



**HAL**  
open science

## **Microseek: an innovative approach to unraveling clinical metagenomic secrets**

Thomas Bigot, Philippe Pérot, Sarah Temmam, Béatrice Regnault, Marc Eloit

### ► **To cite this version:**

Thomas Bigot, Philippe Pérot, Sarah Temmam, Béatrice Regnault, Marc Eloit. Microseek: an innovative approach to unraveling clinical metagenomic secrets. 8th International Conference on Clinical Metagenomics, Nov 2023, Geneva, Switzerland. pasteur-04337913

**HAL Id: pasteur-04337913**

**<https://pasteur.hal.science/pasteur-04337913v1>**

Submitted on 12 Dec 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Microseek: an innovative approach to unraveling clinical metagenomic secrets

Thomas Bigot<sup>1§</sup>, Philippe Pérot<sup>2§</sup>, Sarah Temmam<sup>1</sup>, Béatrice Regnault<sup>1</sup> & Marc Eloit<sup>1,3</sup>

1. Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, 75015 Paris, France  
 2. Institut Pasteur, Université Paris Cité, Pathogen Discovery Laboratory, 75015 Paris, France  
 3. École Nationale Vétérinaire d'Alfort, 94700 Maisons-Alfort, France  
 § These authors contributed equally.



## BACKGROUND

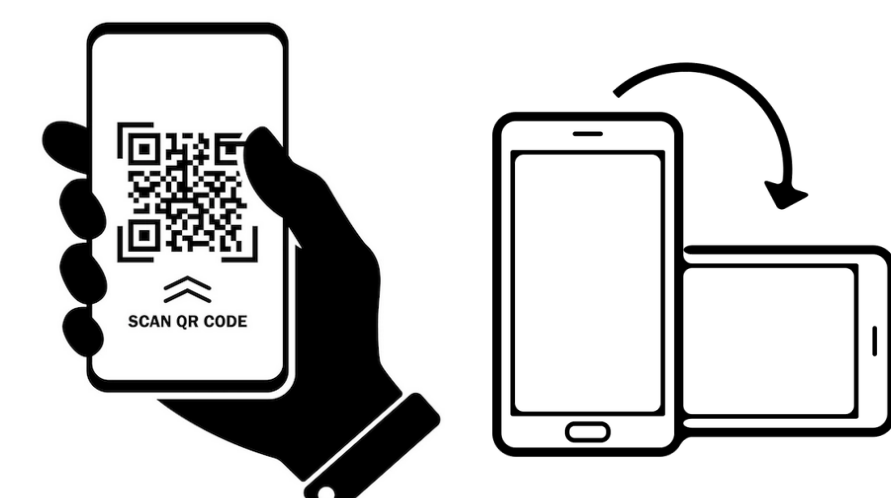
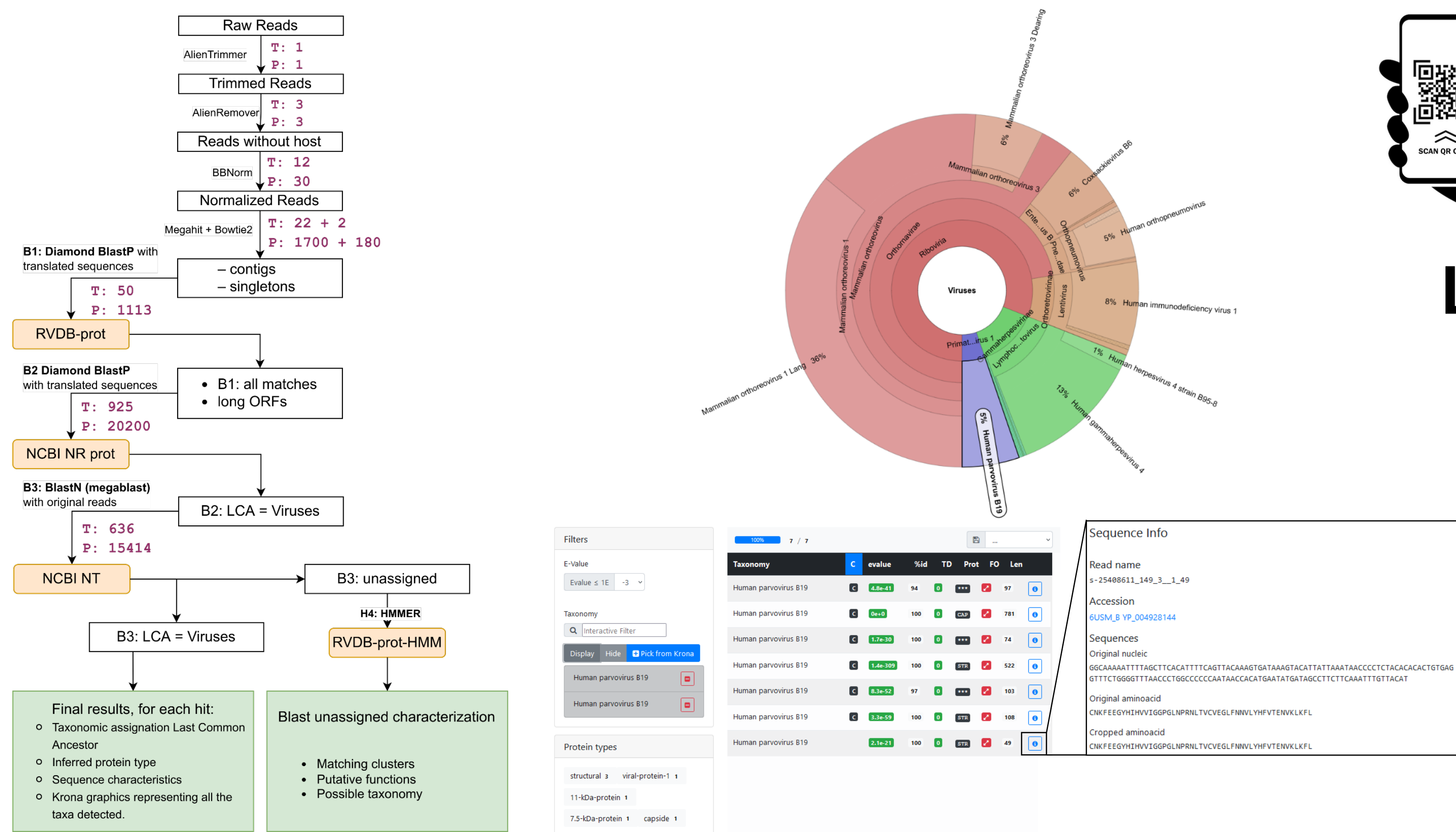
We present Microseek, a pipeline for virus identification and discovery, integrated with RVDB-prot, a comprehensive, curated and regularly updated database of viral proteins.

## METHODS

Microseek analyses metagenomic Next Generation Sequencing (mNGS) raw data by performing quality steps, de novo assembly, and by scoring the Lowest Common Ancestor (LCA) from translated reads and contigs. Microseek runs on a local computer. The outcome of the pipeline is displayed through a user-friendly and dynamic graphical interface.

### The pipeline & its user-friendly interface

**Figure 1. Microseek pipeline steps and output examples.** A/ Step details of the pipeline. For each operation, software and database names are indicated. The times on the right represent the cumulated duration of each step (in minutes), respectively for the Tissues (T) and Plasma (P) datasets and corresponding to spiked experiments at d1. For instance, for Plasma dataset, B1 step would take 1113 minutes if it was run with a single CPU as a unique chunk. B/ Example of a result browser webpage. The Parvovirus B19 taxon was selected from the Krona chart as an example. C/ Resulting list of hits of the Tissues dataset after filtering on Human Parvovirus B19 with e-value  $\leq 10^{-3}$ , showing information such as the existence of contigs (C), e-value, %ID, length of the sequence and protein names. Information boxes associated to each hit give the nt and aa sequences and allow direct access to the corresponding NCBI entries.



LIVE DEMO  
SCAN ME!



Pérot, P.; Bigot, T.; Temmam, S.; Regnault, B.; Eloit, M. Microseek: A Protein-Based Metagenomic Pipeline for Virus Diagnostic and Discovery. *Viruses* 2022

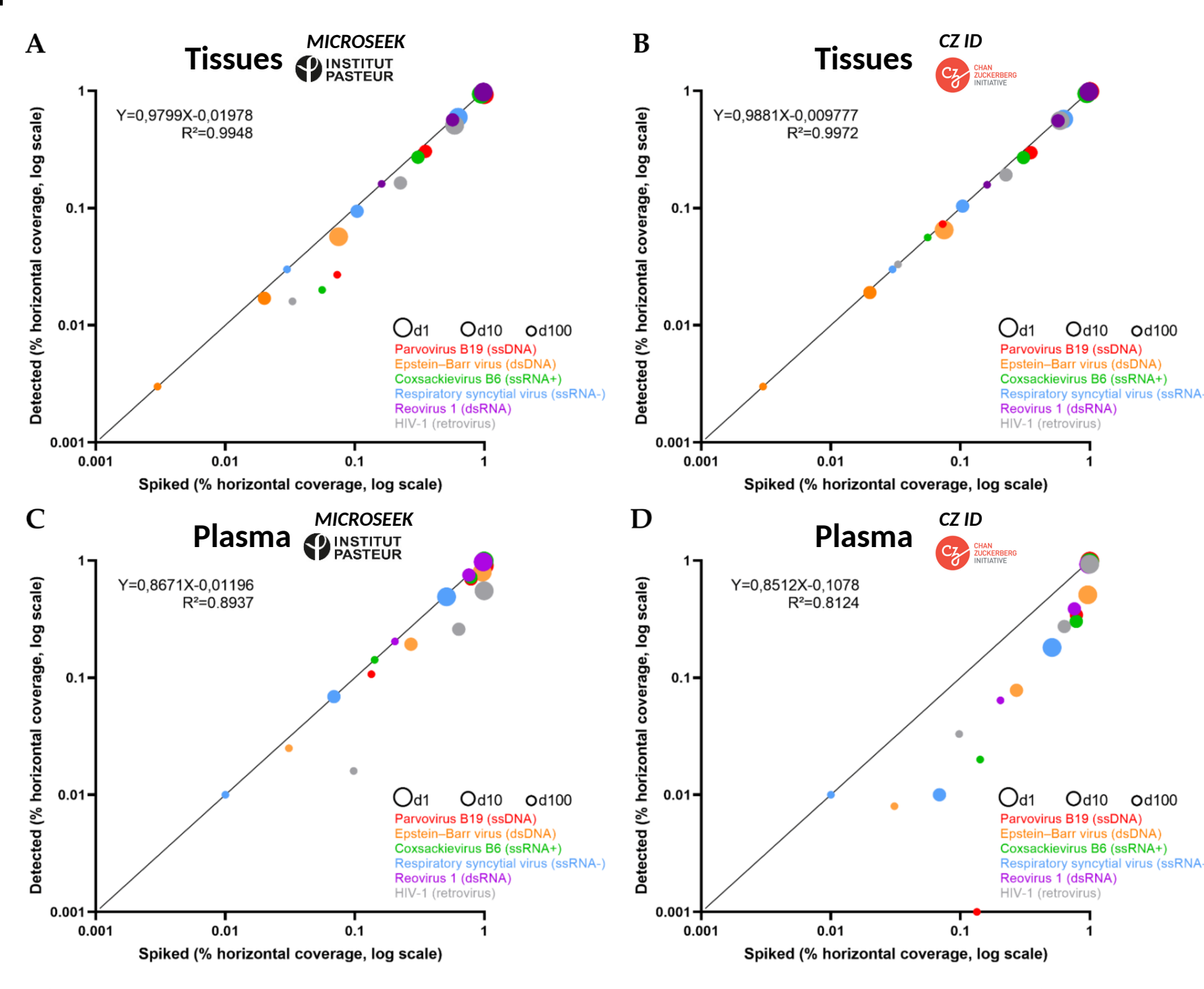
## RESULTS

Drawing from two representative mNGS datasets obtained from human tissue and plasma specimens, we present its performance. In-silico spikes of known viral sequences, as well as spikes of fabricated Neopneumovirus viral sequences generated with varying evolutionary distances from known members of the Pneumoviridae family, were employed. Results were compared to Chan Zuckerberg ID (CZ ID), a benchmark cloud-based mNGS pipeline.

### Microseek versus CZ ID

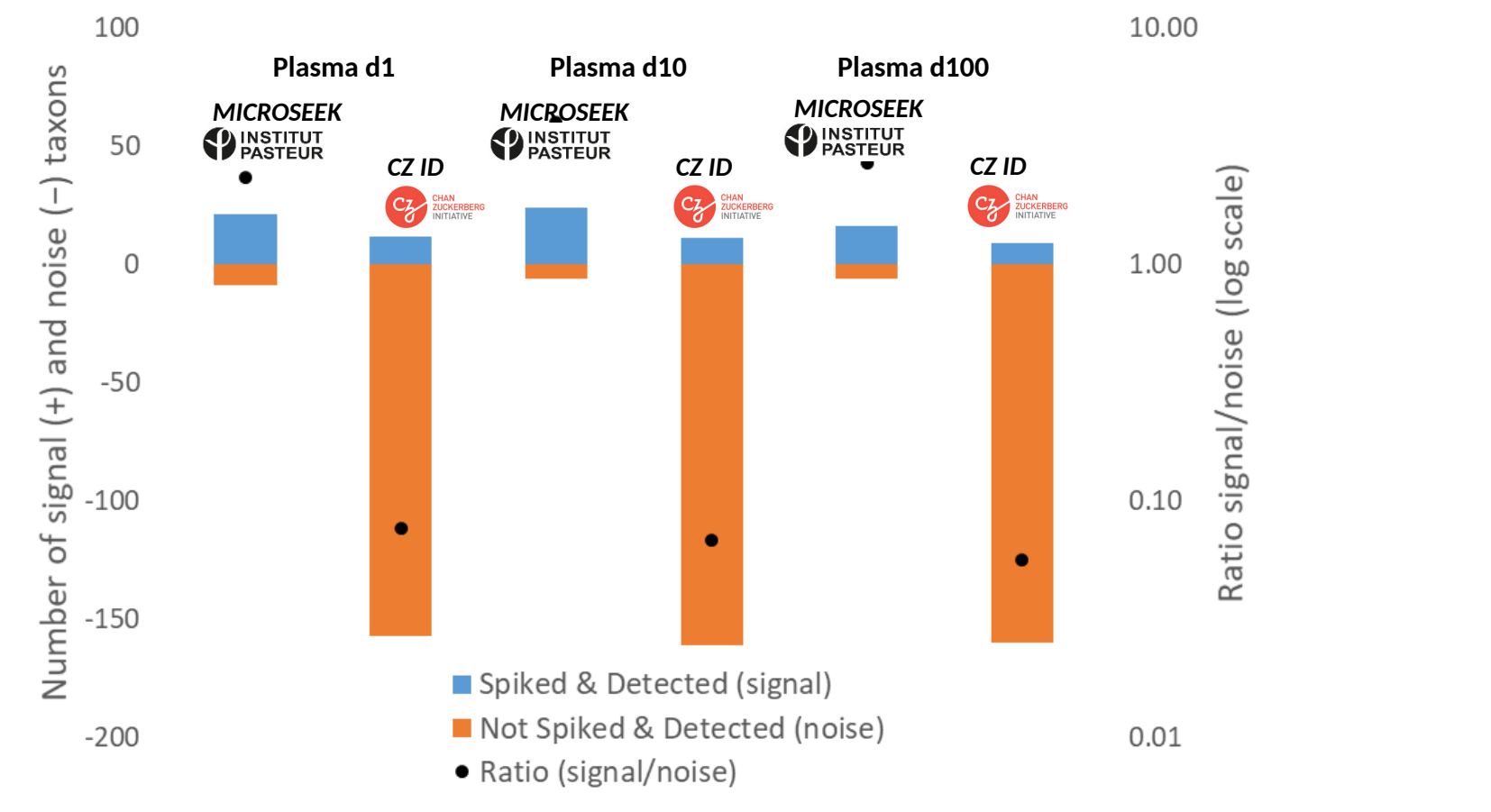
**Table 1:** Six known human viruses representing six groups of the Baltimore classification were selected as input reference sequences for spiking experiments. ssDNA: Human parvovirus B19; dsDNA: EBV; ssRNA+: Coxsackievirus B6; ssRNA-: Human respiratory syncytial virus A (HRSV-A); dsRNA: Mammalian orthoreovirus type 1; and retrovirus: Human immunodeficiency virus type 1 (HIV-1).

Color	Genome Type	Virus
Red	ssDNA	Parvovirus B19
Orange	dsDNA	Epstein-Barr virus
Green	ssRNA+	Coxsackievirus B6
Blue	ssRNA-	Respiratory syncytial virus
Purple	dsRNA	Reovirus 1 (10 segments)
Grey	Retrovirus	HIV-1

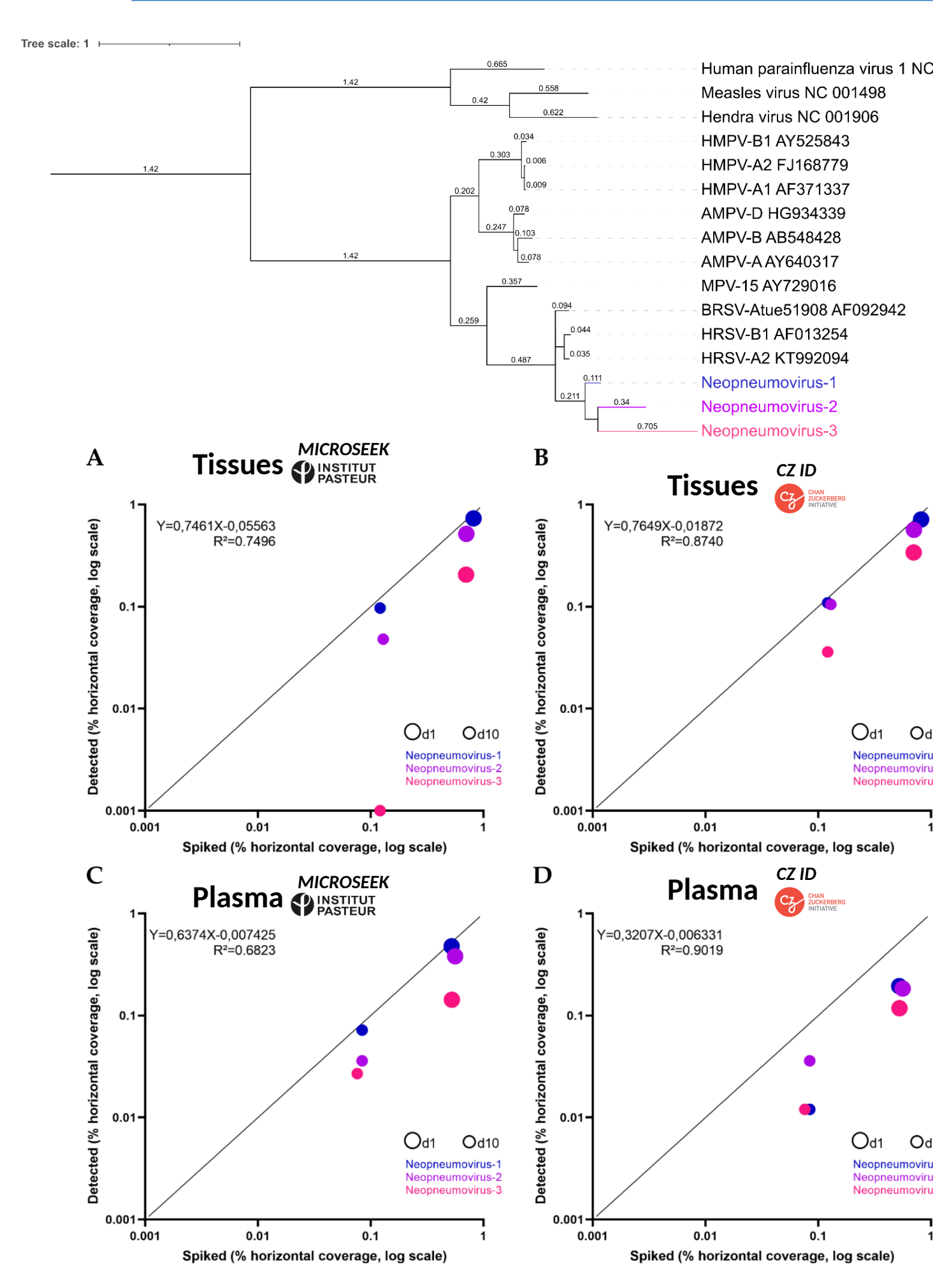


**Figure 3.** Detection of six known viruses (spiked vs. detected). Scatter plots of spiked (X axis) vs. detected (Y axis) percentage horizontal coverage for the 6 known viruses of the study, for the Tissues (A, B) and Plasma (C, D) experiments. Each virus is depicted by 3 points, corresponding to the d1 (large circles), d10 (intermediate circles) and d100 (small circles) experiments, and is associated with a specific color (red: Parvovirus B19; orange: Epstein-Barr virus; green: Coxsackievirus B6; blue: HRSV; purple: Reovirus-1; grey: HIV-1). The identity line (Y = X) is represented in black.

**Figure 5.** Signal-to-noise ratio in the Plasma experiments (known viruses). Histograms below the table represent the signal (blue bars with counts above zero on the primary axis) vs. noise (orange bars with counts below zero on the primary axis) and associated signal-to-noise ratios (black dots on the secondary axis).

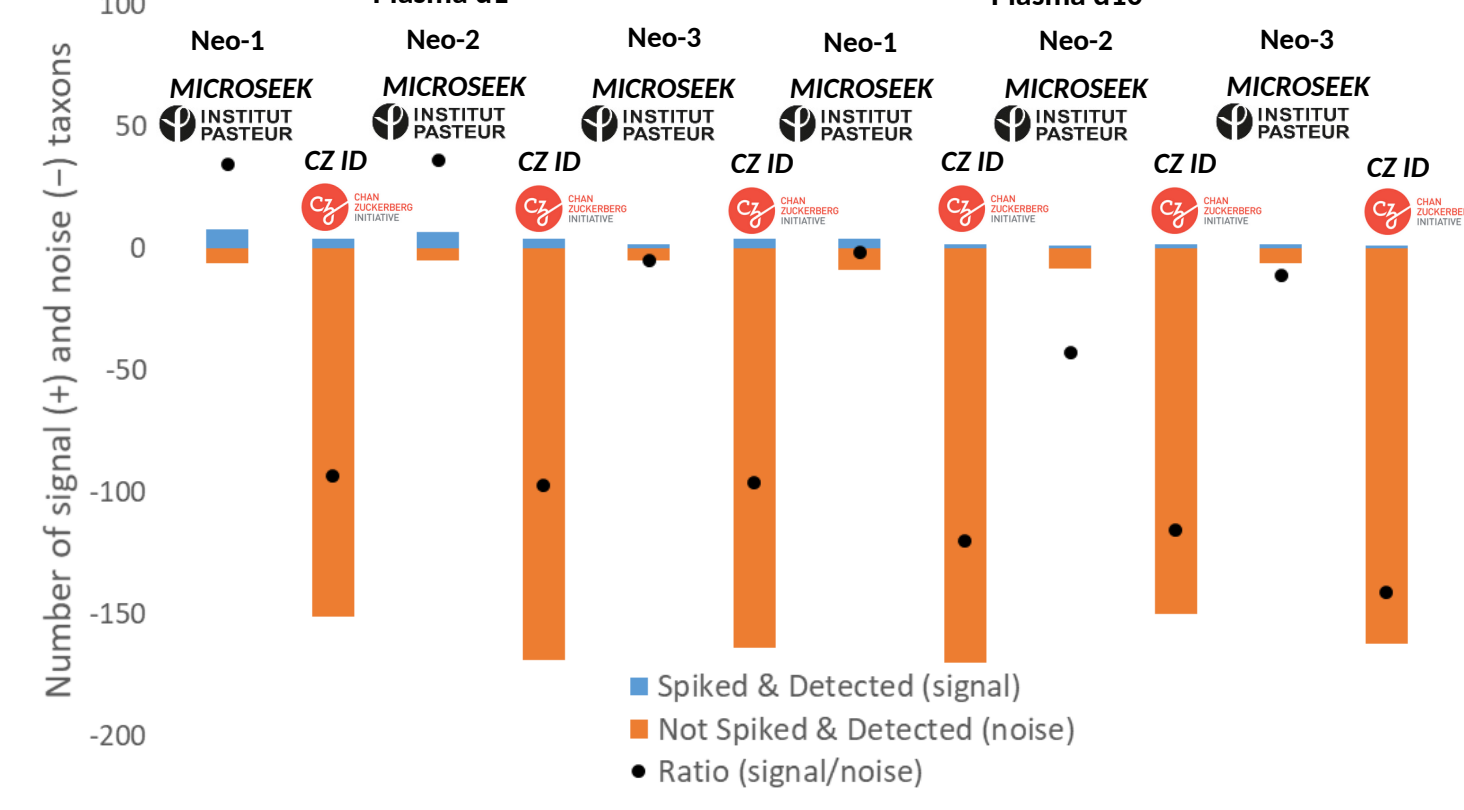


### Detection of novel viruses



**Figure 2.** Phylogeny of Neopneumoviruses. Protein sequences of the polymerases were aligned with MAFFT v7.450, and the phylogeny was reconstructed with IQtree 2.0.6 with model of sub-stitution JTT+F+G4. AMPV: Avian metapneumovirus; HMPV: Human metapneumovirus; MPV: Murine orthopneumovirus; BRSV: Bovine orthopneumovirus; HRSV: Human respiratory syncytial virus.

**Figure 4.** Detection of three Neopneumoviruses (spiked vs. detected). Scatter plots of spiked (X axis) vs. detected (Y axis) percentage horizontal coverage for the 3 fake Neopneumoviruses of the study, for the Tissues (A, B) and Plasma (C, D) experiments. Each virus is depicted by 2 points, corresponding to the d1 (intermediate circles) and d10 (small circles) experiments, and is associated with a specific color (blue: Neopneumovirus-1; purple: Neopneumovirus-2; pink: Neopneumovirus-3). The identity line (Y = X) is represented in black.



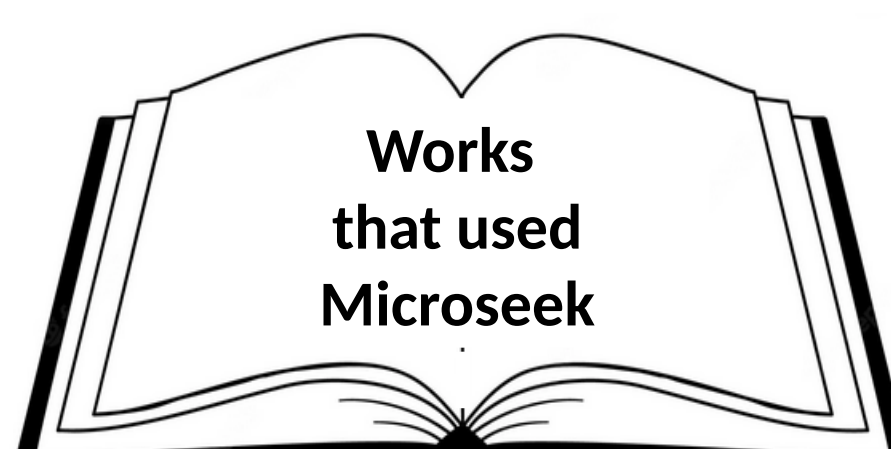
**Figure 6.** Signal-to-noise ratio in the Plasma experiments (novel viruses). Histograms below the table represent the signal (blue bars with counts above zero on the primary axis) vs. noise (orange bars with counts below zero on the primary axis) and associated signal-to-noise ratios (black dots on the secondary axis).

Clinical Infectious Diseases  
**MAJOR ARTICLE**  
 Identification of Umbre Orthobunyavirus as a Novel Zoonotic Virus Responsible for Lethal Encephalitis in 2 French Patients with Hypogammaglobulinemia  
 Sarah Temmam<sup>1</sup>, Fabrice Bouteau<sup>1</sup>, Thomas Bigot<sup>1</sup>, Vincent Faudouzi<sup>1</sup>, Karine Bellou<sup>1</sup>, Delphine Chretien<sup>1</sup>, Patricia Gil<sup>1</sup>, Sarah Collin<sup>1</sup>, George Chikoti<sup>1</sup>, Karim Makani<sup>1</sup>, Jan Kellert<sup>1</sup>, Marie Perennet<sup>1</sup>, Caroline Sautou<sup>1</sup>, Marjolaine Couderc<sup>1</sup>, Anne Beaudou<sup>1</sup>, Sarah Temmam<sup>1</sup>, Thery Hoen<sup>1</sup>, Edward De Souza Cunha<sup>1</sup>, Susana Beliz<sup>1</sup>, Isabelle Pflieger<sup>1</sup>, Marie Bernadette Delisle<sup>1</sup>, Fabrice Bouteau<sup>1</sup>, David Brossat<sup>1</sup>, Claire Fischl<sup>1</sup>, Marine Malgouyres<sup>1</sup>, Charles Doucette<sup>1</sup>, Bertrand Malin<sup>1</sup>, Sophie Demere<sup>1</sup>, Danielle Serrhini<sup>1</sup> and Marc Eloit<sup>1,2,3</sup>

nature  
 About the Journal | Publish with us  
 Article | Published: 16 February 2022  
**Bat coronaviruses related to SARS-CoV-2 and infectious for human cells**  
 Sarah Temmam<sup>1</sup>, Khamsing Vongphayloth<sup>1</sup>, Edward Baquero<sup>1</sup>, Sandie Munier<sup>1</sup>, Maximiliano Bonomi<sup>1</sup>, Béatrice Regnault<sup>1</sup>, Bounsarane Douangboupha<sup>1</sup>, Yesaman Karan<sup>1</sup>, Delphine Chretien<sup>1</sup>, Daoananh Sananayak<sup>1</sup>, Vilakhan Kayapath<sup>1</sup>, Phetpoum Paphaphah<sup>1</sup>, Vincent Lacoste<sup>1</sup>, Sompahavanh Sornke<sup>1</sup>, Khaithong Lakemany<sup>1</sup>, Notbasin Phommavanh<sup>1</sup>, Philippe Pérot<sup>1</sup>, Odane Dehan<sup>1</sup>, Faustine Amara<sup>1</sup>, Flora Donati<sup>1</sup>, Thomas Bigot<sup>1</sup>, Michael Nilgys Felix A. Rey<sup>1</sup>, Sylvie van der Werf<sup>1</sup>, Paul T. Brey<sup>1</sup> & Marc Eloit<sup>1,2,3</sup>

Clinical Infectious Diseases  
**MAJOR ARTICLE**  
 First Case of Lethal Encephalitis in Western Europe Due to European Bat Lyssavirus Type 1  
 Béatrice Regnault<sup>1</sup>, Bruno Enard<sup>1</sup>, Isabelle Pflieger<sup>1</sup>, Laurent Decheux<sup>1</sup>, Eric Trésac<sup>1</sup>, Pascal Cazotte<sup>1</sup>, Delphine Chretien<sup>1</sup>, Mathilde Decheux<sup>1</sup>, Marie Michel<sup>1</sup>, Anne Janet<sup>1</sup>, Marianne Lopez<sup>1</sup>, Philippe Pérot<sup>1</sup>, Hervé Bourry<sup>1</sup>, Marc Eloit<sup>1,2,3</sup> and Danielle Serrhini<sup>1,2,3</sup>

frontiers  
 in Microbiology  
**Monitoring Silent Spillovers Before Emergence: A Pilot Study at the Tick/Human Interface in Thailand**  
 Sarah Temmam<sup>1</sup>, Delphine Chretien<sup>1</sup>, Thomas Bigot<sup>1,2</sup>, Evelynne Dufour<sup>1</sup>, Stéphane Pitres<sup>1</sup>, Marc Desquesnes<sup>1,3,4</sup>, Etienne Devillers<sup>1</sup>, Marine Dumarest<sup>1</sup>, Léna Youssif<sup>1</sup>, Setthaporn Kitpipongchai<sup>1</sup>, Ananika Karachonkarnong<sup>1</sup>, Kitpipong Chairat<sup>1</sup>, Léa Gagnaire<sup>1</sup>, Jean-François Cosson<sup>1</sup>, Muriel Vayssières-Tausat<sup>1</sup>, Serge Morand<sup>1,2</sup>, Sara Moutaillier<sup>1,2</sup> and Marc Eloit<sup>1,2,3</sup>



**CONCLUSION**  
 Microseek reliably identifies known viral sequences and performs well for the detection of distant pseudoviral sequences, especially in complex samples such as in human plasma, while minimizing non-relevant hits.