Peer Community Journal

Section: Infections

Review article

Published
2025-10-03

Cite as
Damien Richard and Nils Poulicard (2025) *Data mining of public genomic repositories: harnessing off-target reads to expand microbial pathogen genomic resources*, Peer Community Journal, 5: e110.

Correspondence
damien.richard@ird.fr

Peer-review
Peer reviewed and recommended by PCI Infections, https://doi.org/10.24072/pci.infections.100248

Data mining of public genomic repositories: harnessing off-target reads to expand microbial pathogen genomic resources

Damien Richard [1] and Nils Poulicard [1]

Volume 5 (2025), article e110

https://doi.org/10.24072/pcjournal.637

Abstract

As sequencing technologies become more affordable and genomic databases expand continuously, the reuse of publicly available sequencing data emerges as a powerful strategy for studying microbial pathogens. Indeed, raw sequencing reads generated for the study of a given organism often contain reads originating from the associated microbiota. This review explores how such off-target reads can be detected and used for the study of microbial pathogens. We present genomic data mining as a method to identify relevant sequencing runs from petabase-scale databases, highlighting recent methodological advances that allow efficient database querying. We then briefly outline methods designed to retrieve relevant data and associated metadata, and provide an overview of common downstream analysis pipelines. We discuss how such approaches have (i) expanded the known genetic diversity of microbial pathogens, (ii) enriched our understanding of their spatiotemporal distribution, and (iii) highlighted previously unrecognized ecological interactions involving microbial pathogens. However, these analyses often rely on the completeness and accuracy of accompanying metadata, which remain highly variable. We detail common pitfalls, including data contamination and metadata misannotations, and suggest strategies for result interpretation. Ultimately, while data mining cannot replace dedicated studies, it constitutes an essential and complementary tool for microbial pathogen research. Broader utility will depend on improved data standardization and systematic genomic monitoring across ecosystems.

[1] PHIM Plant Health Institute, Univ Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Introduction

As sequencing goes down in price, public genomic databases go up in size (Sayers et al., 2023). The most widely used databases are the DNA Data Bank of Japan, GenBank, and the European Nucleotide Archive, all involved in the International Nucleotide Sequence Database Collaboration. These repositories contain assembled genomes and metagenomes, as well as raw sequencing reads. These sequencing reads have sometimes not yet fully been explored outside of the scientific context for which they were generated. Notably, pathogens are often represented in off-target reads of studies initially designed to sequence their host or their host's transcriptome in which a given pathogen may have gone unnoticed due to symptomless infections, co-infection, or lack of interest/knowledge from the researchers and institutions that initiated the sequencing effort. Specificities in the biological origin (lab-grown vs. wild organism), wet-lab methods (tissue-specific sequencing, enrichment or purification steps…) or type of sequencing (Whole Genome, Exome, RNA, Amplicon, or Single-Cell Sequencing for example) influence the presence and relevance of such off-target reads for a given pathogen. The wealth of unexplored data present in raw read databases is increasingly leveraged by researchers, whether to complement their own datasets or to conduct entirely database-based studies (such as e.g. Kawasaki et al., 2023; Lagzian et al., 2024). This data source is valuable not only for pathogen discovery but also for supplementing datasets of available genomes, which are frequently biased because they were generated in response to disease outbreaks and not as part of continuous monitoring programs. For molecular pathologists, a straightforward approach consists of re-analysing all available sequencing runs corresponding to the host(s) infected by the pathogen they study (Jones et al., 2025). However, more efficient, exhaustive, and assumption-free approaches also exist. In this review, we focus on case studies of viral and bacterial pathogens that involve the screening of genomic datasets from NCBI databases (Sayers et al., 2025). We briefly review the existing methods associated with genomic data mining and show some of the benefits and limitations of using publicly available genomic resources for the study of microbial pathogens.

Methodological aspects of genomic data mining

Genomic data mining, i.e. here referring to the process of identifying and analysing sequencing projects of interest among databases, poses a series of challenges, the greatest of which is technical and lies in the amount of data to be screened to find relevant sequencing runs (in February 2024, the Short Read Archive (SRA) database contained 53 petabases, i.e. $5.3x10^{16}$ bases). Recent methodological and computational advances, as well as the demand for genomic resources to characterize the recent SARS-CoV-2 pandemic, have fuelled the development of methods aimed at indexing large-scale sequence databases (Katz et al., 2021; Edgar et al., 2022; Chikhi et al., 2024; Karasikov et al., 2024; Shiryev & Agarwala, 2024). While building the index is computationally demanding, querying it is very efficient. Depending on the indexing method and the study case, the query can be a sequence (e.g., a transcript or an antimicrobial gene), or a taxon (a pathogen, a host, or a vector). Research groups developing these indexing methods sometimes precompute database indexes and make them available online to the research community (Table 1). This pre-filtering step offers to the researcher the advantage of not having to download and process all the sequencing runs, but rather of focusing computing resources on the analysis of a relevant subset. The downside of using such online query tools is that the index does not always contain the entirety of the SRA database nor an up-to-date version of it, and that queries are sometimes limited to taxa present in GenBank RefSeq (Table 1). These limitations can sometimes be fully or partially overcome by combining the use of tools relying on precomputed databases with manual retrieval of SRA runs that are not represented in the indexes. The choice

of the data mining method must be informed and mainly depends on (i) the target database (metagenomes, human, all SRA), (ii) the query organism, (iii) the availability of a RefSeq genome (referring to a GenBank reference sequence) and (iv) the research question addressed. Regardless of the tool used, one should end up with sequencing run accessions matching the query, which can be further downloaded and processed locally. In summary, the research community can now query petabase-scale databases in search for sequencing runs useful to address its research interests.

Downstream analyses of obtained sequencing data

Processing the sequencing runs identified by the data mining approach usually relies on a series of bioinformatic steps. These include database-specific methods to download the sequences and the associated metadata, but also approaches commonly used in Next-Generation Sequencing (NGS) and metagenomics studies such as mapping-based variant calling, *de novo* assembly, or taxonomic assignation. Sequencing runs can be downloaded from NCBI SRA using the SRA Toolkit (https://github.com/ncbi/sra-tools). The associated metadata are stored in the dedicated BioSample database which can be interrogated manually (https://www.ncbi.nlm.nih.gov/biosample) or programmatically (NCBI E-utilities https://www.ncbi.nlm.nih.gov/books/NBK179288/, ffq tool (Gálvez-Merchán et al., 2023)) for larger datasets. Next, in the context of microbial pathogens, subsequent analyses can include (i) spatiotemporal and ecological distribution by metadata analysis, (ii) phylogenetics by reference-based study, (iii) structural genomics through *de novo* assembly, and/or (iv) pathobiome community characterization. Each of these requires multiple analytical steps (Figure 1). Ultimately, public sharing of data derived from data mining strategies (such as assembled genomes) is a crucial aspect of the approach. Since these datasets are often repurposed for organisms other than those originally targeted, it is important to carefully evaluate database requirements to determine the appropriate submission pathway prior to submission. In the case of NCBI, genomes must be submitted to the Third Party Annotation (TPA) database under a new BioProject and BioSample, while explicitly referencing the original SRA accession and associated metadata (GenBank Submissions Staff, personal communication, August 2025). This practice enables reproducibility, preserves a clear link to the source material and maximizes the value of existing datasets by making hidden microbial diversity accessible for future studies.

Table 1 - Overview of tools for query-based screening of sequencing runs in public genomic databases.

Tool	Publication	Nature of the query	Pre-indexed database(s)	Nature of the database index	Use-case example	Online access
STAT	Katz et al., 2021	Taxon (NCBI RefSeq taxid1)	SRA (all)	K-mer based taxonomic assignments	Searching unreported hosts for a known bacterial pathogen species	www.ncbi.nlm.nih.gov/sra/docs/sra-taxonomy-analysis-tool/
Metagraph	Karasikov et al., 2024	Sequence	SRA-Fungi, SRA-MetaGut, SRA-Metazoa, SRA-Microbe, SRA-Mouse	de Bruijn graph index of reads.	Explore the association between gut resistance genes and phage species	https://metagraph.ethz.ch/search/sra_metagut
Pebblescout	Shiryev & Agarwala, 2024	Sequence	SRA microbe, SRA metagenomic	K-mer based index of reads.	Find new species related to known pathogenic bacterial species	https://pebblescout.ncbi.nlm.nih.gov/
Serratus	Edgar et al., 2022	RNA virus family, RefSeq GenBank accession2, or SRA run identifier3	SRA (all)	Alignment of database reads to RNA viral reference sequences	Explore RNA viral diversity across hosts and ecosystems	https://serratus.io/explorer/
Logan	Chikhi et al., 2024	Sequence	SRA (all)	de novo assemblies of database reads	Explore the spatio-temporal distribution of a newly discovered microbial pathogen	https://logan-search.org/

[1] NCBI RefSeq taxid refers to the unique taxonomy identifier assigned by the NCBI Reference Sequence (RefSeq) database for standardized classification of organisms; [2] RefSeq GenBank accession refers to the unique identifier assigned to a reference genomic, transcript, or protein sequence in the NCBI GenBank database, as part of the curated RefSeq collection; [3] SRA run identifier refers to a unique ID assigned to a specific sequencing run in the NCBI Sequence Read Archive (SRA), representing raw sequencing data from a single experiment.

https://doi.org/10.24072/pcjournal.637

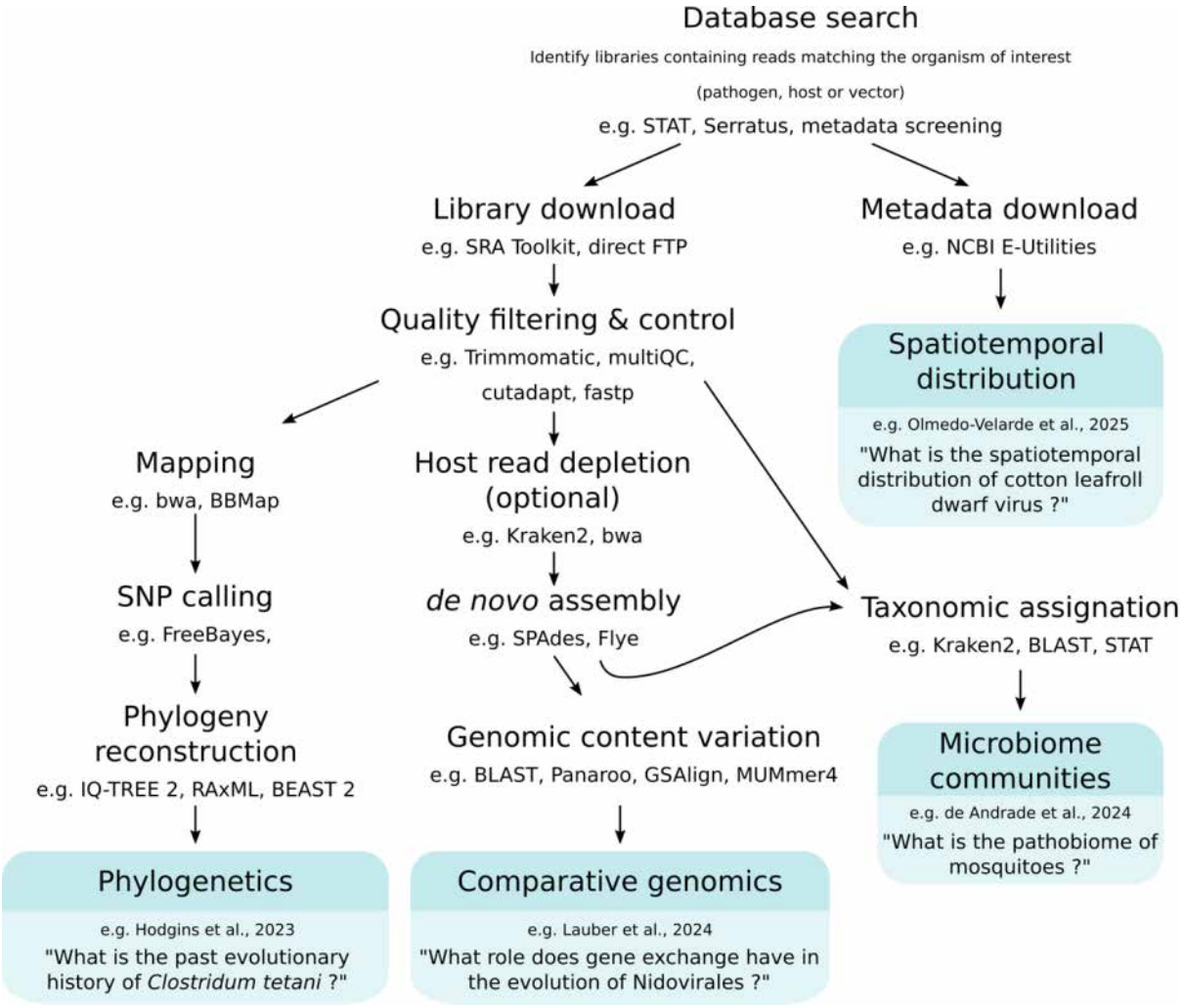

Figure 1 - Schematic of simplified possible analysis pipelines following a genomic data mining approach. Tools provided as examples include STAT (Katz et al., 2021) and Serratus (Edgar et al., 2022) for database search; SRA Toolkit (National Center for Biotechnology Information, 2024) for library download; NCBI E-Utilities (Kans, 2013) for metadata download; Trimmomatic (Bolger et al., 2014), multiQC (Ewels et al., 2016), cutadapt (Martin, 2011) and fastp (Chen et al., 2018) for quality filtering and control; Kraken2 (Wood et al., 2019) and bwa (Li & Durbin, 2009) for host read depletion; bwa (Li & Durbin, 2009), BBMap (Bushnell, 2014) for mapping; Kraken2 (Wood et al., 2019), BLAST (Altschul et al., 1990), and STAT (Katz et al., 2021) for taxonomic assignation; FreeBayes (Garrison & Marth, 2012) and GATK (McKenna et al., 2010) for SNP calling; SPAdes (Bankevich et al., 2012), Velvet (Zerbino & Birney, 2008) and Flye (Kolmogorov et al., 2019) for *de novo* assembly; IQ-TREE 2 (Minh et al., 2020), RaxML-NG (Kozlov et al., 2019) and BEAST 2 (Bouckaert et al., 2019) for Phylogenetics; BLAST (Altschul et al., 1990), Panaroo (Tonkin-Hill et al., 2020), GSAlign (Lin & Hsu, 2020) and MUMmer4 (Marçais et al., 2018) for comparative genomics.

Expanding known genetic diversity of microbial pathogens

The availability of genomic databases and the advancement of tools for their efficient analysis have highlighted that the characterized portion of the pathobiome is merely the tip of the iceberg. Indeed, new taxa, but also previously uncovered genetic diversity of known taxa are both frequently

recovered using data mining approaches. A striking example is that, while only a few dozen viral species are currently reported to infect rice, an exhaustive study of all the 17,115 rice RNA sequencing runs available on NCBI SRA uncovered hundreds of new rice viruses (Zhu et al., 2025). More generally, multiple methods targeting the hallmark RNA-dependent RNA polymerase gene have led to the discovery of thousands of novel RNA viruses from databases (Edgar et al., 2022; Hou et al., 2024). Identifying new taxa improves and expands the taxonomic classification of pathogens (Rosani et al., 2023; Reddy & Sidharthan, 2024; Briand et al., 2025; Koonin & Lee, 2025) but can also highlight pathogens presenting with a potential (Kawasaki et al., 2021) or established (Yan et al., 2025) risk of emergence. Expanding genomic resources at different taxonomical levels also benefits our understanding of the evolutionary history of microbial pathogens. For example, a data mining-based study has suggested that gene exchange between Nidovirales (the order including *Coronaviridae*, a viral family that has recently received a lot of attention) might be more frequent than previously thought and may facilitate host jumps (Lauber et al., 2024). At the species level, ancient DNA analysis from neurotoxigenic *Clostridium tetani* in archaeological human samples revealed a subgroup from South America that produces an unknown tetanus neurotoxin variant (Hodgins et al., 2023). Research questions benefiting from expanded genomic resources of microbial pathogens are numerous, and listing them goes beyond the scope of this paper. They include the characterization of the structuring of the genetic diversity, variability in host/pathogen interactions, the lineage organization, the depiction of evolutionary events, genome organization, or delineation of the pangenome (for a review, see Vello et al., 2024). Beyond fundamental studies, the characterization of the intraspecific diversity is a crucial early step in the study of a pathogen, as it conditions the establishment of disease management and control strategies. For viral and bacterial pathogens, popular detection methods include various molecular and serological methods (Rajapaksha et al., 2019). Although their susceptibility to genetic and antigen diversity varies greatly, their development relies on a good knowledge of the pathogen's diversity to ensure sensitivity across all lineages. The same goes for vaccine development, which needs to target conserved antigens. Ultimately, while expanding genomic resources of microbial pathogens enhances both our fundamental understanding of pathogen evolution and our ability to anticipate and respond to pathogen emergence, the genomic resources brought by and to the scientific community through their sharing on public databases can be even more useful when accompanying metadata are present.

In silico-based epidemiological surveillance : spatiotemporal distribution of microbial pathogens

Extensive efforts are dedicated to the characterization of the spatiotemporal distribution of pathogens. Initiatives such as EPPO (https://gd.eppo.int/) or CABI (https://www.cabidigitallibrary.org/) for plant pathogens, Atlas ECDC (http://atlas.ecdc.europa.eu/public/index.aspx), GISAID (https://gisaid.org) or Nextstrain (https://nextstrain.org/) for human pathogens aggregate occurrence data from multiple sources and aim to provide up-to-date distributions of pathogens. This knowledge is crucial to focus epidemiological surveillance efforts, but also to better understand the migration of pathogens, their emergence, and the factors driving it. Pathogen distributions are subject to ongoing changes due to new emergences –following the concept of biotic homogenization (Bebber et al., 2014)– and their detection and the reporting of newly infested territories based on direct observation can therefore experience some delays. Current distributions might be enriched by indirect occurrence data derived from genomic data mining, possibly uncovering previously undetected occurrences. This approach relies on the metadata accompanying genomic dataset. For example, this approach was used to expand the known geographic distribution of the cotton leafroll dwarf virus (Olmedo-Velarde et al., 2024) and the Solanum nigrum ilarvirus 1 (Rivarez et al., 2023). In the same manner,

the date of isolation of the samples can be very informative, both directly by providing evidence of the presence of the pathogen at a given date (Olmedo-Velarde et al., 2024), or indirectly by enriching inferences of the past evolutionary history of pathogens (Ferreira et al., 2024a). Authors have also exhaustively screened the SRA database to confirm the endemicity of a bacterial genus to New Zealand (Power et al., 2024). To finish, spatiotemporal data obtained through the genomic screening of public repositories represent a powerful complement to traditional epidemiological surveillance monitoring methods, but are also useful to enrich models of pathogen spread and evolution.

Uncover ecological interactions: microbial communities, hosts and vectors

Microbial pathogens are involved in complex relationships with their biotic environment, which data mining approaches can help shed light on. By definition, microbial pathogens interact with the organisms they infect, but they sometimes also interact with vectors (most commonly insects) and microbial communities. Host range remains one of the key epidemiological characteristics of pathogens. Indeed, more than half of human pathogens are zoonotic, and those having a broad host range are most likely to cause disease emergence (Woolhouse & Gowtage-Sequeria, 2005). As such, genomic data mining is often useful to define or expand pathogens' host range (including the identification of pathogen reservoirs), identify broad host range pathogens, uncover host jump events, or identify the genetic determinants of host specificity (Sidharthan et al., 2024; Reddy & Sidharthan, 2024; Thava Prakasa Pandian et al., 2024). Although the cases aforementioned identified the host species in host–pathogen interactions from metadata, we speculate that, in contexts such as adaptation or resistance studies, this resolution could be extended to the host genotype. Querying sequencing databases with vectors or vector-borne pathogens can also enrich known pathogen-vector associations by identifying novel pathogens of known vectors (Lin & Pascall, 2024; de Andrade et al., 2024) or by expanding the potential vector range of known pathogens (Ferreira et al., 2024b). Indeed, the vector range of vector-borne diseases is a key factor for disease management because control measures often target vectors rather than the transmitted pathogen (e.g., malaria, dengue, and most phytopathogenic viruses). Besides the pathogens they transmit, the characterization of the vectors' microbiome is also valuable to identify their own pathogens, which might potentially serve as biocontrol agents (Debat et al., 2024; de Andrade et al., 2024). For instance, Gupta et al. identified a novel virus likely infecting the neurotropic parasite *Toxoplasma gondii* through the screening of human neuronal genomic transcriptome datasets (Gupta et al., 2024). As ecological interactions involving microbial pathogens can be complex and numerous, increasing the sample size and the heterogeneity of a genomic dataset using data mining is a good strategy to deepen our understanding of pathogens' ecological interactions.

Challenges in using public genomic data

While genomic data recovered from data mining strategies are informative *per se* for genetic diversity studies for example, further studies often rely on the accompanying metadata. The most commonly provided –and used– information includes host, location, and date of isolation, but sequencing runs with all three fields filled are unfortunately scarce. When absent, information can be tediously gathered by manually going through the associated publications –a process that might soon benefit from automation, given the recent advances in natural language processing algorithms. When present, metadata must still be taken with caution, because several factors can make them misleading or even inaccurate. First, the date metadata field should represent the date of sampling rather than the date of sequencing, but special cases (e.g. "sampling" a leaf on a

herbarium plant specimen, for example) can lead to date misattribution by the submitter. Depending on the study, flagging laboratory-maintained organisms might also be appropriate because the rate and directionality of their evolution might be specific to the artificial environment of the laboratory. Moreover, for pathogens, the relevant host species to consider can derive from the "host" metadata field or the "organism" metadata field, depending on whether the sequencing run is dedicated to the study of the pathogen or its host. Furthermore, detection of pathogen-derived sequences in an organism's dataset does not necessarily indicate infection of the organism itself, but may instead reflect infection of an associated insect, fungus, or other microbiome constituent. Metadata inaccuracy can also stem from contamination, a term referring to genetic material within a sample that did not originate from that specific biological sample and to which the metadata therefore does not apply. External contamination comes from the sample's surrounding environment, including laboratory equipment, library preparation kits, sequencing platforms, and researchers themselves (Eisenhofer et al., 2019; Mahillon et al., 2024). Cross-sample contamination can originate from well-to-well contamination, index switching, or sample bleeding (Ballenghien et al., 2017; Vigne et al., 2018; Lou et al., 2023). Strategies employed to tackle this problem include specific laboratory practices, the addition of negative controls, and the *in silico* removal of contaminants (Eisenhofer et al., 2019). These approaches are unfortunately poorly adapted to the reuse of publicly available data (*a posteriori* implementation is mostly impossible), but also to the study of pathogens, which are often only represented at low frequencies among host or vector reads (complicating the *in silico* removal of contaminants). In some specific cases, it is possible to suspect cross-sample contamination and hence exclude the corresponding sequencing runs from a study. Indeed, as samples sequenced in the same batch often share the same BioProject accession on NCBI, the detection of the same genotype of a pathogen in two runs within a BioProject calls for a careful check. Given the difficulties in distinguishing contaminants from genuine presence in publicly available sequencing runs, one should remain cautious when interpreting results obtained from data mining studies. A good strategy is to consider these results preliminary and to confirm them with other data sources. This can be done by (i) checking that the results obtained rely on multiple independently generated datasets, (ii) making sure that the results obtained from data mining are compatible with their data mining-free corollary (for example, relying only on curated published genomes), (iii) collaborating with the authors of the relevant sequencing runs to resequence the sample if it still exists, or (iv) re-collecting similar samples with a dedicated study.

Conclusion and perspectives

Genomic data mining has proven to be a powerful approach to uncover previously unrecognized aspects of microbial pathogen biology, ecology, and evolution. By repurposing publicly available sequencing data, researchers can deepen our understanding of the epidemiology and the evolutionary history of microbial pathogens without generating new sequencing data. Moreover, reanalyzing publicly available data helps prevent redundant efforts and maximizes the efficient use of limited research resources. While the variability in quality, completeness, and accuracy of the associated metadata often prevents any conclusive study based solely on data mining, the approach represents a powerful springboard for hypothesis generation and preliminary analyses. Notably, databases remain heavily biased towards few taxa or locations (68% and 70% of all SRA sequences come from 5 species and ten countries, respectively (Vince et al., 2025)). Specifically, the scarcity of systematic and metadata-rich sequencing efforts at ecological scales continues to constrain the full potential of the approach. Filling this gap could open the door to addressing new challenges, such as going from the sole detection of the presence of a taxon to the understanding of its ecological role by uncovering the interactions it has with other organisms. Furthermore, genomic data can be combined with data

from other fields of study. For example, geospatial data (e.g. altitude, land use, soil, atmospheric properties, degree of anthropization, etc.) can enhance our understanding of the effect of abiotic factors on pathogens; text mining data (screening of scientific articles, search engine queries, social network posts, wikipedia traffic, etc.) can provide spatiotemporal distributions of pathogens for epidemiosurveillance programs; physiological, phenotypic, or yield data can be useful to characterize the phenotype of a disease (incidence, severity, ...). While this review has focused on viral and bacterial pathogens, future work could explore the applicability of the method to other pathogen types, such as fungi, protists, or archaea. To finish, the willingness of the researchers to make their data and associated metadata publicly available is a crucial enabler for data mining approaches and deserves commendation. Equally important is the collaborative work dedicated to database interoperability –as exemplified by initiatives like the International Nucleotide Sequence Database Collaboration– which has significantly amplified the usability of genomic databases. Looking ahead, broader adoption of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles would enhance even further the utility of genomic data mining.

Acknowledgments

The authors thank Eugénie Hébrard (PHIM) for her thoughtful comments on the draft of the paper. Preprint version 3 of this article has been peer-reviewed and recommended by PCI Infections (https://doi.org/10.24072/pci.infections.100248; Merino, 2025).

Funding

This work has been publicly funded by the ANR (the French National Research Agency) under the project SPADYVA (ANR-20-CE35-0008-01).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Author contributions

Writing – original draft: Damien Richard & Nils Poulicard; Writing – review and editing: Damien Richard & Nils Poulicard.

Data Availability

No dataset was generated nor used for this review study.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ballenghien M, Faivre N, Galtier N (2017) Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology*, **15**, 25. https://doi.org/10.1186/s12915-017-0366-6

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD (2012) SPAdes: a new genome assembly algorithm and its applications

to single-cell sequencing. *Journal of computational biology*, **19**, 455–477. https://doi.org/10.1089/cmb.2012.0021

Bebber DP, Holmes T, Gurr SJ (2014) The global spread of crop pests and pathogens. *Global Ecology and Biogeography*, **23**, 1398–1407. https://doi.org/10.1111/geb.12214

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PloS computational biology*, **15**, e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Briand M, Jacques M-A, Dittmer J (2025) The hidden life of Xylella: Mining the NCBI Sequence Read Archive reveals potential new species, host plants and infected areas for this elusive bacterial plant pathogen. *BioRxiv*, 2025.04.18.649540. https://doi.org/10.1101/2025.04.18.649540

Bushnell B (2014) *BBMap: A Fast, Accurate, Splice-Aware Aligner*. Lawrence Berkeley National Laboratory.

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chikhi R, Raffestin B, Korobeynikov A, Edgar R, Babaian A (2024) Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. *BioRxiv*, 2024.07.30.605881. https://doi.org/10.1101/2024.07.30.605881

de Andrade AAS, Brustolini O, Grivet M, Schrago CG, Vasconcelos ATR (2024) Predicting novel mosquito-associated viruses from metatranscriptomic dark matter. *NAR Genomics and Bioinformatics*, **6**, lqae077. https://doi.org/10.1093/nargab/lqae077

Debat H, Farrher ES, Bejerman N (2024) Insights into the RNA Virome of the Corn Leafhopper Dalbulus maidis, a Major Emergent Threat of Maize in Latin America. *Viruses*, **16**. https://doi.org/10.3390/v16101583

Edgar RC, Taylor B, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, Banfield JF, de la Peña M, Korobeynikov A, Chikhi R, Babaian A (2022) Petabase-scale sequence alignment catalyses viral discovery. *Nature*, **602**, 142–147. https://doi.org/10.1038/s41586-021-04332-2

Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, **27**, 105–117. https://doi.org/10.1016/j.tim.2018.11.003

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Ferreira RC, Alves GV, Ramon M, Antoneli F, Briones MRS (2024) Reconstructing Prehistoric Viral Genomes from Neanderthal Sequencing Data. *Viruses*, **16**. https://doi.org/10.3390/v16060856

Ferreira LYM, de Sousa AG, Silva JL, Santos JPN, Souza DG do N, Orellana LCB, de Santana SF, de Vasconcelos LBCM, Oliveira AR, Aguiar ERGR (2024) Characterization of the Virome Associated with the Ubiquitous Two-Spotted Spider Mite, Tetranychus urticae. *Viruses*, **16**. https://doi.org/10.3390/v16101532

Gálvez-Merchán Á, Min KH (Joseph), Pachter L, Booeshaghi AS (2023) Metadata retrieval from sequence databases with ffq. *Bioinformatics*, **39**, btac667. https://doi.org/10.1093/bioinformatics/btac667

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. https://doi.org/10.48550/arXiv.1207.3907

Gupta P, Hiller A, Chowdhury J, Lim D, Lim DY, Saeij JPJ, Babaian A, Rodriguez F, Pereira L, Morales-Tapia A (2024) A parasite odyssey: An RNA virus concealed in Toxoplasma gondii. *Virus Evolution*, **10**, veae040. https://doi.org/10.1093/ve/veae040

Hodgins HP, Chen P, Lobb B, Wei X, Tremblay BJM, Mansfield MJ, Lee VCY, Lee P-G, Coffin J, Duggan AT, Dolphin AE, Renaud G, Dong M, Doxey AC (2023) Ancient Clostridium DNA and variants of tetanus neurotoxins associated with human archaeological remains. *Nature Communications*, **14**, 5475. https://doi.org/10.1038/s41467-023-41174-0

Hou X, He Y, Fang P, Mei S-Q, Xu Z, Wu W-C, Tian J-H, Zhang S, Zeng Z-Y, Gou Q-Y, Xin G-Y, Le S-J, Xia Y-Y, Zhou Y-L, Hui F-M, Pan Y-F, Eden J-S, Yang Z-H, Han C, Shu Y-L, Guo D, Li J, Holmes EC, Li Z-R, Shi M (2024) Using artificial intelligence to document the hidden RNA virosphere. *Cell*, **187**, 6929-6942.e16. https://doi.org/10.1016/j.cell.2024.09.027

Jones M, Rastas P, Chacón-Duque JC, DiLeo MF, Nair A, Oostra V, Saastamoinen M, Duplouy A (2025) Recovering ecological interactions by mining non-target data from whole genome re-sequencing projects. *BioRxiv*, 2025.01.17.633498. https://doi.org/10.1101/2025.01.17.633498

Kans J (2013) Entrez Direct: E-utilities on the Unix Command Line. In: *Entrez Programming Utilities Help [Internet]* , p. . National Center for Biotechnology Information (US), Bethesda (MD).

Karasikov M, Mustafa H, Danciu D, Zimmermann M, Barber C, Rätsch G, Kahles A (2024) Indexing All Life's Known Biological Sequences. *BioRxiv*, 2020.10.01.322164. https://doi.org/10.1101/2020.10.01.322164

Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C (2021) STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biology*, **22**, 270. https://doi.org/10.1186/s13059-021-02490-0

Kawasaki J, Kojima S, Tomonaga K, Horie M (2021) Hidden Viral Sequences in Public Sequencing Data and Warning for Future Emerging Diseases. *MBio*, **12**, 10.1128/mbio.01638-21. https://doi.org/10.1128/mbio.01638-21

Kawasaki J, Tomonaga K, Horie M (2023) Large-scale investigation of zoonotic viruses in the era of high-throughput sequencing. *Microbiology and Immunology*, **67**, 1–13. https://doi.org/10.1111/1348-0421.13033

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, **37**, 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koonin EV, Lee BD (2025) Diversity and evolution of viroids and viroid-like agents with circular RNA genomes revealed by metatranscriptome mining. *Nucleic Acids Research*, **53**, gkae1278. https://doi.org/10.1093/nar/gkae1278

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A (2019) RaxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**, 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Lagzian A, Ghorbani A, Tabein S, Riseh RS (2024) Genetic variations and gene expression profiles of Rice Black-streaked dwarf virus (RBSDV) in different host plants and insect vectors: insights from RNA-Seq analysis. *BMC Genomics*, **25**, 736. https://doi.org/10.1186/s12864-024-10649-9

Lauber C, Zhang X, Vaas J, Klingler F, Mutz P, Dubin A, Pietschmann T, Roth O, Neuman BW, Gorbalenya AE, Bartenschlager R, Seitz S (2024) Deep mining of the Sequence Read Archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates. *PloS pathogens*, **20**, e1012163. https://doi.org/10.1371/journal.ppat.1012163

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lin H-N, Hsu W-L (2020) GSAlign: an efficient sequence alignment tool for intra-species genomes. *BMC genomics*, **21**, 1–10. https://doi.org/10.1186/s12864-020-6569-1

Lin Y, Pascall DJ (2024) Characterisation of putative novel tick viruses and zoonotic risk prediction. *Ecology and Evolution*, **14**, e10814. https://doi.org/10.1002/ece3.10814

Lou YC, Hoff J, Olm MR, West-Roberts J, Diamond S, Firek BA, Morowitz MJ, Banfield JF (2023) Using strain-resolved analysis to identify contamination in metagenomics data. *Microbiome*, **11**, 36. https://doi.org/10.1186/s40168-023-01477-2

Mahillon M, Brodard J, Schoen R, Botermans M, Dubuis N, Groux R, Pannell JR, Blouin AG, Schumpp O (2024) Revisiting a pollen-transmitted ilarvirus previously associated with angular mosaic of grapevine. *Virus research*, **344**, 199362. https://doi.org/10.1016/j.virusres.2024.199362

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: A fast and versatile genome alignment system. *PloS computational biology*, **14**, e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, **17**, 10–12. https://doi.org/10.14806/ej.17.1.200

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303. https://doi.org/10.1101/gr.107524.110

Merino, S. (2025) Review on the methodological aspects and potential uses of genomic data mining. *Peer Community in Infections, 100248.* https://doi.org/10.24072/pci.infections.100248

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, **37**, 1530–1534. https://doi.org/10.1093/molbev/msaa015

National Center for Biotechnology Information (2024) NCBI SRA Toolkit.

Olmedo-Velarde A, Shakhzadyan H, Rethwisch M, West-Ortiz MJ, Waisen P, Heck ML (2024) Data mining redefines the timeline and geographic spread of cotton leafroll dwarf virus. *Plant Disease*. https://doi.org/10.1094/PDIS-06-24-1265-SC

Power JF, Carere CR, Welford HE, Hudson DT, Lee KC, Moreau JW, Ettema TJG, Reysenbach A-L, Lee CK, Colman DR, Boyd ES, Morgan XC, McDonald IR, Craig Cary S, Stott MB (2024) A genus in the bacterial phylum Aquificota appears to be endemic to Aotearoa-New Zealand. *Nature Communications*, **15**, 179. https://doi.org/10.1038/s41467-023-43960-2

Rajapaksha P, Elbourne A, Gangadoo S, Brown R, Cozzolino D, Chapman J (2019) A review of methods for the detection of pathogenic microorganisms. *Analyst*, **144**, 396–411. https://doi.org/10.1039/C8AN01488D

Reddy TS, Sidharthan VK (2024) Three-fold expansion of the genetic diversity of blunerviruses through plant (meta)transcriptome data-mining. *Virology*, **599**, 110210. https://doi.org/10.1016/j.virol.2024.110210

Rivarez MPS, Faure C, Svanella-Dumas L, Pecman A, Tušek-Žnidaric M, Schönegger D, De Jonghe K, Blouin A, Rasmussen DA, Massart S, Ravnikar M, Kutnjak D, Marais A, Candresse T (2023) Diversity and Pathobiology of an Ilarvirus Unexpectedly Detected in Diverse Plants and Global Sequencing Data. *Phytopathology®*, **113**, 1729–1744. https://doi.org/10.1094/PHYTO-12-22-0465-V

Rosani U, Gaia M, Delmont TO, Krupovic M (2023) Tracing the invertebrate herpesviruses in the global sequence datasets. *Frontiers in Marine Science*, **10**. https://doi.org/10.3389/fmars.2023.1159754

Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, Feldgarden M, Fine AM, Funk K, Hoffman J, Kannan S, Kelly C, Klimke W, Kim S, Lathrop S, Marchler-Bauer A, Murphy TD, O'Sullivan C, Schmieder E, Skripchenko Y, Stine A, Thibaud-Nissen F, Wang J, Ye J, Zellers E, Schneider VA, Pruitt KD (2025) Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research*, **53**, D20–D29. https://doi.org/10.1093/nar/gkae979

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, Karsch-Mizrachi I (2023) GenBank 2023 update. *Nucleic Acids Research*, **51**, D141–D144. https://doi.org/10.1093/nar/gkac1012

Shiryev SA, Agarwala R (2024) Indexing and searching petabase-scale nucleotide resources. *Nature Methods*, **21**, 994–1002. https://doi.org/10.1038/s41592-024-02280-z

Sidharthan VK, Reddy V, Kiran G, Rajeswari V, Baranwal VK, Kumar MK, Kumar KS (2024) Probing of plant transcriptomes reveals the hidden genetic diversity of the family Secoviridae. *Archives of Virology*, **169**, 150. https://doi.org/10.1007/s00705-024-06076-6

Thava Prakasa Pandian R, Bhavishya, Kavi Sidharthan V, Rajesh MK, Babu M, Sharma SK, Nirmal Kumar BJ, Chaithra M, Hegde V (2024) From the discovery of a novel arepavirus in diseased arecanut palms (Areca catechu L.) in India to the identification of known and novel arepaviruses in bee and plant transcriptomes through data-mining. *Virology*, **600**, 110256. https://doi.org/10.1016/j.virol.2024.110256

Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, Frost SDW, Corander J, Bentley SD, Parkhill J (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, **21**, 180. https://doi.org/10.1186/s13059-020-02090-4

Vello F, Filippini F, Righetto I (2024) Bioinformatics Goes Viral: I. Databases, Phylogenetics and Phylodynamics Tools for Boosting Virus Research. *Viruses*, **16**. https://doi.org/10.3390/v16091425

Vigne E, Garcia S, Komar V, Lemaire O, Hily J-M (2018) Comparison of Serological and Molecular Methods With High-Throughput Sequencing for the Detection and Quantification of Grapevine Fanleaf Virus in Vineyard Samples. *Frontiers in Microbiology*, **9**. https://doi.org/10.3389/fmicb.2018.02726

Vince O, Oldach P, Pereno V, Leung MHY, Greco C, Minto-Cowcher G, Ur-Rehman S, Kam KYK, Chow W, Bolton E, Mwambingu BR, Greenhalgh N, Knot I, Christoffersen L, Clark M, Pecoraro R, Kollasch AW, Bohnuud T, Bakalar M, Lorenz P, Gowers G (2025) Breaking Through Biology's Data Wall: Expanding the Known Tree of Life by Over 10x using a Global Biodiscovery Pipeline. *BioRxiv*, 2025.06.11.658620. https://doi.org/10.1101/2025.06.11.658620

Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. *Genome biology*, **20**, 1–13. https://doi.org/10.1186/s13059-019-1891-0

Woolhouse MEJ, Gowtage-Sequeria S (2005) Host range and emerging and reemerging pathogens. *Emerging infectious diseases*, **11**, 1842–1847. https://doi.org/10.3201/eid1112.050997

Yan W, Zhu Y, Zou C, Liu W, Jia B, Niu J, Zhou Y, Chen B, Li R, Ding S-W, Wu Q, Guo Z (2025) Virome Characterization of Native Wild-Rice Plants Discovers a Novel Pathogenic Rice Polerovirus With World-Wide Circulation. *Plant, cell & environment*, **48**, 1005–1020. https://doi.org/10.1111/pce.15204

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, **18**, 821–829. https://doi.org/10.1101/gr.074492.107

Zhu Y, Raza A, Bai Q, Zou C, Niu J, Guo Z, Wu Q (2025) In-depth analysis of 17,115 rice transcriptomes reveals extensive viral diversity in rice plants. *Nature Communications*, **16**, 1559. https://doi.org/10.1038/s41467-025-56769-y