



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^{1§}, Philippe Pérot^{2§}, Sarah Temmam¹, Béatrice Regnault¹ & Marc Eloit^{1,3}

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

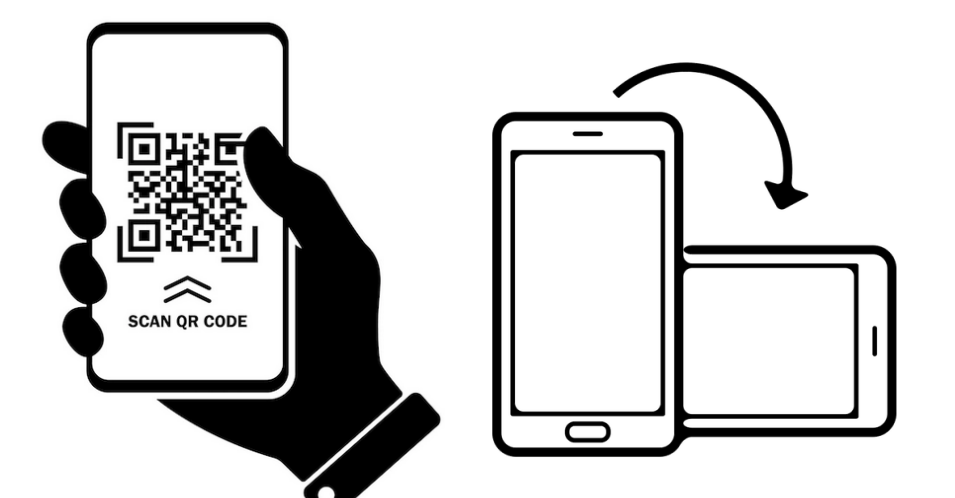
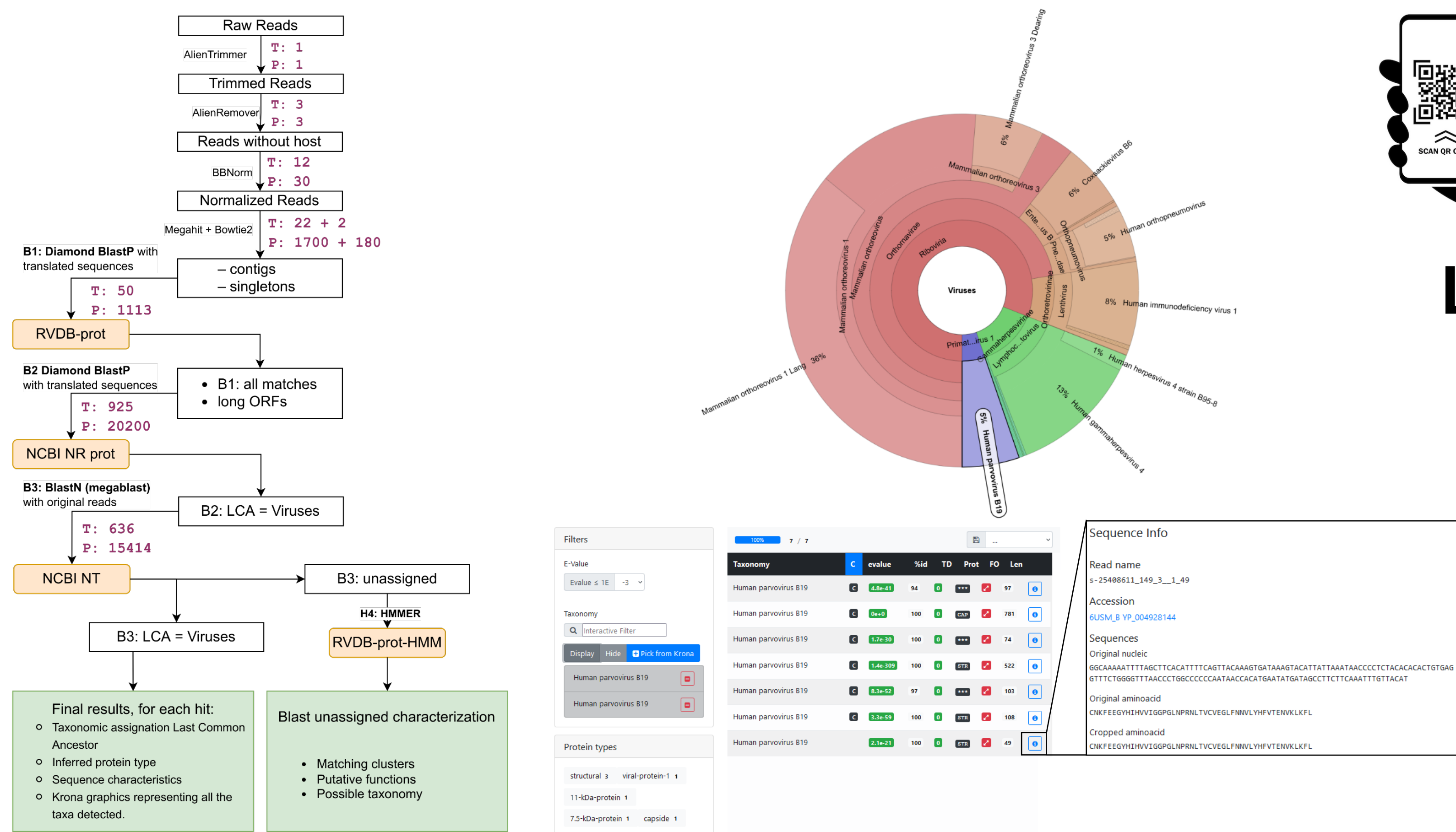
We present Microseek, a pipeline for virus identification and discovery, integrated with RVDB-prot, a comprehensive, curated and regularly updated database of viral proteins.

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).

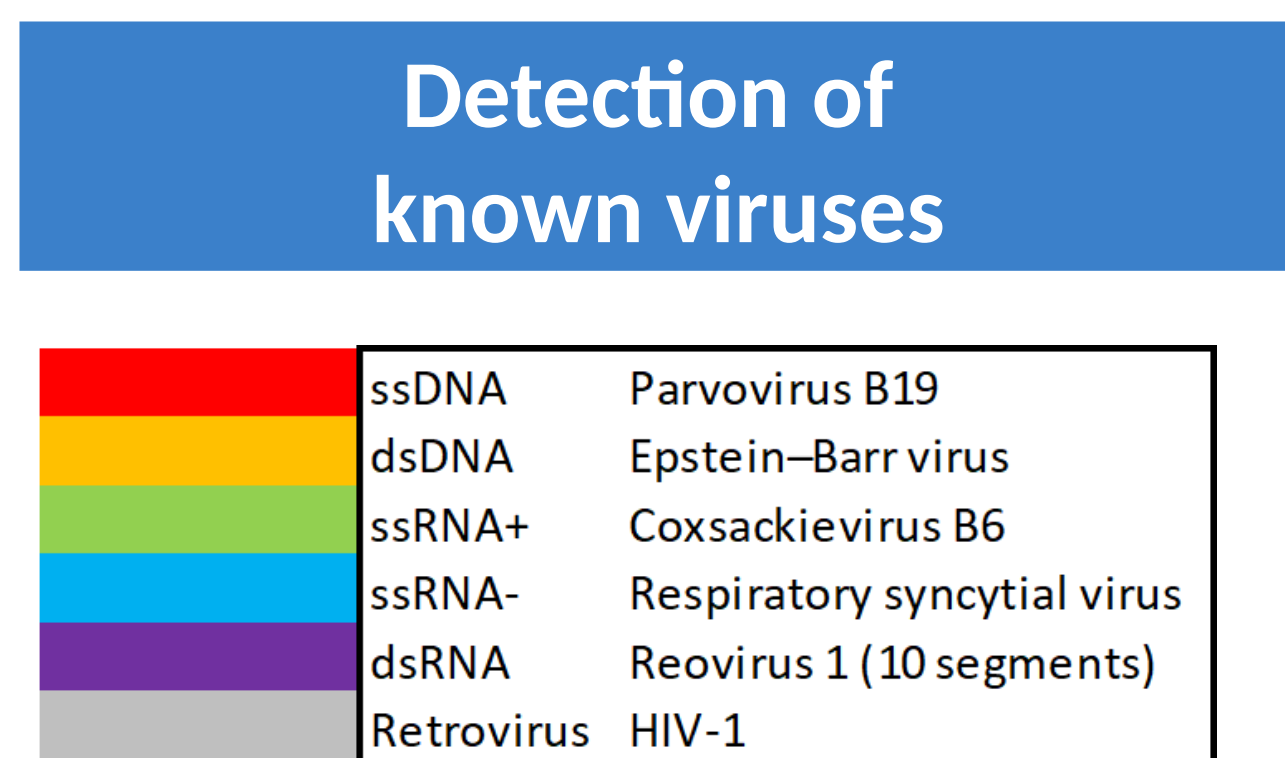
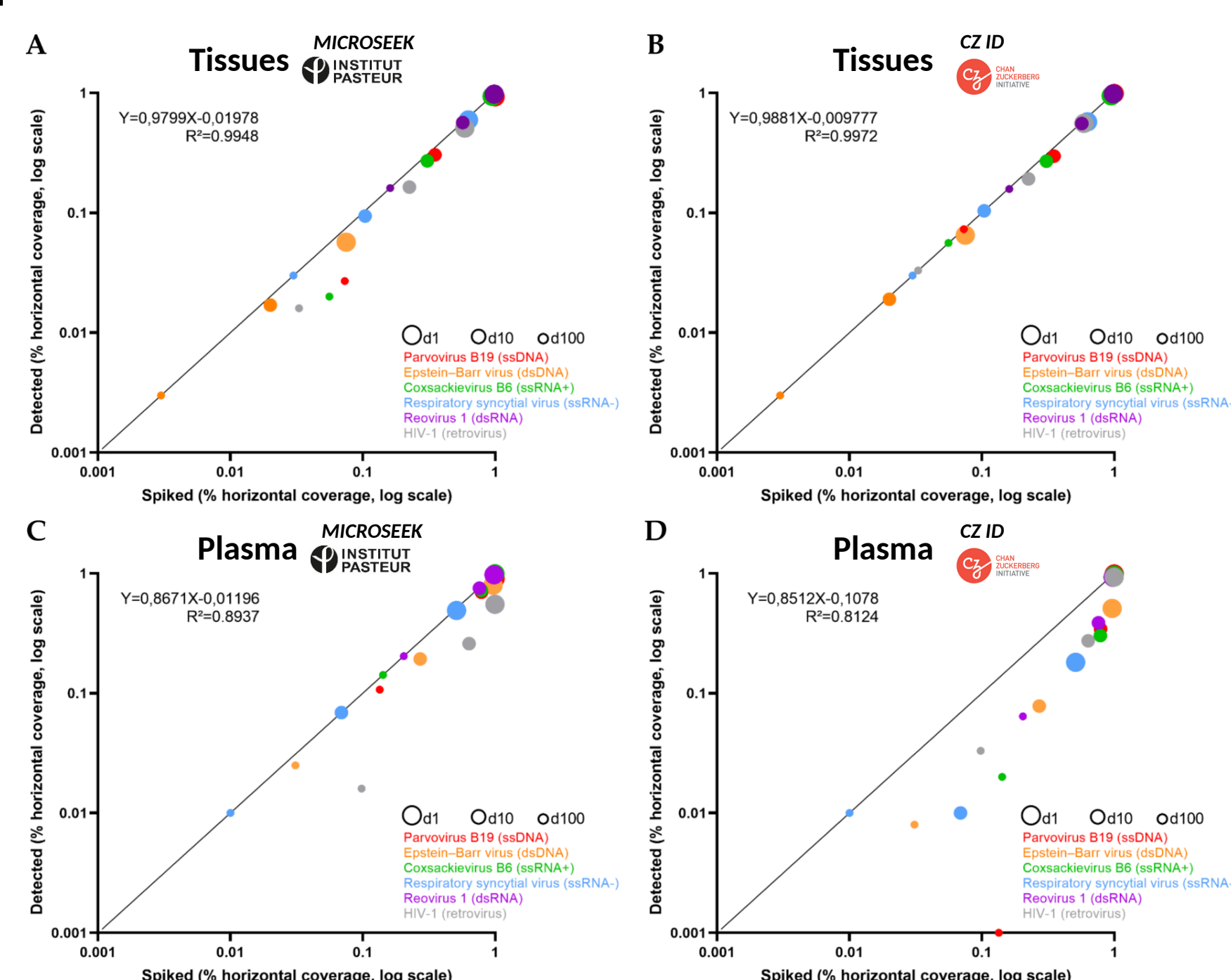


Figure 2. Phylogeny of Neopneumoviruses. Protein sequences of the polymerases were aligned with MAFFT v7.450, and the phylogeny was reconstructed with IQtree 2.0.6 with model of sub-stitution JTT+F+G4. AMPV: Avian metapneumovirus; HMPV: Human metapneumovirus; MPV: Murine orthopneumovirus; BRSV: Bovine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

Figure 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.



Detection of novel viruses

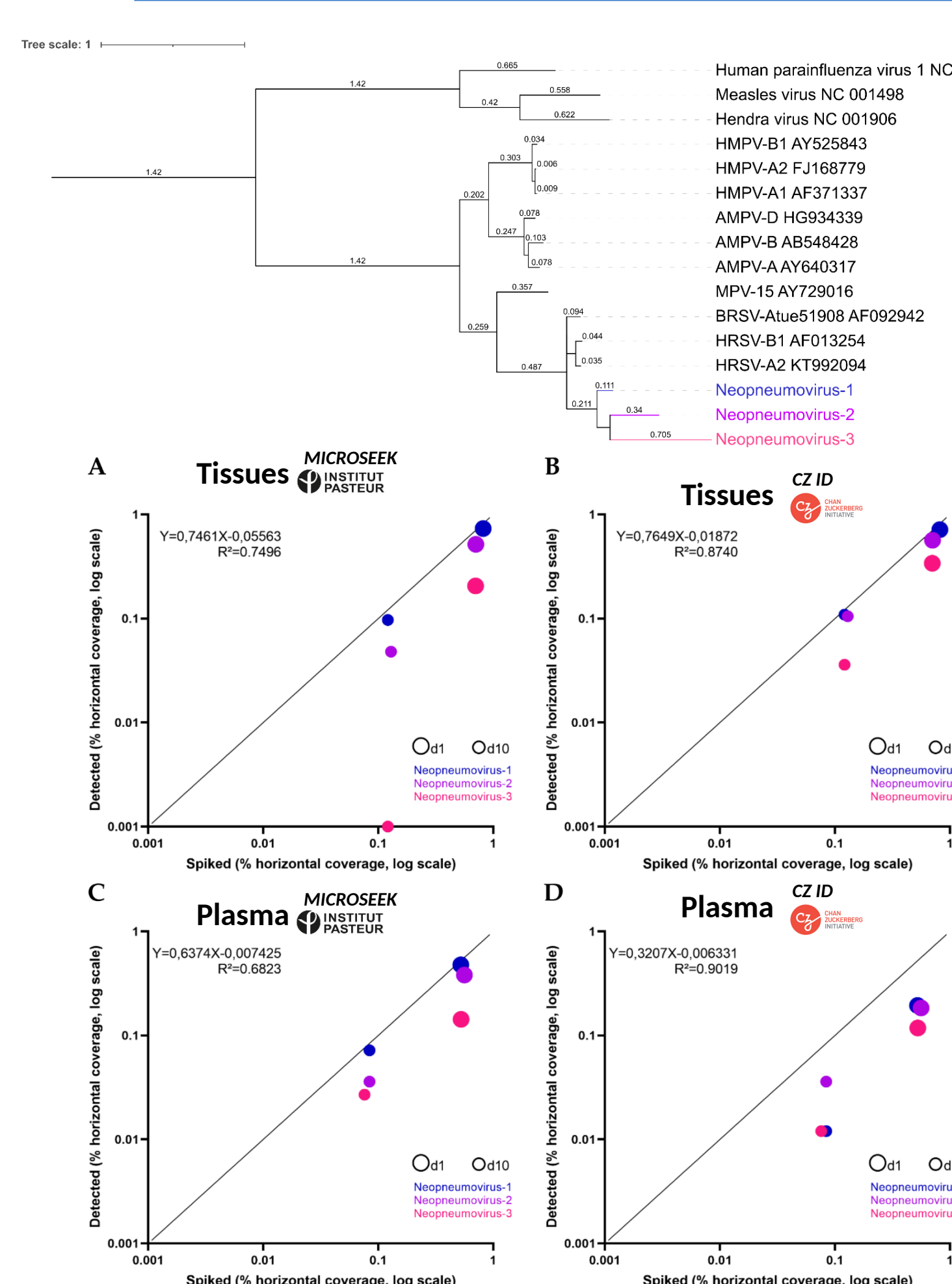


Figure 4. Detection of three Neopneumoviruses (spiked vs. detected). Scatter plots of spiked (X axis) vs. detected (Y axis) percentage horizontal coverage for the 3 fake Neopneumoviruses of the study, for the Tissues (A,B) and Plasma (C,D) experiments. Each virus is depicted by 2 points, corresponding to the d1 (large circles) and d10 (intermediate 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 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).

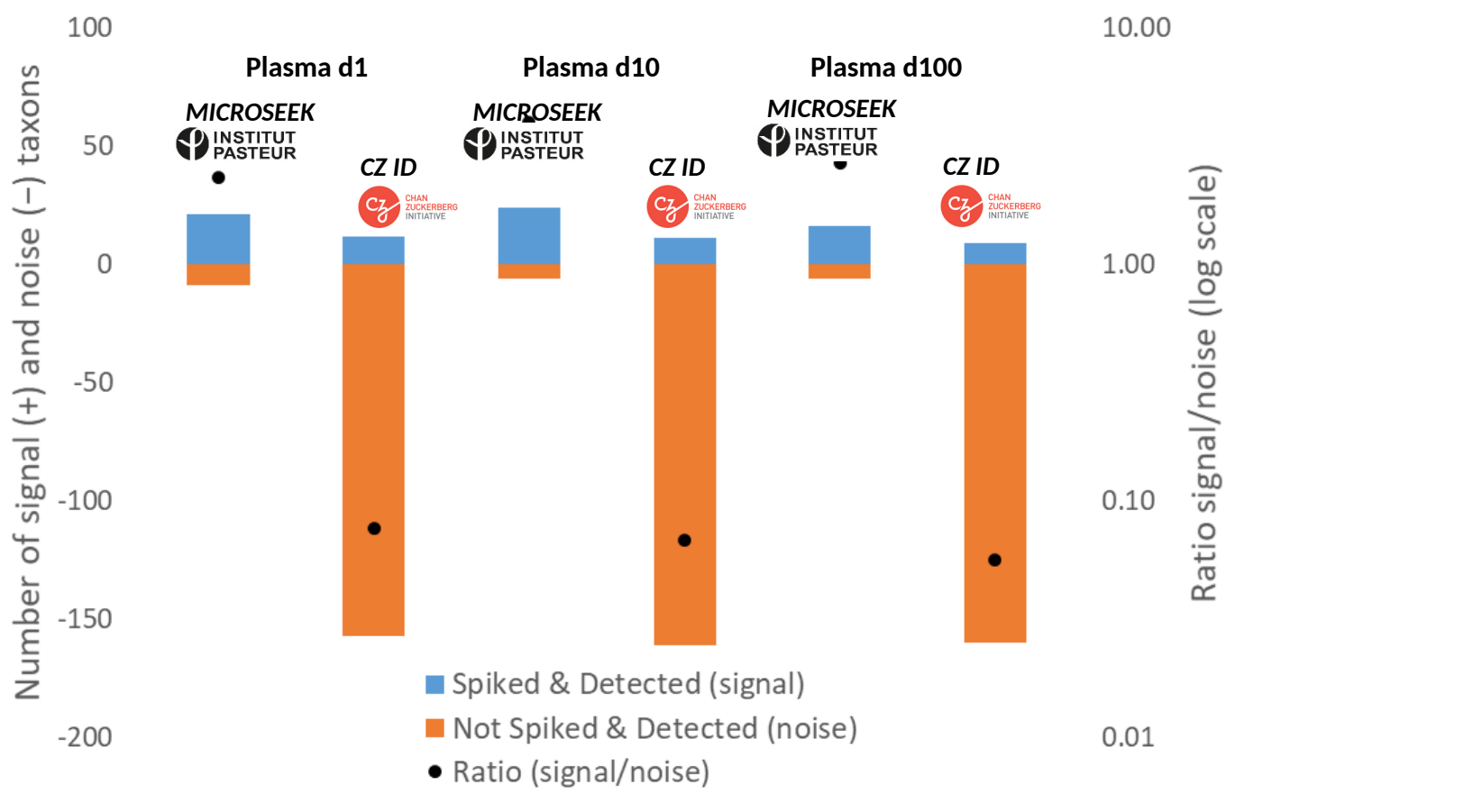


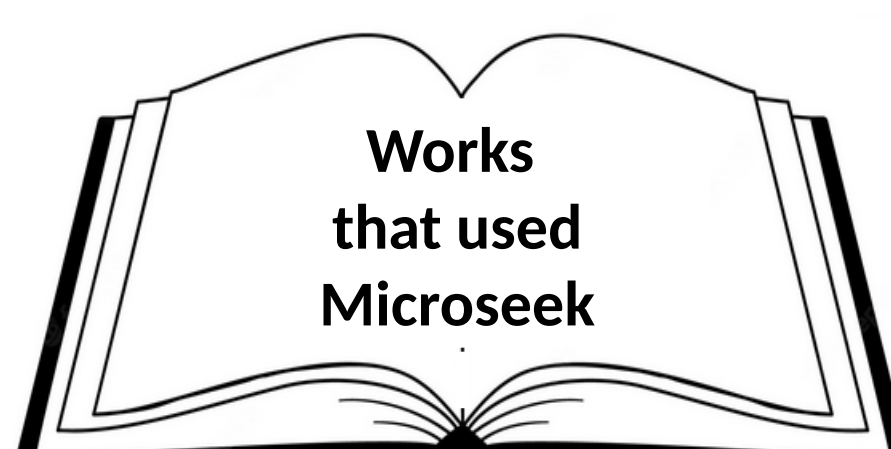
Figure 6. Signal-to-noise ratio in the Plasma experiments (novel viruses). Histograms below the table represent the signal (blue bars with counts above zero on the primary axis) vs. noise (orange bars with counts below zero on the primary axis) and associated signal-to-noise ratios (black dots on the 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 Temmam¹, Thomas Bigot^{1,2}, Vincent Faudouzi¹, Karine Bellou¹, Delphine Chretien¹, Patricia Gil¹, Sarah Collin^{1,2}, George Chikoti¹, Karim Makani¹, Jan Kellert¹, Marie Perennet¹, Caroline Sautou¹, Marjolaine Couderc¹, Anne Bernard de Verennes¹, Sarah Temmam¹, Thierry Couderc¹, Edward De Souza Cunha¹, Susana Beliz^{1,2}, Isabelle Pita^{1,2}, Marie Bernadette Delisle¹, Fabrice Bouteiller¹, David Brossat¹, Claire Fischl¹, Marine Malphe¹, Charles Doucette¹, Bertrand Malin^{1,2}, Sophie Demere¹, Danielle Serrano^{1,2,3} and Marc Eloit^{1,2,3}

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 Temmam¹, Khamsing Vongphayloth¹, Edward Baquero¹, Sandie Munier¹, Maximiliano Bonomi¹, Béatrice Regnault¹, Bounsarane Douangboupha¹, Yesaman Karan¹, Delphine Chretien¹, Daoananh Sananayak¹, Vilakhan Kayapath¹, Phetpoum Paphaphah¹, Vincent Lacoste¹, Sompahavanh Sornke¹, Khaithong Lakomyay¹, Notbasin Phommavanh¹, Philippe Pérot¹, Odane Dehan¹, Faustine Amara¹, Flora Donati¹, Thomas Bigot¹, Michael Nilgys Felix A. Rey¹, Sylvie van der Werf¹, Paul T. Brey¹ & Marc Eloit^{1,2}

Clinical Infectious Diseases
MAJOR ARTICLE
First Case of Lethal Encephalitis in Western Europe Due to European Bat Lyssavirus Type 1
Béatrice Regnault¹, Bruno Enard¹, Isabelle Pita¹, Laurent Decheux¹, Eric Trésac¹, Pascal Cazotte¹, Delphine Chretien¹, Mathilde Decheux¹, Marie Michel¹, Anne Janet¹, Marianne Lopez¹, Philippe Pérot¹, Hervé Bourry¹, Marc Eloit^{1,2,3} and Danielle Serrano^{1,2,3}

frontiers
in Microbiology
Monitoring Silent Spillovers Before Emergence: A Pilot Study at the Tick/Human Interface in Thailand
Sarah Temmam¹, Delphine Chretien¹, Thomas Bigot^{1,2}, Evelynne Dufour¹, Stéphane Pitres¹, Marc Desquesnes^{1,2,3}, Elodie Devillers¹, Marine Dumarest¹, Léna Youssif¹, Setthaporn Kitpipongchai¹, Ananika Karachonkarnong¹, Kitpipong Chairat¹, Léa Gagnaire¹, Jean-François Cosson¹, Muriel Vayssières-Tausat¹, Serge Morand^{1,2}, Sara Moutaillou¹ and Marc Eloit^{1,2,3}



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.