RdRp Summit 2023

22 - 23 May, 2023

ADEIT – Fundación Universidad Empresa de la Universitat de Valéncia

The 1st International Conference for Interoperability in RNA virus discovery

Sponsored by

Chan Zuckerberg Initiative

frontiers

Satellite event to

ViBioM 2023

Contents

Rationale of the RdRp Summit

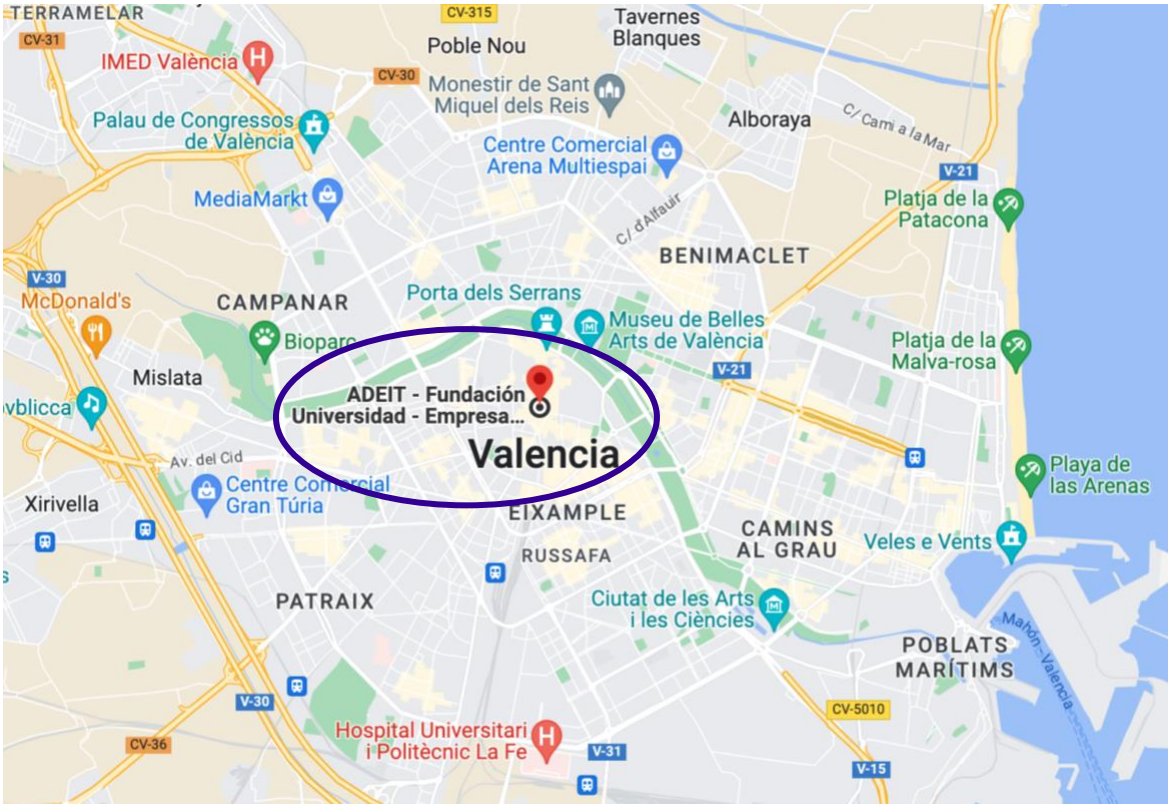
RNA virus discovery has seen an explosive increase over the past years, and there are no signs of slowing down. Yet, there is a rapidly growing schism in how this growing influx of RNA virus data should be identified, analysed, quantified, reported, and shared. The identification and analysis of viral RNA-dependent RNA polymerases (RdRps) is at the centre of this. Current divergences in methodological approaches make it difficult to evaluate each other's findings and build on each other's work. To promote reproducibility, facilitate collaboration, and establish interoperable frameworks for environmental/uncultivated RNA virology, we decided to create a discussion-centric event for computational RNA virologists around the globe. We are delighted to see this idea come to fruition and host the first ever RdRp Summit.

Venue

Dates: 22–23 May, 2023

Venue: Departamento de Congresos y Organización de Actividadaes, ADEIT – Fundación Universidad Empresa de la Universitat de Valéncia (www.adeituv.es)

Plaça Mare de Déu de la Pau, 3, 46001 Valéncia, Spain [Google maps]



Sponsors

Chan-Zuckerberg Initiative – CZI Global Voice Travel Award

We are incredibly grateful to the Chan-Zuckerberg Initiative for their generous support towards the RdRp Summit. The CZI Global Voice Travel Award is designed to increase the accessibility of the Summit to researchers from the Global South and to hear their voices in these important discussions on RdRp interoperability. The participation of computational virologists from the Global South is vital in ensuring the conference truly reflects the diversity of the scientific community, and help to promote inclusivity, equity, and fairness in the field.

CZI's contributions have allowed four researchers from the Global South to join us in-person at the RdRp Summit in Valencia, including our keynote speaker.

Frontiers in Virology

We would like to express our gratitude to Frontiers in Virology for their sponsorship of our conference. Apart from financial support that helps us realise the summit, the journal will be hosting a special issue with the proceedings of the RdRp Summit. The journal editorial has generously vouched to waive several associated publishing charges (APCs) for a few selected manuscripts. For this special issue, Professors Nakagawa, Babaian, and Roux will be acting topic editors, with other members of the organising committee as topic coordinators. We greatly appreciate the journal's commitment to advancing this growing field inside RNA virology, and we look forward to a productive partnership with them in the future.

ViBioM 2023

International Virus Bioinformatics Meeting 2023

The International Virus Bioinformatics Meeting (ViBioM) is being held in Valencia, Spain, from May 24th to 26th, 2023, also at ADEIT. It is jointly organised by the European Virus Bioinformatics Centre, the Institute of Agrochemistry and Food Technology, and the University of Valencia. We are extremely grateful to the ViBioM organising committee, who have provided crucial assistance with event logistics and advertising for the RdRp Summit.

As the RdRp Summit is a satellite event to ViBioM, there will be a welcome get together at Tyris On Tap [Google maps] for participants of both conferences. This will commence at 7:30pm on May 23rd after the conclusion of the Summit. Everyone is welcome to join!

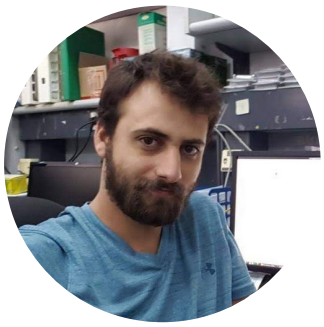
RdRp Summit Organising Committee

Justine Charon

@VRustine

National Research Institute for Agriculture, Food and Environment, France

Event Coordinator



Uri Neri

@UriNeri2

Tel Aviv University, Israel

Event Coordinator/Secretary

Artem Babaian

@RNA_Life

University of Toronto, Canada

Treasurer

So Nakagawa

@sounaka

Tokyo University School of Medicine, Japan

Logistics

Shoichi Sakaguchi

@sho1sakaguchi

Osaka Medical and Pharmaceutical University, Japan

IT and Logistics

Sabrina Sadiq

@RoyalFizzbin

The University of Sydney, Australia

Social Media Coordinator

Heather Curtis

Ohio State University, USA

Network Logistics

Simon Roux

@simroux_virus

The Joint Genome Institute, USA

Community Outreach

Guillermo Domínguez Huerta

@g_dom_huerta

Ohio State University, USA

Graphic Design

Code of Conduct

The primary goal of the RdRp Summit is to facilitate discussion, promote collaboration, and encourage networking across the field of computational virology. This meeting welcomes people from diverse cultural backgrounds and disciplines at any stage in their academic and professional career. We expect participants to be able to enjoy an inclusive and respectful environment.

Expected behaviour

1. Treat everyone with respect and kindness

   Listen attentively during presentations and save questions until after the talk. Do not interrupt or speak over others during the discussion sessions and be thoughtful before speaking.

2. Respect others' privacy.

   Seek permission before taking and sharing photographs/ recordings of participants and their work. If someone indicates they do not wish to be photographed or recorded, or have their work shared outside of the RdRp Summit, this must be immediately respected.

3. Use your true professional identity both in-person and over Zoom

4. Respect the rules and policies of the meeting venue

5. Adhere to the terms of use and policies of online platforms (eg., Twitter, Zoom)

By participating in the RdRp Summit - in-person or online - you agree to follow this code of conduct. Failure to do so will result in being removed from the event. **No refund will be provided for a breach of the Code of Conduct.**

Code of Conduct

Examples of unacceptable behaviour include but are not limited to:

- Harassment, bullying, or discrimination relating to gender, sexual orientation, disability, race, ethnicity, religion (or lack thereof), age, gender identity or expression, or physical appearance
- Any physical or verbal abuse – this includes attacking different ideas or viewpoints as opposed to professional disagreeing dialogue
- Inappropriate and/or unwanted physical contact, sexual language or imagery, or repeated requests for dates or contact information
- Stalking, deliberate intimidation, or threatening participants in-person or online
- Unwanted photography or recording
- Sustained disruption of presentations or discussions
- Breaching policies or terms of use of the venue or any online platforms
- Criminal offences – this will not only result in immediate expulsion from the RdRp Summit, local authorities will also be contacted

If you experience or witness harassment, inappropriate behaviour, or any other breach of the Code of Conduct, please reach out to a member of the organising committee immediately. Reports will be handled in complete confidence. We are here to ensure you are treated with dignity and respect at the RdRp Summit.

Timetable

Day 1 – Monday, 22nd May 2023

Zoom link: https://us02web.zoom.us/j/83166880825 Password: 854567

Session	Session topic	Presenter	Title	Start time
Session 1	Opening remarks	RdRp Summit Team	Welcome to the first ever RdRp Summit!	09:00
	Keynote presentation	Humberto Debat	Ten years of virus discovery based on high-throughput sequencing data: a perspective from the Global South	09:30
	Cookie break			10:00
	Speed-dating with the red flags of RNA virus interoperability pt. 1	Discussion - shuffling tables and breakout rooms		10:30
	Lunch			12:30
	Speed-dating with the red flags of RNA virus interoperability pt.2	Discussion - shuffling tables and breakout rooms		13:30
Cookie break				15:45
Session 2	Fundamentals of RNA virus operability	Katy Brown	Identifying potential RdRp contaminants in the NCBI protein database	16:15
		Michelle Wille	Addressing Reporting Discrepancies in Virome-Scale Studies: Minimum Requirements for Transparent Communication of Findings	16:30
		Discussion - chaired by Michelle Wille		16:45
End Day One				19:00

Day 2 - Tuesday, 23rd May 2023

Zoom link: https://us02web.zoom.us/j/86909272569 Password: 113188

Session	Session topic	Presenter	Title	Start time
Session 3	Opening remarks	RdRp Summit Team	Welcome back	09:00
	RNA virus categorization and data centralization	Yuto Chiba (**Remotely**)	How to report the divided RdRps	09:15
		Erin Harvey	The utility of RdRp phylogenetic trees in virus discovery: Host association, taxonomy and disease potential	09:30
		Discussion - chaired by Erin Harvey		09:45
Cookie break				11:15
Session 4	Methods for divergent virus discovery	Rachid Tahzima	Generic Ensembles of RdRp Modules on the Anvil of Evolution	11:30
		Marco Forgia	Playing hide and seek: looking for new structurally conserved RdRps in fungal metatranscriptomes	11:45
		Discussion - chaired by Marco Forgia		12:00
Lunch				13:30
Session 5	Consensus forming	Discussion - split into tables and breakout rooms by topic of interest		14:30
	Cookie break			16:30
	Future obstacle charting	Discussion - all hands!		17:00
	Roadmap for future work, announcements of prize laureates, send off	RdRp Summit Team	Closing remarks	19:00
End Day Two				19:10
Drinks with ViBiom participants !				19:30

Presentation abstracts

Keynote speaker: Humberto Debat

National Institute of Agricultural Technology, INTA, Argentina

Ten years of virus discovery based on high-throughput sequencing data: a perspective from the Global South

Viruses are the most prevalent biological entities on earth. Notably, conservative estimates suggest that over 99.9% of the virosphere remains elusive. Data mining of publicly available databases is a valuable tool for virus discovery and an efficient and sustainable strategy of secondary analyses. Here I will present some strategies oriented to uncover the global viral dark matter based on open sequencing data. I will summarize our experience in virus hunting from our very first 454 and sRNA libraries to our most recent Nanopore runs, and the use of NCBI resources such as TSA and SRA to identify viruses. Some examples of robust detection and characterization of novel viruses hosted by neurons of gastropods, venom glands of spiders, lantern organs of fireflies, zombie fly fungus, marihuana, frog brains, malaria-vector mosquitoes from Amazonas or kidneys of Argentinian mice will be described. The challenges in virus discovery from Latin America and potential impact in virus emergence and pandemic prediction will be discussed.

Katy Brown
University of Cambridge, UK

Identifying potential RdRp contaminants in the NCBI protein database

Regardless of the methodology used to detect RdRp, it is useful to be able to generate some measure of the proportion of false positives that will be detected, that is, sequences which are identified as RdRp but are actually from another, usually cellular, source. A common approach is to run the detection tool on a database of proteins of known non-viral origin, assuming that any significant hits are false positives. However, this approach is hampered by the presence in many databases of numerous proteins which, while they have been identified in a cellular organism, are derived from viral RdRp. This is especially common amongst proteins assigned to a particular organism but denoted as “unknown protein” or similar. There is currently no specific annotation applied to these proteins, instead the name is assigned by the individual researcher.

To quantify the extent of this problem, I have identified over 230 million proteins in the NCBI protein database which are labelled as “unknown” or similar and as an organism outside of the Orthornavirae. By screening these for similarity to viral RdRp using both pHMM and BLAST-based approaches and performing phylogenetic analysis, proteins of potential RNA viral origin can be identified. This will allow these proteins to be excluded from future false positive screens. This analysis may also uncover unknown viruses in this large and under-studied dataset.

Michelle Wille
The University of Sydney, Australia

Improving the reporting of metagenomic virome-scale data

Wei-Shan Chang[1], Erin Harvey[1], Jackie Mahar[1], Jemma L. Geoghegan[2], Cadhla Firth[3], Mang Shi[4] , Etienne Simon-Loriere[5], Edward C. Holmes[1], Michelle Wille[1,6]
[1]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia
[2]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand; Institute of Environmental Science and Research, Wellington, New Zealand.
[3]EcoHealth Alliance, New York, USA.
[4]Sun Yat-Sen University, Shenzhen campus of Sun Yat-Sen University, Shenzhen, China
[5]Evolutionary Genomics of RNA Viruses, Institut Pasteur, Université Paris Cité, Paris, France
[6]Department of Microbiology and Immunology, at the Peter Doherty Institute for Infection and Immunity. The University of Melbourne, Melbourne, Victoria, Australia.

Over the last decade, metagenomic sequencing has facilitated an ever-increasing number of virome-scale studies, in turn leading to an exponentional expansion in understanding of virus diversity. A central concern associated with this remarkable increase in the number of virome-scale studies is the lack of broadly accepted “gold standards” for reporting the data and results generated. This is of particular importance for animal virome studies as there are a multitude of nuanced approaches for both data presentation and analysis, all of which significantly impact the results. As such, the results of some published studies cannot be contextualised and are of limited utility due to reporting deficiencies. Herein, we aim to address these reporting discrepancies by outlining reporting recommendations for the presentation of virome data, encouraging a transparent communication of findings that can be interpreted in an evolutionary and ecological context.

Yuto Chiba
University of Tsukuba, Japan

How to report the divided RdRps

Yuto Chiba[1,2], Yosuke Nishimura[1]
[1]*Research Center for Bioscience and Nanoscience (CeBN), JAMSTEC, Yokosuka, Japan*
[2]*School of Agriculture, Meiji University, Kawasaki, Japan*

Divided RdRps, encoded by two ORFs on separate genomic segments, have recently been discovered (Chiba *et al.*, 2020, 2021). This finding altered the view that RdRp must be encoded by a single ORF and has revealed greater flexibility of RdRp than was expected. However, the taxonomic distribution and sequence diversity of divided RdRp in the virosphere is still poorly understood. In recent years, our knowledge of the RdRp diversity has been greatly expanded by metatranscriptomic analyses. In these analyses, the divided RdRp is detected as a fragment of RdRp, and it is difficult to identify the paired fragment. To identify the pair, we used the FLDS method, which allows the construction of multi-segmented viral genomes based on the conserved terminal sequence. The FLDS method is a promising tool for detecting many more divided RdRps in the future. At the RdRp summit, we would like to discuss the framework for reporting and sharing divided RdRps in the RNA virus community.

References

Chiba, Y. *et al.* (2020) Discovery of divided RdRp sequences and a hitherto unknown genomic complexity in fungal viruses. *Virus Evolution* **7**: veaa101.
Chiba Y. *et al.* (2021) Splitting of RNA-dependent RNA polymerase is common in *Narnaviridae*: Identification of a type II divided RdRp from deep-sea fungal isolates. *Virus Evolution* **7**: veab095.

Erin Harvey
The University of Sydney, Australia

The utility of RdRp phylogenetic trees in virus discovery: Host association, taxonomy and disease potential

Metatranscriptomics has revolutionised RNA virus discovery, taking the field from labour intensive and prone to failure to something anyone with a laptop and internet access can partake in. But this revolution has evolved into a double-edged sword. We are now able to characterise the viromes of diverse environments with essentially no sample being off limits, but the abundance of low hanging fruit has led to publication of studies which do not provide the bare minimum information for a virome study. Without providing sequence information for at least the RdRp segment of a virus and a phylogenetic tree to place this sequence with some context, a virome study is of little to know consequence to the field. As more and more low quality entries are added to databases such as NCBI, it is impossible to rely simply on the suggested taxonomy of the closest blast hit as many sequences are misassigned by the author or even NCBI themselves. Using a meta-transcriptomic faecal virome study of Australian small mammals, I illustrate the importance of phylogenetic analysis in determining taxonomy and host association of highly divergent virus species.

Rachid Tahzima
University of Liège (Gembloux Agro-Bio Tech), Belgium

Generic Ensembles of RdRp Modules on the Anvil of Evolution
Conformational Dynamics Heterogeneity and Liquid Condensed States Modulate Sequence-Ensembles Nature of RdRp Tethered Folded Domains

Rachid Tahzima[1,2], Nikolay Simankov[1,3], Annelies Haegaman[2], Kris De Jonghe[2], Sébastien Fransceschini[3], Hélène Soyeurt[3], Sébastien Massart[1]
[1]Laboratory of Plant Pathology – TERRA - Gembloux Agro-BioTech – University of Liège (ULg), 5030 Gembloux, Belgium
[2]Flanders Research Institute for Agriculture, Plant Sciences, Fisheries and Food (ILVO), 9820 Merelbeke, Belgium
[3]Laboratory of Statistics, Informatics and Modelling Applied to Bioengineering, Agrobiochem Department – TERRA - Gembloux Agro-BioTech – University of Liège (ULg), 5030 Gembloux, Belgium
Contact: rachid.tahzima@uliege.be

The affordability of high-throughput omics and the wealth of newly discovered RNA viruses is transforming how evolutionary virologists approach metagenomic research. Hence, the RdRp-based identification and functional classification of RNA viruses remains a conceptual hurdle to biological questions, whereas standardized bioinformatic approaches to support analysis of high-dimensional proteomics data are currently lacking, particularly for divergent or poorly described hosts samples. Despite the effectiveness of RdRp-based methods and while several computational tools are available to perform motif-based predictions of RdRps, much less is known about the sequence-ensembles, conformational propensities and biophysical principles underlying host-specific viral replication. In this perspective, we developed an evolutionary-informed approach to investigate generic sequence-determinants of viral replication in a wide representative set of RdRp modules via their conformational heterogeneity landscape, and captured relevant functional features, such as intrinsically disorder, liquid phase separation propensities, folding-promoting properties fostering prediction of context-dependent replication across biologically diverse and important viral lineages. Finally, this comprehensive study provides directly interpretable and fundamental insights into the vast repertoire of context-sensitive functionalities conferred by taxon-specific conserved RdRp motifs while facilitating assignation of novel or ill-characterized RdRp-sequences into pre-classified or newly defined Riboviria taxa together with the robust prediction of their biological affinities.

Marco Forgia
Institute for Sustainable Plant Protection (IPSP), CNR, Italy

Playing hide and seek: looking for new structurally conserved RdRps in fungal metatranscriptomes

Marco Forgia[1] , Stefania Daghino[1] , Massimo Turina[1,2]
[1]*Institute for Sustainable Plant Protection, National Research Council of Italy; Torino, Italy*
[2]*Institute for Sustainable Plant Protection, National Research Council of Italy; Brescia, Italy*

Fungal-derived metatranscriptomes have been an outstanding source of viral diversity since the advent of high throughput sequencing methodologies. Indeed, our recent studies allowed the characterization of several “hidden” RdRps from these sources as the one hosted by splipamiviruses (the first example of splitpalm-domains on two different proteins), by ormycoviruses (characterized by high diversity in the conserved motifs in the palm-domain), by ambiviruses, and finally by the fist example of a permutated RdRp for mycoviruses infecting oomycetes. All of these discoveries would not be possible with simple blast similarity approaches. Our approach relies instead on the identification of the contigs encoding for protein without homology to the nr database. The Illumina cDNA libraries are built with a stranded approach, allowing us to assess quantitatively reads mapping on both strands of the sequences of interest (indicating the presence of dsRNA typical of viral replication). Lastly, the putative proteins from the contigs studied are investigated through protein modelling and folded structure comparison. The availability of data from similar sources allows the search for homologs of the putative viruses identified, creating new clusters of ORFans and strengthening the hypothesis of the viral nature of the sequences.

CZI Global Voice Travel Award Recipients

We are incredibly grateful to the Chan-Zuckerberg Initiative for their generous support towards the RdRp Summit. The CZI Global Voice Travel Award is designed to increase the accessibility of the Summit to researchers from the Global South and to hear their voices in these important discussions on RdRp interoperability. The participation of computational virologists from the Global South is vital in ensuring the conference truly reflects the diversity of the scientific community, and help to promote inclusivity, equity, and fairness in the field.

CZI's contributions have allowed four researchers from the Global South to join us in-person at the RdRp Summit in Valencia, including our keynote speaker.

Chan Zuckerberg Initiative

Humberto Debat

@humbertodebat · Twitter

Humberto Debat was born, raised and educated in Córdoba, Argentina. Trained in biology and virology from the National University of Córdoba, and a background in science management from Ajou University, Humberto is a tenured research scientist at the Institute of Plant Pathology in the Center of Agronomic Research of the National Institute of Agricultural Technology (CIAP-INTA) in Argentina. Humberto studies the interface of viruses and crops from a systems biology perspective, is interested in novel approaches to reduce losses associated to plant diseases and passionate about understanding an expanding global virosphere. Humberto has been a member of the Board of Directors of UFyMA-Institute and technical manager of CIAP-INTA for the National System of Genomic Data. During the COVID-19 pandemic, he has participated in the Argentine SARS-CoV-2 Genomics Project. Humberto is a member of the Advisory Committee on Open and Citizen Science of the Argentine Ministry of Science. Humberto has been a fellow and member of ASAPbio and ambassador of the eLife community. Humberto is an affiliate of the bioRxiv preprint server and co-creator of PanLingua, a multilingual preprint discovery and reading tool in the biomedical sciences. Humberto has published more than 60 peer-reviewed scientific articles, dozens of preprints, served diverse editorial roles in several academic journals, has given multiple talks and presentations at conferences, teaches undergraduate and graduate courses, and has given hundreds of interviews in general media.

Ricardo Rivero

@RicardoRH_IDB · Twitter

Ricardo Rivero is a researcher from Colombia and is currently a Ph.D. student in Stephanie Seifert's lab at Washington State University. He earned his B.Sc in Biotechnology from Institución Universitaria Colegio Mayor de Antioquia (Medellín, Colombia). He has 6 years of experience as a research assistant at Universidad de Córdoba where he worked in the development of molecular, serological, and computational approaches for the study of zoonotic viruses in bats, mosquitoes, and rodents. His Ph.D. research is focused on predicting the cross-species potential of novel viruses from bats and mosquitoes from -omics data

Fhilmar Raj Jayaraj Mallika

fhilmarraj (F.R) · GitHub

@plantvirome · Twitter

Fhilmar Raj Jayaraj Mallika is a research fellow for the virus-indexing laboratory at ICAR - National Research Center for Banana in Tiruchirappalli, India. His research activities are focused on advanced virus diagnosis in fruit crops for the assessment and implementation of the High throughput sequencing (HTS) pipeline(s) into the workflow of virus detection protocols. His interest in RNA virology started during MSc research while studying RNA viruses infecting raspberries in Finland under the guidance of Prof. Jari P.T. Valkonen. His thesis explained the genetic diversity of Black raspberry necrosis virus, a low-titer virus, which possesses a remarkable Alkylation B homolog in the RNA-dependent RNA polymerase coding region. Moreover, the study included small RNA (sRNA) data analysis and endpoint PCR for the validation of virus strain types using bioinformatics tools. At ICAR-NRCB, he is working on developing a genome surveillance method using MinION-generated long-read data to locate the scaffolds of Banana streak virus, a pararetrovirus, which is found as integrated into the B genomes of Musa sp. He is interested in topics such as virus ecology, phylogeography, and the taxonomical classification of viruses.

His educational experience spans a period of 12 years in the discipline of agriculture with a bachelor's and a master's degree. He is an Erasmus Mundus Scholar and an alumnus of the Indo-EU BRAVE (BReeding for plAnt resistance to VirusEs) project. He is a recipient of the Young Minds Award from the International Society of Horticultural Sciences (ISHS). He boasts that plant viruses are good examples for introducing the subject of virology to students with onsite examples.

San Emmanuel James

Dr. San is a bioinformatician, molecular biologist and data scientist with specialist expertise in genomics, phylogenetics, statistical genetics and software development. He is currently a postdoctoral researcher at the University of KwaZulu Natal, Durban, South Africa at the KwaZulu Natal Research and Innovation Sequencing Platform (KRISP). He holds a PhD in Medicine and research experience in infectious diseases with a strong track record of combining molecular epidemiology with statistical analysis of pathogen genomic variation and the development of bioinformatics software. He is a member of several professional research bodies and collaborations including the Network for Genomic Surveillance in South Africa (NGS-SA), The Public Health Alliance for Genomic for Genomic Epidemiology (PHA4GE), Africa COVID-19 Genomics consortium, Climate Amplified Diseases and Epidemics (CLIMADE) consortium, Africa Centers for Disease Control and Prevention (CDC) Pathogen Genome Initiative (PGI) Vaccine Preventable Diseases Focus Group (VPD FG), Phylogenetic Analysis for Generalized Epidemics in Africa (PANGEA_HIV) consortium and a fellow of the H3Africa Bioinformatics Network (H3ABionet). Dr. San has published over 40 manuscripts in high-ranking international journals (including *Nature, Science, The Lancet, Lancet Infectious Diseases)*. His work has been cited over 7,000 times. He has an H-index of over 30. His current research interests include the use of phylogenetic models to understand the drivers of infection transmission of SARS-CoV-2, HIV, TB, Arboviruses and vaccine preventable diseases in Africa. He is also interested in the microbiome, transcriptomics and metagenomics to characterize niche viral and bacterial communities and to investigate their impact on host responses and response to treatment.

Info slides and video briefs

There will be very limited time allotted to speakers during the RdRp Summit as we want to maximise the time for discussion and collaboration. As such, we strongly encouraged all participants to submit a data slide or video brief, regardless of abstract submission. This was to help attendees get acquainted with and educated about the key topics ahead of time, as well as form the basis for the discussion sessions.

Submitted data slides and video briefs are available to peruse on our website at rdrp.io/info-slides until the summit. Data slides submitted up until May 15th are also included here in the program booklet below; however, slides will be accepted all the way up until the conference commences on May 23rd. Many thanks to all who contributed!

VIRUS DISCOVERY BASED ON DATA MINING IS CRUCIAL TO IMPROVE VIRUS TAXONOMY

Nicolas Bejerman- Humberto Debat
UFyMA-INTA-CONICET
@nicobejerman
@humbertodebat

Novel resources, such as Serratus, are crucial to speed up the process of virus hunting illustrating the key importance of open-science practices to gain knowledge on the unexplored virosphere.

I will present our analysis of publically available data on neglected plant virus families, as well as on underexplored hosts, such as gymnosperms, mosses, liverworts and ferns, in order to fill gaps in the global virus "dark matter" linked to plants.

Unlocking the Hidden Genetic Diversity of Varicosaviruses, the Neglected Plant Rhabdoviruses

This data mining effort resulted in two taxonomy proposals recently sent to the *Rhabdoviridae* ICTV study group to create two novel genera to accommodate monosegmented viruses linked to gymnosperms

Two novel flavi-like viruses shed light on the plant infecting koshoviruses

This data mining effort resulted in a taxonomy proposals recently sent to the *Flaviviridae* ICTV study group to create a novel genus to accommodate plant-associated flavi-like viruses

CONCLUSION

The increased number of virus sequences deposited in publically available databases will aid to develop new tools to ease the discovery of distant viruses, but virus discovery of highly divergent viruses through the palmID tool needs to be improved to use as query highly divergent viruses, such as Snake River alfalfa virus

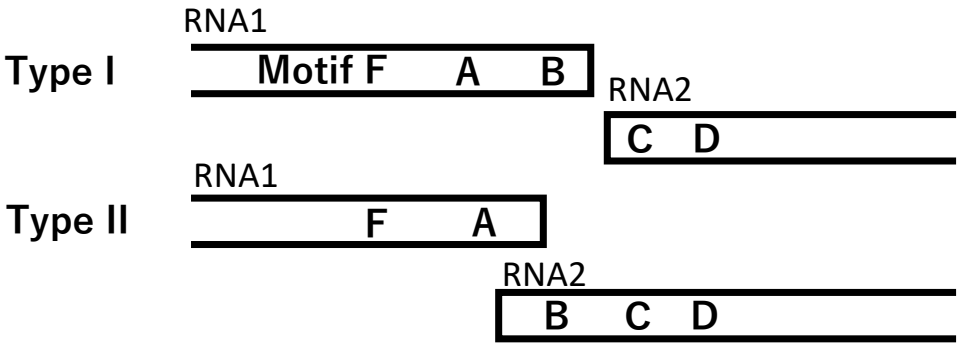
Bejerman, Diztgen and Debat (2022). doi: 10.3390/pathogens 11101127

Debat, Bejerman (2022). bioRxiv https://doi.org/10.1101/2023.02.06.527325

How to report the divided RdRps

Yuto Chiba[1,2], Yosuke Nishimura[1]
[1] Research Center for Bioscience and Nanoscience (CeBN), JAMSTEC, [2]School of Agriculture, Meiji University

① Divided RdRps



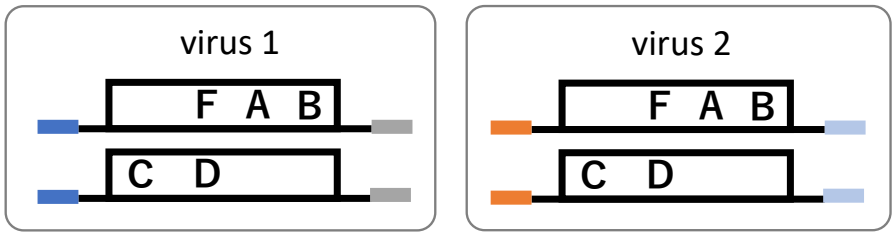
Chiba et al, 2020, 2021

- Divided RdRp is encoded in two ORFs on separate genomic segments.
- Reported divided RdRps are limited to fungal narnaviruses.

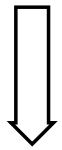
② Inclusion of potential divided RdRps in RNA virome studies

Paper	Neri et al. 2022	Zayed et al. 2022	Chen et al. 2022
source	RdRp contigs	RdRP contigs	PRJNA716119
initial contigs	2,658,344	44779	6546
Narnaviridae contigs	57049	6478	266
divided RdRps (RNA1)	**1796**	**248**	**0**

③ Difficulty in identification using metatranscriptome



Each virus has conserved terminal sequences.



Metatranscriptomic assembly

The termini are frequently lost.

If terminal sequence information is missing, we could not identify the pairs of divided RdRps.

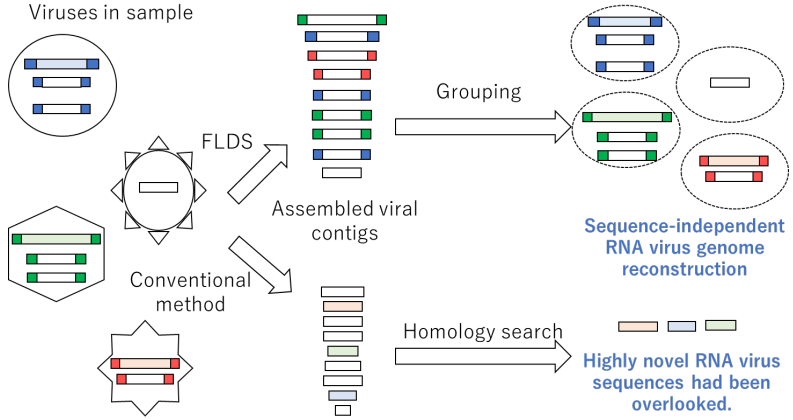
We would like to discuss the framework for reporting and sharing divided RdRps in the RNA virus community.

RdRp Summit
urayama.shunichi.gn@u.tsukuba.ac.jp

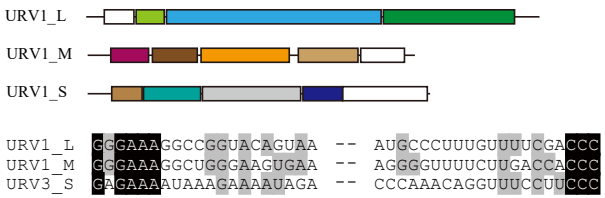
Truly “novel” RdRp identified from potential complete RNA virus genomes obtained by using FLDS.

Fragmented and primer-Ligated DsRNA Sequencing

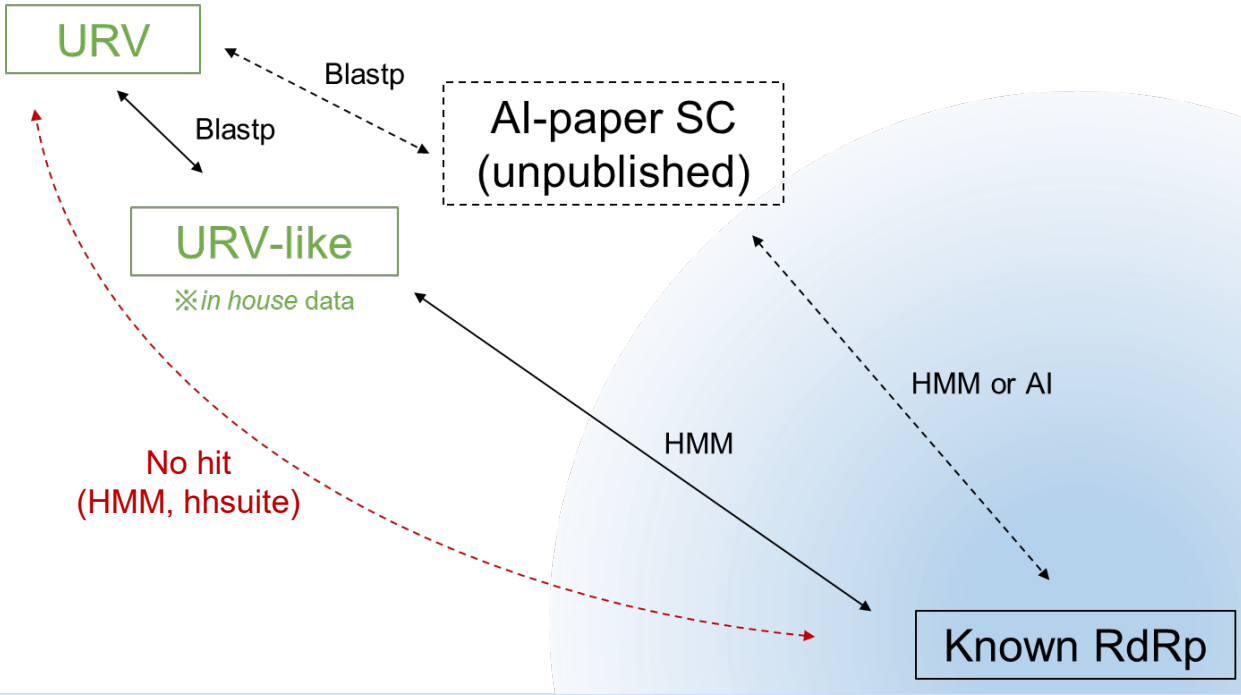
1. FLDS enables us to obtain complete dsRNA sequences, and subsequently to construct RNA virus genomes.



2. Manual identification of RdRp sequences from RNA virus-like genomes that are overlooked by conventional informatics.



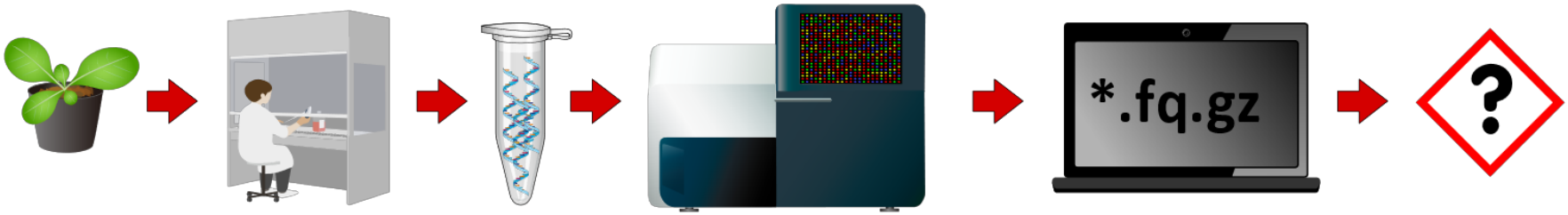
3. The RdRp in URV genome was manually identified by two step homology search.



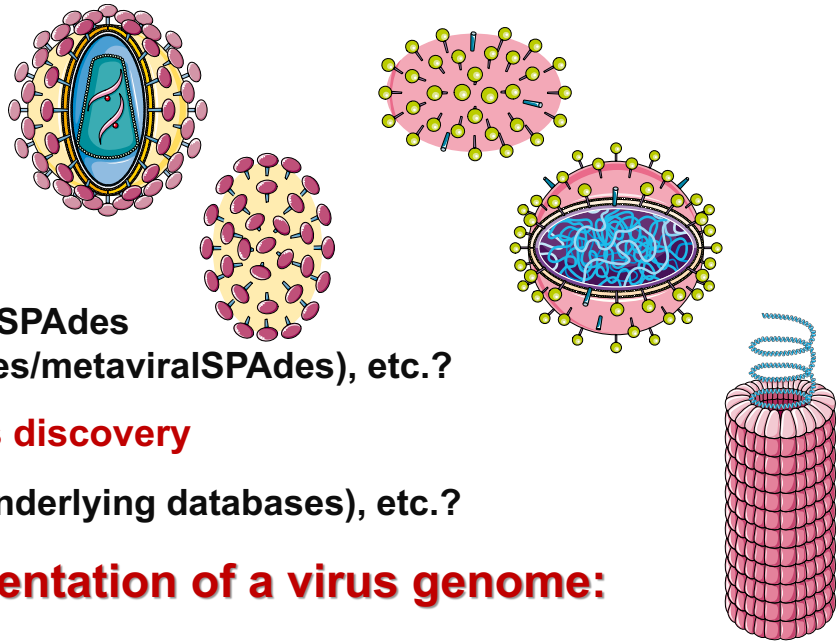
Conclusion: There are un-detectable RdRp using conventional methods.
FLDS can overcome the technical limitations.

Lack of standardized practices for metatranscriptomic virus discovery

Nikita Zrelovs (nikita.zrelovs@biomed.lu.lv)
Twitter: @NZrelovs
Latvian Biomedical Research and Study Centre (LV BMC)



Different practices = differing results:



- **Raw data preprocessing practices**

  e.g., what, how, to which extent to trim, etc.?

- ***de novo* assembly tool selection**

  e.g., Trinity, SOAPdenovo-trans, TransABYSS, SPAdes (just SPAdes or rnaSPAdes/rnaviralSPAdes/metaviralSPAdes), etc.?

- **ORF prediction and functional annotation, virus discovery**

  e.g., Cenote-Taker2, Virsorter2, Prokka (what underlying databases), etc.?

RdRp-containing contig =/= faithful representation of a virus genome:

- **Monopartite genome completeness**

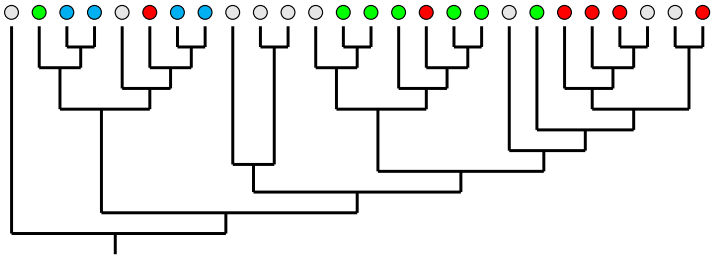
  e.g., are monopartite viral «genomes» (contigs) really complete without RACE? (Rapid Amplification of cDNA Ends)?

- **Multipartite genomes**

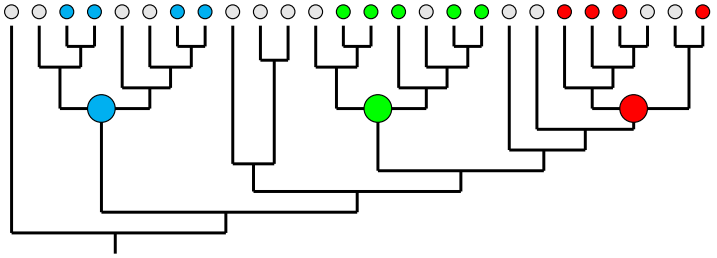
  e.g., what other non-RdRp-containing contigs comprise the genome of this virus; how to identify which other segments belong to which RdRp contig?

Semi-automated taxonomy assignment

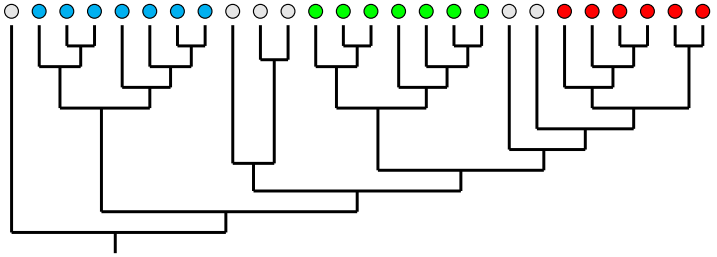
Yuri Wolf, NCBI
wolf@ncbi.nlm.nih.gov



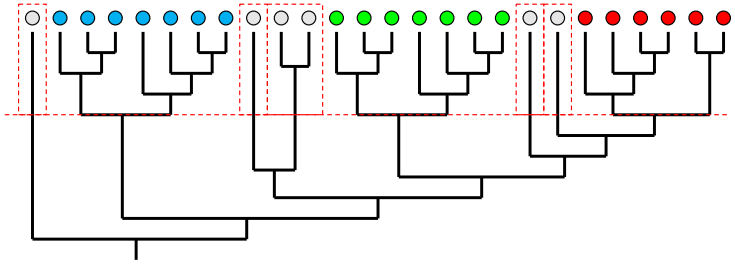
Map existing taxonomic labels of the given rank on the reference tree.



Identify the principal nodes for each taxon using the precision- and recall-like criteria (ignoring unlabeled leaves); identify “intruder” and “outlier” conflicts; remove taxonomic labels from leaves in conflict.



Assign all unlabeled descendants of the principal nodes to the corresponding taxon.



Identify the local characteristic depth of the given rank using the neighboring principal nodes; dissect the tree at the indicated depth; assign new taxa to unlabeled subtrees.

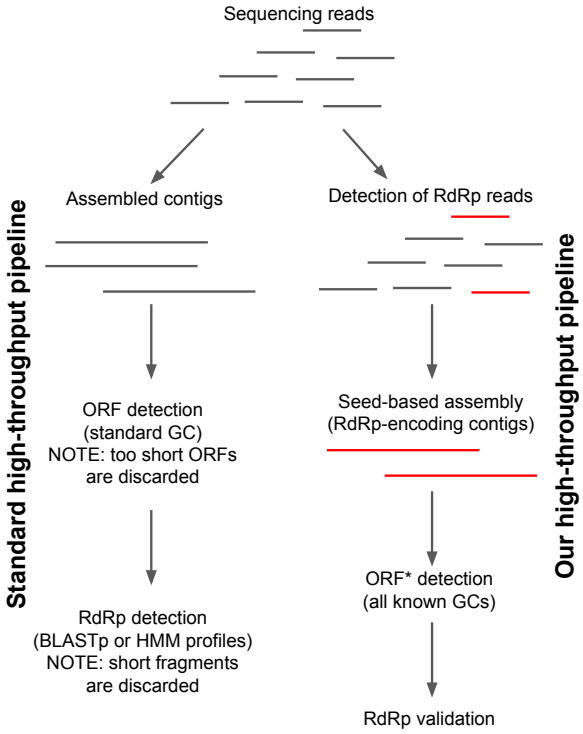
RdRp Summit

rdrp.summit@gmail.com
@RdRp_Summit

Non-standard genetic code usage needs to be considered when detecting the RdRp

Liubov Chuprikova[1,2], Li Chuin Chong[3], Chris Lauber[3], Stefan Seitz[1,2]

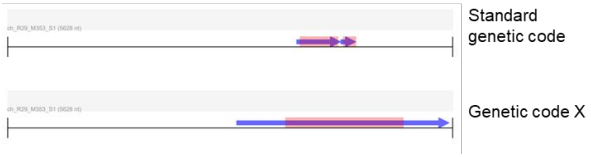
dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer



*In the current study: ORF – a coding region between two stop codons

Finding 1:

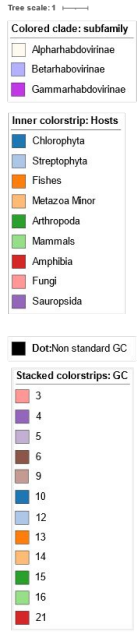
- in some cases, we were able to "repair" RdRp when non-standard GC was used:

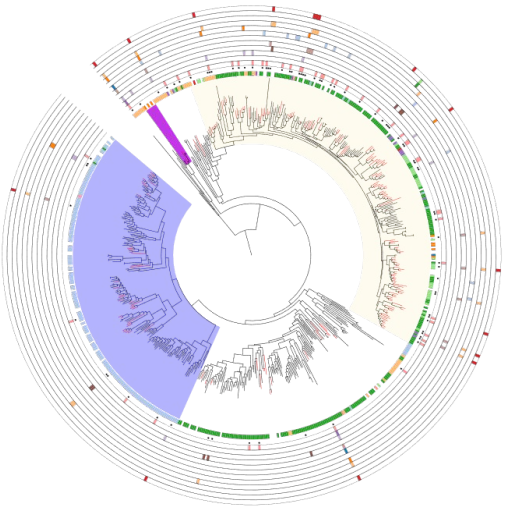


Finding 2: in the case of RefSeq genomes,

- we had the same proportion of longer RdRp(s), when using non-standard GC, as in the case of novel viruses;
- the number of "repaired" RdRp(s) was ten-folds lower, suggesting of inability to detect them with standard pipelines

***Negarnaviricota* sequences**	# total	# longer RdRp (by at least 10 a.a.)	# "repaired"
novel	2809	828 (~29%)	82 (~3%)
RefSeq	691	184 (~27%)	2 (~0.3%)





Finding 3:

- **The association of certain GC with certain hosts is not clear,** which can be due to the overlap in non-standard codons over different non-standard GCs.

Conclusions:

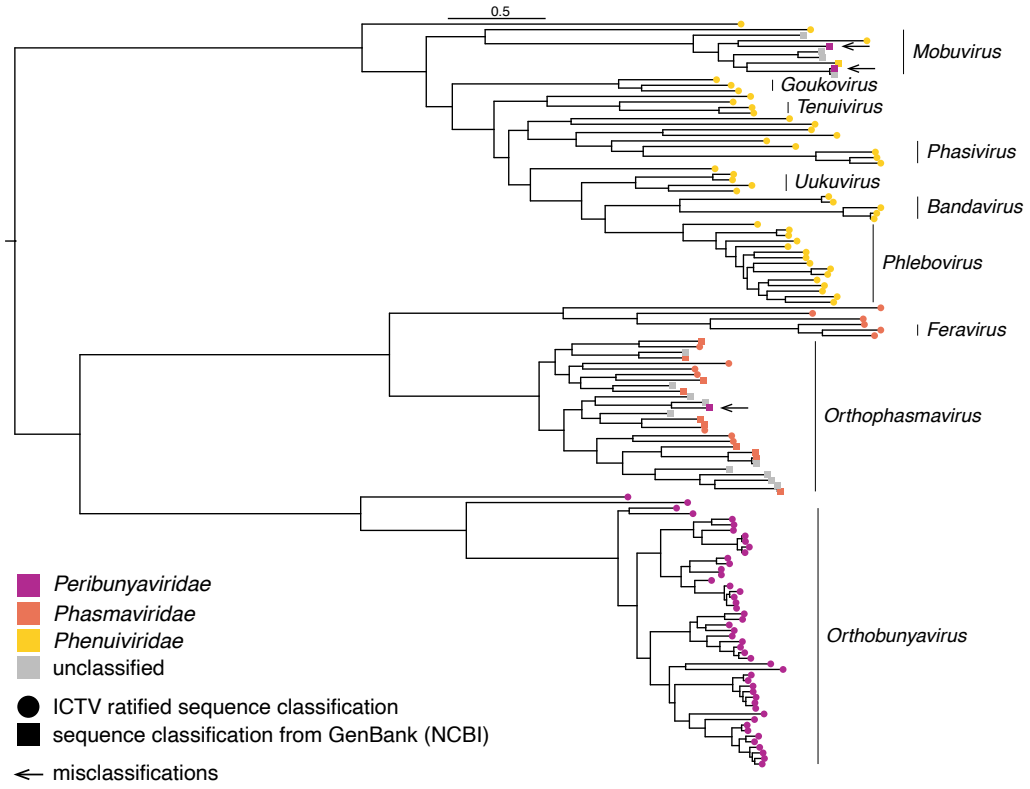
- **Pipelines need to include step to consider all known GCs.**
- **Possible alternative – GC-agnostic protein detection.**

[1] Division of Virus-Associated Carcinogenesis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[2] Department of Infectious Diseases, Molecular Virology, University of Heidelberg, 69120 Heidelberg, Germany
[3] Institute of Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, a Joint Venture between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), 30625 Hannover, Germany

* This project is supported by Helmholtz Corona Virus Pathogen (CoViPa) consortium
COVIPA HELMHOLTZ Corona Virus Pathogenesis

Databases often contain misclassifications for (RNA) viruses sequenced from metagenomic studies

Phylogenetic tree based on the RdRP of the Bunyavirales order as example.

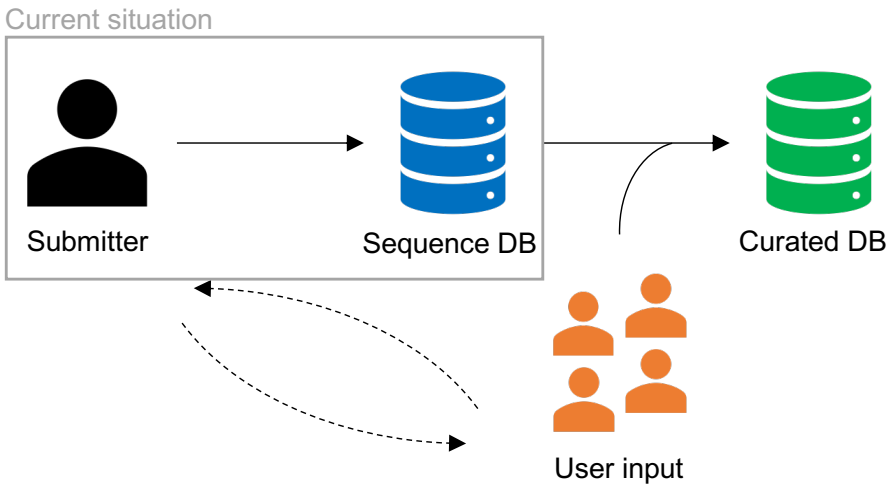


Among other things, the fast changing rate of virus classifications and the discovery of new RNA viruses leave older entries in databases (eg. GenBank) with wrong/outdated classifications entered by the sequence submitters. These entries could lead to problems when using these databases for your analyses and currently it's cumbersome to edit them.

Suggestions:

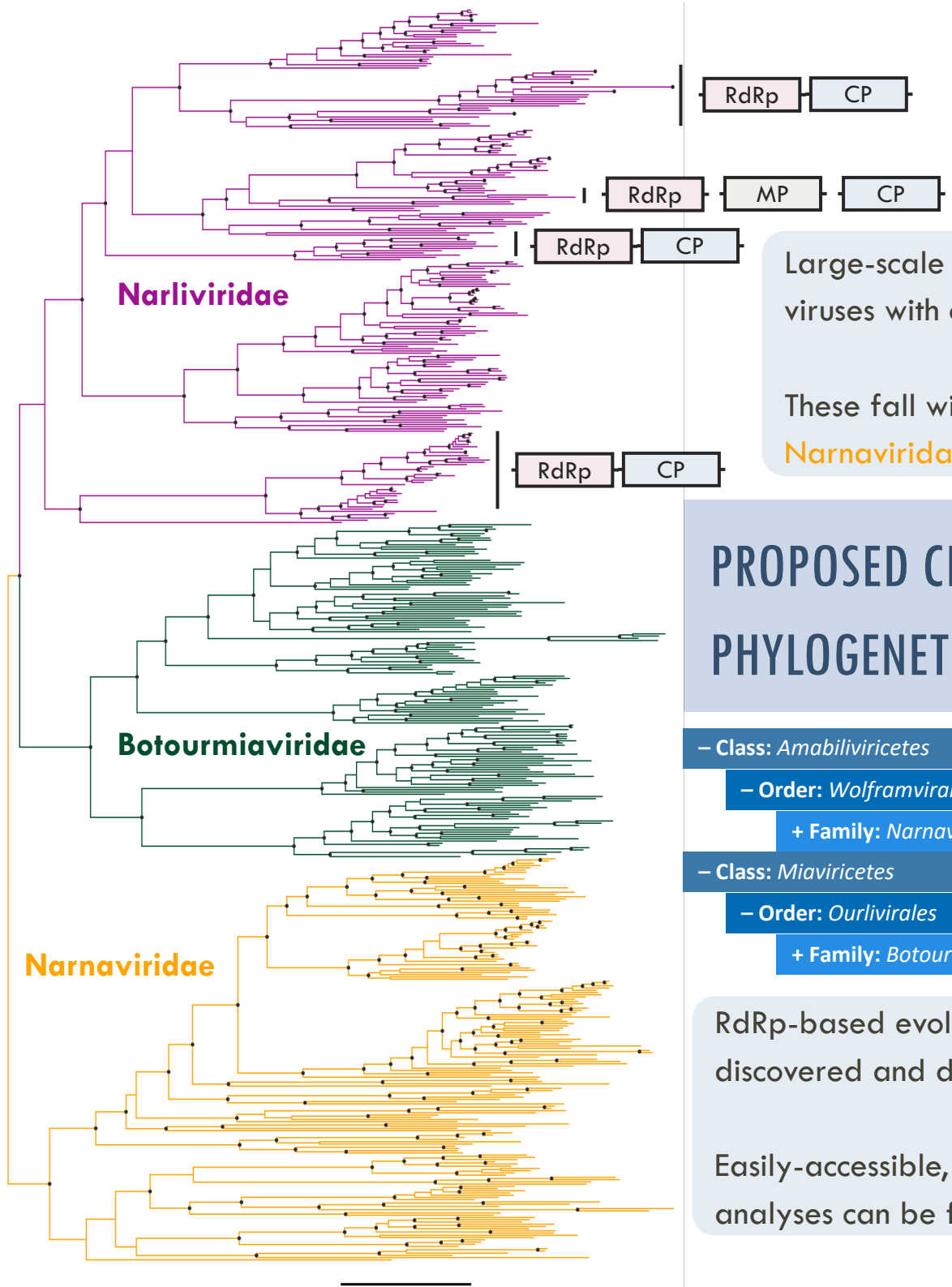
1. New guidelines for RNA virus classification should be integrated and enforced during sequence submission in databases.

2. It should be easier to edit (old) entries in GenBank (or other databases) by its users. As a suggestion, this could follow a similar approach as Wikipedia, where anyone can make well-motivated changes that will be peer-reviewed by the sequence submitter(s) and other users.



Lander De Coninck

SABRINA SADIQ
THE UNIVERSITY OF SYDNEY
Twitter • @RoyalFizzbin



Large-scale meta-transcriptomics projects have identified "narna-like" viruses with capsid genes

These fall within newly classified Narliviridae, which is distinct from Narnaviridae and a sister clade to Botourmiaviridae

PROPOSED CHANGES TO TAXONOMY BASED ON PHYLOGENETIC ANALYSES OF THE *LENARVIRICOTA*

– **Class:** *Amabiliviricetes*
  – **Order:** *Wolframvirales*
    + **Family:** *Narnaviridae*
– **Class:** *Miaviricetes*
  – **Order:** *Ourlivirales*
    + **Family:** *Botourmiaviridae*

→

– **Class:** *Amabiliviricetes*
  – **Order:** *Wolframvirales*
    + **Family:** *Narnaviridae*
  – **Order:** *Ourlivirales*
    + **Family:** *Botourmiaviridae*
    + **Family:** *Narliviridae*

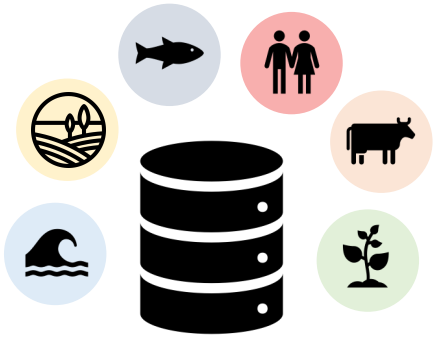
RdRp-based evolution is constantly fluctuating as new RNA viruses are being discovered and deposited at increasingly rapid rates

Easily-accessible, well-classified RdRp sequences crucial so evolutionary analyses can be frequently updated and remain relevant for longer

Sadiq *et al.*, 2022 doi.org/10.1093/ve/veac055

RdRp Summit

The importance of centralizing the viral RdRp big data



Metatranscriptome data sets

- SRA/ENA
- Mgnify
- Zenodo
- RVMT
- ftp://
- etc…

Sequence repositories

Chen et al 2022

NCBI

Zayed et al 2022

Neri et al 2022

zenodo

PALMdb

Edgar et al 2021

ftp://

Wolf et al 2020

Hou 2023

Contig comparison to RdRp sequences, HMMs and structures

Profile repositories

GitHub

Sakaguchi et al 2022

Charon et al 2022

Olendraite et al 2023

Structure repositories

AlphaFold Protein Structure Database

ESM Metagenomic atlas

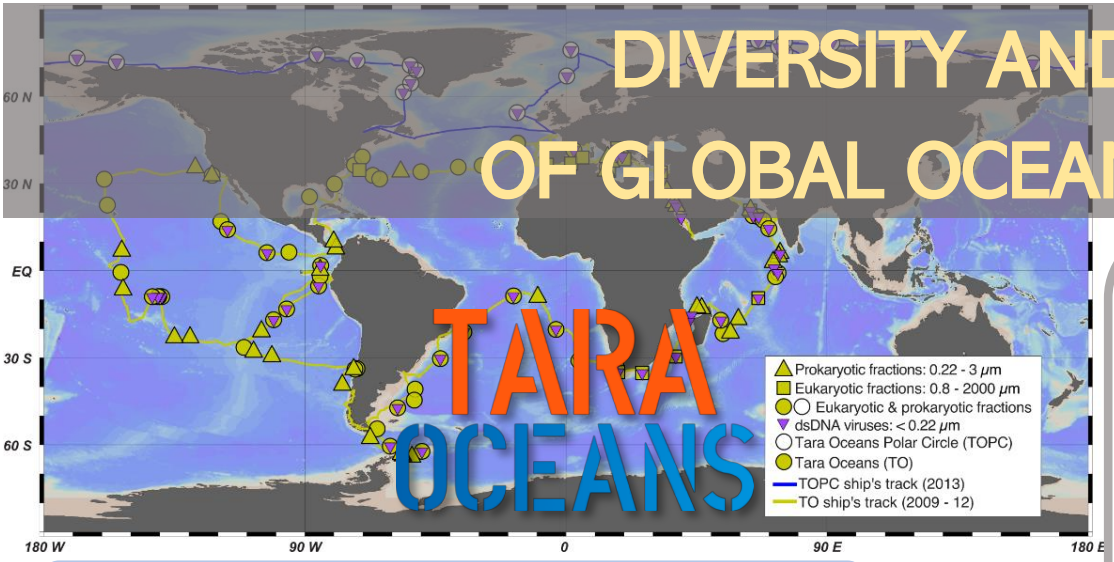
RNA virus discovery

➔ Important need to **centralize the flow of RdRp sequences, HMM-profiles and 3D-structures** from viral metagenomic studies into one single repository

Justine Charon – Post-Doc / INRAE France

DIVERSITY AND ECOLOGY OF GLOBAL OCEAN RNA VIRUSES

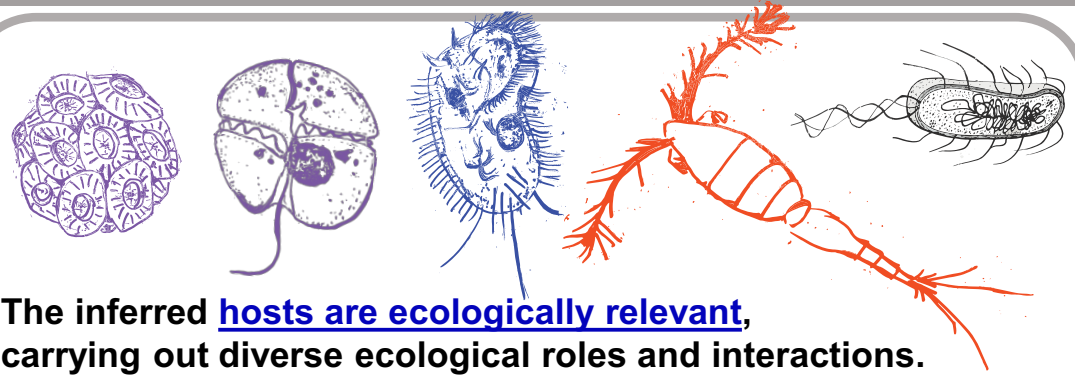
Guillermo Domínguez-Huerta
The Ohio State University
Twitter: @g_dom_huerta



771 metaTs, ~28 Terabases

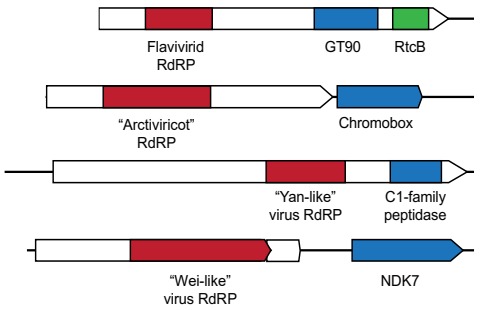
RdRps

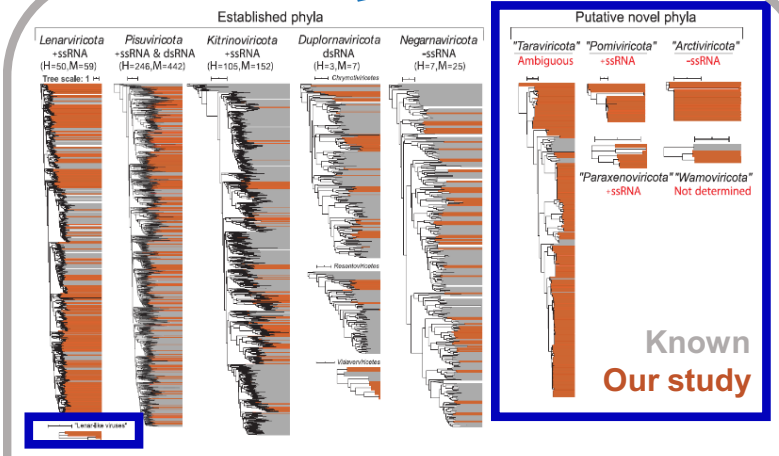
vOTUs



The inferred hosts are ecologically relevant, carrying out diverse ecological roles and interactions. They are involved in processes of the biological carbon pump.

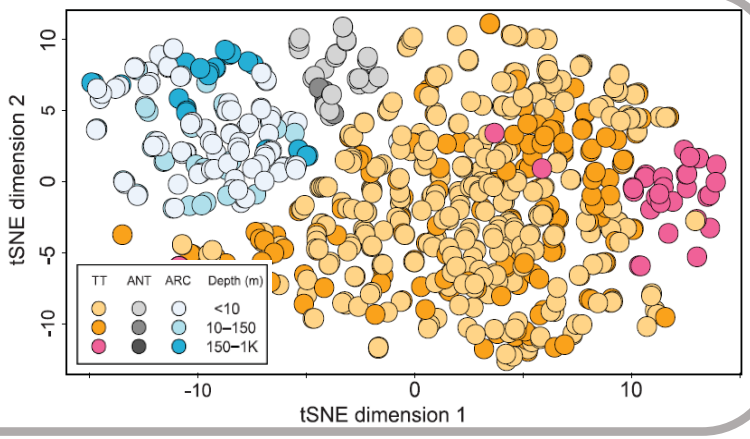
Functional annotation of RNA viral genomes revealed auxiliary metabolic genes, putatively acquired from host mRNA and hypothesized to reprogram the host metabolism to maximize the viral cycle.





Putative expansion of the established RNA viral megataxonomy, from five to ten phyla.

First report on ecological patterns of RNA viruses across Global Oceans. Meta-community analysis revealed four ecological zones for RNA viruses, paralleling those from cosampled prokaryotic dsDNA viruses and prokaryotic plankton.

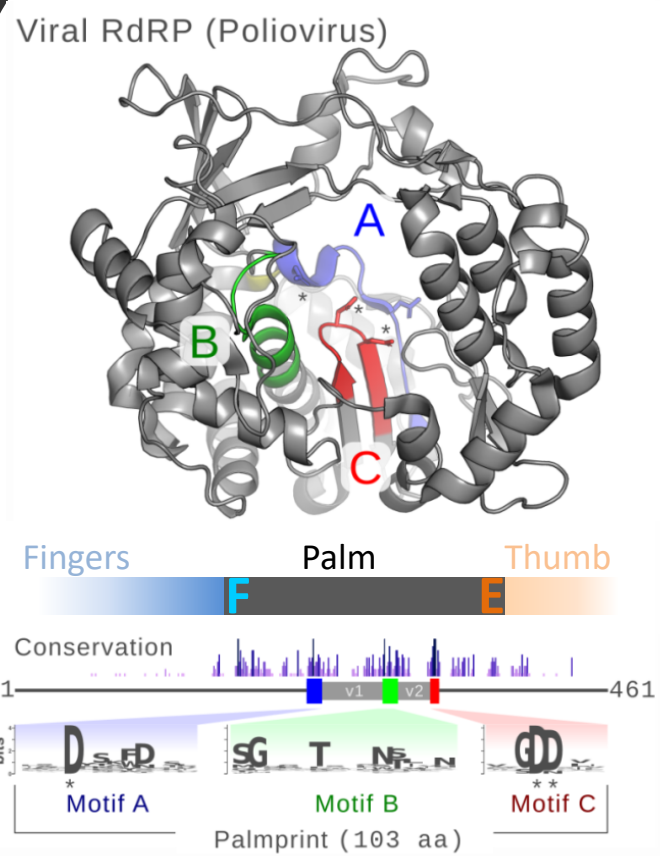


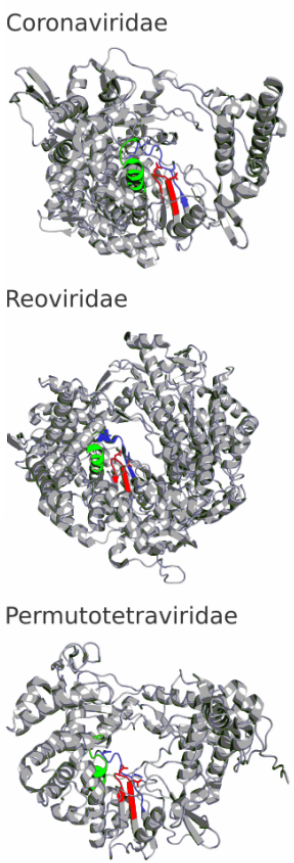
Zayed, Wainaina & Domínguez-Huerta *et al*. (2022) *Science* 10.1126/science.abm5847

Domínguez-Huerta & Zayed *et al*. (2022) *Science* 10.1126/science.abn6358

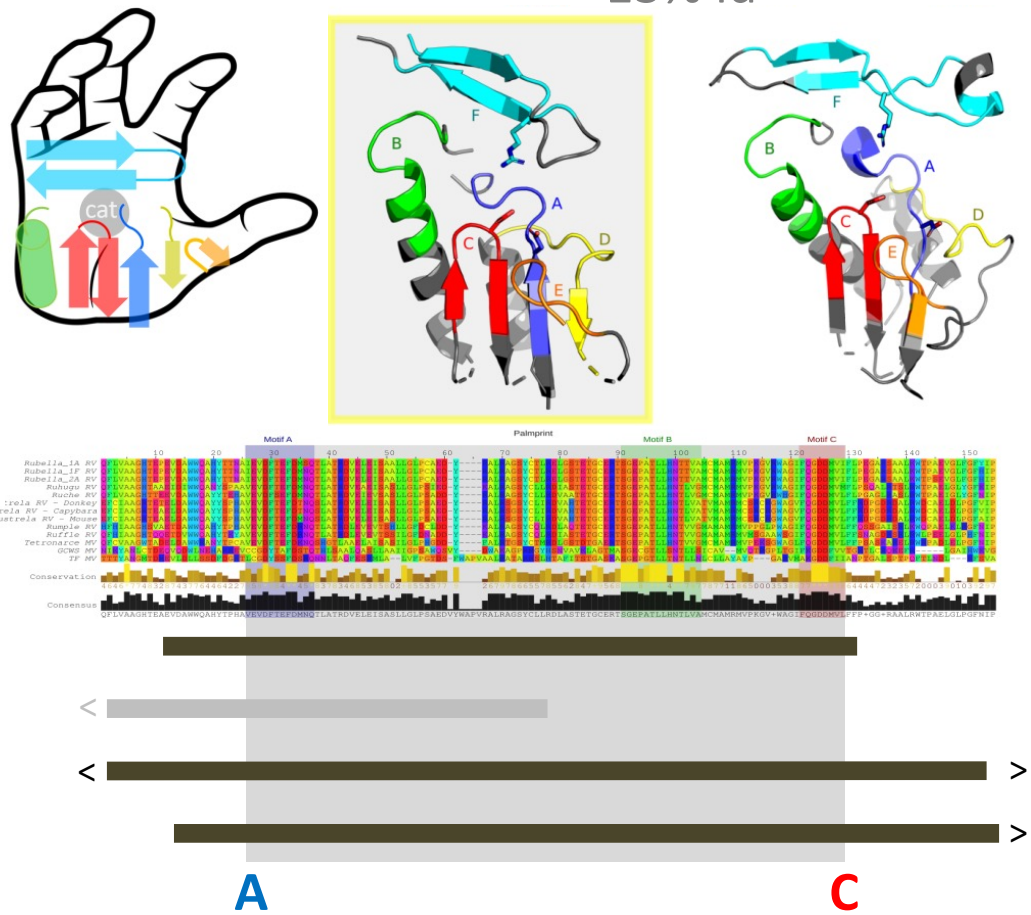
Structural motifs are our "Anchors" to define OTU

A. Babaian





13% id

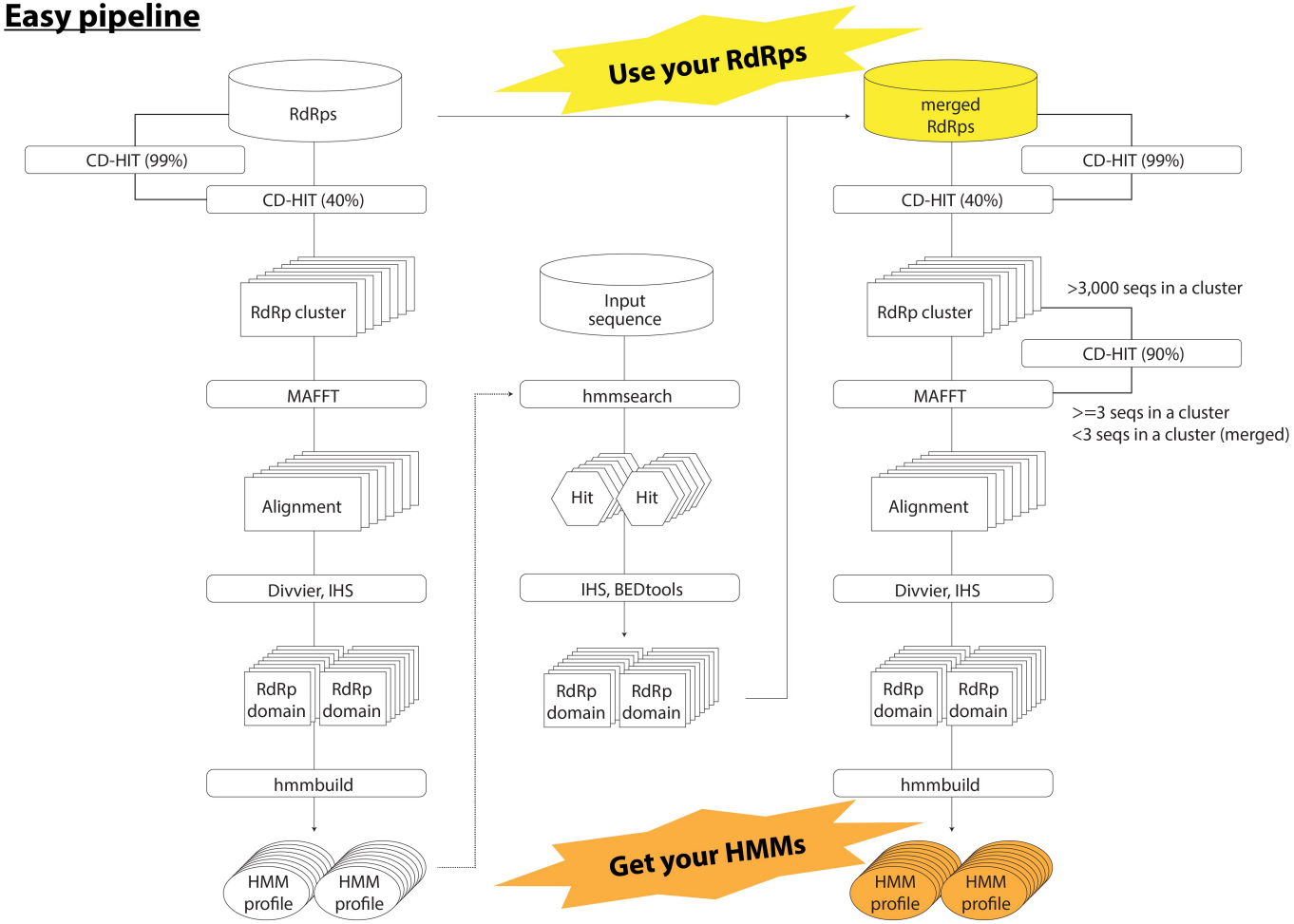


Enforces a Global Alignment %

NeoRdRp - Create and Update Your HMMs for RdRp search!

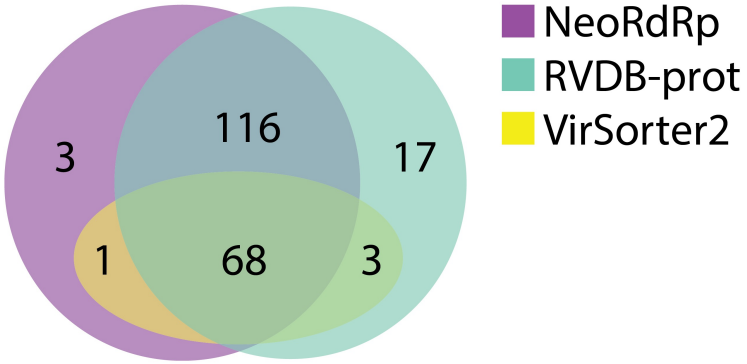
Shoichi Sakaguchi Osaka Medical and Pharmaceutical University
So Nakagawa Tokai University

Easy pipeline



Contains important motifs

Good performance



https://github.com/shoichisakaguchi/NeoRdRp sho1sakaguchi sounaka