Ruminant microbiome data are skewed and unFAIR, undermining their usefulness for improving sustainable production

Abimael Ortiz-Chura
Université Clermont Auvergne, INRAE, VetAgro Sup, UMR 1213 Herbivores Unit
Milka Popova
Université Clermont Auvergne, INRAE, VetAgro Sup, UMR 1213 Herbivores Unit
Diego P. Morgavi (diego.morgavi@inrae.fr)
Université Clermont Auvergne, INRAE, VetAgro Sup, UMR 1213 Herbivores Unit

Research Article

Keywords: Ruminant microbiome, metadata, global representativeness, metagenome, ontology

Posted Date: October 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3384050/v1

Abstract

The ruminant microbiome plays a key role in the health, feed utilization and environmental impact of ruminant production systems. Microbiome research provides insights to reduce the environmental footprint and improve meat and milk production from ruminants. However, the microbiome composition depends on the ruminant species, habitat and diet, highlighting the importance of having a good representation of ruminant microbiomes in their local environment to translate research findings into beneficial approaches. This information is currently lacking. In this study, we explored the metadata of microbiome studies from farmed ruminants to determine global representativeness and summarized information according to ruminant species, geographic location, body site, and host information. We accessed data from the International Nucleotide Sequence Database Collaboration through the National Center for Biotechnology Information database. We recovered 47,628 sample metadata with cattle accounting for over two-thirds of the samples. In contrast, goats with a worldwide population similar to cattle were markedly underrepresented, making up less than 4% of the total samples. Most samples originated in Western Europe, North America, Australasia and China but countries with large ruminant populations in South America, Africa, Asia, and Eastern Europe were underrepresented. Microbiomes from the gastrointestinal tract were the most frequently studied comprising about 87% of all samples. Additionally, the number of samples from other body sites such as the respiratory tract, milk, skin, reproductive tract, and fetal tissue, has markedly increased over the past decade. More than 40% of the samples lacked basic information and many were retrieved from generic taxonomic classifications where the ruminant species was manually recovered. The lack of information on diet, production system, age, or breed limits the reusability of the data for reanalysis and follow-up studies. Taxonomic assignment of the ruminant host and a minimum set of metadata attributes using accepted ontologies adapted to host-associated microbiomes are prerequisites for this. Public repositories are encouraged to require this information. The results from this survey highlight the need to encourage studies of the ruminant microbiome from underrepresented ruminant species and underrepresented countries worldwide.

Background

The ruminant livestock sector is central to global food security and human nutrition. According to the FAO [1], 17% of calories and 33% of the protein consumed in the world come from animal sources, and a large proportion of these come from ruminants. Likewise, ruminants improve the livelihoods and food security of millions of smallholders [2]. Compared to 2021, global demand for meat and milk is projected to increase by about 15% by 2031 [3]. The higher milk and meat production are projected to come largely from the global expansion of cattle herds mostly in Africa (+ 13%), Latin America (+ 5%), and India (+ 3%), which are already home to the largest concentration of ruminants. This would lead to potentially adverse environmental consequences, such as greater greenhouse gas emissions, changes in land use, and negative effects on water use and quality [4].

To improve the sustainability of the ruminant livestock sector, rather than to increase herd size it is necessary to improve productivity (especially in regions with low productivity) by improving feed

efficiency while preserving animal health and mitigating the environmental impact of production [5]. However, these productive traits vary widely around the world and depend on many factors including production system, animal genetics, husbandry practices, pasture and forage quality, and the use of feed supplements [6]. Additionally, the genetic potential of indigenous ruminant species and breeds would also help to address some of these major challenges, particularly in dry and tropical regions in developing countries, where population growth is expected to be higher [3, 7, 8].

Microbiomes associated with host animals are essential for the adaptation of the holobiont “the host and associated microbes”, to the environment [9] and there is growing interest in ruminant microbiomes. Fueled by recent advances in amplicon sequencing, metagenomics, metabolomics, and other omics technologies [10], there is a better understanding of the key role of the microbiome in ruminant health [11], performance [12] and environmental impact [13].

Rangeland systems dominate ruminant production worldwide, covering 36% of the world's land area, mostly in arid areas unsuitable for crop production [14]. Furthermore, each region or country has distinctive singularities because the type, quality, and quantity of rangelands are widely variable around the world [15]. Similarly, feeding management and diet, including the presence of plant secondary compounds, potentially influence the rumen microbial ecosystem and affect animal performance and health. For example, in tropical and subtropical regions of Latin America and the Caribbean, Africa, Asia, and northern Australia, *Leucaena leucocephala* is used as forage for cattle, but its secondary metabolites tannins and mimosine have antimethanogenic and toxic effects [16, 17]. Nevertheless, ruminants can develop adaptive microbial mechanisms to neutralize the toxic effects of plant secondary metabolites, thereby developing a gradual tolerance to these compounds in feedstuffs [18]. This singularity is naturally observed in extensive farming systems with adapted native livestock.

Ruminant microbiomes differ among species [19], breeds [20], and body sites [21], and they also differ among geographic locations as the feeding and husbandry conditions are different, as described above. To enhance our understanding of the functions, diversity, and interactions of the microbiome with the host, a robust and global reference genomic database representing all these situations is needed. The issue of representativeness has been addressed by projects such as the Global Rumen Census [22], and the Hungate1000 [23]. Although the Hungate collection represents a global effort, limitations remain because 93% of microbial cultures come from traditional livestock (cattle, sheep, and goats) that originated predominantly from developed countries (91%). Additionally, more recently, efforts to expand the database with culture-free metagenome-assembled genomes have been reported in Europe with local cattle [24] and in Africa with indigenous cattle [25]. However, they still do not represent the diversity of the rumen ecosystem from other geographic locations and other ruminant species.

To date, there are no reports on the global representativeness of studies on the ruminant microbiome, nor is it known which ruminant species and body location are the most studied. To address these information gaps, we explored and summarized information on the ruminant microbiome research metadata according to animal species, geographic location, body site, information about age, sex, and breed of the

host, and system of production using databases from the International Nucleotide Sequence Database Collaboration. We also compared the country of origin of samples with the ruminant population as a proxy to assess the representativeness of regional production systems.

Methods

Data search and processing

This study was performed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [26]. We focused on the ten most important farmed ruminant species: cattle (*Bos taurus*), sheep (*Ovis aries*), goat (*Capra hircus*), yak (*Bos grunniens*), buffalo (*Bubalus bubalis*), bison (*Bison bison*), and the Old World (*Camelus dromedarius* and *Camelus bactrianus*) and New World (*Lama glama* and *Vicugna pacos*) camelids. Metadata available for these ten species were exported using the search query "txid[Organism] AND biosample sra[filter] AND "public"[filter]" in the NCBI BioSample database (https://www.ncbi.nlm.nih.gov/biosample), date of access: 28.07.2022. For instance, using the search category "bovine gut metagenome" in the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi), date of access: 28.07.2022, we obtained the taxonomy identifier for the search field "txid506599 [organism]". Then, the search query requests all samples classified in this search category. This procedure was repeated for each ruminant species combining or not with body site (gut, oral, skin, vaginal, lung, nasopharyngeal, feces, reproductive system, blood, milk, urinary tract, tracheal, eye and semen) and the word metagenome (*e.g.*, sheep gut metagenome).

Following the initial search, we found only three search categories available in the NCBI taxonomy. However, we found other generic categories nested under "gut metagenome" and "metagenome" that were not explicitly tagged as cattle, sheep or goats but contained many ruminant related samples (Table 1). Sample identifiers and all associated tags were loaded into a full XML file format. The XML files were converted into a single data frame format using the XML and xml2 packages in R software version 4.2.0 [27], which allowed extraction of the information in the principal nodes (publication date, submission date, id, project name, and attributes) and in the subnodes of the attribute node (host, geolocation and source of the sample, among others).

Table 1
Search categories available in the NCBI taxonomy and search query in the NCBI BioSample database

Ruminant species	Search category	Taxonomy	Search query
Cattle	Bovine gut metagenome	txid506599	txid506599[Organism] AND biosample sra[filter] AND "public"[filter]
	Bovine metagenome	txid1218275	txid1218275[Organism] AND biosample sra[filter] AND "public"[filter]
Sheep	Sheep gut metagenome	txid1904483	txid1904483[Organism] AND biosample sra[filter] AND "public"[filter]
	Sheep metagenome	txid1898966	txid1898966[Organism] AND biosample sra[filter] AND "public"[filter]
Goat	Goat gut metagenome	txid2809082	txid2809082[Organism] AND biosample sra[filter] AND "public"[filter]
Generic	Gut metagenome	txid749906	txid749906[Organism] AND biosample sra[filter] AND "public"[filter]
Generic	Metagenome	txid256318	txid256318[Organism] AND biosample sra[filter] AND "public"[filter]

Data from the two generic categories were analyzed to find samples associated with the ten ruminant species. For this, we manually checked the "host" attribute and, if it was empty, we checked the rest of the attributes and added any information explicitly indicating that the sample was from one of the ruminants of interest for this study. For buffalo, yak, bison, and all camelids' species metadata were only retrieved from the generic search categories, as we did not find any specific taxonomy identifier associated with search categories. For cattle, sheep, and goats, a total of 5,567, 2,607, and 1,656 samples, respectively, were retrieved from generic search categories and included in the analysis.

Prior to the final sample count for each ruminant species, we filtered out those samples that were from the environment (*e.g.*, soil, drinking water, air, cages), associated with animal samples that were processed industrially (*e.g.*, cheese) or included in the experimental design but not obtained from the ruminant animal (*e.g.*, negative control and mock). Therefore, we considered only samples coming directly from the animal's body. The result is a dataset containing 47,628 sample metadata from multiple body sites.

In analyzing the data, we found that a large proportion of the samples lacked basic information about the host attributes, such as age, sex, and breed. To retrieve this information, we reverified the metadata according to the sample and bioproject identifier in the NCBI database. If the information was not found in the bioproject description, we performed a literature search to find metadata associated with the bioproject identifier linked to the samples. Due to the high heterogeneity of the data, we recategorized some attributes to render the information contained in the dataset more meaningful. The age of cattle was categorized into calves (birth to 1 year), yearlings (> 1 to 2 years), and adults (> 2 years); for sheep

and goats, lamb or kid (birth to 5 months), yearlings (> 5 months to 1 year) and adults (> 1 year). Although there is no specific attribute for in vivo or in vitro samples in the metadata set, we were able to separate in vitro from in vivo samples by manually searching for those samples associated with reactor, culture, in vitro, and RUSITEC. Likewise, in the cattle metadata, we added the attribute production system, associating it directly with the breed, *e.g.*, breed specialized in milk production such as Holstein, so it was assigned to the dairy production system. Finally, the sequencing technique employed was not explicitly described in the available attributes of each sample, although a few had tags referring to 16S rRNA amplicon and shotgun metagenomics. Therefore, this information was not taken into account in this study.

Descriptive analysis and representative proportion

For the general descriptive analysis and for each ruminant species, we created the pivot table of the Excel file considering the attributes, biosample ID, ruminant species, date, body site (categories: oral [subcategory: oral, tonsil and saliva], gut [esophageal, rumen, reticulum, omasum, abomasum, duodenum, jejunum, ileum, cecum, colon, rectum, and anus], feces [feces], respiratory system [nasal, lung, larynx and trachea], milk [milk and colostrum], fetal tissue [liver, placenta, kidney, ileum, amniotic fluid, cecum, meconium, allantoic fluid, rumen, fetal gut, and umbilical cord], skin [skin, foot, udder skin, and ventral skin], reproductive system [uterus, vagina, and penis], liver [liver], mammary gland [udder and teat], blood [blood], eye [eye], musculoskeletal system [muscle and joint] and ears [ears]), sample type, country, breed, sex, age, and production system (for cattle only). The bar, alluvial, and donut charts were generated with the ggplot2 package [28] using R Software.

Cattle and sheep were the only species considered to estimate the patterns of over- or underrepresentation by country in relation to its global cattle and sheep population, because they were the species with the highest number of samples (~ 90%). For this purpose, we downloaded the total population of cattle and sheep per country for 2020 using the FAOSTAT database [29] (https://www.fao.org/faostat/fr/#data/QI), date of access: 26.10.2022. Consequently, the representation index was estimated with data from the country's share of the world population (of cattle or sheep) and the country's share of the microbiome samples following the methodology of [30]. Briefly, for countries with a percentage of samples greater than the percentage of cattle or sheep populations, we divided the former by the latter to obtain a number indicating how many times more samples are present than expected. For countries where the percentage of samples was less than the percentage of cattle or sheep populations, we took the negative reciprocal of this number. The provisional result leaves overrepresented countries with positive scores and underrepresented countries with negative scores. After removing scores for countries with fewer than ten samples, we scaled the positive scores to be between 0 and 100 and separately scaled the negative scores to be between 0 and − 100. The R package maps and ggplot2 were used to graphically display the representativeness maps. To add more variation to the color coding of countries, the scaled representativity indices were transformed to $\log_{10}$.

Results and discussion

Global metadata distribution of ruminant microbiome samples

A dataset with 47,628 sample metadata was obtained from ten farmed ruminant species (Fig. 1). Cattle (including *Bos taurus* and *Bos taurus indicus*) represented 71.2% of the samples followed by sheep at 18.9%. Other species were goat (3.9%), yak (2.7%), and buffalo (2.1%). The rest of the samples (~ 1.2%) are from four camelid species and from bison. Samples from live animals were dominant compared to those from in vitro experiments (93.5% vs. 6.5%, respectively). Present estimates of the worldwide farmed ruminant population are about 4.2×10^9 heads, including yak [31] and bison [32] populations that are not counted in FAOSTAT [29]. Cattle (36.45%), sheep (30.17%) and goats (26.95%) account for the largest populations, followed by buffalo (4.86%), camelids (0.92% Old World and 0.21% New World), yak (0.42%) and bison (0.01%). A comparison between the proportion of samples and head numbers for each ruminant species to identify gaps in the global research effort in regards to some ruminant populations related to others is prone to criticism. Factors such as economic and regional importance should be considered for a finer interpretation. The use of head numbers or livestock units [33] will also modify the results. Nevertheless, samples from sheep, goats and buffalo seem clearly underrepresented. This is even more evident considering that these three species are particularly abundant and economically important in African and Asian countries [34], which have a low overall contribution of samples (see below).

Figure 1.

Geographic location was a frequent metadata attribute that allowed us to identify the country of origin of the sample. We identified a total of 52 countries with China, the USA and Canada contributing more than half of the samples. Other countries contributing more than 1% of the samples were nine European countries, New Zealand, Australia, Israel, Brazil, and Japan. The remaining 31 countries contributed 5.6% of the samples (Supplementary Table 1).

For a better understanding of the microbiome metadata representation and given that cattle and sheep represent about 90% of the total samples, we analyzed the data separately for each of these two species. We then used the geographic location attribute, and along with information on cattle populations in countries worldwide, we evaluated the representativeness of sampling efforts on a global scale. To obtain a clear picture, we filtered the dataset by removing in vitro samples and countries that had a low number of samples (< 10). The latest available data for the worldwide cattle population is 1.53×10^9 heads [29]. One animal out of four in the world is from only two countries, Brazil and India. Other countries with large cattle populations are the USA (6.1%), Ethiopia (4.3%), China (3.9%), Argentina (3.5%), Pakistan (3.4%), Mexico (2.4%), Chad (2.2%) and Sudan (2.1%). However, the samples mainly originated from the USA (25.4%), Canada (13.2%), China (12.1%), Austria (6.5%), the UK (5.9%), and Israel (5.1%) (Supplementary Fig. 1). Countries with a low to moderate cattle population, for example, Israel, Austria, Denmark, Finland, Sweden, Canada, Japan, and the UK, were overrepresented. In contrast, out of the 25 countries with the

largest cattle populations, 21 are underrepresented (Fig. 2A). Furthermore, out of the 190 countries reported with cattle populations, 144 have zero samples reported in this database.

As for cattle, the geographic location and information for the worldwide sheep population were analyzed. Our results showed that, although the sheep population from the USA, Canada, New Zealand, and Ireland did not exceed 3% of the total, 55% of the sheep microbiome samples originated from these four countries. Consequently, these countries were overrepresented (Fig. 2B). China has the largest sheep population in the world (13.7%) and accounts for 32.3% of the total samples; hence, as well as the UK and France, they are considered well-represented countries. In contrast, 7 of the top ten countries in sheep populations (not including China, India, or the UK) did not register any samples (Supplementary Fig. 2). Likewise, India, Brazil, South Africa, Spain, and Egypt were ranked as the most underrepresented countries. Remarkably, out of the 173 countries reported with sheep populations, 127 have zero samples reported in this database.

Figure 2.

Sample metadata information from the three most abundant ruminant species

Regarding the body site of origin, the vast majority of samples (~ 87%) come from the gastrointestinal tract (GIT), particularly from rumen and feces, and were prevalent in all ten ruminant species. Other body sites and biological matrices represented about 13% of the samples. These are in decreasing order of importance from respiratory system, milk, fetal tissue, skin, and reproductive system categories (Table 2). Samples from body sites other than the gut and feces were mainly found in cattle and sheep. Minor categories represented less than 1% of the total samples (listed in Supplementary Table 2).

Table 2
Sample metadata distribution by body site and ruminant species

Category/subcategory	Sample counts	%	Ruminant species
Gut	30452	63.9	C, S, G, Y, Bu, BC, DC, A, LL and Bi
Esophageal	5		C
Rumen	26652		C, S, G, Y, Bu, BC, DC, A, LL and Bi
Reticulum	131		C, S, G, Y and Bu
Omasum	150		C, S, G, Y and Bu
Abomasum	252		C, S, G, Y, Bu and BC
Duodenum	374		C, S, G, Y, Bu and A
Jejunum	567		C, S, G, Y, Bu and A
Ileum	496		C, S, G, Y, Bu, BC and A
Cecum	405		C, S, G, Y, Bu and A
Colon	658		C, S, G, Y, and Bu
Rectum	525		C, S, G, Y, Bu and DC
Anus	73		S and G
Gut*	164		C, BC, and DC
Feces	10825	22.7	C, S, G, Y, Bu, BC, DC, A, LL and Bi
Respiratory system	1759	3.7	C, S, Y and DC
Milk	1389	2.9	C, S and Bu
Fetal tissue	1001	2.1	C and S
Skin	752	1.6	C and S
Reproductive system	624	1.3	C and S

*Sample metadata tagged as gut.

C = Cattle, S = Sheep, G = Goat, Y = Yak, Bu = Buffalo, BC = Bactrian camel, DC = Dromedary camel, A = Alpaca, LL = Llama, and Bi = Bison.

Cattle represented 71% of all sample metadata, and the body site was the attribute where the information was most complete. However, the information was not straightforward, and it was only recovered after refining the search on the attribute "description" of the bioproject or by manually searching the associated publications. We found 13 categories for the body site attribute. The categories Gut and Feces were also

dominant, representing about 8 out of 10 samples (Fig. 3A). Other relevant categories were: respiratory system, fetal tissue, milk, reproductive system, skin, liver, oral, mammary gland, blood, eye and musculoskeletal system (Supplementary Table 3). The breed is an important descriptive information in any animal study but it was not reported in the majority of sample metadata (57.3%). In spite of the limited availability of breed attribute data, Holstein was the dominant breed (70.0%), followed by Aberdeen Angus, Angus × Hereford crossbreed, Holstein × Jersey crossbreed, and Black Japanese (which refers mainly to the Wagyu breed) (Fig. 3B). Similar to breed, fundamental attributes for reusability and reinterpretation of sequencing data such as production system, age, and sex were poorly completed. No information on these attributes was found in 40 to 58% of the samples. The available data should be interpreted with caution but there is a predominance of sample metadata from dairy versus meat production systems (74% vs. 26%, respectively) (Fig. 3C), which is opposite to the global cattle structure, 17% for dairy cattle and 83% for beef cattle [29, 35]. Furthermore, samples from adult animals are higher than those from calves but otherwise they can be considered equilibrated (Fig. 3D). Whereas, the female category (Fig. 3E) is more abundant than the male category, which seems logical considering the sex ratio in commercial cattle herds.

Figure 3.

For sheep, a total of 9,003 sample metadata were found. As in cattle, the gut and feces categories of the body site predominated (90.9%) over the other categories (Fig. 4A) (Supplementary Table 3). Likewise, for the breed attribute, there was a high percentage of missing data (56.3%). We found a total of 31 breeds, and the most abundant were the Lacaune (20.2%), Romney (14.5%), and Hu sheep (14.0%) breeds. Most breeds were poorly represented (< 1%) (Fig. 4B). Finally, for the attributes age and sex, although there was a high percentage of samples with missing data, lambs and adults were the most represented categories (Fig. 4C), and similar proportions were observed for males and females (Fig. 4D).

Figure 4.

Goat results showed only two body site categories, gut and feces (Supplementary Table 3). Although 29 breeds were identified, about 50% of the samples lacked this attribute (Supplementary Table 4). The predominant breeds were: Liuyang black, Boer, Black fattening, and Xiangdong black. Approximately half of the breeds that were informed in the metadata have a Chinese origin, as 90% of the samples originated from China (Supplementary Table 5). Seventeen other countries registered samples, but they represented less than 10%. We found no or few samples from countries with large populations of goats (*e.g.*, India, Nigeria, Pakistan, Bangladesh, and Ethiopia). Finally, although the kids and female categories predominated in the age and sex attributes, respectively, there was a higher percentage of missing data (45.5 to 67.8%) (Supplementary Tables 6 and 7).

Sample metadata information from minor ruminant species

Outside of the major ruminant species, the number of total samples from other ruminants (yak, buffalo, camel, camelid, and bison) were equilibrated compared to their worldwide population (~ 6%). Regardless

of the ruminant species, the gut and feces categories were the most prevalent among these seven ruminant species (Supplementary Table 8). Likewise, some respiratory system and milk samples were reported from yaks, camels, and buffaloes. Sample metadata originated mainly from the Asian continent (91%). China and India had the largest number of samples (Supplementary Table 9); China was highlighted by the number of samples of yak (1,280 samples), and both countries contributed 916 samples of buffalo. For the Dromedary camel, India, Egypt, Iran, and other countries contributed 151, 108, 44, and 11 samples, respectively. There were 79 samples from Bactrian camels originating from Russia, China, Italy, and Denmark. Likewise, for bison, 58 samples were reported from the USA, Canada, and Mexico. It is noted that for New World camelids most samples were from outside the main geographic area of production and origin. There were 123 alpaca samples from the USA and New Zealand, and only eight llama samples, six from Argentina and two from France.

Database representation and FAIR principles

Our results, based on the number of scientific papers (Fig. 5A) and sample metadata evolution (Fig. 5B), suggest a growing interest in ruminant microbiome studies with the aim of understanding the function of the holobiont organism and its linkages with animal health, production efficiency, and environmental impact [11, 13]. Additionally, advances, and cost reductions, in high-throughput sequencing technologies have contributed to the increased data volume in the last decade [10]. The results indicate that the GIT is the most studied body site in farmed ruminants (Supplementary Fig. 3). This is explained by the importance of the GIT microbiota to the major challenges facing ruminant production, namely reducing greenhouse gas emissions, increasing feed efficiency, and preserving animal health [36–38]. In addition, the number of samples from the respiratory tract, milk, skin, reproductive tract, and fetal tissue has increased exponentially over the past decade, reflecting the increased interest in better understanding how resident microbiota are associated with health problems, such as mastitis [39], lameness [40] and respiratory disease [41].

Figure 5.

The quality and depth of the microbiome data from farmed ruminants is steadily improving, allowing us to explore their connection to essential biological processes relevant to production and health. Several projects and international initiatives [*e.g.*, 22,23] are contributing data, expanding the ruminant microbiome. However, the existing metadata and samples mainly originated from production systems prevalent in high-income countries, and there is still a large number of regions with large ruminant populations where metadata were scarce or nonexistent, *e.g.*, countries from South America and the Caribbean, Western Asia, Eastern Europe, and the African continent.

It is, therefore, urgent to rethink and encourage ruminant microbiome studies in underrepresented countries worldwide. It is imperative to obtain information from indigenous breeds and less represented ruminants reared under harsh environmental conditions from low- and middle-income countries where they contribute to food security [7]. These regions are where ruminant populations are increasing and where ruminants contribute the most to the economic and environmental sustainability (adaptation and

mitigation to climate change) of local human populations. We also consider that the vast but underexplored genetic diversity of ruminant microbiomes could be mined for the discovery of new genes and potentially valuable new microbial products for the biotechnology industry [42]. Finally, a better understanding of pathogenic microbes and their interactions with other microbiomes in ruminants and their environment may contribute not only to the development of healthy and sustainable livestock, but also to improved public health following the “One Health” approach [43, 44].

A main result of this study was the poor quality of the available metadata. For instance, there was no global consensus for the taxonomic assignment of the sample metadata to a ruminant species since much of the data were manually retrieved from generic taxonomies such as metagenome or gut metagenome, which include the vast majority of animal species. Likewise, we found samples of sheep and yak in the bovine metagenome and bovine gut metagenome taxonomies. All of this made it much more difficult to find and retrieve metadata. A further issue when refining the metadata information was the difficulty of distinguishing the nature of the samples. For instance, samples from in vitro studies were difficult to distinguish from in vivo because these were not explicitly defined in the metadata. Therefore, we classified samples as in vitro when they were associated with the reactor, culture, RUSITEC, or in vitro, and the remaining samples were considered in vivo. It is also important to know that in vitro anaerobic culture samples are taken from bottles or tubes, which often come from three or four individual animals or their mixture [45]. For this reason, it was important to exclude them from the proportional representativeness analysis as they do not truly represent a sample from an individual animal *per se*. Similarly, it is likely that some samples come from longitudinal studies, as this type of information was not found in the list of attributes. Given the growing interest in studying the long-term impact of dietary interventions and the gut microbiome in early life [46–48], it is therefore likely that the number of longitudinal samples will increase, and it is important that the nature of the samples be clearly defined in the metadata. An additional key point regarding data quality was incomplete (basic, but essential) host information. Although the associated bioprojects in the literature and those with more information on their attributes allowed us to complete basic host information, most of the samples did not have complete information on breed, sex, age, and production system, which was missing in more than 40% of samples. Therefore, our results related to host attributes, except for ruminant species, country, and body site, which did contain complete information, are partial and should be interpreted accounting for this caveat.

The completeness and standardization of metadata using a common language (ontology) are essential not only to ensure the quality of the available data, but also to ensure transparency, reproducibility, and reusability of data for secondary studies (meta-analyses and reviews, among others) [49]. To address these issues, there is a checklist with the minimum information about any (x) sequence (MIxS) required to be completed in the repositories [50], and international initiatives are underway to improve the quality of metadata, *e.g.*, The National Microbiome Data Collaborative (NMDC) [51], the Genomic Standards Consortium (GSC) [52], and the Agricultural Microbiome Data [53]. However, we did not observe major progress, even in more recent studies, toward incorporating these recommendations into metadata information from ruminant microbiome research. Although some issues related to metadata quality could

be related to legal concerns (*e.g.*, intellectual property protection), we believe that the major drawback is the lack of a common ontology that correctly describes the host organism and that insufficient emphasis is placed on metadata as an indissociable element of the sequencing data to follow FAIR (findable, accessible, interoperable, and reusable) principles [54]. Finding the correct ontology of animal-associated microbiomes to submit metadata is therefore a challenge to improve metadata quality. One possibility to facilitate the search for nonredundant ontology is to hierarchize the data structure for the ruminant microbiome, as was suggested for the plant-associated microbiome [49], and to adopt some categories of metadata (*i.e.*, production system, productive and health traits, sampling method, processing and storage for host samples and sequenced materials) suggested in the checklist of the Agricultural Microbiome Data [53]. Host information on the (ruminant) species, breed, age, and sex are obvious basic information that should be a minimum prerequisite to deposit microbiome sequencing data. Furthermore, adopting and using livestock-specific ontologies that define animals in their environment, such as the Animal Trait Ontology of Livestock (www.atol-ontology.com), and others related to productive and health traits such as the Animal QTLdb database (https://www.animalgenome.org/QTLdb), would provide much-needed information for data reuse. Given that it is well known that the GIT microbiota is modulated primarily by the type and quality of the diet [55], further information on the type of diet and its possible associations with productive and health traits in the global microbiome database would be interesting. The animal research microbial community should improve its compliance with open data and FAIR principles that are required by international and national funding agencies. Training focused on quality standards, FAIR principles, and ontology for microbiome data could help promote adoption.

A recent work compiled public animal metagenome data (which included pigs, horses, cattle, sheep, and wild animals) from the NCBI database [56]. These authors used a different approach to data searching and reported 3.6 times fewer cattle samples than we found in this work. This indicates that there are samples of animal metagenome incorrectly deposited in generic taxonomies, stressing the need for the correct identification of samples to the animal taxonomy. Nevertheless, we found some similarities, *e.g.*, the samples mainly came from the GIT, and from countries such as the USA, China, Canada, the UK and Austria, although they included other animal species. It is also important to note that our results are limited to databases from the International Nucleotide Sequence Database Collaboration [57], which includes the EMBL-EBI European Nucleotide Archive [58], the GenBank database of the NCBI [59] from the USA, and the DNA Data Bank of Japan [60]. Therefore, it is likely that different global representation patterns exist in other databases, such as Metagenomics RAST (MG-RAST) [61], Genome Sequence Archive (GSA) [62], Global Catalogue of Metagenomics (gcMeta) [63], and Genomes Online Database (GOLD) [64], although their orders of magnitude are small, and redundant in some cases (*e.g.*, GSA and GOLD) compared to the International Nucleotide Sequence Database Collaboration.

Conclusions

This study demonstrates that ruminant microbiome sequencing data in public repositories are accompanied by incomplete metadata, thereby hampering their reusability. Users can take some easy steps to improve metadata information when submitting data for ruminant metagenomes. The first step

is to assign the correct taxonomic classification. Additional measures should ideally involve using customized ontologies that can be accessed from public repositories for metadata collection. This could greatly improve the collection of metadata information. Repositories should require basic sample metadata information as a prerequisite for acceptance.

Moreover, certain ruminant species, such as goats, which have significant populations and regional importance, are underrepresented in the dataset, and many countries with large cattle, sheep and goat populations from South America, the Caribbean, Africa, South and Western Asia, and Eastern Europe are underrepresented or did not register any samples in public repositories.

Declarations

Acknowledgements

All the authors gratefully acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101000213-HoloRuminant.

Authors' contributions

AOC: Performed data collection and analysis, prepared figures and tables, writing the original draft. **MP and DM:** Funding acquisition, study design, critical review of manuscript**.** All the authors read and approved the final manuscript.

Funding

We gratefully acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101000213-HoloRuminant.

Availability of data and material

Not applicable.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. FAO. World Livestock: Transforming the livestock sector through the Sustainable Development Goals. Rome: FAO; 2018.
2. FAO. The Global Dairy Sector: Facts. 2016. Available from: https://www.fil-idf.org/wp-content/uploads/2016/12/FAO-Global-Facts-1.pdf.
3. OECD/FAO, OECD-FAO Agricultural. Outlook 2022–2031. Paris: OECD; 2022. Available from: https://www.oecd-ilibrary.org/agriculture-and-food/oecd-fao-agricultural-outlook-2022-2031_f1b0b29c-en.
4. OECD. Making Better Policies for Food Systems. Paris: OECD. ; 2021. Available from: https://www.oecd-ilibrary.org/agriculture-and-food/making-better-policies-for-food-systems_ddfba4de-en.
5. Mottet A, de Haan C, Falcucci A, Tempio G, Opio C, Gerber P, Livestock. On our plates or eating at our table? A new analysis of the feed/food debate. Glob Food Sec. 2017;14:1–8.
6. Gerber PJ, Mottet A, Opio CI, Falcucci A, Teillard F. Environmental impacts of beef production: Review of challenges and perspectives for durability. Meat Sci. 2015;109:2–12.
7. Cawthorn D-M, Hoffman LC. The role of traditional and non-traditional meat animals in feeding a growing and evolving world. Anim Front. 2014;4:6–12.
8. Rahimi J, Fillol E, Mutua JY, Cinardi G, Robinson TP, Notenbaert AMO, et al. A shift from cattle to camel and goat farming can sustain milk production with lower inputs and emissions in north sub-Saharan Africa's drylands. Nat Food. 2022;3:523–31.
9. Peixoto RS, Harkins DM, Nelson KE. Advances in Microbiome Research for Animal Health. Annu Rev Anim Biosci. 2021;9:289–311.
10. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013;95:401–14.
11. O'Hara E, Neves ALA, Song Y, Guan LL. The Role of the Gut Microbiome in Cattle Production and Health: Driver or Passenger? Annu Rev Anim Biosci. 2020;8:199–220.
12. Matthews C, Crispie F, Lewis E, Reid M, O'Toole PW, Cotter PD. The rumen microbiome: a crucial consideration when optimizing milk and meat production and nitrogen utilization efficiency. Gut Microbes. 2019;10:115–32.
13. Mizrahi I, Wallace RJ, Moraïs S. The rumen microbiome: balancing food security and environmental impacts. Nat Rev Microbiol. 2021;19:553–66.
14. Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, et al. Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. Bioscience. 2001;51:933–8.
15. OECD. The contribution of the ruminant livestock sector to the triple challenge. Mak Better Policies Food Syst. Paris: OECD; 2021.
16. Derakhshani H, Corley SW, Al Jassim R. Isolation and characterization of mimosine, 3, 4 DHP and 2, 3 DHP degrading bacteria from a commercial rumen inoculum. J Basic Microbiol. 2016;56:580–5.

17. Ku-Vera JC, Jiménez-Ocampo R, Valencia-Salazar SS, Montoya-Flores MD, Molina-Botero IC, Arango J et al. Role of Secondary Plant Metabolites on Enteric Methane Mitigation in Ruminants. Front Vet Sci. 2020;7.
18. Smith GS. Toxification and Detoxification of Plant Compounds by Ruminants: An Overview. J Range Manag. 1992;45:25.
19. Liu X, Gao J, Liu S, Cheng Y, Hao L, Liu S, et al. The uniqueness and superiority of energy utilization in yaks compared with cattle in the highlands: A review. Anim Nutr. 2023;12:138–44.
20. McLoughlin S, Spillane C, Campion FP, Claffey N, Sosa CC, McNicholas Y, et al. Breed and ruminal fraction effects on bacterial and archaeal community composition in sheep. Sci Rep. 2023;13:3336.
21. Lin L, Lai Z, Zhang J, Zhu W, Mao S. The gastrointestinal microbiome in dairy cattle is constrained by the deterministic driver of the region and the modified effect of diet. Microbiome. 2023;11:10.
22. Henderson G, Cox F, Ganesh S, Jonker A, Young W, Abecia L, et al. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep. 2015;5:14567.
23. Seshadri R, Leahy SC, Attwood GT, Teh KH, Lambie SC, Cookson AL, et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol. 2018;36:359–67.
24. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37:953–61.
25. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, et al. 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. Genome Biol. 2020;21:229.
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;n71.
27. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
28. Wickham H. ggplot2. New York, NY: Springer New York; 2009.
29. FAOSTAT. FAOSTAT [Internet]. 2022 [cited 2022 Oct 13]. Available from: https://www.fao.org/faostat/fr/#data/QI.
30. Abdill RJ, Adamowicz EM, Blekhman R. Public human microbiome data are dominated by highly developed countries. PLOS Biol. 2022;20:e3001536.
31. Ayalew W, Chu M, Liang C, Wu X, Yan P. Adaptation Mechanisms of Yak (Bos grunniens) to High-Altitude Environmental Stress. Animals. 2021;11:2344.
32. Freese CH, Aune KE, Boyd DP, Derr JN, Forrest SC, Cormack Gates C, et al. Second chance for the plains bison. Biol Conserv. 2007;136:175–84.
33. Benoit M, Veysset P. Calcul des Unités Gros Bétails: proposition d'une méthode basée sur les besoins énergétiques pour affiner l'étude des systèmes d'élevage. INRAE Prod Anim. 2021;34:139–60.

34. Miller BA, Lu CD. Current status of global dairy goat production: an overview. Asian-Australasian J Anim Sci. 2019;32:1219–32.
35. Medeiros I, Fernandez-Novo A, Astiz S, Simões J. Historical Evolution of Cattle Management and Herd Health of Dairy Farms in OECD Countries. Vet Sci. 2022;9:125.
36. Capper JL, Bauman DE. The Role of Productivity in Improving the Environmental Sustainability of Ruminant Production Systems. Annu Rev Anim Biosci. 2013;1:469–89.
37. Herrero M, Havlík P, Valin H, Notenbaert A, Rufino MC, Thornton PK, et al. Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems. Proc Natl Acad Sci. 2013;110:20888–93.
38. Løvendahl P, Difford GF, Li B, Chagunda MGG, Huhtanen P, Lidauer MH, et al. Review: Selecting for improved feed efficiency and reduced methane emissions in dairy cattle. Animal. 2018;12:336–49.
39. Derakhshani H, Fehr KB, Sepehri S, Francoz D, De Buck J, Barkema HW, et al. Invited review: Microbiota of the bovine udder: Contributing factors and potential implications for udder health and mastitis susceptibility. J Dairy Sci. 2018;101:10605–25.
40. Caddey B, Orsel K, Naushad S, Derakhshani H, De Buck J. Identification and Quantification of Bovine Digital Dermatitis-Associated Microbiota across Lesion Stages in Feedlot Beef Cattle. Metcalf JL, editor. mSystems. 2021;6.
41. Zeineldin M, Lowe J, Aldridge B. Contribution of the Mucosal Microbiota to Bovine Respiratory Health. Trends Microbiol. 2019;27:753–70.
42. Cowan DA. Microbial genomes – the untapped resource. Trends Biotechnol. 2000;18:14–6.
43. Trinh P, Zaneveld JR, Safranek S, Rabinowitz PM. One Health Relationships Between Human, Animal, and Environmental Microbiomes: A Mini-Review. Front Public Heal. 2018;6.
44. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020;8:103.
45. Yáñez-Ruiz DR, Bannink A, Dijkstra J, Kebreab E, Morgavi DP, O'Kiely P, et al. Design, implementation and interpretation of in vitro batch culture experiments to assess enteric methane mitigation in ruminants—a review. Anim Feed Sci Technol. 2016;216:1–18.
46. Meale SJ, Popova M, Saro C, Martin C, Bernard A, Lagree M, et al. Early life dietary intervention in dairy calves results in a long-term reduction in methane emissions. Sci Rep. 2021;11:3003.
47. Saro C, Hohenester UM, Bernard M, Lagrée M, Martin C, Doreau M et al. Effectiveness of Interventions to Modulate the Rumen Microbiota Composition and Function in Pre-ruminant and Ruminant Lambs. Front Microbiol. 2018;9.
48. Yáñez-Ruiz DR, Abecia L, Newbold CJ. Manipulating rumen microbiome and fermentation through interventions during early life: a review. Front Microbiol. 2015;6.
49. Cernava T, Rybakova D, Buscot F, Clavel T, McHardy AC, Meyer F, et al. Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis. Environ Microbiome. 2022;17:33.

50. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29:415–20.
51. Vangay P, Burgin J, Johnston A, Beck KL, Berrios DC, Blumberg K et al. Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities. mSystems. 2021;6.
52. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The Genomic Standards Consortium. PLoS Biol. 2011;9:e1001088.
53. Dundore-Arias JP, Eloe-Fadrosh EA, Schriml LM, Beattie GA, Brennan FP, Busby PE, et al. Community-Driven Metadata Standards for Agricultural Microbiome Research. Phytobiomes J. 2020;4:115–21.
54. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
55. Newbold CJ, Ramos-Morales E, Review. Ruminal microbiome and microbial metabolome: effects of diet and ruminant host. Animal. 2020;14:78–86.
56. Hu R, Yao R, Li L, Xu Y, Lei B, Tang G, et al. A database of animal metagenomes. Sci Data. 2022;9:312.
57. Karsch-Mizrachi I, Takagi T, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2018;46:D48–51.
58. Silvester N, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. The European Nucleotide Archive in 2017. Nucleic Acids Res. 2018;46:D36–40.
59. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank Nucleic Acids Res. 2018;46:D41–7.
60. Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, et al. DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Res. 2018;46:D30–5.
61. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016;44:D590–4.
62. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence Archive *. Genomics Proteom Bioinf. 2017;15:14–8.
63. Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, et al. gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res. 2019;47:D637–48.
64. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, et al. Genomes OnLine Database (GOLD) v.8: overview and updates. Nucleic Acids Res. 2021;49:D723–33.

Figures

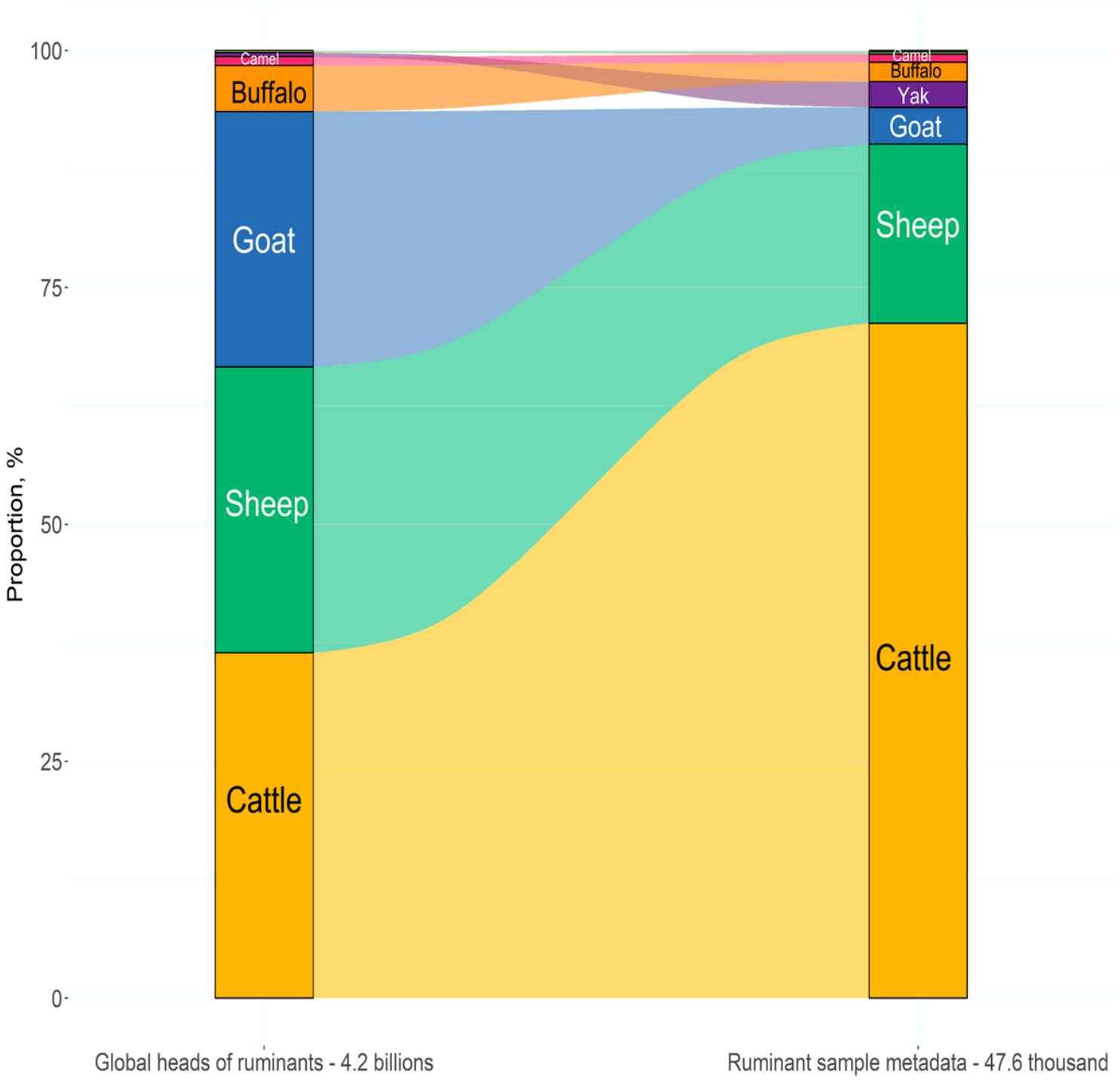

Figure 1

Comparison of proportion (%) between worldwide heads and sample metadata of ruminant species.

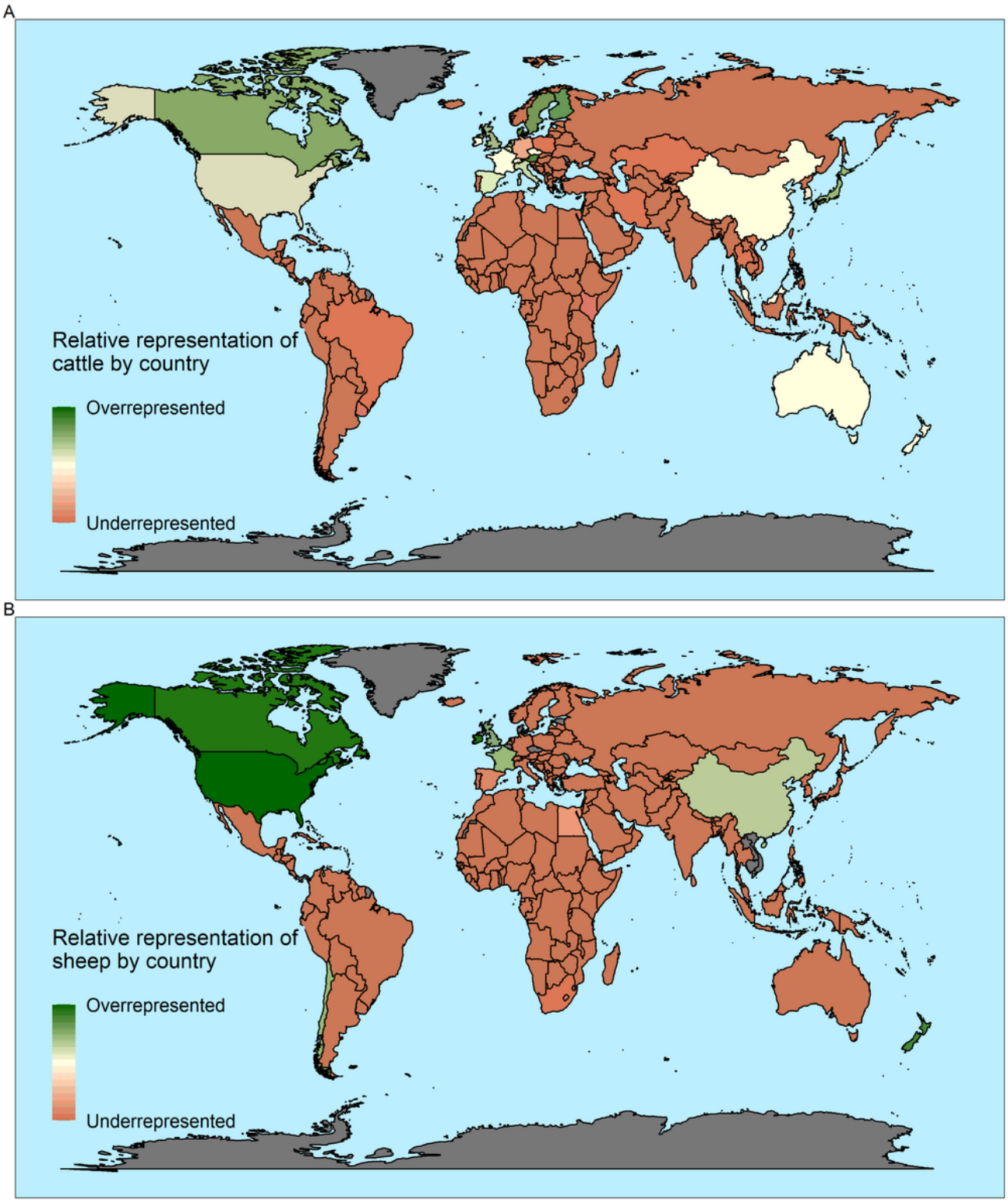

Figure 2

Data from cattle (A) and sheep (B) associated microbiome relative to abundance of livestock population in the world. Green hues mark countries where microbiome samples are overrepresented relative to their cattle or sheep populations, and red hues mark countries that are underrepresented or that have no sample metadata. Countries with no data on cattle or sheep populations in the FAOSTAT database (date of access 26.10.2022) are marked in gray.

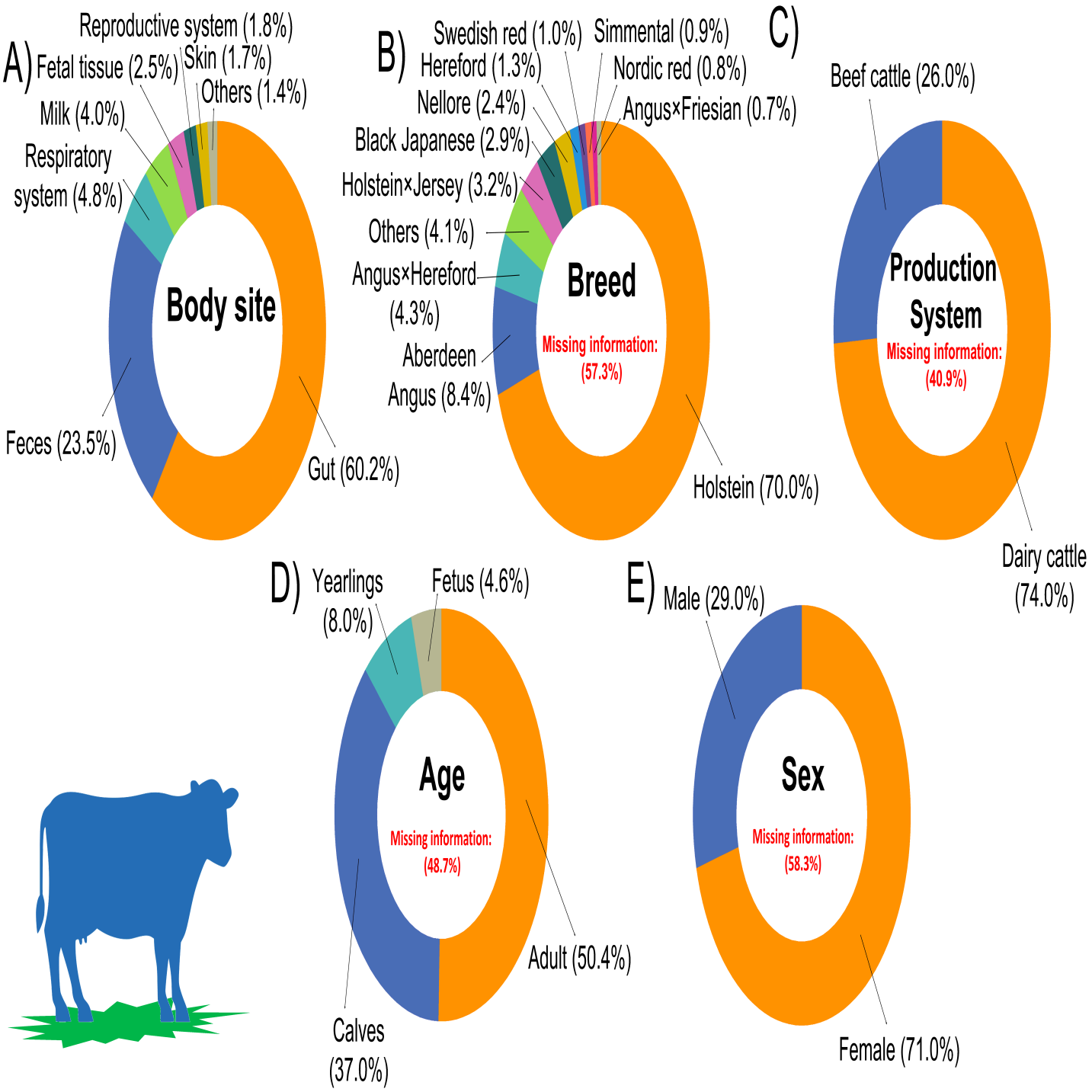

Figure 3

Cattle sample distribution according to five different categories: body site (A), breed (B), production system (C), age (D), and sex (E). For the body site and breed categories, body sites and breeds with less than 1% and 0.3% representation, respectively, were grouped in the subcategory others. Missing information on breed, production system, age and sex were not included as subcategories in the figures.

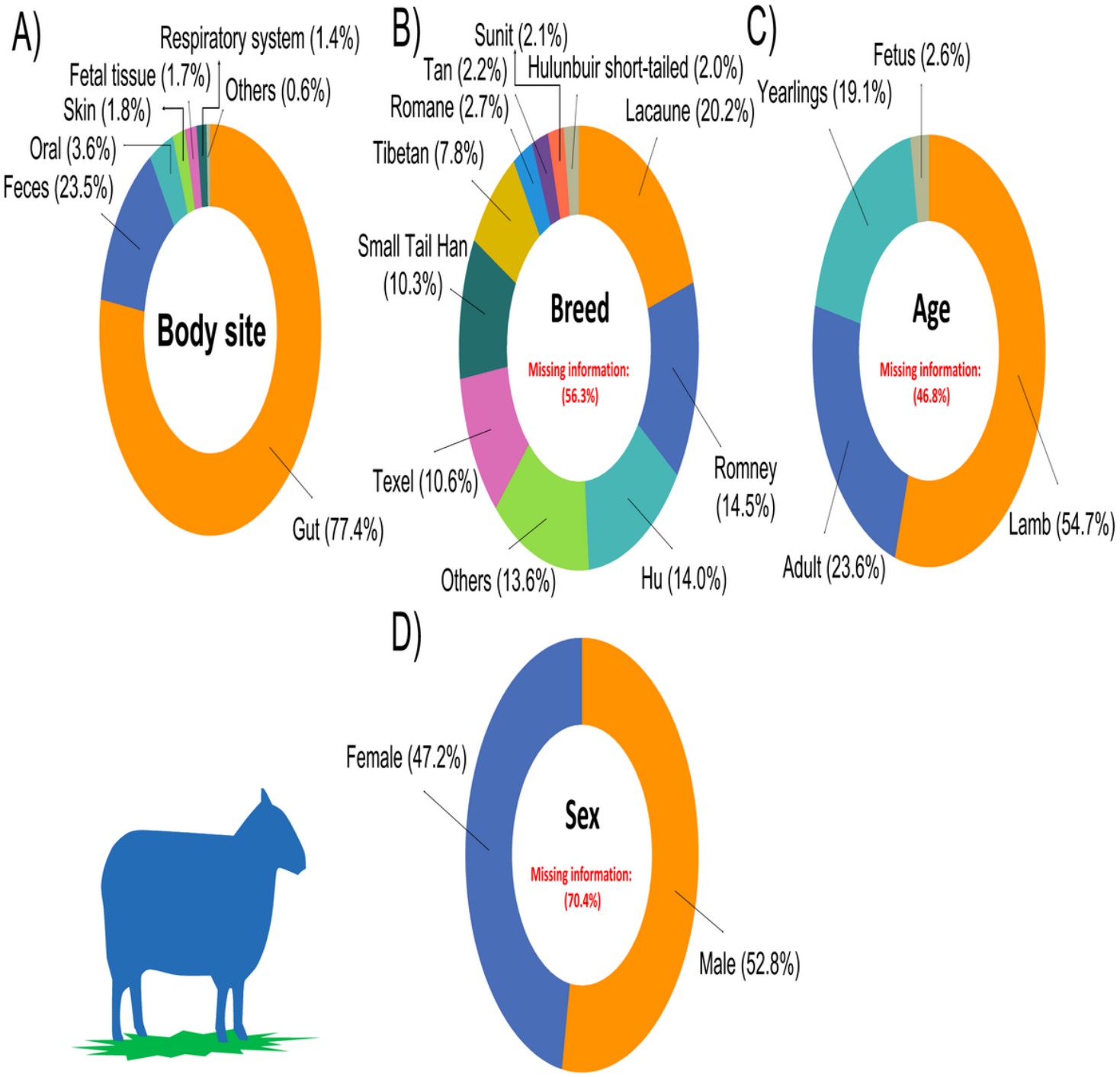

Figure 4

Sheep sample distribution according to four different categories: body site(A), breed (B), age (C) and sex (D). For the body site and breed categories, body sites and breeds with less than 1% and 2%, representation, respectively, were grouped in the subcategory others. Missing information on breed, age and sex were not included as subcategories in the figures.

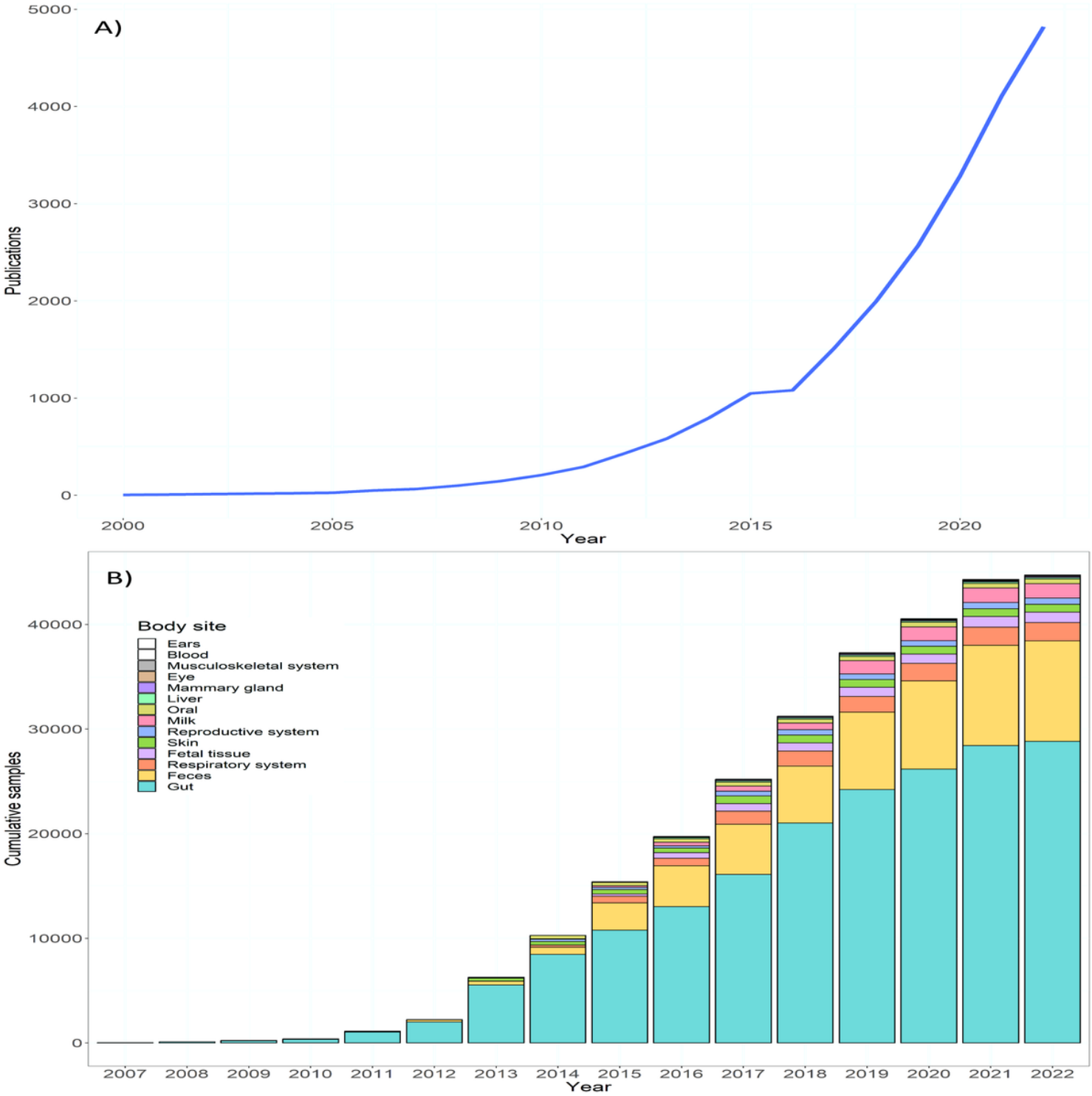

Figure 5

Timeline evolution of the ruminant microbiome studies. A) Cumulative number of published papers related to ruminant microbiome (PubMed search query: microbiome OR microbiota OR metagenome AND cow OR cattle OR sheep OR lamb OR rumen OR ruminants OR camels OR camelids OR Buffalo OR Bison [cumulative total = 4,820]) from 2000 to 2022 (up to October 26th). B) Cumulative evolution of total sample metadata by body site attribute. Bar chart plots were made using body site data of cattle, sheep and goats, including in vivo and in vitro samples. Metadata for 2022 is up to June.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure1.tif
- SupplementaryFigure2.tif
- SupplementaryFigure3.tif
- OrtizChuraetal2023SupplementaryTables.doc