



Mining livestock genome datasets for an unconventional characterization of animal DNA viromes

Samuele Bovo, Giuseppina Schiavo, Matteo Bolner, Mohamad Ballan, Luca Fontanesi*

Department of Agricultural and Food Sciences, Division of Animal Sciences, University of Bologna, Viale Giuseppe Fanin 46, 40127 Bologna, Italy

ARTICLE INFO

Keywords:

Database
DNA sequence
Farm animal
Genomics
High-throughput sequencing
Next generation sequencing
Pathogen
Virus
Zoonosis

ABSTRACT

Whole genome sequencing (WGS) datasets, usually generated for the investigation of the individual animal genome, can be used for additional mining of the fraction of sequencing reads that remains unmapped to the respective reference genome. A significant proportion of these reads contains viral DNA derived from viruses that infected the sequenced animals. In this study, we mined more than 480 billion sequencing reads derived from 1471 WGS datasets produced from cattle, pigs, chickens and rabbits. We identified 367 different viruses among which 14, 11, 12 and 1 might specifically infect the cattle, pig, chicken and rabbit, respectively. Some of them are ubiquitous, avirulent, highly or potentially damaging for both livestock and humans. Retrieved viral DNA information provided a first unconventional and opportunistic landscape of the livestock viromes that could be useful to understand the distribution of some viruses with potential deleterious impacts on the animal food production systems.

1. Introduction

The study of the animal virome is particularly relevant in the context of a One Health approach that includes humans, wild and domestic animals and their bound to the health of the ecosystems in which they live [1]. Many viruses can easily cross species barriers and can act as zoonotic agents that cause significant global health risks and burdens. Most zoonotic viruses have wildlife origins and derive from mammalian or avian hosts [2–6]. Some of them also infect livestock species which, sometimes, can act as amplifier hosts or reservoirs in the transmission to humans through their frequent and close contacts [7,8]. A few examples of these viral diseases are the Middle East respiratory syndrome (MERS) [9], Avian influenza [10,11], Rift valley fever (RVF) [12], and Swine influenza [11,13,14], among several others that have local or global diffusion.

Livestock species are also affected by many newly emerging and re-emerging specific viruses that do not infect humans but that largely impact the livestock production sectors causing direct losses through increased mortality and reduced productivity of the animals as well as indirect losses associated with cost of control and prevention, loss of trade, decreased market values and food insecurity [15–18]. Some examples of these viruses are the Bovine viral diarrhea virus (BVDV) that has large negative economic impacts in the dairy industry [18,19],

Porcine epidemic diarrhea virus (PEDV), Betaarterivirus suid 1 (also known as Porcine reproductive and respiratory syndrome viruses; PRRSV), African swine fever virus (ASF), Ungulate protoparvovirus 1 (also known as Porcine parvovirus; PPV) and Porcine circoviruses (PCVs) that hit the pork production sector with dramatic effects [16,20–24], Gallid alphaherpesvirus 2 (GaAHV2) (also known as Marek's disease virus; MDV) that is a global threat for the poultry industry [25], Myxomatosis caused by Myxoma virus (MyxV) and Rabbit haemorrhagic disease virus (RHDV) that have devastating effects on European rabbit populations [26,27].

Next-generation sequencing (NGS) has been extensively used as discovery, diagnostic and surveillance tool for viral diseases in both humans and livestock, including the exploitation of metagenomic approaches based on clinical and informative specimens with targeted and untargeted designs aimed to detect viral sequences [28–30]. NGS has been applied for many other purposes in different disciplines, including livestock genomics, where whole genome sequencing (WGS) is producing massive datasets at an increasing pace. Whole animal genome datasets that are generated for the main original objectives, that usually include the investigation of the targeted genome of individual animals for the detection of variants, can be considered interesting resources for additional data mining of the fraction of sequence reads that remain unmapped to the respective reference genome used for the sequence

* Corresponding author.

E-mail address: luca.fontanesi@unibo.it (L. Fontanesi).

<https://doi.org/10.1016/j.ygeno.2022.110312>

Received 26 July 2021; Received in revised form 16 January 2022; Accepted 6 February 2022

Available online 10 February 2022

0888-7543/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

alignment [31,32]. We will hereafter refer to them simply as unmapped reads. A significant proportion of all sequence data in these studies, about 2–10%, is left uncharacterized as it might be derived from unassembled regions of the animal genome, bacterial, archaea, non-human eukaryotic and viral DNA that is sequenced randomly by the shot-gun sequencing approach [31–33]. Part of this DNA can derive from contaminated reagents or laboratory components used in the high-throughput sequencing reactions and preparation steps of NGS libraries, from other environmental sources or from infecting agents of the organism that is the main target of the sequencing analysis [31–33]. Contaminating DNA can sometime create noise in the interpretation of the sequence information derived by pathogen agents [33].

Sequence data originated from infecting agents that can be detected within massive NGS products can provide interesting information for the identification of new potential threats and for monitoring purposes. For example, these by-products derived by massive WGS have been exploited to obtain an opportunistic characterization of the human blood virome of European and Chinese populations [33,34]. The mining of these WGS datasets obtained indications on the prevalence of infections of some viruses, a few of which unexpected or with different diffusion between human populations, sex or age cohorts [33,34]. We recently applied a similar approach in the pig using NGS data generated by shotgun sequencing of pig DNA pooled from 100 individual archived samples and identified several viruses of the *Parvoviridae* family, re-dating back their first occurrence in *Sus scrofa* [31].

In this study, we enlarged the mining of WGS data produced in a few livestock species for the identification of virus sequences by exploiting a total of about 1500 publicly available NGS datasets generated from cattle, pigs, chickens and rabbits. The results provided a first i) global (embracing a high number of breeds, lines and specimens sequenced over the last years at the global scale from the targeted livestock species) and ii) opportunist (derived from a serendipity-oriented approach) catalogue of livestock DNA viromes that could be useful to understanding the diffusion of some viruses and the composition of viromes in these animal species. The outcome can be useful to design a novel and unconventional monitoring strategy for viruses infecting livestock that could be part of a One Health approach against new potential virus-derived pandemics.

2. Materials and methods

2.1. Whole genome sequencing datasets from several animal species

Genomic information (i.e. WGS data) produced from cattle, pigs, chickens, rabbits, related wild populations of the same targeted species and other phylogenetically close species were retrieved from the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena/browser/>) [35]. REST APIs available through the ENA taxonomy services were used to retrieve the deposited information (metadata: e.g. taxonomy identifier, project identifier, sequencing data) considering the following taxonomic elements and identifiers (txid): (i) 9903 (genus *Bos*; cattle datasets), 9822 (genus *Sus*; pig datasets), (ii), 9331 (species *Gallus gallus*; chicken datasets), and 9984 (genus *Oryctolagus*; rabbit datasets). For the final evaluation, we retained only datasets presenting the following metadata tags: library_source = GENOMIC, library_layout = PAIRED and library_strategy = WGS. Then, based on the number of sequenced nucleotides and the size of the reference genomes, only the subset of datasets presenting an estimated depth of sequencing against the corresponding animal reference genome (see below) greater than 5× were retained for further analyses. Sequencing data were locally download via the Aspera ascp command line client (<https://www.ibm.com/products/aspera>). A summary of the datasets used in this study is presented in Table 1.

This process led to download a total of 1471 datasets, including more than 30 Tera bytes (TB) of WGS data (503 datasets from the *Bos* genus including *Bos taurus*, *Bos javanicus* and *Bos grunniens*; 464 datasets from

Table 1
Datasets analyzed in this study.

Information on the datasets	Cattle	Pig	Chicken	Rabbit
No. of projects	9	27	43	5
No. of datasets ¹	503	464	437	67
No. of populations ²	35	52	15	31
Stored data (TB) ³	13.0	9.5	5.8	1.8

¹ It includes whole genome sequencing datasets from paired-end sequencing with sequencing depth of the host genome >5 × .

² Approximate information as detailed metadata were not always available.

³ Total amount of downloaded and analyzed data in terabyte.

the *Sus* genus including *S. scrofa*, *Sus cebifrons*, *Sus barbatus* and *Sus verrucosus*; 437 *G. gallus* datasets; and 67 *Oryctolagus cuniculus* datasets) related to 84 sequencing projects and 137 farmed and wild populations. To simplify subsequent analyses and the presentation of the results, datasets were then referred to the cattle (genus *Bos*), pig (genus *Sus*), avian (*G. gallus*) and rabbit (genus *Oryctolagus*). In many cases, one WGS dataset corresponded to genome sequences from one individual animal. However, this was not true for some WGS datasets that derived from DNA pooling strategies. Clear indications on the number of individuals included in each dataset was not always available as metadata were incomplete. Therefore, we used the term dataset to indicate identifiable WGS data both obtained from a single individual animal (that represented most of the reported cases) or from a DNA pool from more than one animal. Information about the genome sequencing project was available for all datasets whereas information of the breed, sex and the tissue used for DNA extraction was not always available and was not used in this study to summarize the data. Information on all evaluated datasets is provided in Table S1.

2.2. Sequence alignment to the animal reference genomes (host reference genomes)

Downloaded sequencing data were initially quality checked with FASTQC v.0.11.7 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Then, reads were mapped on the host reference genome using the BWA-MEM algorithm v.0.7.17 and Samtools v.1.10 [36,37] and the parameters for paired-end data. The following animal reference genomes were used: ARS-UCD1.2 (cow), Sscrofa11.1 (pig), GRCg6a (chicken) and OryCun2.0 (rabbit).

2.3. Building a comprehensive virus database

The list of NCBI accessions of known viral genomes was retrieved from the NCBI Viruses resource (<https://www.ncbi.nlm.nih.gov/genome/viruses/>; file:taxid10239.nbr) [38]. The DNA sequence of each reference viral genome (representative genome) and its strains (neighbor genomes) was download in FASTA format using the Entrez Programming Utilities (*esearch* and *efetch* utilities; <https://www.ncbi.nlm.nih.gov/books/NBK25501/>). The final dataset included 238,353 viral genomes. The NCBI Viruses resource also provided additional information, including taxonomy name of the virus and its lineage (genus, family, species), that were used to evaluate the results. Where possible, these metadata were verified and crosschecked with the International Committee on Taxonomy of Viruses (ICTV; <https://talk.ictvonline.org/>) resources. In particular, we made use of the latest release of the Virus Metadata Resource (MRS version October 19, 2021; MSL36; <https://talk.ictvonline.org/taxonomy/vmr/>) to extract and evaluate the species name (included genus and family) approved by the ICTV as well as the virus name(s) and virus name abbreviation(s). The taxonomy/virus name will be used through the text hereafter. The traditional nomenclature/name was also reported for a few viruses, to facilitate cross-reference based on the literature that also refers to alternative names.

2.4. In silico identification of virus sequences

The virome was explored by adopting two strategies (Fig. 1): (i) read mapping and (ii) read assembly. Both procedures started with a step that implied the extraction of sequencing reads not mapped on the host reference genome (unmapped reads). Unmapped reads pairs were subsequently extracted with Samtools v1.10 [37] with the options -f 12 and -f 256.

2.4.1. Read mapping

Unmapped read pairs were initially re-mapped over the virus database. In this initial step, BWA-MEM v.0.7.17 was used to speed up the identification process and reads pairs were separated and treated as single entities to maximize the mapping. Duplicated reads were then removed with Picard v.2.1.1 (<https://broadinstitute.github.io/picard/>). The remaining read identifiers were collected (considering both paired and unpaired reads) and DNA sequences of both pairs were retrieved and aligned back over the same resource (virus database) by using BLAST+ v.2.7.1 (algorithm *blastn*; [39]) in order to refine the results. As several strains are available for a given virus reference genome, for each read we retained all the alignments presenting an E -value $\leq 10^{-20}$, sequence coverage $\geq 97\%$ and a sequence identity $\geq 75\%$. Reads mapped simultaneously over different isolates or strains belonging to different representative (reference) viral genomes were discarded. Read pairs were interrogated for pointing out the same representative viral genome: pairs mapped on different representative viral genomes were discarded. Only viral genomes having at least three mapped read pairs were considered as characterizing the sample. Read counts were lifted over the representative genome. Reads mapping to the different strains were globally retrieved and re-aligned with BWA-MEM v.0.7.17 to the representative genome to compute their coverage in the whole population.

2.4.2. Read assembly

Following our previous work [40], unmapped reads were assembled with MEGAHIT v.1.1.3 [41] with default parameters except the “meta-

large” option that forced the usage of a k-mer list equal to 27,37,47,57,67,77,87. For each assembled contig, we retained all the alignments presenting an E -value ≤ 0.01 , sequence coverage $\geq 50\%$ and a sequence identity $\geq 75\%$. Also in this case, contigs mapped simultaneously over strains belonging to different reference virus genomes were discarded. The detection of one viral DNA sequence (contig) was considered sufficient for declaring the presence of a given virus.

3. Results

3.1. Characteristics of the mined animal (host) genome datasets

More than 480 billion reads from a total of 1471 WGS datasets were aligned with BWA on the selected host reference genomes. After having pruned duplicated reads, on average a reference animal (host) genome was covered at least for its $\sim 95\%$ with a depth of sequencing of $\sim 16\times$ (Table S2). The rabbit datasets had, on average, the minimum values of sequencing depth and coverage ($12.39\times$ and 94.50% , respectively). The cattle datasets had the maximum average value of sequencing coverage (99.68%) whereas the chicken datasets had the maximum average value of sequencing depth ($21.53\times$). The pig datasets had intermediate average values for these two parameters (95.43% and $12.73\times$).

3.2. Summary of viruses identified within the unmapped reads

Unmapped reads (i.e. reads not mapped on the reference host genomes) were used to identify viral sequences. Statistics of unmapped reads are given in Table S2. The average percentage of unmapped reads was $< 2\%$ in all species. A few cattle datasets had a percentage of unmapped reads higher than 10% due to the hybrid origin of the sequenced bovine genomes (*B. taurus* \times *Bos indicus*) or because they were obtained from *B. indicus* (while the reference genome that we used was from *B. taurus*). Two chicken datasets (originally labelled as obtained from *G. gallus*), actually, were highly contaminated (or mixed with reads of different origin) and constituted two outliers.

The first step of analysis based on BWA alignment of unmapped reads

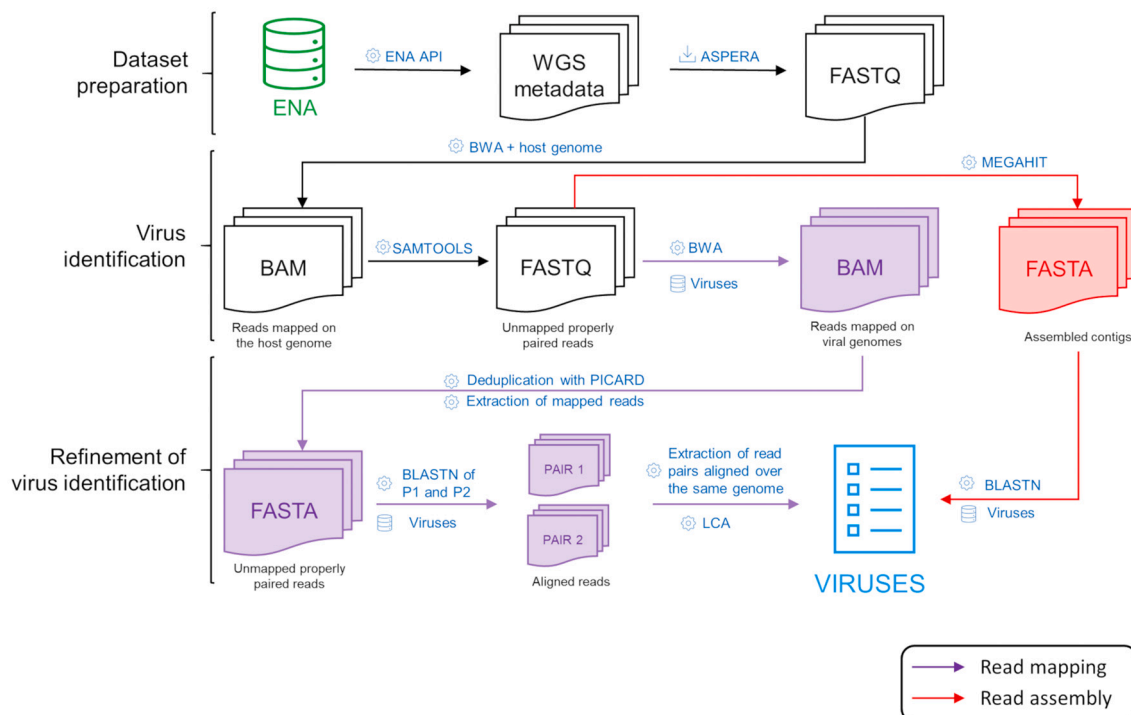


Fig. 1. The flowchart that summarizes the study design and the bioinformatic pipeline applied to identify virus sequences in whole genome sequencing datasets derived from the four livestock species.

against the virus database identified an averaged percentage of matched reads that ranged from 7.33% (pig datasets) to 15.18% (cattle datasets). Subsequent refinement based on BLASTn analysis, that removed false/spurious alignments, indicated that reads that truly mapped to a viral genome ranged between 15.89% (pig datasets) to 36.30% (cattle datasets) of the reads previously assigned to virus sequences with BWA. The median values, less affected by some outlier datasets, ranged from 0.07% (rabbit datasets) to 4.89% (cattle datasets), with intermediate median values for the pig datasets (0.37%) and the chicken datasets (0.44%).

The read mapping approach identified a total of 510 matches with viral genomes (represented by a minimum of three read pairs), if the identification is considered independently in the four targeted host genome species. The largest number was detected in the cattle datasets (Table 2), which is similar, in terms of number of investigated entries, to the pig and chicken datasets (whereas the rabbit datasets were the smallest ones in terms of number of sequenced individuals). Some virus genomes were identified in more than one host dataset. The Venn diagram reported in Fig. 2 shows the degree of co-presence of viruses in the datasets of the four livestock species. Therefore, considering unique identifications, a total of 367 viruses was detected (Table S3). Based on the information available, few of these viruses (7.9%) could be considered most specific, as reported below (Table 3).

On average, each sample contained 1–4 different viruses (Table S2). Datasets from which no virus sequences were retrieved were 91 in cattle (18.1%), 166 in pig (35.8%), 101 in chicken (23.0%) and 22 in rabbit (32.8%). Most of these datasets probably represented WGS data that were filtered before the deposition in the public repository. This information was not available in the metadata that annotated these datasets. Distribution of detected viruses over the different analyzed livestock species and sequencing projects is given in Fig. 3a.

Table S4 reports the number of samples in which the identification of reads assigned to viruses demonstrated the presence of the detected virus families. *Myoviridae* (a family of bacteriophages) was the virus family accounting for the highest percentage of positive samples (~18% of the investigated samples), followed by the *Herpesviridae* and *Siphoviridae* families (Table S4). Within the chicken datasets, the *Retroviridae* was detected in 18.6% of the samples whereas *Microviridae* (a family of bacteriophages) was detected in ~30% of the investigated samples.

3.3. Viruses infecting bacteria, fungi, plants and invertebrates

The identified virus sequences covered different classes of viruses infecting bacteria, fungi, plants and animals of several taxa (including several invertebrates), reflecting an extensive environmental contamination of the sequenced DNA samples or contamination derived from the reagents used in the different steps of the sequencing analysis (Fig. 3b).

Among the contaminating viruses, we observed that ~40% of the datasets (Table S3) contained traces of the *Escherichia* phage phiX174. This result was expected as DNA of this phage is used as control in the Illumina sequencing. The Semliki forest virus (SFV) was the second most present virus (~24% of samples over all datasets; Table S3). The identification of this virus seems due to mis-annotation in the virus database of sequences belonging to vectors used to express recombinant proteins that contaminate reagents used in the different analytical steps [42]. A high number of bacteriophages (n. 251) were retrieved, with some prevalence in the cattle datasets (n. 174). This also explains why the

Table 2
Number of viruses identified in the four livestock datasets.

Viruses	Cattle	Pig	Chicken	Rabbit
Total no. of viruses	223	160	105	22
Number of host specific viruses ¹	14	11	12	1

¹ Viruses mainly infecting the indicated livestock species. This count includes retroviruses.

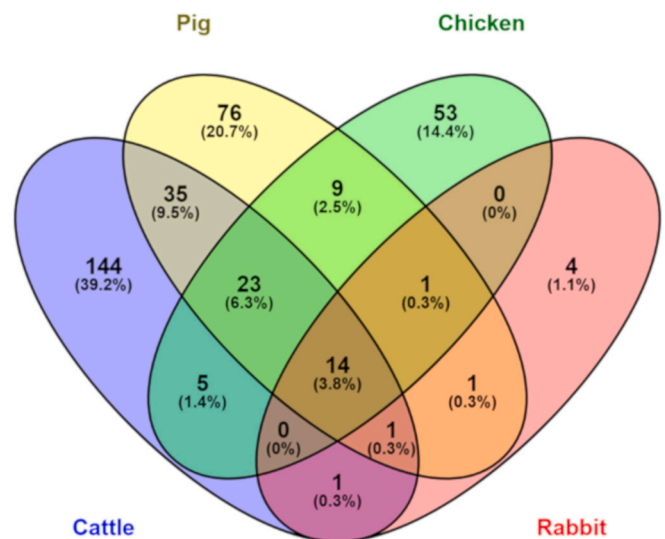


Fig. 2. Venn diagram showing the degree of co-presence of viruses in the datasets of the four livestock species

Myoviridae and the *Microviridae* families were those detected in the greatest part of the analyzed samples (Table S4).

Again, this is a clear indirect indication of contamination derived from bacteria [33,43]. Most of these phages derive from *Escherichia coli*, *Staphylococcus* spp. and other Enterobacteriaceae (Table S3). The contamination from invertebrate viruses was relevant and, again, mainly due to one virus that is commonly used for engineering plasmids and molecular biology constructs: the baculovirus *Autographa californica* multiple nucleopolyhedrovirus (AcMNPV). This virus was detected in the datasets of all animal species (11.9% of all datasets), with the highest frequency in the cattle datasets (25.8%). Among the other 14 viruses classified to infect invertebrates, the most frequent was a shrimp virus (White spot syndrome virus) identified only in 0.7% of all samples (Table S3), probably due to a laboratory specific contamination. Other invertebrate viruses, detected only sporadically in one to three datasets, mainly infect insects of different species (Table S3). A total of 22 viruses infecting plants were identified: 10 datasets contained DNA of *Citrus exocortis* viroid whereas all other viruses were identified in one to three datasets. The presence of fungi viruses was very limited: only three datasets contained viruses classified to affect these organisms.

3.4. Vertebrate viruses

Sequences from a total of 68 viruses affecting one or more vertebrate species were identified among the unmapped reads (Table S2; Fig. 3c). These viruses were detected in the datasets of all four livestock species. They can be distinguished in the following groups: 1) viruses known to infect only humans, 2) viruses known to infect both humans and other vertebrates, 3) viruses known to infect non-human vertebrates but usually not the livestock species targeted in this study, 4) retroviruses and 5) viruses known to infect the targeted livestock species.

Among the first four groups, again, many viruses could be considered contaminants of the reagents or contaminants of the laboratory environments (where workers could contribute to contaminate laboratory consumables and equipment) and few could be considered potential infecting viruses of the targeted livestock species. Moreover, in some cases matches reported on a specific strain resulted to be due to engineered regions containing vectors or other artificially introduced constructs in the reference viral genomes reported in the virus database. Some examples of these viruses are described below. Information about the viruses of the fifth group, that could be considered to infect only (or mainly) cattle, pigs, avian and rabbits, is reported in Table 3.

Table 3
Details of some vertebrate specific viruses detected in the investigated whole genome sequencing datasets.

Host species	Virus ¹	NCBI ID	No. of datasets	No. of projects	% of positive datasets	No. of sequencing reads per dataset		Sequencing depth of viral genome (×)		
						Max	Total reads	Median	Max	
Cattle	Bovine gammaherpesvirus 6	NC_024303	73	5	14.5	1008	5878	4	21	
	Ungulate erythroparvovirus 1	NC_037053	14	3	2.8	16,319	21,979	67	6662	
	<i>Bos taurus</i> polyomavirus 1	NC_001442	6	2	1.2	426	7372	240	299	
	Deltapapillomavirus 4	NC_001522, NC_030795	2	2	0.4	12	18	<1	1	
	Bovine gammaherpesvirus 4	NC_002665	2	1	0.4	8	14	<1	1	
	Bovine papular stomatitis virus	NC_005337	2	1	0.4	8	14	<1	1	
	Pseudocowpox virus	NC_013804	2	1	0.4	14	20	<1	3	
	Betapapillomavirus 2	NC_001596	1	1	0.2	26	26	<1	3	
	Xipapillomavirus 1	NC_004197	1	1	0.2	6	6	<1	1	
	Bovine adenovirus E	NC_020074	1	1	0.2	6	6	<1	1	
	Bovine polyomavirus 2	NC_025811	1	1	0.2	24	24	<1	2	
	Xipapillomavirus 2	NC_004197	1	1	0.2	6	6	<1	1	
	Bovine papillomavirus	NC_035208	1	1	0.2	24	24	<1	2	
	Bovine leukemia virus*	NC_001414	1	1	0.2	8	8	<1	1	
Pig	Suid gammaherpesvirus 3	NC_038264	61	12	13.1	474	1996	2	10	
	Suid gammaherpesvirus 4	NC_038265	42	10	9.1	480	1404	2	9	
	Suid betaherpesvirus 2	NC_022233	15	9	3.2	3492	4514	3	15	
	Torque teno sus virus 1a	NC_014070	9	4	1.9	120	220	2	16	
	Torque teno sus virus 1b	NC_027059	7	3	1.5	50	110	<1	20	
	Ungulate copiparvovirus 4	NC_023860	6	4	1.3	14,120	14,326	371	511	
	Ungulate tetraparvovirus 2	NC_038546	3	3	0.6	8688	8720	287	364	
	Porcine partetravirus ²	NC_022104	2	2	0.4	350	374	<1	161	
	Ungulate protoparvovirus 1	NC_001718	1	1	0.2	12	12	<1	4	
	Porcine parvovirus 5	NC_023020	1	1	0.2	10	10	<1	3	
	Porcine polyomavirus	NC_040714	1	1	0.2	6	6	<1	2	
	Avian	Avian endogenous retrovirus, EAV-HP*	NC_005947, AJ623292.1 ³	202	34	45.9	602	15,618	<1	2176
		Gallid alphaherpesvirus 3	NC_002577	40	18	9.1	2470	8126	4	17
Gallid alphaherpesvirus 2		NC_002229	24	13	5.5	5442	16,912	9	26	
Avian leukosis virus*		NC_001408, NC_015116	23	9	5.4	34	338	<1	100	
Avian gyrovirus 2		NC_015396	8	4	1.8	187,798	233,350	7908	8007	
Gyrovirus GyV3		NC_017091	8	5	1.8	124	274	17	38	
Gallid alphaherpesvirus 1		NC_006623	3	2	0.7	478	556	<1	9	
Fowlpox virus		NC_002188	2	2	0.5	210	216	<1	4	
Duck hepatitis B virus		NC_005890, NC_005950, NC_005888, NC_001344	2	2	0.5	14	26	1	5	
Turkey siadenovirus A		NC_001958	1	1	0.2	12	12	<1	3	
Meleagrid alphaherpesvirus 1		NC_002641	1	5	0.2	6	6	<1	2	
Reticuloendotheliosis virus*		NC_006934	1	1	0.2	6	6	<1	2	
Rabbit		Myxoma virus	NC_001132	3	1	4.5	3468	3506	1	8

¹ Retroviruses are indicated with an asterisk. Virus names have been retrieved from the NCBI and ICTV resources.

² A recently classification grouped Porcine partetravirus with Ungulate tetraparvovirus 2. We however reported separately information from reads matching the two reference genomes corresponding to the two viruses (or strains) as they were included separately in the NCBI Virus Reference Genome database. Reads were assigned to one of the two viruses according to the highest BLASTn *E*-value.

³ This EAV-HP strain (provirus), which has only a few domains of the *pol* gene, matched better the identified reads. It was used to calculate median and maximum coverage statistics.

3.4.1. Human and other non-livestock specific vertebrate viruses

A total of 197 datasets across three livestock species (19.7%, 9.3% and 12.5% of the cattle, pig and avian datasets, respectively) contained sequences from the Human betaherpesvirus 5 (HuBHV5; NCBI: NC_006273), also known as Human herpes virus 5 (HHV-5) or Human cytomegalovirus (HCMV). The detection of this virus should be considered a signature of contamination from reagents as cytomegalovirus (CMV) regulatory elements are used in expression vectors and could remain in molecular biology reagents [44]. Human alphaherpesvirus 1 (HuAHV1; NCBI:NC_001806), also known as Human herpesvirus type 1 or Herpes simplex virus type 1 (HSV1), was identified in a total of 123 datasets derived from all four livestock species (8.3%) and was the second most frequent human virus that we detected. Again, this virus could indicate contamination from expression vectors that are used to produce reagents like recombinant proteins and enzymes [44]. Moreover, it is worth to note that the pig dataset counted 51 samples positive for either the HuBHV5 or the HSV1 virus, 13 of which presented

the co-occurrence of these two viruses (Table S5). Other reads, which identified human/vertebrate infecting viruses could indicate potential contaminations of sequencing reagents [33]. For example, Human mastadenovirus C (HAdV-C; also known as Human Adenovirus C; NCBI: NC_001405) was identified in 1.6% of all datasets (and in datasets from all four livestock species) suggesting that it could be derived from contaminated reagents, as already indicated [33]. Reads matching the Hepatitis B virus (HBV) sequences (NCBI:NC_003977) were identified in two different pig sequencing projects (with a total of five datasets) and in one chicken sequencing project (two chicken datasets). *Macaca mulatta* polyomavirus 1 [also known as Simian virus 40 (SV40), NCBI: NC_001669] was identified in only two pig datasets. Reads from Influenza A virus matching different segments of a H1N1 strain were identified in the cattle and pig datasets (Fig. S1). This is an unexpected result considering that Influenza A is an RNA virus and should not be detected in DNA derived sequencing datasets, suggesting a potential contamination from reagents [33]. Matches to different Influenza A segments

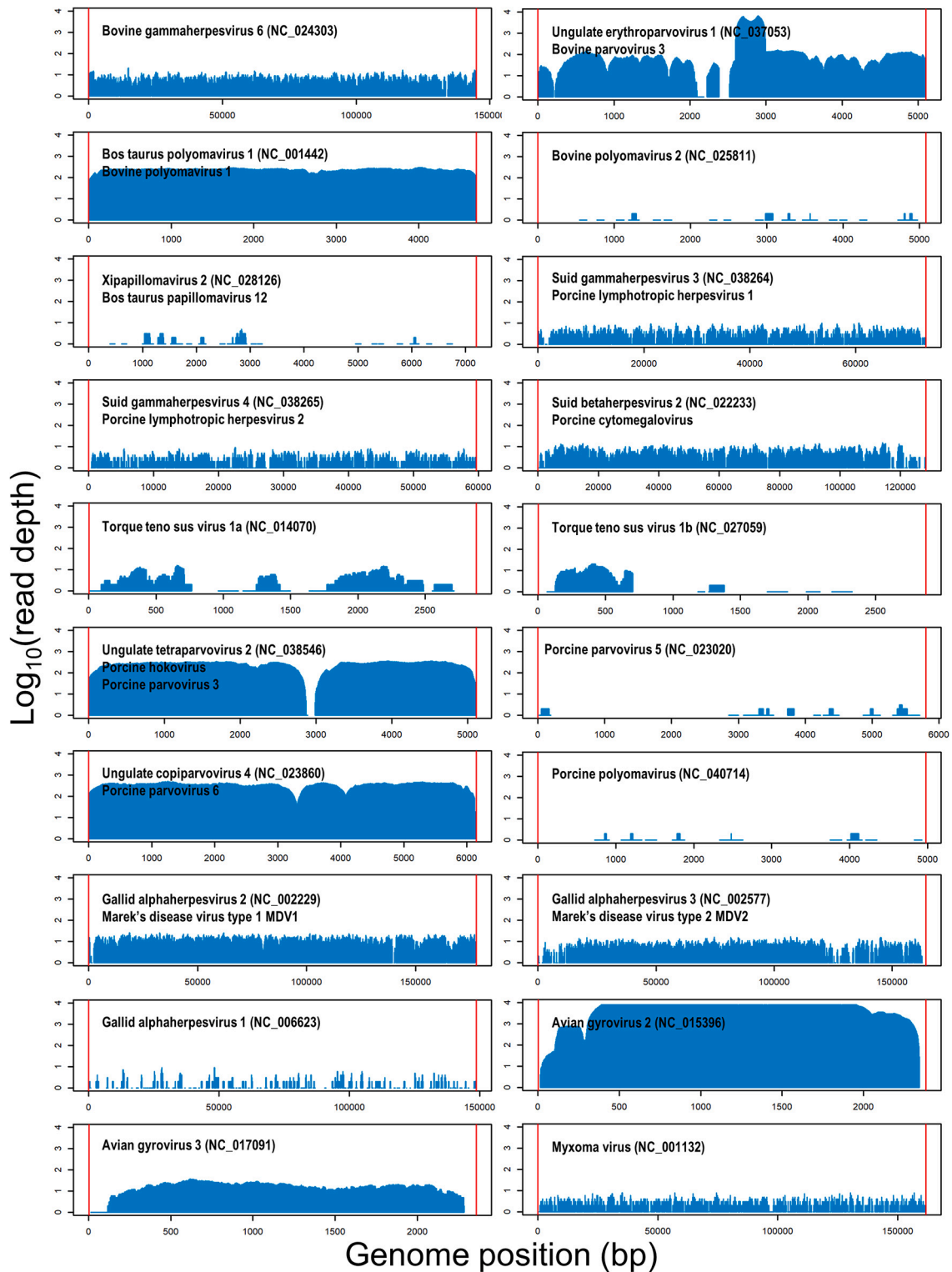


Fig. 3. Genome coverage of some detected viruses

were identified (from nine to one cattle and pig datasets), with usually consistent results derived from the different segments. The pig positive datasets derived from two different sequencing projects whereas the cattle positive datasets derived from the same sequencing project and all positive animals were from the same breed. It is also worth mentioning that not all samples from these pig and cattle sequencing projects were positive for this virus, raising some questions on the source of this virus. Another puzzling result was from the identification of reads assigned to

another RNA virus (Crimean-Congo hemorrhagic fever orthonairovirus; NCBI:NC_005301), known to affect both humans and several animal species (mainly ruminants), that was detected in one chicken dataset.

Again, another group of viruses, known to infect other vertebrate species (but usually not humans), could be considered contaminants. For example, Canine morbillivirus (CDV; also known as Canine distemper virus; NCBI:NC_001921) was identified in a total of 36 datasets from three livestock species (cattle, pig and chicken). This virus is known to

mainly infect species of the Carnivora order and, according to what is currently known, it does not infect the targeted livestock species. Reads assigned to Equid alphaherpesvirus 1 (EHV-1; NCBI:NC_001491) were identified in the cattle, pig and chicken datasets and reads matching Equine infectious anemia virus (EIAV; NCBI:NC_001450) were identified in a chicken dataset. Detailed analysis of the matches identified on the Equid alphaherpesvirus 1 reported to an engineered region of the Equid alphaherpesvirus 1 strain Ab4p_attB_delta_VP22 (GenBank: LC193725.1) including the *tpsL*-neo cassette.

In the case of Human gammaherpesvirus 4 (HuGHV4), it was not clear if its presence in the datasets was due to some sort of contamination. Reads assigned to this virus [also known as Epstein–Barr virus (EBV); NCBI:NC_007605] were identified in four pig datasets (from two different sequencing projects) with a total viral genome coverage >40% and mean depth of sequencing equal to 1.9×. Distribution of reads across the genome is provided in Fig. S2. Very few reads assigned to this virus were also identified in two chicken datasets.

3.4.2. Retroviruses

Retroviruses were considered all together even if in some cases they could be specific for one of the four livestock species. Owing that sequencing depth was on average not very high, it was not possible to retrieve enough information and reads that could provide indication of viral integration on the host genome. For that reason, we did not go into the details of this analysis. Considering all four livestock datasets, a total of 13 different retroviruses were identified. This low number can be in part explained by the fact that sequence reads from endogenous retroviruses mainly mapped on the host genomes. As consequence, this led to underestimate the fraction of retroviruses.

One retrovirus (Avian endogenous retrovirus EAV-HP) was prevalent (50%) in the chicken datasets. Another avian retrovirus (Avian leukosis virus; ALV) was identified almost exclusively in the chicken datasets: 23 out of 24 positive datasets were from chicken whereas one positive dataset was from the pig. Among the six retroviruses that were identified only in the chicken datasets, five (Murine leukemia-related retroviruses, Moloney murine sarcoma virus, Squirrel monkey retrovirus, *Mus musculus* mobilized endogenous polytropic provirus and Equine infection anemia virus) could be considered contaminants derived from cellular lines or vectors used to produce recombinant enzymes commonly utilized in the sequencing steps (analytical contaminants) or in other biotechnology derived products. Another detected retrovirus, Reticuloendotheliosis virus (REV), was an avian specific retrovirus that causes predominately neoplastic disease of turkeys, ducks, chickens and other birds [45,46]. This retrovirus was however identified only in one chicken dataset.

Bovine leukemia virus (BLV), an oncogenic member of the *Deltaretrovirus* genus that is the causative agent of enzootic bovine leukosis (EBL), was identified in one cattle dataset. Four other retroviruses were identified in more than one species: Murine leukemia virus (MLV), a potential contaminant that was identified in 2.8% of all samples; Jaagsiekte sheep retrovirus (JSRV) identified in four cattle samples from two different sequencing projects in different countries and in five chicken samples from three sequencing projects in three different countries; Porcine Type-C oncovirus, identified in one cattle and four Chinese chickens from two sequencing projects; Enzootic nasal tumour virus of goats (ENTV-2; NCBI:NC_004994), identified in three cattle samples (from two sequencing projects in two different countries and three different breeds) and in one pig sample.

3.4.3. Cattle viruses

A total of 13 viruses (14, if we add to this list the retrovirus Bovine Leukemia Virus mentioned above) were detected only in cattle datasets and, according to what is known for these infecting agents, they could be considered specific for the bovine or close ruminant species. Bovine gammaherpesvirus 6 (BoGHV6; NCBI:NC_024303; Fig. 3), also known as Bovine lymphotropic virus (BLV), was identified in 14.5% datasets of

this livestock species, derived from several breeds and different sequencing projects around the world, suggesting a widespread diffusion of this virus. BoGHV6 has been already reported to be ubiquitous in healthy adult cattle and calves [47–49]. Ungulate erythroparvovirus 1, also known as Bovine parvovirus 3 (BPV3) (NCBI:NC_037053), was identified in 2.8% of cattle datasets derived from several cosmopolitan and local breeds of different continents (Table S3). This is a parvovirus that has been previously identified in apparently healthy Brazilian cattle [50]. *Bos taurus* polyomavirus 1 (BPyV1), also known as Bovine polyomavirus 1 (NCBI:NC_001442; Fig. 3) was identified in six cattle datasets, from two continents (Europe and North America) and different breeds and Bovine polyomavirus 2 (BPyV2; NCBI:NC_025811; Fig. 3) was detected in one Holstein cattle dataset from the North America. BPyV1, a prevalent virus in several cattle populations without any relevant pathogenic effects, has been first reported to be a contaminant of bovine serum used in cell culture and has been subsequently used as indicator of environment contamination from bovine excretions [51–54]. BPyV2, identified by few studies, instead, has been recently reported to be the most plausible agent of a novel non-suppurative encephalitis in cattle [55].

Other cattle specific viruses (or known to mainly affect cattle), which can cause some diseases or that can exist in healthy animals, have been identified in few cattle datasets and for that reason they could be considered derived from sporadic infected cases, at least from the data that could be retrieved from our mined datasets. However, some of them have been already reported to be widespread viruses. For example, a few viruses (some of which identified with few reads and for that reason are not reported in Fig. 3) were detected in two datasets: Deltapapillomavirus 4, a small oncogenic papillomavirus with marked tropism for squamous epithelia; Bovine gammaherpesvirus 4, considered a ruminant specific virus causing sub-clinical infections in most cases but also respiratory and clinical reproductive infections in cattle; two poxviruses not in the same cattle samples: Pseudocowpox virus, which causes a common mild infection of the udder and teats of cows with transmission to humans; Bovine papular stomatitis virus that, when in common with Pseudocowpox virus may determine similar lesions in calves and young cattle. It is worth to note that the co-occurrence of Bovine gamma-herpesvirus 4 with Equid alphaherpesvirus 1 was 100% (3 samples out of 3; Table S5).

Several other viruses, sometimes reported to cause infecting diseases, were detected in just one dataset (different for each virus): Xipapillomavirus 1, Xipapillomavirus 2 (NCBI:NC_028126; Fig. 3), Bovine adenovirus E (also known as Bovine adenovirus 6; BAdV-6), Bovine papillomavirus. A surprise was due to a cattle dataset that was positive for Bovine astrovirus (NCBI:NC_024297). Astroviruses are small non-enveloped RNA viruses that several reports have indicated to be infecting agents causing diarrhea or neurological diseases in cattle [56–59].

3.4.4. Pig viruses

A total of 11 porcine specific viruses were identified. Three porcine herpesviruses were quite frequent in the pig datasets. Suid gamma-herpesvirus 3 [SuGHV3; also known as Porcine lymphotropic herpesvirus 1 (PLVH1); NCBI:NC_038264; Fig. 3] and Suid gammaherpesvirus 4 [SuGHV4; also known as Porcine lymphotropic herpesvirus 2 (PLHV2); NCBI:NC_038265; Fig. 3], which are usually considered harmless, were identified in 13.1% and 9.1% sequenced pigs, respectively. Co-occurrence of these two viruses, observed in 11 datasets (Table S5; 18% of the positive PLVH1 and 26% of the positive PLHV2 datasets), was not statistically significant if compared to the occurrence of these viruses in the whole pig datasets (2 × 2 contingency table, fisher exact test: $P > 0.05$). Suid betaherpesvirus 2 [SuBHV2; also known as Porcine cytomegalovirus (PCMV); NCBI:NC_022233; Fig. 3], which causes lifelong latent infections in pigs and occasionally rhinitis in piglets and reproductive disorders in pregnant sows with relevant impacts in the porcine breeding industry and concerns in

xenotransplantation [60,61], was identified in 3.2% investigated pigs. Two virus of the genus *Iotatorquevirus* (Fig. 3), Torque teno sus virus 1a (TTSV1a; NCBI:NC_014070) and Torque teno sus virus 1b (TTSV1b; NCBI:NC_027059), described to be ubiquitous in swine populations and sometimes associated with several pathological conditions [62–65] were identified in 1.9% and 1.5% of the pig datasets, respectively. Two chicken datasets, however, contained a few reads of these viruses. Several parvoviruses [Ungulate protoparvovirus 1, also known as Porcine parvovirus 1 (PPV1); Ungulate tetraparvovirus 2 (PPARV4; NCBI:NC_038546; Fig. 3), also known as Porcine parvovirus 3 (PPV3) or Porcine hokovirus; Porcine partetravirus, recently grouped within the PPV3 group; Porcine parvovirus 5 (PPV5; NCBI:NC_023020; Fig. 3); Ungulate copiparvovirus 4, also known as Porcine parvovirus 6 (PPV6); NCBI:NC_023860; Fig. 3] were identified in the unmapped reads of few sequenced pigs (from six to one) despite the relatively high prevalence of most of them in the pig populations around the world [66–69]. For example, PPV1, considered to be one of the major causes of reproductive failure in pig breeding herds [70], was identified in just one Swiss Large White pig. Porcine polyomavirus (NCBI:NC_040714; Fig. 3), recently described for the first time by a metagenomic sequencing of pooled nasal swabs from pre-weaned pigs with unexplained respiratory disease raised in a Minnesota farm [71], was identified within the unmapped reads of a Chinese Anqing Six-end-white pig [72].

3.4.5. Avian viruses

Nine additional viruses could be considered avian-specific in addition to the three retroviruses already mentioned above (i.e. Avian endogenous retrovirus, Avian leukosis virus and Reticuloendotheliosis virus). Three of these viruses identified only in the chicken datasets were alphaherpesviruses members of the genus *Mardivirus*. These viruses included the devastating pathogenic Marek's disease virus type 1 MDV1, reclassified as Gallid alphaherpesvirus 2 (GaAHV2; NCBI:NC_002229; Fig. 3), the avirulent type 2 (Marek's disease virus serotype 2 MDV2), renamed Gallid alphaherpesvirus 3 (GaAHV3; NCBI:NC_002577; Fig. 3) and the Meleagrid alphaherpesvirus 1 (MeAHV1; NCBI:NC_006623), also known as Turkey herpesvirus (HVT), that is the serotype 3 (MDV-3), which includes avirulent strains of turkeys that can also replicate in chickens [73–76].

These viruses were reported in a total of 24 (5.5%), 40 (9.1%) and 1 (0.2%) of the chicken datasets. The Gallid alphaherpesvirus 1 (NCBI:NC_00662; Fig. 3), the agent of the avian infectious laryngotracheitis [77], was identified in three chickens. Close inspections of the reads derived from the pig and cattle datasets and matching sequences assigned to this virus indicated matches with an engineered strain suggesting that Gallid Alphaherpesvirus 1 was detected only in the sequenced chickens.

Avian gyrovirus 2 (AGyV2; NCBI:NC_015396; Fig. 3), first discovered in diseased chicken in Brazil and now considered a significant potential threat to humans and poultry due to its global dissemination and infectiousness [78–80], was identified in eight sequenced chickens (seven from China and one from Germany). Two of the positive Chinese chickens, despite having a comparable number of sequenced reads, had a number of AGyV2 reads that was 2 to 3 orders of magnitude higher than the other positive samples, suggesting very high viral loads in these two birds. Another virus of the *Gyrovirus* genus, Gyrovirus GyV3 (Gyrovirus homsa1; NCBI:NC_017091; Fig. 3), was identified only in chickens. Eight sequenced Chinese chickens were positive for this virus that is closely related to AGyV2 and that was first identified in human stool and chicken meat and subsequently reported to be ubiquitous in Chinese broiler [81–83]. AGyV2 and Gyrovirus GyV3, however, were detected in different chickens. Other three viruses (Fowlpox virus; Duck hepatitis B virus; and Turkey siadenovirus A, also known as Turkey adenovirus 3) mainly affecting avian species have been identified in just one or two chickens (Fig. S3).

3.4.6. Rabbit viruses

The only rabbit specific virus that was identified in the unmapped reads derived by the sequencing of the genome of this species was the Myxoma virus (NCBI:NC_001132; Fig. 3) that, in *O. cuniculus*, is the causative agent of the fulminant disease myxomatosis [84]. Reads assigned to this virus were detected in only three rabbits (4.5% of the whole rabbit dataset). All positive rabbits were wild rabbits, two sampled in Spain and one sampled in France. One of the two Spanish wild rabbits had a very high viral load compared to the other two rabbits.

4. Discussion

Whole genome sequencing datasets that are generated for many species, including livestock, with the main aims to analyze and characterize their genome features contain additional unexploited information that can be mined for different purposes than those that have been originally envisaged [85–87]. The fraction of reads that does not map to a reference genome contains information on the pathogens that infected the individuals and their tissues from which DNA was extracted [31–34,88,89].

In this study, we mined WGS datasets produced from cattle, pig, chicken and rabbit and retrieved from public databases to obtain a first global landscape picture of the animal virome contained in these resources. Thus far, these WGS resources have not been investigated for this purpose.

The developed bioinformatic pipeline relied on the ENA Browser Data APIs and the ASPERA tool to rapidly evaluate and download WGS (meta) data on an HTC infrastructure. More than 30 TB of DNA sequences were processed in about 1.5 months. The most time consuming and storage demanding step was the mapping of reads on the host genome, the first step of our pipeline. Two approaches – read mapping and de novo assembly – were developed and tested for our purposes. The read mapping approach took advantage of BWA to rapidly screen reads associated to viral genomes whereas the BLASTn based step allowed a more reliable virus detection. Moreover, aiming at deeply characterizing the virome, the reference database of viral sequences (NCBI Refseq Viral Genomes [38]) was expanded to include also non-reference viral genomes (mainly derived from different strains). Despite a direct alignment of unmapped reads via BLASTn over the NCBI nt resource would be resulted appropriate to detect viral sequences solely while avoiding spurious alignments, it would be prohibiting in terms of computational demands. Results obtained by applying the two approaches let emerge the low sensitivity of the latter. This was expected as a high viral load is necessary to produce enough sequencing reads that only through their overlap can generate contigs. For this reason, the virome characterization proceeded mainly via the evaluation of single reads with the read mapping approach. It is worth to note that, as these approaches rely on sequence homology against reference genomes, viruses here detected are only those closely related to already described virus taxa. Therefore, we might have missed novel viruses and virus strains that are quite divergent from those already characterized. Alternative sequence mining approaches (e.g. profile Hidden Markov Models [90]) could be potentially useful to detect some of these missed viruses. A few other biases, which might affect the general interpretation of the obtained results, could be also derived from the way in which we set the bioinformatic pipeline and from intrinsic characteristics of the mined WGS datasets deposited in the public repository. Some false positive results might be due to the relaxed threshold (75%) for sequence identity that we set to claim for the presence of viral DNA in the WGS datasets, which on the other hand, could counterbalance the problem mentioned above related to the quite low sequence identity between different virus strains. A low viral load (or an early stage of virus replication) in the original sample from which DNA sequences were obtained, would prevent the identification of viral DNA, especially if WGS datasets have a low depth of sequencing. Selection of WGS with a higher depth of

sequencing than what we defined here for our mining could be useful to obtain more reliable statistics on the diffusion and distribution of the viruses. Moreover, if WGS datasets underwent a quality control step that removed non-animal DNA sequences before the deposition in the public database, we could not obtain any viral DNA from the mining. Again, the frequency of occurrence of the viruses would not reflect their real diffusion and distribution in the investigated livestock species.

We identified a total of 38 viruses that could be classified as specific infecting agents of the four livestock species or close species (four retroviruses and 34 other viruses). Many other viral sequences could be derived from different sources of contamination as already reported in studies that investigated the human virome or that systematically evaluated the contribution of various laboratory components on viral sequences in high-throughput sequencing datasets [32,33,43]. The presence of bacteriophages and other recombinant viruses can provide information on how the WGS datasets were obtained and could be considered the hallmarks of the sequencing facilities and reagents. However, we did not investigate these details as metadata attached to the mined datasets did not report all needed information to further exploit this DNA fingerprinting. These contaminating viral sequences created, in general, only some noise in the analysis. It was also needed to evaluate the alignment of the reads that for several viruses matched engineered regions that were not annotated in the corresponding virus entry deposited in the virus database used in the bioinformatic identification. For example, *A. californica* multiple nucleopolyhedrovirus, Human betaherpesvirus 5, Human alphaherpesvirus 1, Canine morbillivirus and Equid alphaherpesvirus 1, among several other viruses were reported because the strains deposited in the virus database are engineered with some constructs that are commonly used in molecular biology. In few cases, the presence of human viruses could shed some doubts on the interpretation as being derived from potential unknown contamination sources, especially when zoonotic agents or when some of these viruses could rarely infect livestock. Detailed analyses of some of these cases would be needed, combing information from the coverage of the virus genome, on the virus strains and evaluation of the origin of the WGS datasets, including the tissue from which DNA was extracted.

To disentangle the hypotheses about the presence of some viruses in the livestock datasets, detailed metadata attached to the WGS projects and datasets would be needed. However, in most datasets metadata were incomplete: few datasets reported information on the tissue from which DNA, on the geographical origin of the animals and on the breed; almost none of the datasets reported information of the year of specimen collection, sequencing facilities that generated the sequencing data, protocols and reagents used in the sequencing process and the sex of the animals. In few cases these details could be retrieved or deduced from the original publications in which the datasets were first produced and used. However, the heterogeneity or the lack of several information could not make it possible to construct detailed viral epidemiological information from the identified viruses. It will be important to define new rules to deposit WGS data in public databases that should ask as compulsory metadata useful to better interpret viral data in unmapped reads.

Despite these problems, results were useful to have a preliminary picture of the viral distribution in the four livestock species over many regions and animal populations. The first concern that could be raised derives from the detection of several pathogenic viruses. It is not known if the positive datasets derived from healthy or diseased animals or if this information was available to the investigators. From the analysis of the literature from which these datasets were first described, this information was not reported. This fact can have a few implications from different perspectives: 1) the safety of the personnel who collected, managed and then processed the original specimens, in case of zoonotic viruses; 2) the health state of the animals that are recruited for WGS studies that is unknown in most cases; 3) the real pathogenicity of the detected viruses, shedding some additional lights on this aspect for different strains; 4) the viral load needed to show a diseased state in the

sequenced animals; 5) the relevance of the tissue from which the DNA used in the sequencing project was extracted that could create heterogeneity in the abundance across positive animals; 6) the genetic resistance to viral diseases of some positive animals; 7) the possibility that some animals according to the age or physiological stage (not known in our study) could act as reservoir of viral infections. All these issues can open interesting questions about the epidemiology of several viruses.

Among the 14 cattle specific viruses, Bovine gammaherpesvirus 6 was the most frequently present in the cattle datasets. A few studies that in cattle mined unmapped reads from WGS or transcriptome datasets already reported this virus [86–89], confirming its widespread diffusion in cattle populations [47–49]. Only another virus, Bovine gammaherpesvirus 4, was already identified in cattle transcriptomic data [89]. To our knowledge, none of the other viruses have been previously detected within WGS datasets. Except Ungulate erythroparvovirus 1, which was detected in about 3% of the cattle datasets, all other cattle viruses were reported in very few datasets (one to two animals). As several of them can cause diseases in the infected cattle, it could be possible to speculate that their low frequency in the cattle datasets may indicate that most of the remaining animals that provided specimens for WGS were healthy and not infected and for just few of them this status was doubtful. An RNA virus, not expected considering that the analyzed datasets were from WGS projects, was identified. Its origin or source is not known, and further studies are needed to demonstrate why reads assigned to this virus were present in just one positive cattle dataset.

Two main groups of viruses were also detected in the pig datasets: viruses that were quite frequently detected and viruses that were detected only in few datasets. Among the first group, Suid gammaherpesvirus 3 (PLHV1) and Suid gammaherpesvirus 4 (PLHV2), known to be usually harmless [91,92] were detected in about 10% of the datasets. It was however surprising that Suid gammaherpesvirus 5 (PLHV3) was not detected, considering the high prevalence of this virus reported by epidemiological surveys that specifically targeted Suid gammaherpesviruses [92–94]. This could be derived to a tissue effect (i.e. the tissues usually used for DNA extraction in pig WGS experiments are usually muscle or blood where the PLHV3 load would not be high) or that, actually, the prevalence of PLHV3 in the pig populations is much lower than previously reported. However, as previously pointed out, it is also possible that low viral loads or data filtering procedures led to an underestimation of the prevalence of this virus.

A total of 3.2% of the pig datasets contained reads from Porcine cytomegalovirus (Suid betaherpesvirus 2) demonstrating that it would be possible to use WGS as diagnostic method to identify pigs infected with this immunosuppressive virus that raises a specific attention in xenotransplantation [95]. The identification of Porcine polyomavirus 2 in a WGS datasets of a Chinese pig generated in 2020 [72] indicates that this virus has been recently expanded in another continent from the original identification in 2018 in USA [71]. This result demonstrates in practice the usefulness of our approach to monitor the diffusion of potential harmful viruses.

A total of 12 avian viruses were identified. Among the three avian retrovirus, Reticuloendotheliosis virus (REV), detected in one chicken dataset is quite peculiar. REV was derived from mammalian retroviruses that were accidentally introduced into avian hosts during experimental studies of a malaria parasite in the late 1930s [95]. REVs have subsequently inserted into the genomes of two large DNA viruses that infect birds, generating chimeric viruses that now circulate naturally in poultry and wild birds [96]. It will be also interesting to monitor WGS datasets from other avian species and evaluate if this retrovirus has been reached other birds. Several other highly pathogenic and/or potentially zoonotic avian viruses (Marek's disease virus type 1 known also as Gallid alphaherpesvirus 2; Gallid alphaherpesvirus 1; Avian gyrovirus 2; and Gyrovirus GyV3) were identified in three or more datasets. Gyrovirus GyV3 was the only virus detected exclusively in Chinese chicken whereas the other two viruses confirmed the worldwide distribution as they were detected in animals from more than one continent [74–83].

Only one virus (Myxoma virus) was detected in the rabbit datasets. This might be due to the lower number of the datasets that we mined in this species (~100) compared to the number of datasets analyzed in other three livestock species (~400–500 samples each). In addition, another bias could be derived by the limited number of sequenced rabbit virus genomes that are available compared to the number of sequenced genomes infecting other livestock species. According to virus host database (<https://www.genome.jp/virushostdb/>; accessed on 14/01/2022), only 25 virus species are described to infect *O. cuniculus*, whereas a much higher number of viruses have been described to infect the cattle (n. 204), the pig (n. 185) and the chicken (n. 109). This is probably related to the lower economic relevance of the rabbit and the limited number of studies carried out in this species than in other more relevant species. Thus, it is possible that the identification of only one virus in the rabbit dataset could be also due to the lack of knowledge in this field for the rabbit, for which only a low number of rabbit-infecting viruses have been identified and, in turn, analyzed at the genome level thus far. Myxoma virus in rabbits is one of the best documented examples of host-virus co-evolution [84]. This virus was introduced into wild rabbit populations in Australia, Europe and Great Britain with the aim to control their populations, considered as pests in some countries. This virus was initially highly lethal, killing more than 99% of infected rabbits. Development of resistance was then favored by the emergence of attenuated virus strains which allowed the survival of some rabbits [97]. The identification of three wild rabbits that were positive for this virus might confirm the epidemiological trajectory of the host-virus co-existence. It would be interesting to characterize the strains of the detected Myxoma viruses and the genomic structure of the positive rabbits to evaluate if their genome could have signatures of adaptation to this disease [98].

These examples indicate that viral DNA sequences retrieved mining unmapped reads of WGS datasets, combined with available metadata attached to these datasets, can produce useful epidemiological hints to understand the distribution, diffusion and prevalence of viruses infecting livestock, if the biases discussed above are carefully considered. Datasets from other livestock species can provide additional information useful for this purpose. We will continue to mine new WGS that will be produced over the next years. In addition, to further improve this strategy based on the mining of WGS datasets, we are also developing pipelines to mine transcriptomic datasets (derived from host tissue RNA), which might be interesting sources of RNA viruses.

Early indication of transboundary diffusion of viral diseases can be important to establish reliable countermeasures that can limit the potential damaging impact on livestock production systems and on human health. To this aim, it would be useful to include this mining as a routine approach to monitor viral infectious diseases in livestock with this unconventional approach. This approach could be also applied to mine WGS datasets produced from wild animal species to complete a monitoring strategy against viruses that could infect both domestic and wild species.

Data availability

All publicly available data used for alignment are available via the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena/browsers/>). Supplementary Table S1 reports sample and project identifiers for each evaluated dataset.

Author statement

SB, LF conceived and designed the study. SB performed bioinformatic analyses. SB, GS, MBo, MBa curated and analyzed the data. SB, LF interpreted the results. LF, SB drafted the manuscript. SB, LF, GS, MBo, MBa revised the manuscript. All authors read and approved the final manuscript.

The authors declare that they have no competing interests.

Declaration of Competing Interest

The authors declare that they have no competing interest.

Acknowledgments

This work has received funding from the University of Bologna RFO 2020-2021 programs, from the European Open Science Cloud (EOSC) Secretariat, project “Application of animal genomics and data mining to predict and monitor novel coronavirus potential infections (VirAnimalOne)”, the EGI call for COVID-19 research projects (AnGen1H project) and from the POR FESR Emilia-Romagna 2014-2020 (actions 1.1.4 and 1.2.2—Bando per sostenere progetti di ricerca ed innovazione per lo sviluppo di soluzioni finalizzate al contrasto dell’epidemia da COVID-19—Project LIVESTOCK-STOP-COVI).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jgeno.2022.110312>.

References

- [1] OIE, One Health. <https://www.oie.int/en/forthe-media/onehealth/>, 2020.
- [2] M.E.J. Woolhouse, S. Gowtage-Sequeria, Host range and emerging and reemerging pathogens, *Emerg. Infect. Dis.* 11 (2005) 1842–1847, <https://doi.org/10.3201/eid1112.050997>.
- [3] M. Woolhouse, E. Gaunt, Ecological origins of novel human pathogens, *Crit. Rev. Microbiol.* 33 (2007) 231–242, <https://doi.org/10.1080/10408410701647560>.
- [4] A.D. Luis, D.T.S. Hayman, T.J. O’Shea, P.M. Cryan, A.T. Gilbert, J.R.C. Pulliam, J. N. Mills, M.E. Timonin, C.K.R. Willis, A.A. Cunningham, A.R. Fooks, C. E. Rupprecht, J.L.N. Wood, C.T. Webb, A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proc. Biol. Sci.* 280 (2013) 20122753, <https://doi.org/10.1098/rspb.2012.2753>.
- [5] B.H. Bird, J.A.K. Mazet, Detection of emerging zoonotic pathogens: an integrated one health approach, *Ann. Rev. Anim. Biosci.* 6 (2018) 121–139, <https://doi.org/10.1146/annurev-animal-030117-014628>.
- [6] D. Carroll, P. Daszak, N.D. Wolfe, G.F. Gao, C.M. Morel, S. Morzaria, A. Pablos-Méndez, O. Tomori, J.A.K. Mazet, The global virome project, *Science* 359 (2018) 872–874, <https://doi.org/10.1126/science.aap7463>.
- [7] M. Liverani, J. Waage, T. Barnett, D.U. Pfeiffer, J. Rushton, J.W. Rudge, M. E. Loveinsohn, I. Scoones, R.D. Smith, B.S. Cooper, L.J. White, S. Goh, P. Horby, B. Wren, O. Gundogdu, A. Woods, R.J. Coker, Understanding and managing zoonotic risk in the new livestock industries, *Environ. Health Perspect.* 121 (2013) 873–877, <https://doi.org/10.1289/ehp.1206001>.
- [8] G. Klous, A. Huss, D.J.J. Heederik, R.A. Coutinho, Human-livestock contacts and their relationship to transmission of zoonotic pathogens, a systematic review of literature, *One Health* 2 (2016) 65–76, <https://doi.org/10.1016/j.onehlt.2016.03.001>.
- [9] E.I. Azhar, S.A. El-Kafrawy, S.A. Farraj, A.M. Hassan, M.S. Al-Saeed, A.M. Hashem, T.A. Madani, Evidence for camel-to-human transmission of MERS coronavirus, *N. Engl. J. Med.* 370 (2014) 2499–2505, <https://doi.org/10.1056/NEJMoa1401505>.
- [10] T.T.-Y. Lam, B. Zhou, J. Wang, Y. Chai, Y. Shen, X. Chen, C. Ma, W. Hong, Y. Chen, Y. Zhang, L. Duan, P. Chen, J. Jiang, Y. Zhang, L. Li, L.L.M. Poon, R.J. Webby, D. K. Smith, G.M. Leung, J.S.M. Peiris, E.C. Holmes, Y. Guan, H. Zhu, Dissemination, divergence and establishment of H7N9 influenza viruses in China, *Nature* 522 (2015) 102–105, <https://doi.org/10.1038/nature14348>.
- [11] G.J.D. Smith, D. Vijaykrishna, J. Bahl, S.J. Lycett, M. Worobey, O.G. Pybus, S. K. Ma, C.L. Cheung, J. Raghvani, S. Bhatt, J.S.M. Peiris, Y. Guan, A. Rambaut, Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic, *Nature* 459 (2009) 1122–1125, <https://doi.org/10.1038/nature08182>.
- [12] E. Tambo, O.A. Olalubi, M. Sacko, Rift valley fever epidemic in Niger near border with Mali, *Lancet Infect. Dis.* 16 (2016) 1319–1320, [https://doi.org/10.1016/S1473-3099\(16\)30477-7](https://doi.org/10.1016/S1473-3099(16)30477-7).
- [13] D. Henritzi, P.P. Petric, N.S. Lewis, A. Graaf, A. Pessia, E. Starick, A. Breithaupt, G. Strebelow, C. Luttermann, L.M.K. Parker, C. Schröder, B. Hammerschmidt, G. Herrler, E.G. Beilage, D. Stadlbauer, V. Simon, F. Krammer, S. Wacheck, S. Pesch, M. Schwemmler, M. Beer, T.C. Harder, Surveillance of European domestic pig populations identifies an emerging reservoir of potentially zoonotic Swine Influenza A viruses, *Cell Host Microbe* 28 (2020) 614–627.e6, <https://doi.org/10.1016/j.chom.2020.07.006>.
- [14] H. Sun, Y. Xiao, J. Liu, D. Wang, F. Li, C. Wang, C. Li, J. Zhu, J. Song, H. Sun, Z. Jiang, L. Liu, X. Zhang, K. Wei, D. Hou, J. Pu, Y. Sun, Q. Tong, Y. Bi, K.-C. Chang, S. Liu, G.F. Gao, J. Liu, Prevalent Eurasian avian-like H1N1 swine influenza virus with 2009 pandemic viral genes facilitating human infection, *Proc. Natl. Acad. Sci. U. S. A.* 117 (2020) 17204–17210, <https://doi.org/10.1073/pnas.1921186117>.

- [15] D. Holtkamp, H. Rotto, R. Garcia, The economic cost of major health challenges in large US swine production systems, in: *Proceedings of the American Association of Swine Veterinarians Annual Meeting, Orlando, FL, USA, 2007, 3–6 March*.
- [16] L.L. Schulz, G.T. Tonsor, Assessment of the economic impacts of porcine epidemic diarrhoea virus in the United States, *J. Anim. Sci.* 93 (2015) 5111–5118, <https://doi.org/10.2527/jas.2015-9136>.
- [17] A.K. Wiethoelter, D. Beltrán-Alcudro, R. Kock, S.M. Mor, Global trends in infectious diseases at the wildlife-livestock interface, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) 9662–9667, <https://doi.org/10.1073/pnas.1422741112>.
- [18] V. Richter, K. Lebl, W. Baumgartner, W. Obritzhauser, A. Käsbohrer, B. Pinior, A systematic worldwide review of the direct monetary losses in cattle due to bovine viral diarrhoea virus infection, *Vet. J.* 220 (2017) 80–87, <https://doi.org/10.1016/j.tvjl.2017.01.005>.
- [19] H. Houe, Economic impact of BVDV infection in dairies, *Biologicals* 31 (2003) 137–143, [https://doi.org/10.1016/s1045-1056\(03\)00030-7](https://doi.org/10.1016/s1045-1056(03)00030-7).
- [20] D.J. Holtkamp, J.B. Kliebenstein, E. Neumann, J.J. Zimmerman, H. Rotto, T. K. Yoder, C. Wang, P. Yeske, C.L. Mowrer, C.A. Haley, Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers, *J. Swine Health Prod.* 21 (2013) 72–84.
- [21] R. Palinski, P. Piñeyro, P. Shang, F. Yuan, R. Guo, Y. Fang, E. Byers, B.M. Hause, A novel Porcine Circovirus distantly related to known Circoviruses is associated with porcine dermatitis and nephropathy syndrome and reproductive failure, *J. Virol.* 91 (2017), <https://doi.org/10.1128/JVI.01879-16> e01879–16.
- [22] I. Galindo, C. Alonso, African swine fever virus: a review, *Viruses* 9 (2017) E103, <https://doi.org/10.3390/v9050103>.
- [23] I. Mészáros, F. Olasz, A. Csáögla, P. Tijssen, Z. Zádori, Biology of porcine parvovirus (ungulate parvovirus 1), *Viruses* 9 (2017) E393, <https://doi.org/10.3390/v9120393>.
- [24] C. Zhou, S. Liu, W. Song, S. Luo, G. Meng, C. Yang, H. Yang, J. Ma, L. Wang, S. Gao, J. Wang, H. Yang, Y. Zhao, H. Wang, X. Zhou, Characterization of viral RNA splicing using whole-transcriptome datasets from host species, *Sci. Rep.* 8 (2018) 3273, <https://doi.org/10.1038/s41598-018-21190-7>.
- [25] P.M. Biggs, V. Nair, The long view: 40 years of Marek's disease research and avian pathology, *Avian Pathol.* 41 (2012) 3–9, <https://doi.org/10.1080/03079457.2011.646238>.
- [26] G. Le Gall-Reculé, F. Zwillingstein, S. Boucher, B. Le Normand, G. Plassiart, Y. Portejoie, A. Decors, S. Bertagnoli, J.-L. Guérin, S. Marchandeau, Detection of a new variant of rabbit haemorrhagic disease virus in France, *Vet. Rec.* 168 (2011) 137–138, <https://doi.org/10.1136/vr.d697>.
- [27] S. Bertagnoli, S. Marchandeau, *Myxomatosis, Rev. Sci. Tech.* 34 (549–556) (2015) 539–547.
- [28] S. Temmam, B. Davoust, J.-M. Berenger, D. Raoult, C. Desnues, Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection? *Int. J. Mol. Sci.* 15 (2014) 10377–10397, <https://doi.org/10.3390/ijms150610377>.
- [29] J.L. Gardy, N.J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system, *Nat. Rev. Genet.* 19 (2018) 9–20, <https://doi.org/10.1038/nrg.2017.88>.
- [30] K.T.T. Kwok, D.F. Nieuwenhuijse, M.V.T. Phan, M.P.G. Koopmans, Virus metagenomics in farm animals: a systematic review, *Viruses* 12 (2020) E107, <https://doi.org/10.3390/v12010107>.
- [31] S. Bovo, G. Mazzoni, A. Ribani, V.J. Utzeri, F. Bertolini, G. Schiavo, L. Fontanesi, A viral metagenomic approach on a non-metagenomic experiment: mining next generation sequencing datasets from pig DNA identified several porcine parvoviruses for a retrospective evaluation of viral infections, *PLoS One* 12 (2017), e0179462, <https://doi.org/10.1371/journal.pone.0179462>.
- [32] A. Moustafa, C. Xie, E. Kirkness, W. Biggs, E. Wong, Y. Turpaz, K. Bloom, E. Delwart, K.E. Nelson, J.C. Venter, A. Telenti, The blood DNA virome in 8,000 humans, *PLoS Pathog.* 13 (2017), e1006292, <https://doi.org/10.1371/journal.ppat.1006292>.
- [33] M. Asplund, K.R. Kjartansdóttir, S. Mollerup, L. Vinner, H. Fridholm, J.A. R. Herrera, J. Friis-Nielsen, T.A. Hansen, R.H. Jensen, I.B. Nielsen, S.R. Richter, A. Rey-Iglesia, M.L. Matey-Hernandez, D.E. Alquezar-Planas, P.V.S. Olsen, T. Sicheritz-Pontén, E. Willerslev, O. Lund, S. Brunak, T. Mourier, L.P. Nielsen, J.M. G. Izarzugaza, A.J. Hansen, Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries, *Clin. Microbiol. Infect.* 25 (2019) 1277–1285, <https://doi.org/10.1016/j.cmi.2019.04.028>.
- [34] S. Liu, S. Huang, F. Chen, L. Zhao, Y. Yuan, S.S. Francis, L. Fang, Z. Li, L. Lin, R. Liu, Y. Zhang, H. Xu, S. Li, Y. Zhou, R.W. Davies, Q. Liu, R.G. Walters, K. Lin, J. Ju, T. Korneliusen, M.A. Yang, Q. Fu, J. Wang, L. Zhou, A. Krogh, H. Zhang, W. Wang, Z. Chen, Z. Cai, Y. Yin, H. Yang, M. Mao, J. Shendure, J. Wang, A. Albrechtsen, X. Jin, R. Nielsen, X. Xu, Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history, *Cell* 175 (2018) 347–359.e14, <https://doi.org/10.1016/j.cell.2018.08.016>.
- [35] P.W. Harrison, A. Ahamed, R. Aslam, B.T.F. Alako, J. Burgin, N. Buso, M. Courtot, J. Fan, D. Gupta, M. Haseeb, S. Holt, T. Ibrahim, E. Ivanov, S. Jayathilaka, V. Balavenkataraman Kadhirvelu, M. Kumar, R. Lopez, S. Kay, R. Leinonen, X. Liu, C. O' Cathail, A. Paksresht, Y. Park, S. Pesant, N. Rahman, J. Rajan, A. Sokolov, S. Vijayaraja, Z. Waheed, A. Zyoud, T. Burdett, G. Cochrane, The European nucleotide archive in 2020, *Nucleic Acids Res.* 49 (2020) D82–D85, <https://doi.org/10.1093/nar/gkaa1028>.
- [36] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, *Bioinformatics* 25 (2009) 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324>.
- [37] 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, *Bioinformatics* 25 (2009) 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352>.
- [38] J.R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource, *Nucleic Acids Res.* 43 (2015) D571–D577, <https://doi.org/10.1093/nar/gku1207>.
- [39] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications, *BMC Bioinform.* 10 (2009) 421, <https://doi.org/10.1186/1471-2105-10-421>.
- [40] S. Bovo, V.J. Utzeri, A. Ribani, R. Cabbri, L. Fontanesi, Shotgun sequencing of honey DNA can describe honey bee derived environmental signatures and the honey bee hologenome complexity, *Sci. Rep.* 10 (2020) 9279, <https://doi.org/10.1038/s41598-020-66127-1>.
- [41] D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, *Bioinformatics* 31 (2015) 1674–1676, <https://doi.org/10.1093/bioinformatics/btv033>.
- [42] M.L. MacDonald, S.W. Polson, K.H. Lee, k-mer-based metagenomics tools provide a fast and sensitive approach for the detection of viral contaminants in biopharmaceutical and vaccine manufacturing applications using next-generation sequencing, *MSphere* 6 (2021), <https://doi.org/10.1128/mSphere.01336-20> e01336–20.
- [43] X. Chen, D. Li, Sequencing facility and DNA source associated patterns of virus-mappable reads in whole-genome sequencing data, *Genomics* 113 (2021) 1189–1198, <https://doi.org/10.1016/j.ygeno.2020.12.004>.
- [44] M.J. Strong, E. Blanchard, Z. Lin, C.A. Morris, M. Baddoo, C.M. Taylor, M.L. Ware, E.K. Flemington, A comprehensive next generation sequencing-based virome assessment in brain tissue suggests no major virus - tumor association, *Acta Neuropathol. Commun.* 4 (2016) 71, <https://doi.org/10.1186/s40478-016-0338-z>.
- [45] H.G. Purchase, R.L. Witter, The reticuloendotheliosis viruses, *Curr. Top. Microbiol. Immunol.* 71 (1975) 103–124, https://doi.org/10.1007/978-3-642-66193-8_3.
- [46] A.M. Fadly, G. Zavala, R.L. Witter, *Reticuloendotheliosis*, in: Y.M. Saif (Ed.), *Diseases of Poultry*, John Wiley Sons, New York, 2011.
- [47] J. Rovnak, S.L. Quackenbush, R.A. Reyes, J.D. Baines, C.R. Parrish, J.W. Casey, Detection of a novel bovine lymphotropic herpesvirus, *J. Virol.* 72 (1998) 4237–4242, <https://doi.org/10.1128/JVI.72.5.4237-4242.1998>.
- [48] J.K. Collins, C. Bruns, T.L. Vermedahl, A.L. Schiebel, M.T. Jessen, P.C. Schultheiss, G.M. Anderson, R.P. Dinsmore, R.J. Callan, J.C. DeMartini, Malignant catarrhal fever: polymerase chain reaction survey for ovine herpesvirus 2 and other persistent herpesvirus and retrovirus infections of dairy cattle and bison, *J. Vet. Diagn. Investig.* 12 (2000) 406–411, <https://doi.org/10.1177/104063870001200503>.
- [49] P. Kubis, M. Materniak, J. Kuźmak, Comparison of nested PCR and qPCR for the detection and quantitation of BoHV6 DNA, *J. Virol. Methods* 194 (2013) 94–101, <https://doi.org/10.1016/j.jviromet.2013.08.006>.
- [50] W.M. de Souza, T. Dennis, M.J. Fumagalli, J. Araujo, G. Sabino-Santos, F.G. M. Maia, G.O. Acrani, De A.O.T. Carrasco, M.F. Romeiro, S. Modha, L.C. Vieira, T. Ometto, L.H. Queiroz, E.L. Durigon, M.R.T. Nunes, L.T.M. Figueiredo, R. J. Gifford, Novel parvoviruses from wild and domestic animals in Brazil provide new insights into parvovirus distribution and diversity, *Viruses* 10 (2018) E143, <https://doi.org/10.3390/v10040143>.
- [51] M. Reissig, T.J. Kelly, R.W. Daniel, S.R. Rangan, K.V. Shah, Identification of the stump-tailed macaque virus as a new papovavirus, *Infect. Immun.* 14 (1976) 225–231, <https://doi.org/10.1128/iai.14.1.225-231.1976>.
- [52] W. Coackley, D. Maker, V.W. Smith, A possible bovine polyomavirus, *Arch. Virol.* 66 (1980) 161–166, <https://doi.org/10.1007/BF01314984>.
- [53] D.G. Westcott, J. Ticehurst, M. Chaplin, J.R. Lukey, M. Lucas, The isolation of a virus resembling a polyomavirus from normal calves, *Vet. Microbiol.* 15 (1987) 175–180, [https://doi.org/10.1016/0378-1135\(87\)90142-8](https://doi.org/10.1016/0378-1135(87)90142-8).
- [54] R. Schuurman, B. van Steenis, A. van Strien, J. van der Noordaa, C. Sol, Frequent detection of bovine polyomavirus in commercial batches of calf serum by using the polymerase chain reaction, *J. Gen. Virol.* 72 (Pt 11) (1991) 2739–2745, <https://doi.org/10.1099/0022-1317-72-11-2739>.
- [55] M.M. Hierweger, M.C. Koch, T. Seuberlich, Bovine polyomavirus 2 is a probable cause of non-suppurative encephalitis in cattle, *Pathogens* 9 (2020) E620, <https://doi.org/10.3390/pathogens9080620>.
- [56] C.P. Sharp, W.F. Gregory, C. Mason, De B.M.C. Bronsvort, P.M. Beard, High prevalence and diversity of bovine astroviruses in the faeces of healthy and diarrhoeic calves in South West Scotland, *Vet. Microbiol.* 178 (2015) 70–76, <https://doi.org/10.1016/j.vetmic.2015.05.002>.
- [57] T. Seuberlich, D. Wüthrich, S. Selimovic-Hamza, C. Drögemüller, A. Oevermann, R. Bruggmann, I. Bouzalas, Identification of a second encephalitis-associated astrovirus in cattle, *Emerg. Microb. Infect.* 5 (2016), e71, <https://doi.org/10.1038/em.2017.56>.
- [58] S. Selimovic-Hamza, S. Sanchez, H. Philibert, E.G. Clark, T. Seuberlich, Bovine astrovirus infection in feedlot cattle with neurological disease in western Canada, *Can. Vet. J.* 58 (2017) 601–603.
- [59] F. Giannitti, R.D. Caffarena, P. Pesavento, F.A. Uzal, L. Maya, M. Fraga, R. Colina, M. Castells, The first case of bovine astrovirus-associated encephalitis in the southern hemisphere (Uruguay), uncovers evidence of viral introduction to the Americas from Europe, *Front. Microbiol.* 10 (2019) 1240, <https://doi.org/10.3389/fmicb.2019.01240>.
- [60] A.H. Corner, D. Mitchell, R.J. Julian, E.B. Meads, A generalized disease in piglets associated with the presence of cytoplasmic inclusions, *J. Comp. Pathol.* 74 (1964) 192–199, [https://doi.org/10.1016/s0368-1742\(64\)80024-2](https://doi.org/10.1016/s0368-1742(64)80024-2).

- [61] A.W. Tucker, D. Galbraith, P. McEwan, D. Onions, Evaluation of porcine cytomegalovirus as a potential zoonotic agent in xenotransplantation, *Transplant. Proc.* 31 (1999) 915, [https://doi.org/10.1016/s0041-1345\(98\)01833-8](https://doi.org/10.1016/s0041-1345(98)01833-8).
- [62] N.E. McKeown, M. Fenaux, P.G. Halbur, X.J. Meng, Molecular characterization of porcine TT virus, an orphan virus, in pigs from six different countries, *Vet. Microbiol.* 104 (2004) 113–117, <https://doi.org/10.1016/j.vetmic.2004.08.013>.
- [63] T. Kekkarainen, M. Sibila, J. Segalés, Prevalence of swine Torque teno virus in post-weaning multisystemic wasting syndrome (PMWS)-affected and non-PMWS-affected pigs in Spain, *J. Gen. Virol.* 87 (2006) 833–837, <https://doi.org/10.1099/vir.0.81586-0>.
- [64] T. Kekkarainen, J. Segalés, Torque teno sus virus in pigs: an emerging pathogen? *Transbound. Emerg. Dis.* 59 (Suppl. 1) (2012) 103–108, <https://doi.org/10.1111/j.1865-1682.2011.01289.x>.
- [65] Y. Lee, C.-M. Lin, C.-R. Jeng, H.-W. Chang, C.-C. Chang, V.F. Pang, The pathogenic role of torque teno sus virus 1 and 2 and their correlations with various viral pathogens and host immunocytes in wasting pigs, *Vet. Microbiol.* 180 (2015) 186–195.
- [66] D. Cadar, M. Lórinçz, T. Kiss, D. Novosel, K. Podgórska, Z. Beeskei, T. Tuboly, A. Cságola, Emerging novel porcine parvoviruses in Europe: origin, evolution, phylogenetics and phylogeography, *J. Gen. Virol.* 94 (2013) 2330–2337, <https://doi.org/10.1099/vir.0.055129-0>.
- [67] J. Sun, L. Huang, Y. Wei, Y. Wang, D. Chen, W. Du, H. Wu, C. Liu, Prevalence of emerging porcine parvoviruses and their co-infections with porcine circovirus type 2 in China, *Arch. Virol.* 160 (2015) 1339–1344, <https://doi.org/10.1007/s00705-015-2373-7>.
- [68] S.F. Cotmore, M. Agbandje-McKenna, M. Canuti, J.A. Chiorini, A.-M. Eis-Hubinger, J. Hughes, M. Mietzsch, S. Modha, M. Ogliastrò, J.J. Péznes, D.J. Pintel, J. Qiu, M. Soderlund-Venermo, P. Tattersall, P. Tijssen, null Ictv report consortium, ICTV virus taxonomy profile: Parvoviridae, *J. Gen. Virol.* 100 (2019) 367–368, <https://doi.org/10.1099/jgv.0.001212>.
- [69] D. Milek, A. Woźniak, M. Guzowska, T. Stadejek, Detection patterns of porcine parvovirus (PPV) and novel porcine parvoviruses 2 through 6 (PPV2-PPV6) in polish swine farms, *Viruses* 11 (2019) E474, <https://doi.org/10.3390/v11050474>.
- [70] A.F. Streck, U. Truyen, Porcine parvovirus, *Curr. Issue Mol. Biol.* 37 (2020) 33–46, <https://doi.org/10.21775/cimb.037.033>.
- [71] B.M. Hause, C. Smith, B. Bishop, C. Stewart, R. Simonson, Complete genome sequence of a porcine polyomavirus from nasal swabs of pigs with respiratory disease, *Genom. Announc.* 6 (2018), <https://doi.org/10.1128/genomeA.00344-18>.
- [72] W. Zhang, M. Yang, M. Zhou, Y. Wang, X. Wu, X. Zhang, Y. Ding, G. Zhao, Z. Yin, C. Wang, Identification of signatures of selection by whole-genome resequencing of a chinese native pig, *Front. Genet.* 11 (2020), 566255, <https://doi.org/10.3389/fgene.2020.566255>.
- [73] A.J. Davison, R. Eberle, B. Ehlers, G.S. Hayward, D.J. McGeoch, A.C. Minson, P. E. Pellett, B. Roizman, M.J. Studdert, E. Thiry, The order Herpesvirales, *Arch. Virol.* 154 (2009) 171–177, <https://doi.org/10.1007/s00705-008-0278-4>.
- [74] K.A. Schat, V. Nair, Marek's disease, in: D.E. Swayne, J.R. Glisson, L.R. McDougald, L.K. Nolan, D.L. Suarez (Eds.), *Diseases of Poultry*, Wiley-Blackwell, 2013, pp. 513–674.
- [75] R.D. Chacón, C.S. Astolfi-Ferreira, M.B. Guimarães, L.N. Torres, D.I. De la Torre, L. R.M. de Sá, A.J. Piantino Ferreira, Detection and molecular characterization of a natural coinfection of Marek's disease virus and reticuloendotheliosis virus in Brazilian backyard chicken flock, *Vet. Sci.* 6 (2019) E92, <https://doi.org/10.3390/vetsci6040092>.
- [76] L.D. Bertzbach, A.M. Conradie, Y. You, B.B. Kaufer, Latest insights into Marek's disease virus pathogenesis and tumorigenesis, *Cancers (Basel)* 12 (2020) E647, <https://doi.org/10.3390/cancers12030647>.
- [77] W. Fuchs, J. Veits, D. Helferich, H. Granzow, J.P. Teifke, T.C. Mettenleiter, Molecular biology of avian infectious laryngotracheitis virus, *Vet. Res.* 38 (2007) 261–279, <https://doi.org/10.1051/vetres:200657>.
- [78] F.A.M. Rijsewijk, H.F. Dos Santos, T.F. Teixeira, S.P. Cibulski, A.P.M. Varela, D. Dezen, A.C. Franco, P.M. Roehe, Discovery of a genome of a distant relative of chicken anemia virus reveals a new member of the genus Gyrovirus, *Arch. Virol.* 156 (2011) 1097–1100, <https://doi.org/10.1007/s00705-011-0971-6>.
- [79] J. Ye, X. Tian, Q. Xie, Y. Zhang, Y. Sheng, Z. Zhang, C. Wang, H. Zhu, Y. Wang, H. Shao, A. Qin, Avian Gyrovirus 2 DNA in fowl from live poultry markets and in healthy humans, China, *Emerg. Infect. Dis.* 21 (2015) 1486–1488, <https://doi.org/10.3201/eid2108.150203>.
- [80] S. Yao, T. Tuo, X. Gao, C. Han, Y. Li, Y. Gao, Y. Zhang, C. Liu, X. Qi, H. Gao, Y. Wang, X. Wang, Avian gyrovirus 2 in poultry, China, 2015–2016, *Emerg. Microb. Infect.* 5 (2016), e112, <https://doi.org/10.1038/emi.2016.113>.
- [81] D.K.W. Chu, L.L.M. Poon, S.S.S. Chiu, K.H. Chan, E.M. Ng, I. Bauer, T.K. Cheung, I. H.Y. Ng, Y. Guan, D. Wang, J.S.M. Peiris, Characterization of a novel gyrovirus in human stool and chicken meat, *J. Clin. Virol.* 55 (2012) 209–213, <https://doi.org/10.1016/j.jcv.2012.07.001>.
- [82] T.G. Phan, L. Li, M.G. O'Ryan, H. Cortes, N. Mamani, I.J.O. Bonkougou, C. Wang, C.M. Leutenegger, E. Delwart, A third gyrovirus species in human faeces, *J. Gen. Virol.* 93 (2012) 1356–1361, <https://doi.org/10.1099/vir.0.041731-0>.
- [83] G. Li, S. Yuan, M. He, M. Zhao, X. Hao, M. Song, L. Zhang, C. Qiao, L. Huang, L. Zhang, C. Li, G. Wang, Z. Cheng, Emergence of gyrovirus 3 in commercial broiler chickens with transmissible viral proventriculitis, *Transbound. Emerg. Dis.* 65 (2018) 1170–1174.
- [84] P.J. Kerr, S.M. Best, Myxoma virus in rabbits, *Rev. Sci. Tech.* 17 (1998) 256–268, <https://doi.org/10.20506/rst.17.1.1081>.
- [85] H. Tae, E. Karunasena, J.H. Bavarva, L.J. McIver, H.R. Garner, Large scale comparison of non-human sequences in human sequencing data, *Genomics* 104 (2014) 453–458, <https://doi.org/10.1016/j.ygeno.2014.08.009>.
- [86] 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, <https://doi.org/10.1186/s12864-015-2313-7>.
- [87] M. Sangiovanni, I. Granata, A.S. Thind, M.R. Guarracino, From trash to treasure: detecting unexpected contamination in unmapped NGS data, *BMC Bioinform.* 20 (2019) 168, <https://doi.org/10.1186/s12859-019-2684-x>.
- [88] T. Usman, F. Hadlich, W. Demasius, R. Weikard, C. 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 (2017) 36–42, <https://doi.org/10.1016/j.ygeno.2016.11.009>.
- [89] L. Buggiotti, Z. Cheng, D.C. Wathes, null GplusE consortium, Mining the unmapped reads in bovine RNA-Seq data reveals the prevalence of Bovine Herpes Virus-6 in European dairy cows and the associated changes in their phenotype and leucocyte transcriptome, *Viruses* 12 (2020) E1451, <https://doi.org/10.3390/v12121451>.
- [90] L.S. Oliveira, A. Gruber, Rational design of profile hidden Markov models for viral classification and discovery, in: I. Helder, N. (Eds.), *Bioinformatics, Exon Publications*, Brisbane (AU), 2021, <https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch9>.
- [91] X.J. Meng, Emerging and re-emerging swine viruses, *Transbound. Emerg. Dis.* 59 (Suppl. 1) (2012) 85–102, <https://doi.org/10.1111/j.1865-1682.2011.01291.x>.
- [92] G. Franzo, M. Drigo, M. Legnardi, L. Grassi, M.L. Menandro, D. Pasotto, M. Cecchinato, C.M. Tucciarone, Porcine Gammaherpesviruses in Italian commercial swine population: frequent but harmless, *Pathogens* 10 (2021) 47, <https://doi.org/10.3390/pathogens10010047>.
- [93] K.J. McMahon, D. Minihan, E.M. Campion, S.T. Loughran, G. Allan, F. McNeilly, D. Walls, Infection of pigs in Ireland with lymphotropic gamma-herpesviruses and relationship to postweaning multisystemic wasting syndrome, *Vet. Microbiol.* 116 (2006) 60–68, <https://doi.org/10.1016/j.vetmic.2006.03.022>.
- [94] B. Chmielewicz, M. Goltz, T. Franz, C. Bauer, S. Brema, H. Ellerbrok, S. Beckmann, H.-J. Rziha, K.-H. Lahrmann, C. Romero, B. Ehlers, A novel porcine gammaherpesvirus, *Virology* 308 (2003) 317–329, [https://doi.org/10.1016/s0042-6822\(03\)00006-0](https://doi.org/10.1016/s0042-6822(03)00006-0).
- [95] J. Denner, Xenotransplantation and porcine cytomegalovirus, *Xenotransplantation* 22 (2015) 329–335, <https://doi.org/10.1111/xen.12180>.
- [96] A.M. Niewiadomska, R.J. Gifford, The extraordinary evolutionary history of the reticuloendotheliosis viruses, *PLoS Biol.* 11 (2013), e1001642, <https://doi.org/10.1371/journal.pbio.1001642>.
- [97] P.J. Kerr, I.M. Cattadori, J. Liu, D.G. Sim, J.W. Dodds, J.W. Brooks, M.J. Kennett, E. C. Holmes, A.F. Read, Next step in the ongoing arms race between myxoma virus and wild rabbits in Australia is a novel disease phenotype, *Proc. Natl. Acad. Sci. U. S. A.* 114 (2017) 9397–9402, <https://doi.org/10.1073/pnas.1710336114>.
- [98] J.M. Alves, M. Carneiro, J.Y. Cheng, A. Lemos de Matos, M.M. Rahman, L. Loog, P. F. Campos, N. Wales, A. Eriksson, A. Manica, T. Strive, S.C. Graham, S. Afonso, D. J. Bell, L. Belmont, J.P. Day, S.J. Fuller, S. Marchandeu, W.J. Palmer, G. Queney, A.K. Surridge, F.G. Vieira, G. McFadden, R. Nielsen, M.T.P. Gilbert, P.J. Esteves, N. Ferrand, F.M. Jiggins, Parallel adaptation of rabbit populations to myxoma virus, *Science* 363 (2019) 1319–1326, <https://doi.org/10.1126/science.aau7285>.