Genomics 117 (2025) 111108

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Research Paper

Unmapped reads from whole-genome sequencing data reveal pathogen diversity in European and African cattle breeds

Daniil Ruvinskiy [a,b], Kisun Pokharel [a,*], Rodney Okwasiimire [a,b], Rayner Gonzalez-Prendes [c], Catarina Ginja [d,e], Nasser Ghanem [f], Donald R. Kugonza [g], Mahlako L. Makgahlela [h,j], Heli Lindeberg [i], Melak Weldenegodguad [a], Juha Kantanen [a], Martijn Derks [c], Richard P.M.A. Crooijmans [c,*]

[a] Natural Resources Institute Finland (Luke), Tietotie 4, FI-31600 Jokioinen, Finland
[b] Department of Agricultural Sciences, University of Helsinki, Finland
[c] Animal Breeding and Genomics, Wageningen University & Research, Wageningen, the Netherlands
[d] lBIOPOLIS, Program in Genomics, Biodiversity and Land Planning, University of Porto, CIBIO, Vairão Campus, 4485-661 Vairão, Portugal
[e] CIISA, Centre for Interdisciplinary Research in Animal Health, Faculty of Veterinary Medicine, University of Lisbon, 1300-477 Lisbon, Portugal
[f] Animal Production Department, Faculty of Agriculture, Cairo University, Giza, Egypt
[g] Department of Animal and Range Sciences, College of Agricultural and Environmental Sciences, Makerere University, Kampala, Uganda
[h] Agricultural Research Council-Animal Production, Irene, South Africa
[i] Natural Resources Institute Finland (Luke), Halolantie 31A, FI-71750 Maaninka, Finland
[j] Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein, South Africa

ARTICLE INFO

Keywords:
Whole-genome sequencing
Unmapped reads
de novo assembly
Pathogens

ABSTRACT

Climate change is impacting the global spread of infectious diseases, altering pathogen distribution and transmission, and threatening human and animal health. This study investigates the presence of potential pathogens in blood within unmapped reads obtained from whole-genome sequencing (WGS) data of various cattle breeds across geographically diverse regions, including South Africa, Uganda, Egypt, Portugal, The Netherlands, and Finland. Unmapped reads were extracted, assembled into contigs, and subjected to taxonomic analysis based on an extensive literature search. The analysis revealed significant geographic variation in pathogen composition, with breeds in the Southern Hemisphere (Uganda, Egypt, and South Africa) showing higher pathogen alignment counts while Northern breeds (particularly from Finland) exhibited lower diversity and counts. Portugal, representing a transition zone, exhibited a higher burden of parasites and tick-borne related pathogens than their Northern counterparts, which were also prevalent in Southern Hemisphere breeds such as *Theileria parva*, *Anaplasma platys*, *Theileria orientalis*, and *Babesia bigemina,* which is in line with the known capacity of these breeds to cope with local pathogens. Dutch breeds were found to harbor *Escherichia coli O157*, a known public health concern. The study provided key insights into emerging disease risks influenced by climate change and livestock management practices, but also on the need to investigate possible adaptive responses underlying disease resistance in some breeds. This study highlights the potential for climate-driven variations in disease ecology and transmission, emphasizing the need for integrating genomic and environmental data, and is currently the most comprehensive study to date investigating the microbial diversity present in unmapped reads obtained from WGS data of cattle populations.

1. Introduction

Livestock production is a cornerstone of the agricultural sector in many parts of the world, providing essential resources such as meat, milk, and fiber, while also contributing to livelihoods, particularly in rural communities. This is especially true in regions such as sub-Saharan Africa, where livestock plays a critical role in both local economies and food security [1]. However, cattle are susceptible to a range of infectious diseases, many of which are zoonotic, meaning they can also affect humans. The risk of zoonotic diseases is expected to increase in the

* Corresponding authors.
E-mail addresses: kisun.pokharel@luke.fi (K. Pokharel), richard.crooijmans@wur.nl (R.P.M.A. Crooijmans).

https://doi.org/10.1016/j.ygeno.2025.111108
Received 4 April 2025; Received in revised form 25 August 2025; Accepted 9 September 2025
Available online 10 September 2025

context of climate change, with important implications for public health [2]. Changes in temperature, precipitation, and habitat suitability directly influence the ecology and behavior of vectors, reservoirs, hosts, and pathogens, leading to shifts in disease prevalence, distribution, and seasonality [3–5]. For instance, vector-borne diseases, including those caused by ticks, are especially sensitive to climatic conditions. As temperatures rise, the geographic range of these vectors is expected to expand, increasing the risk of disease transmission in new areas [6]. Consequently, infectious diseases can have negative effects on livestock productivity, leading to reduced yields, trade restrictions, and increased veterinary costs [7]. The ongoing changes in global climate patterns are expected to exacerbate these challenges by altering the distribution, abundance, and transmission dynamics of pathogens.

Vector-borne and pathogen-induced diseases in cattle are a major constraint on livestock production across Africa. These diseases can lead to reduced milk and meat production, increased mortality and morbidity, elevated costs due to veterinary care (e.g., tick control), vaccination programs, as well as trade restrictions and compromised food security [8].

Similarly, in Europe, cattle farming is a crucial component of agricultural systems. For instance, cattle production not only supports rural economies but also plays an essential role in the conservation of certain breeds that are adapted to local environmental conditions. Breeds such as the Barrosã are well-adapted to the mountainous terrain and semi-extensive farming systems practiced in Northern Portugal. These breeds are integral to the rural economy, providing high-quality meat products and maintaining traditional grazing practices that contribute to biodiversity conservation. However, like many other regions, Southern Europe is expected to experience more pronounced temperature increases, leading to longer periods of vector activity and an expansion of suitable habitats for disease vectors [9]. Commercial breeds of cattle such as the Holstein are more susceptible to non-native environmental conditions and local pathogens [7].

With the advent of high-throughput sequencing (HTS) technologies, it is now possible to investigate the microbial diversity present in biological samples. HTS allows researchers to sequence vast amounts of genetic material from environmental or biological samples, providing insights into the presence of pathogens, even those that are not well represented in traditional reference databases [10]. One of the aspects of HTS is its resultant “unmapped reads”- those sequences that do not align to the reference genome being used (i.e. to the target species), meaning that they may be representative of other organisms such as pathogens [11]. By focusing on the analysis of unmapped reads, researchers can uncover a broader range of microbial diversity and identify potential pathogens that could be driving disease dynamics in livestock populations. This is particularly important in the context of climate change, as it allows for the detection of emerging or shifting pathogen populations that may be influenced by environmental factors. Integrating genomic data with environmental and geographic information is a critical next step in understanding how climate-sensitive pathogens might spread and establish in new geographic areas [12]. A few studies have explored the potential of these unmapped reads generated from whole genome sequencing (WGS) of blood, tissue and sperm (e.g., [13,57]) to uncover viral and bacterial pathogens, providing valuable information on the microbial diversity of both environmental and biological samples. A study in Black Pied cattle found bovine parvovirus 3 and *Mycoplasma* species [13]. Outside of cattle, unmapped DNA and RNA reads of songbirds have been analyzed to find a variety of pathogenic species, including *Plasmodium* and *Trypanosoma* [14].

We focused on the analysis of unmapped reads in blood taken from the tail region and the jugular vein of a range of cattle breeds across the Northern and Southern hemispheres. Our aim was to investigate global pathogen trends and patterns, and the disease backgrounds of native cattle on a north-south transect. Unmapped reads were used to infer the pathogen content using a combination of custom alignment filtering and k-mer count analysis. This is a novel and very much first-of-its-kind foray into the potential of unmapped reads as a method of understanding and studying diseases of cattle and the occurrence of pathogens globally.

2. Materials and methods

2.1. Data acquisition and preprocessing

Whole-genome sequencing (WGS) data were obtained from the LEAP-Agri-project OPTIBOV (ENA Accessions PRJEB90914, PRJEB90816, and PRJEB76602; supplementary file 1) on the Genetic characterization of cattle populations for optimized performance in African ecosystems [15]. OPTIBOV aimed to characterize native cattle breeds adapted to various ecosystems in Europe and Africa by sequencing the whole genomes of 27 native breeds from Europe (Finland, the Netherlands, Portugal) and Africa (Egypt, Uganda, South Africa; Fig. 1). Commercial Holstein Friesian samples from each of the six countries were also included for comparison. Altogether 552 samples representing 23 breeds in 6 countries and 1 breed (Holstein) present in every country have been included in our study (Fig. 1).

The blood samples of Finnish, Dutch, South African, Egyptian, Ugandan, and Portuguese cattle (including the Holstein) were collected from the jugular vein using 9 ml vacuum collection tubes containing K3-EDTA as an anticoagulant (Greiner Bio-one reference n. 455,036).

Total DNA was extracted following routine procedures common to all partners of the OPTIBOV consortium [16], and the genomic libraries were sequenced (paired-end, 150 bp) using the Illumina Novaseq 6000 platform yielding approximately 10× coverage [16]. WGS data were aligned to the ARS-UCD1.2 bovine reference genome and all the sequences that did not align to the reference genome (i.e., unmapped reads) were analyzed in this study. Unmapped reads are often indicative of unknown or unexpected microbial content, including potential pathogens.

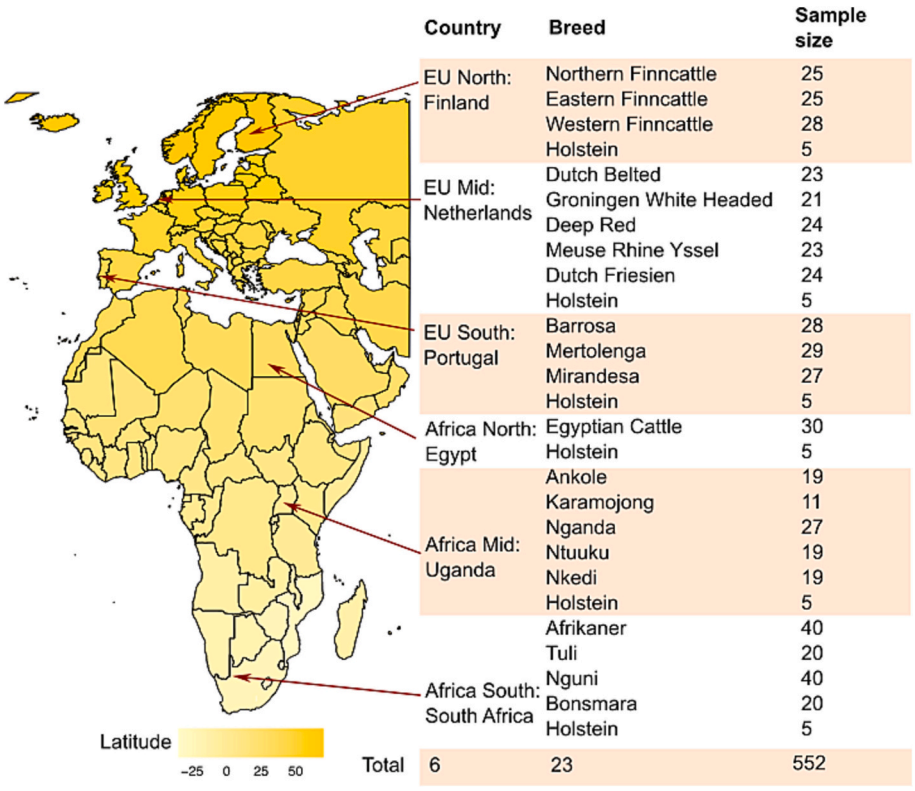
Unmapped reads were extracted from the BAM files using the Samtools (version 1.16) software [17]. The samtools view -f 4 command was used to extract all unmapped reads from each BAM file. These reads were subsequently converted to a FASTQ format using the samtools bam2fq command, ensuring compatibility with the downstream analysis pipeline.

2.2. De novo assembly of unmapped reads

The unmapped reads in FASTQ format were assembled into contiguous sequences (contigs) using SPAdes (version 3.13.0)[18], a *de novo* assembler optimized for short-read data [18]. SPAdes employs de Bruijn graph algorithms to assemble overlapping reads into contigs, allowing us to identify novel microbial sequences not represented in the reference genome. Assembly was performed using the default parameters, and the resulting contigs were subjected to taxonomic and pathogen classification (see details below, section 2.4). Pathogens are defined by the Food and Agriculture Organization of The United Nations (FAO) as microorganisms causing disease, including viruses, bacteria, fungi, and helminths [19]. Contigs were further filtered to retain only sequences with >95 % for both percentage identity and coverage and a minimum length of 500 bp.

2.3. Principal components analysis of unmapped reads

We performed a principal component analysis (PCA) on assembled contigs derived from unmapped reads to investigate the genetic variability across cattle breeds. Contigs were constructed using SPAdes [18] and k-mer frequencies were computed with $k = 9$. This k-mer size was chosen as it balances sensitivity and specificity for genomic sequence analysis, particularly for viral pathogen sequences. A k-mer size of 9 provides the resolution necessary to detect patterns in longer genomic regions typical of pathogens, as shorter k-mers may lack sufficient



Country	Breed	Sample size
EU North: Finland	Northern Finncattle	25
	Eastern Finncattle	25
	Western Finncattle	28
	Holstein	5
EU Mid: Netherlands	Dutch Belted	23
	Groningen White Headed	21
	Deep Red	24
	Meuse Rhine Yssel	23
	Dutch Friesien	24
	Holstein	5
EU South: Portugal	Barrosa	28
	Mertolenga	29
	Mirandesa	27
	Holstein	5
Africa North: Egypt	Egyptian Cattle	30
	Holstein	5
Africa Mid: Uganda	Ankole	19
	Karamojong	11
	Nganda	27
	Ntuuku	19
	Nkedi	19
	Holstein	5
Africa South: South Africa	Afrikaner	40
	Tuli	20
	Nguni	40
	Bonsmara	20
	Holstein	5
Total	6	23	552

Fig. 1. List of samples used and respective geographic locations. Samples included in the study represent 23 cattle breeds from six countries (Finland, the Netherlands, Portugal, Egypt, Uganda, and South Africa) across two continents (Europe and Africa).

discriminatory power, while larger k-mers can lead to a far too sparse dataset, reducing potential signal detection. We selected k = 9 for our k-mer analysis as it represents an optimal. According to Zhang et al. [20], k = 9 was found to be a suitable feature length in large-scale genomic analyses, where cumulative relative entropy (CRE) and other metrics indicated that smaller k-mer values capture sufficient diversity while maintaining computational feasibility for highly diverse datasets [20]. This value ensures robust phylogenomic comparisons while minimizing artifacts introduced by excessive feature lengths. To analyze genetic differences among breeds and countries, k-mers of size 9 were generated from unmapped reads using a sliding window approach. FASTA files containing the filtered contigs were processed to calculate k-mer frequencies for each contig. A sparse matrix representation of k-mer counts was generated using Python scripts and stored in .npz format. Low-frequency k-mers in fewer than 10 sequences were excluded to reduce noise and computational complexity. This pre-processing step resulted in a set of k-mer matrices, one for each breed, with each matrix capturing the distribution of filtered k-mers across the contigs. 4^k possible k-mers (262,144 for k = 9) provide an effective compromise between capturing meaningful biological features and maintaining computational feasibility. K-mer frequency matrices were generated for each contig using a custom Python script employing the scipy and pandas libraries. The high dimensionality of the k-mer space necessitated dimensionality reduction through PCA, implemented using the scikit-learn library. Sparse matrices were used to optimize memory usage during the computation. The first principal components, explaining the majority of variance, were used for clustering and visualization of contigs by breed, highlighting genetic relationships and potential pathogen-driven variations among global cattle populations.

Clustering was performed to explore genetic relationships at both the country and breed levels. K-means clustering was applied to the PCA-transformed data for 27 clusters (breeds including Holsteins presented separately on a country basis). The silhouette score was used to determine the optimal number of principal components. For 27 clusters, the optimal number was 30 principal components. These dimensionalities were subsequently used for all clustering and visualization analyses.

PCA scatter plots were generated to visualize clustering patterns for 27 clusters and PCA results with 30 components were visualized, highlighting breed-specific clusters. To further aid interpretation, an additional PCA visualization was performed where each breed was represented by a single point, derived from the mean PCA scores of all contigs belonging to the breed. To aid comprehension, country and breed names were overlaid on the plots to identify their positions within the PCA space.

2.4. Taxonomic analysis using KronaCharts and BLAST

The assembled contigs were subjected to taxonomic analysis to identify potential pathogens. First, BLAST (Basic Local Alignment Search Tool) searches [21] were conducted against the non-redundant NCBI nucleotide database to identify homologous sequences. The output from BLAST was then processed using KronaTools [22] to visualize the taxonomic composition of the sequences [23]. KronaCharts provides interactive, hierarchical visualizations of the microbial taxa present in each sample at various levels (e.g., kingdom, phylum, genus, species), allowing for detailed exploration of pathogen diversity. BLAST results were filtered using the pathogen database generated from the literature search, and contigs matching known pathogens of interest were prioritized for further analysis. The pathogen database and NCBI TaxIDs from TaxonKit [24] ensured that identified pathogens were relevant to the breeds and regions studied.

2.5. Pathogen database and literature search

To facilitate the identification of relevant pathogens, a comprehensive pathogen database was created based on a systematic literature search in NCBI (see supplementary file 2). The search focused on pathogens known to affect cattle in the regions studied, including both

endemic and emerging infectious agents. Searches were conducted across peer-reviewed literature using keywords related to cattle pathogens, specific breeds, and geographic regions. For example, for a literature search of Ugandan pathogens, the keywords "Ugandan cattle pathogens, disease" were used. Articles from major veterinary, microbiology, and zoonotic disease journals were included in this database, using 12 articles in total. Each identified pathogen was cross-referenced with available NCBI Taxonomy identifiers (TaxIDs) to ensure consistency and to facilitate downstream analysis. This database was then used to filter the assembled contigs for potential pathogens of interest.

2.6. Taxonomic resolution using TaxonKit

To ensure that pathogen identification was accurate and up to date, we used TaxonKit v. 0.19.0 [24], a command-line toolkit for taxonomic data manipulation. TaxonKit was employed to map scientific names of the pathogens identified in our literature search to their respective NCBI Taxonomy identifiers (TaxIDs). TaxonKit commands such as taxonkit name2taxid, were used to convert the names of identified pathogens into their corresponding TaxIDs as well as the TaxIDs of subspecies and variants of these pathogens. Additionally, where the literature did not provide a TaxID, TaxonKit's taxonkit list function was used to retrieve and verify the correct TaxIDs from the NCBI taxonomy database. Pathogen genera were used to get taxids of all potential related species of the pathogen. These identifiers were then used to match against the taxonomic output from the BLAST searches, ensuring that the identified pathogens were accurate and up to date with the latest taxonomic revisions. In total 89,785 potential pathogens were identified as a result.

2.7. Comparative pathogen profiling

The identified pathogens from the taxonomic and statistical analyses were further compared across regions to identify patterns of geographic variation using Python 3.10.6. Pathogen counts were correlated with environmental and climatic factors in each region to explore the potential influence of climate on pathogen diversity and prevalence. Additionally, the relative abundance of pathogens within each breed was examined to assess whether certain cattle breeds exhibited greater exposure to pathogens. The bioinformatics pipeline for pathogen sequence detection in unmapped reads is presented in Fig. 2 and scripts are available on Github at https://github.com/druvinskii/unmapped-reads/tree/main.

3. Results

3.1. Results from de novo assembly

The assembly results were analyzed to assess the quality and characteristics of contigs before and after filtering. Prior to analysis there were 1,327,848,784 reads in all unmapped read files with an average 249,127 reads per sample (Supplementary file 3). In total, SPAdes assemblies across all datasets generated 3,358,783 unfiltered contigs (supplementary file 4), with a cumulative length of 1,558,751,633 bp and an average GC content of 46.72 %. The N50 length of the unfiltered assemblies ranged from 233 bp to 4247 bp, while the largest contig reached a length of 962,072 bp, reflecting the assembly's ability to capture large genomic regions. Filtering significantly improved the quality of the assemblies by removing shorter (< 500 bp) and lower confidence (>95 % coverage and sequence identity) contigs. After filtering, the total number of contigs was reduced to 506,868, corresponding to a cumulative length of 873,470,670 bp. The N50 length increased to a range of 740 bp to 10,532 bp, indicating improved contiguity in the filtered datasets. Additionally, the GC content remained consistent at 46.42 %, further validating the accuracy of the assembly process. Among the datasets, the largest filtered contig measured 962,072 bp, and the number of high-quality contigs (≥1000 bp)

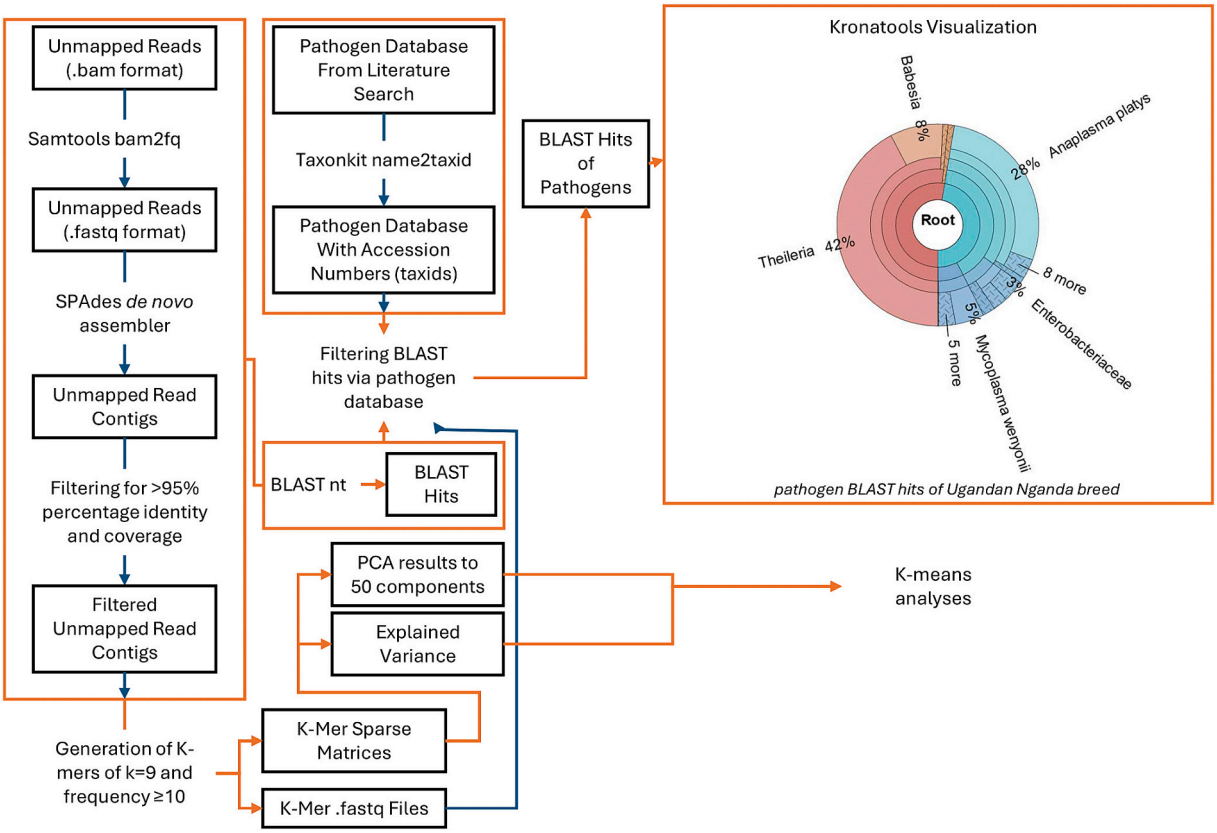


Fig. 2. Flowchart showing the pipeline used for analyzing unmapped reads.

increased significantly post-filtering.

3.2. Principal component analysis of unmapped reads

Principal Component Analysis (PCA) was performed on k-mer frequency matrices derived from assembled contigs of unmapped reads across cattle breeds. The PCA, calculated to 50 principal components, revealed that the variance was distributed across multiple components. This is consistent with the high dimensionality and complexity of a dataset of unmapped read contigs. The first principal component (PC1) explained 8.62 % of the total variance, while the first 10 principal components collectively accounted for 18.19 % of the variance. All 50 calculated PCs had a cumulative variance at ~22.35 % (supplementary file 5). This distribution of variance indicates a more complex set of data with no completely dominating trend. Despite this, some clustering patterns were evident when visualizing the breeds (supplementary Fig. 1) and countries in the PCA space. A scatterplot of the first two principal components (PC1 and PC2) revealed both clustering and overlap among countries (Fig. 3), reflecting genetic diversity and potential shared pathogen-related sequences. Specific breeds such as the Ugandan Nkedi formed distinct clusters, while others overlapped (Supplementary Fig. 1). This is indicative of shared genetic or environmental factors influencing pathogen presence. Most interestingly, Holstein samples appear across multiple clusters rather than forming just one group. The scattering of Holstein suggests that the unmapped k-mer profiles are influenced by geographic factors. This suggests that unmapped sequences are not simply a result of the host genome and are strongly affected by external factors that vary by geographic location.

3.3. Total pathogen counts across breeds and regions

A comparison of total pathogen alignment counts across cattle breeds [23] and regions revealed significant variation, indicating potentially differing levels of exposure among the studied populations. A variety of studies from respective countries in the Northern and Southern hemispheres were used to build a database for filtering pathogens (supplementary file 1). In total 89,785 potential pathogens were identified.

The total number of pathogens related sequences detected varied significantly across regions, with some regions showing much more alignments to pathogen related sequences compared to others. Uganda had the highest total pathogen count, with 527 pathogen related sequences identified across Ugandan cattle breeds (Table 1). South Africa, the Netherlands and Portugal also exhibited high pathogen counts, with 396, 236 and 119 pathogens detected, respectively. Finland and Egypt reported 59 and 58 total pathogen related sequences (Table 1, supplementary file 6).

At the breed level, several breeds showed greater pathogen counts, particularly in regions where environmental conditions are favourable for pathogen survival. The Nganda cattle from Uganda had the highest total pathogen count, with 121 pathogens detected (Table 1), suggesting that these cattle are exposed to a broad range of pathogens. In contrast, the Eastern Finncattle from Finland exhibited the lowest pathogen count, with only 11 pathogens detected (Table 1). Dutch Belted cattle from the Netherlands showed a moderate pathogen sequence presence, with 52 pathogens detected (Table 1), potentially reflecting the temperate climate in this region rather than controlled farming practices as it is a year-round grazing breed.

Overall, the breeds from southern regions such as South Africa (despite greater sample sizes for some breeds) and Uganda showed higher mean pathogen counts (often >3 per animal), and therefore, pathogen loads, likely due to environmental conditions that favor pathogen proliferation and transmission, while breeds from temperate regions exhibited lower pathogen counts. European breeds (Finland, Netherlands, Portugal) generally had low mean pathogen counts (<1 per animal), except Holstein in Finland (~2.6) and Netherlands Friesian (~2.3). Holsteins varied a lot depending on the country — in Africa they seem to carry more pathogens on average than in Europe, which may suggest that environmental exposure plays a big role. The absence of pathogens in Portuguese Holsteins is particularly notable. This could reflect enhanced biosecurity measures such as antiparasitic and antibiotics' treatments, or potential sampling or sequencing biases.

3.4. Pathogen diversity across breeds and regions

A wide variety of pathogens were identified across the cattle breeds studied, primarily consisting of bacteria and a few eukaryotic pathogens. Table 1 provides a summary of the unique pathogen counts for each breed, reflecting the diversity of pathogens identified in all samples in a breed.

In Uganda, indigenous breeds such as Nkedi (108 unique pathogens)

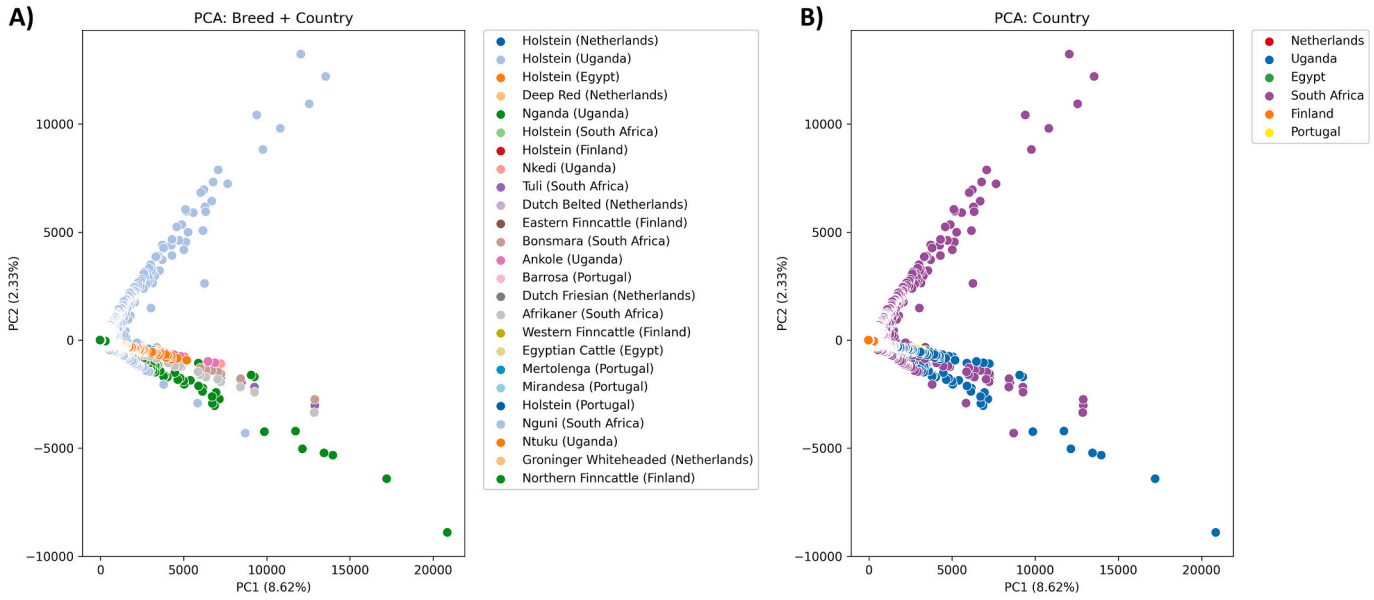


Fig. 3. Principal Component Analysis visualization of cattle breed k-mers of size 9. This visualization shows the clustering according to breed labelled on a country basis.

Table 1
Total number of blast hits post-filtering, and total and unique pathogen taxa counts for countries and breeds.

Countries	Breeds	Total blast hit count	Total pathogen count	Mean pathogen count	Unique pathogen count	Unique pathogen kingdoms
Finland	$N = 83$	481	59			
	Eastern Finncattle ($N = 25$)	59	11	0.44	3	2× Bacteria; 1× Eukaryota
	Western Finncattle ($N = 28$)	55	15	0.54	4	3× Eukaryota; 1× Bacteria
	Northern Finncattle (N = 25)	197	20	0.8	4	3× Bacteria; 1× Eukaryota
	Holstein ($N = 5$)	170	13	2.6	3	3× Eukaryota
Netherlands	$N = 120$	28,204	236			
	Dutch Belted ($N = 23$)	4227	52	2.26	14	14 x Bacteria
	Dutch Friesian ($N = 24$)	4795	57	2.38	18	15× Bacteria; 3× Eukaryota
	Deep Red (N = 24)	15,263	37	1.54	13	11× Bacteria; 2× Eukaryota
	Groninger White Headed ($N = 21$)	3095	46	2.19	13	9× Bacteria; 4× Eukaryota
	Holstein (N = 5)	179	14	2.8	3	3× Bacteria
	Meuse-Rhine-Issel (N = 23)	708	30	1.30	10	7× Bacteria; 3× Eukaryota
Portugal	$N = 89$	18,965	119			
	Barrosã (N = 28)	12,897	17	0.61	3	2× Bacteria; 1× Eukaryota
	Mertolenga ($N = 29$)	5362	30	1.03	14	9× Bacteria; 5× Eukaryota
	Mirandesa ($N = 27$)	707	19	0.70	5	4× Bacteria; 1× Eukaryota
	Holstein (N = 5)	0	0	0	0	
Egypt	$N = 35$	7893	58			
	Egyptian cattle ($N = 30$)	3958	36	1.20	18	14× Bacteria; 4× Eukaryota
	Holstein (N = 5)	3935	22	4.40	3	3× Bacteria
Uganda	$N = 100$	37,534	527			
	Ankole ($N = 19$)	7463	70	3.68	4	3× Bacteria; 1× Eukaryota
	Holstein (N = 5)	545	23	4.60	3	3× Bacteria
	Karamojong ($N = 11$)	4121	95	8.64	11	7× Bacteria; 4× Eukaryota
	Nganda (N = 27)	4126	121	4.48	31	29 Bacteria; 2× Eukaryota
	Nkedi (N = 19)	17,248	108	5.68	8	6× Eukaryota; 2× Bacteria
	Ntuku (N = 19)	4108	110	5.79	21	14× Bacteria; 7× Eukaryota
South Africa	$N = 125$	28,603	396			
	Afrikaner ($N = 40$)	6261	91	2.28	6	4× Bacteria; 2× Eukaryota
	Bonsmara ($N = 20$)	8373	87	4.35	9	5× Eukaryota; 4× Bacteria
	Holstein (N = 5)	1774	35	7.00	4	3× Bacteria; 1× Eukaryota
	Nguni (N = 40)	6509	95	2.38	18	13× Bacteria; 5× Eukaryota
	Tuli (N = 20)	5720	88	4.40	12	7× Bacteria; 5× Eukaryota

and Nganda (121 unique pathogens) demonstrated the highest pathogen diversity. The imported Holstein breed, commonly used for dairy production, showed a much lower pathogen diversity, with only 3 unique pathogens identified.

Supplementary file 2 provides an overview of the unique pathogens identified across different cattle breeds and countries. Each breed is associated with a distinct set of pathogens, reflecting regional variations in pathogen exposure.

3.5. Pathogen composition across breeds and regions

3.5.1. Pathogen profiles in the Finnish cattle breeds

There are three native breeds in Finland: Western Finncattle, Northern Finncattle, and Eastern Finncattle. These breeds are adapted to the cold climate and are typically raised in small scale farming systems. While vector-borne diseases are less prevalent in colder climates, there is concern that rising temperatures may allow new pathogens to emerge in Northern Europe [6]. Among the four Finnish cattle breeds—Eastern Finncattle, Holstein, Northern Finncattle, and Western Finncattle—several pathogens were found to be common across all breeds. Notably, the following pathogens were identified in all four breeds: *Candidatus Mycoplasma haemominutum 'Birmingham 1', Mycoplasma haemocanis* str. Illinois, *Mycoplasma haemofelis Ohio2, Mycoplasma ovis* str. Michigan, and *Mycoplasma parvum* str. Indiana. These common pathogens indicate a shared exposure across Finnish breeds, which may be reflective of environmental factors or common management practices in Finnish cattle farming. The presence of these specific *Mycoplasma* species across all breeds suggests that these pathogens might be prevalent in the Finnish cattle population and could be a focus for disease management programs. In addition to common pathogens, each Finnish breed also exhibited a distinct set of unique pathogens, indicating breed-specific pathogen-host interactions. Eastern Finncattle was uniquely associated with pathogens such as *Eimeria maxima* and *Mycoplasma haemocanis*. These pathogens suggest exposure to parasitic infections, with *Eimeria maxima* being particularly associated with intestinal coccidiosis in avian species[25]. Northern Finncattle exhibited unique pathogens including *Candidatus Mycoplasma haemobos*. The presence of *Candidatus Mycoplasma haemobos* is interesting as it has previously been confirmed in Korean and German cattle [26,27]. Western Finncattle showed pathogens such as *Alentia gelatinosa* and *Eimeria tenella*, the latter being a well-known cause of coccidiosis in broilers [28]. This finding suggests a potential higher risk of parasitic infections in this breed.

3.5.2. Pathogen profiles in the Dutch cattle breeds

The Netherlands is home to several dual-purpose breeds but selected mainly for dairy production, including the Groningen White Headed and Dutch Frisian breeds. The Deep Red and Dutch Belted are selected as meat type animals. Dairy type traditional cattle are typically raised in more intensive farming systems, where biosecurity measures are critical to controlling the spread of infectious diseases, while the Dutch belted breed is outside year-round. However, the increasing occurrence of extreme weather events poses new risks for Dutch cattle populations by potentially altering the distribution of disease vectors [4]. From the

analysis of unmapped reads across multiple breeds of Dutch cattle, several pathogens were identified as common across all Dutch breeds. These pathogens include *Anaplasma phagocytophilum, Pseudomonas aeruginosa, and Escherichia coli.* The presence of these pathogens suggests a widespread occurrence in the Dutch cattle population, indicating potential endemicity or common exposure to similar environmental or management conditions.

In addition to the common pathogens, there were also those found to be unique to specific breeds. In the Dutch Belted cattle, a variety of *Salmonella enterica* subspecies were aligned, including *Salmonella enterica* subsp. *enterica* serovar *Newport, Bareilly, Kentucky* and *Stanley*. A number of *E. coli* strains were also detected, including *O18, O157, O1*, and *O99*. A similar profile can be observed in the Deep Red, Holstein, and the Friesian with the latter showing the presence of *Klebsiella oxytoca* and *Pseudomonas aeruginosa*- gram negative bacterium known to cause mastitis [29]. In the Meuse-Rhines-Yssel *Candidatus Mycoplasma haemobos* was also detected. This pathogen infects the red blood cells of cattle [30]. In addition, *Spirometra erinaceieuropaei* and *Mycoplasma yeatsii* were also identified in this breed. The former is a tapeworm typically infecting carnivore and amphibian hosts as well as humans, but cattle are not considered a reservoir thus far; the latter is a *Mycoplasma* that typically affects goats [31], but has previously been located in cattle ear canals [32].

3.5.3. Pathogen profiles in the Portuguese cattle breeds

Portuguese cattle breeds, including Barrosã, Mertolenga, and Mirandesa, are integral to the country's rural economy with their meat products certified by the European Union as Protected Designation of Origin. These cattle are valued for their meat, but they were developed as multi-purpose breeds, traditionally used as draught animals and, e.g., Barrosã also for milk. The country, of course, also boasts a commercial Holstein population. Barrosã is raised in small herds in the mountainous regions of northwestern Portugal with high-precipitation Atlantic coastal climate. The medium-sized Mirandesa cattle herds are found in the northeastern highland pastures, a region characterized by its rigorous winters and hot summers. Whereas the Mertolenga cattle are raised in extensive systems in the Southern flatlands of Alentejo with hot summers and Continental Mediterranean climate. The potential impact of climate-sensitive pathogens is an emerging threat to these cattle populations, as southern Europe is expected to face rising temperatures and increased vector activity in the coming decades [9]. The analysis of pathogen profiles in these Portuguese native cattle —Barrosã, Mertolenga Mirandesa, and Portuguese Holstein—revealed both common and unique pathogen occurrences among breeds. The Portuguese Holstein breed, notably, did not show any alignments to pathogens in the BLAST analysis. This absence could be indicative of several factors: potentially lower pathogen exposure in the commercial environment due to specific veterinary treatments. This lack of detectable pathogens in the Portuguese Holstein which are raised under intensive conditions presents an interesting contrast to the other Portuguese breeds. The remaining breeds had some pathogens in common: *Theileria orientalis*, *Anaplasma marginale* and *ovis,* as well as a range of *Mycoplasma* species. *Theileria orientalis* has already been characterized in Portuguese cattle, specifically in the Mirandesa [33]. In addition, this pathogen has been detected in asymptomatic cattle as well [34]. *Anaplasma marginale* has been detected in *Rhipicephalus bursa* ticks in Portugal [35] and *Anaplasma ovis* has been detected in wildlife vectors, sheep and cattle in other areas of Europe. In the Barrosã breed, three unique pathogens were identified, including *Anaplasma phagocytophilum* strains (Norway variant 1 and variant 2) and *Schistosoma mattheei*. The Mertolenga breed showed a broader range of unique pathogens, including B*abesia bigemina, Candidatus Mycoplasma haemolamae* str. Purdue*, and Harmothoe impar*. Similarly, the Mirandesa breed showed unique pathogen profiles, including *Anaplasma marginale* (Gypsy Plains and St. Maries strains)*, Escherichia coli, and Theileria equi*. These findings indicate a clear diversity in pathogen exposure among the different Portuguese breeds, with Mertolenga and Mirandesa showing the highest number of unique pathogens at 14 and 5 (supplementary file 2). The absence of pathogens in the Portuguese Holstein breed provides a significant point of contrast and suggests potential differences in pathogen resistance or environmental exposure. We also observed several common pathogens shared between at least two of the breeds. These common pathogens included: *Anaplasma ovis* str. Haibei*, Anaplasma phagocytophilum* str. Dog2*, Schistosoma curassoni, Anaplasma marginale* str. Florida*,* and *Anaplasma centrale* str. Israel*.* The presence of *Anaplasma* species across these breeds suggests a significant exposure to tick-borne diseases, as *Anaplasma* is commonly associated with diseases like anaplasmosis in cattle [36]. The detection of *Schistosoma curassoni*, a parasitic flatworm, indicates potential exposure to schistosomiasis, a disease typically associated with tropical and subtropical regions [37,38]. The shared presence of these pathogens across multiple breeds underscores the common environmental and epidemiological pressures faced by these cattle breeds in Portugal.

3.5.4. Pathogen profiles in the Egyptian cattle breeds

Egypt's cattle populations, including Egyptian cattle and Holstein, are primarily raised in intensive farming systems along the Nile River and its delta. Egypt's hot and arid climate poses unique challenges for livestock management, with diseases such as *Pasteurella* infections and Anaplasmosis being of particular concern. The increasing frequency of extreme weather events, such as heatwaves, is likely to exacerbate these challenges by increasing pathogen load and vector populations [6]. Egyptian cattle and Egyptian Holstein revealed both shared and unique pathogens across the two groups. The Egyptian cattle breed exhibited a wider diversity of pathogens compared to the Egyptian Holstein, with a total of 3 unique pathogens identified. Among these, notable pathogens included *Anaplasma marginale*, *Babesia bigemina, Listeria grayi, Mycoplasma wenyonii, parvum* and *ovis* and *Schistosoma margrebowiei*. These pathogens are associated with various cattle diseases such as anaplasmosis, babesiosis, and listeriosis, which may have significant impacts on the health and productivity. *Babesia bigemina* in particular is a serious problem in Egypt and is considered one of the most worrying endemic parasitic diseases affecting cattle there [39,40]. In contrast, the Egyptian Holstein breed showed a more restricted pathogen profile, with only three unique pathogens: *Anaplasma phagocytophilum* str. Norway variant1, *Pasteurella multocida,* and *Ehrlichia ruminantium* str. Welgevonden. *Pasteurella multocida* is known to be a causative agent of respiratory diseases in cattle [41], while *Anaplasma phagocytophilum* variants are associated with tick-borne diseases [42]. Both breeds shared several common pathogens, including *Theileria annulata*, *Anaplasma marginale* (with multiple strains), *Pseudomonas aeruginosa, Mycoplasma wenyonii,* and *Theileria parva*. These common pathogens suggest a similar exposure to endemic diseases in the Egyptian environment.

3.5.5. Pathogen profiles in the Ugandan cattle breeds

Uganda's cattle industry is vital, with breeds such as Karamojong, Ntuku, and Ankole forming the backbone of rural livelihoods. Uganda's climate, characterized by tropical conditions and seasonal rainfall, creates an ideal environment for the spread of vector-borne diseases such as East Coast fever, caused by *Theileria parva*. As temperatures continue to rise in East Africa, the risk of these diseases spreading to new areas is a growing concern [1]. The analysis of pathogen presence across Ugandan cattle breeds revealed both similarities and differences between their pathogen profiles.

Among the Ugandan breeds analyzed, seven pathogens were shared: *Babesia bigemina, Escherichia coli, Mycoplasma ovis* (str. Michigan)*, Mycoplasma wenyonii* (str. Massachusetts)*, Pseudomonas aeruginosa, Pseudomonas aeruginosa PA96,* and *Theileria parva* (str. Muguga). These pathogens, shown to be prevalent across all breeds, indicate a pervasive risk of tick-borne diseases and other bacterial infections in Uganda's cattle populations. All these breeds, regardless of management or location, are exposed to similar pathogen pressures. Four of these common

pathogens are vector diseases- *Babesia* and *Theleria* are transmitted by ticks (*Rhipicephalus boophilus micropus* and *Rhipicephalus appendiculatus* respectively) [40]; *Mycoplasma wenyonii* and *ovis* are both transmitted through various blood sucking insects, biting flies as well as contaminated equipment (Amadou [27,32]). What is interesting about the aforementioned *T. parva* Muguga strain is that it is often used in a vaccine dubbed "the Muguga cocktail", at times responsible for carrier status in individuals [43]. Having both *Pseudomonas aeruginosa* and *Pseudomonas aeruginosa PA96* is an interesting addition here. As well as exhibiting the more widely known strain of the pathogen, the common pathogen profile also exhibits *PA96*, which is multidrug-resistant. Resistant strains have been prevalent in Uganda at a household and farm level, but the latest research as of 2023 is yet to embark on multilocus sequence typing to provide more precise information on the identity among different strains [44].

The analysis of Ugandan cattle breeds revealed distinct pathogen profiles across the various breeds as well. The Ankole breed was found to host *Brucella* sp. *09RB8471, Theileria parva* glutamine rich membrane protein mRNA*, Mycobacterium avium paratuberculosis* and *Brucella melitensis. Brucella* species are zoonotic, with *B. melitensis* being the most common species of brucella in human illnesses and typically associated with sheep and goats as a specific animal host, and more rarely in camels and cattle in some regions with extensive small ruminant populations [45]. Ugandan cattle breeds have been known to be infected with *B. melitensis* [46]. *Theileria parva parva* glutamine rich membrane protein mRNA presence could be due simply to the BLAST database annotating a genomic region to a known transcript due to a close representation to coding exon regions. *Mycobacterium avium* has also been described in Ugandan cattle and has been a source of concern [47]. In the Holstein breed, unique pathogens included *Candidatus Mycoplasma haemobos,* which has been observed in Ugandan cattle and can develop clinical signs including anemia, transient fever, lymphadenopathy and anorexia, though in most cases remaining subclinical [48]. The Karamojong breed exhibited a diverse array of unique pathogens, including *Candidatus Brucella* strains*, Trichobilharzia regenti, and Anaplasma marginale* str. South Idaho, *Salmonella and Mycobacterium bovis* variant bovis BCG strain. These pathogens are associated with diseases like anaplasmosis and babesiosis, which can cause significant morbidity in cattle due to parasitic and bacterial infections [49]. For the Nganda breed, unique pathogens such as *Salmonella enterica*, *Escherichia coli* O157:H7, *Brucella* sp. MAB-22, and a variety of *Proteus* species were identified. These pathogens suggest the potential for exposure to both tick-borne diseases and bacterial infections in this breed. In the Nkedi breed, notable pathogens included *Trypanosoma vivax, brucei and cruzi, Prototheca wickerhamii, Mycoplasma haemocanis.* These pathogens are associated with trypanosomiasis, a vector-borne disease transmitted by tsetse flies (a pest which covers 70–75 % of Uganda's landmass), and protothecosis, which can affect both animals and humans [50]. Indeed, the Nkedi breed is rarely sprayed with acaricide as the animals seldom have ticks. It was assumed that the results would show significant absence of tick-borne pathogens and indeed, that appears to be the case. The Ntuku breed showed a unique profile with pathogens such as *Trypanosoma equiperdum*, *Schistosoma mattheei*, *Escherichia coli* O1:H42 *Escherichia coli* O125ac:K+:H10, *Escherichia coli* O15:H12, and *Salmonella enterica*. The presence of *Schistosoma* is interesting as this has been a problem for Ugandan cattle, especially in the West of the country [38]. It is important to note that Nganda and Nkedi are kept in areas with heavy human population density so a build up of pathogens is more likely than for the other breeds. The individuals sampled are from near the Lake Victoria basin with more rainfall and humid conditions, while the other 3 breeds (Ankole, Ntuku and Karamojong) are from much drier zones.

3.5.6. Pathogen profiles in South African cattle breeds

South Africa is home to several indigenous and commercial cattle breeds that play an essential role in the country's agricultural economy. Breeds such as Afrikaner, Bonsmara, and Nguni are well-adapted to the region's diverse climates, ranging from semi-arid areas to temperate zones. However, these breeds are also vulnerable to a variety of vector-borne diseases, including those transmitted by ticks and mosquitoes, which are prevalent in warmer climates [5]. There were common pathogens among the South African breeds such as *Theileria parva* strain Muguga, *Babesia bigemina & bovis*, *Mycoplasma* species*,* and *Anaplasma centrale & marginale* including strains St. Maries, Jaboticabal*,* and Palmeira. As observed for Uganda, South African cattle exhibit a range of tick-borne pathogens including *Theileria, Babesia, Anaplasma* and *Ehrlichia*. The theileria strain seen here is commonly vaccine related much like in the Ugandan case. Both *Babesia bigemina* and *bovis* have been diagnosed in south African cattle in multiple provinces and are known to cause anemia, fever and hemoglobinuria [51]. The *Anaplasma* species included *Marginale* of strains Dawn, St Maries and Florida, as well as *Ovis* of strain Haibei.

The analysis of South African cattle breeds—Afrikaner, Bonsmara, Holstein, Nguni, and Tuli—revealed distinct pathogen profiles for each breed, with unique pathogens identified in all breeds. The Afrikaner breed was found to have unique pathogens such as *Sarcocystis hjorti*, *Candidatus Mycoplasma erythrocervae*, and *Theileria parva parva*. *Sarcocystis* species are protozoan parasites known to affect muscle tissues, while *Theileria parva* is associated with East Coast fever, a tick-borne disease that can be lethal for cattle [52]. The Bonsmara breed exhibited several unique pathogens, including *Candidatus Anaplasma turritanum*, *Arenicola marina, Anaplasma* sp. NS108*,* and *Trypanosoma congolense* IL3000. These pathogens are associated with diseases like anaplasmosis and trypanosomiasis, both of which can cause significant cattle morbidity and mortality due to blood-borne infections [53]. The Holstein breed was found to harbor unique pathogens such as *Anaplasma marginale* str. Washington Okanogan, *Babesia* sp. *Xinjiang-2005, and Mycoplasma* sp. *China-1*. *Anaplasma marginale* is a well-known pathogen responsible for anaplasmosis, while *Babesia* species are associated with babesiosis, a parasitic disease that affects red blood cells [53]. The Nguni breed displayed unique pathogens including *Brucella pseudintermedia, Mycoplasma putrefaciens, Anaplasma phagocytophilum* str. HZ2*,* and *Brucella anthropi*. These pathogens suggest potential exposure to brucellosis and anaplasmosis, both of which can have significant health implications for cattle. Lastly, the Tuli breed demonstrated a distinct pathogen profile with unique species such as *Neospora caninum* Liverpool, *Pseudomonas aeruginosa* SJTD-1, *Corynebacterium bovis* DSM 20582, *Trypanosoma theileria*, and *Klebsiella oxytoca*. *Neospora caninum* is particularly concerning as it is associated with reproductive issues and abortions in cattle [54], while *Trypanosoma* species can lead to trypanosomiasis, a major cattle disease in Africa [53,55].

4. Discussion

This is the most comprehensive study to date investigating the microbial diversity present in unmapped reads obtained from WGS data of cattle populations. Unmapped read data was analyzed across geographically diverse regions in Europe and Africa. These regions were selected to capture a wide range of environmental conditions and farming systems, providing a comprehensive view of how pathogen diversity may vary in response to geographic and climatic factors. We have been studying locally adapted cattle breeds in all these regions to generate novel data on climatic and environmental adaptation of livestock, therefore these regions are bioclimatically diverse ranging from as low as −30 °C in Finland up to +50 °C in South Africa and Egypt, therefore presenting an interesting landscape to investigate pathogen presence globally. We have developed a bioinformatics pipeline to identify pathogens in unmapped reads and filter against a custom database informed by literature searches pertinent to countries and their pathogen profiles.

One of the most striking findings of this study is the sharing of pathogens traditionally associated with African cattle breeds and cattle from southern European regions, particularly in Portugal. Pathogens

such as *Theileria parva, Anaplasma platys, Theileria orientalis,* and *Babesia bigemina,* commonly found in Uganda, South Africa, and Egypt, were also detected in Portuguese breeds. This supports the expectation that warmer European countries, like Portugal, may be increasingly exposed to pathogens typically associated with tropical and subtropical climates. The presence of these pathogens in Portugal, a country experiencing rising temperatures, supports the hypothesis that climate change is facilitating the northward spread of vector-borne diseases. In contrast, these pathogens were absent or present at much lower frequencies in northern European countries such as Finland and the Netherlands, where cooler climates likely restrict vector activity. The stark difference between pathogen profiles in Northern versus Southern Europe underscores the importance of climate in shaping pathogen distribution [56]. With Southern Europe projected to experience even warmer and more humid conditions in the coming decades, the risk of pathogen spillover is likely to intensify, posing a growing threat to livestock and human health in the region.

The study revealed considerable variation in pathogen diversity across cattle breeds. Breeds from tropical and subtropical regions (e.g., Ugandan and South African breeds) showed higher pathogen counts compared to those from temperate regions. This aligns with the well-established relationship between warmer climates and the proliferation of disease vectors such as ticks and mosquitoes. For example, Ugandan breeds such as Nkedi and Nganda demonstrated a high diversity of pathogens, including *Trypanosoma* species and *Brucella*, which are known to thrive in warm, vector-rich environments.

In contrast, Finnish breeds such as Eastern and Northern Finncattle exhibited far lower pathogen diversity. This may reflect not only the cooler climate but also the small-scale, biosecure farming practices typical of northern Europe. However, it is notable that certain pathogens, such as *Mycoplasma haemominutum* and *Mycoplasma ovis*, were consistently found across all Finnish breeds, indicating that some pathogens are well-adapted to colder climates. What is most peculiar in Finnish breeds is the presence of various avian coccidiosis causing pathogens. This may be due to environmental exposure. This could very well be the case with other breeds and pathogens, especially in Dutch breeds that exhibited multiple *E. coli* and *Salmonella* species.

Nevertheless, this study is limited by its initial design. The filtered sequences attributed to pathogens are the result of a BLAST alignment, which compares our nucleotide sequences to those available in a database and attributes a statistical likelihood to these matches. This by no means is a reliable indicator of disease due to the presence of a pathogen. In addition, the study is limited by the literature search, which is biased in its own way. By stringed filtering of the output, we expect to reduce the false positive results. It was found that it was much more straightforward to search for comprehensive literature about endemic disease of Ugandan cattle breeds than it was for their Finnish counterparts, for example. In addition, the k-mer analysis is based on a design for the search of viral genomes and failed to produce a high degree of clustering. This means that while visually there are some patterns, statistically their presence cannot be confirmed. K-means clustering and ARI scores have highlighted this unfortunate trend in further investigations. It is also very important to note that the sample collection also could have not been free from contamination. The pathogens found could very well not be associated with the individual sampled and could be associated with the environment of the animal and validation of the results in blood serum samples of these animals for the specific pathogens will improve benchmarking of this approach. What this study offers is correlation between field data collected across various production environments and the potential development and use of a new approach in identification of the disease history of an animal. Further investigations are needed in unmapped read and pathogen data analyses. Complementary data, such as RNA-sequences are currently being investigated for the same purpose and could provide illuminating results.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2025.111108.

CRediT authorship contribution statement

Daniil Ruvinskiy: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Kisun Pokharel:** Writing – review & editing, Visualization, Supervision, Methodology. **Rodney Okwasiimire:** Writing – review & editing. **Rayner Gonzalez-Prendes:** Writing – review & editing. **Catarina Ginja:** Writing – review & editing. **Nasser Ghanem:** Writing – review & editing. **Donald R. Kugonza:** Writing – review & editing. **Mahlako L. Makgahlela:** Writing – review & editing. **Heli Lindeberg:** Writing – review & editing. **Melak Weldenegodguad:** Writing – review & editing. **Juha Kantanen:** Writing – review & editing, Supervision, Conceptualization. **Martijn Derks:** Writing – review & editing. **Richard P.M.A. Crooijmans:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Acknowledgements

This study was funded by the Long-term EU-Africa Research and Innovation Partnership on Food and Nutrition Security and Sustainable Agriculture (LEAP-Agri) as part of the OPTIBOV project (LEAP-Agri-326), and by the European Union's Horizon 2020 Research and Innovation Program under grant agreement No. 727715. We thank all members of the OPTIBOV consortium for their invaluable contributions to sample collection, sequencing, and expert input.

For Finland, we express our gratitude to the Research Council of Finland (No. 319987) and the national organizations managing the LEAP-Agri funding for their financial support. We also thank the breeders and breed associations for preserving these important local genetic resources and providing access to animals. We are grateful to the Natural Resources Institute Finland (LUKE), especially Tiina Reilas and Tuula-Marjatta Hamama, for their essential role in coordinating sampling and supporting the work involving Eastern Finncattle, Northern Finncattle, Western Finncattle, and Holstein breeds. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. Finnish Cultural Foundation is acknowledged for providing a PhD grant to DR. For Egypt, we gratefully acknowledge Rania Agamy and Mohamed Hamada Elsawy from the Animal Production Department at Cairo University and the Department of Cattle from the Animal Production Research Institute for their efforts in collecting Egyptian samples for the OPTIBOV project. For South Africa, the valuable contribution of Dr. Avhashoni Zwane, and her laboratory personnel Mr. Khanyisani Nxumalo and Mr. Maano Malima in assembling the requisite OPTIBOV research samples is greatly appreciated. Much gratitude is due to Dr. Morris Agaba for domesticating the laboratory protocols for the Ugandan component of the wider study and for leading the sampling teams for cattle breeds. The contribution of Damian Munyiirwa for his across-breed inputs into the sample collection, preparation, assembling and processing is very gratefully acknowledged. Dr. Barbara Mugwanya Zawedde and Ms. Christine Nakkazi of the National Agricultural Research Organization (NARO) enabled the team to sample animals from the only existing nucleus Nganda breeding herd in Uganda. Mr. Emmanuel Tayebwa of the Ankole Cattle Conservation Scheme, Mr. Bomera Asante of Butungama Multipurpose Cooperative in Ntoroko; and Mr. Charles Kasoro of Butuku Cattle Farmers Association, Ntoroko are all gratefully acknowledged. For the Netherlands, we acknowledge the Dutch Belted Breeders' Association (Vereniging lakenvelder Runderen), Groninger Whiteheaded (Blaarkop Stichting), the MRY Study (MRY Studieverenigingen Zuid en Oost), Deep Red Cattle (Vereniging Het Brandrode Rund), and Dutch Friesian (Fries Hollands Rundvee Stamboek). We gratefully acknowledge Bert Dibbits and Kimberley Laport from Animal Breeding and Genomics (WUR) for their technical

support in the laboratory. For Portugal, we acknowledge funding from the Portuguese Science Foundation (FCT) through contract grants 2020.02754.CEECIND/CP1601/CP1649/CT0008 and Leap Agri-326/LEAPAgri/0003/2017 (CG). We gratefully acknowledge the collaboration of the breeders and breed associations in Portugal for maintaining these local genetic resources and providing access to the animals: the Mirandesa Breeders' Association (Associação dos Criadores de Bovinos de Raça Mirandesa) and Valter Raposo; the Mertolenga Breeders' Association (Associação de Criadores de Bovinos Mertolengos) and José Pais and Nuno Henriques; and AMIBA (Associação dos Criadores de Bovinos de Raça Barrosã) and José Leite and Rui Dantas. We also thank the Escola Profissional de Agricultura e Desenvolvimento Rural de Vagos for facilitating access to Holstein-Friesian cattle, and CTM – Centro de Testagem Molecular, CIBIO, Vairão for their support in sample collection across Barrosã, Mirandesa, Mertolenga, and Holstein breeds.

Data availability

Whole-genome sequencing (WGS) data were obtained from the LEAP-Agri-project OPTIBOV. All associated raw sequencing data have been deposited to European Nucleotide Archive under ENA Accession codes: PRJEB90914, PRJEB90816, and PRJEB76602. Scripts are made available on GitHub through https://github.com/druvinskii/unmapped-reads

References

[1] P.K. Thornton, J. van de Steeg, A. Notenbaert, M. Herrero, The impacts of climate change on livestock and livestock systems in developing countries: A review of what we know and what we need to know, Agr. Syst. 101 (2009) 113–127.
[2] M. Shafique, M. Khurshid, S. Muzammil, M.I. Arshad, I.R. Malik, M.H. Rasool, A. Khalid, R. Khalid, R. Asghar, Z. Baloch, B. Aslam, Traversed dynamics of climate change and One Health, Environ. Sci. Eur. 36 (2024) 135.
[3] J. Kantanen, P. Løvendahl, E. Strandberg, E. Eythorsdottir, M.-H. Li, A. Kettunen, P. Berg, T. Meuwissen, S. Wang, Utilization of farm animal genetic resources in a changing agro-ecological environment in the Nordic countries, Front. Genet. 6 (2015).
[4] A.J. McMichael, Globalization climate change, and human health, N. Engl. J. Med. 368 (2013) 1335–1343.
[5] D.J. Rogers, S.E. Randolph, Climate change and vector-borne diseases, in: S.I. Hay, A. Graham, David J. Rogers (Eds.), Advances in Parasitology, Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications, Academic Press, 2006, pp. 345–381.
[6] J.A. Patz, D. Campbell-Lendrum, T. Holloway, J.A. Foley, Impact of regional climate change on human health, Nature 438 (2005) 310–317.
[7] A. Nardone, B. Ronchi, N. Lacetera, U. Bernabucci, Climatic effects on productive traits in livestock, Vet. Res. Commun. 30 (2006) 75–81.
[8] T.J.D. Knight-Jones, J. Rushton, The economic impacts of foot and mouth disease – What are they, how big are they and where do they occur? Prev. Vet. Med. 112 (2013) 161–173.
[9] J.C. Semenza, J.E. Suk, Vector-borne diseases and climate change: a European perspective, FEMS Microbiol. Lett. 365 (2018) fnx244.
[10] H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool, R.P. Thakare, S. Banday, A.K. Mishra, G. Das, S.K. Malonia, Next-generation sequencing technology: current trends and advancements, Biology 12 (2023) 997.
[11] L.K. Whitacre, P.C. Tizioto, J. Kim, T.S. Sonstegard, S.G. Schroeder, L.J. Alexander, J.F. Medrano, R.D. Schnabel, J.F. Taylor, J.E. Decker, What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual, BMC Genomics 16 (2015) 1114.
[12] K.E. Jones, N.G. Patel, M.A. Levy, A. Storeygard, D. Balk, J.L. Gittleman, P. Daszak, Global trends in emerging infectious diseases. Global trends in emerging infectious diseases, Nature 451 (2008) 990–993.
[13] G.B. Neumann, P. Korkuć, M. Reißmann, M.J. Wolf, K. May, S. König, G. A. Brockmann, Unmapped short reads from whole-genome sequencing indicate potential infectious pathogens in German Black Pied cattle, Vet. Res. 54 (2023) 95.
[14] V.N. Laine, T.I. Gossmann, K. van Oers, M.E. Visser, M.A.M. Groenen, Exploring the unmapped DNA and RNA reads in a songbird genome, BMC Genomics 20 (2019) 19.
[15] R.P.M.A. Crooijmans, OPTIBOV: Genetic characterization of cattle populations for optimized performance in African ecotypes: the Netherlands, Uganda, Finland, Egypt, South Africa and Portugal, in: Presented at the LEAP-Agri Project's Kick-off Meeting of ERA-Net LEAP-Agri Funded Projects, 2018.
[16] R. Gonzalez-Prendes, C. Ginja, J. Kantanen, N. Ghanem, D. Kugonza, M. Makgahlela, M. Groenen, R. Crooijmans, Integrative QTL mapping and selection signatures in Groningen White Headed cattle inferred from whole-genome sequences, PloS One 17 (2022) e0276309.
[17] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools, Bioinform. Oxf. Engl. 25 (2009) 2078–2079.
[18] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V. M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.
[19] FAO, WHO, Foodborne Antimicrobial Resistance – Compendium of Codex Standards, First revision, Codex Alimentarius Commission, Rome, 2023.
[20] Q. Zhang, S.-R. Jun, M. Leuze, D. Ussery, I. Nookaew, Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer, Sci. Rep. 7 (2017) 40712.
[21] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[22] B.D. Ondov, N.H. Bergman, A.M. Phillippy, Interactive metagenomic visualization in a Web browser, BMC Bioinform. 12 (2011) 385.
[23] D. Ruvinskiy, K. Pokharel, Kronatools Charts of Blast Analyses of Unmapped Reads. https://figshare.com/articles/figure/Kronatools_charts_of_Blast_analyses_of_Unmapped_Reads/28730078, 2025, https://doi.org/10.6084/m9.figshare.28730078.v1.
[24] W. Shen, H. Ren, TaxonKit: A practical and efficient NCBI taxonomy toolkit, J. Genet. Genom. 48 (2021) 844–850. Special issue on Microbiome.
[25] E. Jebessa, L. Guo, X. Chen, S.F. Bello, B. Cai, M. Girma, O. Hanotte, Q. Nie, Influence of Eimeria maxima coccidia infection on gut microbiome diversity and composition of the jejunum and cecum of indigenous chicken, Front. Immunol. 13 (2022) 994224.
[26] K. Hoelzle, M. Winkler, M.M. Kramer, M.M. Wittenbrink, S.M. Dieckmann, L. E. Hoelzle, Detection of *Candidatus Mycoplasma* haemobos in cattle with anaemia, Vet. J. Lond. Engl. 187 (3) (2011) 408–410.
[27] Y. Kim, H. Kim, J.-H. Choi, H.-C. Cho, M.-J. Ji, Y.-J. Park, J. Park, K.-S. Choi, Preliminary report of Mycoplasma Wenoynii and *Candidatus Mycoplasma* haemobos infection in Korean native cattle, BMC Vet. Res. 20 (2024) 121.
[28] J. Choi, H. Ko, Y.H. Tompkins, P.-Y. Teng, J.M. Lourenco, T.R. Callaway, W.K. Kim, Effects of eimeria tenella infection on key parameters for feed efficiency in broiler chickens, Animals 11 (2021) 3428.
[29] J. Massé, S. Dufour, M. Archambault, Characterization of *Klebsiella* isolates obtained from clinical mastitis cases in dairy cattle, J. Dairy Sci. 103 (2020) 3392–3400.
[30] L. De Souza Ferreira, P.L. Ruegg, *Graduate Student Literature Review:* Hemotropic mycoplasmas in cattle, J. Dairy Sci. 107 (2024) 3185–3196.
[31] M.J. Calcutt, B.N. Kent, M.F. Foecking, Complete genome sequence of *Mycoplasma yeatsii* strain GM274B (ATCC 43094), Genome Announc. 3 (2015) e00328-15.
[32] A. Sery, C.A.K. Sidibe, M. Kone, B. Sacko, A.K. Bouare, M. Niang, Isolation and identification of mycoplasma strains in the inner ear of cattle and small ruminants in Mali, World J. Biol. Pharm. Health Sci. 20 (2024) 171–178.
[33] C. Brígido, I.P. da Fonseca, R. Parreira, I. Fazendeiro, V.E. do Rosário, S. Centeno-Lima, Molecular and phylogenetic characterization of *Theileria* spp. parasites in autochthonous bovines (Mirandesa breed) in Portugal, Vet. Parasitol. 123 (2004) 17–23.
[34] J. Gomes, R. Soares, M. Santos, G. Santos-Gomes, A. Botelho, A. Amaro, J. Inácio, Detection of Theileria and Babesia infections amongst asymptomatic cattle in Portugal, Ticks Tick-Borne Dis. 4 (2013) 148–151.
[35] J. Ferrolho, S. Antunes, A.S. Santos, R. Velez, L. Padre, A. Cabezas-Cruz, M. M. Santos-Silva, A. Domingos, Detection of Theileria and Babesia infections amongst asymptomatic cattle in Portugal, Ticks and Tick-Borne Dis 7 (2016) 443–448.
[36] A.S. Santos, F. Bacellar, J.S. Dumler, A 4-year study of Anaplasma phagocytophilum in Portugal, Clin. Microbiol. Infect. 15 (2009) 46–47.
[37] E. Léger, A. Garba, A.A. Hamidou, B.L. Webster, T. Pennance, D. Rollinson, J. P. Webster, Introgressed animal Schistosomes Schistosoma curassoni and S. bovis naturally infecting humans, Emerg. Infect. Dis. 22 (2016) 2212–2214.
[38] D. Namirembe, T. Huyse, R. Wangalwa, J. Tumusiime, C.U. Tolo, Liver fluke and schistosome cross-infection risk between livestock and wild mammals in Western Uganda, a One Health approach, Int. J. Parasitol. Parasites Wildl. 25 (2024) 101022.
[39] F.K. Adham, E.M. Abd-El-Samie, R.M. Gabre, H.El. Hussein, Detection of tick blood parasites in Egypt using PCR assay I—Babesia bovis and Babesia bigemina, Parasitol. Res. 105 (2009) 721–730.
[40] H.Y.A.H. Mahmoud, A.A. Rady, T. Tanaka, Molecular detection and characterization of *Theileria annulata, Babesia bovis,* and *Babesia bigemina* infecting cattle and buffalo in southern Egypt, Parasite Epidemiol. Control 25 (2024) e00340.
[41] K. Namba, A. Terao, Y. Ueno, C. Teratani, I. Yamamoto, R. Kaji, S. Tamaboko, C. Mizukami, A. Hashida, M. Ikezawa, S. Tanaka, K. Kimura, Pathological and bacteriological investigations of *Pasteurella multocida*-induced epididymitis in calves, Vet. Microbiol. 304 (2025) 110445.
[42] T.T. Apaa, H. McFadzean, S. Gandy, K. Hansford, J. Medlock, N. Johnson, Anaplasma phagocytophilum ecotype analysis in cattle from Great Britain, Pathogens 12 (2023) 1029.
[43] S. Oligo, A. Nanteza, J. Nsubuga, A. Musoba, A. Kazibwe, G.W. Lubega, East coast fever carrier status and theileria parva breakthrough strains in recently ITM vaccinated and non-vaccinated cattle in Iganga District, Eastern Uganda, Pathogens 12 (2023) 295.
[44] B. Badawy, S. Moustafa, R. Shata, M.Z. Sayed-Ahmed, S.S. Alqahtani, M.S. Ali, N. Alam, S. Ahmad, N. Kasem, E. Elbaz, H.S. El-Bahkiry, R.M. Radwan, A. El-Gohary, M.M. Elsayed, Prevalence of multidrug-resistant pseudomonas aeruginosa

Isolated from dairy cattle, milk, environment, and workers' hands, Microorganisms 11 (2023) 2775.
[45] E.D. Aparicio, Epidemiology of brucellosis in domestic animals caused by Brucella melitensis, Brucella suis and Brucella abortus, Rev. Sci. Tech. 32 (1) (2013).
[46] R. Miller, J.L. Nakavuma, P. Ssajjakambwe, P. Vudriko, N. Musisi, J.B. Kaneene, The prevalence of brucellosis in cattle, goats and humans in rural Uganda: A comparative study, Transbound. Emerg. Dis. 63 (2016) e197–e210.
[47] J.B. Okuni, C.I. Dovas, P. Loukopoulos, I.G. Bouzalas, D.P. Kateete, M.L. Joloba, L. Ojok, Isolation of Mycobacterium avium subspecies paratuberculosis from Ugandan cattle and strain differentiation using optimised DNA typing techniques, BMC Vet. Res. 8 (2012) 99.
[48] B. Byamukama, M.A. Tumwebaze, D.S. Tayebwa, J. Byaruhanga, M.K. Angwe, J. Li, E.M. Galon, M. Liu, Y. Li, S. Ji, P.F.A. Moumouni, A. Ringo, S.-H. Lee, P. Vudriko, X. Xuan, First molecular detection and characterization of hemotropic Mycoplasma Species in cattle and goats from Uganda, Anim. Open Access J. 10 (2020) 1624.
[49] H. Fesseha, M. Mathewos, E. Eshetu, B. Tefera, Babesiosis in cattle and ixodid tick distribution in Dasenech and Salamago Districts, southern Ethiopia, Sci. Rep. 12 (2022) 6385.
[50] S. Ramadán, L. Bulacio, H. Dalmaso, G. Sepúlveda, M. Sortino, F. Fay, C. Misto, M. F. Salvador, A. Etchecopaz, M.L. Cuestas, First report of canine prototheca wickerhamii in Argentina. Brief literature review, Rev. Argent. Microbiol. (2025).
[51] M.S. Mtshali, P.S. Mtshali, Molecular diagnosis and phylogenetic analysis of Babesia bigemina and Babesia bovis hemoparasites from cattle in South Africa, BMC Vet. Res. 9 (2013) 154.
[52] A.A. Surve, J.Y. Hwang, S. Manian, J.O. Onono, J. Yoder, Economics of East Coast fever: a literature review, Front. Vet. Sci. (2023) 10.
[53] M.S. Paoletta, L. López Arias, S. de la Fournière, E.C. Guillemi, C. Luciani, N. F. Sarmiento, J. Mosqueda, M.D. Farber, S.E. Wilkowsky, Epidemiology of *Babesia*, *Anaplasma* and *Trypanosoma* species using a new expanded reverse line blot hybridization assay, Ticks and Tick-Borne Dis. 9 (2018) 155–163.
[54] J.P.A. Haddad, I.R. Dohoo, J.A. VanLeewen, A review of Neospora caninum in dairy and beef cattle — a Canadian perspective, Can. Vet. J. 46 (2005) 230–243.
[55] K. Rascón-García, B. Martínez-López, G. Cecchi, C. Scoglio, E. Matovu, D. Muhanguzi, Prevalence of African animal trypanosomiasis among livestock and domestic animals in Uganda: a systematic review and meta-regression analysis from 1980 to 2022, Sci. Rep. 13 (2023) 20337.
[56] C. Ainsworth, Tropical diseases move north, Nature (2023).
[57] Tahir Usman, Frieder Hadlich, Wiebke Demasius, Weikard Rosemarie, Christa Kühn, Unmapped reads from cattle RNAseq data: A source for missing and misassembled sequences in the reference assemblies and for detection of pathogens in the host, Genomics 109 (1) (2017) 36–42, https://doi.org/10.1016/j.ygeno.2016.11.009.