Research on the genome of microorganisms: ethical considerations and recommendations regarding the incidental bystander sequencing of human genetic material

Koen Bartholomeeusen, Pieter Monsieurs, Jan Van Den Abbeele, Pieter Moris, Odin Goovaerts, Wim Pinxten, Jef Verellen, Oren Tzfadia, Kristien Verdonck, Raffaella Ravinetto

In genomic research primarily targeting microorganisms (or pathogens), a substantial risk exists that the presence of human genetic bycatch is not sufficiently recognised, and that the potential harm of unwarranted analysis, access, or sharing of human genetic bystander data is also insufficiently acknowledged or mitigated. In this Personal View, we contend that mandatory risk mitigation measures are necessary, more so in view of the likely increase of sharing of materials and pathogen sequence information under the WHO Pandemic Agreement and the related Pathogen Access and Benefit Sharing framework. Based on a joint reflection of the Institutional Review Board and individual researchers at the Institute of Tropical Medicine in Antwerp, Belgium, we propose a four-step approach to mitigate such risks: prevention or early removal of human genetic sequences, secure storage of samples and data, adaptation of informed consent, and targeted ethics review. This approach should contribute to maintaining ethical integrity, protect the rights of individuals and communities, and bolster public trust in the expanding use of untargeted sequencing in global health research.

Lancet Microbe 2026; 7: 101341

https://doi.org/10.1016/j.lanmic.2025.101341

Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium (K Bartholomeeusen PhD, P Monsieurs PhD, Prof J Van Den Abbeele PhD, P Moris PhD, O Tzfadia PhD); **Department of Clinical Sciences, Institute of Tropical Medicine, Antwerp, Belgium** (O Goovaerts PhD), **Quality Assurance Unit, Institute of Tropical Medicine, Antwerp, Belgium** (J Verellen MSc); **Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium** (K Verdonck PhD, R Ravinetto PhD); **Institutional Review Board, Institute of Tropical Medicine, Antwerp, Belgium** (K Bartholomeeusen PhD, Prof J Van Den Abbeele PhD, O Goovaerts PhD, Prof W Pinxten PhD, O Tzfadia PhD, K Verdonck PhD, R Ravinetto PhD); **Faculty of Medicine and Life Sciences, University of Hasselt, Hasselt, Belgium** (Prof W Pinxten)

Correspondence to:
Dr Raffaella Ravinetto,
Department of Public Health,
Institute of Tropical Medicine,
Antwerp 2000, Belgium
rravinetto@itg.be

Introduction

Genomic research offers major opportunities for scientific advancement. Ethical issues related to research that directly targets the human genome include the benefit–harm ratio of biobanks and research programmes, informed consent, return-of-results policies, tensions between data privacy and sharing, and inequities stemming from genomic health care and research. According to the report of the Nuffield Council titled *Ethics in Genomics Healthcare & Research*,[1] those working in genomic research end up negotiating ethics issues on a continuous basis.

Genomic research also targets microorganisms, such as bacteria, single-cell parasites, and viruses. Genetic information from microorganisms is essential for public health surveillance and biomedical research, particularly in infectious diseases[2] and outbreak response. The importance of microbial genetic information is recognised in WHO's Pandemic Agreement, approved in April, 2025, and adopted by the Seventy-eighth World Health Assembly in May, 2025.[3] The Agreement underscores the importance of sharing materials and sequence information on pathogens with pandemic potential, in addition to the relevant benefits. The Agreement establishes a multilateral system for access and benefit sharing under the Pathogen Access and Benefit Sharing framework, which is developed as an annex.[3]

Untargeted sequencing of microorganism genomes and human genetic bycatch

Untargeted sequencing approaches to characterise whole microbial genomes or transcriptomes can yield detailed insights into the biology of microorganisms and into their associated virulence, disease transmission, and resistance mechanisms. Highly sensitive and informative in-depth sequencing technologies are becoming a mainstay in tropical medicine, global health research, and outbreak investigation.[4] Ethical reflection should keep pace with these developments, which are increasingly accessible and attainable. Johnson and Parker[5] identified some key ethical issues related to pathogen sequencing. These authors called for innovative and critical thinking on the use of pathogen sequence data, implying the adaptation of standards for informed consent, data collection, use, and sharing—underpinned by a strong commitment to justice and global health equity. The importance of their recommendations cannot be downplayed, in view of the potential harm caused by the identification of origin or transmission routes of outbreaks, at the level of both individuals and communities.

We believe that, given the often untargeted nature and high sensitivity of sequencing approaches, additional considerations are needed. The use of human-derived or environment-derived biosamples for extraction and analysis of microbial genetic material carries a potential for incidental bystander sequencing of human genetic material (ie, human genetic bycatch). Although some genomic sequences (eg, highly conserved housekeeping genes) might not permit individual identification, others, even in small fragments, can allow for re-identification of individuals or relatives, especially when cross-referenced with expanding human genomic repositories. Even comparably scarce human genomic sequence information can allow for unique identification, reconstruction of genetic ancestry and disease susceptibility, or re-identification of individuals or family members.[6–18] Risks of re-identification increase as human genomic repository sizes grow, organisational safeguards are absent or insufficient,[19–21] and the information, even when not publicly accessible, can be accessed or seized by governmental or law-enforcement actors.[22] Rapid progress in artificial intelligence and bioinformatics can even lower the technical threshold for re-identification, intensifying the associated risks.

Therefore, a careful reflection is needed, not only on microorganism (pathogen) sequence data, but also on the concurrent presence of bystander human sequences. Some researchers have looked at this topic from a technical perspective,[12,13,16,17] but this issue has rarely been explicitly addressed from an ethics standpoint. Even when addressed, the focus is on environmental human DNA. A few researchers highlighted the ethical, legal, and privacy concerns surrounding the inadvertent capture and analysis of human genetic material from environmental samples (ie, issues of consent, surveillance, and identifiability).[23–26] But the risks of unwarranted sharing or analysis, or both, of human genetic bycatch data from microorganism (pathogen) research, are, to the best of our knowledge, still insufficiently addressed, even though the WHO guiding principles for pathogen genome data sharing of 2022 briefly indicate that "human genomic data that may inadvertently be included in raw reads should be removed before submission to or release of the sequence on public databases".[27]

Data can be qualified as anonymous only when full certainty exists that any remaining human reads will be indistinguishable from other sequences or pathogen sequences. The inherent impossibility of rendering human DNA sequence information anonymous can cause privacy breaches, both for the individuals who provide the biosamples and for their relatives or the broader community whose information is partly contained in the same biosamples. Risks are magnified when sequences are shared in public repositories that provide free and unrestricted access to unfiltered bulk DNA and RNA sequences and related information and metadata. At a massive level of sequencing, with transmission chains recreated by analysing pathogen sequences, caution should be exercised to prevent stigmatisation or blaming of people or communities. To achieve this aim, only pathogen sequence data void of any human DNA sequences should be shared, unless explicit informed consent and ethics approval have been obtained for sharing human genomic information too. Regarding identifiability, the European General Data Protection Regulation (GDPR) states that "account should be taken of all means reasonably likely to be used, such as singling out to identify a natural person directly or indirectly". In practice, singling out boils down to determining the minimum amount of genetic data that is still, alone or in combination with other data sources, reasonably likely to identify an individual. This risk-based assessment is context specific and can depend on a combination of various elements, such as the amount and uniqueness of the DNA sequences, availability of resources (eg, cross-referenceable datasets), incentives to re-identify for a given purpose, and so forth.

Given the non-revocable nature of publication of an individual's genetic information, and its familial and transgenerational consequences, we believe that detailed consideration and vigilant caution are warranted to avoid inadvertent sharing. A precautionary approach is dictated by the risks at stake, such as the rapidly advancing and more powerful technologies, the exponential proliferation of available DNA data in gene banks, and the notion that human DNA does not change throughout the donor's life span. In practice, all human-derived genetic data should be dealt with—and protected—as personal data altogether. Unfortunately, if research focuses on the genome of microorganisms, then the risk of human genetic bycatch can go unrecognised in the donor's consent documents or in the protocol submitted to the ethics committee(s), or both. Mitigation measures are feasible and necessary, but inadequate awareness of the risks associated with human genetic bycatch will result in insufficient safeguards.

A four-step approach to mitigate the risks of human genetic bycatch

The WHO *Guidance for Human Genome Data Collection, Access, Use, and Sharing*[28] recognises the right of individuals to make informed decisions on their genomic data: misuse of human genetic data should be prevented, with mechanisms in place to hold individuals, institutions, and stakeholders accountable for failure to adhere to scientifically and ethically sound processes. The application of these principles and processes to human genetic bycatch is crucial, including in outbreak response, and in view of implementation of the WHO Pandemic Agreement. Compliance is a shared responsibility of research institutions, researchers, public health actors (ie, Ministries of Health, disease control programmes, surveillance programmes), regulators, and ethics committees. Other stakeholders include funders, sequencing providers, repositories, administrators, and editors of medical journals. Collectively, all of them together should ensure protection of human genetic data and transparency and accountability to the public.

We propose that the relevant stakeholders act at four interconnected levels: prevention or early removal of human genetic sequences, secure storage of samples and data, adaptation of informed consent, and targeted ethics review (figure).

Prevention or early removal of human genetic sequences

To prevent unintended human genomic data analysis and sharing, researchers should assess the risk for gathering human genetic data inherent to the methods used (ie, untargeted *vs* targeted sequencing; human-derived or environment-derived biosamples *vs* culture-grown microorganisms). Enrichment methods for genetic material of microorganisms (eg, selective whole-genome amplification or hybridisation capture such as SureSelect, SeqCap EZ, or Twist) are often used for technical reasons, to improve efficiency and specificity in whole-genome sequencing of the organism of interest. By contrast, in some cases (partial) in-vitro depletion of host DNA is feasible (eg, the depletion of white blood cells in fresh whole-blood samples using filtration methods). However, aside from incomplete

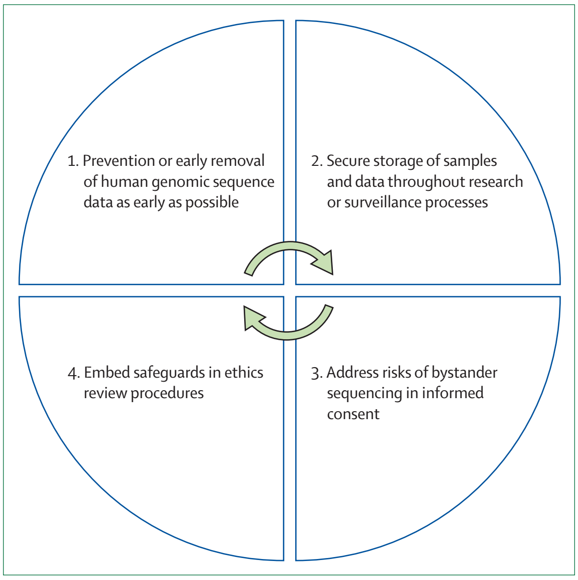

***Figure:* A four-step approach to protection of human genetic data from incidental bystander sequencing**

removal of host DNA, these methods can sometimes lead to a loss of microorganism DNA, cause a bias in the collected target DNA, or increase the time and cost of sample processing. Additionally, culturing can introduce some biases, and some microorganisms (eg, *Plasmodium vivax* or *Mycobacterium leprae*) cannot be cultured under laboratory conditions, making direct sequencing the only feasible option to obtain some pathogen genomes. But these methods do not necessarily remove the non-targeted genetic material completely, and the chosen sequencing approach might be sensitive enough to detect bystander genetic material.

Thus, an upfront risk assessment is needed, and the bioinformatic strategy for human sequence removal should be documented in the research or surveillance protocol. The two such main methods are alignment-based filtering and classification-based filtering.[16,17] Alignment-based filtering strategies include subtractive filtering (align all reads to the human genome and retain unmapped reads),[29] exclusive retention (align reads to the target genome and discard unmapped reads),[30] and comparative alignment (align reads to both genomes and bin by best match).[31–33] Classification-based filtering methods rely on k-mers to predict the taxonomic origin of each read using a database of reference genomes, allowing human-derived reads to be removed;[34–37] this method is a common strategy for metagenomic samples, wherein different organisms or groups need to be distinguished, on top of removal of human contamination. Of note, exclusive retention and classification-based filtering strategies are typically preferred in studies targeting a known microorganism, whereas the subtractive filtering approach is more suitable for metagenomic analyses, particularly when the identity of the microorganism is not known upfront. These strategies can be combined[38–42] and further adapted to achieve optimal balance between sensitivity (stringent removal of all suspected human reads) and specificity (retaining sufficient microorganism-derived reads for downstream analyses), while simultaneously controlling the computational costs of the selected method.

Of note, bioinformatic strategies do not add much complexity overall, since they either rely on the same steps that are required for processing sequencing data (alignment-based filtering) or they are implemented by several commonly used software packages (classification-based filtering). However, the burden of human DNA removal can be higher in low-income settings with scarce infrastructure for library preparation and bioinformatic analysis. Thus, the upgrade of these capacities should be incorporated in capacity-building programmes, in addition to outbreak-preparedness programmes, to help to overcome gaps in the protection of individuals and communities between high-income and low-income countries.

Secure storage of samples and data

Samples or leftovers containing human materials, genomic sequences, metadata, and any accompanying personal or medical data (as per ethics approval) need to be stored and shared on secured platforms, in compliance with data protection laws (eg, the European GDPR and the South African Protection of Personal Information Act) and with ethics guidelines (eg, the Declaration of Helsinki,[43] Declaration of Taipei,[44] and Nagoya Protocol[45]). All protection measures need to be described in the protocol and data-management plan submitted to the ethics committee(s) and other bodies, in all countries where samples are collected, used, and stored in the long term.

For more on the **Protection of Personal Information Act** see https://popia.co.za/

The protocol needs to detail the measures taken to avoid inadvertent sharing of human genetic bycatch with others, such as partners in microorganism research, service providers, sequences repositories, and medical journals. Unauthorised usage of any bystander human sequences should be explicitly forbidden in contractual agreements with sequencing providers. The sharing of sequencing data for future research, pooled analyses, or other uses, should only happen after the elimination of human reads. The same applies for submission to public repositories, such as the European Nucleotide Archive[46] or the National Center for Biotechnology Information Sequencing Read Archive (SRA),[47] which indeed encourage the removal of human reads. SRA even provides basic tools to assist in the removal of human sequences.[48]

Adequate governance measures are also needed for internal use of samples and data (ie, at the same institution). For instance, institutions' governance needs to include the principle of finality (ie, ensure that use of samples stays close to the goal for which they were collected) and strategic policies and procedures (including scientific and ethics review) to ensure adequate data protection, including for human genetic bycatch.

Adaptation of informed consent

Prospective participants or donors (or their legal representatives) in research or surveillance activities involving analyses of microorganism genomes should be informed, in lay language, of the risk of bystander sequencing and how the risk is being mitigated. The information should be included in the participant information sheet. Given the technical complexity of the subject, the information sheet should be discussed upfront with representatives of the communities concerned, and pretested in the target group, so as to address the local circumstances (eg, in terms of literacy, research literacy, and foreseeable concerns).

The subject complexity, and the fact that a research project primarily targets genomic information of microorganisms, cannot justify lack or paucity of information: everybody has the right to be transparently informed about what happens with their biological samples and genetic information, and the associated risks.

Targeted ethics review

Ethics committee(s) or institutional review board(s) (IRBs) should include at least one member with expertise relevant to the topic, or seek ad-hoc expertise as needed. The members should be, individually and collectively, aware of the risks related to genomic studies on microorganisms. The members should also require that protocols include adequate procedures for early removal of human genetic data; that protocols describe how samples and data will be securely stored; that the consent documents include clear and comprehensive information, in lay language, on the risks of bystander sequencing processes and how they are mitigated; that contractual agreements with sequencing providers prohibit unauthorised usage of bystander human sequences; and that reads from human material are excluded before submitting sequencing data to public repositories or sharing them for pooled analyses or future research.

Such aspects should be included in the standard operating procedures of ethics committee(s) or IRBs, to ensure a consistent approach across protocols.

Conclusions

Our proposal results from a joint reflection of the IRB, Quality Assurance Unit, and researchers at the Institute of Tropical Medicine (Antwerp, Belgium), in response to emerging challenges encountered in our activities. The inclusion of biologists and laboratory experts as IRB members was key to broaden the expertise of the IRB, and to allow identification of the challenges related to human genetic bycatch and define an approach to face them.

Human genetic data cannot be processed, used, or stored in the absence of robust justification, explicit and meaningful informed consent, ethics approval from the countries concerned, and transparent governance mechanisms. When the primary focus of genomic research is a microorganism, a risk remains that the potential presence of human genetic bycatch and its consequences is not (sufficiently) recognised. Mitigation measures are necessary, even more so in view of increased sharing of materials and sequence information on pathogens with pandemic potential. We hope that the proposed four-step approach—ie, prevention or removal of human sequences, secured data-storage solutions, adaptation of informed consent, and targeted ethics review—will contribute to maintaining ethical integrity, protect the rights of individuals and communities, and bolster public trust in the expanding use of untargeted genetic sequencing in global health research.

Contributors

KB, PMon, and RR conceptualised the Personal View. KB and RR were in charge of project administration and wrote the original draft. RR and WP supervised the writing. KV was in charge of visualisation. PMon, RR, JVDA, PMor, OG, WP, JV, OT, and KV reviewed and edited the manuscript. All the authors approved the final manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgments

This proposal results from the joint reflection of the Institutional Review Board (IRB), the Quality Assurance Unit, and other individual researchers at the Institute of Tropical Medicine in Antwerp, Belgium, in response to emerging ethics challenges encountered in our routine activities. We acknowledge the other members of the IRB who contributed to the reflection that led to the writing of this Personal View: Joannes Clerinx, Tanyth De Goyer, Nicole Berens-Riha, Grace Marie Ku, and Arsenii Alenichev. We also acknowledge the IRB secretary, Ann Caron, for her continuous support.

References

1 Nuffield Council on Bioethics. Ethics in genomics healthcare & research: building connections & sharing best practice. July, 2023. https://cdn.nuffieldbioethics.org/wp-content/uploads/FINAL-version-Genomics-workshop-summary-and-analysis.pdf (accessed April 23, 2025).

2 Ribeiro CDS, van Roode MY, Haringhuizen GB, Koopmans MP, Claassen E, van de Burgwal LHM. How ownership rights over microorganisms affect infectious disease control and innovation: a root-cause analysis of barriers to data sharing as experienced by key stakeholders. *PLoS One* 2018; **13:** e0195885.

3 WHO. Seventy-eighth World Health Assembly concludes: historic outcomes, consequential highlights. May 28, 2025. https://www.who.int/news/item/28-05-2025-seventy-eighth-world-health-assembly-concludes–historic-outcomes–consequential-highlights (accessed Jan 22, 2026).

4 Geller G, Dvoskin R, Thio CL, et al. Genomics and infectious disease: a call to identify the ethical, legal and social implications for public health and clinical practice. *Genome Med* 2014; **6:** 106.

5 Johnson S, Parker M. Ethical challenges in pathogen sequencing: a systematic scoping review. *Wellcome Open Res* 2020; **5:** 119.

6 Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science* 2004; **305:** 183.

7 Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008; **4:** e1000167.

8 Cassa CA, Schmidt B, Kohane IS, Mandl KD. My sister's keeper?: genomic research and the identifiability of siblings. *BMC Med Genomics* 2008; **1:** 32.

9 Schadt EE. The changing privacy landscape in the era of big data. *Mol Syst Biol* 2012; **8:** 612.

10 Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science* 2013; **339:** 321–24.

11 Lippert C, Sabatini R, Maher MC, et al. Identification of individuals by trait prediction using whole-genome sequencing data. *Proc Natl Acad Sci USA* 2017; **114:** 10166–71.

12 Tomofuji Y, Sonehara K, Kishikawa T, et al. Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nat Microbiol* 2023; **8:** 1079–94.

13 Guccione C, Patel L, Tomofuji Y, et al. Incomplete human reference genomes can drive false sex biases and expose patient-identifying information in metagenomic data. *Nat Commun* 2025; **16:** 825.

14 von Thenen N, Ayday E, Cicek AE. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics* 2019; **35:** 365–71.

15 Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. *Science* 2018; **362:** 690–94.

16 Bush SJ, Connor TR, Peto TEA, Crook DW, Walker AS. Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microb Genom* 2020; **6:** mgen000393.

17 Forbes M, Ng DYK, Boggan RM, et al. Benchmarking of human read removal strategies for viral and microbial metagenomics. *Cell Rep Methods* 2025; **5:** 101218.

18 Sarhan MS, Antonello G, Weissensteiner H, et al. Human mitochondrial DNA in public metagenomes: opportunity or privacy threat? *Cell* 2025; **188:** 2561–66.

19 Aziz MMA, Ghasemi R, Waliullah M, Mohammed N. Aftermath of Bustamante attack on genomic beacon service. *BMC Med Genomics* 2017; **10** (suppl 2)**:** 43.

20 Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. *Am J Hum Genet* 2015; **97:** 631–46.

21 Oliva A, Kaphle A, Reguant R, et al. Future-proofing genomic data and consent management: a comprehensive review of technology innovations. *Gigascience* 2024; **13:** giae021.

22 Guerrini CJ, Kalokairinou L, Robinson JO, et al. Investigative genetic genealogy practices warranting policy attention: results of a modified policy Delphi. *PLoS Genet* 2025; **21:** e1011520.

23 Doi H, Kelly RP. Ethical considerations for human sequences in environmental DNA. *Nat Ecol Evol* 2023; **7:** 1334–35.

24 de Groot NF. Capturing human environmental DNA: ethical challenges. *DISO* 2023; **2:** 43.

25 Whitmore L, McCauley M, Farrell JA, et al. Inadvertent human genomic bycatch and intentional capture raise beneficial applications and ethical concerns with environmental DNA. *Nat Ecol Evol* 2023; **7:** 873–88.

26 Ram N. The ethics of human sequences in environmental samples. *Nat Ecol Evol* 2023; **7:** 796–97.

27 WHO. WHO guiding principles for pathogen genome data sharing. Nov 8, 2022. https://www.who.int/publications/i/item/9789240061743 (accessed Oct 10, 2025).

28 WHO. Guidance for human genome data collection, access, use and sharing. Nov 20, 2024. https://www.who.int/publications/i/item/9789240102149 (accessed Oct 10, 2025).

29 Constantinides B, Hunt M, Crook DW. Hostile: accurate decontamination of microbial host sequences. *Bioinformatics* 2023; **39:** btad728.

30 Hunt M, Swann J, Constantinides B, Fowler PW, Iqbal Z. ReadItAndKeep: rapid decontamination of SARS-CoV-2 sequencing reads. *Bioinformatics* 2022; **38:** 3291–93.

31 Kluin RJC, Kemper K, Kuilman T, et al. XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. *BMC Bioinformatics* 2018; **19:** 366.

32 Khandelwal G, Girotti MR, Smowton C, et al. Next-generation sequencing analysis and algorithms for PDX and CDX models. *Mol Cancer Res* 2017; **15:** 1012–16.

33 BBtools. BBMap. https://sourceforge.net/projects/bbmap/ (accessed July 11, 2025).

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

35 Zentgraf J, Rahmann S. Fast lightweight accurate xenograft sorting. *Algorithms Mol Biol* 2021; **16:** 2.

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

37 Constantinides B, Lees J, Crook DW. Deacon: fast sequence filtering and contaminant depletion. *bioRxiv* 2025; published online June 12. https://doi.org/10.1101/2025.06.09.658732 (preprint).

38 Rumbavicius I, Rounge TB, Rognes T. HoCoRT: host contamination removal tool. *BMC Bioinformatics* 2013; **24:** 371.

39 Lataretu M, Krautwurst S, Huska MR, et al. Targeted decontamination of sequencing data with CLEAN. *NAR Genom Bioinform* 2025; **7:** lqaf105.

40 Seidel J, Kaipf C, Straub D, Nahnsen S. nf-core/detaxizer: a benchmarking study for decontamination from human sequences. *NAR Genom Bioinform* 2025; **7:** lqaf125.

41 Liu D, Hu J, Zhang D, et al. Use of the CHM13-T2T genome improves metagenomic analysis by minimizing host DNA contamination. *mSystems* 2025; **10:** e0084025.

42 Hall MB, Coin LJM. Pangenome databases improve host removal and mycobacteria classification from clinical metagenomic data. *GigaScience* 2024; **13:** giae010.

43 Bulut B, Başcı AB, Olcay HE. Declaration of Helsinki: comparison of the 2024 version and the 2013 version. *Infect Dis Clin Microbiol* 2025; **7:** 112–13.

44 Ballantyne A. Adjusting the focus: a public health ethics approach to data research. *Bioethics* 2019; **33:** 357–66.

45 Convention on Biological Diversity. The Nagoya Protocol on access and benefit-sharing. https://www.cbd.int/abs (accessed April 23, 2025).

46 European Nucleoide Archive. ENA Browser. https://www.ebi.ac.uk/ena/browser/home (accessed April 23, 2025).

47 National Center for Biotechnology Information Sequencing Read Archive (SRA). https://www.ncbi.nlm.nih.gov/sra (accessed April 23, 2025).

48 Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol* 2021; **22:** 270.