., Garnatje, T., Germain, J., Niell, M., Palero, F., Pons, J., Puigdomènech, P., Cuevas-Caballé, C., Obiol, J. F., Gut, I., Gut, M., Hidalgo, O., ... Guigó, R. (2024). The Catalan initiative for the Earth BioGenome Project: contributing local data to global biodiversity genomics. *NAR Genomics and Bioinformatics*, 6(3).
<https://doi.org/10.1093/nargab/lqae075>

Cox, E., Tsuchiya, M. T. N., Ciufu, S., Torcivia, J., Falk, R., Anderson, W. R., Holmes, J. B., Hem, V., Breen, L., Davis, E., Ketter, A., Zhang, P., Soussov, V., Schoch, C. L., & O'Leary, N. A. (2024). NCBI Taxonomy: enhanced access via NCBI Datasets. *Nucleic Acids Research*.
<https://doi.org/10.1093/nar/gkae967>

Crampton-Platt, A., Yu, D. W., Zhou, X., & Vogler, A. P. (2016). Mitochondrial metagenomics: Letting the genes out of the bottle. In *GigaScience* (Vol. 5, Issue 1). BioMed Central Ltd.
<https://doi.org/10.1186/s13742-016-0120-y>

Crates, R., Stojanovic, D., & Heinsohn, R. (2023). The phenotypic costs of captivity. *Biological Reviews*, 98(2), 434–449.
<https://doi.org/10.1111/brv.12913>

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* (Vol. 7, Issue 9, pp. 1008–1018). British Ecological Society.
<https://doi.org/10.1111/2041-210X.12574>

Curto, M., Veríssimo, A., Riccioni, G., Santos, C. D., Ribeiro, F., Jentoft, S., Alves, M. J., & Gante, H. F. (2024). Improving whole biodiversity monitoring and discovery with environmental DNA metagenomics. *BioRxiv*.
<https://doi.org/10.1101/2024.12.17.628643>

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2).
<https://doi.org/10.1093/gigascience/giab008>

Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2000). Emerging Infectious Diseases of Wildlife-Threats to Biodiversity and Human Health. *Science*, 287, 443–449.
<https://doi.org/10.1126/science.287.5452.443>

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* (Vol. 12, Issue 7, pp. 499–510).
<https://doi.org/10.1038/nrg3012>

Davison, H. R., Pilgrim, J., Wybouw, N., Parker, J., Pirro, S., Hunter-Barnett, S., Campbell, P. M., Blow, F., Darby, A. C., Hurst, G. D. D., & Siozios, S. (2022). Genomic diversity across the *Rickettsia* and 'Candidatus Megaira' genera and proposal of genus status for the Torix group. *Nature Communications*, 13(1).
<https://doi.org/10.1038/s41467-022-30385-6>

de Flamingh, A., Ishida, Y., Pečnerová, P., Vilchis, S., Siegismund, H. R., van Aarde, R. J., Malhi, R. S., & Roca, A. L. (2023). Combining methods for non-invasive fecal DNA enables whole genome and metagenomic analyses in wildlife biology. *Frontiers in Genetics*, 13.
<https://doi.org/10.3389/fgene.2022.1021004>

Deagle, B. E., Chiaradia, A., McInnes, J., & Jarman, S. N. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: Is what goes in what comes out? *Conservation Genetics*, 11(5), 2039–2048.
<https://doi.org/10.1007/s10592-010-0096-6>

Deblock, S., & Combes, C. (1965). Contribution à l'étude des Microphallidae, Tranassos, 1920 (Trematoda). X. *Maritrema pyrenaica* n. sp., parasite du Desman (Mammifère Insectivore). Essai de clé diagnostique des espèces du genre. *Bulletin de La Société Zoologique de France*, 90, 101–117.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. In *Molecular Ecology* (Vol. 26, Issue 21, pp. 5872–5895). Blackwell Publishing Ltd.
<https://doi.org/10.1111/mec.14350>

Destoumieux-Garzón, D., Mavingui, P., Boetsch, G., Boissier, J., Darriet, F., Duboz, P., Fritsch, C., Giraudoux, P., Roux, F. Le, Morand, S., Paillard, C., Pontier, D., Sueur, C., & Voituren, Y. (2018). The one health concept: 10 years old and a long road ahead. In *Frontiers in Veterinary Science* (Vol. 5, Issue FEB). Frontiers Media S.A.
<https://doi.org/10.3389/fvets.2018.00014>

Diamanka, A., Loch, T. P., Cipriano, R. C., Winters, A. D., & Faisal, M. (2014). Infection of sea lamprey with an unusual strain of *Aeromonas salmonicida*. *Journal of Wildlife Diseases*, 50(2), 159–170.
<https://doi.org/10.7589/2013-01-026>

Díaz, S., Demissew, S., Carabias, J., Joly, C., Lonsdale, M., Ash, N., Larigauderie, A., Adhikari, J. R., Arico, S., Báldi, A., Bartuska, A., Baste, I. A., Bilgin, A., Brondizio, E., Chan, K. M. A., Figueroa, V. E., Duraiappah, A., Fischer, M., Hill, R., ... Zlatanova, D. (2015). The IPBES Conceptual Framework - connecting nature and people. *Current Opinion in Environmental Sustainability*, 14, 1–16.
<https://doi.org/10.1016/j.cosust.2014.11.002>

Dilthey, A. T., Jain, C., Koren, S., & Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature Communications*, 10(1).
<https://doi.org/10.1038/s41467-019-10934-2>

Dorris, M., De Ley, P., & Blaxter, M. L. (1999). Molecular Analysis of Nematode Diversity and the Evolution of Parasitism. *Parasitology Today*, 15(5), 188–193.

Duszynski, D. W. (2021). Biodiversity of the Coccidia (Apicomplexa: Conoidasida) in vertebrates: what we know, what we do not know, and what needs to be done. *Folia Parasitologica*, 66, 1–18.
<https://doi.org/10.14411/FP.2021.001>

Duszynski, D. W., & Upton, S. J. (2000). Coccidia (Apicomplexa: Eimeriidae) of the Mammalian Order Insectivora.
<https://digitalcommons.unl.edu/parasitologyfacpubs/196>

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10).
<https://doi.org/10.1371/journal.pcbi.1002195>

Egloff, C., Labrosse, A., Hebert, C., & Crump, D. (2009). A nondestructive method for obtaining maternal DNA from avian eggshells and its application to embryonic viability determination in herring gulls (*Larus argentatus*). *Molecular Ecology Resources*, 9(1), 19–27. <https://doi.org/10.1111/j.1755-0998.2008.02214.x>

Elyasigorji, Z., Izadpanah, M., Hadi, F., & Zare, M. (2023). Mitochondrial genes as strong molecular markers for species identification. *Nucleus (India)* (Vol. 66, Issue 1, pp. 81–93). Springer. <https://doi.org/10.1007/s13237-022-00393-4>

ENETWILD-consortium, Queirós, J., Caballero, J., Blanco-Aguir, J. A., Bocanegra, I., Torres, M. J., Acevedo, P., Guerrasio, T., Apollonio, M., Berdión, O., Carro, F., Casaer, J., Carvalho, J., Csányi, S., Ferroglio, E., Fonseca, C., Gačić, D., Gavier-Widen, D., Galán, V. G., ... Vicente, J. (2023). A pilot on integrated wildlife monitoring at European scale: environmental detection of selected pathogens in the European Observatory of Wildlife. *EFSA Supporting Publications*, 20(10). <https://doi.org/10.2903/sp.efsa.2023.EN-8241>

Epps, C. W., Palsbøll, P. J., Wehausen, J. D., Roderick, G. K., Ramey, R. R., & McCullough, D. R. (2005). Highways block gene flow and cause a rapid decline in genetic diversity of desert bighorn sheep. *Ecology Letters*, 8(10), 1029–1038. <https://doi.org/10.1111/j.1461-0248.2005.00804.x>

Escoda, L., & Castresana, J. (2021). The genome of the Iberian desman and the effects of bottlenecks and inbreeding on the genomic landscape of an endangered species. *Evolutionary Applications*, 14(7), 1898–1913. <https://doi.org/10.1111/eva.13249>

Escoda, L., Fernández-González, Á., & Castresana, J. (2019). Quantitative analysis of connectivity in populations of a semi-aquatic mammal using kinship categories and network assortativity. *Molecular Ecology Resources*, 19(2), 310–326. <https://doi.org/10.1111/1755-0998.12967>

Escoda, L., González-Esteban, J., Gómez, A., & Castresana, J. (2017). Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal *Galemys pyrenaicus*. *Molecular Ecology*, 26(13), 3343–3357. <https://doi.org/10.1111/mec.14133>

Esnaola, A., Arrizabalaga-Escudero, A., González-Esteban, J., Elosegi, A., & Aihartza, J. (2018). Determining diet from faeces: Selection of metabarcoding primers for the insectivore Iberian desman (*Galemys pyrenaicus*). *PLoS ONE*, 13(12). <https://doi.org/10.1371/journal.pone.0208986>

Esnaola, A., González-Esteban, J., Elosegi, A., Arrizabalaga-Escudero, A., & Aihartza, J. (2018). Need for speed: Preference for fast-flowing water by the endangered semi-aquatic Iberian desman (*Galemys pyrenaicus*) in two contrasting streams. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 28(3), 600–609. <https://doi.org/10.1002/aqc.2893>

Esnaola, A., Mata, V., Arrizabalaga-Escudero, A., González-Esteban, J., Larrañaga, A., Rebelo, H., Elosegi, A., & Aihartza, J. (2021). It is the ambience, not the menu. Prey availability does not drive habitat selection by the endangered Iberian desman. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31(7), 1859–1872. <https://doi.org/10.1002/aqc.3563>

Ewen, J. G., Armstrong, D. P., Parker, K. A., & Seddon, P. J. (2012). *Reintroduction Biology: Integrating Science and Management*. Wiley Blackwell. <https://doi.org/10.1002/9781444355833.fmatter>

Feehery, G. R., Yigit, E., Oyola, S. O., Langhorst, B. W., Schmidt, V. T., Stewart, F. J., Dimalanta, E. T., Amaral-Zettler, L. A., Davis, T., Quail, M. A., & Pradhan, S. (2013). A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA. *PLoS ONE*, 8(10). <https://doi.org/10.1371/journal.pone.0076096>

Feng, X., Cheng, H., Portik, D., & Li, H. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, 19(6), 671–674. <https://doi.org/10.1038/s41592-022-01478-3>

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., ... Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology and Evolution* (Vol. 37, Issue 3, pp. 197–202). <https://doi.org/10.1016/j.tree.2021.11.008>

Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q., Schaalje, G. B., Clement, M. J., Crandall, K. A., & Johnson, W. E. (2013). Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Research*, 23(10), 1721–1729. <https://doi.org/10.1101/gr.150151.112>

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology* (Vol. 26, Issue 20, pp. 5369–5406). Blackwell Publishing Ltd. <https://doi.org/10.1111/mec.14264>

Gaywood, M. J., Ewen, J. G., Hollingsworth, P. M., & Moehrensclager A. (Eds.). (2022). *Conservation Translocations*. Cambridge University Press.
<https://doi.org/https://doi.org/10.1017/9781108638142>

Gibson, K. M., Nguyen, B. N., Neumann, L. M., Miller, M., Buss, P., Daniels, S., Ahn, M. J., Crandall, K. A., & Pukazhenth, B. (2019). Gut microbiome differences between wild and captive black rhinoceros – implications for rhino health. *Scientific Reports*, 9(1).
<https://doi.org/10.1038/s41598-019-43875-3>

Gillet, F., Le Roux, B., Blanc, F., Bodo, A., Fournier-Chambrillon, C., Fournier, P., Jacob, F., Lacaze, V., Némoz, M., Aulagnier, S., & Michaux, J. R. (2016). Genetic monitoring of the endangered Iberian desman (*Galemys pyrenaicus*) in the Aude River, France. *Belgian Journal of Zoology*, 146(1), 44–52.
<https://doi.org/https://doi.org/10.26496/bjz.2016.38>

Gillet, F., Tiouchichine, M. L., Galan, M., Blanc, F., Némoz, M., Aulagnier, S., & Michaux, J. R. (2015). A new method to identify the endangered Iberian desman (*Galemys pyrenaicus*) and to study its diet, using next generation sequencing from faeces. *Mammalian Biology*, 80(6), 505–509.
<https://doi.org/10.1016/j.mambio.2015.08.002>

Gisbert, J., & García-Perea, R. (2014). History of the decline of the Iberian desman *Galemys pyrenaicus* (É. Geoffroy Saint-Hilaire, 1811) in the Central System (Iberian Peninsula). *Munibe Monographs. Nature Series*, 3, 19–35.

Gisbert, J., García-Perea, R., Fernández-González, Á., & Quaglietta, L. (2024). El desmán ibérico (*Galemys pyrenaicus*): Revisión del estado de conservación. Málaga, España: UICN.

Gissi, C., Iannelli, F., & Pesole, G. (2008). Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. In *Heredity* (Vol. 101, Issue 4, pp. 301–320).
<https://doi.org/10.1038/hdy.2008.62>

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D.-H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., ... Murphy, T. D. (2024). NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*.
<https://doi.org/10.1093/nar/gkae1038>

González-Esteban, J., Villate, I., & Castién González-Esteban, E. (2003). Sexual identification of *Galemys pyrenaicus*. *Acta Theriologica*, 48(4), 571–573.

Gu, W., Deng, X., Lee, M., Sucu, Y. D., Arevalo, S., Stryke, D., Federman, S., Gopez, A., Reyes, K., Zorn, K., Sample, H., Yu, G., Ishpuniani, G., Briggs, B., Chow, E. D., Berger, A., Wilson, M. R., Wang, C., Hsu, E., ... Chiu, C. Y. (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nature Medicine*, 27(1), 115–124.
<https://doi.org/10.1038/s41591-020-1105-z>

Gu, W., Miller, S., & Chiu, C. Y. (2019). Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. *Annual Review of Pathology: Mechanisms of Disease*, 14, 319–338.
<https://doi.org/https://doi.org/10.1146/annurev-pathmechdis-012418-012751>

Guz, L., Nowakiewicz, A., Puk, K., Zięba, P., Gnat, S., & Matuszewski, Ł. (2021). Virulence and antimicrobial resistance pattern of *aeromonas spp.* Colonizing european pond turtles emys orbicularis and their natural environment. first study from poland. *Animals*, 11(10).
<https://doi.org/10.3390/ani1102772>

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E. M., & Troedsson, C. (2014). Characterization of the 18s rRNA gene for designing universal eukaryote specific primers. *PLoS ONE*, 9(2).
<https://doi.org/10.1371/journal.pone.0087624>

Hall, G. A., Speed, T. P., & Woodruff, C. J. (2020). Strain-level sample characterisation using long reads and MAPQ scores. *BioRxiv*, 2020.10.18.344739.
<https://doi.org/10.1101/2020.10.18.344739>

Hammerl, J. A., Ulrich, R. G., Imholt, C., Scholz, H. C., Jacob, J., Kratzmann, N., Nöckler, K., & Al Dahouk, S. (2017). Molecular Survey on Brucellosis in Rodents and Shrews – Natural Reservoirs of Novel *Brucella* Species in Germany? *Transboundary and Emerging Diseases*, 64(2), 663–671.
<https://doi.org/10.1111/tbed.12425>

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and molecular biology reviews*, 68(4), 669–685.
<https://doi.org/10.1128/MBR.68.4.669-685.2004>

Hendier, A., Chatelain, C., Du Pasquier, P. E., Paris, M., Ouattara, K., Koné, I., Croll, D., & Zuberbühler, K. (2021). A new method to determine the diet of pygmy hippopotamus in Taï National Park, Côte d'Ivoire. *African Journal of Ecology*, 59(4), 809–825.
<https://doi.org/10.1111/aje.12888>

Haring, V., Jacob, J., Walther, B., Trost, M., Stubbe, M., Mertens-Scholz, K., Melzer, F., Scuda, N., Gentil, M., Sixl, W., Schäfer, T., Stanko, M., Wolf, R., Pfeffer, M., Ulrich, R. G., & Obiegala, A. (2023). White-Toothed Shrews (Genus *Crocidura*): Potential Reservoirs for Zoonotic *Leptospira* spp. and Arthropod-Borne Pathogens? *Pathogens*, 12(6).
<https://doi.org/10.3390/pathogens12060781>

Hawltischek, O., Fernández-González, A., Balmori-de la Puente, A., & Castresana, J. (2018). A pipeline for metabarcoding and diet analysis from fecal samples developed for a small semi-aquatic mammal. *PLoS ONE*, 13(8).
<https://doi.org/10.1371/journal.pone.0201763>

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270(1512), 313–321.
<https://doi.org/10.1098/rspb.2002.2218>

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular Ecology*, 30(1), 62–82.
<https://doi.org/10.1111/mec.15720>

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., Crandall, K. A., & Johnson, W. E. (2014). PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, 2(1).
<https://doi.org/10.1186/2049-2618-2-33>

Höss, M., Kohn, M., Pääbo, S., Knauer, F., & Schröder, W. (1992). Excrement analysis by PCR. *Nature*, 359, 199.
<https://doi.org/https://doi.org/10.1038/359199a0>

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3), 377–386.
<https://doi.org/10.1101/gr.5969107>

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., Fitzgerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., ... White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207–214.
<https://doi.org/10.1038/nature11234>

Igea, J., Aymerich, P., Fernández-González, A., González-Esteban, J., Gómez, A., Alonso, R., Gosálbez, J., & Castresana, J. (2013). Phylogeography and postglacial expansion of the endangered semi-aquatic mammal *Galemys pyrenaicus*. *BMC Evolutionary Biology*, 13(1).
<https://doi.org/10.1186/1471-2148-13-115>

Jakovlić, I., Zou, H., Ye, T., Zhang, H., Liu, X., Xiang, C. Y., Wang, G. T., & Zhang, D. (2023). Mitogenomic evolutionary rates in bilateria are influenced by parasitic lifestyle and locomotory capacity. *Nature Communications*, 14(1).
<https://doi.org/10.1038/s41467-023-42095-8>

Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., Van Immerseel, F., Vandewynckel, Y. P., De Tré, G., & De Spiegeleer, B. (2018). Disbiome database: Linking the microbiome to disease. *BMC Microbiology*, 18(1).
<https://doi.org/10.1186/s12866-018-1197-5>

Jayson, S., Ferguson, A., Goetz, M., Routh, A., Tapley, B., Harding, L., Michaels, C. J., & Dawson, J. (2018). Comparison of the nutritional content of the captive and wild diets of the critically endangered mountain chicken frog (*Leptodactylus fallax*) to improve its captive husbandry. *Zoo Biology*, 37(5), 332–346.
<https://doi.org/10.1002/zoo.21442>

Johri, S., Doane, M. P., Allen, L., & Dinsdale, E. A. (2019). Taking advantage of the genomics revolution for monitoring and conservation of chondrichthyan populations. *Diversity*, 11(4).
<https://doi.org/10.3390/d11040049>

Jourdane, J. (1979). Le cycle biologique de *Maritrema pyrenaica* Deblock et Combes, 1965, parasite de micromammifères insectivores pyrénéens. *Annales de Parasitologie Humaine et Comparée*, 54, 449–456.
<https://doi.org/10.1051/parasite/1979544449>

Justus, W., Valle, S., Barton, O., Gresham, A., & Shannon, G. (2024). A review of bovine tuberculosis transmission risk in European wildlife communities. *Mammal Review*, 54(3), 325–340).
<https://doi.org/10.1111/mam.12347>

Kang, D., Choi, J. H., Kim, M., Yun, S., Oh, S., Yi, M. H., Yong, T. S., Lee, Y. A., Shin, M. H., & Kim, J. Y. (2024). Optimization of 18 S rRNA metabarcoding for the simultaneous diagnosis of intestinal parasites. *Scientific Reports*, 14(1).
<https://doi.org/10.1038/s41598-024-76304-1>

Kans, J. (2013). Entrez Direct: E-utilities on the Unix Command Line. In *Entrez Programming Utilities Help* (Updated 2024 Oct 21). National Center for Biotechnology Information (US).
<https://www.ncbi.nlm.nih.gov/books/NBK179288/>

Karbowiak, G., Stanko, M., Smahol, K., Werszko, J., & Rychlik, L. (2023). Parasitic Arthropods of Soricinae Shrews in North-Eastern Poland. *Animals*, 13(18). <https://doi.org/10.3390/ani13182960>

Khan, S., Nabi, G., Ullah, M. W., Yousaf, M., Manan, S., Siddique, R., & Hou, H. (2016). Overview on the Role of Advance Genomics in Conservation Biology of Endangered Species. *International Journal of Genomics* (Vol. 2016). Hindawi Limited. <https://doi.org/10.1155/2016/3460416>

Klein, S. L. (2004). Hormonal and immunological mechanisms mediating sex differences in parasite infection. *Parasite Immunology*, 26(6–7), 247–264. <https://doi.org/10.1111/j.0141-9838.2004.00710.x>

Klymus, K. E., Marshall, N. T., & Stepien, C. A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLoS ONE*, 12(5). <https://doi.org/10.1371/journal.pone.0177643>

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11), 1103–1110. <https://doi.org/10.1038/s41592-020-00971-x>

Kołodziej-Sobocińska, M. (2019). Factors affecting the spread of parasites in populations of wild European terrestrial mammals. *Mammal Research*, 64(3), 301–318. <https://doi.org/10.1007/s13364-019-00423-8>

Kołodziej-Sobocińska, M., Demiaszkiewicz, A. W., Lachowicz, J., Borowik, T., & Kowalczyk, R. (2016). Influence of management and biological factors on the parasitic invasions in the wild-spread of blood-sucking nematode *Ashworthius sidemi* in European bison (*Bison bonasus*). *International Journal for Parasitology: Parasites and Wildlife*, 5(3), 286–294. <https://doi.org/10.1016/j.ijppaw.2016.09.005>

Kołodziej-Sobocińska, M., Pyziel, A. M., Demiaszkiewicz, A. W., Borowik, T., & Kowalczyk, R. (2016). Pattern of parasite egg shedding by European bison (*Bison bonasus*) in the Białowieża Primeval Forest, Poland. *Mammal Research*, 61(3), 179–186. <https://doi.org/10.1007/s13364-016-0270-4>

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). *ANGSD: Analysis of Next Generation Sequencing Data*. <http://www.biomedcentral.com/1471-2105/15/356>

Kounosu, A., Murase, K., Yoshida, A., Maruyama, H., & Kikuchi, T. (2019). Improved 18S and 28S rDNA primer sets for NGS-based parasite detection. *Scientific Reports*, 9(1). <https://doi.org/10.1038/s41598-019-52422-z>

Kraus, R. H. S., vonHoldt, B., Cocchiararo, B., Harms, V., Bayerl, H., Kühn, R., Förster, D. W., Fickel, J., Roos, C., & Nowak, C. (2015). A single-nucleotide polymorphism-based approach for rapid and cost-effective genetic wolf monitoring in Europe based on noninvasively collected samples. *Molecular Ecology Resources*, 15(2), 295–305. <https://doi.org/10.1111/1755-0998.12307>

Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution*, 30(1), 25–35. <https://doi.org/10.1016/j.tree.2014.10.008>

Kriska, G. (2023). *Freshwater Invertebrates in Central Europe* (2nd ed.). Springer Cham. <https://doi.org/https://doi.org/10.1007/978-3-030-95323-2>

Kumar, A., Choudhury, B., Dayanandan, S., & Khan, M. L. (2022). Molecular Genetics and Genomics Tools in Biodiversity Conservation. *Molecular Genetics and Genomics Tools in Biodiversity Conservation*. Springer Nature. <https://doi.org/10.1007/978-981-16-6005-4>

Langmead, B. (2017). A tandem simulation framework for predicting mapping quality. *Genome Biology*, 18(1). <https://doi.org/10.1186/s13059-017-1290-3>

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>

Lauková, A., Stropfiová, V., Kandričáková, A., Ščerbová, J., Semedo-Lemsaddek, T., Miltko, R., & Belzecki, G. (2015). Virulence factors genes in enterococci isolated from beavers (*Castor fiber*). *Folia Microbiologica*, 60(2), 151–154. <https://doi.org/https://doi.org/10.1007/s12223-014-0357-8>

Lehmitz, R., & Decker, P. (2017). The nuclear 28S gene fragment D3 as species marker in oribatid mites (Acari, Oribatida) from German peatlands. *Experimental and Applied Acarology*, 71(3), 259–276. <https://doi.org/10.1007/s10493-017-0126-x>

Leonard, J. A. (2008). Ancient DNA applications for wildlife conservation. *Molecular Ecology*, 17(19), 4186–4196. <https://doi.org/10.1111/j.1365-294X.2008.03891.x>

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, E., Thomas, M., Gilbert, P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., ... Earth BioGenome Project: Sequencing life for the future of life. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.*, 115. <https://doi.org/10.1073/pnas.1720115115>

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10), 1674–1676.
<https://doi.org/10.1093/bioinformatics/btv033>

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint*. <http://arxiv.org/abs/1303.3997>

Li, H. (2018). *Seqtk: a toolkit for processing sequences in FASTA/Q formats*. GitHub.
<https://github.com/lh3/seqtk>

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 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. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
<https://doi.org/10.1093/bioinformatics/btp352>

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851–1858.
<https://doi.org/10.1101/gr.078212.108>

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, 50(D1), D912–D917.
<https://doi.org/10.1093/nar/gkab1107>

López-de Sancha, A., Roig, R., Aymerich, P., Vila-Gispert, A., & Guasch, H. (2022). Trophic competition in a guild of insectivorous semi-aquatic vertebrates in a Pyrenean headwater stream: diet specialisation in the endangered *Galemys pyrenaicus*. *Mammalian Biology*, 102(5–6), 1673–1683.
<https://doi.org/10.1007/s42991-022-00265-z>

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner’s guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23), 5966–5993.
<https://doi.org/10.1111/mec.16077>

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, 2017(1).
<https://doi.org/10.7717/peerj-cs.104>

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L., & Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*, 17(12), 2815–2839.
<https://doi.org/10.1038/s41596-022-00738-y>

Lukoschus, F. S. (1969). *Eadidea desmaninae* spec. Nov. (Acarina: Myobiidae) von *Galemys pyrenaicus*. *Acarologia*, 11, 575–584.

Lynggaard, C., Frøslev, T. G., Johnson, M. S., Olsen, M. T., & Bohmann, K. (2024). Airborne environmental DNA captures terrestrial vertebrate diversity in nature. *Molecular Ecology Resources*, 24(1).
<https://doi.org/10.1111/1755-0998.13840>

Martchenko, D., & Shafer, A. B. A. (2023). Contrasting whole-genome and reduced representation sequencing for population demographic and adaptive inference: an alpine mammal case study. *Heredity*, 131(4), 273–281.
<https://doi.org/10.1038/s41437-023-00643-4>

Martínez-Guijosa, J., Martínez-Carrasco, C., López-Olvera, J. R., Fernández-Aguilar, X., Colom-Cadena, A., Cabezón, O., Mentaberre, G., Ferrer, D., Velarde, R., Gassó, D., Garel, M., Rossi, L., Lavín, S., & Serrano, E. (2015). Male-biased gastrointestinal parasitism in a nearly monomorphic mountain ungulate. *Parasites and Vectors*, 8(1).
<https://doi.org/10.1186/s13071-015-0774-9>

Mas Coma, S., Roset, F., & Mantoliu, I. (1985). *Matovius galemydis* sp. n. (Trematoda: Lecithodendrioidea), a parasite of *Galemys pyrenaicus* (Geoffroy, 1811) (Insectivora: Talpidae) in Spain. Proposal of *Combesia* gen. n. and *Matoviinae* subfam. n. *Helminthologica*, 22, 87–99.

Mc Cartney, A. M., Formenti, G., Mouton, A., De Panis, D., Marins, L. S., Leitão, H. G., Diedericks, G., Kirangwa, J., Morselli, M., Salces-Ortiz, J., Escudero, N., Iannucci, A., Natali, C., Svardal, H., Fernández, R., De Pooter, T., Joris, G., Strazisar, M., Wood, J. M. D., ... Mazzoni, C. J. (2024). The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. *Npj Biodiversity*, 3(1), 28.
<https://doi.org/10.1038/s44185-024-00054-6>

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal*, 6(3), 610–618.
<https://doi.org/10.1038/ismej.2011.139>

Mclain, K. D., Li, J., & Oliver, J. H. (2001). Interspecific and geographical variation in the sequence of rDNA expansion segment D3 of *Ixodes* ticks (Acari: Ixodidae). *Heredity*, 86, 234–242.
<https://doi.org/doi.org/10.1046/j.1365-2540.2001.00835.x>

Megléc, E. (2023). COInr and mkCOInr: Building and customizing a nonredundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline. *Molecular Ecology Resources*, 23(4), 933–945. <https://doi.org/10.1111/1755-0998.13756>

Mehlhorn, H. (2016). *Animal Parasites*. Springer Cham. <https://doi.org/https://doi.org/10.1007/978-3-319-46403-9>

Mengistu, T. S., Garcias, B., Castellanos, G., Seminati, C., Molina-López, R. A., & Darwich, L. (2022). Occurrence of multidrug resistant Gram-negative bacteria and resistance genes in semi-aquatic wildlife - *Trachemys scripta*, *Neovison vison* and *Lutra lutra* - as sentinels of environmental health. *Science of the Total Environment*, 830(154814). <https://doi.org/https://doi.org/10.1016/j.scitotenv.2022.154814>

Mondol, S., Ullas Karanth, K., Samba Kumar, N., Gopalaswamy, A. M., Andheria, A., & Ramakrishnan, U. (2009). Evaluation of non-invasive genetic sampling methods for estimating tiger population size. *Biological Conservation*, 142(10), 2350–2360. <https://doi.org/10.1016/j.biocon.2009.05.014>

Monteiro, N. M., Silva, R. M., Cunha, M., Antunes, A., Jones, A. G., & Vieira, M. N. (2014). Validating the use of colouration patterns for individual recognition in the worm pipefish using a novel set of microsatellite markers. *Molecular Ecology Resources*, 14(1), 150–156. <https://doi.org/10.1111/1755-0998.12151>

Morand, S., Krasnov, B. R., & Poulin, R. (2006). *Micromammals and Macroparasites*. Springer Tokyo. <https://doi.org/https://doi.org/10.1007/978-4-431-36025-4>

Morin, P. A., & Woodruff, D. S. (1992). Paternity exclusion using multiple hypervariable microsatellite loci amplified from nuclear DNA of hair cells. In R. D. Martin, A. D. Dixon, & E. J. Wickings (Eds.), *Paternity in Primates: Genetic tests and theories* (pp. 63–81). Karger.

Murray, K. A., Skerratt, L. F., Speare, R., & McCallum, H. (2009). Impact and dynamics of disease in species threatened by the amphibian chytrid fungus, *batrachochytrium dendrobatidis*. *Conservation Biology*, 23(5), 1242–1252. <https://doi.org/10.1111/j.1523-1739.2009.01211.x>

Nabholz, B., Mauffrey, J. F., Bazin, E., Galtier, N., & Glemin, S. (2008). Determination of mitochondrial genetic diversity in mammals. *Genetics*, 178(1), 351–361. <https://doi.org/https://doi.org/10.1534/genetics.107.073346>

Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A. L., Luk, K. C., Enge, B., Wadford, D. A., Messenger, S. L., Genrich, G. L., Pellegrino, K., Grard, G., Leroy, E., Schneider, B. S., Fair, J. N., Martínez, M. A., ... Chiu, C. Y. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, 24(7), 1180–1192. <https://doi.org/10.1101/gr.171934.113>

Nájera, F. Grande-Gómez, R., Peña, J., Vázquez, A., Palacios, M. J., Rueda, C., Corona-Bravo, A. I., Zorrilla, I., Revuelta, L., Gil-Molino, M., & Jiménez, J. (2021). Disease surveillance during the reintroduction of the Iberian lynx (*Lynx pardinus*) in southwestern Spain. *Animals*, 11(2), 1–19. <https://doi.org/10.3390/ani11020547>

Nájera, F., López, G., del Rey-Wamba, T., Malik, R. A., Garrote, G., López-Parra, M., Fernández-Pena, L., García-Tardío, M., Arenas-Rojas, R., Simón, M. A., Zorrilla, I., Fernández, I., Alcaide, E. M., Ruiz, C., Revuelta, L., Salcedo, J., Hofmann-Lehmann, R., & Meli, M. L. (2024). Long-term surveillance of the feline leukemia virus in the endangered Iberian lynx (*Lynx pardinus*) in Andalusia, Spain (2008–2021). *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-55847-3>

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: An extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20). <https://doi.org/10.1093/nar/gks678>

Nielsen, E. S., Hanson, J. O., Carvalho, S. B., Beger, M., Henriques, R., Kershaw, F., & von der Heyden, S. (2023). Molecular ecology meets systematic conservation planning. *Trends in Ecology and Evolution*, 38(2), 143–155. <https://doi.org/10.1016/j.tree.2022.09.006>

Nores, C. (2017). Desmán ibérico – *Galemys pyrenaicus*. En: Enciclopedia Virtual de los Vertebrados Españoles (A. Salvador & I. Barja, Eds.). Museo Nacional de Ciencias Naturales, Madrid. <http://www.vertebradosibericos.org/>

Norman, A. J., & Spong, G. (2015). Single nucleotide polymorphism-based dispersal estimates using noninvasive sampling. *Ecology and Evolution*, 5(15), 3056–3065. <https://doi.org/10.1002/ece3.1588>

Nunn, C. L., & Altizer, S. M. (2005). The global mammal parasite database: An online resource for infectious disease records in wild primates. *Evolutionary Anthropology: Issues, News, and Reviews*, 14(1), 1–2. <https://doi.org/10.1002/evan.20041>

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*, 27(5), 824–834. <https://doi.org/10.1101/gr.213959.116>

Ogedengbe, M. E., El-Sherry, S., Whale, J., & Barta, J. R. (2014). Complete mitochondrial genome sequences from five *Eimeria* species (Apicomplexa; Coccidia; Eimeriidae) infecting domestic turkeys. *Parasites and Vectors*, 7(1). <https://doi.org/10.1186/1756-3305-7-335>

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2), 292–294. <https://doi.org/10.1093/bioinformatics/btv566>

O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733–D745. <https://doi.org/10.1093/nar/gkv1189>

Oliver-Guimerà, A., Martínez-Carrasco, C., Tvarijonaviciute, A., Ruiz de Ybáñez, M. R., Martínez-Guijosa, J., López-Olvera, J. R., Fernández-Aguilar, X., Colom-Cadena, A., Mentaberre, G., Velarde, R., Gassó, D., Garel, M., Rossi, L., Lavín, S., & Serrano, E. (2017). The physiological cost of male-biased parasitism in a nearly monomorphic mammal. *Parasites and Vectors*, 10(1). <https://doi.org/10.1186/s13071-017-2060-5>

Oliveros, A., Terraube, J., Levengood, A. L., Powell, D., & Frère, C. H. (2023). Influence of scat ageing on the gut microbiome: how old is too old? *BMC Genomics*, 24(1). <https://doi.org/10.1186/s12864-023-09520-0>

Pajdak-Czaus, J., Platt-Samoraj, A., Szweda, W., Siwicki, A. K., & Terech-Majewska, E. (2019). *Yersinia ruckeri*—A threat not only to rainbow trout. In *Aquaculture Research* (Vol. 50, Issue 11, pp. 3083–3096). Blackwell Publishing Ltd. <https://doi.org/10.1111/are.14274>

Palmeirim, J. M., & Hoffmann, R. S. (1983). *Galemys pyrenaicus*. *Mammalian Species*, 207, 1–5

Palomares, F., Godoy, J. A., Piriz, A., O’Brien, S. J., & Johnson, W. E. (2002). Faecal genetic analysis to determine the presence and distribution of elusive carnivores: design and feasibility for the Iberian lynx. *Molecular Ecology*, 11, 2171–2182. <https://doi.org/https://doi.org/10.1046/j.1365-294X.2002.01608.x>

Papudeshi, B., Haggerty, J. M., Doane, M., Morris, M. M., Walsh, K., Beattie, D. T., Pande, D., Zaeri, P., Silva, G. G. Z., Thompson, F., Edwards, R. A., & Dinsdale, E. A. (2017). Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genomics*, 18(1). <https://doi.org/10.1186/s12864-017-4294-1>

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), 1043–1055. <https://doi.org/10.1101/gr.186072.114>

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11), 1533–1542. <https://doi.org/10.1038/s41564-017-0012-7>

Paula, D. P., Linard, B., Andow, D. A., Sujii, E. R., Pires, C. S. S., & Vogler, A. P. (2015). Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. *Molecular Ecology Resources*, 15(4), 880–892. <https://doi.org/10.1111/1755-0998.12364>

Pedersen, A. B., Jones, K. E., Nunn, C. L., & Altizer, S. (2007). Infectious diseases and extinction risk in wild mammals. *Conservation Biology*, 21(5), 1269–1279. <https://doi.org/10.1111/j.1523-1739.2007.00776.x>

Peng, P. Y., Guo, X. G., Ren, T. G., Dong, W. G., & Song, W. Y. (2016). An updated distribution and hosts: trombiculid mites (Acari: Trombidiformes) associated with small mammals in Yunnan Province, southwest China. *Parasitology Research*, 115(5), 1923–1938. <https://doi.org/10.1007/s00436-016-4934-4>

Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, 19(24), 5332–5344. <https://doi.org/10.1111/j.1365-294X.2010.04888.x>

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7(5). <https://doi.org/10.1371/journal.pone.0037135>

Peyre, A. (1955). Intersexualité du tractus génital femelle du desman des pyrénées (*Galemys pyrenaicus* G.). *Bulletin de La Société Zoologique de France*, 80, 132–138.

Peyre, A. (1956). Ecologie et biogeographie du desman (*Galemys pyrenaicus* G.) dans les Pyrénées francaises. *Mammalia*, 20, 405–418.

Peyre, A. (1957). Dimorphisme sexuel de la ceinture pelvienne d'un mammifère Insectivore, *Galemys pyrenaicus* G. *Compte Rendus de l'Académie Des Sciences de Paris*, 244, 118–120.

Peyre, A. (1962). Recherches sur l'intersexualité spécifique chez *Galemys pyrenaicus*, G. (Mammifcre Insectivore). *Archives de Biologie*, 73, 1–174.

Piggott, M. P. (2004). Effect of sample age and season of collection on the reliability of microsatellite genotyping of faecal DNA. *Wildlife Research*, 31(5), 485–493. <https://doi.org/10.1071/WR03096>

Piggott, M. P., & Taylor, A. C. (2003). Extensive evaluation of faecal preservation and DNA extraction methods in Australian native and introduced species. *Australian Journal of Zoology*, 51(4), 341–355. <https://doi.org/10.1071/ZO03012>

Pinto, Y., & Bhatt, A. S. (2024). Sequencing-based analysis of microbiomes. *Nature Reviews Genetics*, 25, 829–845. <https://doi.org/10.1038/s41576-024-00746-6>

Pödra, M., Maran, T., Sidorovich, V. E., Johnson, P. J., & Macdonald, D. W. (2013). Restoration programmes and the development of a natural diet: A case study of captive-bred European mink. *European Journal of Wildlife Research*, 59(1), 93–104. <https://doi.org/10.1007/s10344-012-0653-z>

Poinar, G. (2008). Global diversity of hairworms (Nematomorpha: Gordiaceae) in freshwater. *Hydrobiologia*, 595(1), 79–83. <https://doi.org/10.1007/s10750-007-9112-3>

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, 27(2), 313–338. <https://doi.org/10.1111/mec.14478>

Porter, T. M., & Hajibabaei, M. (2020). Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. *Frontiers in Ecology and Evolution* (Vol. 8) <https://doi.org/10.3389/fevo.2020.00248>

Poulin, R. (2006). *Evolutionary Ecology of Parasites: Second Edition*. Princeton, New Jersey: Princeton University Press.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T., Jensen, L. J., Von Mering, C., & Bork, P. (2012). eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, 40(D1). <https://doi.org/10.1093/nar/gkr1060>

Poyntz-Wright, I. P., Harrison, X. A., Pedersen, S., & Tyler, C. R. (2024). Effectiveness of eDNA for monitoring riverine macroinvertebrates. *Science of the Total Environment*, 941. <https://doi.org/https://doi.org/10.1016/j.scitotenv.2024.173621>

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., ... Zoetendal, E. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285), 59–65. <https://doi.org/10.1038/nature08821>

Quaglietta, L., Gisbert, J., García-Perea, R., & Fernández-González, A. (2024). *Galemys pyrenaicus*. *The IUCN Red List of Threatened Species 2024: e.T8826A227363991*. <https://www.iucnredlist.org/species/8826/227363991>

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(D1). <https://doi.org/10.1093/nar/gks1219>

Querejeta, M., & Castresana, J. (2018). Evolutionary history of the endemic water shrew *Neomys anomalus*: Recurrent phylogeographic patterns in semi-aquatic mammals of the Iberian Peninsula. *Ecology and Evolution*, 8(20), 10138–10146. <https://doi.org/10.1002/ece3.4487>

Querejeta, M., González-Esteban, J., Gómez, A., Fernández-González, A., Aymerich, P., Gosálbez, J., Escoda, L., Igea, J., & Castresana, J. (2016). Genomic diversity and geographical structure of the Iberian desman. *Conservation Genetics*, 17(6), 1333–1344. <https://doi.org/10.1007/s10592-016-0865-y>

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9), 833–844. <https://doi.org/10.1038/nbt.3935>

Quiroga-González, C., Cardenas, L. A. C., Ramírez, M., Reyes, A., González, C., & Stevenson, P. R. (2021). Monitoring the variation in the gut microbiota of captive woolly monkeys related to changes in diet during a reintroduction process. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-85990-0>

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System: Barcoding. *Molecular Ecology Notes*, 7(3), 355–364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>

Ribas, A., Casanova, J. C., Miquel, J., Fons, R., Guisset, C., & Feliu, C. (2005). On the fauna of digenetic trematodes, parasites of small mammals, in the Natural Reserves of Py and Mantet (Oriental Pyrenees, France). *Helminthologia*, 42, 71–75.

Ribas, A., Feliu, C., Fons, R., & Casanova, J. C. (2009). Larval digeneans of mammals in freshwater invertebrates as intermediate hosts. *vie et milieu - life and environment*, 59, 209–213.

Richard, P. B. (1962). Determination de l'âge et de la longévité chez les desman de Pyrénées (*Galemys pyrenaicus*). *Extrait de La Terre et Id Vie, Revue d'Ecologie Appliquée*, 30, 181–192.

Richard, P. B. (1986). Le desman des Pyrénées. Un mammifère inconnu à découvrir. *Editions Le Rocher, Monaco*, 118.

Ripa, A., Díaz-Caballero, J. A., Palacios-González, M. J., Zalba, J., Espinosa, A., García-Zapata, J. L., Gómez-Martín, A., Tkach, V., & Fernández-García, J. L. (2023). Non-Invasive Wildlife Disease Surveillance Using Real Time PCR Assays: The Case of the Endangered *Galemys pyrenaicus* Populations from the Central System Mountains (Extremadura, Spain). *Animals*, 13(7). <https://doi.org/10.3390/ani13071136>

Roca, J. R., Ribas, M., & Baguñà, J. (1992). Distribution, ecology, mode of reproduction and karyology of freshwater planarians (Platyhelminthes; Turbellaria; Tricladida) in the springs of the central Pyrenees. *Ecography*, 15(4), 373–384.

Rohner, S., Wohlsein, P., Prenger-Berninghoff, E., Ewers, C., Waindok, P., Strube, C., Baechlein, C., Becher, P., Wilmes, D., Rickerts, V., & Siebert, U. (2022). Pathological findings in eurasian otters (*Lutra lutra*) found dead between 2015–2020 in Schleswig-Holstein, Germany. *Animals*, 12(1). <https://doi.org/10.3390/ani12010059>

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17. <https://doi.org/10.1016/j.gecco.2019.e00547>

Sabour, S., Azimi, T., Nasser, A., Hadi, N., Mohsenzadeh, A., & Shariati, A. (2022). A global overview of the most important zoonotic bacteria pathogens transmitted from *Rattus norvegicus* to humans in urban environments. In *Infectious Medicine*, 1(3), 192–207. <https://doi.org/10.1016/j.imj.2022.07.002>

Sainsbury, A. W., & Vaughan-Higgins, R. J. (2012). Analyzing Disease Risks Associated with Translocations. *Conservation Biology*, 26(3), 442–452. <https://doi.org/10.1111/j.1523-1739.2012.01839.x>

Sánchez-Barreiro, F., Gopalakrishnan, S., Ramos-Madrigal, J., Westbury, M. V., de Manuel, M., Margaryan, A., Ciucani, M. M., Vieira, F. G., Patramanis, Y., Kalthoff, D. C., Timmons, Z., Sicheritz-Pontén, T., Dalén, L., Ryder, O. A., Zhang, G., Marquès-Bonet, T., Moodley, Y., & Gilbert, M. T. P. (2021). Historical population declines prompted significant genomic erosion in the northern and southern white rhinoceros (*Ceratotherium simum*). *Molecular Ecology*, 30(23), 6355–6369. <https://doi.org/10.1111/mec.16043>

Sánchez-Migallón Guzmán, D., Carvajal, A., García-Marín, J. F., Ferreras, M. C., Pérez, V., Mitchell, M., Urrea, F., & Ceña, J. C. (2008). Aleutian disease serology, protein electrophoresis, and pathology of the European mink (*Mustela lutreola*) from Navarra, Spain. *Journal of Zoo and Wildlife Medicine*, 39(3), 305–313. <https://doi.org/10.1638/2006-0033.1>

Santamarina, J. (1993). Feeding ecology of a vertebrate assemblage inhabiting a stream of NW Spain (Riobo; Ulla basin). *Hydrobiologia*, 252, 175–191.

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., ... Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 49(D1), D10–D17. <https://doi.org/10.1093/nar/gkaa892>

Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., ... Pruitt, K. D. (2025). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research*, 53(D1), D20–D29. <https://doi.org/10.1093/nar/gkae979>

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, 47(D1), D94–D99. <https://doi.org/10.1093/nar/gky989>

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., & Karsch-Mizrachi, I. (2024). GenBank 2024 Update. *Nucleic Acids Research*, 52(D1), D134–D137. <https://doi.org/10.1093/nar/gkad903>

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. In *Genome Research* (Vol. 20, Issue 9, pp. 1165–1173).
<https://doi.org/10.1101/gr.101360.109>

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020a). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* (Vol. 2020). Oxford University Press.
<https://doi.org/10.1093/database/baaa062>

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020b). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* (Vol. 2020). Oxford University Press.
<https://doi.org/10.1093/database/baaa062>

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8), 811–814.
<https://doi.org/10.1038/nmeth.2066>

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5(JUN), 209.
<https://doi.org/10.3389/fpls.2014.00209>

Shaw, L. P., Wang, A. D., Dylus, D., Meier, M., Pogacnik, G., Dessimoz, C., & Balloux, F. (2020). The phylogenetic range of bacterial and viral pathogens of vertebrates. *Molecular Ecology*, 29(17), 3361–3379.
<https://doi.org/10.1111/mec.15463>

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE*, 11(10).
<https://doi.org/10.1371/journal.pone.0163962>

Smith, K. F., Acevedo-Whitehouse, K., & Pedersen, A. B. (2009). The role of infectious diseases in biological conservation. In *Animal Conservation* (Vol. 12, Issue 1, pp. 1–12). <https://doi.org/10.1111/j.1469-1795.2008.00228.x>

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., Schlebusch, S. A., Wall, J. D., Alberts, S. C., Mukherjee, S., Zhou, X., & Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, 203(2), 699–714.
<https://doi.org/10.1534/genetics.116.187492>

Solberg, K. H., Bellemain, E., Drageset, O. M., Taberlet, P., & Swenson, J. E. (2006). An evaluation of field and non-invasive genetic methods to estimate brown bear (*Ursus arctos*) population size. *Biological Conservation*, 128(2), 158–168.
<https://doi.org/10.1016/j.biocon.2005.09.025>

Sonnenberg, R., Nolte, A., & Tautz, D. (2007). An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. *Frontiers in Zoology*, 4.
<https://doi.org/10.1186/1742-9994-4-6>

Srivathsan, A., Ang, A., Vogler, A. P., & Meier, R. (2016). Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Frontiers in Zoology*, 13(1).
<https://doi.org/10.1186/s12983-016-0150-4>

Srivathsan, A., Sha, J. C. M., Vogler, A. P., & Meier, R. (2015). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources*, 15(2), 250–261.
<https://doi.org/10.1111/1755-0998.12302>

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313.
<https://doi.org/10.1093/bioinformatics/btu033>

Stetz, J. B., Smith, S., Sawaya, M. A., Ramsey, A. B., Amish, S. J., Schwartz, M. K., & Gordon, L. (2016). Discovery of 20,000 RAD-SNPs and development of a 52-SNP array for monitoring river otters. *Conservation Genetics Resources*, 8, 299–302.
<https://doi.org/10.1007/s12686-016-0558-3>

Suminda, G. G. D., Bhandari, S., Won, Y., Goutam, U., Kanth Pulicherla, K., Son, Y. O., & Ghosh, M. (2022). High-throughput sequencing technologies in the detection of livestock pathogens, diagnosis, and zoonotic surveillance. *Computational and Structural Biotechnology Journal* (Vol. 20, pp. 5378–5392). Elsevier B.V.
<https://doi.org/10.1016/j.csbj.2022.09.028>

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., De Vos, W. M., Wang, J., Li, J., Doré, J., Dusko Ehrlich, S., ... Bork, P. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12), 1196–1199.
<https://doi.org/10.1038/nmeth.2693>

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, 19(1).
<https://doi.org/10.1186/s13059-018-1520-3>

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford University Press.
<https://doi.org/10.1093/oso/9780198767220.001.0001>

Taberlet, P., & Bouvet, J. (1992). Bear conservation genetics. *Nature*, 358, 197.
<https://doi.org/https://doi.org/10.1038/358197a0>

Taberlet, P., & Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society*, 68(1–2), 41–55.
<https://doi.org/10.1111/j.1095-8312.1999.tb01157.x>

Taylor, R. S., Manseau, M., Redquest, B., Keobouasone, S., Gagné, P., Martineau, C., & Wilson, P. J. (2022). Whole genome sequences from non-invasively collected caribou faecal samples. *Conservation Genetics Resources*, 14(1), 53–68.
<https://doi.org/10.1007/s12686-021-01235-2>

Tedersoo, L., Hosseini Moghaddam, M. S., Mikryukov, V., Hakimzadeh, A., Bahram, M., Nilsson, R. H., Yatsiuk, I., Geisen, S., Schwelm, A., Piwosz, K., Prous, M., Sildever, S., Chmolewska, D., Rueckert, S., Skaloud, P., Laas, P., Tines, M., Jung, J. H., Choi, J. H., ... Anslan, S. (2024). EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. *Database*, 2024.
<https://doi.org/10.1093/database/baee043>

Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P. R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G. R., Godoy, J. A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Secomandi, S., ... Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in Genetics*, 39(7), 545–559.
<https://doi.org/10.1016/j.tig.2023.01.005>

Tian, S., Jiang, B. G., Liu, W. S., Chen, H. R., Gao, Z. H., Pu, E. N., Li, Y. Q., Chen, J. J., Fang, L. Q., Wang, G. L., Du, C. H., & Wei, Y. H. (2023). Zoonotic pathogens identified in rodents and shrews from four provinces, China, 2015–2022. *Epidemiology and Infection*, 151.
<https://doi.org/10.1017/S0950268823001450>

Timon-David, J. (1960). Un trématode parasite du Desman des Pyrénées (*Galemys pyrenaicus* Geoffr.). *Bulletin de La Société D'histoire Naturelle de Toulouse.*, 95, 11–17.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure & genetic diversity from metagenomes. *Genome Research*, 27(4), 626–638.
<https://doi.org/10.1101/gr.216242.116>

Tyagi, A., Khan, A., Thatte, P., & Ramakrishnan, U. (2022). Genome-wide single nucleotide polymorphism (SNP) markers from fecal samples reveal anthropogenic impacts on connectivity: case of a small carnivore in the central Indian landscape. *Animal Conservation*, 25(5), 648–659.
<https://doi.org/10.1111/acv.12770>

Tyagi, A., Yadav, N., Pandit, A., & Ramakrishnan, U. (2024). On the road to losing connectivity: Faecal samples provide genome-wide insights into anthropogenic impacts on two large herbivore species in central India. *Molecular Ecology*, 33(16).
<https://doi.org/10.1111/mec.17461>

Unterköfler, M. S., Schausberger, M., Deutz, A., Gressmann, G., Kübber-Heiss, A., Ferroglio, E., & Joachim, A. (2023). Sarcoptic mange in wild ungulates in the European Alps – A systematic review. *International Journal for Parasitology: Parasites and Wildlife*, 22, 121–125.
<https://doi.org/10.1016/j.ijppaw.2023.10.003>

Vadlejch, J., Kyriánová, I. A., Rylková, K., Zikmund, M., & Langrová, I. (2017). Health risks associated with wild animal translocation: a case of the European bison and an alien parasite. *Biological Invasions*, 19(4), 1121–1125.
<https://doi.org/10.1007/s10530-016-1306-z>

Van Der Valk, T., Díez-Del-Molino, D., Marques-Bonet, T., Guschanski, K., & Dalén, L. (2019). Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. *Current Biology*, 29(1), 165–170.
<https://doi.org/10.1016/j.cub.2018.11.055>

Vaucher, C. (1975). Sur quelques trematodes parasites de Chiroptères et d’Insectivores. *Bulletin de La Société. Neuchâtel Sci. Nat.* , 98, 17–25.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., ... Smith, H. O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304, 66–74.
<https://doi.org/10.1126/science.1093857>

Vidal, O., Pérez-Serra, A., & Pla, C. (2010). A sex determination protocol for the Iberian desman (*Galemys pyrenaicus*) based on a three primer amplification of DBX and DBY fragments with non-invasive samples. *Conservation Genetics*, 11(3), 1185–1187.
<https://doi.org/10.1007/s10592-009-9913-1>

Vulfson, L., Pedersen, K., Chrie Ål, M., Frydendahl, K., Andersen, T. H., Madsen, M., & Dietz, H. H. (2001). Serogroups and antimicrobial susceptibility among *Escherichia coli* isolated from farmed mink (*Mustela vison Schreiber*) in Denmark. *Vet Microbiol*, 79, 143–153.
[https://doi.org/https://doi.org/10.1016/S0378-1135\(00\)00343-6](https://doi.org/https://doi.org/10.1016/S0378-1135(00)00343-6)

Wang, Q., & Cole, J. R. (2024). Updated RDP taxonomy and RDP Classifier for more accurate taxonomic classification. *Microbiology Resource Announcements*, 13(4).
<https://doi.org/10.1128/mra.01063-23>

Wei, F., Wu, Q., Hu, Y., Huang, G., Nie, Y., & Yan, L. (2019). Conservation metagenomics: a new branch of conservation biology. *Science China Life Sciences*, 62(2), 168–178. Science in China Press.
<https://doi.org/10.1007/s11427-018-9423-3>

White, R. J., & Razgour, O. (2020). Emerging zoonotic diseases originating in mammals: a systematic review of effects of anthropogenic land-use change. *Mammal Review*, 50(4), 336–352.
<https://doi.org/10.1111/mam.12201>

Willerslev, E., Gilbert, M. T. P., Binladen, J., Ho, S. Y., Campos, P. F., Ratan, A., Tomsho, L. P., Da Fonseca, R. R., Kuznetsova, T. V., Nowak-Kemp, M., Roth, T. L., Miller, W., & Schuster, S. C. (2009). Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evolutionary Biology*, 9(1).
<https://doi.org/10.1186/1471-2148-9-95>

Williams, C. L., Caraballo-Rodríguez, A. M., Allaband, C., Zarrinpar, A., Knight, R., & Gauglitz, J. M. (2018). Wildlife-microbiome interactions and disease: exploring opportunities for disease mitigation across ecological scales. *Drug Discovery Today: Disease Models*, 28, 105–115.
<https://doi.org/10.1016/j.ddmod.2019.08.012>

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1).
<https://doi.org/10.1186/s13059-019-1891-0>

Wood, D. E., & Salzberg, S. L. (2014). *Kraken: ultrafast metagenomic sequence classification using exact alignments*.
<http://ccb.jhu.edu/software/kraken/>.

Wright, R. J., Comeau, A. M., & Langille, M. G. I. (2023). From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microbial Genomics*, 9(3).
<https://doi.org/10.1099/mgen.0.000949>

Wu, S., Xiong, J., & Yu, Y. Y. (2015). Taxonomic resolutions based on 18S rRNA Genes: A case study of subclass Copepoda. *PLoS ONE*, 10(6).
<https://doi.org/10.1371/journal.pone.0131498>

Wylezich, C., Nies, G., Mylnikov, A. P., Tautz, D., & Arndt, H. (2010). An evaluation of the use of the LSU rRNA D1-D5 domain for DNA-based taxonomy of eukaryotic protists. *Protist*, 161(3), 342–352.
<https://doi.org/10.1016/j.protis.2010.01.003>

Xu, R., Rajeev, S., & Salvador, L. C. M. (2023). The selection of software and database for metagenomics sequence analysis impacts the outcome of microbial profiling and pathogen detection. *PLoS ONE*, 18(4 April).
<https://doi.org/10.1371/journal.pone.0284031>

Yan, Y., Nguyen, L. H., Franzosa, E. A., & Huttenhower, C. (2020). Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine*, 12(1).
<https://doi.org/10.1186/s13073-020-00765-y>

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. In *Cell* (Vol. 178(4), 779–794).
<https://doi.org/10.1016/j.cell.2019.07.010>

Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., Stoica, I., Karp, R. M., & Sittler, T. (2011). Faster and More Accurate Sequence Alignment with SNAP.
<http://arxiv.org/abs/1111.5572>

Zhang, F., Yu, J., Wu, S., Li, S., Zou, C., Wang, Q., & Sun, R. (2017). Keeping and breeding the rescued Sunda pangolins (*Manis javanica*) in captivity. *Zoo Biology*, 36(6), 387–396.
<https://doi.org/10.1002/zoo.21388>

IX.

ANNEX

Table S1. Samples of *Galemys pyrenaicus* used in this study, indicating the main river associated with the sub-basins where the subpopulations are located, as well as the corresponding subpopulations.

Sample #	Sample	Main river	Subpopulation
1	BC2042	Tormes	Aravalle
2	BC2045	Tormes	Endrinal
3	BC2280	Adaja	Adaja
4	BC2360	Adaja	Adaja
5	BC2378	Tormes	Endrinal
6	BC2381	Tormes	Endrinal
7	BC2408	Tormes	Aravalle
8	BC2416	Adaja	Adaja
9	BC2765	Tormes	Becedillas
10	BC2856	Tormes	Becedillas
11	BC2865	Tormes	Endrinal
12	BC2876	Tormes	Aravalle
13	BC2934	Adaja	Adaja
14	BC2989	Tormes	Becedillas
15	BC3015	Tormes	Aravalle
16	BC3267	Tormes	Aravalle
17	BC3343	Tormes	Aravalle
18	BC3345	Tormes	Endrinal
19	BC3876	Tormes	Endrinal
20	BC3895	Tormes	Endrinal
21	BC3899	Tormes	Endrinal
22	BC3903	Tormes	Endrinal
23	BC3906	Tormes	Endrinal

Table S2. Species of the genus *Yersinia* according to NCBI Taxonomy, along with the size of the reference genome used. Their pathogenicity is also indicated.

Species	Genome size (pb)	Source	Pathogenicity
<i>Yersinia aldovae</i>	4,471,090	NCBI Taxonomy	
<i>Yersinia aleksiciae</i>	4,526,044	NCBI Taxonomy	
<i>Yersinia alsatica</i>	4,901,396	NCBI Taxonomy	
<i>Yersinia artesianae</i>	4,520,664	NCBI Taxonomy	
<i>Yersinia bercovieri</i>	4,428,721	NCBI Taxonomy	
<i>Yersinia canariae</i>	4,710,154	NCBI Taxonomy	
<i>Yersinia enterocolitica</i>	4,548,822	NCBI Taxonomy	X
<i>Yersinia entomophaga</i>	4,275,406	NCBI Taxonomy	
<i>Yersinia frederiksenii</i>	4,941,872	NCBI Taxonomy	
<i>Yersinia hibernica</i>	4,803,435	NCBI Taxonomy	
<i>Yersinia intermedia</i>	4,928,910	NCBI Taxonomy	
<i>Yersinia kristensenii</i>	4,733,508	NCBI Taxonomy	
<i>Yersinia massiliensis</i>	5,050,276	NCBI Taxonomy	
<i>Yersinia mollaretii</i>	4,603,534	NCBI Taxonomy	
<i>Yersinia nurmii</i>	4,143,134	NCBI Taxonomy	
<i>Yersinia pekkanenii</i>	5,046,986	NCBI Taxonomy	
<i>Yersinia pestis</i>	4,658,411	NCBI Taxonomy	X
<i>Yersinia proxima</i>	4,616,898	NCBI Taxonomy	
<i>Yersinia pseudotuberculosis</i>	4,839,430	NCBI Taxonomy	X
<i>Yersinia rochesterensis</i>	4,448,883	NCBI Taxonomy	
<i>Yersinia rohdei</i>	4,372,253	NCBI Taxonomy	
<i>Yersinia ruckeri</i>	3,894,226	NCBI Taxonomy	X
<i>Yersinia similis</i>	4,964,409	NCBI Taxonomy	
<i>Yersinia thracica</i>	4,420.333	NCBI Taxonomy	
<i>Yersinia vastinensis</i>	4,540,093	NCBI Taxonomy	
<i>Yersinia wautersii</i>	4,858,848	NCBI Taxonomy	

Table S3. Species of pathogenic bacteria analyzed and the size of their reference genomes.

Species	Synonym	Genome size (bp)
<i>Acinetobacter baumannii</i>		3,980,230
<i>Actinobacillus lignieresii</i>		2,236,520
<i>Aeromonas hydrophila</i>		4,733,702
<i>Aeromonas salmonicida</i>		4,954,811
<i>Aeromonas sobria</i>		4,683,669
<i>Aeromonas veronii</i>		4,561,870
<i>Anaplasma marginale</i>		1,202,435
<i>Anaplasma phagocytophilum</i>		1,481,598
<i>Bacillus anthracis</i>		5,503,926
<i>Bacillus cereus</i>		5,836,971
<i>Bartonella bacilliformis</i>		1,411,655
<i>Bartonella henselae</i>		1,905,383
<i>Bartonella quintana</i>		1,588,683
<i>Bordetella pertussis</i>		4,088,701
<i>Borrelia afzelii</i>		906,136
<i>Borrelia burgdorferi</i>		1,321,434
<i>Borrelia garinii</i>		905,692
<i>Brucella abortus</i>		3,278,307
<i>Brucella melitensis</i>		3,294,931
<i>Brucella microti</i>		3,337,369
<i>Brucella neotomae</i>		3,329,628
<i>Brucella ovis</i>		3,275,590
<i>Brucella suis</i>		3,315,175
<i>Burkholderia cepacia</i>		8,366,868
<i>Burkholderia pseudomallei</i>		7,085,397
<i>Campylobacter coli</i>		1,716,536
<i>Campylobacter fetus</i>		1,804,582
<i>Campylobacter jejuni</i>		1,641,481
<i>Chlamydia psittaci</i>		1,179,220
<i>Chlamydia trachomatis</i>		1,042,519
<i>Citrobacter freundii</i>		5,171,093
<i>Clostridioides difficile</i>	<i>Clostridium difficile</i>	4,095,894
<i>Clostridium botulinum</i>		3,903,260
<i>Clostridium chauvoei</i>		2,889,569
<i>Clostridium histolyticum</i>		2,740,791
<i>Clostridium novyi</i>		2,499,078
<i>Clostridium perfringens</i>		3,275,424
<i>Clostridium septicum</i>		3,404,718
<i>Clostridium tetani</i>		2,873,333

Table S3 (cont.)

<i>Corynebacterium diphtheriae</i>	2,406,896
<i>Corynebacterium striatum</i>	2,904,831
<i>Coxiella burnetii</i>	2,032,807
<i>Edwardsiella ictaluri</i>	3,844,237
<i>Edwardsiella piscicida</i>	3,819,771
<i>Edwardsiella tarda</i>	3,720,168
<i>Elizabethkingia anophelis</i>	4,058,311
<i>Enterobacter asburiae</i>	4,768,325
<i>Enterobacter bugandensis</i>	4,734,039
<i>Enterobacter cancerogenus</i>	4,736,684
<i>Enterobacter chengduensis</i>	5,218,125
<i>Enterobacter chuandaensis</i>	4,634,324
<i>Enterobacter cloacae</i>	5,023,439
<i>Enterobacter hormaechei</i>	4,855,498
<i>Enterobacter kobei</i>	4,773,801
<i>Enterobacter ludwigii</i>	4,952,770
<i>Enterobacter mori</i>	4,844,012
<i>Enterobacter oligotrophicus</i>	4,476,585
<i>Enterobacter roggenkampii</i>	4,899,997
<i>Enterobacter sichuanensis</i>	4,711,389
<i>Enterobacter soli</i>	5,012,132
<i>Enterococcus faecalis</i>	2,870,381
<i>Enterococcus faecium</i>	2,919,198
<i>Enterococcus hirae</i>	2,845,651
<i>Escherichia coli</i>	5,594,605
<i>Flavobacterium psychrophilum</i>	2,830,557
<i>Francisella tularensis</i>	1,870,206
<i>Fusobacterium necrophorum</i>	2,288,480
<i>Haemophilus influenzae</i>	1,846,259
<i>Helicobacter pylori</i>	1,624,458
<i>Klebsiella oxytoca</i>	5,879,076
<i>Klebsiella pneumoniae</i>	5,682,322
<i>Kluyvera intermedia</i>	4,938,529
<i>Kosakonia oryzendophytica</i>	5,132,919
<i>Kosakonia oryziphila</i>	4,814,900
<i>Legionella pneumophila</i>	3,504,074
<i>Leptospira interrogans</i>	4,630,763
<i>Listeria innocua</i>	2,922,148
<i>Listeria ivanovii</i>	2,919,548
<i>Listeria monocytogenes</i>	2,944,528

Table S3 (cont.)

<i>Listeria seeligeri</i>		2,797,636
<i>Mannheimia haemolytica</i>		2,757,078
<i>Mesomycoplasma conjunctivae</i>	<i>Mycoplasma conjunctivae</i>	846,214
<i>Mesomycoplasma hyopneumoniae</i>	<i>Mycoplasma hyopneumoniae</i>	921,093
<i>Morganella morganii</i>		3,906,921
<i>Mycobacterium bovis</i>	<i>M. tuberculosis variant bovis</i>	4,411,814
<i>Mycobacterium caprae</i>	<i>M. tuberculosis variant caprae</i>	4,324,961
<i>Mycobacterium microti</i>	<i>M. tuberculosis variant microti</i>	4,369,915
<i>Mycobacterium tuberculosis</i>		4,411,532
<i>Mycoplasma mycoides</i>		1,084,586
<i>Mycoplasma pneumoniae</i>	<i>Mycoplasma pneumoniae</i>	823,017
<i>Mycoplasma agalactiae</i>	<i>Mycoplasma agalactiae</i>	916,806
<i>Neisseria gonorrhoeae</i>		2,171,755
<i>Neisseria lactamica</i>		2,200,224
<i>Neisseria meningitidis</i>		2,157,444
<i>Neisseria polysacchara</i>		2,029,584
<i>Pasteurella multocida</i>		2,335,516
<i>Photobacterium damsela</i>		4,427,003
<i>Phytobacter massiliensis</i>		5,003,540
<i>Pluralibacter gergoviae</i>		5,408,082
<i>Providencia alcalifaciens</i>		3,990,106
<i>Pseudomonas aeruginosa</i>		6,264,404
<i>Pseudomonas putida</i>		6,156,701
<i>Rickettsia conorii</i>		1,268,755
<i>Rickettsia monacensis</i>		1,353,450
<i>Rickettsia prowazekii</i>		1,109,804
<i>Rickettsia rickettsii</i>		1,257,710
<i>Rickettsia typhi</i>		1,112,372
<i>Salmonella bongori</i>		4,773,537
<i>Salmonella enterica</i>		4,951,383
<i>Serratia marcescens</i>		5,238,537
<i>Shewanella algae</i>		4,990,025
<i>Shigella dysenteriae</i>		5,192,674
<i>Shigella flexneri</i>		4,828,820
<i>Shigella sonnei</i>		4,762,774
<i>Staphylococcus aureus</i>		2,821,361
<i>Staphylococcus pseudintermedius</i>		2,629,596
<i>Stenotrophomonas maltophilia</i>		4,481,118
<i>Streptobacillus moniliformis</i>		1,673,280

Table S3 (cont.)

<i>Streptococcus agalactiae</i>		2,079,123
<i>Streptococcus equi</i>		2,154,778
<i>Streptococcus iniae</i>		2,078,360
<i>Streptococcus mutans</i>		2,028,032
<i>Streptococcus pneumoniae</i>		2,154,362
<i>Streptococcus pyogenes</i>		1,746,380
<i>Streptococcus suis</i>		2,170,808
<i>Treponema pallidum</i>		1,139,330
<i>Vibrio cholerae</i>		4,138,412
<i>Vibrio fluvialis</i>		4,827,733
<i>Vibrio harveyi</i>		5,869,008
<i>Vibrio metschnikovii</i>		3,831,334
<i>Vibrio parahaemolyticus</i>		5,165,770
<i>Vibrio vulnificus</i>		5,117,890
<i>Yersinia enterocolitica</i>		4,548,822
<i>Yersinia pestis</i>		4,658,411
<i>Yersinia pseudotuberculosis</i>		4,839,430
<i>Yersinia ruckeri</i>		3,894,226

Table S4 The upper section shows DNA concentrations in the samples (ng/μL), the number of Illumina reads obtained for each sample, the percentage of duplicated reads filtered, the number of retained reads and their percentage, and the most frequent insert size in each sample. The lower section displays statistics for the analysis of the results: average, standard deviation (STD), coefficient of variation (CV), minimum value (Min), maximum value (Max), and median.

Sample	DNA (ng/μl)	Total raw reads	Fastp filtering			
			Duplication rate (%)	Retained reads (n)	Retained reads (%)	Insert size peak
BC2042	0.416	77,911,584	10.8143	66,397,220	85	202
BC2045	0.142	85,621,556	14.5819	63,295,728	74	168
BC2280	0.436	93,152,950	12.3679	77,518,202	83	173
BC2360	0.456	71,959,282	11.7864	60,255,746	84	271
BC2378	0.392	78,752,260	11.6913	62,977,472	80	160
BC2381	0.246	76,509,430	12.3964	61,028,108	80	169
BC2408	0.384	108,021,396	10.5968	84,184,200	78	172
BC2416	0.340	83,776,698	10.9693	71,018,018	85	155
BC2765	0.478	66,132,750	10.374	56,312,202	85	172
BC2856	1.81	100,940,134	11.7107	82,903,094	82	178
BC2865	0.274	88,948,006	12.1693	71,110,558	80	161
BC2876	0.478	79,788,742	11.0164	67,644,992	85	186
BC2934	0.182	91,058,572	11.2636	75,000,748	82	183
BC2989	0.254	82,428,140	12.865	66,424,834	81	228
BC3015	2.28	52,030,508	9.25385	42,603,532	82	167
BC3267	0.334	93,826,662	12.7939	78,037,646	83	180
BC3343	0.294	70,503,076	10.6645	58,363,984	83	179
BC3345	0.272	81,209,102	12.5661	66,333,060	82	215
BC3876	0.338	98,439,460	12.0939	79,548,714	81	183
BC3895	0.416	86,496,334	12.4508	72,157,126	83	224
BC3899	3.48	59,816,578	8.71211	42,235,116	71	156
BC3903	0.192	64,088,746	11.1205	50,908,242	79	152
BC3906	0.476	75,769,478	11.5143	63,112,800	83	183
Average	0.62	81,181,802	12	66,059,624	81	183
STD	0.80	13,516,925.66	1.25	11,332,586.31	3.51	28.12
CV (%)	128.26	16.65	10.81	17.16	4.31	15.34
Min	0.14	52,030,508	9	42,235,116	71	152
Max	3.48	108,021,396	15	84,184,200	85	271
Median	0.38	81,209,102	12	66,397,220	82	178

Table S5. Results of the processing of endogenous-origin reads. For each sample, the table shows the total number of reads after quality filtering, the number of reads aligned to the *G. pyrenaicus* nuclear genome, and their respective depth and breadth of coverage values. Additionally, it includes the percentage of reads filtered due to low entropy (complexity) values and the number of retained reads. The table also shows the number of reads aligned to the mitochondrial genome and its depth and breadth of coverage values. Finally, it reports the number of exogenous reads in each sample and their mean length (bp).

Sample	Filtered reads (n)	Nuclear genome			BBduk filtering		Mitogenome			Exogenous reads	
		Reads (n)	Depth (X)	Breadth (%)	Entropy (%)	Retained reads (n)	Reads (n)	Depth (X)	Breadth (%)	Total reads (n)	Mean lenght (pb)
BC2042	66,397,220	47,582	0.004	0.17	6.95	61,739,996	2,308	19.35	98.34	61,736,980	149.351
BC2045	63,295,728	12,092,536	0.971	41	2.02	50,169,980	33,480	247.28	99.99	50,124,430	144.638
BC2280	77,518,202	17,004,552	1.359	54.02	1.38	59,677,178	19,986	164.62	99.99	59,649,976	146.892
BC2360	60,255,746	1,019,544	0.084	5.98	5.34	56,071,220	13,766	111.71	100	56,053,206	150.499
BC2378	62,977,472	12,598,510	1.006	43.73	3.54	48,598,022	47,040	368.64	99.85	48,532,104	144.569
BC2381	61,028,108	220,114	0.018	1.04	0.89	60,264,562	6,014	47.61	99.61	60,256,352	147.092
BC2408	84,184,200	2,862,668	0.229	12.74	4.18	77,922,810	12,758	102.14	100	77,905,478	146.147
BC2416	71,018,018	2,382,304	0.19	10.53	2.68	66,796,844	7,142	58.34	99.53	66,786,794	145.737
BC2765	56,312,202	6,056,388	0.48	23.76	6.38	47,047,892	21,416	176.75	99.76	47,017,792	146.784
BC2856	82,903,094	197,424	0.015	0.82	1.52	81,449,238	2,406	19.64	95.37	81,444,940	146.836
BC2865	71,110,558	38,876,690	3.088	79.04	5.43	30,484,268	54,522	431.66	100	30,409,082	144.337
BC2876	67,644,992	539,906	0.044	2.69	1.75	65,930,596	2,796	23.18	98.8	65,926,618	148.439
BC2934	75,000,748	876,468	0.065	3.15	1.74	72,837,992	26,306	209.95	100	72,800,924	147.37
BC2989	66,424,834	362,066	0.028	1.62	3.24	63,923,150	9,336	74.17	100	63,910,586	148.803
BC3015	42,603,532	1,132,576	0.088	5.02	0.8	41,137,140	3,232	26.17	97.54	41,132,864	145.834
BC3267	78,037,646	816,462	0.066	3.93	2.81	75,048,408	8,592	70.14	99.53	75,036,328	147.748
BC3343	58,363,984	12,710,970	1.029	45.73	9.17	41,466,972	20,210	164.01	100	41,438,942	146.748
BC3345	66,333,060	10,573,752	0.862	41.22	3.47	53,822,292	125,918	991.4	100	53,654,356	148.875
BC3876	79,548,714	151,170	0.012	0.56	0.61	78,914,314	13,960	109.66	99.9	78,895,038	146.793
BC3895	72,157,126	1,832,636	0.146	8.59	6.03	66,080,668	14,048	116.5	100	66,061,256	149.241
BC3899	42,235,116	11,314,712	0.852	35.71	3.82	29,738,802	14,332	112.52	99.85	29,719,132	144.347
BC3903	50,908,242	4,590,572	0.37	19.52	6.08	43,500,472	15,920	125.7	100	43,476,990	144.355
BC3906	63,112,800	3,374,458	0.271	14.99	4.97	56,767,170	14,040	113.26	99.74	56,744,902	147.269

Table S6. The upper section shows the number of reads assigned to each category: nuclear genome and mitochondrial genome of *G. pyrenaicus*, Metazoa, Fungi, Viridiplantae, SAR, Other Eukaryota, Bacteria, Archaea, Viruses, and Unclassified. The lower section presents general statistics for each category, including the total reads across all samples, the weighted average number of reads (W. AV.), the weighted standard deviation (W. STD), the coefficient of variation (CV), and the maximum (Max) and minimum (Min) values.

Sample	Nuc. Gen. (n)	Mitogen. (n)	Metazoa (n)	Fungi (n)	Viridip. (n)	SAR (n)	Other Euk. (n)	Bacteria (n)	Archaea (n)	Viruses (n)	Unclassified (n)
BC2042	47,582	2,308	11,386,360	152,456	1,279,728	112,452	767,206	5,531,050	32,992	56,914	47,028,172
BC2045	12,092,536	33,480	7,368,002	125,270	748,208	615,448	581,038	9,371,812	35,564	112,612	32,211,758
BC2280	17,004,552	19,986	6,803,802	128,752	1,271,068	101,730	468,240	8,516,742	35,456	68,600	43,099,274
BC2360	1,019,544	13,766	3,249,024	93,780	887,330	255,160	328,072	6,083,870	17,594	39,036	48,268,570
BC2378	12,598,510	47,040	4,008,048	105,616	671,704	112,566	237,352	5,161,276	19,656	49,252	39,966,452
BC2381	220,114	6,014	8,007,674	123,458	1,168,840	85,100	603,556	19,229,912	40,614	140,172	31,402,654
BC2408	2,862,668	12,758	6,221,534	215,436	4,904,016	240,166	562,130	6,528,644	28,118	70,668	62,538,062
BC2416	2,382,304	7,142	11,997,110	133,872	826,674	63,996	578,650	1,143,180	19,358	50,996	53,814,736
BC2765	6,056,388	21,416	3,436,000	87,898	505,294	241,068	244,502	3,639,160	15,826	34,936	42,029,714
BC2856	197,424	2,406	2,455,536	237,142	479,212	124,400	190,600	15,311,118	56,294	38,480	63,810,482
BC2865	38,876,690	54,522	3,739,544	67,502	548,808	224,042	199,768	2,159,598	9,988	17,322	25,212,774
BC2876	539,906	2,796	4,809,408	143,380	793,112	59,246	290,748	5,713,162	24,478	90,336	55,178,420
BC2934	876,468	26,306	4,635,914	153,568	2,430,842	78,558	341,590	15,869,432	32,852	65,126	50,490,092
BC2989	362,066	9,336	5,067,266	117,588	468,030	124,634	323,112	6,964,714	24,744	58,318	52,905,026
BC3015	1,132,576	3,232	1,690,652	67,396	326,720	29,978	115,858	5,048,092	14,816	47,350	34,126,862
BC3267	816,462	8,592	8,744,846	174,756	1,924,150	147,954	523,780	4,521,538	28,520	58,650	61,088,398
BC3343	12,710,970	20,210	7,086,344	93,298	1,064,534	162,156	445,496	1,760,368	15,150	35,708	34,969,750
BC3345	10,573,752	125,918	3,800,606	94,414	911,980	101,466	251,050	8,486,096	17,898	40,472	41,929,408
BC3876	151,170	13,960	3,150,152	195,606	847,106	40,914	210,978	13,285,070	41,080	71,598	61,541,080
BC3895	1,832,636	14,048	11,584,792	178,064	1,967,646	147,670	757,370	5,495,048	24,708	52,292	50,102,852
BC3899	11,314,712	14,332	2,378,334	58,734	249,776	84,530	127,714	4,095,378	11,540	18,812	23,881,254
BC3903	4,590,572	15,920	4,726,340	87,368	851,978	131,212	301,574	1,792,460	14,780	33,368	38,362,670
BC3906	3,374,458	14,040	6,672,180	132,576	853,952	289,888	447,300	4,235,558	24,280	51,786	47,016,782
Total	141,634,060	489,528	133,019,468	2,967,930	25,980,708	3,574,334	8,897,684	159,943,278	586,306	1,302,804	1,040,975,242
W. AV.	6079810.8	21320.0	5926054.9	135497.4	1216066.1	155459.0	397125.3	7241293.4	26650.2	57691.7	46662243.0
W. STD	8912563.7	25953.3	2924610.1	47440.2	1041102.6	118169.8	184494.7	4757682.3	11278.6	26614.6	11457279.2
CV (%)	146.6	121.7	49.4	35.0	85.6	76.0	46.5	65.7	42.3	46.1	24.6
Min	47582	2308	1690652	58734	249776	29978	115858	1143180	9988	17322	23881254
Max	38876690	125918	11997110	237142	4904016	615448	767206	19229912	56294	140172	63810482

Table S7. Sex determination. For each sample, the number of high-quality reads aligned to the nuclear genome (NG) of *G. pyrenaicus* and the Y chromosome (Y) is presented, along with their corresponding mean depth of coverage, the ratio of the mean depth of coverage, and the sex assignment based on this ratio.

Sample	Filtered NG reads (n)	NG depth (X)	Y reads (n)	Y depth (X)	Y:NG (depth (X))	Sex
BC2042	35,110	0.003	0	0	0.00	Female
BC2045	11,993,118	0.963	940	0.583	0.61	Male
BC2280	16,886,448	1.35	0	0	0.00	Female
BC2360	1,010,720	0.083	72	0.045	0.54	Male
BC2378	12,475,390	0.996	966	0.595	0.60	Male
BC2381	215,298	0.017	0	0	0.00	Female
BC2408	2,830,224	0.226	234	0.143	0.63	Male
BC2416	2,359,328	0.189	174	0.106	0.56	Male
BC2765	6,007,454	0.477	532	0.322	0.68	Male
BC2856	192,954	0.015	0	0	0.00	Female
BC2865	38,632,796	3.069	0	0	0.00	Female
BC2876	525,316	0.043	44	0.028	0.65	Male
BC2934	853,930	0.063	0	0	0.00	Female
BC2989	356,290	0.028	0	0	0.00	Female
BC3015	1,122,226	0.088	80	0.05	0.57	Male
BC3267	807,100	0.065	68	0.042	0.65	Male
BC3343	12,566,154	1.018	0	0	0.00	Female
BC3345	10,507,858	0.857	794	0.499	0.58	Male
BC3876	135,168	0.01	12	0.008	0.80	Male
BC3895	1,817,618	0.145	180	0.112	0.77	Male
BC3899	11,204,000	0.845	988	0.57	0.67	Male
BC3903	4,547,826	0.367	356	0.222	0.60	Male
BC3906	3,336,506	0.268	268	0.169	0.63	Male

Table S8. Statistics of the final set of consensus sequences after quality filtering. The table shows the minimum, maximum, and mean values for the following parameters: number of reads comprising the consensus sequence (n), depth of coverage (X), breadth of coverage (%), consensus sequence length (bp), total percentage identity (%), query coverage (%), and e-value.

Statistics	Reads (n)	Depht cov (X)	Breadth cov (%)	Lenght (pb)	Total Pident (%)	Qcovs (%)	e-value
Min	4	0.02	1.84	500	90.15	50	0
Max	55090	1717.29	100	16329	100	100	4.75e-57
Average	132.6	7.8	41.4	992.6	99.0	97.8	7.70e-61

Table S9. Depth of coverage (X) values for each detected diet element.

Diet Element	BC2042	BC2045	BC2280	BC2360	BC2378	BC2381	BC2408	BC2416	BC2765	BC2856	BC2865	BC2876	BC2934	BC2989	BC3015	BC3267	BC3343	BC3345	BC3876	BC3895	BC3899	BC3903	BC3906
Trichoptera	0.73	4.98	0.48	0.66	1.30	0.37	1.09	66.20	1.14	0.37	0.39	1.29	0.70	2.46	1.48	18.26	15.02	0.61	1.20	12.39	1.43	1.50	5.53
Ephemeroptera	16.88	0.83	1.20	0.16	0.85	3.98	0.85	1.74	1.22	2.82	0.40	1.96	18.11	0.79	0.90	2.99	5.59	0.76	0.43	13.01	0.00	2.53	0.71
Plecoptera	0.00	0.04	0.10	0.00	0.00	0.00	4.05	0.00	0.00	0.19	0.00	2.18	0.06	0.49	0.54	6.79	0.00	0.00	0.05	0.04	0.69	0.04	0.00
Odonata	0.00	0.40	0.11	0.00	0.00	0.24	0.00	0.00	0.00	0.00	3.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Megaloptera	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.00	0.35	0.00	0.00	0.00	0.00
Diptera	0.08	2.12	0.32	0.00	0.33	0.37	1.32	0.00	0.06	0.53	0.05	0.00	0.24	0.78	1.21	4.08	0.20	0.43	0.15	1.52	8.16	0.00	0.00
Hemiptera	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.82	0.00	1.04	0.39	0.00	0.00	0.00	0.28	0.00	0.07	0.05	0.00
Coleoptera	0.07	0.00	0.00	0.12	0.04	0.14	0.34	0.00	0.09	0.07	0.43	0.00	0.27	0.86	0.28	0.00	0.00	0.00	0.00	0.05	0.00	0.98	0.56
Lepidoptera	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.00	13.14	0.94	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.84
Hymenoptera	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.28	0.00	0.00	0.48	0.00	0.00	0.00	0.94	1.53	0.00	0.00	0.00	0.00	0.15	0.00
Orthoptera	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51	0.00	0.00	0.04
Tricladida	2.58	0.82	0.95	0.46	0.00	0.00	0.00	0.00	10.21	12.96	0.78	26.15	2.71	33.82	7.33	0.00	0.35	0.32	1.12	1.36	1.94	1.60	1.54
Gordioida	1.14	0.44	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.00	1.99	0.00	0.00	0.61	0.34	0.00	0.00	0.00
Bivalvia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45
Gastropoda	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.86
Oligochaeta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Isopoda	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Araneae	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.47	0.00	0.00	1.17
Opiliones	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Salmoniformes	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	2.00	0.87	0.23	0.00	0.00	1.23	0.00	0.00	0.00	0.00	0.00	0.46	0.00
Caudata	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table S10. Depth of coverage (X) values for each detected Parasite.

Parasite	BC2042	BC2045	BC2280	BC2360	BC2378	BC2381	BC2408	BC2416	BC2765	BC2856	BC2865	BC2876	BC2934	BC2989	BC3015	BC3267	BC3343	BC3345	BC3876	BC3895	BC3899	BC3903	BC3906
Coccidia	2.59	3.31	1.99	8.83	4.32	1.81	10.81	1.61	6.51	4.46	3.37	3.00	2.57	6.22	2.50	3.82	7.86	4.59	1.10	4.94	4.74	5.49	8.46
Cestoda	0.00	0.00	0.00	0.00	0.00	0.00	8.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mermithida	1.34	0.00	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Rhabditida	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Trombidiformes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.58	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.44
Mesostigmata	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sarcoptiformes	0.00	0.00	0.00	0.00	0.05	1.93	0.00	0.00	0.00	0.00	0.00	0.00	1.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table S11. Values of number of reads and breadth of coverage (%) obtained from aligning reads of each sample with each genome of Yersinia species using the Bowtie2 end-to-end aligner.

Yersinia Number of Reads	BC2042	BC2045	BC2280	BC2360	BC2378	BC2381	BC2408	BC2416	BC2765	BC2856	BC2865	BC2876	BC2934	BC2989	BC3015	BC3267	BC3343	BC3345	BC3876	BC3895	BC3899	BC3903	BC3906
Y. aldovae	86	100	30	22	104	262	288	6	230	10	44	468	308	164	224	126	58	13790	338	32	272	12378	76
Y. aleksiciae	278	4048	1470	1120	1718	7118	2900	60	2498	46	1502	2788	12538	1792	2490	92	266	20016	6566	148	1462	624	384
Y. alsatica	50	90	42	20	102	70	228	2	322	8	44	558	332	88	1188	10	24	5636	404	2	250	406	92
Y. artesia	298	4600	1664	1256	1970	7798	3200	68	2866	56	1676	3090	14842	1958	2822	98	286	19082	7702	150	1610	868	426
Y. bercovieri	282	4308	1542	1204	1826	7528	3110	62	2562	50	1546	2826	13418	1782	2610	98	268	19044	6850	148	1516	588	364
Y. canariae	24	34	70	10	86	76	220	2	252	4	30	492	110	56	244	8	30	5700	360	6	216	390	84
Y. enterocolitica	38	84	82	20	128	108	256	4	336	6	64	604	1364	40	316	14	34	6352	456	8	326	976	116
Y. entomophaga	260	110	50	18	164	446	188	0	132	4	50	202	446	88	212	112	28	2494	16966	0	100	122	38
Y. frederiksenii	36	126	92	32	122	138	262	12	334	6	32	514	454	92	332	18	22	6310	442	6	266	386	88
Y. hibernica	24	284	50	30	86	68	186	2	226	4	36	378	300	58	208	12	32	5876	292	4	196	324	70
Y. intermedia	832	214	148	164	4724	1166	9908	14	22870	24	562	32400	314	120	10780	76	878	29610	9128	456	17292	904	7162
Y. kristensenii	40	44	20	12	64	80	218	6	234	4	24	428	138	44	206	72	26	8670	410	2	208	414	84
Y. massiliensis	286	4582	1706	1318	1944	7540	3188	62	2588	54	1586	2872	14250	1856	2702	100	278	19978	7410	156	1518	508	368
Y. mollaretii	28	64	30	8	92	112	222	6	278	6	24	488	210	54	218	10	32	4878	330	2	234	326	94
Y. nurmii	304	5234	1842	1368	1916	7864	2946	62	2550	50	1640	2546	16304	2078	2564	102	264	14782	9348	146	1478	316	334
Y. pekkanenii	300	4324	1544	1178	1810	7500	2862	64	2538	50	1632	2774	13532	2006	2584	102	256	19672	6938	142	1494	760	374
Y. pestis	28	74	38	14	66	58	174	4	122	4	22	184	226	44	126	6	24	4816	204	0	138	200	42
Y. proxima	290	4754	1610	1284	1894	7750	3192	64	2668	56	1592	3020	14136	1908	3172	100	282	19378	7222	154	1606	990	412
Y. pseudotuberculosis	26	62	38	20	76	78	132	6	122	4	32	202	1020	40	122	6	24	6172	306	0	146	202	46
Y. rochesterensis	26	86	50	28	90	106	216	6	266	2	26	454	246	34	200	10	22	5282	318	2	222	386	82
Y. rohdei	24	102	36	30	92	74	184	6	316	4	28	460	370	48	248	12	28	4880	380	6	244	412	90
Y. ruckeri	92	110	26	10	100	74	154	2	78	4	20	214	224	62	154	20	78	3998	1650	2	90	138	40
Y. similis	34	48	12	4	66	44	218	4	146	4	28	238	94	50	132	4	16	4686	192	0	128	218	38
Y. thracica	276	3784	1406	1086	1696	7498	2876	60	2360	42	1610	2752	12316	1786	2502	94	268	19714	6402	138	1578	660	352
Y. vastinensis	308	5076	1772	1332	1978	7840	3090	58	2810	56	1778	2986	15352	2104	2724	96	284	18480	7754	156	1608	532	382
Y. wautersii	182	2970	1040	898	1308	5658	1754	48	2142	38	932	1508	9926	1258	1452	62	150	13868	4812	82	992	418	218

Table S11 (cont.)

<i>Yersinia</i> Breadth of coverage	BC2042	BC2045	BC2280	BC2360	BC2378	BC2381	BC2408	BC2416	BC2765	BC2856	BC2865	BC2876	BC2934	BC2989	BC3015	BC3267	BC3343	BC3345	BC3876	BC3895	BC3899	BC3903	BC3906
<i>Y. albiovae</i>	0.19	0.1	0.04	0.05	0.16	0.17	0.52	0.01	0.37	0.02	0.08	0.77	0.26	0.34	0.37	0.27	0.13	1.5	0.49	0.08	0.44	19.03	0.16
<i>Y. aleksici</i>	0.13	0.16	0.14	0.11	0.25	0.22	0.48	0.06	0.52	0.06	0.13	0.85	0.26	0.18	0.45	0.09	0.15	0.96	0.5	0.08	0.47	0.66	0.26
<i>Y. alsatica</i>	0.1	0.07	0.05	0.03	0.16	0.09	0.34	0	0.46	0.01	0.03	0.87	0.14	0.1	2.25	0.02	0.05	0.88	0.42	0	0.38	0.56	0.18
<i>Y. artesisana</i>	0.17	0.19	0.12	0.12	0.28	0.41	0.5	0.07	0.6	0.06	0.14	0.88	0.54	0.17	0.63	0.11	0.15	1.12	0.76	0.09	0.52	1.01	0.28
<i>Y. bercovieri</i>	0.14	0.15	0.13	0.12	0.25	0.22	0.51	0.06	0.56	0.06	0.13	0.82	0.22	0.17	0.46	0.1	0.13	0.95	0.49	0.08	0.49	0.61	0.26
<i>Y. canariae</i>	0.05	0.05	0.04	0.01	0.13	0.12	0.32	0	0.4	0.01	0.04	0.78	0.12	0.06	0.39	0.02	0.07	0.94	0.5	0.01	0.36	0.55	0.17
<i>Y. enterocolitica</i>	0.08	0.07	0.12	0.03	0.19	0.14	0.39	0.01	0.52	0.02	0.07	0.91	2.34	0.06	0.53	0.03	0.08	1.19	0.48	0.02	0.51	1.54	0.23
<i>Y. entomophaga</i>	0.67	0.09	0.06	0.02	0.28	1	0.24	0	0.17	0.01	0.05	0.32	0.14	0.11	0.34	0.27	0.06	0.4	30.63	0	0.14	0.2	0.08
<i>Y. frederiksenii</i>	0.06	0.07	0.06	0.04	0.17	0.15	0.34	0.02	0.45	0.01	0.03	0.73	0.15	0.1	0.46	0.04	0.04	0.96	0.37	0.01	0.4	0.52	0.16
<i>Y. hibernica</i>	0.04	0.13	0.04	0.05	0.15	0.1	0.28	0	0.34	0.01	0.05	0.57	0.19	0.07	0.33	0.03	0.07	0.8	0.35	0.01	0.31	0.46	0.14
<i>Y. intermedia</i>	1.8	0.34	0.27	0.4	8.47	2.44	18.36	0.02	34.52	0.05	0.99	45.38	0.39	0.22	18.78	0.16	1.79	37.85	15.81	1.03	27.54	1.47	13.69
<i>Y. kristensenii</i>	0.08	0.04	0.03	0.03	0.1	0.13	0.31	0.01	0.37	0.01	0.04	0.72	0.13	0.08	0.35	0.05	0.06	0.8	0.47	0	0.33	0.57	0.18
<i>Y. massiliensis</i>	0.12	0.13	0.1	0.09	0.18	0.15	0.32	0.05	0.33	0.05	0.11	0.56	0.18	0.13	0.31	0.09	0.12	0.73	0.36	0.06	0.31	0.43	0.19
<i>Y. mollaretii</i>	0.06	0.07	0.06	0.02	0.14	0.12	0.36	0.01	0.45	0.01	0.04	0.82	0.13	0.09	0.37	0.02	0.07	0.87	0.38	0	0.4	0.48	0.2
<i>Y. nurmii</i>	0.19	0.18	0.15	0.11	0.2	0.28	0.27	0.06	0.23	0.05	0.13	0.35	0.2	0.2	0.29	0.12	0.13	0.48	3.35	0.07	0.22	0.28	0.15
<i>Y. pekkannenii</i>	0.16	0.13	0.1	0.09	0.19	0.18	0.41	0.06	0.49	0.05	0.11	0.84	0.22	0.15	0.4	0.08	0.11	0.98	0.4	0.07	0.44	0.7	0.25
<i>Y. pestis</i>	0.06	0.06	0.03	0.02	0.09	0.08	0.21	0.01	0.18	0.01	0.03	0.3	0.1	0.07	0.18	0.01	0.05	0.55	0.21	0	0.19	0.3	0.08
<i>Y. proxima</i>	0.15	0.16	0.12	0.15	0.24	0.21	0.43	0.06	0.53	0.06	0.14	0.88	0.48	0.17	1.23	0.1	0.15	1.09	0.55	0.07	0.47	1.23	0.27
<i>Y. pseudotuberculosis</i>	0.05	0.05	0.03	0.03	0.1	0.09	0.18	0.01	0.18	0.01	0.04	0.31	0.71	0.06	0.18	0.01	0.05	0.57	0.23	0	0.19	0.29	0.08
<i>Y. rochesterensis</i>	0.06	0.06	0.04	0.04	0.13	0.17	0.35	0.01	0.42	0	0.04	0.78	0.17	0.06	0.33	0.02	0.05	0.88	0.43	0	0.39	0.61	0.18
<i>Y. rohdei</i>	0.05	0.07	0.04	0.04	0.13	0.12	0.31	0.01	0.45	0.01	0.04	0.75	0.16	0.08	0.41	0.02	0.07	0.93	0.43	0.02	0.41	0.7	0.19
<i>Y. ruckeri</i>	0.25	0.08	0.03	0.01	0.17	0.14	0.21	0	0.13	0.01	0.04	0.41	0.15	0.11	0.26	0.05	0.21	0.59	3.64	0	0.17	0.23	0.09
<i>Y. similis</i>	0.07	0.04	0.02	0.01	0.1	0.07	0.26	0.01	0.2	0.01	0.04	0.32	0.08	0.05	0.18	0.01	0.04	0.54	0.2	0	0.19	0.28	0.07
<i>Y. thracica</i>	0.15	0.14	0.11	0.11	0.22	0.24	0.47	0.06	0.54	0.05	0.11	0.83	0.2	0.18	0.49	0.09	0.16	1.03	0.56	0.08	0.58	0.7	0.26
<i>Y. vastinensis</i>	0.15	0.15	0.14	0.12	0.26	0.23	0.44	0.05	0.57	0.06	0.14	0.84	0.22	0.19	0.47	0.1	0.15	0.97	0.53	0.07	0.46	0.56	0.26
<i>Y. wautersii</i>	0.13	0.15	0.12	0.12	0.21	0.28	0.36	0.04	0.44	0.04	0.1	0.42	0.2	0.14	0.27	0.06	0.1	0.7	0.44	0.05	0.27	0.39	0.15

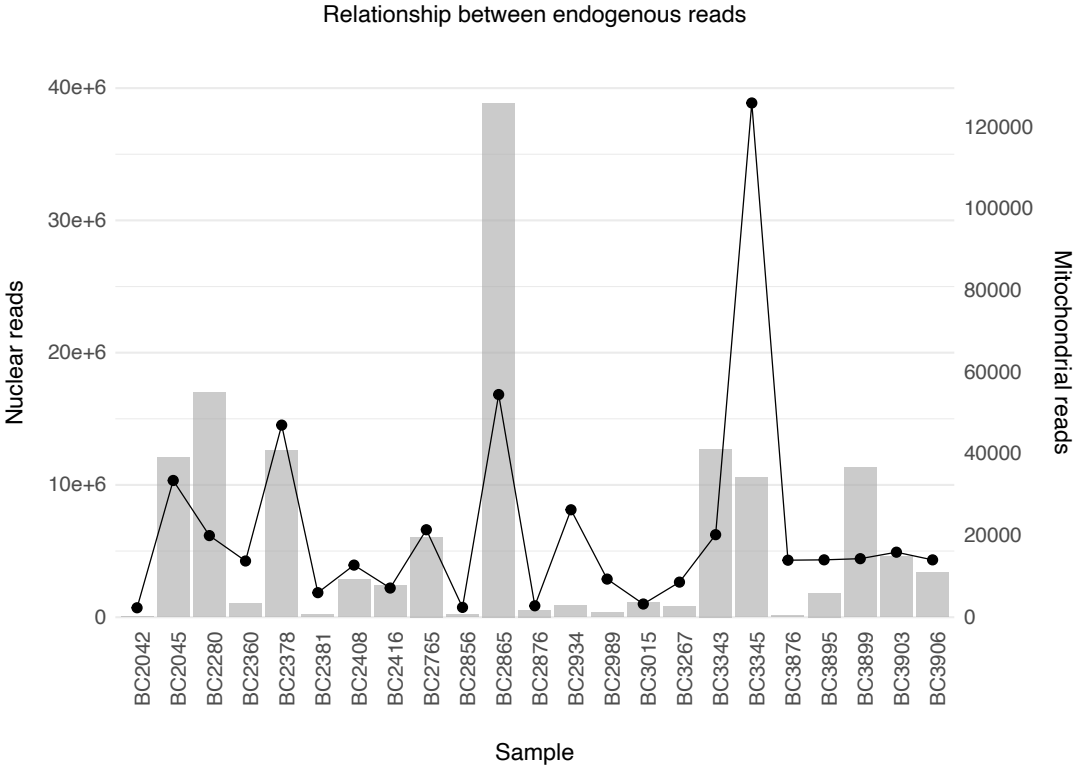


Figure S1. Number of reads from each sample aligned to the nuclear and mitochondrial genome of *G. pyrenaicus*. Bars represent nuclear genome reads, whereas mitochondrial genome reads are shown as points connected by a line.

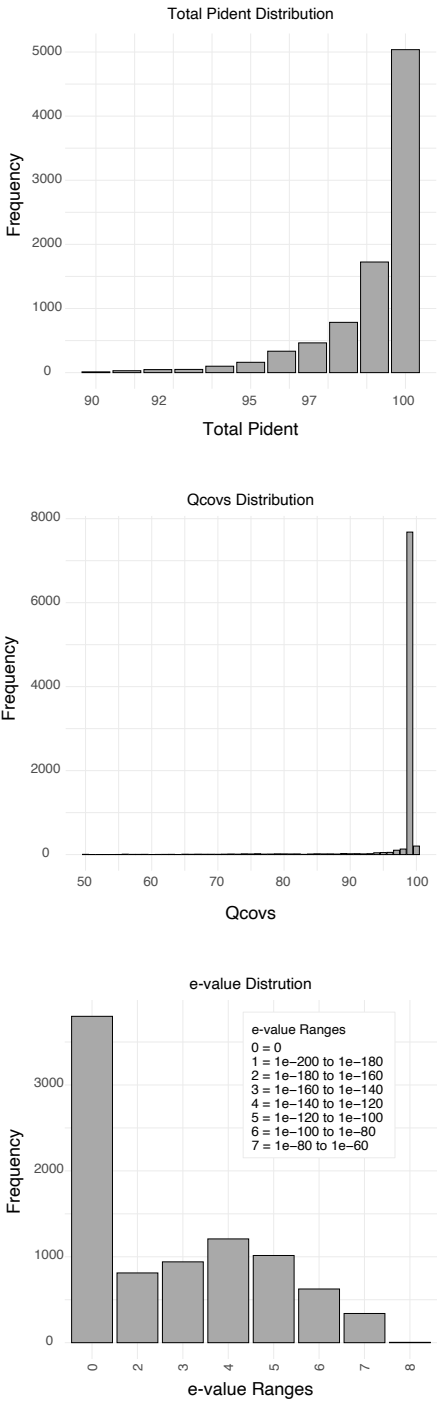


Figure S2. Distribution of parameter values: Total percentage of identity (Total Pident), query coverage (Qcovs), and e-value for the consensus sequences.

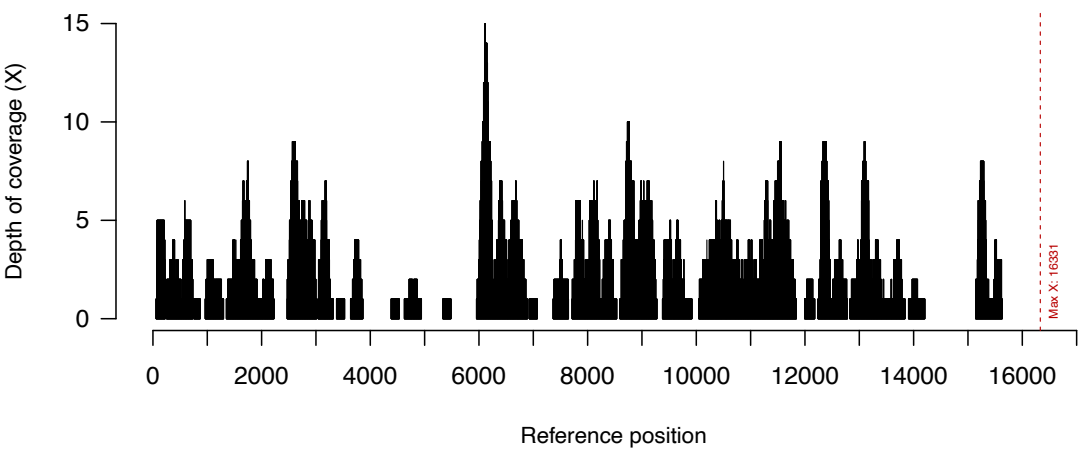


Figure S3. Depth of coverage across the complete mitogenome of the consensus sequence assigned to *Salamandra salamandra* in the sample BC2876 (Max X (x-axis): reference sequence length).

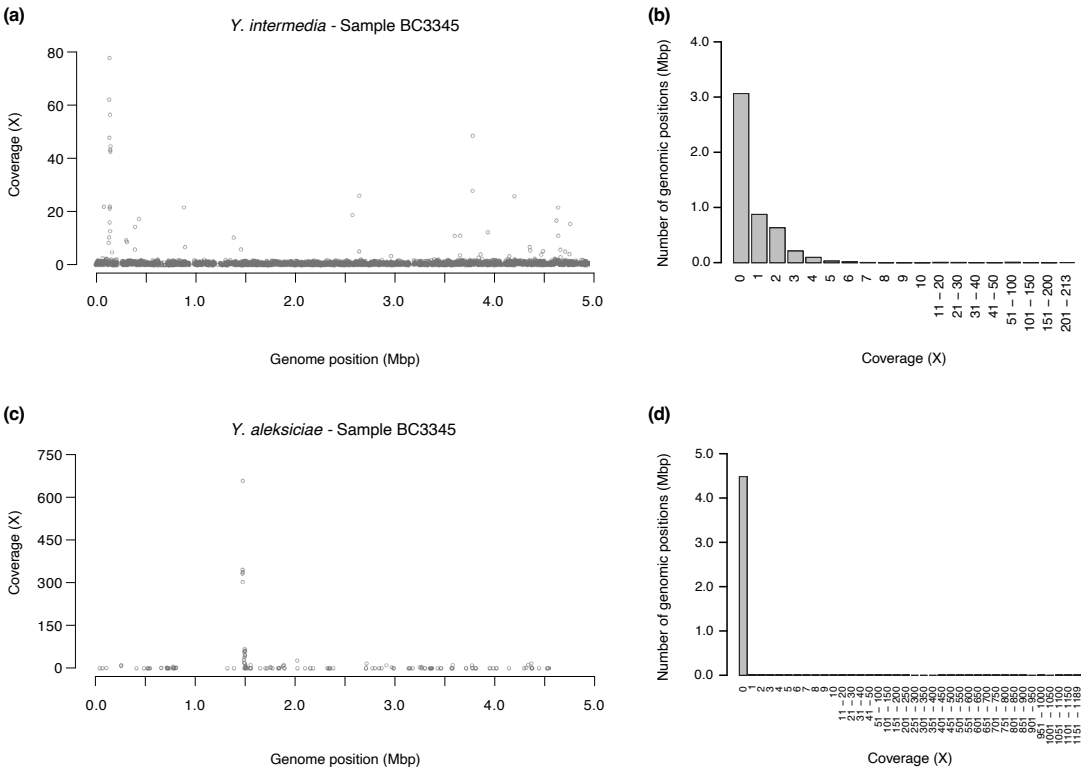


Figure S4. Comparison of depth of coverage across the reference genomes of *Y. intermedia* and *Y. aleksiciae* and the corresponding depth of coverage barplots for sample BC3345. (a) Coverage across the reference genome of *Y. intermedia* in 4,000 windows and (b) coverage barplot, in which 29,610 reads were assigned, resulting in a mean depth coverage of 0.90X and a breadth of coverage of 37.85%. (c) Coverage across the reference genome of *Y. aleksiciae* and (d) coverage barplot, in which 20,016 reads were assigned, resulting in a mean depth coverage of 0.66X and a breadth of coverage of 0.96%. In (a) and (c), windows with zero coverage are not plotted. Note the different coverage scales in (a) and (c).

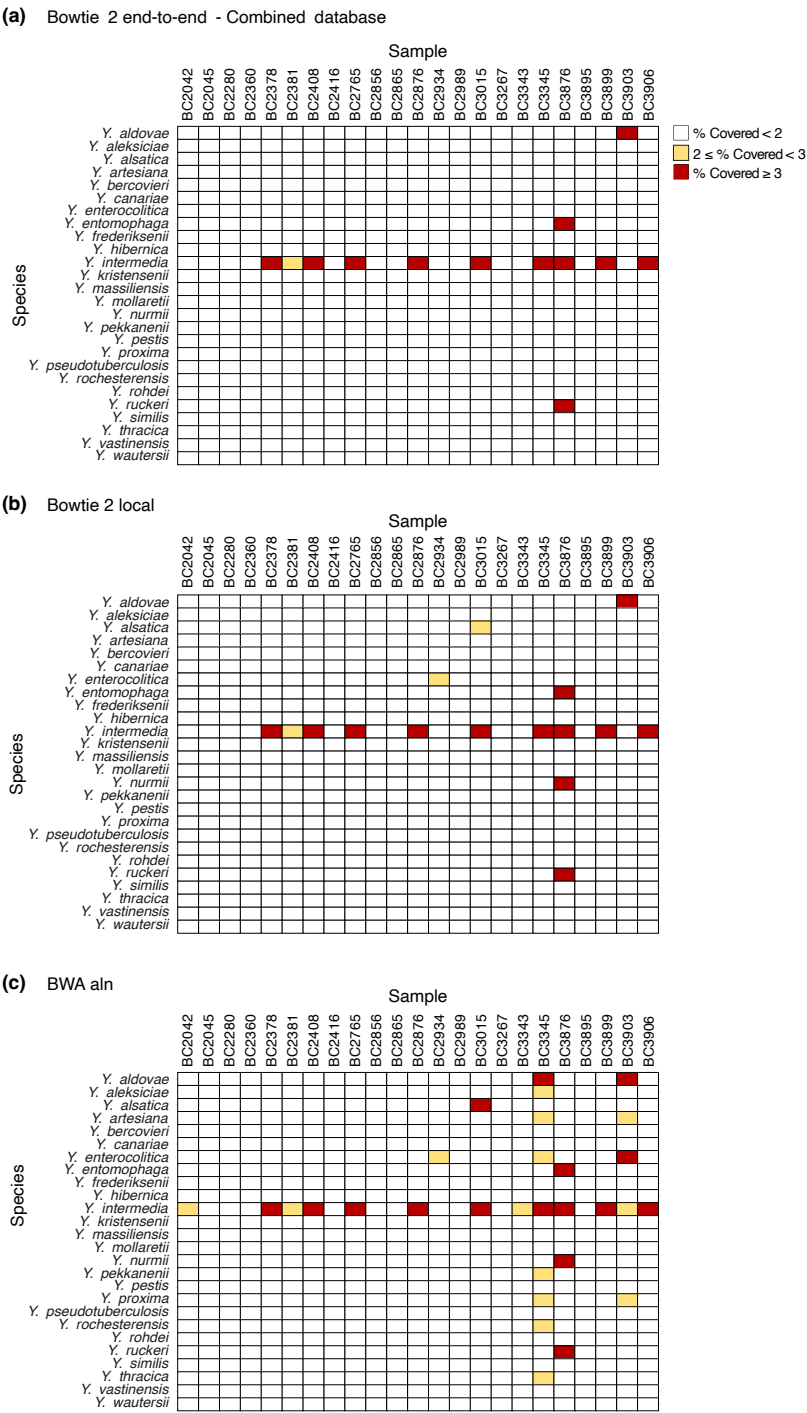


Figure S5. Identification of *Yersinia* species in 23 samples of Iberian desman using different approaches. (a) Identification with a combined database for all species using Bowtie2 in end-to-end mode. (b) Identification with single-species databases using Bowtie2 in default local mode. (c) Identification with single-species databases using BWA-aln with default parameters.

DISSEMINATION OF RESULTS

The results corresponding to the analysis of endogenous DNA, the global DNA composition, and the pipeline for the analysis of diet and parasites were presented at the Society for Molecular Biology and Evolution Meeting 2023, as detailed below. These results are currently compiled into a manuscript titled "Methodological advances in the analysis of the metagenome of Iberian desman feces: monitoring population genomics, ecology, and health", authored by Román Sapino, Ángel Fernández-González, and José Castresana.

Román Sapino, Lídia Escoda, Ángel Fernández-González and José Castresana. Methodological advances in the analysis of the metagenome of Iberian desman feces. Society for Molecular Biology and Evolution Meeting (SMBE23). July 23–27, 2023. Ferrara, Italy. Format: Poster presentation.

The results related to the development of the bioinformatic pipeline for the identification of pathogenic bacteria were submitted for publication to a peer-reviewed journal and deposited in a digital repository:

Sapino, R., Fernández-González, A., and Castresana, J. (2025). Development of metagenomic methods for non-invasive health monitoring of endangered species: Unveiling hidden microbial threats in fecal samples. BioRxiv. <https://doi.org/10.1101/2025.01.21.633432>

