Virus Evolution, 2024, **10(1)**, veae099
DOI: https://doi.org/10.1093/ve/veae099
Advance Access Publication 23 November 2024
Reflections

Endogenous viral elements: insights into data availability and accessibility

Muriel Ritsch[1,2,‡], Nadja Brait[2,3,‡], Erin Harvey[2,4], Manja Marz[1,2,5,6,7,8], Sebastian Lequime[2,3,*]

[1]RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, Jena 07743, Germany
[2]European Virus Bioinformatics Center, Leutragraben 1, Jena 07743, Germany
[3]Cluster of Microbial Ecology, Groningen Institute for Evolutionary Life Sciences, University of Groningen, P.O. Box 11103, Groningen 9700 CC, The Netherlands
[4]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia
[5]German Center for Integrative Biodiversity Research (iDiv), Puschstrasse 4, Halle-Jena-Leipzig 04103, Germany
[6]Michael Stifel Center Jena, Ernst-Abbe-Platz 2, Jena 07743, Germany
[7]Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, Fürstengraben 1, Jena, Thüringen 07745, Germany
[8]Fritz Lipmann Institute-Leibniz Institute on Aging, Beutenbergstraße 11, Jena 07745, Germany

[‡]These authors contributed equally to this work.
*Corresponding author. European Virus Bioinformatics Center, Leutragraben 1, Jena 07743, Germany. Cluster of Microbial Ecology, Groningen Institute for Evolutionary Life Sciences, University of Groningen, P.O. Box 11103, Groningen 9700 CC, The Netherlands. E-mail: s.j.j.lequime@rug.nl

Abstract

Endogenous viral elements (EVEs) are remnants of viral genetic material endogenized into the host genome. They have, in the last decades, attracted attention for their role as potential contributors to pathogenesis, drivers of selective advantage for the host, and genomic remnants of ancient viruses. EVEs have a nuanced and complex influence on both host health and evolution, and can offer insights on the deep evolutionary history of viruses. As an emerging field of research, several factors limit a comprehensive understanding of EVEs: they are currently underestimated and periodically overlooked in studies of the host genome, transcriptome, and virome. The absence of standardized guidelines for ensuring EVE-related data availability and accessibility following the FAIR ('findable, accessible, interoperable, and reusable') principles obstructs our ability to gather and connect information. Here, we discuss challenges to the availability and accessibility of EVE-related data and propose potential solutions. We identified the biological and research focus imbalance between different types of EVEs, and their overall biological complexity as genomic loci with viral ancestry, as potential challenges that can be addressed with the development of a user-oriented identification tool. In addition, reports of EVE identification are scattered between different subfields under different keywords, and EVE sequences and associated data are not properly gathered in databases. While developing an open and dedicated database might be ideal, targeted improvements of generalist databases might provide a pragmatic solution to EVE data and metadata accessibility. The implementation of these solutions, as well as the collective effort by the EVE scientific community in discussing and setting guidelines, is now drastically needed to lead the development of EVE research and offer insights into host–virus interactions and their evolutionary history.

Keywords: endogenous viral element; EVE; endogenous retroviral element; ERV; non-retroviral EVE; nrEVE; database; data availability; data accessibility

Introduction

The exponential increase in data production facilitated by multi-omics approaches leads to difficulties in establishing a standardized and effective framework for (meta)data sharing and management. To address this growing problem, FAIR guiding principles were outlined in 2016, stating that biological data should be 'findable, accessible, interpretable and reusable' (Wilkinson et al. 2016). These guidelines are increasingly being enforced by third-party repositories, publishers, and scientific communities at large, as part of an effort toward Open Science. Some scientific fields, especially emerging ones, have not yet been able to implement all guidelines due to field-specific challenges, such as a lack of well-established practices and widely accepted standards. One such emerging field is the research on endogenous viral elements (EVEs).

EVEs are remnants of viral genetic material endogenized into the host genome. They arise from the integration of viral DNA (or DNA copies of viral RNA) into the host's germ cells genome, and can be inherited by its progeny. Although different terms for EVE subsets exist, they can be classified into two main categories: endogenous retroviruses (ERVs) and non-retroviral endogenous viral elements (nrEVEs) (Fig. 1). Retroviruses are characterized by a mandatory integration into their hosts' genome as part of

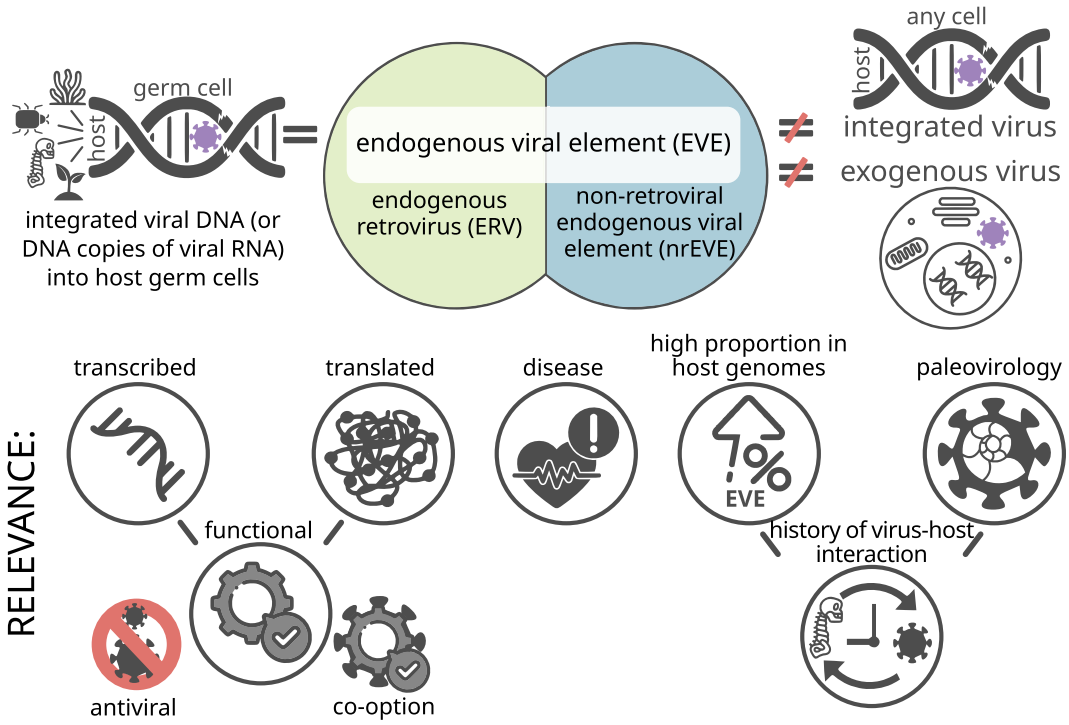

Figure 1. EVE is a generic term, encompassing "endogenous retrovirus" (ERV) and "non-retroviral endogenous viral element" (nrEVE). EVEs can be confused with integrated viruses (found in any cell type) and exogenous viruses (located outside or inside the cells but not in the host's genome). EVEs can be transcribed and translated and can play a significant role in host–virus interactions. Some EVEs are linked to antiviral immune response or provide new functions to the host through co-option. ERVs have also been associated with a range of diseases, such as cancer, autoimmune diseases, and neurological disorders. In addition, EVEs can outnumber host protein-coding genes in the genome and hold significant value for paleovirological studies.

their replication cycle, facilitating their potential endogenization (for an overview, see Johnson 2019). ERVs may constitute a substantial portion of the host's genome: in humans, for example, they make up approximately 9%, three times more than cellular protein-coding genes (Lander et al. 2001, Katzourakis and Tristem 2005). Additionally, EVEs can also arise from non-retroviral RNA or DNA viruses, collectively known as nrEVEs (Zhdanov 1975, Klenerman et al. 1997, Geuking et al. 2009). Here, endogenization is likely facilitated by reverse transcription and integration through retrotransposable cellular (e.g. LINE) elements or non-homologous recombination (Young and Samulski 2001, Horie et al. 2010). While EVEs derived from non-retroviruses are less common, they have been found in almost all eukaryotic organisms, including unicellular eukaryotes (Nelson et al. 2021, Bellas et al. 2023), plants (Jakowitsch et al. 1999, Chiba et al. 2011; see review Takahashi et al. 2019), vertebrates (Katzourakis et al. 2010, Horie et al. 2010; see review Kapusta and Suh 2017), and arthropods (Palatini et al. 2017, Tassetto et al. 2019, Whitfield et al. 2017; see reviews Blair et al. 2020, Wallau 2022). This widespread presence provides a unique perspective on the long-lasting relationships of viruses with their hosts, allowing the exploration of their deep evolutionary history (Fig. 1) (Aiewsakun and Katzourakis 2015, Aswad and Katzourakis 2012; Holmes 2011) (see reviews Horie and Tomonaga 2019, Barreat and Katzourakis 2022).

In addition, EVEs have been shown to have a variable and complex influence on host health and evolution. Elevated expressions of ERVs have shown connections with some forms of cancer (Weiss 2006; see review Kassiotis 2014), autoimmune diseases (see reviews Morandi et al. 2017, Greenig 2019, Latifi et al. 2022), and neurological disorders such as Alzheimer's disease and schizophrenia (Sankowski et al. 2019, Jönsson et al. 2021). EVEs may also provide new functions to the host, a phenomenon known as co-option (Fig. 1). A prominent example is Syncytin, derived from an expressed ERV's envelope gene, now playing a pivotal role in placental development in mammals (Boyd et al. 1993, Venables et al. 1995, Sha et al. 2000). EVEs have also been shown to be involved in the host defense mechanisms against viral infections in insects, plants, and mammals (see reviews: Aswad and Katzourakis 2012, Pooggin 2018, Broecker and Moelling 2019, Ophinni et al. 2019, Bonning and Saleh 2021, Rosendo Machado et al. 2021).

The absence of standardized guidelines for ensuring data availability and accessibility is a significant barrier to deepening our understanding of EVEs as important components of host–virus interactions. Without data being made available following FAIR principles, there are significant challenges in determining if specific EVEs are present at particular genomic loci, how diverse they are within a specific host or host group, and whether specific viral taxa are more likely to endogenize. Our current inability to address these questions impedes our understanding of the biological function and the broader ecological and evolutionary implications of EVEs. Here, we outline and discuss the existing challenges and limitations in data availability and accessibility in EVE research and propose strategies to overcome them.

Data availability

We refer to data availability as the consistent and reliable presence of research data and resources to be used by the scientific community. It plays a pivotal role in implementing FAIR principles by ensuring that genomic information has been reliably

detected and processed as well as establishing clear and standardized metadata. However, challenges persist in young fields like EVE research, where early-stage data collection and processing already pose significant challenges.

Challenge: the imbalance between ERVs and nrEVEs

ERVs are more prominently featured in scientific research and discussions compared to nrEVEs (for an overview see Tugnet et al. 2013, Kassiotis 2014, Morandi et al. 2017, Küry et al. 2018, Latifi et al. 2022). The abundance and clear genomic characteristics of ERVs, a common occurrence in host genomes, led to an earlier discovery in the 1960s (Weiss 1967) and increasing awareness in the scientific community. In contrast, nrEVEs have been discovered only relatively recently, with first reports dating back to the late 1990s and early 2000s (Bejarano et al. 1996, Crochu et al. 2004). They are also more diverse, potentially spanning the complete virosphere, except, by definition, retroviruses (Katzourakis et al. 2010). In addition, nrEVEs have attracted less attention, possibly due to low abundance in the human genome, and currently no known associations with health-related conditions.

The discrepancy in abundance between nrEVEs and ERVs prompts the question of whether this is due to the rarity of nrEVE integration, the increased difficulty in their detection, their lower visibility in the scientific community, or a combination of these factors. It is most likely that there are significantly more ERVs than nrEVEs due to the necessity for retroviruses to integrate into the host genome to complete their life cycle, as well as the recurrent copying of ERVs into host genomes through duplications or horizontal gene transfer. ERVs use a "copy-and-paste" mechanism by transcription, followed by reverse transcription, and reintegration into the host genome, which increases their copy number.

A large-scale study in 2010 detected nearly 500 nrEVEs in 40 vertebrate and 4 invertebrate genomes (Katzourakis et al. 2010). Three years later, nearly 90,000 ERVs from 60 vertebrate host genomes have been identified in a single study (Hayward et al. 2013). This apparent bias leads to an overall lack of awareness or reports of the existence of nrEVEs and regular exclusions of newly detected nrEVEs as simple contaminants in virus discovery or genome assemblies (Mifsud et al. 2022).

Challenge: the biological complexity of EVEs

While ERVs contain conserved retrovirus-specific sequences, making their in silico identification by similarity-based detection approaches easier, the initial discovery of nrEVEs can be difficult due to frequent partial endogenization, and high sequence and genomic structure diversity. In addition, after endogenization, selective pressures on the EVE's sequence can be altered. Indeed, EVEs are subjected to a slower evolutionary rate than their exogenous virus predecessors, with around 10^{-9} substitutions per base pair per year for mammals (Kumar and Subramanian 2002). Except for EVEs serving or acquiring a function within the host, there is typically little to no selective pressure acting to maintain their sequence integrity and functionality. As a result, these sequences accumulate mutations disrupting existing viral open reading frames in the form of frameshifts, premature stop codons, and missense protein mutations. This leads to an ever-increasing divergence from the viral predecessors and degradation of the original viral genomic structure and sequence, which may result in EVEs appearing more similar to host noncoding regions than to any exogenous viruses. In addition, EVE sequences (Palatini et al. 2020) or even their presence (Crava et al. 2021) may vary between host populations.

However, depending on the time since endogenization, these mutations may not have accumulated sufficiently to differentiate the endogenized sequence from an exogenous virus, which may also contain frameshifts and premature stop codons (Hill and Brierley 2023). The inability to recover a virus's complete genome may indicate an EVE but may also indicate low exogenous viral abundance. Moreover, errors introduced during sequencing or data processing can lead to sequence contamination, which makes distinguishing genuine EVEs from artifactual sequences difficult. Flanking host genes are a reliable indicator of an EVE sequence but may not always be present in the sequenced transcript or reconstructed contig. The best indicator of an EVE would be its annotation in a high-quality, well-annotated host genome assembly, but these are often unavailable for nonmodel organisms, and variations between individuals or populations may also exist. All these difficulties lead to uncertainty in accurately identifying, describing, and annotating EVEs, which might discourage researchers from submitting them to public databases.

Challenge: computational tools for EVE detection

Until now, researchers have used varying methodologies and criteria for identifying and characterizing EVEs, including Southern blot, RT-PCRs, fluorescence *in situ* hybridization, and sequencing (Bejarano et al. 1996, Crochu et al. 2004, Filloux et al. 2015, Da Fonseca et al. 2016, Palatini et al. 2022). The lack of distinguishing characteristics demands specialized and nuanced approaches for data acquisition, classification, and annotation. There is currently no field-wide preferred bioinformatic tool or recommended analysis pipeline for identifying EVEs in genomic or metagenomic datasets. This is also of concern for metatranscriptomic virus discovery which has grown rapidly as a field in the past decade (Harvey and Holmes 2022).

For nrEVEs identification within genomes, multiple approaches have been developed (Kryukov et al. 2019, Ter Horst et al. 2019, Nelson et al. 2021, Kinsella et al. 2022, Palatini et al. 2022, Pienaar et al. 2022, Kinsella and van der Hoek 2023), but most lack directly shareable pipelines and only two provide readily available tools (Fig. 2). The Database Integrated Genome Screening has recently been employed to identify EVEs in metazoan genomes (Zhu et al. 2018, Blanco-Melo et al. 2024), though this tool is yet to be used outside of the group that developed it. For a more targeted approach, CAULIFINDER presents an intuitive tool for the identification, annotation, grouping, and classification of EVEs derived specifically from the *Caulimoviridae* family in plant genomes (Vassilieff et al. 2022). Similarly, RepeatMasker, a tool that can identify and mask repeated DNA sequences, e.g. LINEs, SINEs, and ERVs, is typically used for ERV detection at the genome level (Smit et al. 1996). Currently, there is no pipeline or tool designed for the specific identification of EVEs at a transcriptomic level.

Most tools aiming at identifying recent viral integration sites in somatic cells in the context of disease, such as ViFi (Nguyen et al. 2018), Virus-Clip (Ho et al. 2015), or VirusFinder (Wang et al. 2013), or tools used for or aiming at identifying exogenous viruses, such as Kraken2 (Wood et al. 2019), Centrifuge (Kim et al. 2016), or VirusFinder (Wang et al. 2013), can be used to detect EVEs. However, the vast majority do not consider EVEs in their analysis, despite their high risk of being misinterpreted as, respectively, a recently integrated virus or as exogenous viruses (Brait et al. 2024). For example, Edgar et al. (2022) estimated that approximately 1% of RNA virus sequences in their wide-scale metatranscriptomic analysis could be EVEs (Edgar et al. 2022). Developing a pipeline or a specialized tool for categorizing novel virus sequences as either likely exogenous or likely endogenous from transcriptomic

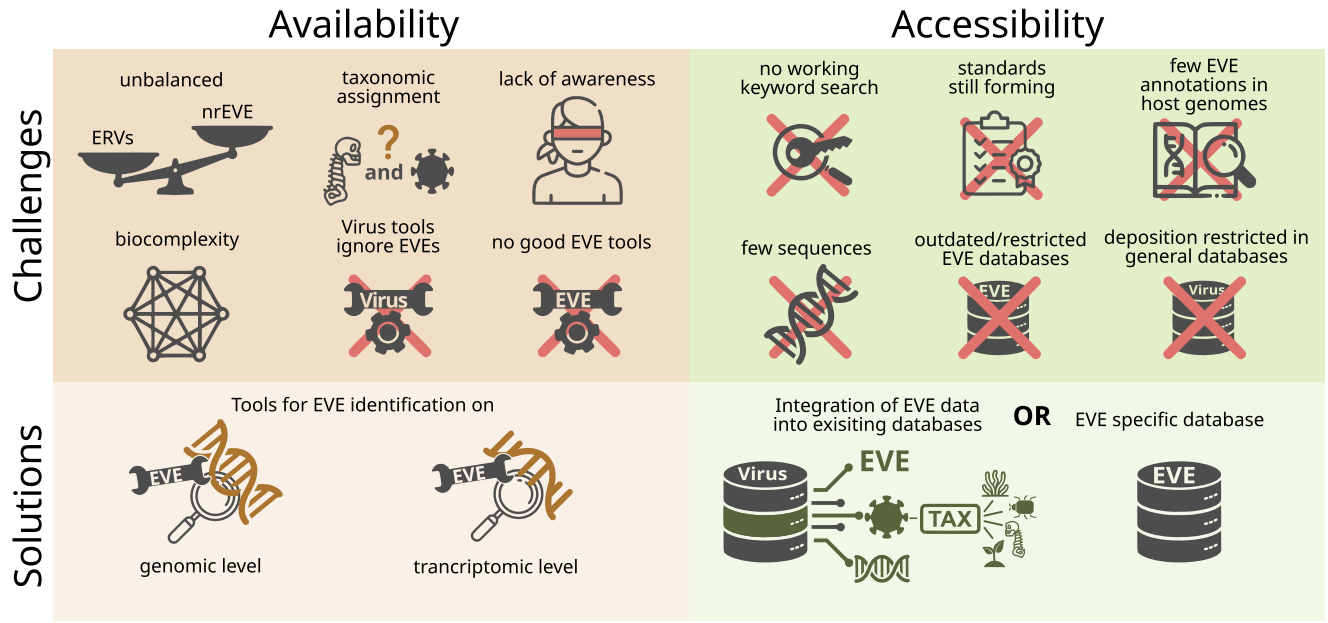

Figure 2. Challenges (above) are present in the context of EVE data availability and accessibility, to which we propose solutions (below). (Left): Regarding data availability, a notable obstacle is the significant disparity between data related to ERVs and nrEVEs. In addition, the taxonomic assignment of EVEs as part host or as part virus is unclear. EVEs are often seen as contaminants and thus excluded from analyses or mistakenly assigned as being of host or exogenous virus origin. While there are some methods for genome-level EVE identification, there are currently no dedicated tools for identifying EVEs in eukaryotic meta-transcriptomic data. As a proposed solution, we advocate for the development of intuitive and accessible tools for both genomic and transcriptomic EVE identification. (Right): Data accessibility is key, as unfound data is akin to nonexistent data. There are no effective keyword searches for scientific papers or databases, and standards for presenting and storing EVE data are still forming. Moreover, annotations in host genomes for EVEs are rare, and there are only a few accessible EVE sequences. This is primarily due to the lack of updated and accessible EVE databases and the challenge of ensuring proper access to EVE sequences in more generalist databases. To address this, we propose two solutions: (i) improved integration of EVE data into existing databases, which involves tagging EVE sequences and assigning taxonomic associations for hosts and viruses, and (ii) a less realistic but potentially transformative implementation of a dedicated EVE-specific database that incrementally meets the needs of the EVE research community.

data would assist both researchers involved in virus discovery and studying EVEs.

Finally, some tools for assessing the completeness of an exogenous virus transcript could be used for EVE detection, such as CheckV (Nayfach et al. 2021), VIBRANT (Kieft et al. 2020), and viralComplete (Antipov et al. 2020), but these were primarily designed for identifying environmental viruses and are less suitable for eukaryotic virus discovery. CheckV, for example, relies on high-quality reference host genomes and similar virus sequences, which limits its use in identifying highly divergent viral sequences.

Solutions: intuitive tools in EVE characterization

To address the challenges outlined above effectively, computational tools that cater to different challenges are required. It is crucial to distinguish between transcriptomic and genomic data, as well as between ERVs and nrEVE detection. Common standards for both ERVs and nrEVEs identification could include approaches that can detect EVEs with high sensitivity and specificity, and account for variations in sequence, genomic structure, and integration site. However, tools for ERV detection should be tailored to detect characteristic features such as long terminal repeats (LTRs), high copy numbers, and retroviral genes (e.g. gag, pol, env). In contrast, tools designed for nrEVE identification should be flexible enough to account for high viral diversity and primarily partial genomic integration. Moreover, it should provide standardized metadata containing information on the predecessor virus taxonomic group, the host, and the integration site, thus assisting in EVE classification and annotation. This could help for future seamless integration into existing genomic databases.

Finally, EVE identification tools should be intuitive and accessible to enable researchers with varying levels of expertise to harness the tool's capabilities effectively. The development and wide use of optimal EVE identification tools would significantly advance EVE research, enhance data availability, and foster a more standardized approach to EVE characterization. This would lead to an improvement of our understanding of EVEs as important components of host–virus interactions, as well as harness them as invaluable markers for studying the deep evolutionary history of viruses.

Data accessibility

Here, we define accessibility, within the FAIR framework, as data being not only available but also easily locatable, retrievable, and usable by the scientific community. The inaccessibility of data is akin to its nonexistence. Accessibility promotes inclusivity, allowing researchers of diverse expertise levels and scientific backgrounds to access and use this data effectively, enhancing interdisciplinary research efforts. Improving accessibility involves standardization efforts, as well as designing or modifying user-oriented databases and repositories tailored to EVE-related genomic (meta)data, which are major challenges for EVE research.

Challenge: retrievability of EVE studies and associated data

The need for precise and widely adopted keywords concerning EVEs is a recurring challenge (Fig. 2). Current literature presents

several classifications and nomenclatures, leading to ambiguities. Proposals for EVE nomenclature have been suggested (Gifford et al. 2018), but they are not currently widely used. While "endogenous viral element" has become popular, alternative terms exist, for example, the outdated "non-retroviral integrated RNA virus sequences" (NIRVs), which have not gained traction due to a lack of clear relationship to the generic "EVEs." Most publications on nrEVEs and ERVs are discoverable only through the "EVE" keyword rather than more specific ones, which will mostly yield ERV studies due to the previously described imbalance between the two types. Some terms may also pertain to particular hosts, viruses, or viral proteins, such as "EBLN" (endogenous bornavirus-like nucleoprotein), and these studies are often not discoverable with the keywords "EVE." We believe widely adopted and precise keywords such as "ERVs" and "nrEVEs" could easily solve this issue.

In addition, standardized communication practices regarding EVE sequences and associated metadata are still evolving. This includes aspects such as the methodology used for their identification, their position within the host genome, details of the host genome assembly used, and their association with exogenous viral sequences. Finally, EVEs may only be a minor publication focus, leading to limited data sharing: EVE sequences may not be included in associated data or only presented in tabular form specifying contigs and positions. Even if sufficient information is provided to retrieve associated EVE sequences, they are, as of now, not indexed as a searchable resource, and metadata cannot be easily uploaded. As a result, direct access to the data is not possible and can only be obtained by a thorough and time-consuming review of the associated publication and its supplementary material.

Challenge: current EVE-specific databases

A search through both literature and database-containing platforms, such as "re3data.org," "FAIRsharing," "The Database Commons," "ELIXIR bio.tools," and "Integbio Database Catalog," using keywords like "EVEs," "ERVs," and "endogenous" (Ison et al. 2016, Sansone et al. 2019, Ma et al. 2022) identified six EVE-specific databases, of which only four were still accessible at the time of writing (Table 1).

Three of these databases are restricted in scope, such as DbHERV-Res and HERVd on human ERVs, or limited, such as gEVE on only 20 host genomes. On the other hand, the pEVE database contains predicted EVEs, primarily DNA virus-derived, in over 4000 eukaryotic genomes (Kryukov et al. 2019). Except for HERVd, no database had been updated since publication, with two (ERE and FabriEVEs) becoming inaccessible, and none allowed external contributions. The lack of a centralized, updated, and open EVE database poses a challenge (Fig. 2).

Challenge: limitations in generalist databases

Without satisfactory EVE databases, generalist ones might offer effective solutions for storing EVE-related data. Many EVE sequences can be found in the International Nucleotide Sequence Database Collaboration (INSDC), but provide limited standardized associated metadata and keywords. Retrieving one or all EVE sequences for a particular organism or derived from a specific virus taxon is, therefore, challenging. In addition, the deposition of EVE sequences in the INSDC is complicated by its requirement of standardized formatting and comprehensive annotations to maintain data accuracy and quality. These restrictive rules conflict with the biological nature of EVEs: many only present partial open-reading frames or include complex features such as stop codons or frameshifts which can lead to a series of tedious resubmissions for validation. These hurdles prevent submission and lead to EVEs being mainly described in manuscripts or compiled in associated supplementary tables. A less constraining solution would be to deposit EVEs in host genomic annotations; however, this relies on the availability and quality of a reference genome. Even if they can be annotated, it does not guarantee the effortless retrieval of EVEs or associated viral metadata across annotated genomes.

A potential solution would be the incorporation of EVEs in a general virus database. Currently, however, no centralized virus database has fulfilled the necessary requirements, including comprehensive coverage of all viruses, adherence to FAIR principles, user-friendliness, and meticulous curation (Ritsch et al. 2023).

Solutions: integration versus dedicated databases

The lack of an active and open EVE database or reasonable alternative limits researchers' ability to assess the known presence of EVEs in a particular host species. For example, *Parvoviridae*, *Filoviridae*, and *Bornaviridae* EVEs have been thoroughly described in marsupials (Harding et al. 2021), but these sequences remain difficult to retrieve due to the lack of an appropriate database. This limits the possibility of incorporating these sequences into subsequent research or identifying similar or identical EVE sequences in transcriptomic research or virus discovery analysis (Fig. 2).

We propose two possible solutions: improved integration in generalist databases, such as GenBank, or the creation of a dedicated EVE database (Fig. 2). The latter would provide an optimal design for EVE-related data, but the former would probably be a more practical and reasonable approach. In addition, the use of databases familiar to the broader research community could enhance their adoption, and establish additional links between already deposited virus, host, and EVE data.

Table 1. List of accessible EVE databases

Name (reference)	Host screened	EVE type	Limitations	Last update	Number of EVEs
ERE (Kao et al. 2012)	Inaccessible				
gEVE (Nakagawa and Takahashi 2016)	20 Genomes-19 mammalia	All	Open Reading Frame-related	2016	736,771
DbHERV-Res (Ito et al. 2017)	Human	ERVs	Just the regulation elements	2017	445
FabriEVEs (Zhong et al. 2019)	Inaccessible				
pEVE database (Kryukov et al. 2019)	4,102 Eukaryotic genomes	nrEVE	Mainly dsDNA	2020	6,300,132
HERVd (Paces 2002)	Human	Mainly ERVs	-	2021	565,471

The minimum requirement for an existing database is a tag or field that clearly defines an entry as an EVE. Another relatively simple enhancement would be incorporating two taxonomic identifications, assigning the EVE to the host and the virus based on sequence similarity to the current closest relative. Viral taxonomy itself can be challenging and may lead to issues in assigning EVEs to particular viral species. As an alternative, the taxonomic assignment could be on a higher taxonomic level, such as genus, family, or higher. Such dual classification acknowledges that EVEs can be classified as host elements with viral ancestry or viruses residing within a host genome. Most databases, however, do not support setting two taxonomic identifiers to a single entry. A pragmatic solution would be to denote the second taxonomic identifier (e.g. of the virus) as a tag, meaning the EVE entry can be retrieved in virus-centered database searches. Additional information could be incorporated for each entry, such as linking to the appropriate viral sequence used for EVE detection, facilitating backtracking, indicating which part of the virus has been integrated, and integrating motifs or similarities to retrotransposons.

This information can and should be incorporated into a centralized, dedicated, and curated EVE database, enhancing the feasibility and efficiency of meta-analyses. Conducted at the host level, it could allow the investigation of the genomic distribution and conserved integration patterns within the host's genomic landscape. Studies at the virus level could reveal commonly endogenized viral segments or open reading frames, assess EVE abundance within genomes, and identify evolutionary constraints. Such a database would revolutionize our understanding of EVEs by preventing redundancy, opening new research lines, and enhancing collaboration.

Conclusion

The study of EVEs is inherently interdisciplinary and requires collaboration between virology, genomics, and bioinformatics. EVEs are important components of the complex interaction between hosts and their viruses, and they offer insights into their evolutionary history (Aiewsakun and Katzourakis 2015, Frank and Feschotte 2017, Barreat and Katzourakis 2022). However, research on EVEs is limited by challenges in data availability and accessibility, such as the lack of intuitive identification tools or databases to store and retrieve EVE-related (meta)data. Here, we discussed two ideal solutions, a dedicated, user-oriented EVE detection tool and a centralized EVE database meeting the community's diverse needs. The feasibility of the latter in a limited funding landscape is, at best, uncertain. More pragmatically, improvements to existing generalist databases, such as the use of an EVE tag or the possibility of assigning both host and virus taxonomic identifications to the record, could already address some of these challenges.

Improving the availability and accessibility of EVE-related data also contributes to limiting the risk of errors. Indeed, inaccuracies or errors in EVE sequences or associated metadata have the potential to lead to the misclassification of EVEs, exogenous viruses, or host sequences, which are difficult to correct. This can lead to erroneous biological insights and propagate errors that become increasingly entangled, drastically impacting any subsequent study. The identification and correction of these errors can only be achieved with access to substantial data for validation, a critical element currently only fulfilled for some EVEs.

Due to the rapid pace of technological and bioinformatics advancements, certain standards governing genetic data, but also more specifically viral (meta)data, have become outdated. Failure to update or reassess these standards in light of technological progress can impede discovery and innovation. For instance, the requirement to demonstrate cellular infection and the presence of viral particles led to the initial rejection of complete genome RNA virus discoveries from metatranscriptomic data. Similarly, *in silico* detected EVEs often required validation by PCR and current formatting standards imposed by INSDC prevent a fast and easy upload of partial and/or mutation-prone EVEs. It is therefore important to review and revise standards to foster innovation in emerging biological fields.

Finally, mandating standard "user-friendly" pipelines may temporarily improve inclusiveness as suggested. While we propose solutions for some of the challenges we highlight, we believe that the establishment of standardized practices following FAIR guidelines can only come from a collective effort of the EVE scientific community. Targeted conferences and collaborative initiatives can promote discussions, set guidelines, and disseminate knowledge. Furthermore, long-term progress in the field will depend on training scientists to develop computational skills in line with technological advances such as machine learning and AI. Enhancing the availability and accessibility of EVE-related (meta)data is now crucial to allow research to bring together the currently scattered reports. This would strengthen our understanding of the biology of EVEs and harness the valuable insights they offer into host–virus interactions.

Acknowledgements

The authors would like to thank Thomas Hackl for a helpful discussion during the early stages of this research paper.

Author contributions

Conceptualization, M.R., N.B., E.H., and S.L; Investigation, M.R., N.B., E.H., and S.L.; Writing—Original Draft Preparation, M.R., N.B., and E.H.; Writing—Review & Editing, M.R., N.B., E.H., M.M., and S.L.; Writing—Polishing, N.B., S.L.; Visualization, M.R.; and Supervision, M.M. and S.L.

Conflict of interest: None declared.

Funding

This research was funded by the TMWWDG grant "DigLeben" number 5575/10-9 TMWBDG (M.M. and M.R.); scholarship from the Studienstiftung des deutschen Volkes (M.R.); the EU Horizon 2020 grant "VIROINF" number 955974 (M.M.) and AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China (E.H.).

Data availability

No data was generated for this study.

References

Aiewsakun P, Katzourakis A. Endogenous viruses: connecting recent and ancient viral evolution. *Virology* 2015;**479–480**:26–37.

Antipov D, Raiko M, Lapidus A *et al.* Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics (Oxford, England)* 2020;**36**:4126–29.

Aswad A, Katzourakis A. Paleovirology and virally derived immunity. *Trends Ecol Evol* 2012;**27**:627–36.

Barreat JGN, Katzourakis A. Paleovirology of the DNA viruses of eukaryotes. *Trends Microbiol* 2022;**30**:281–92.

Bejarano ER, Khashoggi A, Witty M *et al*. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc Natl Acad Sci* 1996;**93**:759–64.

Bellas C, Hackl T, Plakolb M-S *et al*. Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proc Natl Acad Sci USA* 2023;**120**:e2300465120.

Blair CD, Olson KE, Bonizzoni M. The widespread occurrence and potential biological roles of endogenous viral elements in insect genomes. *Curr Issues Mol Biol* 2020;**34**:13–30.

Blanco-Melo D, Campbell MA, Zhu H *et al*. A novel approach to exploring the dark genome and its application to mapping of the vertebrate virus fossil record. *Genome Biol* 2024;**25**:120.

Bonning BC, Saleh M-C. The interplay between viruses and RNAi pathways in insects. *Annu Rev Entomol* 2021;**66**:61–79.

Boyd MT, Bax CMR, Bax BE *et al*. The human endogenous retrovirus ERV-3 is upregulated in differentiating placental trophoblast cells. *Virology* 1993;**196**:905–09.

Brait N, Hackl T, Morel C *et al*. A tale of caution: how endogenous viral elements affect virus discovery in transcriptomic data. *Virus Evol* 2024;**10**:vead088.

Broecker F, Moelling K. Evolution of immune systems from viruses and transposable elements. *Front Microbiol* 2019;**10**:51.

Chiba S, Kondo H, Tani A *et al*. Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathog* 2011;**7**:e1002146.

Crava CM, Varghese FS, Pischedda E *et al*. Population genomics in the arboviral vector *Aedes aegypti* reveals the genomic architecture and evolution of endogenous viral elements. *Mol Ecol* 2021;**30**:1594–611.

Crochu S, Cook S, Attoui H *et al*. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of aedes spp. mosquitoes. *J Gen Virol* 2004;**85**:1971–80.

Da Fonseca GC, De Oliveira LFV, De Morais GL *et al*. Unusual RNA plant virus integration in the soybean genome leads to the production of small RNAs. *Plant Sci* 2016;**246**:62–69.

Edgar RC, Taylor J, Lin V *et al*. Petabase-scale sequence alignment catalyses viral discovery. *Nature* 2022;**602**:142–47.

Filloux D, Murrell S, Koohapitagtam M *et al*. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. *Virus Evol* 2015;**1**:vev002.

Frank JA, Feschotte C. Co-option of endogenous viral sequences for host cell function. *Curr Opin Virol* 2017;**25**:81–89.

Geuking MB, Weber J, Dewannieux M *et al*. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science* 2009;**323**:393–96.

Gifford RJ, Blomberg J, Coffin JM *et al*. Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* 2018;**15**:59.

Greenig M. HERVs, immunity, and autoimmunity: understanding the connection. *PeerJ* 2019;**7**:e6711.

Harding EF, Russo AG, Yan GJH *et al*. Ancient viral integrations in marsupials: a potential antiviral defence. *Virus Evol* 2021;**7**:veab076.

Harvey E, Holmes EC. Diversity and evolution of the animal virome. *Nat Rev Microbiol* 2022;**20**:Article6.

Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proc Natl Acad Sci USA* 2013;**110**:20146–51.

Hill CH, Brierley I. Structural and functional insights into viral programmed ribosomal frameshifting. *Annu Rev Virol* 2023;**10**:217–42.

Ho DW, Sze KM, Ng IO. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget* 2015;**6**:20959–63.

Holmes EC. The evolution of endogenous viral elements. *Cell Host Microbe* 2011;**10**:368–77.

Horie M, Honda T, Suzuki Y *et al*. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 2010;**463**:84–87.

Horie M, Tomonaga K. Paleovirology of bornaviruses: what can be learned from molecular fossils of bornaviruses. *Virus Res* 2019;**262**:2–9.

Ison J, Rapacki K, Ménager H *et al*. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res* 2016;**44**:D38–D47.

Ito J, Sugimoto R, Nakaoka H *et al*. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet* 2017;**13**:e1006883.

Jakowitsch J, Mette MF, van der Winden J *et al*. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc Natl Acad Sci* 1999;**96**:13241–46.

Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* 2019;**17**:355–70.

Jönsson ME, Garza R, Sharma Y *et al*. Activation of endogenous retroviruses during brain development causes an inflammatory response. *EMBO J* 2021;**40**:e106423.

Kao D, Hsu K, Chiu S *et al*. ERE database: a database of genomic maps and biological properties of endogenous retroviral elements in the C57BL/6J mouse genome. *Genomics* 2012;**100**:157–61.

Kapusta A, Suh A. Evolution of bird genomes—a transposon's-eye view. *Ann NY Acad Sci* 2017;**1389**:164–85.

Kassiotis G. Endogenous retroviruses and the development of cancer. *J Immunol* 2014;**192**:1343–49.

Katzourakis A, Gifford RJ, Malik HS. Endogenous viral elements in animal genomes. *PLoS Genet* 2010;**6**:e1001191.

Katzourakis A, and Tristem M. Phylogeny of human endogenous and exogenous retroviruses. In: Sverdlov ED (ed.), *Retroviruses and Primate Genome Evolution*. Boca Raton: CRC Press. 2005, 186–203.

Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;**8**:90.

Kim D, Song L, Breitwieser FP *et al*. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;**26**:1721–29.

Kinsella CM, Deijs M, Becker C *et al*. Host prediction for disease-associated gastrointestinal cressdnaviruses. *Virus Evol* 2022;**8**:veac087.

Kinsella CM, van der Hoek L. Vertebrate-tropism of a cressdnavirus lineage implicated by poxvirus gene capture. *Proc Natl Acad Sci USA* 2023;**120**:e2303844120.

Klenerman P, Hengartner H, Zinkernagel RM. A non-retroviral RNA virus persists in DNA form. *Nature* 1997;**390**:298–301.

Kryukov K, Ueda MT, Imanishi T *et al*. Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. *Virus Res* 2019;**262**:30–36.

Kumar S, Subramanian S. Mutation rates in mammalian genomes. *Proc Natl Acad Sci USA* 2002;**99**:803–08.

Küry P, Nath A, Créange A *et al*. Human endogenous retroviruses in neurological diseases. *Trends Mol Med* 2018;**24**:379–94.

Lander ES, Linton LM, Birren B *et al*. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.

Latifi T, Zebardast A, Marashi SM. The role of human endogenous retroviruses (HERVs) in multiple sclerosis and the plausible interplay between HERVs, Epstein-Barr virus infection, and vitamin D. *Mult Scler Relat Disord* 2022;**57**:103318.

Ma L, Zou D, Liu L *et al*. Database commons: a catalog of worldwide biological databases. *Genom Proteom Bioinform* 2022;**21**1054–8.

Mifsud JCO, Costa VA, Petrone ME *et al*. Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians. *Virus Evol* 2022;**9**:veac124.

Morandi E, Tanasescu R, Tarlinton RE *et al*. The association between human endogenous retroviruses and multiple sclerosis: a systematic review and meta-analysis. *PloS One* 2017;**12**:e0172415.

Nakagawa S, Takahashi MU. gEVE: A genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database* 2016;**2016**:baw087.

Nayfach S, Camargo AP, Schulz F *et al*. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnol* 2021;**39**:578–85.

Nelson DR, Hazzouri KM, Lauersen KJ *et al*. Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. *Cell Host Microbe* 2021;**29**:250–266.e8.

Nguyen ND, Deshpande V, Luebeck J *et al*. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res* 2018;**46**:3309–25.

Ophinni Y, Palatini U, Hayashi Y *et al*. piRNA-guided CRISPR-like immunity in eukaryotes. *Trends Immunol* 2019;**40**:998–1010.

Paces J. HERVd: database of human endogenous retroviruses. *Nucleic Acids Res* 2002;**30**:205–06.

Palatini U, Masri RA, Cosme LV *et al*. Improved reference genome of the arboviral vector aedes albopictus. *Genome Biol* 2020;**21**:215.

Palatini U, Miesen P, Carballar-Lejarazu R *et al*. Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors aedes aegypti and aedes albopictus. *BMC Genom* 2017;**18**:512.

Palatini U, Pischedda E, Bonizzoni M. Computational methods for the discovery and annotation of viral integrations. *Methods Mol Biol* 2022;**2509**:293–313.

Pienaar RD, Gilbert C, Belliardo C *et al*. First evidence of past and present interactions between viruses and the black soldier fly, hermetia illucens. *Viruses* 2022;**14**:1274.

Pooggin MM. Small RNA-omics for plant virus identification, virome reconstruction, and antiviral defense characterization. *Front Microbiol* 2018;**9**:2779.

Ritsch M, Cassman NA, Saghaei S *et al*. Navigating the landscape: a comprehensive review of current virus databases. *Viruses* 2023;**15**:1834.

Rosendo Machado S, van der Most T, Miesen P. Genetic determinants of antiviral immunity in dipteran insects—compiling the experimental evidence. *Dev Comp Immunol* 2021;**119**:104010.

Sankowski R, Strohl JJ, Huerta TS *et al*. Endogenous retroviruses are associated with hippocampus-based memory impairment. *Proc Natl Acad Sci* 2019;**116**:25982–90.

the FAIRsharing Community, Sansone S-A, McQuilton P, Rocca-Serra P *et al*. FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnol* 2019;**37**:358–67.

Sha M, Lee X, Li X *et al*. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 2000;**403**:785–89.

Smit AFA, Hubley R, Green P. *RepeatMasker Open*-3.0. 1996. Computer software. http://www.repeatmasker.org (8 November 2024, date last accessed).

Takahashi H, Fukuhara T, Kitazawa H *et al*. Virus latency and the impact on plants. *Front Microbiol* 2019;**10**:2764.

Tassetto M, Kunitomi M, Whitfield ZJ *et al*. Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. *eLife* 2019;**8**:e41244.

Ter Horst AM, Nigg JC, Dekker FM *et al*. Endogenous viral elements are widespread in arthropod genomes and commonly give rise to PIWI-interacting RNAs. *J Virol* 2019;**93**:e02124–18.

Tugnet N, Rylance P, Roden D *et al*. Human Endogenous Retroviruses (HERVs) and autoimmune rheumatic disease: is there a link? *Open Rheumatol J* 2013;**7**:13–21.

Vassilieff H, Haddad S, Jamilloux V *et al*. CAULIFINDER: a pipeline for the automated detection and annotation of caulimovirid endogenous viral elements in plant genomes. *Mobile DNA* 2022;**13**:31.

Venables PJ, Brookes SM, Griffiths D *et al*. Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function. *Virology* 1995;**211**:589–92.

Wallau GL. RNA virus EVEs in insect genomes. *Curr Opin Insect Sci* 2022;**49**:42–47.

Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* 2013;**8**:e64465.

Weiss R. Spontaneous virus production from "non-virus producing" Rous sarcoma cells. *Virology* 1967;**32**:719–23.

Weiss RA. The discovery of endogenous retroviruses. *Retrovirology* 2006;**3**:67.

Whitfield ZJ, Dolan PT, Kunitomi M *et al*. The diversity, structure, and function of heritable adaptive immunity sequences in the aedes aegypti genome. *Curr Biol* 2017;**27**:3511–3519.e7.

Wilkinson MD, Dumontier M, Aalbersberg IJ *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**:257.

Young SM, Samulski RJ. Adeno-associated virus (AAV) site-specific recombination does not require a rep-dependent origin of replication within the AAV terminal repeat. *Proc Natl Acad Sci* 2001;**98**:13525–30.

Zhdanov VM. Integration of viral genomes. *Nature* 1975;**256**:471–73.

Zhong J, Zhu J, Xu Z *et al*. FabriEVEs: A dedicated platform for endogenous viral elements in fishes, amphibians, birds, reptiles and invertebrates. *Evol Biol* 2019.

Zhu H, Dennis T, Hughes J *et al*. Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database. 2018.

Virus Evolution, 2024, 10(1), veae099 , DOI: https://doi.org/10.1093/ve/veae099, Advance Access Publication 23 November 2024, Reflections