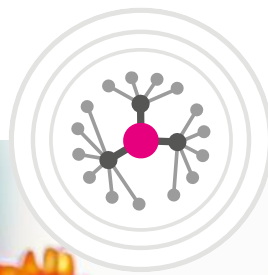


# DATA ANALYSIS FOR THE COVID-19 RESEARCH

Contributions of the German Network  
for Bioinformatics Infrastructure



## FOREWORD

# Dear Reader,

This brochure contains articles describing de.NBI-related COVID-19 projects, which are carried out by de.NBI members using de.NBI resources. In total, there are 15 projects showing how various de.NBI resources can be used in the analysis of large data sets. This booklet demonstrates that without delay the existing diverse bioinformatics infrastructure of the de.NBI network has been employed to address new research questions.

In addition to the extensive COVID-19 research section, the brochure also contains basic information on the de.NBI network. It describes the location and thematic orientation of

This COVID-19 brochure is strongly research-oriented and is intended to introduce our colleagues in Europe to the de.NBI related COVID-19 activities in Germany. An interview has also been included, which gives the coordinator of the de.NBI network and the Head of Node of ELIXIR Germany the opportunity to report on research activities in the COVID-19 field in Germany and Europe.



Prof. Dr. Andreas Tauch

the eight service centres. Furthermore, the basic tasks of the network, namely service and training, are delineated. Finally, a presentation of the federated de.NBI Cloud and the newly founded de.NBI Industrial Forum is given. Closely connected to the de.NBI network is the German node of the European ELIXIR network. The participation of the German ELIXIR node in COVID-19 research questions is also part of the booklet.



Prof. Dr. Alfred Pühler

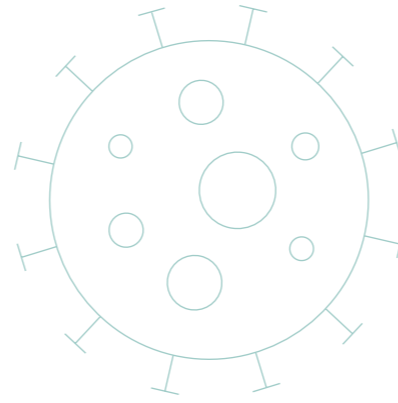
The authors of this brochure now hope that the COVID-19 research linked to the de.NBI network will receive great response. We wish all readers an interesting reading.

Andreas Tauch  
Head of Node of ELIXIR Germany

Alfred Pühler  
de.NBI Coordinator

# CONTENT

FOREWORD	3
CONTENT	4



THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE (de.NBI)	6
de.NBI TRAINING FOR LIFE SCIENTISTS	8
de.NBI CLOUD – COMPUTE POWER FOR YOUR PROJECT	10



<b>BEYOND THE VIRUS SURFACE – MULTI-OMICS APPROACHES LEAD TO INSIGHTS INTO SARS-CoV-2 AND COVID-19</b>	12
--	----

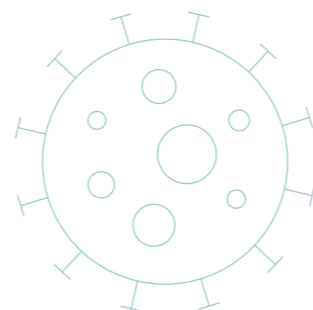
MANAGING OMICS DATA IN THE CONTEXT OF THE COVID-19 PANDEMIC	14
---	----

SARS-CoV-2 GENOMES FOR EPIDEMIOLOGICAL SURVEILLANCE – MUTATIONAL SIGNATURES ARE HIGHLY INFORMATIVE FOR TRACING INFECTION CHAINS AND VIRUS EVOLUTION	18
---	----

RNA BIOINFORMATICS TO ANALYZE SARS-CoV-2 – THE CAUSATIVE AGENT OF COVID-19	24
--	----

IDENTIFYING MOTIF MIMICRY IN SARS-CoV-2 HOST PATHOGEN INTERACTIONS	30
--	----

FRIEND OR FOE? SARS-CoV-2 HIJACKS HOST BIOLOGY TO LETHAL EFFECT	34
---	----



<b>BIOINFORMATICS TOOLS FOR ANALYZING COVID-19 DATA</b>	40
---	----

ANALYZING SARS-CoV-2 CHROMATOGRAMS AND MUTATIONS USING THE GENOME ANALYSIS SERVER GEAR	42
--	----

COMPARATIVE GENOME AND METAGENOME ANALYSIS OF CORONAVIRUS-POSITIVE CLINICAL SAMPLES	48
---	----

VIRUS-INDUCED LUNG INJURY: PATHOBIOLOGY AND NOVEL THERAPEUTIC STRATEGIES	54
--	----

OPEN DATA, SOFTWARE AND ANALYTICS AS A RESPONSE TO EMERGING PATHOGEN THREATS	60
--	----

BioInfra.Prot SUPPORTS PUBLICATION OF PROTEOMICS DATASETS AS WELL AS STUDY DESIGN AND DATA ANALYSIS FOR COVID-19 RESEARCH PROJECTS	68
--	----

FAST, ADAPTED, CURATED FAIR DATA FOR COVID-19 RESEARCH	72
--	----



<b>BIOINFORMATICS ASSISTS DISCOVERY OF DRUGS AGAINST SARS-CoV-2</b>	78
---	----

COVID-19 DRUG RESEARCH – GAINING INSIGHTS WITH PROTEINSPUS	80
--	----

BIOINFORMATICS TOOLS REVEAL SINGLE-CELL TRANSCRIPTOME DYNAMICS OF SARS-CoV-2 INFECTION	86
--	----

VIRTUAL SCREENING FOR SARS-CoV-2 DRUG DEVELOPMENT USING OPEN RESEARCH AND COMPUTE INFRASTRUCTURES	90
---	----

IDENTIFY POTENTIAL DRUGS AND DRUG TARGETS AGAINST SARS-CoV-2 BY HOST FACTOR SIRNA SCREENING	96
---	----

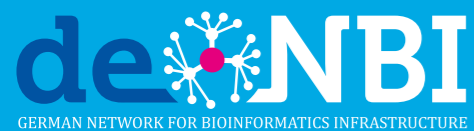


INTERVIEW WITH THE de.NBI COORDINATOR AND HEAD OF NODE OF ELIXIR GERMANY	102
--	-----

THE GERMAN NODE WITHIN ELIXIR EUROPE	106
--------------------------------------	-----

ELIXIR SUPPORT TO COVID-19 RESEARCH	108
-------------------------------------	-----

IMPRINT	110
---------	-----



# THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE – de.NBI

The German Network for Bioinformatics Infrastructure (de.NBI) is a national, academic and non-profit infrastructure which was launched by the German Ministry of Education and Research (BMBF) in March 2015. The network consists of eight Service Centers which are specialized in different omics fields assuring excellent services and high level of expertise. With its wide range of bioinformatics know-how, the de.NBI network is aimed to deliver high standard bioinformatics services, comprehensive training, powerful computing capacities (de.NBI Cloud) as well as connections to industrial companies. The de.NBI network assists researchers to more effectively exploit their data and contributes to the advancement of bioinformatics research in Germany and Europe.

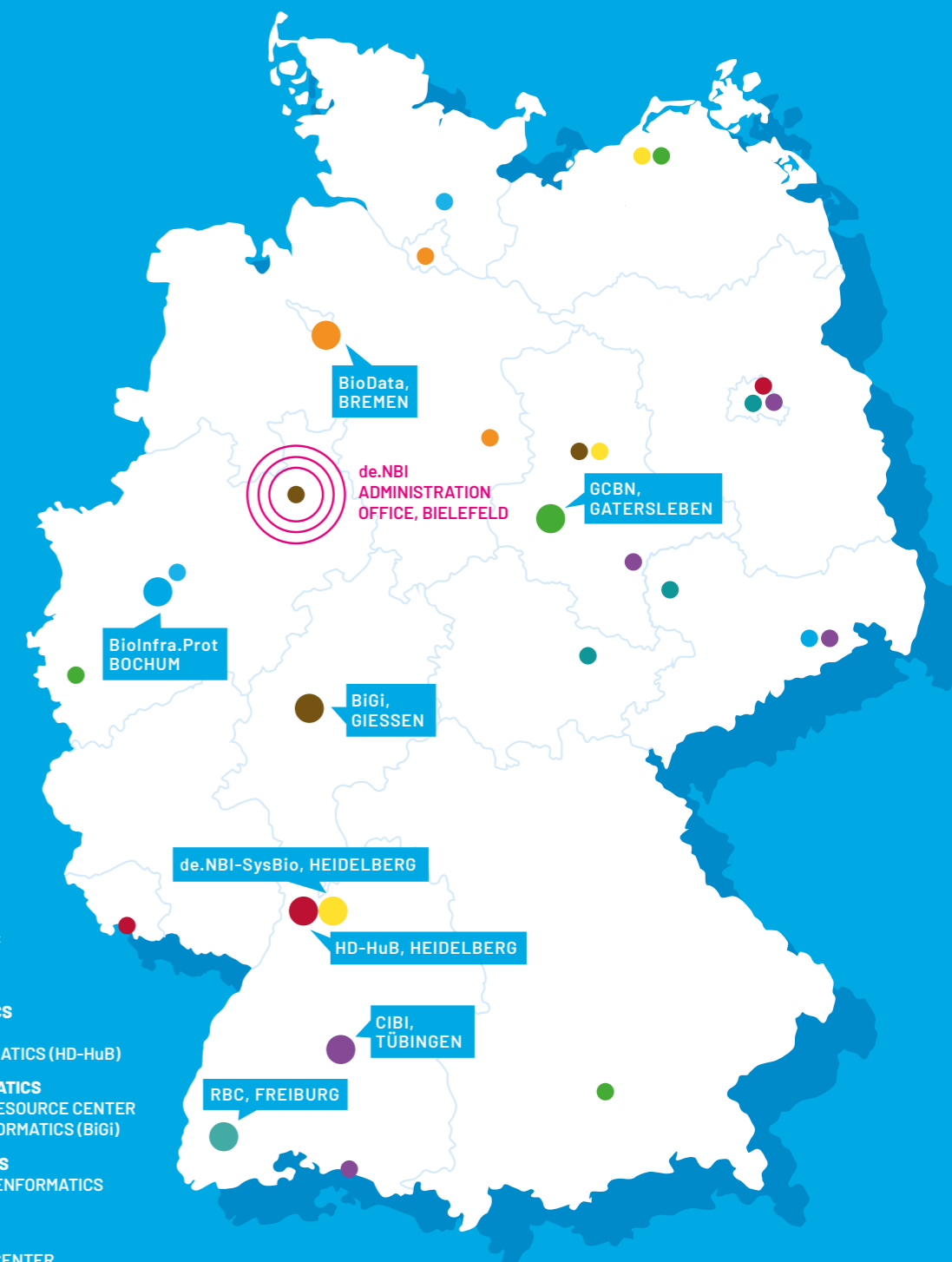


- TOOLS
- WORKFLOWS
- DATABASES
- CONSULTING

- TRAINING COURSES
- SUMMER SCHOOLS
- HACKATHONS
- WEBINARS

- INFRASTRUCTURE
- PLATFORM AND SOFTWARE AS A SERVICE

- SOFTWARE SOLUTIONS
- CONSULTING
- NETWORKING



## THEMATIC FOCUSES & SERVICE CENTERS:

- HUMAN BIOINFORMATICS  
HEIDELBERG CENTER FOR HUMAN BIOINFORMATICS (HD-HuB)
- MICROBIAL BIOINFORMATICS  
BIELEFELD-GIESSEN RESOURCE CENTER FOR MICROBIAL BIOINFORMATICS (BiGi)
- PLANT BIOINFORMATICS  
GERMAN CROP BIOGREENFORMATICS NETWORK (GCBN)
- RNA BIOINFORMATICS  
RNA BIOINFORMATICS CENTER (RBC)
- PROTEOME BIOINFORMATICS  
BIOINFORMATICS FOR PROTEOMICS (BioInfra.Prot)
- INTEGRATIVE BIOINFORMATICS  
CENTER FOR INTEGRATIVE BIOINFORMATICS (CIBI)
- BIODATABASES  
CENTER FOR BIOLOGICAL DATA (BioData)
- DATA MANAGEMENT/SYSTEMS BIOLOGY  
de.NBI SYSTEMS BIOLOGY SERVICE CENTER (de.NBI-SysBio)

- LOCATIONS OF SERVICE CENTERS
- LOCATIONS OF PARTNERS



# de.NBI TRAINING for life scientists

The de.NBI network provides a high-quality, coherent, timely, and impactful training program across its eight Service Centers. Current developments in the field of bioinformatics are also addressed in de.NBI symposia, special workshops and annual summer schools. Life scientists learn how to handle and analyze biological big data more effectively by applying tools, standards, and compute services provided by de.NBI.

**WORKFLOW E-LEARNING** **TRAINING** **SOFTWARE HACKATHON** **RESEARCH** **WORKSHOP** **USER MEETING**

**LEARNING DATABASES** **TOOL** **WEBINAR**

Online



“de.NBI has evolved to be a leading contact point for bioinformatics training in Germany and Europe and guarantees high quality standards.”

**Daniel Wibberg**  
de.NBI Training Coordinator  
contact@denbi.de  
www.denbi.de/training

## Examples from de.NBI's training portfolio

**de.NBI - CeBiTec NANOPORE WORKSHOP (ONLINE)**  
Best Practice and SARS-CoV-2 Applications

**INTRODUCTION TO THE de.NBI CLOUD (ONLINE)**  
de.NBI Cloud User Meeting

**DEPLOYING WEB SERVICES FOR COVID-19 RESEARCH IN THE de.NBI CLOUD (ONLINE)**  
Practical Application in the de.NBI Cloud Infrastructure

**INTRODUCTION TO METAGENOME DATA ANALYSIS**  
Metagenomics Bioinformatics

**INTRODUCTION TO RNAseq DATA ANALYSIS**  
RNAseq and High Throughput Omics Analysis for Plants

**ADVANCED METHODS FOR DIFFERENTIAL ANALYSIS OF PROTEOMICS DATA**  
Analysis of Quantitative Proteomics Data Using R

**BASIC BIOINFORMATICS TRAINING FOR BIOLOGISTS**  
Analysis, Visualization, and Integration of Multi-Level Omics Data

**PRIMARY MEANS TO STORE BIOLOGICAL DATA**  
Ontologies - Statistics, Biases, Tools, Networks, and Interpretation

**ANALYZING SCIENTIFIC DATA USING MACHINE LEARNING ALGORITHMS**  
Machine Learning Using Galaxy

**DATA CARPENTRY WORKSHOP**  
Domain-Specific training Covering the Full Lifecycle of Data-Driven Research

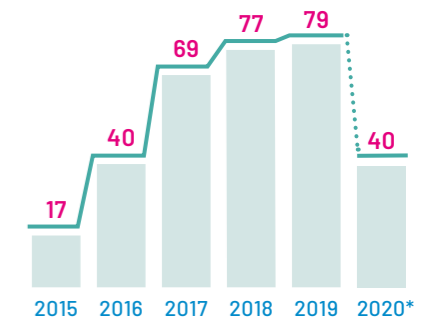
**GLOBAL GALAXY COURSE**  
Wide Variety of GTN (Galaxy Training Network) Tutorials

and many more at:  
[www.denbi.de/training](http://www.denbi.de/training)

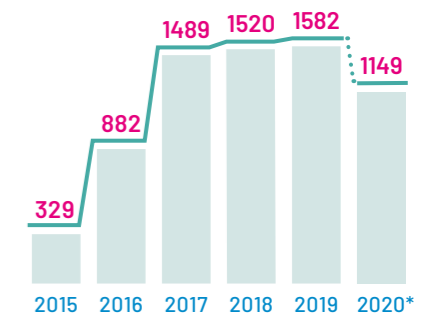


## Training courses 2015-2020

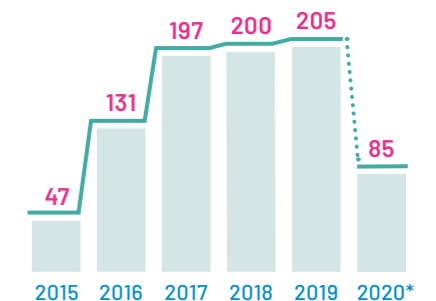
Number of training courses



Number of participants



Number of training days



\* In 2020 only online courses were conducted.



# de.NBI CLOUD

## Compute Power for Your Project

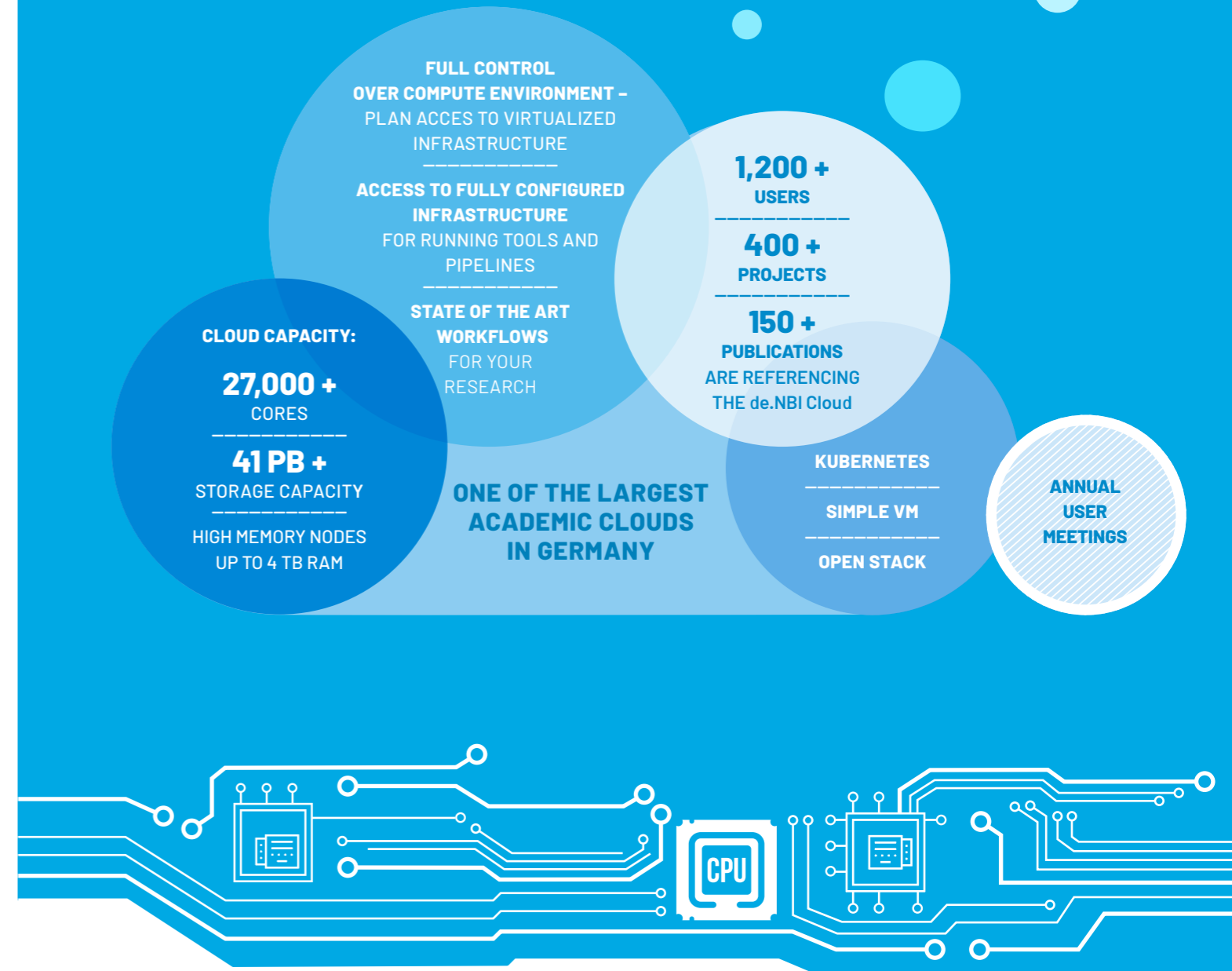
In today's life sciences the handling, analysis and storage of enormous amounts of data is a challenging issue. An appropriate IT infrastructure is crucial to perform analyses with such large datasets and to ensure secure data access and storage. The de.NBI Cloud is an excellent solution to enable integrative analyses and the efficient use of data in research and application. Researchers from the life sciences in Germany can use the de.NBI Cloud free of charge. User meetings are regularly organized to ensure that the requirements of the de.NBI community are taken into account for the future development of the de.NBI Cloud.



“The possibilities of cloud computing open up new perspectives for data processing and analysis in the life sciences.”

**Peter Belmann**  
de.NBI Cloud Governance  
cloud@denbi.de  
<https://cloud.denbi.de/get-started/>

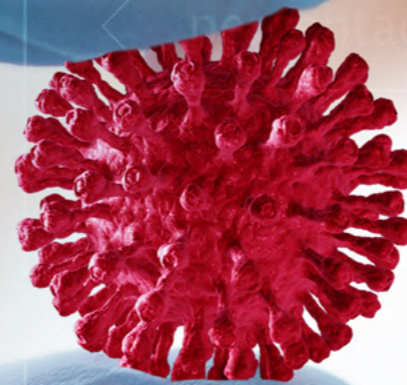
## Compute und Storage Resources



**Project applications are welcome!**  
**Free of charge for German academia!**

# BEYOND THE VIRUS SURFACE – MULTI-OMICS APPROACHES LEAD TO INSIGHTS INTO SARS-CoV-2 AND COVID-19

COVID-19 has become omnipresent in our lives, and researchers are working at full speed to understand the virus infection and spread. Bioinformatics analysis of multi-omics data on SARS-CoV-2 infection offers a great potential to unravel the puzzles of the SARS-CoV-2 pathogenicity, and ways forward for therapeutic innovation.



11010010001110010111100000010000110100100011100101111000000100001101001000111001011110000  
00110100100011100101111000000100001101001000111001011110000001000011010010001110010111100  
11100101111000000100001101001000111001011110000001000011010010001110010111100000010000110  
10010001110010111100000010000110100100011100101111000000100001101001000111001011110000001  
11010010001110010111100000010000110100100011100101111000000100001101001000111001011110000  
0011010010001110010111100000010000110100100011100101111000000100001101001000111001011110000  
0011010010001110010111100000010000110100100011100101111000000100001101001000111001011110000  
0011010010001110010111100000010000110100100011100101111000000100001101001000111001011110000

# MANAGING OMICS DATA IN THE CONTEXT OF THE COVID-19 PANDEMIC



In the past years, substantial efforts have been made in Germany to set up bioinformatics infrastructure, with de.NBI spearheading and coordinating these initiatives. Already early on during the COVID-19 pandemic, the value of platforms such as de.NBIs was proven: operational data infrastructures, cloud computing resources, robust software tools, and experienced staff were already available to efficiently process relevant data. This existing backbone of activities enabled us to set up operational infrastructures and protocols for data capturing, managing, and the analysis of omics data generated in the context of SARS-CoV-2 infections. Here, we illustrate (1) how de.NBI, ELIXIR and NFDI have joined forces to create an active bridge to international activities in order to provide a data management and analysis infrastructure, as well as (2) the value of such infrastructures for an agile response to emerging challenges.

## OMICS DATA ARE ESSENTIAL FOR THE UNDERSTANDING OF COVID-19 ON THE MOLECULAR LEVEL

Since the COVID-19 pandemic caused the ongoing global health crisis, it was widely recognized to be essential to gain a better understanding of the molecular mechanisms driving the infection, and - more critically - to understand the different degrees of severity in patients. Different types of COVID-19 related omics data have thus been collected in local, but increasingly concerted efforts in Germany, such as the German COVID-19 OMICS Initiative (DeCOI; [www.decoi.eu](http://www.decoi.eu)). These include virus and patient genomes, as well as transcriptomics, proteomics data revealing the expressed proteins, and associated phenotypic/clinical data describing disease progression, therapies, and outcomes. Although the initiative described here focuses on the omics data, we would like to point out that linking these data to associated clinical data (e.g., as collected in clinical studies such as LEOSS or NAPKON) will be critical to interpret the omics data or to build predictive models.

Clearly, sequencing the SARS-CoV-2 genome is important to understand

the virus, but also to trace infection chains and detect changes in the virus that could indicate changes in its infectiousness. As with other viruses, it is essential to study SARS-CoV-2 in context, by understanding patients' genomes, and critical aspects of the host immune response. Different patients experience drastic differences in severity of a SARS-CoV-2 infection, which is in part related to their genetic characteristics. Exome sequencing, whole-genome sequencing, epigenome or transcriptome sequencing and single cell sequencing in patient samples can be essential to predict severity and progression, and thus choose appropriate therapeutic options. Crucial added value is generated by integrating these data types: for example, host genomes and virus genomes, as well as clinical progress and outcomes are critical to build integrated models or to understand mechanistic details. Since data quality is key for the success of such modeling efforts, this requires a concerted effort to establish data standards, workflows and approaches to harmonize data on a national and European level. These efforts profit tremendously from central data and computing infrastructures such as GHGA and the de.NBI-Cloud.

## COMPLEMENTARY AND COMPATIBLE INFRASTRUCTURES ENGAGING IN COVID-19 OMICS DATA GENERATION, PROCESSING AND SHARING

### GHGA

Human genome sequencing and other omics data modalities are of critical importance for biomedical research and the future development of healthcare. The German Human Genome Phenome Archive (GHGA, <http://www.ghga.de>), a novel national infrastructure that will be established as part of the national research data infrastructure initiative will provide the means to meet both the desire to handle omics data in an open and FAIR manner and the need to keep personal data safe and secure. Distinct from existing European infrastructures, setup as a National consortium, GHGA will be able to effectively address the legal requirements specific to Germany, enabling German researchers to help shape future international standards for omics data exchange, also in the context of COVID-19 data. The employment of the de.NBI Cloud for data processing and thereby avoiding the transfer of sensitive data to (inter)national partners will help to mitigate some of the risks associated with working with sensitive data (Figure 1).



### DeCOI

DeCOI (<http://www.decoi.eu>) is a German national network founded in March 2020, which aims to generate next-generation sequencing (NGS)-based omics data for COVID-19 research. It has been established as a network to make a specific contribution to the COVID-19 crisis through the use of NGS technologies. DeCOI currently involves more than 88 members from 45 institutions in 22 cities across Germany that bring together expertise in NGS data generation (Figure 2). GHGA is one of the founding consortia of DeCOI, which in collaboration with de.NBI offers extended expertise in data management and data analysis for the COVID-19 omics data generated.

### de.NBI/ELIXIR-DE Task Force for human COVID-19 omics data

de.NBI and GHGA is working together with the European Bioinformatics Institute (EBI) and its European Genome-Phenome Archive (EGA), and with other collaborators within the ELIXIR-CONVERGE project, to establish a European COVID-19

Data Portal. Submission of COVID-19 omics data to the COVID-19 Data Portal will be coordinated by the de.NBI/ELIXIR-DE task force for human COVID-19 omics data. Data stewards will provide assistance during data submission ('helpdesk') and metadata curation for COVID-19 data in a human context, to enable these data to become interoperable before they are uploaded into the COVID-19 Data Portal. While we envision that initially data will come primarily from DeCOI, the de.NBI/ELIXIR-DE task force will offer submission assistance and metadata curation also for other related datasets from Germany.

Bioinformatics tools that will support this activity range from the whole spectrum of omics-related bioinformatics tools available through de.NBI, from sequence alignment to read quantification, differential gene expression analysis etc.

### SERVICES AND ACTIVITIES

GHGA, DeCOI and de.NBI have joined forces and established a task force that

will bridge national activities related to COVID-19 research to international activities and in particular the COVID-19 Data Portal. The task force has the mission to facilitate the consistent curation of human omics data and will provide extended submission support via dedicated data stewards. Generated datasets will be submitted to both national (GHGA) and international (EGA) archives, to facilitate uptake and secondary use. Nationally, the COVID-19 human omics task force will foster connected activities, the sharing of omics data and the deployment of novel analytical methods, including AI, to fully exploit the data being generated. Data submitted to GHGA will additionally be made available to authorized users in the de.NBI Cloud. This provision of infrastructure and data will make it possible for all researchers to investigate their ideas independent of local infrastructure constraints (a single human genome for example requires about 150 GB of storage).



FIGURE 2: DeCOI (The German COVID-19 OMICS Initiative, DeCOI, which combines >88 groups from >45 institutions based in Germany).

- GHGA = German human genome/phenome archive, <https://ghga.dkfz.de/>
- is a new national research data infrastructure (NFDI) for human omics data from Germany
  - German hub of the federated EGA (European Genome/Phenome Archive network), which will be implemented and carried out by ELIXIR-CONVERGE
  - has been authorized to assemble German COVID-19 data and share them in the European context via the European COVID-19 Data Portal
- NFDI = national research data infrastructure, [www.nfdi.de/en-gb](http://www.nfdi.de/en-gb)
- has the objective to collect and share data from research within Germany in a centralized manner
  - among the medical NFDI consortia GHGA (collection of omics data) and NFDI4Health (collection of personal health data) will use synergies and work together in the COVID-19 crisis
- DeCOI = German COVID-19 OMICS initiative, [www.decoi.eu](http://www.decoi.eu)
- is the national network for omics data in Germany
  - has authorized GHGA to assemble German COVID-19 data and share them in the European context via the European COVID-19 Data Portal
- European COVID-19 Data Portal, [www.covid19dataportal.org/](http://www.covid19dataportal.org/)
- central portal from the European Commission which collects and shares all European COVID-19 data from the individual countries
  - is operated at EMBL-EBI and uses the infrastructure of the federated EGA network
- de.NBI/ELIXIR-DE Task Force for human COVID-19 omics data
- is a newly established organ to coordinate and assist submission of COVID-19 omics data to the European COVID-19 Data Portal
- de.NBI-Cloud, <https://cloud.denbi.de>
- federated national research cloud with six different sites financed by the German federal ministry of education and research (BMBF)
  - provides hardware infrastructure (IaaS) as well as software (SaaS) and support

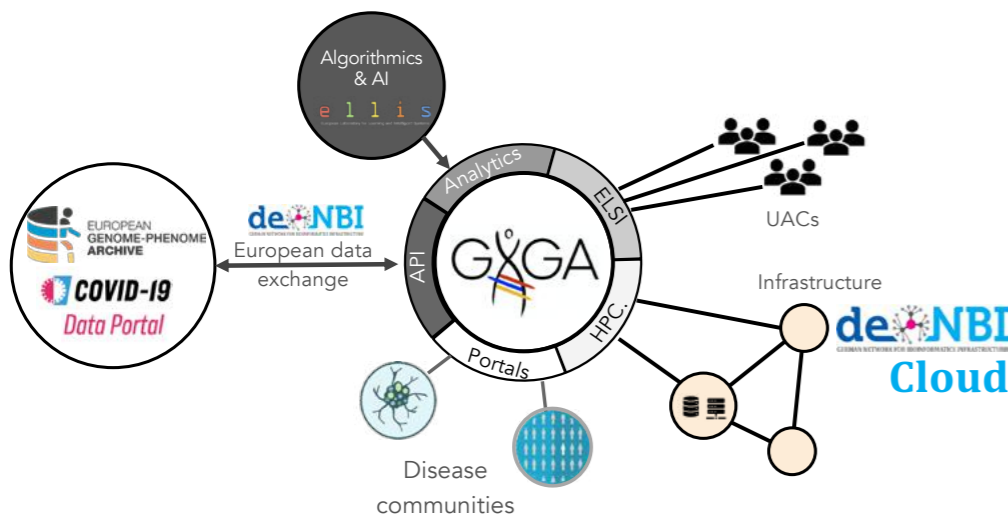


FIGURE 1: GHGA and de.NBI/ELIXIR-DE. Interaction between GHGA, de.NBI and the COVID-19 Data Portal.

**AUTHORS:** Sina Barysch<sup>1</sup>, Peer Bork<sup>1</sup>, Ivo Buchhalter<sup>2</sup>, Oliver Kohlbacher<sup>3,4</sup>, Jan Korbel<sup>1</sup>, Jens Krüger<sup>3</sup>,

Joachim L. Schultze<sup>5</sup>, Oliver Stegle<sup>1,2</sup> and the GHGA-Consortium<sup>6</sup>

<sup>1</sup> European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg

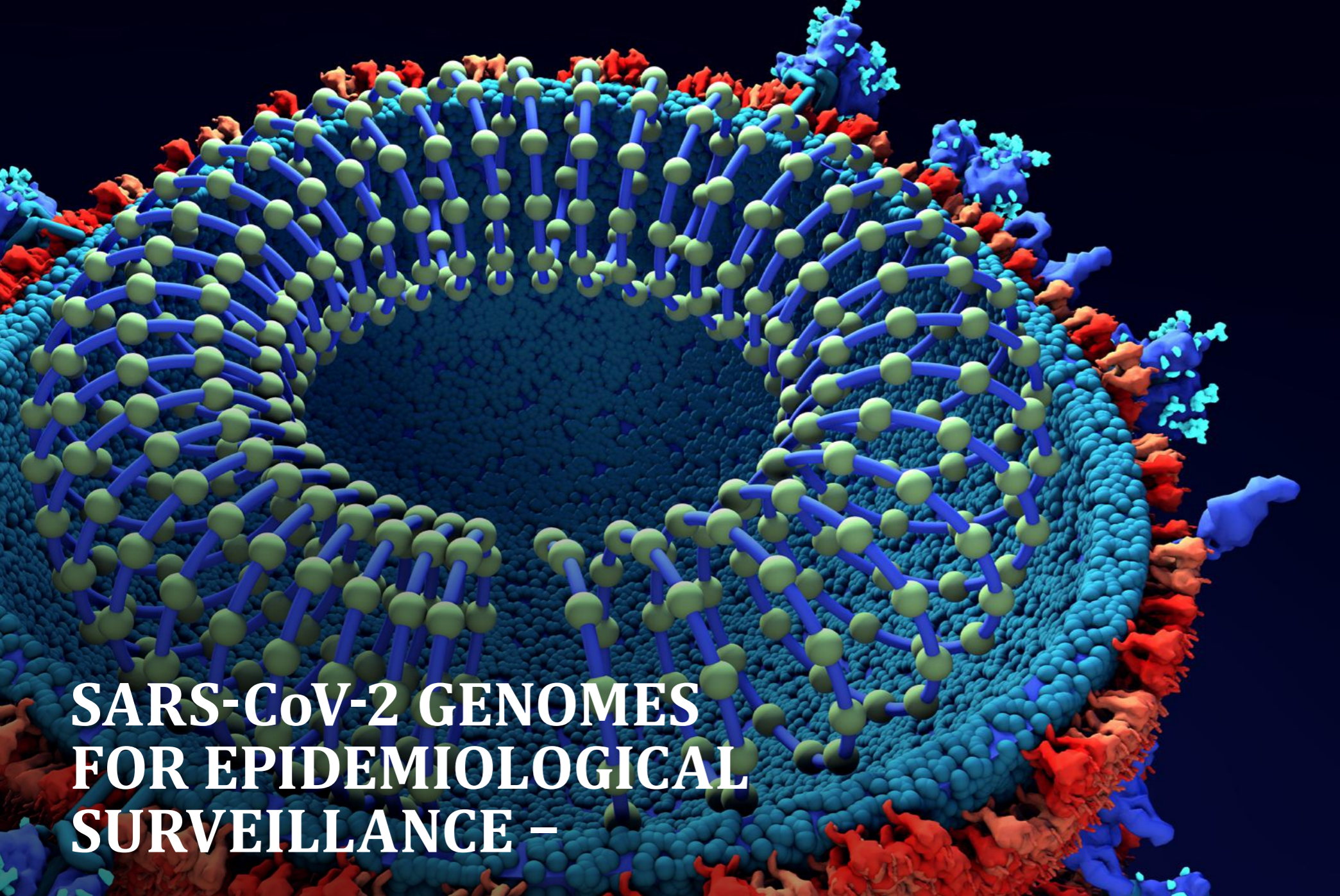
<sup>2</sup> Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg

<sup>3</sup> Eberhard Karls Universität Tübingen, Sand 14, 72076 Tübingen

<sup>4</sup> University Hospital Tübingen, Geissweg 3, 72076 Tübingen

<sup>5</sup> Deutsches Zentrum für Neurodegenerative Erkrankungen e. V. (DZNE), Venusberg-Campus 1/99, 53127 Bonn

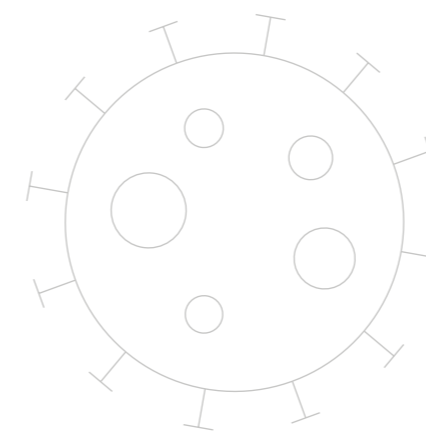
<sup>6</sup> German Human Genome-Phenome Archive, [www.ghga.de](http://www.ghga.de)



# SARS-CoV-2 GENOMES FOR EPIDEMIOLOGICAL SURVEILLANCE –

## Mutational signatures are highly informative for tracing infection chains and virus evolution

The spread of COVID-19 is monitored by testing for SARS-CoV-2 RNA in the human nose and throat. However, the isolated virus RNA is not routinely sequenced, leaving the highly informative pattern of virus mutations unexplored. We here present our workflow for establishing high-quality complete genome sequences. Comparative analyses with SARS-CoV-2 genomes in databases and from presumed members of infection chains or infection clusters yield critical new information to be used for epidemiological surveillance and, on the longer run, on virus evolution.



The deadly COVID-19 pandemic arose in China in late 2019 and spread over the whole planet within weeks. As of 12/2020 more than 67 million infections have been reported and over 1.5 million COVID-related fatalities. The genome of the COVID-19 causing SARS-CoV-2 consists of single stranded RNA and has a length of 29,903 nucleotides. Testing for the virus is routinely done by quantitative reverse transcriptase real-time PCR (qRT-PCR), and by measuring the abundance of RNA for regions of specific virus genes. Most frequently, the virus genes N and E are targeted, and a detection of these transcripts in levels higher than a threshold means a COVID-positive diagnosis. In fact, since the ct-value (cycle exceeding threshold) of qRT-PCR is inversely proportional to the original relative transcript amount, low ct-values mean higher virus loads.

Beside its mere presence, the SARS-CoV-2 RNA has much more information to offer. The multiple copying of virus RNA in infected human cells leads to 'copying errors' (mutations). Since the frequency of these mutations is low and does not occur in any patient, transmission of the virus normally conserves a mutation pattern and an infection chain can be traced back by identical or very similar mutations.

As the mutation pattern of the virus provides a further measure to validate infection chains and identify 'hot spots' or problematic environmental settings in which larger numbers of transmissions occur, SARS-CoV-2 genome analysis has emerged as a powerful tool for epidemiological surveillance in many countries. However, as compared to routine testing by qRT-PCR, today only a tiny fraction of SARS-CoV-2 genomes have been sequenced.

In addition to this, isolated virus RNA is often discarded by the testing labs after the assay and not comprehensively available for the sequencing labs.

### NANOPORE SEQUENCING OF CoV-2 GENOMES USING THE ARTIC NETWORK PIPELINE

Any medical testing is critically dependent on standardisation. One of the early attempts to provide a standardized workflow for SARS-CoV-2 genome analysis came from the ARTIC network (<https://artic.network/>). This network aims for a 'real-time molecular epidemiology for outbreak response' for several emerging viral diseases. As such, it provides an end-to-end solution from the virus outbreak itself up to epidemiological information that is interpretable and actionable by public health bodies in a near real-time manner. The workflow is assembled around a 'real-time' single molecule sequencing technology, nanopore sequencing. The sequencing devices are small, portable and affordable. They can be used not only in the lab, but also in the field and in other low-tech settings.

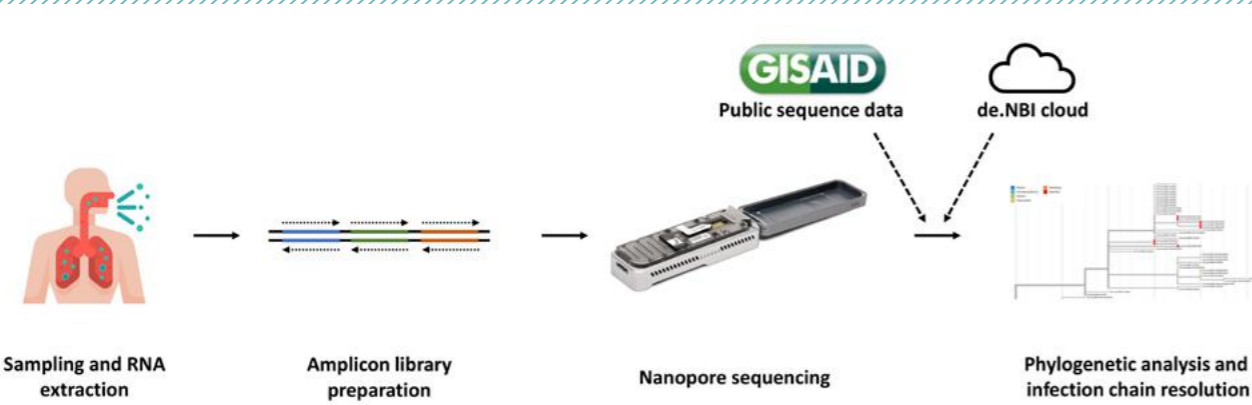
We originally adopted the ARTIC network protocols for SARS-CoV-2 sequencing and slightly modified them in the course of our work to reduce workload and processing times (Figure 1). Starting from a qRT-PCR-positive RNA sample, the processing steps in the lab comprise quality and quantity control by capillary electrophoresis, reverse transcription of single-stranded virus RNA into its double-stranded cDNA equivalent and in generating so-called amplicons for nanopore sequencing from the cDNA by standard PCR. The amplicons are generated in pools and cover the whole virus genome. In the ARTIC network protocols, two sets of 49 primer pairs each ensure that

the whole virus genome is covered in an overlapping fashion similar to roof tiles. From these amplicon pools, a sequencing library is generated and sequenced on a nanopore sequencer, in our case a GridION X5 (Oxford Nanopore Technologies Inc.). Since sequencing is very fast, it only takes minutes to a few hours to generate sufficient sequence information to be analysed for mutations. Due to the inherent error rate of the nanopore sequencing technology, a certain coverage of reads (~50fold) is needed to establish the correct amplicon sequences. The ARTIC network protocols also include software running on the raw nanopore data for experiment control and another software pipeline for matching the amplicon sequence reads to a SARS-CoV-2 reference genome. Putative mutations have

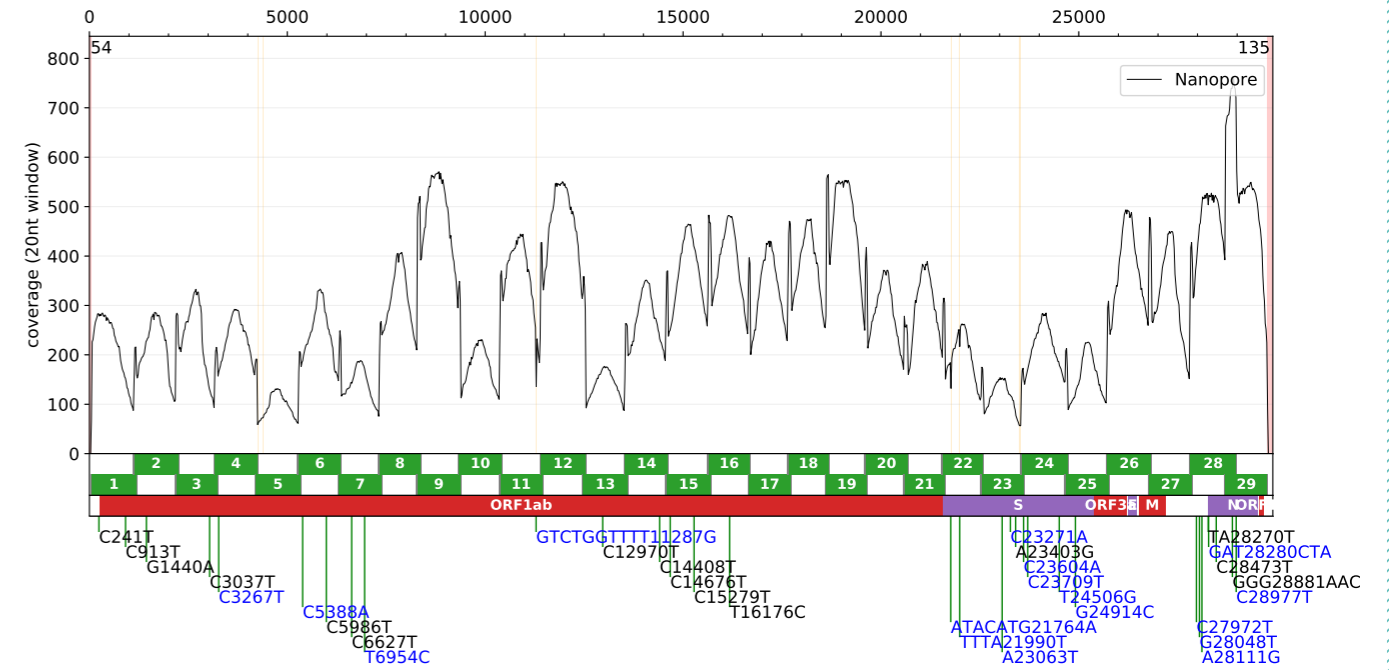
to pass several filtering steps and need to be identified by more than one method to avoid false positive calls.

Although the protocols work fine in general, there are several sources of problems, especially prominent for samples with low virus RNA concentrations. In addition, the original workflow took us 8 hours for one sequencing cycle, however, with the possibility to process 24 samples at the same time. We were able to simplify the protocol and to shorten the processing time considerably by using fewer but longer amplicons which also enable the use of a faster library preparation protocol. With the experimental modifications, including new sets of PCR primers, we are now down to 4 hours processing time for 12 samples at the maximum.

Further problems in the workflow could arise by a drop-out or low sequence coverage of individual amplicons that have to be redone. As mentioned above, these problems are the more frequent, the lower the virus RNA concentration is. At last, we aim at individually validating every called mutation by at least one complementary technology, either classical Sanger sequencing of amplicons or by transcriptome sequencing using the Illumina sequencing technology. In addition, we have established our own project management and quality control pipeline around the ARTIC network software, allowing us to integrate all sequencing and mutation detection information and providing a real-time view on all samples in the sequencing project (Figure 2).



**FIGURE 1:** The SARS-CoV-2 sequencing workflow. Extracted virus RNA is obtained from local clinics and subjected to amplicon library preparation using the ARTIC protocol. Amplified DNA is then sequenced using nanopore sequencing and analyzed by comparing the Wuhan reference sequence and publicly available SARS-CoV-2 sequences using de.NBI Cloud computing resources.



**FIGURE 2:** Detailed view on a sequenced SARS-CoV-2 genome from the B.1.1.7 lineage. Sample summary from our internal quality control tool. Tiled amplicon locations are visualized using green boxes on top of the SARS-CoV-2 genome annotation. Mutations are displayed at the bottom, with green lines indicating an approved mutation call after manual curation and blue lettering used for mutations defining the B.1.1.7 lineage.

### LEARNING ABOUT INFECTION CHAINS AND VIRUS EVOLUTION FROM THE PATTERNS OF MUTATIONS

The workflow described above mostly leads to full-length high-quality SARS-CoV-2 genomes. It has to be mentioned, however, that mutations at the ultimate ends of the virus genome (30-50 nucleotides) can not be called because of the use of specific primers for amplicon generation. Nevertheless, the remaining virus genome gives not only a bar code for tracing virus infection chains, but in the longer run also for tracing virus evolution.

Our specific interest is in a survey of cases in a smaller region in eastern Northrhine-Westfalia that is called 'Ostwestfalen' (OWL) and there, specifically, the districts of Bielefeld and neighboring

Gütersloh. We obtained samples from a Bielefeld clinic (Evangelisches Klinikum Bethel) and a local testing lab (MVZ Diamedis GmbH). The survey started with samples obtained in March 2020 up to May 2020, when this original low prevalence region had an infection hot spot in a local meat processing plant affecting more than 2,000 workers and people in the surrounding area (Figure 3).

Although especially the mentioned outbreak attracted international attention and scientific interest [1], there is ample information also specifically in the longer time regional survey. Virus genome comparisons not only show the spread of viruses with additional mutations from the 'meat processing plant outbreak' into the next district, but also short infection chains that appear to have been successfully contained. In this line, we

have observed mutations that have never been found elsewhere. Also, the plethora of samples we have sequenced so far does not support the idea of a long range transmission from other parts of the world to our region. This, however, is expected to change in the course of the recently ending holiday season.

Besides providing surveillance data, we will learn a lot on virus evolution through comprehensive and temporarily resolved mutation patterns.

**It can only be hoped that government and science funding bodies provide resources to establish these data sets in the near future.**

Important building blocks that are critical for success are well-annotated databases (GISAID for example; <https://www.gisaid.org/>) and powerful bioinformatics tool sets such as Nextstrain (<https://nextstrain.org/>, [2]). In addition, it is critical to involve not only the virus sequencing labs but to build national and international alliances between scientists from a broad spectrum of fields. Initiatives like DeCOI (German COVID-19 Omics Initiative; <https://decoi.eu/>) or the European Leoss network (<https://leoss.net/>) are a step in the right direction and also a better way to address governmental bodies or policy makers.

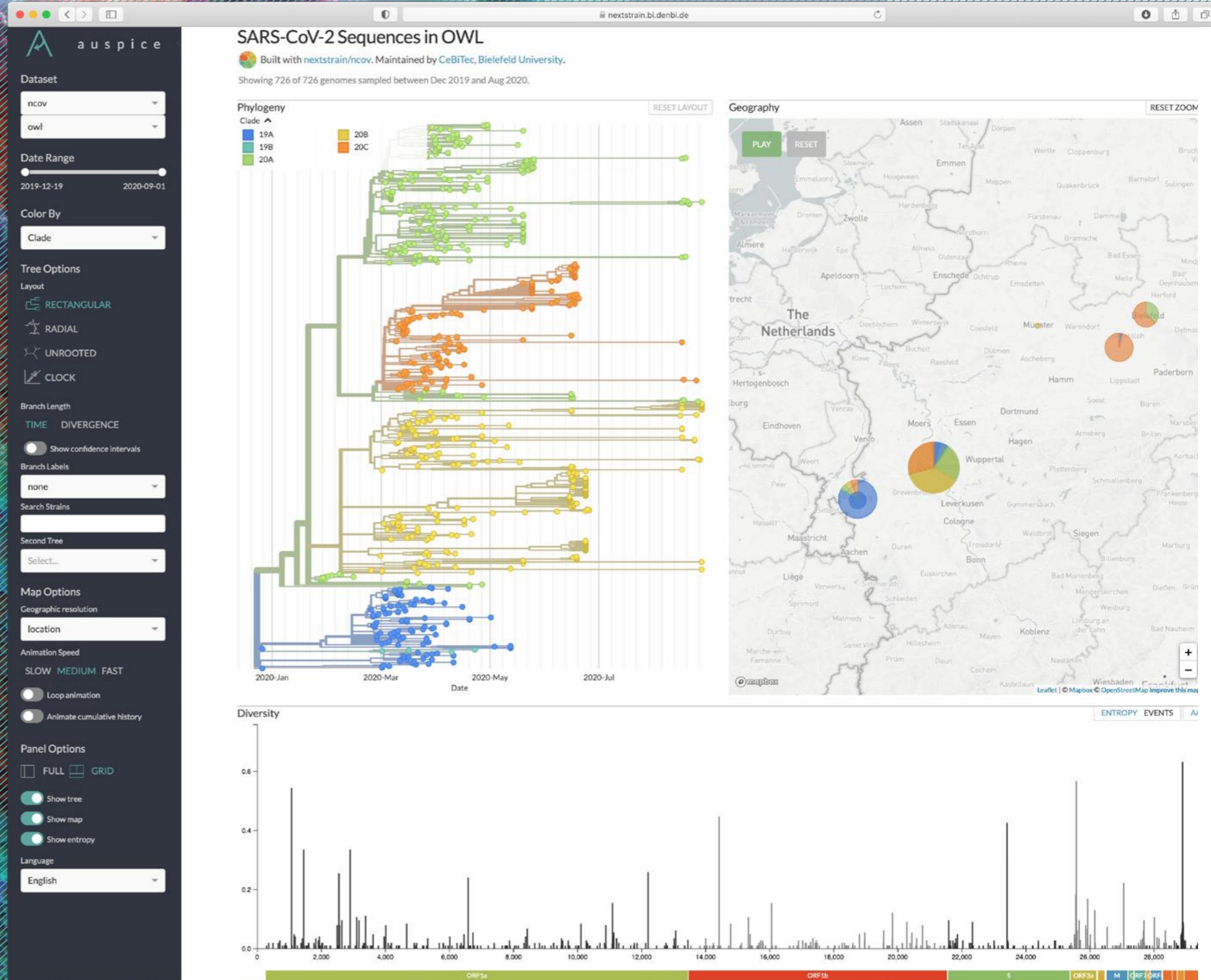
### CONCLUSIONS

The continuous and comprehensive survey of SARS-CoV-2 genomes and analyses of their mutational patterns has proven successful for epidemiological surveillance and will deliver also critical data on virus evolution in the longer run. Therefore, despite being a threat to humankind, the COVID-19 pandemic might speed up the establishment of functional genome-based epidemiological surveillance systems in general.

### ACKNOWLEDGEMENTS

We are most grateful for receiving high-quality RNA samples from our cooperation partners Evangelisches Klinikum Bielefeld Bethel (Dr. Christiane Scherer) and VMZ Diamedis GmbH (Dr. Thomas Diedrich). For partial funding, the de.NBI initiative and the Federal Ministry of Education and Research (BMBF) is thankfully acknowledged.

**FIGURE 3:** Comparison of SARS-CoV-2 genomes with the Nextstrain toolkit. The main Nextstrain interface consists of three linked panels – a phylogenetic tree (upper left), geographic transmissions (upper right) and the genetic diversity across the genome (bottom).



**REFERENCES:** [1] SSRN 2020; DOI: <https://doi.org/10.2139/ssrn.3654517>. [2] *Bioinformatics* 2018;34(23):4121-4123. DOI: <https://doi.org/10.1093/bioinformatics/bty407>.

**AUTHORS:** Jörn Kalinowski<sup>1</sup>, David Brandt<sup>1</sup>, Tobias Busche<sup>1</sup>, Markus Haak<sup>1</sup>, Levin-Joe Klages<sup>1</sup>, Marina Simunovic<sup>1</sup>, Svenja Vinke<sup>1</sup>, Alexander Sczyrba<sup>1</sup>  
<sup>1</sup> Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstrasse 27, 33615 Bielefeld

# RNA BIOINFORMATICS TO ANALYZE SARS-CoV-2 – the causative agent of COVID-19

SARS-CoV-2 is a positive-strand RNA virus. Hence, its single-stranded RNA genome can act directly as mRNA. Therefore, structured RNA elements are crucially relevant for the virus to interact with the host cells' systems. Information about these RNA elements is central to understand its pathogenicity and to develop strategies for pharmacological intervention. Functionally relevant RNA elements are known in the 5' and 3' untranslated regions and in a region mediating a programmed ribosomal -1 frame-shifting. These elements are involved in the regulation of viral RNA synthesis, packaging and production of the ORF1ab polyprotein but many additional RNA elements are likely to exist *in vivo*.

Here, we chose the element involved in ribosomal frame-shifting and production of the ORF1ab polyprotein to illustrate

the usefulness of computational approaches in analyzing the RNA biology of SARS-CoV-2. The GLASSgo algorithm was used to identify homologous sequences in more remotely related viruses. LocARNA aligns RNA sequences with unknown structure and predicts a consensus secondary structure, illustrated here for the set of homologous RNA fragments identified by GLASSgo. Finally, we show that the CARNA algorithm can predict pseudoknot structures in the SARS-CoV-2 frame-shift control region that are close to the structures experimentally determined by NMR.

These algorithms are developed, maintained and provided by the 'RNA Bioinformatics Center', a resource for RNA-focused bioinformatics research and teaching. The tools possess substantial potential to facilitate the comparative analysis of RNA elements in SARS-CoV-2.

## STRUCTURED RNA ELEMENTS IN SARS-CoV-2 ARE FUNCTIONALLY RELEVANT

The Severe Acute Respiratory Syndrome Coronavirus [1] (SARS-CoV-2), the etiological agent of the Coronavirus Disease 2019 (COVID-19), was discovered by the end of 2019 in China [1] and has since been developed into a catastrophic global pandemic.

Similar to other human pathogenic viruses such as the coronaviruses SARS [2] and MERS-CoV [3], but also Zika [4], West Nile Virus [5], Dengue virus [6] and many others, SARS-CoV-2 belongs to the group of positive-strand RNA viruses. Hence, their single-stranded RNA genomes can act directly as mRNA.

Segments of RNA fold and behave in peculiar ways that are functionally relevant and may constitute pharmacological intervention targets. That is not different for the Betacoronaviruses to which SARS-CoV-2 belongs.

Therefore, workflows, algorithms, and analytical tools specifically developed to analyze molecular RNA data are crucially relevant for understanding the properties of SARS-CoV-2 and ultimately developing suitable antidotes. In the quest for effective therapies, transcriptional host responses to infections with SARS-CoV-2 are of particular interest. Similar to earlier work on negative-strand RNA viruses [7], a recent study compared the transcriptional host responses to SARS-CoV-2 with the negative-strand human respiratory syncytial (RSV) and influenza A virus (IAV) [8]. The investigation reported a significant attenuation of interferon expression (IFN-1 and IFN-2) in human primary lung cells and transformed alveolar cells while chemokine expression appeared to be upregulated. This specific combination is known from other coronaviruses

[cf. 9,10] and maybe critical to understand and treat excessive inflammatory responses to SARS-CoV-2 infections. An independent study investigating the peripheral blood of COVID-19 patients also reported dysregulation of IFN expression [11]. Notably, coronaviruses encode for a specific endonuclease (EndoU) capable to delay the MDA5 mediated host response by cleaving the viral polyuridine sequence [12]. It has been hypothesized that the permanent shortening of the poly-U tail by EndoU inhibits the formation of stem-loop structures to avoid the recognition by dsRNA sensors [cf. 13]. This mechanism may thus limit the formation of a pathogen-associated molecular pattern (PAMP) and could be a key factor in suppressing the IFN production. Conversely, SARS-CoV-2 has been described as a remarkably structured virus [14]. Under certain conditions, RNA structures have been suggested to play a role in evading the innate defense systems and improve viral persistence [15,16]. For instance, the positive-strand RNA Dengue virus has been shown to produce a subgenomic RNA that directly binds to the ubiquitin ligase TRIM25 [17]. The formation of this nucleic acid sequence is modulated by a pseudoknot in the 3' UTR. The interaction of the subgenomic RNA and TRIM25 significantly represses IFN.

**In the quest for effective therapies, transcriptional host responses to infections with SARS-CoV-2 are of particular interest.**

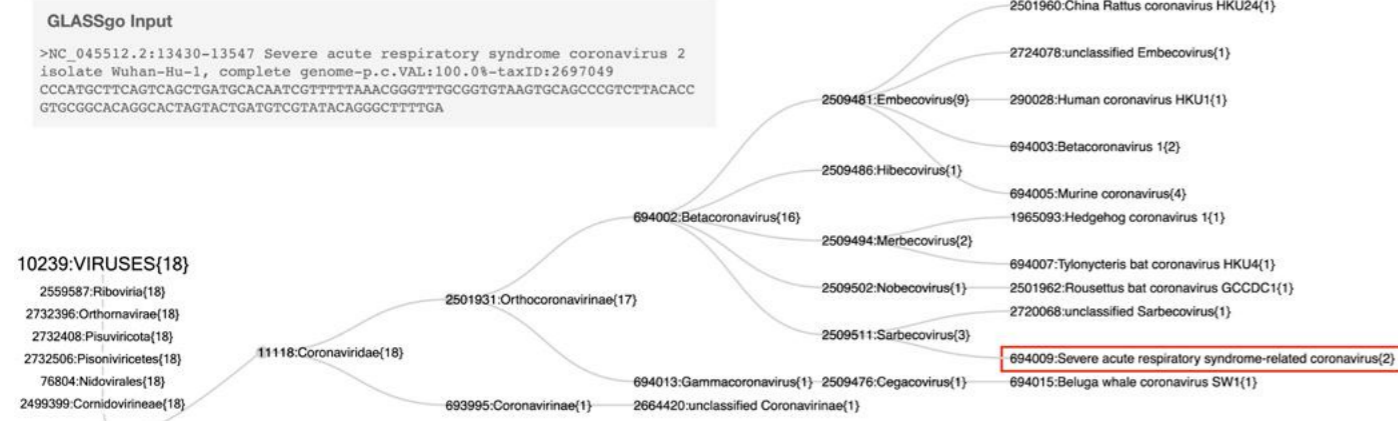
With the first sequenced SARS-CoV-2 isolate, Wuhan-Hu-1 (29,903 nt), at hand [1], it is possible to analyze these structures in detail. Several cis-acting RNA elements were shown previously in related coronaviruses to be functionally important. These include secondary structures in the 5' and 3' untranslated regions (UTRs)

relevant for the regulation of RNA synthesis, viral replication, and packaging [18]. The 3' UTR can form alternative RNA structures that contribute to the control of RNA synthesis at various stages, including a 3' UTR pseudoknot [19]. A complex RNA structure consisting of a three-stemmed pseudoknot [20], an attenuator hairpin, and a so-called 'slippery site' is in SARS-CoV-2 critically relevant to produce the ORF1ab polyprotein via programmed ribosomal -1 frame-shifting [20-23].

It can be expected that many more functionally relevant RNA structures exist in the SARS-CoV-2 genome. There is evidence for secondary structures that form *in vivo* and support the existence of novel regulatory motifs and mechanisms in the SARS-CoV-2 genome [24]. However, such potential novel RNA elements have mostly remained unexplored thus far. Currently, an abundance of primary SARS-CoV-2 genome sequences is being produced, and this can be accessed through the NCBI SARS-CoV-2 Resources interface at <https://www.ncbi.nlm.nih.gov/sars-cov-2/> with 22,981 complete and nearly-complete natural genomes available today (September 08, 2020).

## COMPUTATIONAL TOOLS FOR THE ANALYSIS OF SARS-CoV-2 STRUCTURED RNA ELEMENTS

Therefore, tools for the computational analysis of these sequence data are crucial. Primary genome sequences can be analyzed and integrated with proteomic, cheminformatics, and other datasets on the Galaxy open-source platform (see <https://covid19.galaxyproject.org> and [25]). Complementary to this, cyberinfrastructure for addressing virtually all aspects of RNA-related analyses has been compiled by the 'RNA Bioinformatics Center' (RBC). This center comprises all major bioinformatics groups in Germany that are devoted to the



**FIGURE 1:** Identification of homologous SARS-CoV-2 RNA sequences using the GLASSgo algorithm [33]. The 118 nt sequence 'CCCAUGCUUCAGUCAGCUGAUG-CACAAUCGUUUUUAACGGGU-UUGCGGUGUAAAGUCAGCCCGU-CUUACACCGUGCGGCACAGGCAC-UAGUACUGAUGUCGUUACAG-GCCUUUUGA' derived from the

Wuhan-Hu-1 isolate [1] as given in Figure 1B of the analysis by [22] was used as query (upper left corner, 'input'). The analysis was conducted using the latest RefSeq database from 07-10-2020 ([ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/](http://ftp.ncbi.nlm.nih.gov/refseq/release/viral/)) in combination with GLASSgo version 1.5.2 (<https://github.com/lotts/GLASSgo>). The red box

indicates the position of the newly detected SARS-CoV-2. The numbers in front of each taxon refer to the respective taxonomic identifiers and the numbers in brackets behind it to the number of entries in the database. Please note that the number of SARS-CoV-2 isolates was manually restricted to two.

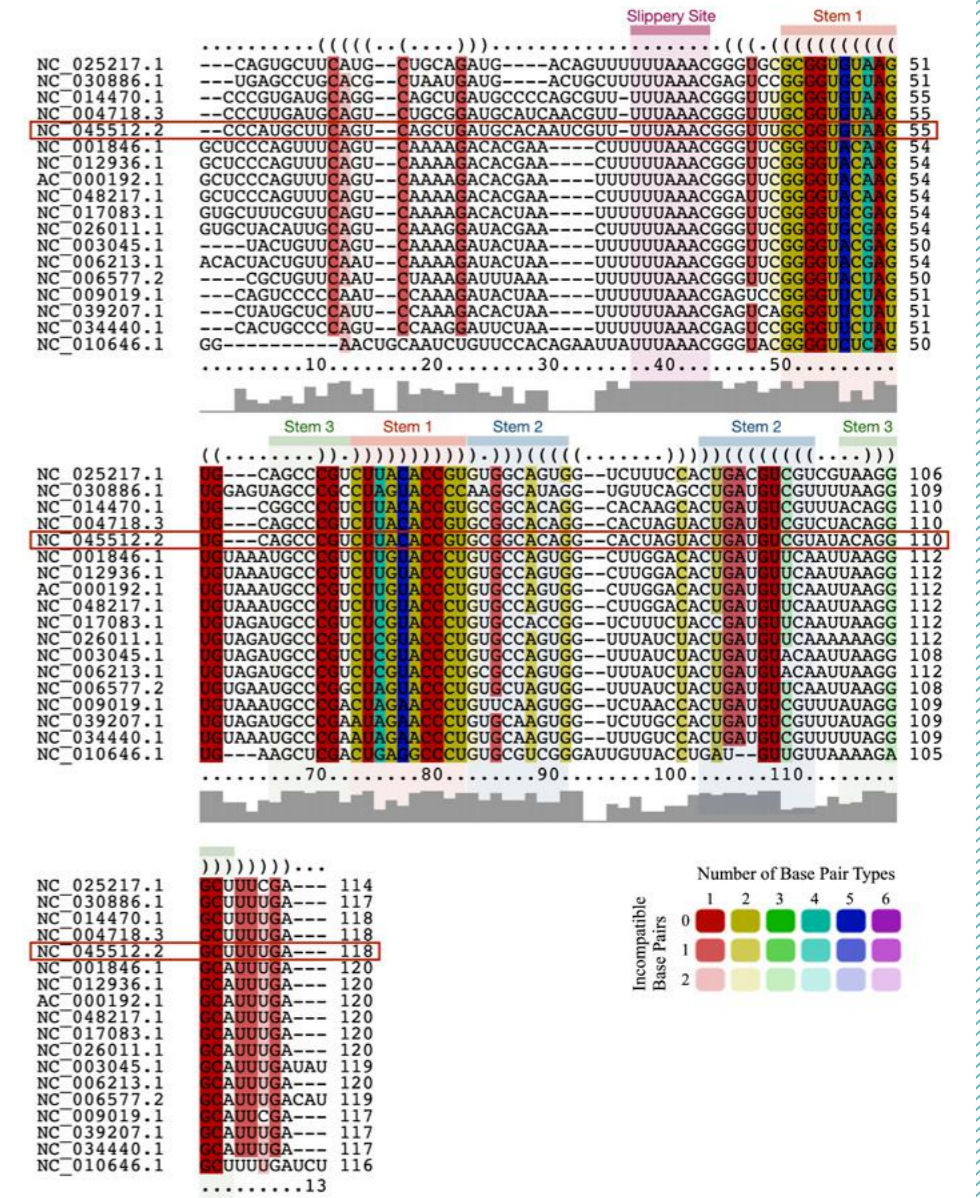
study of RNA. These groups are located at the University of Freiburg, the University of Leipzig, the University of Rostock, the Max Delbrück Center – Berlin, and the Leibniz Institute on Aging in Jena. Although the different groups are geographically dispersed, they closely collaborate and provide a central online resource for RNA-focused bioinformatics research and teaching [26].

The center provides tools for the RNA-related data analysis ranging from transcriptome analysis [27] to RNA sequence and structure analysis [28], the prediction of sRNA and miRNA targets [29,30], the definition and classification of RNA transcripts, the analysis of mutations specifically RNA features [31], the visualization of RNA family models affecting [32] and the analysis of RNA-protein interactions.

Here, we illuminate the power of some of these tools using the sequence from the SARS-CoV-2 genome that controls the programmed ribosomal -1 frame-shifting for the production of the ORF1ab polyprotein. Taking a 118 nt fragment containing the critical RNA elements as the basis, homologous sequences in more remotely related viruses can be identified efficiently using the GLASSgo algorithm [33]. Naturally, homologous sequences can be identified in other coronaviruses but also in more remote viruses (Figure 1). The GLASSgo algorithm was initially developed for the identification of bacterial small regulatory RNAs (sRNAs) homologous to a given query sequence, where it has become very popular [34,35]. However, bacterial sRNAs are short, structured RNA molecules explaining why this tool works so well also in predicting homologous sequences to the SARS-CoV-2

ORF1ab control region. The GLASSgo algorithm has recently been integrated into the Galaxy platform, making its use even more straightforward [36].

Another tool, LocARNA, computes multiple RNA alignments based on their sequence and structure similarity [37]. In Figure 2, the relevant RNA elements in the SARS-CoV-2 control region involved in the programmed ribosomal -1 frame-shifting to produce the ORF1ab polyprotein, including the 'slippery site' [20-23, 38,39] are annotated alongside the LocARNA alignment. In contrast to other multiple alignment tools, such as MARNA [40], LocARNA considers the whole ensemble of secondary structures for each RNA. Thus, LocARNA aligns RNAs with unknown structure and predicts a consensus secondary structure for a set of RNAs. Specification of additional



**FIGURE 2:** LocARNA multiple alignment of the SARS-CoV-2 control region involved in the programmed ORF1ab frame-shifting based on sequence and structure similarities. The multiple sequence alignment was based on the GLASSgo search results (same settings as mentioned in Figure 1) and produced using LocARNA v1.9.1 [37]. The SARS-CoV-2 sequence is boxed in red. The slippery site is annotated. Compatible base pairs are coloured. The predicted secondary structure is indicated in dot-bracketed annotation and refers to the entire alignment as shown. In contrast, the annotated stems 1 to 3 refer only to the SARS-CoV sequences. The hue shows the number of different types C-G, G-C, A-U, U-A, G-U, or U-G of compatible base pairs in the corresponding columns. In this way, the hue shows the sequence conservation of the base pair. The saturation decreases with the number of incompatible base pairs. Thus, it indicates the structural conservation of the base pair.

The RefSeq identifiers refer to the following viruses: NC\_025217.1, Bat Hp-betacoronavirus/Zhejiang013; NC\_030886.1, Rousettus bat coronavirus isolate GCCDC1 356; NC\_014470.1, Bat coronavirus BM48-31/BGR/2008; NC\_004718.3, SARS coronavirus; NC\_045512.2, SARS-CoV-2 isolate Wuhan-Hu-1; NC\_001846.1, Mouse hepatitis virus strain MHV-A59 C12 mutant; NC\_012936.1, Rat coronavirus Parker; AC\_000192.1, Murine hepatitis virus strain JHM; NC\_048217.1, Murine hepatitis virus strain A59; NC\_017083.1, Rabbit coronavirus

HKU14; NC\_026011.1, Betacoronavirus HKU24 strain HKU24-R050051; NC\_003045.1, Bovine coronavirus; NC\_006213.1, Human coronavirus OC43 strain ATCC VR-759; NC\_006577.2, Human coronavirus HKU1; NC\_009019.1, Bat coronavirus HKU4-1; NC\_039207.1, Betacoronavirus Erinaceus/VMC/DEU/2012 isolate ErinaceusCoV/2012-174/GER/2012; NC\_034440.1, Bat coronavirus isolate PREDICT/PDF-2180; NC\_010646.1, Beluga Whale coronavirus SW1.

The GLASSgo algorithm was initially developed for the identification of bacterial small regulatory RNAs (sRNAs)

structural RNAs, in particular, of low sequence similarity. However, similar to many other programs used to predict structured RNA elements, LocARNA cannot consider putative pseudoknots such as those relevant in the SARS-CoV-2 ORF1ab control region.

pseudoknots and represent those in the form of a so-called ‘consensus’ graph [28].

CARNA successfully predicts the pseudoknots in the SARS-CoV-2 frame shift control region (Figure 3). For this result, we used as input only the 69 nt structure-relevant region of five sequences aligned in (Figure 2) and a sequence experimentally investigated by NMR

constraints or even enforcement of fixed input structures is possible. LocARNA is best suited to compare

Another tool for multiple alignments of structured RNA sequences, CARNA, can be employed. CARNA is able to consider

(<https://covid19-nmr.de/rna-results/pseudoknot/>). Indeed, the computationally predicted interacting bases match experimental analysis by NMR (structures in the Biological Magnetic Resonance Data Bank doi:10.13018/BMR50348)[41].

The precise identification of structured RNA elements in the SARS-CoV-2 genome and their functional importance is

currently studied with great effort, highlighted by initiatives such as the international COVID19-NMR project (<https://covid19-nmr.de/>), or the COVID-19 Resources section run by the RFAM database (<https://rfam.xfam.org/covid-19>). The tools maintained, developed, and provided by the RNA Bioinformatics Center substantially facilitate the comparative analysis of RNA elements in SARS-CoV-2.



ACKNOWLEDGEMENT

This work has been supported by the German Federal Ministry of Education and Research (BMBF) through the de.NBI project to R. Backofen (O31A538A) and the de.STAIR project to W. R. Hess and S. Hoffmann (O31L0106).

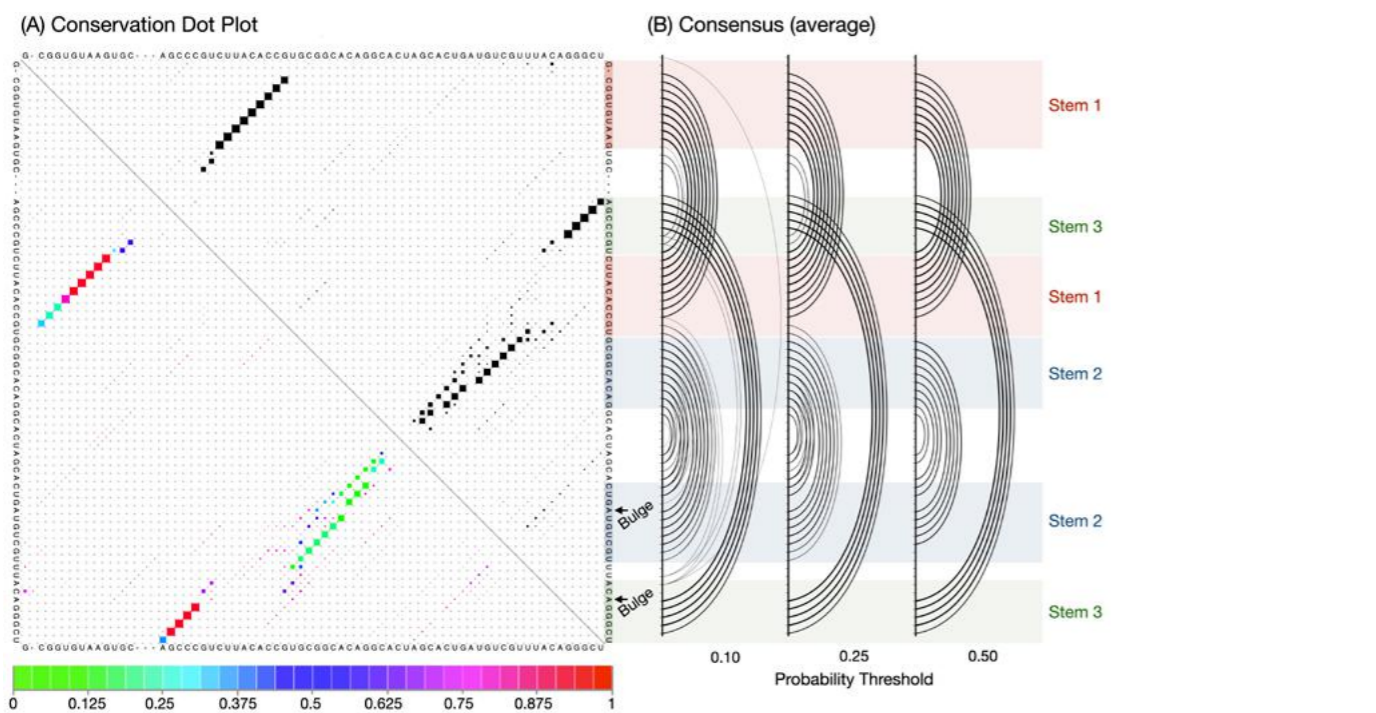


FIGURE 3: Multiple alignments of RNAs considering the formation of pseudoknots using the CARNA algorithm v1.3.3 [28]. A: Dot plot of conservation. The lower left triangle of the dot plots contains the average dot plot colored with variance information. Pure green means maximum variance. Pure red means no variance at all (the dot has the same probability in all sequences). B: Consensus structure graphs referring to the dot plot in panel (A), representing three different probabilities (0.10, 0.25, and 0.50). In addition, the stems of the three stem-loop elements in the frame shifting control element of SARS-CoV-2 are indicated by the differently colored boxes.

THE FOLLOWING SEQUENCES SERVED AS INPUT FOR CARNA:

>NMR  
GGCGGTGTAAGTGCAGCCCGTCTTACACCGTGGCGCACAGGCACTAGTACTGATGTCGTATACAGGGCT  
>NC\_014470.1:13304-13371 BAT CORONAVIRUS BM48-31/BGR/2008  
GCGGTGTAAGTGCAGCCCGTCTTACACCGTGGCGCACAGGCACTAGTACTGATGTCGTTTACAGGGCT  
>NC\_004718.3:13405-13472 SARS CORONAVIRUS  
GCGGTGTAAGTGCAGCCCGTCTTACACCGTGGCGCACAGGCACTAGTACTGATGTCGTCTACAGGGCT  
>NC\_045512.2:13475-13542 SARS-CoV-2 ISOLATE WUHAN-HU-1  
GCGGTGTAAGTGCAGCCCGTCTTACACCGTGGCGCACAGGCACTAGTACTGATGTCGTATACAGGGCT  
>NC\_009019.1:13561-13627 BAT CORONAVIRUS HKU4-1  
TTCTAGTGAATGCCGACTAGAACCCTGTTCAAGTGGTCTAACCCTGATGTCGTTTATAGGGCA  
>NC\_025217.1:13938-14005 BAT HP-BETACORONAVIRUS/ZHEJIANG2013  
GCGGTGTAAGTGCAGCCCGTCTTACACCGTGGCGCACAGGCACTAGTACTGATGTCGTTTACAGGGCT

REFERENCES: [1] Nature 2020;579(7798):265-269. DOI: <https://doi.org/10.1038/s41586-020-2008-3>. [2] N. Engl. J. Med. 2003;348(20):1967-1976. DOI: <https://doi.org/10.1056/NEJMoa030747>. [3] J. Virol. 2013;87(14):7790-7792. DOI: <https://doi.org/10.1128/JVI.01244-13>. [4] The Lancet 2016;387(10015):227-228. DOI: [https://doi.org/10.1016/S0140-6736\(16\)00003-9](https://doi.org/10.1016/S0140-6736(16)00003-9). [5] Clin. Microbiol. Rev. 2012;25(4):635-648. DOI: <https://doi.org/10.1128/CMR.00045-12>. [6] Cell. Mol. Life Sci. 2010;67(16):2773-2786. DOI: <https://doi.org/10.1007/s00018-010-0357-z>. [7] Sci. Rep. 2016;6(1):34589. DOI: <https://doi.org/10.1038/srep34589>. [8] Cell 2020;181(5):1036-1045.e9. DOI: <https://doi.org/10.1016/j.cell.2020.04.026>. [9] Virol. J. 2006;3(1):17. DOI: <https://doi.org/10.1186/1743-422X-3-17>. [10] J. Virol. 2005;79(4):2079-2086. DOI: <https://doi.org/10.1128/JVI.79.4.2079-2086.2005>. [11] medRxiv 2020;[Preprint]. DOI: <https://doi.org/10.1101/2020.07.20.20155507>. [12] Proc. Natl. Acad. Sci. 2020;117(14):8094-8103. DOI: <https://doi.org/10.1073/pnas.1921485117>. [13] J. Virol. 2012;86:2900-2910. DOI: <https://doi.org/10.1128/JVI.05738-11>. [14] mBio 2020;11:e01661-20. DOI: <https://doi.org/10.1128/mBio.01661-20>. [15] RNA 2004;10(9):1337-1351. DOI: <https://doi.org/10.1261/rna.7640104>. [16] Front. Immunol. 2018;9:2097. DOI: <https://doi.org/10.3389/fimmu.2018.02097>. [17] Science 2015;350(6257):217-221. DOI: <https://doi.org/10.1126/science.aab3369>. [18] Methods Mol. Biol. Clifton NJ 2020;2203:1-29. DOI: [https://doi.org/10.1007/978-1-0716-0900-2\\_1](https://doi.org/10.1007/978-1-0716-0900-2_1). [19] RNA 2011;17(9):1747-1759. DOI: <https://doi.org/10.1261/rna.2816711>. [20] PLoS Biol. 2005;3(6):e172. DOI: <https://doi.org/10.1371/journal.pbio.0030172>. [21] PLoS One 2013;8(4):e62283. DOI: <https://doi.org/10.1371/journal.pone.0062283>. [22] J. Biol. Chem. 2020;295(31):10741-10748. DOI: <https://doi.org/10.1074/jbc.AC120.013449>. [23] Front. Biosci. J. Virtual Libr. 2008;13:4873-4881. DOI: <https://doi.org/10.2741/3046>. [24] BioRxiv Prepr. Serv. Biol. 2020;[Preprint]. DOI: <https://doi.org/10.1101/2020.07.10.197079>. [25] Nucleic Acids Res. 2020;48(W1):W395-W402. DOI: <https://doi.org/10.1093/nar/gkaa434>. [26] Nucleic Acids Res. 2018;46(W1):W25-W29. DOI: <https://doi.org/10.1093/nar/gky329>. [27] Bioinformatics 2018;34(6):1066-1068. DOI: <https://doi.org/10.1093/bioinformatics/btx690>. [28] Nucleic Acids Res. 2012;40:W49-53. DOI: <https://doi.org/10.1093/nar/gks491>. [29] Nucleic Acids Res. 2017;45(15):8745-8757. DOI: <https://doi.org/10.1093/nar/gkx605>. [30] Nucleic Acids Res. 2014;42:W119-123. DOI: <https://doi.org/10.1093/nar/gku359>. [31] Nat. Commun. 2019;10(1):2569. DOI: <https://doi.org/10.1038/s41467-019-10489-2>. [32] Bioinformatics 2018;34(15):2676-2678. DOI: <https://doi.org/10.1093/bioinformatics/bty158>. [33] Front. Genet. 2018;9:124. DOI: <https://doi.org/10.3389/fgene.2018.00124>. [34] RNA 2020;26(10):1448-1463. DOI: <https://doi.org/10.1261/rna.076992.120>. [35] Nucleic Acids Res. 2018;46(17):8803-8816. DOI: <https://doi.org/10.1093/nar/gky584>. [36] Bioinformatics 2020;15:4357-4359. DOI: <https://doi.org/10.1093/bioinformatics/btaa556>. [37] RNA 2012;18(5):900-914. DOI: <https://doi.org/10.1261/rna.029041.111>. [38] Virology 2005;332(2):498-510. DOI: <https://doi.org/10.1016/j.virol.2004.11.038>. [39] Virus Res. 2006;119(1):29-42. DOI: <https://doi.org/10.1016/j.virusres.2005.10.008>. [40] Bioinforma. Oxf. Engl. 2005;21(16):3352-3359. DOI: <https://doi.org/10.1093/bioinformatics/bti550>. [41] Nucleic Acids Res 2020;gkaa1013. DOI: <https://doi.org/10.1093/nar/gkaa1013>.

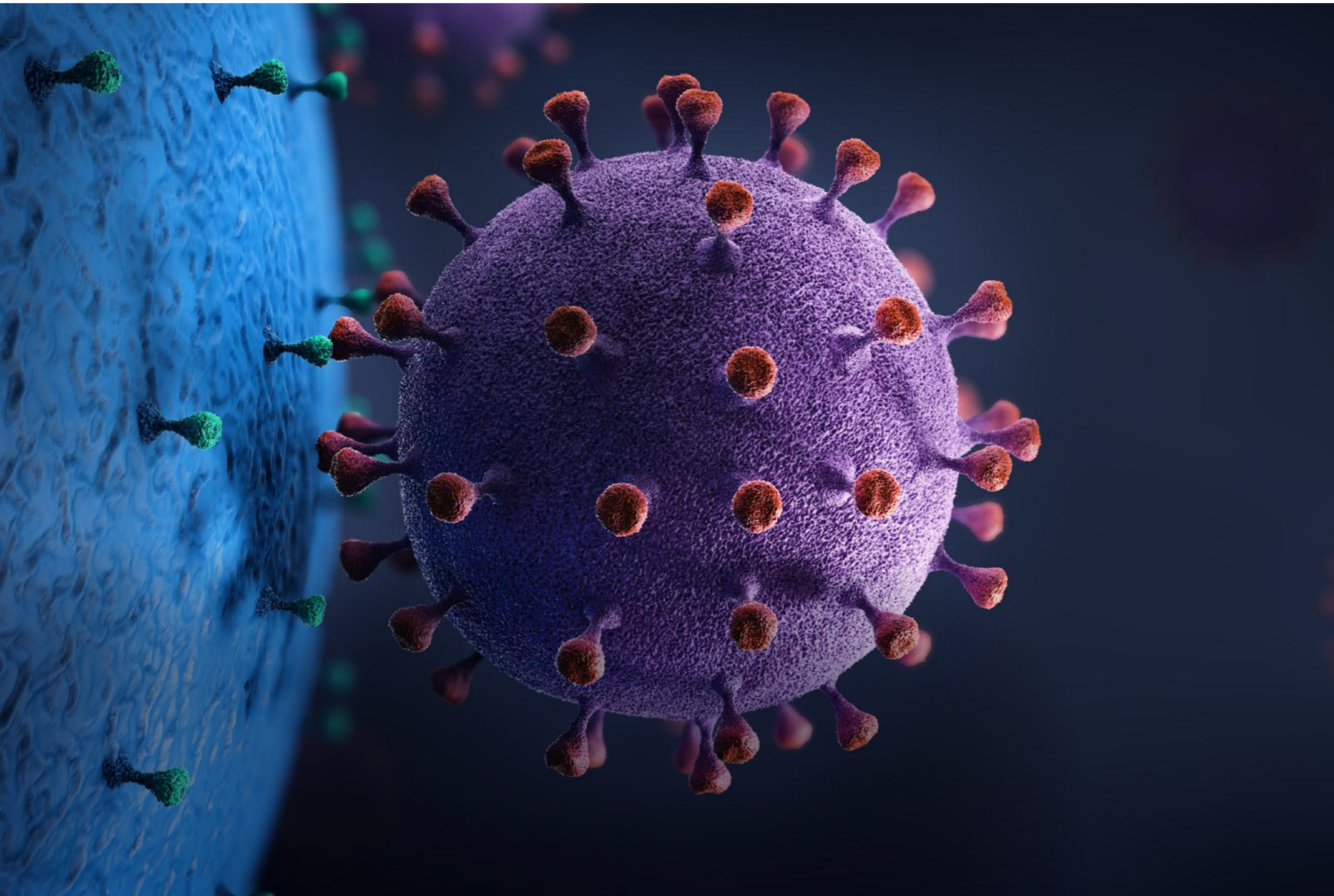
AUTHORS: Wolfgang R. Hess<sup>1</sup>, Steffen C. Lott<sup>1</sup>, Steve Hoffmann<sup>2</sup>, Rolf Backofen<sup>3</sup>

<sup>1</sup> Genetics & Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg

<sup>2</sup> Computational Biology Group, Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstraße 11, 07745 Jena

<sup>3</sup> Bioinformatics Group, Faculty of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79104 Freiburg

# IDENTIFYING MOTIF MIMICRY IN SARS-CoV-2 HOST PATHOGEN INTERACTIONS



Viruses are known to hijack features of human host proteins to alter cellular systems for their own benefit. As it is not yet known whether SARS-CoV-2 exploits similar hijacking mechanisms, we used a battery of sophisticated protein bioinformatics tools to predict them, providing a number of putative mechanistic insights into how the virus alters human systems. Our ranked list of SARS-CoV-2/human interaction mechanisms (often involving short linear motifs) immediately suggest possible therapeutic targets for COVID-19.

Molecular interactions between proteins are central to all biological processes, whether this is the normal human kidney cell, bacterial cell division or the process by which a virus invades and takes over a human host cell. Known the molecular details of how proteins interact can provide a deeper understanding of disease and more importantly can suggest new potential points for therapeutic intervention. At this point, still little is known about the mechanism of how SARS-CoV-2 and human proteins interact, though initial candidates have been identified [1]. In this work, we use preliminary data and a battery of computational techniques to predict the molecular details of how SARS-CoV-2 bind to human proteins, both on the cell surface and within the cell after the virus has penetrated the human cell.

It is well established that viruses often hijack host cellular networks by exploiting natural interaction mechanisms [2]. Often this is via a mimicry of peptide segments, or linear motifs [3, 4], which are short protein stretches used all over biology to facilitate interactions between human proteins [5]. For instance, the NS1 protein of influenza A

H3N2 exploits a mimic of a human histone peptide sequence to direct gene expression in target human cells. Since it is likely that SARS-CoV-2 exploits such strategies when infecting humans, we predicted possible points of interaction between virus and human proteins via a computational pipeline (Figure 1). For each pair, we looked for candidate protein segments that could, in principle, interact with each other.

A major aspect of this was to first establish often distant relationships between SARS-CoV-2 proteins and other proteins of known three-dimensional structure using a sensitive, but computationally intensive, sequence database searching procedure [6]. This provided structural matches for more SARS-CoV-2 segments than would come by traditional sequence comparison. We currently run this program via the de.NBI HD-HuB Cloud.

To identify or predict potential interaction points we used a diversity of methods developed during the last two decades. InterPRETS [7] predicts interactions for a pair of proteins by identifying homologous proteins



in contact inside previously determined structures. We also used motifs curated in the ELM [5] databases to find putative motifs in SARS-CoV-2 proteins and their counterpart domains inside human proteins.

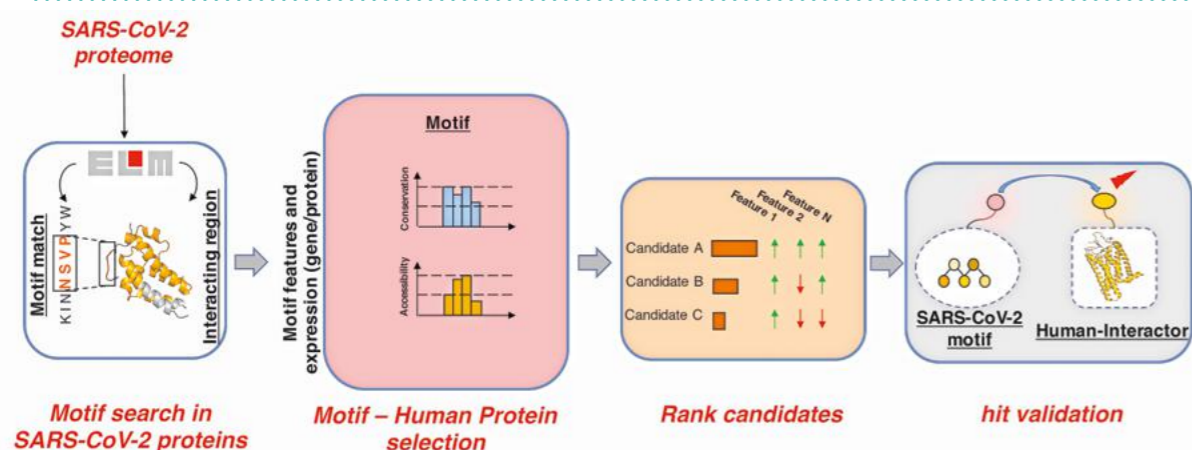
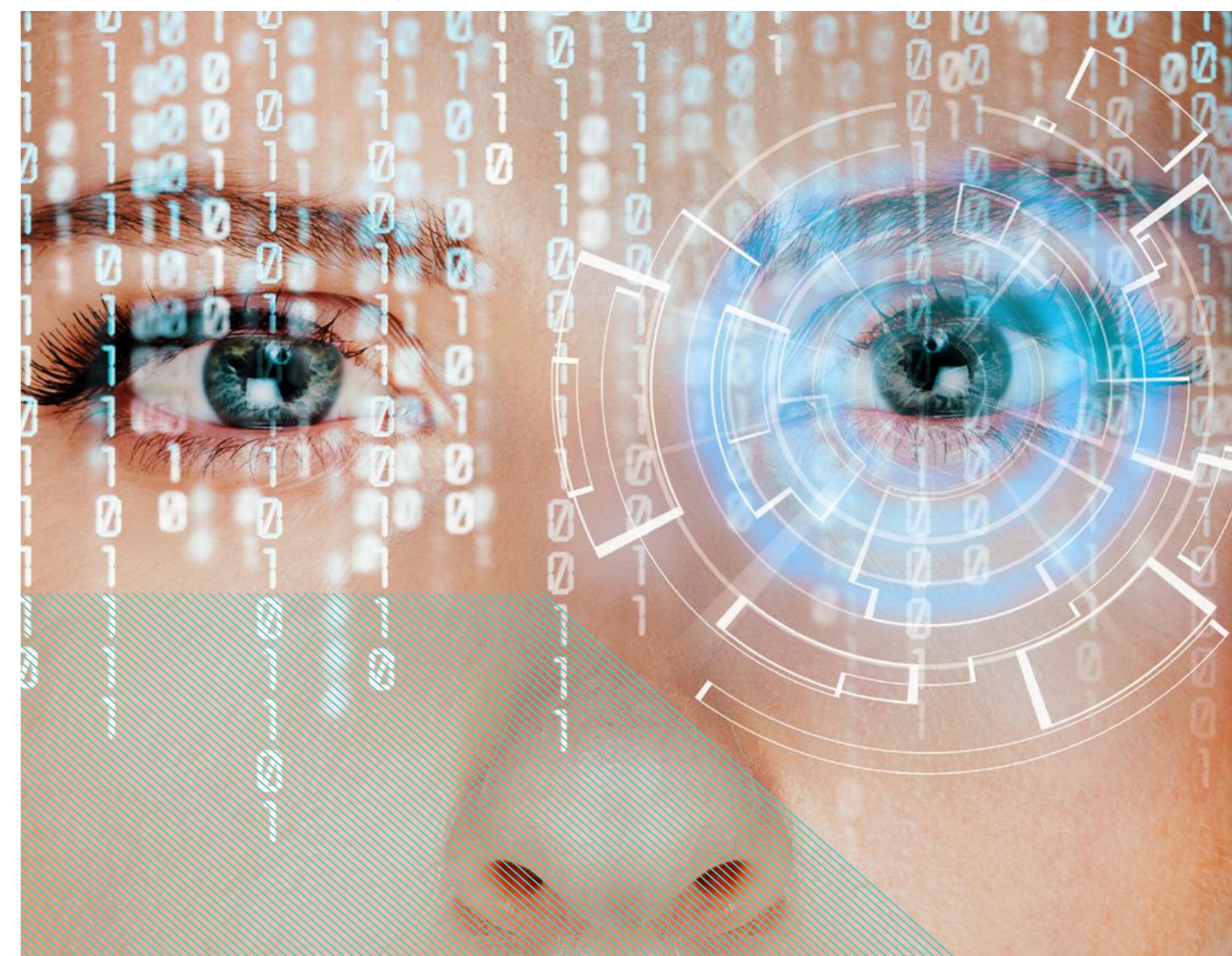
This process identified an initial set of 22140 possible interaction points. We know that the vast majority of these will be false positives, owing to the simplicity of certain peptide motifs, which makes random instances very likely. We therefore sought to use additional, contextual information on protein and gene function to filter out unlikely pairs (e.g. similar to [8]).

This filtering left only a few tens of interaction points, which we then ranked using an assortment of statistical and bioinformatics metrics. Importantly, this process identifies several known interactions, including the SARS-CoV-2 Spike protein binding to human ACE2 [9]. There are also several tantalizing new potential interaction points between SARS-CoV-2 and human proteins. For example, the viral nsp12 protein contains a motif that (in other contexts) interacts with guanylate kinases. These proteins play critical roles in recycling GMP/cGMP, the second messenger that plays key roles

in NO metabolism and vascular muscle control [10, 11]. It is tempting to suggest a potential viral interaction critical aspect of vascular system function. Indeed, other groups have, via very different approaches, suggested guanylate kinase could indeed be a drug target for COVID-19 [12].

In the next 18 months, we plan to update this pipeline using the flood of new SARS-CoV-2 data and also to test key candidates using biophysical and biochemical assays for binding. Validated interaction points between SARS-CoV-2 and human proteins will provide possible novel points for therapeutic intervention, potentially (e.g. as for guanylate kinase) with compounds already available in the clinic.

Overall, this study demonstrates that existing techniques and datasets, when organized into a suitable pipeline and coupled to sufficient compute resources, can provide new insights into a system like SARS-CoV-2 in a relatively short time. We anticipate that our approach can be applied to other human pathogens, or indeed the wider set of acute clinical issues, where mechanistic insights are critical for the discovery of new points of therapeutic intervention.



**FIGURE 1:** Schematic of pipeline to identify the candidate SARS-CoV-2 SLiMs; interacting human proteins.

**REFERENCES:** [1] Science 2020;eabe9403(2020). DOI:10.1126/science.abe9403. [2] Rev. Microbiol. 2009;7(11):787-97. DOI: 10.1038/nrmicro2222. [3] Curr. Opin. Struct. Biol. Apr 2015;32:91-101. DOI: https://doi.org/10.1016/j.sbi.2015.03.004. [4] Mol. Biosyst. 2015;11:2821-2829. DOI: https://doi.org/10.1039/C5MB00301F. [5] Nucleic Acids Res. 2012;40:D242-51. DOI: 10.1093/nar/gkr1064. [6] J. Mol. Biol. 2018;430:2237-2243. DOI: 10.1016/j.jmb.2017.12.007. [7] Proc Natl Acad Sci U S A. 2002;99:5896-5901. DOI: 10.1073/pnas.092147999. [8] Cell. 2007;129(7):P1415-1426. DOI: 10.1016/j.cell.2007.05.052. [9] PLOS Pathog. 2020;16:e1008392. DOI: 10.1371/journal.ppat.1008392. [10] Science 1992;258:1862-1865. DOI: 10.1126/science.1361684. [11] Med. Res. Rev. 2009;29:683-741. DOI: 10.1002/med.20151. [12] Zenodo 2020;V3:1-11. DOI: 10.5281/zenodo.3752641.

**AUTHORS:** Gurdeep Singh<sup>1</sup>, Manjeet Kumar<sup>2</sup>, Toby J Gibson<sup>2</sup> and Robert B Russell<sup>1</sup>  
<sup>1</sup> BioQuant & BZH, Heidelberg University, Im Neuenheimer Feld 267, 69121 Heidelberg  
<sup>2</sup> EMBL Heidelberg, Meyerhofstrasse 1, 69118 Heidelberg



# FRIEND OR FOE? SARS-CoV-2 HIJACKS HOST BIOLOGY TO LETHAL EFFECT

In less than a year COVID-19 has caused over 1 million deaths, making it vital to better understand the biological mechanisms that enable the infection and disease progression. We exploit single cell RNA sequencing to identify target cells of the virus, describe how over-activation of the immune system furthers disease progression, and provide community access to this data *via* the cloud.

The COVID-19 pandemic has had widespread social, healthcare, and economic impact. It has been 100 years since a similar event of comparable scale, the Spanish flu, and just over a decade since the previous outbreak of the SARS coronavirus that was limited to far east Asia. The COVID-19 epidemic has had a dramatic effect on many lives, and as such the global scientific community has mobilized to unravel the mechanisms driving the disease to identify better strategies for clinical management of patients and ultimately the development of a preventative vaccine. One such major effort was led by Professor Roland Eils, the founding director of the Center for Digital Health at BIH and Charité. Working within a multidisciplinary team consisting of physicians, biologists, epidemiologists, bioinformaticians and mathematicians, they identified which cells can be infected and analyzed the host response to SARS-CoV-2 infection at the single-cell level. Together they elucidated surprising findings of how the virus takes advantage of the host immune system to lethal effect.

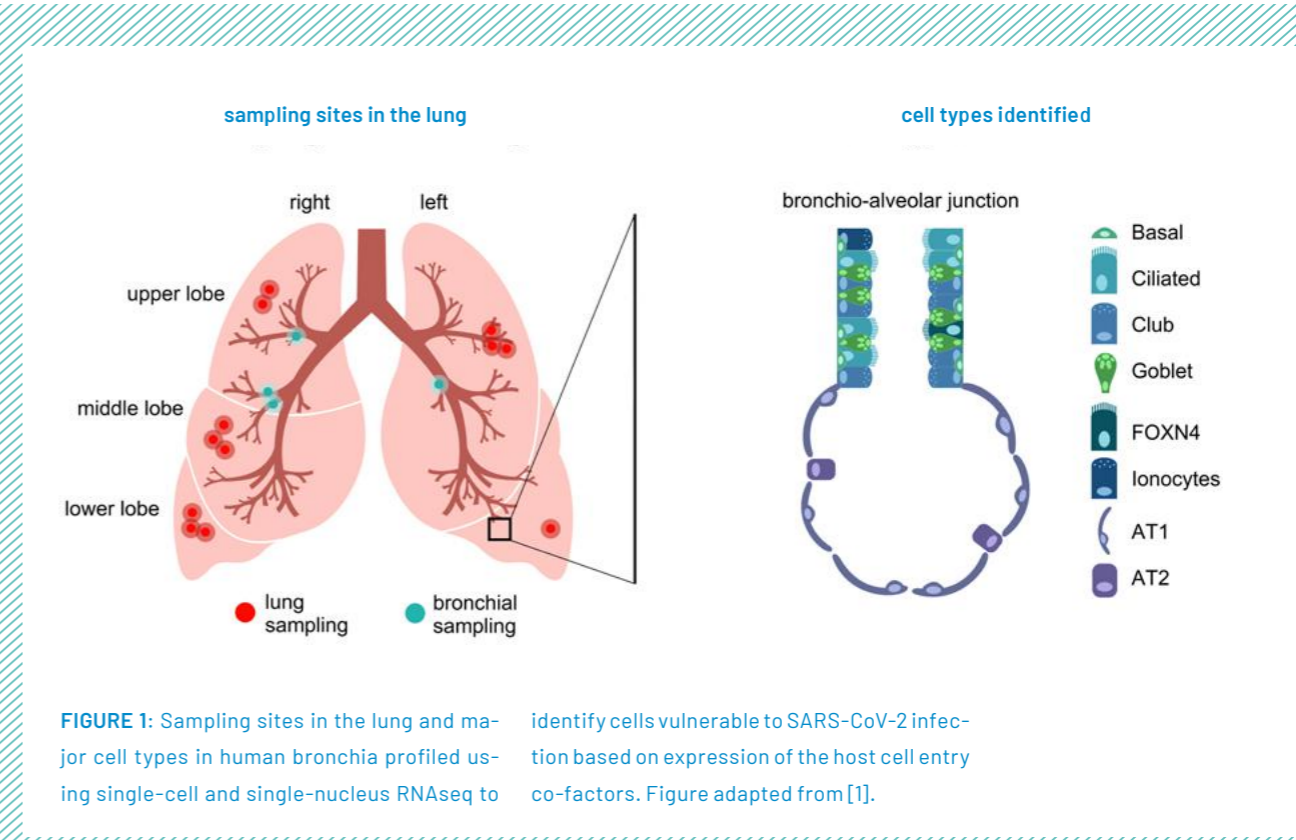
## **ONLY A SUBSET OF LUNG CELLS ARE POTENTIAL TARGETS OF SARS-CoV-2**

The human body contains many thousands of cell types, not all of which are targeted by SARS-CoV-2. The ability of the virus to infect cells is determined by receptor proteins and other co-factors on the surface of cells that the virus 'hijacks' to enter and infect cells. Some of these co-factors facilitating host cell entry are well characterised, such as the cell surface receptor ACE2 or the protease TMPRSS2. In order to identify which cells in the lung could in principle be infected by SARS-CoV-2, a team of scientists operating under the German Center for Lung Research (DZL) from the Charité, BIH and Thorax Clinic at the University Hospital of Heidelberg examined nearly 60,000 individual cells from healthy lung tissue and cells from the subsegmental bronchial branches of the lung cultured in an Air-Liquid Interface (Figure 1). They found that ACE2 and TMPRSS2 transcripts were expressed in very few cells, which were determined to be transient secretory cells that were intermediates between the well-described goblet and ciliated cells.

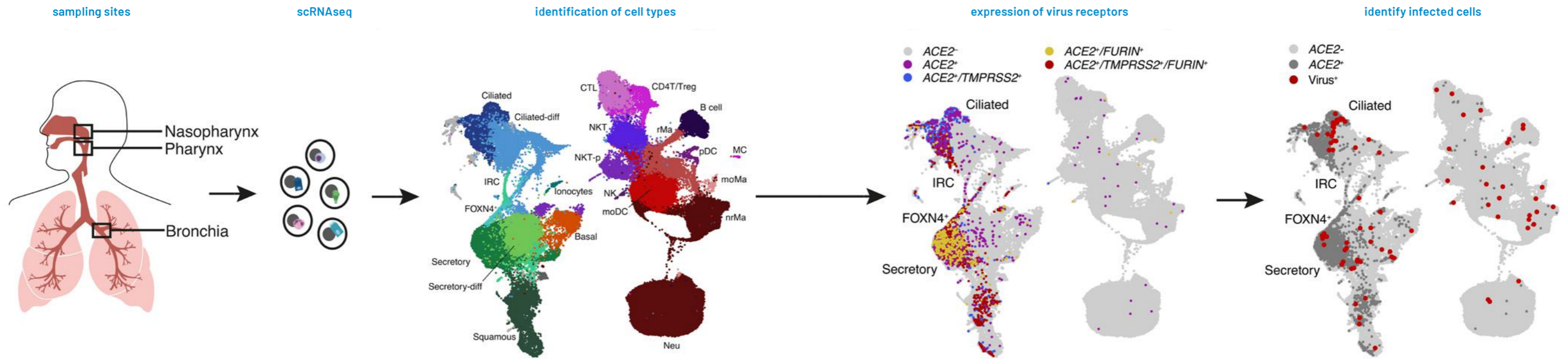
**OVER-ACTIVATION OF IMMUNE RESPONSE ACCELERATES SARS-CoV-2 INFECTION AND TISSUE DAMAGE**

After determining the most vulnerable cells to infection, the team investigated COVID-19 patient samples together with the Department of Virology at Charité and the Clinic for Anaesthesiology and Intensive Therapy at University Clinic Leipzig. The team collected patient samples from the upper airway by nasopharyngeal swabs. These samples represent the first point of contact with SARS-CoV-2 before it descends into the lungs. With this data they analysed ~160,000 cells obtained from both COVID-19 patients and SARS-CoV-2-negative donors to identify the major epithelial and immune cell components of the samples, and

comprehensively analysed how the abundance and transcriptional patterns of these cells are perturbed upon infection. They found that epithelial cells showed an average three-fold increase in expression of the SARS-CoV-2 entry receptor ACE2, which correlated with interferon signals by immune cells (Figure 2). Furthermore, the epithelial cells with the most virus positive cells were ciliated and secretory cells, which also exhibited ACE2 expression. They also found a population of interferon gamma responsive cells that were undergoing a 'differentiation short-cut' to become ciliated cells in response to the infection. This essentially 'added fuel to the fire' since ciliated cells are one of the prime targets of SARS-CoV-2!



**FIGURE 1:** Sampling sites in the lung and major cell types in human bronchia profiled using single-cell and single-nucleus RNAseq to identify cells vulnerable to SARS-CoV-2 infection based on expression of the host cell entry co-factors. Figure adapted from [1].



**FIGURE 2:** Dissection of the transcriptional landscape of cell types and states in the upper airways identifying secretory and ciliated cells to be targets of SARS-CoV-2 infection. Figure adapted from [2].

They also observed increased levels of immune-epithelial signalling in critically affected patients. The hyper-activation of the chemokine and cytokine signalling pathways were postulated to cause the major destruction of lung tissues during infection. In critical COVID-19 cases, they observed induction of CCL2 and CCL3 in macrophages alongside increased expression of their receptor CCR1 - this binding can induce monocyte recruitment into the lung followed by both recruitment and activation of more immune cells and further epithelial damage.

#### CLINICAL IMPACT

The authors further describe possible therapeutic exploitation of these results. While targeting the virus and virus receptors are the obvious targets, they hypothesize that it is the hyper-activation of the immune system that causes the severe and critical COVID-19 responses that may lead to fatalities. At the time there were already efforts focussing on certain aspects of the host immune response and this study elucidates that targeting of certain chemokine receptors is promising.

#### ACCELERATING RESEARCH IN THE COMMUNITY

While the scientific and medical conclusions drawn from the study were of extreme importance, sharing the data generated in such studies is paramount - it enables further in-depth analysis, meta-analysis, test hypotheses and generate new ones. As such, with the de.NBI

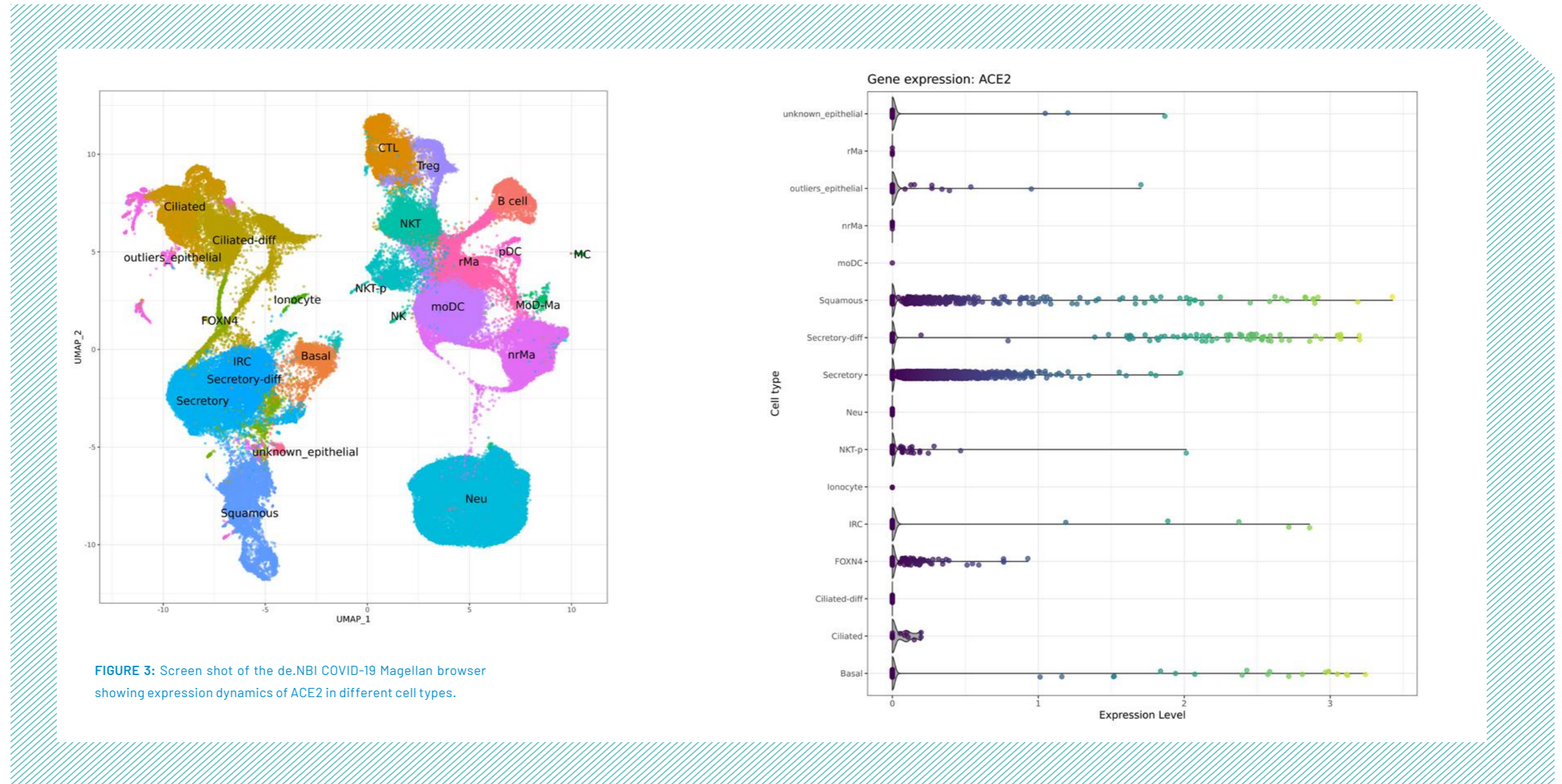
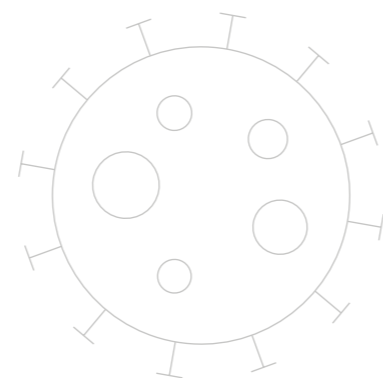


FIGURE 3: Screen shot of the de.NBI COVID-19 Magellan browser showing expression dynamics of ACE2 in different cell types.



cloud team in Berlin, they created an interactive data browser based on Magellan with the goal of providing free access to this valuable data resource to the scientific and medical community. The browser is accessible via <https://digital.bihealth.org>, with announcements of updates available on <http://www.hidih.org/projects/covid-19-sars-cov-2>.

**REFERENCES:** [1] EMBO Journal 2020;39:e105114. DOI: 10.15252/emj.20105114. [2] Nature Biotechnology 2020;38:970-979. DOI: <https://doi.org/10.1038/s41587-020-0602-4>.

**AUTHORS:** Naveed Ishaque<sup>1</sup>, Bianca Henning<sup>1</sup>, Robert Lorenz Chua<sup>1</sup>, Soeren Lukassen<sup>1</sup>, Sven Olaf Twardziok<sup>1</sup>, Juergen Eils<sup>1</sup>, Christian Conrad<sup>1</sup>, Roland Eils<sup>1</sup>  
<sup>1</sup> Center for Digital Health, Berlin Institute of Health (BIH) and Charité - Universitätsmedizin Berlin; HD-HuB, Anna-Louisa-Karsch-Straße 2, 10178 Berlin

```
mirror_mod = modifier_ob.modifiers.new("mirror")
mirror_ob.mirror_object = mirror_ob
mirror_mod.operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
mirror_mod.operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
mirror_mod.operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True

selection at the end -add back the des
mirror_ob.select= 1
mirror_ob.select=1
key.context.scene.objects.active = modifier_ob
key.context.selected + str(modifier_ob) # modi
mirror_ob.select = 0
key.context.selected_objects[0]
key.context.objects[one.name].select = 1

print("please select exactly two objects,")

OPERATOR CLASSES -----
class Operator(Operator):
    def mirror_to_the_selected_object(self):
        key.context.mirror_mirror_x
        mirror_x

    def mirror_to_the_selected_object(self, context):
        if key.context.active_object is not None
```

# BIOINFORMATICS TOOLS FOR ANALYZING COVID-19 DATA

Development of novel bioinformatics solutions to study coronaviruses as well as careful adaptation of existing tools provides answers to a wide range of biological, medical, and epidemiological questions about SARS-CoV-2 and COVID-19. In the future, bioinformatics will take a key role providing researchers the opportunity to better understand the biology of the virus.

# ANALYZING SARS-CoV-2 CHROMATOGRAMS AND MUTATIONS USING THE GENOME ANALYSIS SERVER GEAR

The GEAR genome analysis web server ([www.gear-genomics.com](http://www.gear-genomics.com)), offered through de.NBI via the European Molecular Biology Laboratory (EMBL) in Heidelberg, has been extended in early March 2020 to help address the global COVID-19 health crisis. This article describes how these extensions enable analyses of SARS-CoV-2 chromatogram traces as well as viral mutations.

## GEAR: A GENOME ANALYSIS WEB SERVER

One of the missions of the Genomics Core Facility (GeneCore) and of the Korb group at EMBL Heidelberg is the development of key methods for the analysis of genomic variation – these efforts have led to tools widely used in genetics communities (e.g. [1], [2] and [3]). To complement these command-line tools geared towards bioinformaticians, we launched in 2017 a genome analysis web server (GEAR, <https://www.gear-genomics.com/>, [4]), with online applications to facilitate the use of different DNA sequence-oriented bioinformatics tools for molecular biologists.

GEAR hosts a multitude of applications for wet lab scientists including to analyse

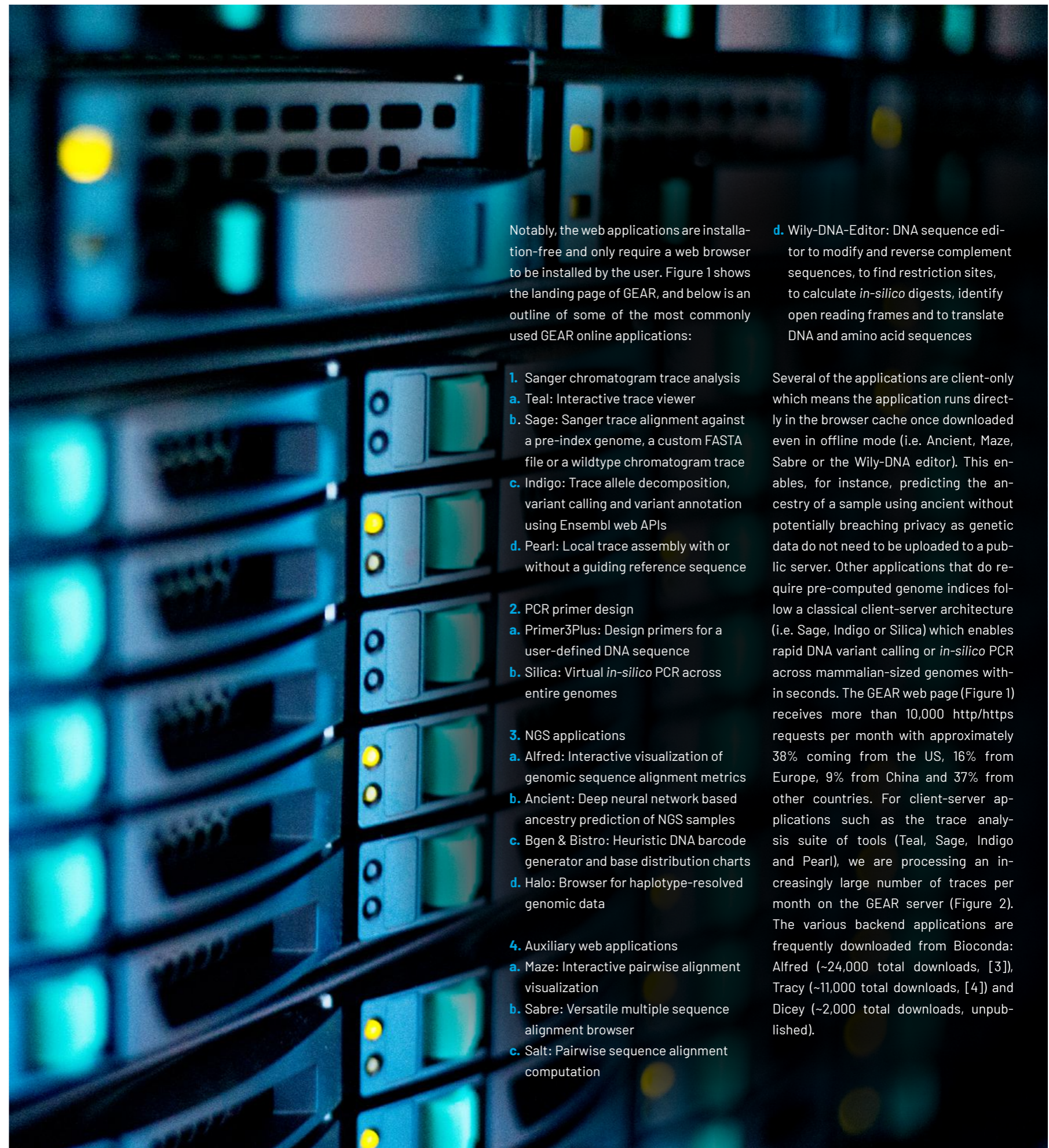
Sanger traces, discover DNA variants, verify engineered CRISPR/Cas9 mutations, design PCR primers and perform DNA sequence analyses (<https://www.gear-genomics.com>). These openly accessible web applications are part of EMBL's de.NBI/ELIXIR-DE offerings, and the web applications are actively developed and maintained by members of GeneCore and the Korb group. All code is open-source and freely available in the GEAR software repositories (<https://github.com/gear-genomics>), including the code for the web server front-end applications and the backend genomics analysis software. Among other command-line applications, the backend code includes Tracy [4], a method to rapidly analyze Sanger chromatogram traces, which is available on Bioconda [5], as a statically linked binary or as a Docker/Singularity container.

Notably, the web applications are installation-free and only require a web browser to be installed by the user. Figure 1 shows the landing page of GEAR, and below is an outline of some of the most commonly used GEAR online applications:

1. Sanger chromatogram trace analysis
  - a. Teal: Interactive trace viewer
  - b. Sage: Sanger trace alignment against a pre-index genome, a custom FASTA file or a wildtype chromatogram trace
  - c. Indigo: Trace allele decomposition, variant calling and variant annotation using Ensembl web APIs
  - d. Pearl: Local trace assembly with or without a guiding reference sequence
2. PCR primer design
  - a. Primer3Plus: Design primers for a user-defined DNA sequence
  - b. Silica: Virtual *in-silico* PCR across entire genomes
3. NGS applications
  - a. Alfred: Interactive visualization of genomic sequence alignment metrics
  - b. Ancient: Deep neural network based ancestry prediction of NGS samples
  - c. Bgen & Bistro: Heuristic DNA barcode generator and base distribution charts
  - d. Halo: Browser for haplotype-resolved genomic data
4. Auxiliary web applications
  - a. Maze: Interactive pairwise alignment visualization
  - b. Sabre: Versatile multiple sequence alignment browser
  - c. Salt: Pairwise sequence alignment computation

d. Wily-DNA-Editor: DNA sequence editor to modify and reverse complement sequences, to find restriction sites, to calculate *in-silico* digests, identify open reading frames and to translate DNA and amino acid sequences

Several of the applications are client-only which means the application runs directly in the browser cache once downloaded even in offline mode (i.e. Ancient, Maze, Sabre or the Wily-DNA editor). This enables, for instance, predicting the ancestry of a sample using ancient without potentially breaching privacy as genetic data do not need to be uploaded to a public server. Other applications that do require pre-computed genome indices follow a classical client-server architecture (i.e. Sage, Indigo or Silica) which enables rapid DNA variant calling or *in-silico* PCR across mammalian-sized genomes within seconds. The GEAR web page (Figure 1) receives more than 10,000 http/https requests per month with approximately 38% coming from the US, 16% from Europe, 9% from China and 37% from other countries. For client-server applications such as the trace analysis suite of tools (Teal, Sage, Indigo and Pearl), we are processing an increasingly large number of traces per month on the GEAR server (Figure 2). The various backend applications are frequently downloaded from Bioconda: Alfred (~24,000 total downloads, [3]), Tracy (~11,000 total downloads, [4]) and Dicey (~2,000 total downloads, unpublished).



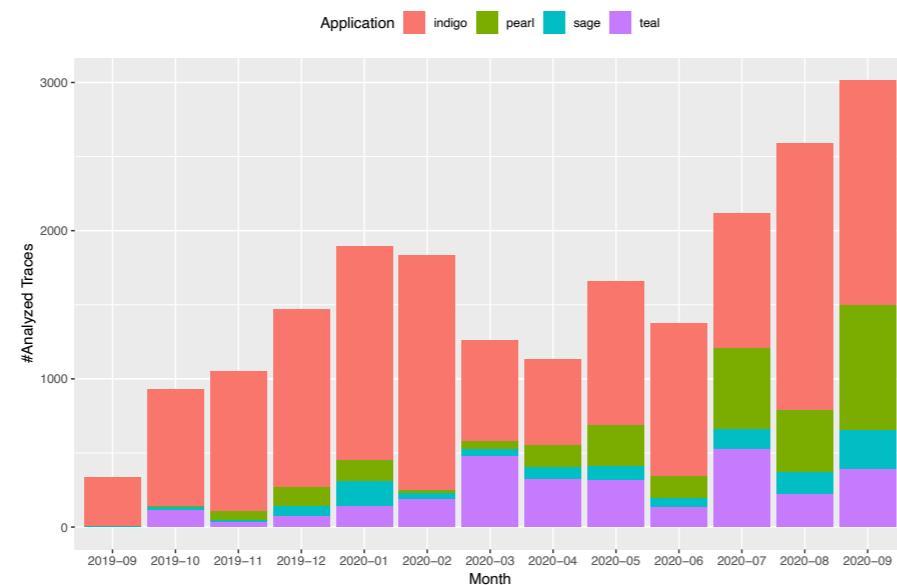


genome analysis server



FIGURE 1: The GEAR homepage with one tile for each online application. Each tile shows the name of the application and a short description of the tool. The header bar includes links to the contact form, the terms of use and the GEAR code repository.

FIGURE 2: Usage graph of the chromatogram trace analysis applications available on GEAR. Each bar shows the number of uploaded and processed chromatogram traces per month.



### RECENT DEVELOPMENTS DURING THE COVID-19 PANDEMIC

To help address the global COVID-19 health crisis, GEAR was extended in early March 2020 to enable direct analyses of SARS-CoV-2 chromatogram trace files. We incorporated the SARS-CoV-2 reference genome to facilitate viral genome related analyses including primer design and SARS-CoV-2 mutation calling. In particular, Indigo can provide utility in this context through its ability to decompose more complex SARS-CoV-2 mutations, such as insertions or deletions. Indigo calls variants with respect to a reference genome and hyperlinks all variants to the original position in the chromatogram trace, which allows users to quickly ex-

plore potentially noisy chromatograms arising from mixtures of viral SARS-CoV-2 genomes. The application also features an alignment of the decomposed alleles in Figure 3. The backend software of Indigo can be used on the command-line to generate standard variant call format files (VCF), which allows the rapid analysis of hundreds of SARS-CoV-2 chromatogram trace files in batch.

Another noteworthy application, in the context of the COVID-19 health crisis, is Pearl, a bioinformatics tool enabling reference-guided assembly of multiple trace files of viral origin. Pearl aligns traces to a specified reference sequence (or an assembled consensus sequence), and

then conveniently highlights conflicts and mismatches between the reference and all uploaded chromatogram trace files (Figure 4). This enables a direct identification of (1) clonal mismatches, colored red, where all traces disagree with the reference, (2) sub-clonal mismatches, colored orange, where some traces disagree, (3) clonal matches, colored green, where all traces agree with the reference and (4) sites without trace data, colored grey. By design, Pearl then allows the user to quickly browse all conflicts (potential SARS-CoV-2 mutations) and users can interactively edit the SARS-CoV-2 reference sequence to establish the 'personalized', sequenced SARS-CoV-2 viral genome that can be exported in FASTA format.

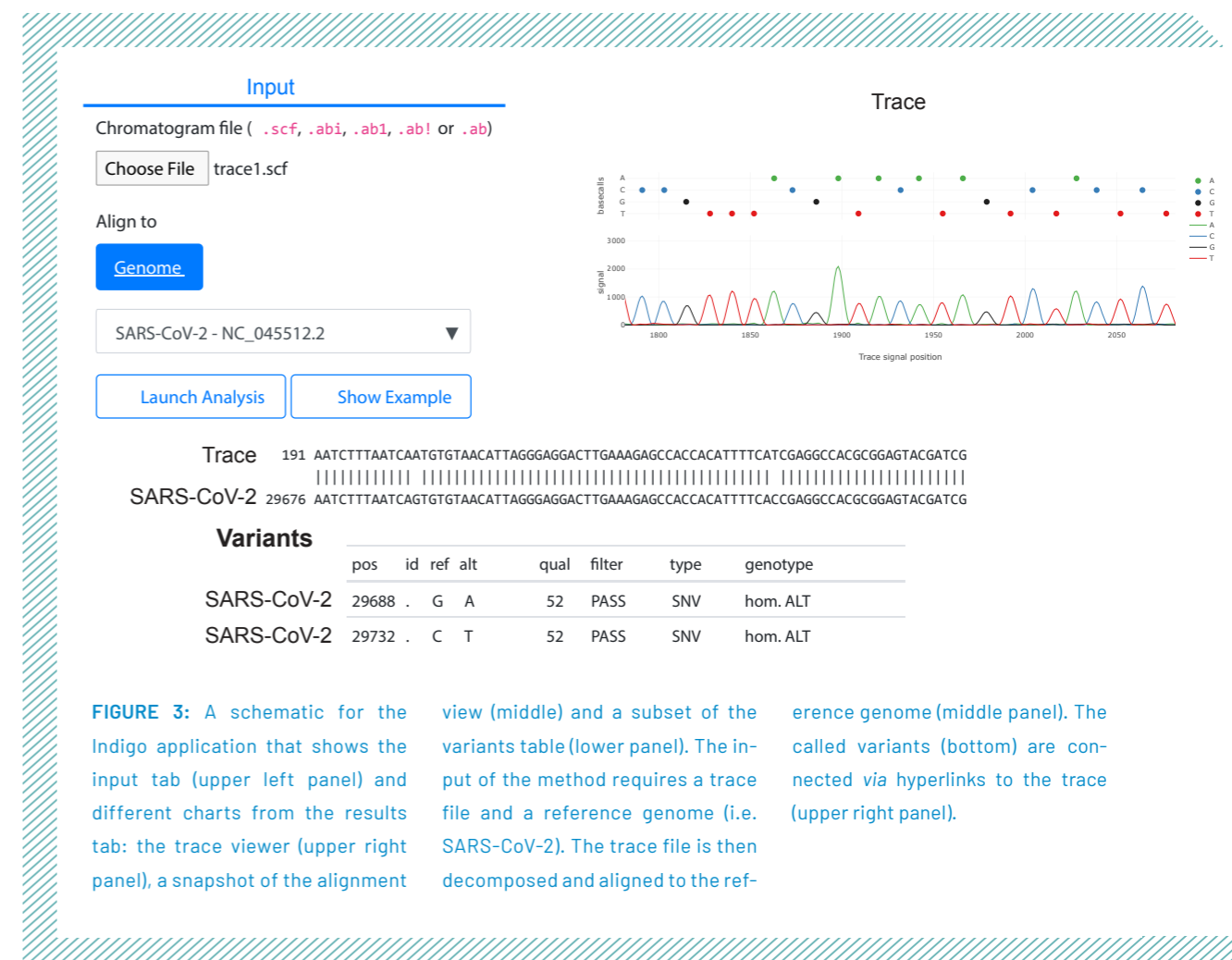


FIGURE 3: A schematic for the Indigo application that shows the input tab (upper left panel) and different charts from the results tab: the trace viewer (upper right panel), a snapshot of the alignment

view (middle) and a subset of the variants table (lower panel). The input of the method requires a trace file and a reference genome (i.e. SARS-CoV-2). The trace file is then decomposed and aligned to the ref-

erence genome (middle panel). The called variants (bottom) are connected via hyperlinks to the trace (upper right panel).







# COMPARATIVE GENOME AND METAGENOME ANALYSIS OF CORONAVIRUS-POSITIVE CLINICAL SAMPLES

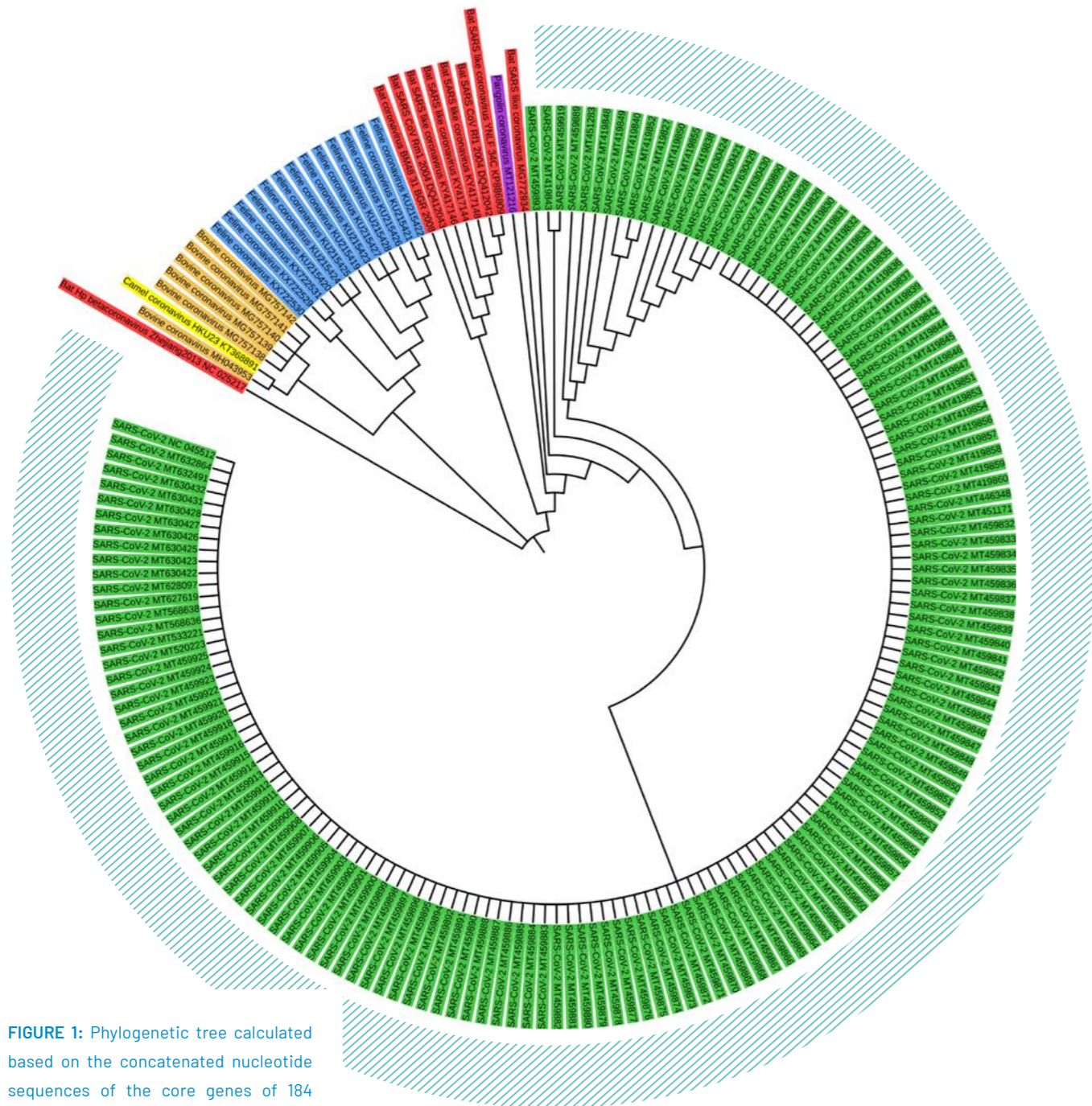
Modern sequencing technologies have enabled fast and efficient generation of sequencing data that has subsequently also led to the development of a variety of specialized bioinformatics tools for detailed data analysis and visualization. We have developed tools like EDGAR and MGX to specifically analyze microbial genomes and metagenomes and have refined these tools to also meet the needs of viral data.

Technical advances in next-generation sequencing (NGS) technologies and specialized bioinformatics solutions for data processing and evaluation have enabled cost-effective high-throughput experiments to address a broad range of biological questions. Over recent years we have successfully used these approaches to access a detailed snapshot of the global genomic, metagenomic or transcriptomic landscape of organisms from different kingdoms and taxa. NGS technologies are also applicable to analyze viral genomes. With the establishment of many viral data sets, it becomes feasible to implement novel approaches that enable comparative views and broader perspectives to better understand the relevance and the role of DNA sequence variations in coding and non-coding regions with respect to the development and progression of a pandemic such as the coronavirus disease (COVID-19).

COVID-19 is an ongoing pandemic first reported in December 2019 in Wuhan, caused by severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2), a novel betacoronavirus. To date (11.01.2021) there are more than 88 million confirmed infections worldwide with 1.921.024 confirmed cases in Germany. The most common clinical symptoms of COVID-19 are fever, fatigue, cough, and dyspnea. While in most cases, the severity of disease is low and the patients recovered, approximately 5% of patients required intensive care. Many of those patients required mechanical ventilation due to acute respiratory distress syndrome (ARDS) as most common complication.

## EDGAR – COMPARATIVE GENOME ANALYSIS FOR MICROBIAL AND VIRAL GENOMES

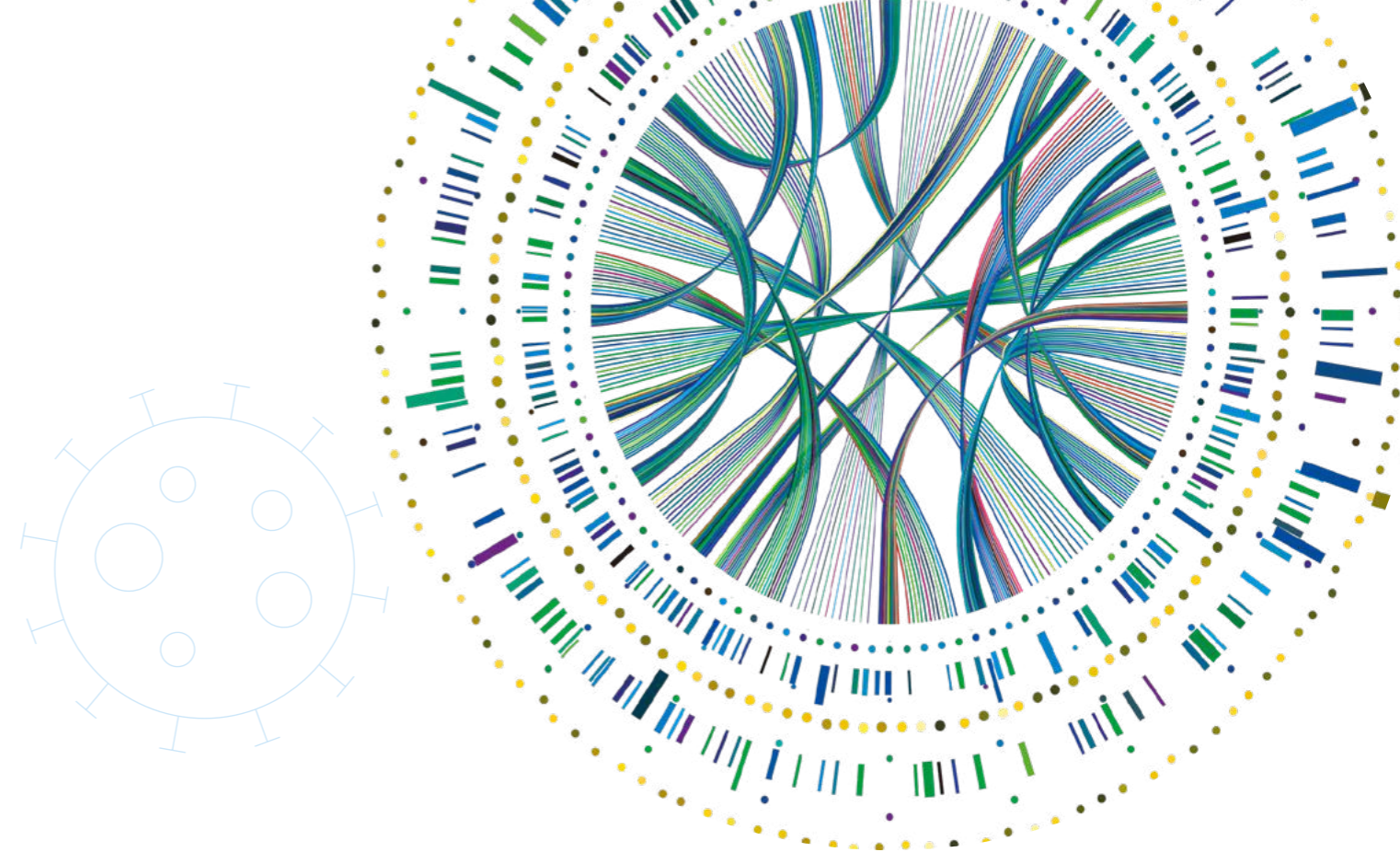
One of the standard tools in bacterial comparative genome analysis and phylogenomics is EDGAR [1]. EDGAR in-



**FIGURE 1:** Phylogenetic tree calculated based on the concatenated nucleotide sequences of the core genes of 184 coronavirus genomes calculated with EDGAR. Strains were of human (green), bat (red), bovine (orange), feline (blue), camel (yellow), and pangolin (purple) origin. The approximate maximum likelihood algorithm FastTree 2.1 was used to construct the phylogeny. Genomes were initially published by [7].

cludes numerous analysis and visualization functions such as the calculation of common and species-specific gene sets, Venn diagrams to display gene sets, circular genome plots or multiple synteny plots. A special focus of the software is on phylogenomics. The web-based software provides access to a wide range of tools for analyzing the lineage and the taxonomic classification of bacterial species. In particular, EDGAR provides methods for calculating lineage trees and genome-to-genome distances such as Average Nucleotide Identity (ANI) or Average Amino Acid Identity (AAI).

The software tool EDGAR and the implemented methods are freely available to scientists as precalculated projects in a public database. There are currently 322 genera with a total of 8,079 genomes accessible in EDGAR. Furthermore, another 226 projects with 4,400 genomes are available in which the type strains of taxonomic families can be analyzed. For custom analyses or unpublished data, more than 1,100 private projects with more than 44,000 genomes were analyzed over the years.



### EDGAR enables to analyze unpublished data in password-protected and private projects.

EDGAR was initially developed for the comparison of bacterial genomes, but it can also be easily used for the analysis of viral genomes as there is also high demand to analyze viral samples (Figure 1). Accordingly, EDGAR is already being used for the analysis of coronaviruses, e.g. in cooperation with the Pukyong National University (PKNU) in Busan in South Korea. In this project, more than 150 coronavirus genomes with different host specificities (human, feline, canine, equine, murine) are analyzed and compared among each other and with genomes of other virus genera.

### FLEXIBLE METAGENOME ANALYSIS USING THE MGX FRAMEWORK – AN EXTENSION FOR VIRUSES

The characterization of microbial communities based on sequencing and analysis of their genetic information has become a popular approach also referred

to as metagenomics; in particular, the recent advances in sequencing technologies have enabled researchers to study even the most complex communities. Metagenome analysis, the assignment of sequences to taxonomic and functional entities, however, remains a computationally exhaustive task. The MGX framework is a graphical solution for the management and processing of such metagenome datasets, featuring a wide range of reproducible workflows for taxonomic classification as well as functional assignment [2](Figure 2).

Employing MGX, researchers are also able to process viral metagenomes and metatranscriptomes in order to identify DNA or RNA virus fragments, or assign viral gene fragments to specialized protein databases such as the ‘Reference Viral DataBase’ (RVDB) or vFAM, the HMMER3 database of profile hidden Markov models (HMMs) built from all the viral proteins present in RefSeq. Moreover, annotated viral genomes can be imported into the MGX application and can be used as a target sequence to perform reference mappings and to create fragment recruitments.

### USING NEXT-GENERATION SEQUENCING TO IDENTIFY GENOMIC MUTATIONS WITHIN SARS-CoV-2 GENOMES AND TO MONITOR MICROBIAL COMMUNITY SHIFTS AFTER SARS-CoV-2 INFECTIONS IN HUMANS

Over the past years, we have applied NGS technologies to generate detailed information on viral and bacterial genome diversity, microbial community characterization, and transcriptomic profiling of coinfecting hosts with their viral and bacterial pathogens or commensals. To study viral pathogens on the genomic level, we have determined the genome sequences of e.g. Ebola virus, Lassa virus and feline coronavirus using RNA genome sequencing, which allows the precise determination of SNPs in viral genomes [3-5].

In the current pandemic, we also embarked on sequencing and analyzing SARS-CoV-2 genomes [6] isolated from COVID-19 patients using NGS technologies. This will allow us to characterize the complete genome sequence, focusing on discriminative mutations among patient isolates. In the long run, SARS-CoV-2 sequence variation will provide insight

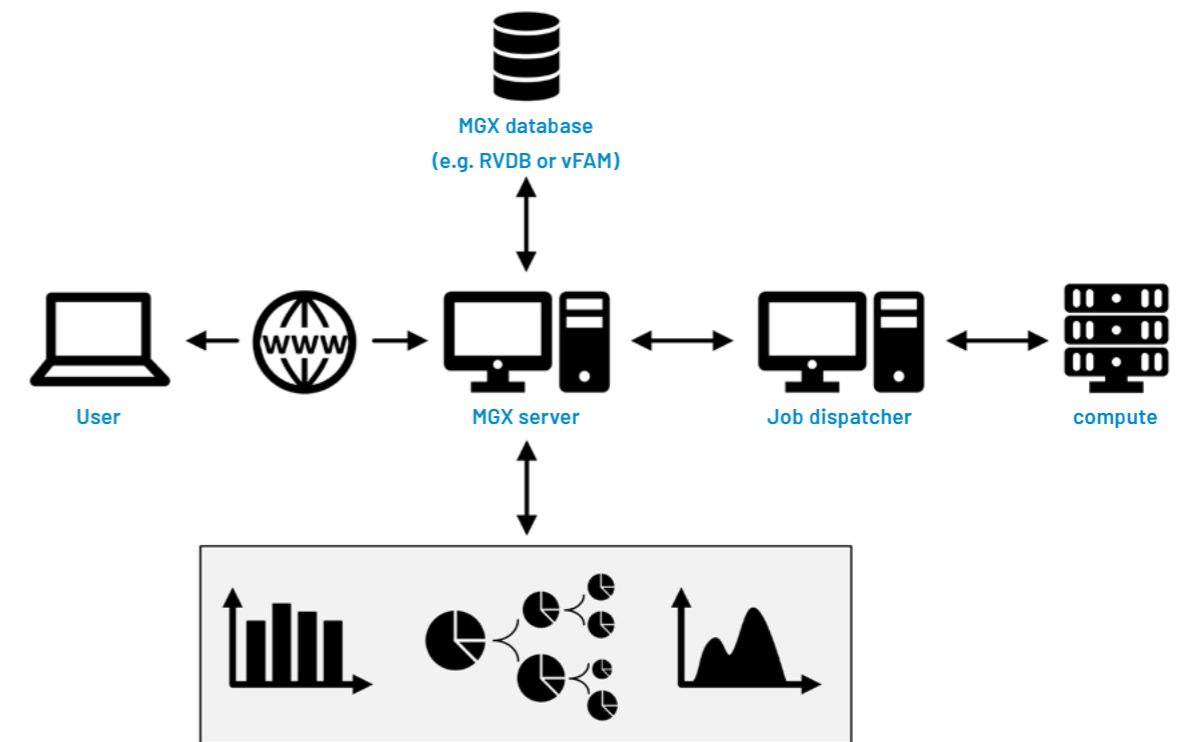
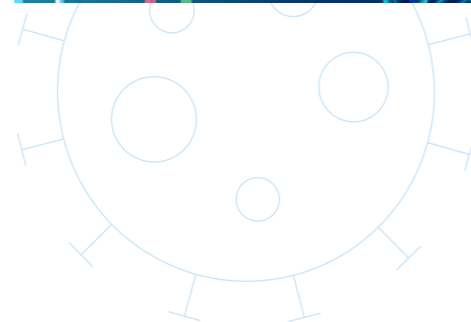
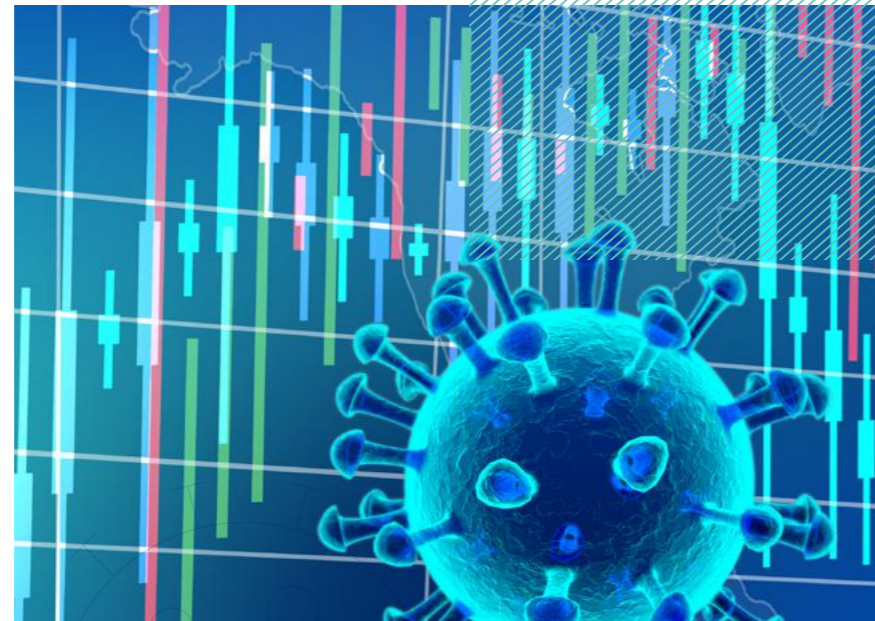
into immune escape mechanisms, viral pathogenesis, resistance to antiviral drugs (if they become available) and it will generate important information for the development of effective vaccines.

Besides genome-based studies we also investigate possible effects of coronavirus infections on the microbiome of the respiratory tract of children and adults. All human coronaviruses (including low-pathogenic 'common cold' coronaviruses) effectively suppress the activation of the host innate immune response, thereby possibly affecting the microbiome and facilitating viral and/or bacterial co- and superinfections. Shifts in the microbial composition are usually monitored by amplicon short read sequencing (e.g. 150 bp) targeting short fragments of the 16S rRNA gene. While this is a well-established and cost-effective method it is limited in taxonomic resolution. To gain a deeper insight into the changes of the microbial composition during SARS-CoV-2 infection, we will apply 'long-read' amplicon sequencing of the complete bacterial rRNA operon with a length of 4,500 bp which will provide genus and species level taxonomic classification. It is currently unknown if (and to what extent) the microbiome modulates coronavirus pathogenesis or affects the predisposition to co-infections or superinfections. Likewise, potential consequences of genomic differences between clinical isolates of alpha- and betacoronaviruses and cell culture-adapted laboratory strains have not been studied in any detail. In addition to the above-mentioned coronaviruses, the study is going to be extended to SARS-CoV-2.

The following major questions will be addressed: I. Do coronavirus infections lead to dysbiosis in the respiratory tract in children and adults? II. How do the different coronaviruses affect the microbiome and can potential differences be

linked to specific genetic loci? Do different coinfections or superinfections occur depending on the respective coronavirus species involved? III. Are there differences between viral mono-infections and multiple infections?

The project will also help answer the question of whether dysbiosis due to a pre-existing chronic lung disease affects the susceptibility to (and severity of) a SARS-CoV-2 infection and whether SARS-CoV-2-induced changes of the airway microbiome of these patients influence short- and long-term disease progression and outcome. To address these questions, various large-scale data sets will be generated that demand for compute sources, data management, specialized tools and bioinformatics support that will be provided by the de.NBI BIGI Service Center.



**FIGURE 2:** MGX system overview. In the MGX framework, each user connects to one or several MGX server instances. Sequence data and related metadata are deposited on the server. Analysis workflows are prioritized and scheduled to compute resources by a job dispatcher. The MGX database includes specific protein databases for viral data e.g. 'Reference Viral DataBase' (RVDB) or vFAM. Data is visualized by one click solutions offering different graphical outputs.

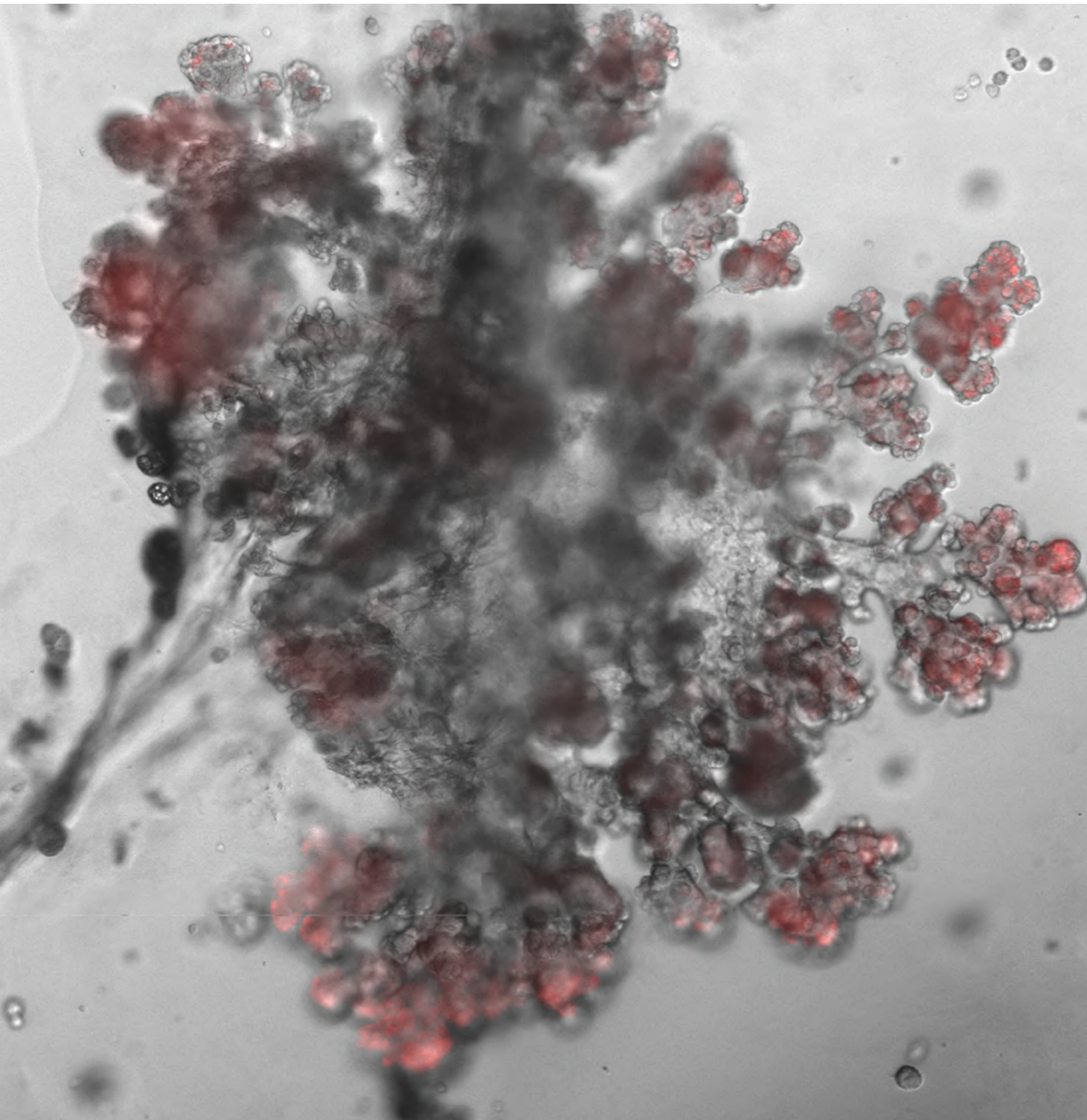
**REFERENCES:** [1] *Nucleic acids research* 2016;44(W1). DOI: 10.1093/nar/gkw255. [2] *Microbiome* 2018;6(1). DOI: 10.1186/s40168-018-0460-1. [3] *Genome Announcements* 2016;4(5). DOI: 10.1128/genomeA.01011-16. [4] *Genome Announcements* 2016;4(5). DOI: 10.1128/genomeA.00938-16. [5] *mBio* 2018;9(4). DOI: 10.1128/mBio.01422-18. [6] *bioRxiv* 2020. DOI: <https://doi.org/10.1101/2020.08.26.266304>. [7] *Genomics* 2020;112(6):4993-5004. DOI: 10.1016/j.ygeno.2020.09.014.

**AUTHORS:** Karina Brinkrolf<sup>1</sup>, Jochen Blom<sup>1</sup>, Sebastian Jaenicke<sup>1</sup>, Markus Weigel<sup>2</sup>, Jan Philipp Mengel<sup>2</sup>, Benjamin Ott<sup>2</sup>, Christian Schüttler<sup>3</sup>, Heiko Slanina<sup>3</sup>, Michael Kracht<sup>4</sup>, John Ziebuhr<sup>3,5</sup>, Torsten Hain<sup>2,5</sup> and Alexander Goesmann<sup>1,5</sup>  
<sup>1</sup> *Bioinformatics & Systems Biology, Justus Liebig University Giessen, Heinrich-Buff-Ring 58, 35392 Giessen*  
<sup>2</sup> *Medical Microbiology, Justus Liebig University Giessen, Schubertstraße 81, 35392 Giessen*  
<sup>3</sup> *Medical Virology, Justus Liebig University Giessen, Schubertstraße 81, 35392 Giessen*  
<sup>4</sup> *Rudolf Buchheim Institute of Pharmacology, Justus Liebig University Giessen, Schubertstraße 81, 35392 Giessen*  
<sup>5</sup> *German Center for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Schubertstraße 81, 35392 Giessen*

# VIRUS-INDUCED LUNG INJURY:

---

## Pathobiology and Novel Therapeutic Strategies



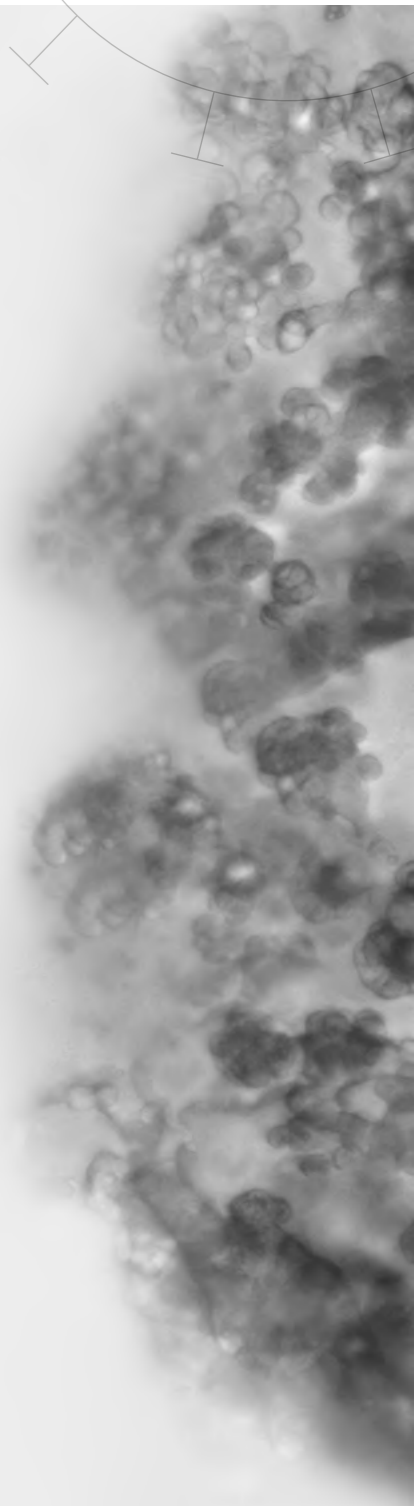
Single cell RNA sequencing has become a valuable method to study transcriptomic changes on the cellular level. It is also applicable to monitor effects of viral infection on eukaryotic cells. Within a clinical research unit, we are analyzing such effects in a virus-host system involving murine lung organoids. For evaluation of the resulting data, the user-friendly software WASP has been developed.

Acute infections of the lower respiratory tract represent an increasing problem of the public health worldwide and mortality rates remained unchanged over the past 50 years. Up to now, there is a lack of pharmacological treatments for the most devastating clinical course of pulmonary infection, acute lung injury (ALI) or acute respiratory distress syndrome (ARDS). Antiviral treatments are currently only available for influenza virus (IV) infection, but they are of very limited efficacy and bear the risk of rapid acquisition of viral resistance. In addition, there is increasing evidence that respiratory virus infection frequently predisposes for severe gram-positive bacterial superinfections. Emergence of novel respiratory viruses such as pH1N1/2009 pandemic IV and highly pathogenic avian IV strains, or the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and the recent out-

break of SARS-CoV-2 highlight the need for a better understanding of the underlying pathobiology of virus infections. Against this background, the Clinical Research Unit (CRU) program 'KF0309 - Virus-induced lung injury: pathobiology and novel therapeutic strategies' at Justus Liebig University Giessen is dedicated to unraveling the mechanisms driving anti-viral host defense to dissecting the cellular and molecular contributors to the tissue damage at the virus-host interface in the distal lung and to defining pathways and mediators of organ regeneration in the context of viral infection.

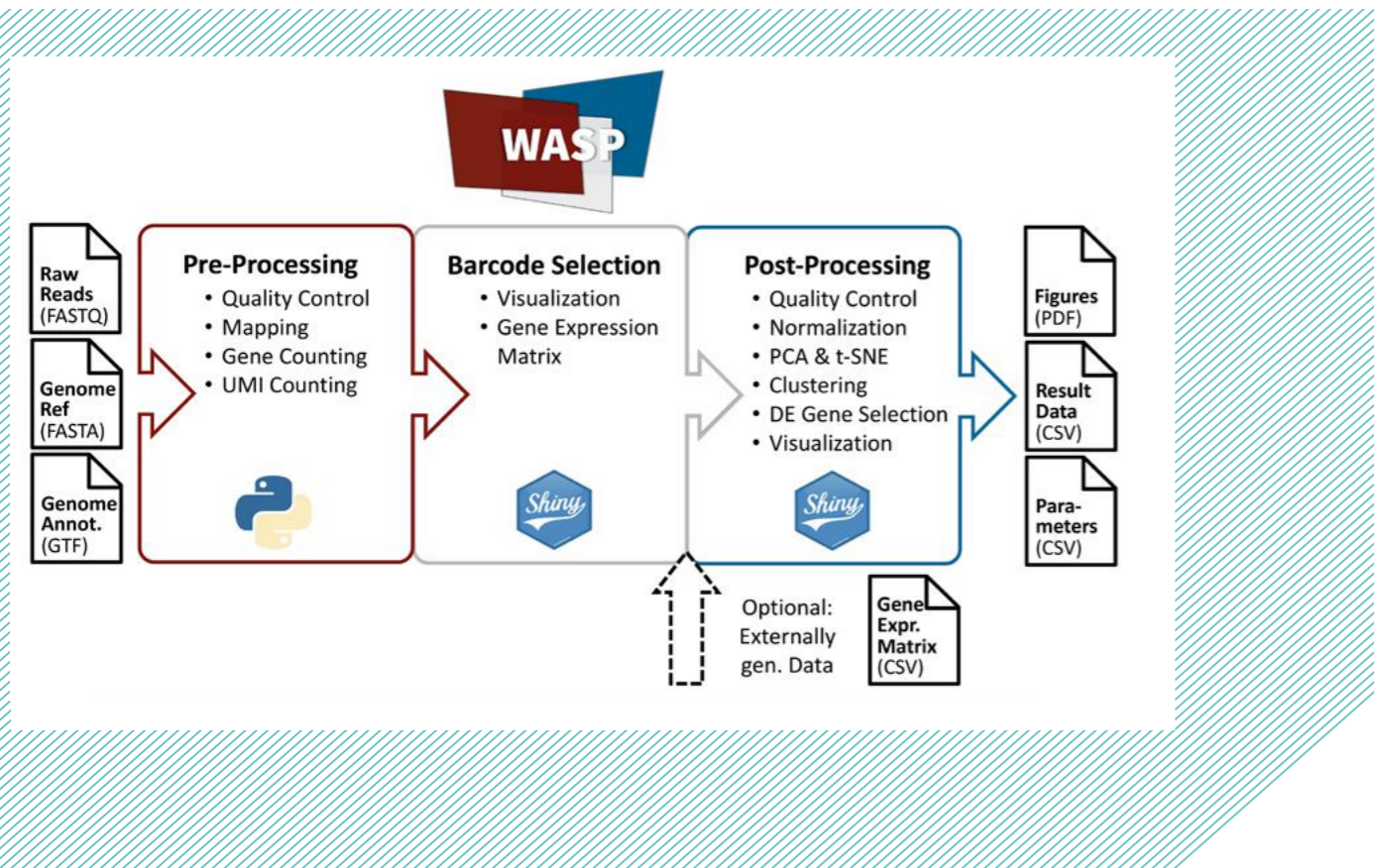
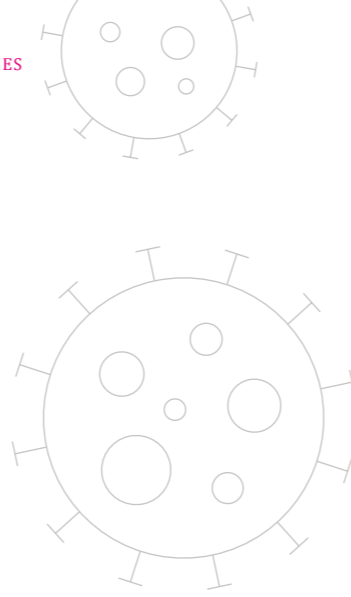
### PROCESSING SINGLE CELL RNA SEQUENCING DATA WITH WASP

In recent years, single cell RNA sequencing (scRNA-Seq) has become more and more popular to analyze transcriptomic



changes on the single cell level. It allows for an unprecedented insight into cellular heterogeneity and enables identification and characterization of cellular populations and sub-populations in detail. The technique of scRNA-Seq is applicable to a wide range of medical and biological research questions. However, new techniques in next-generation sequencing are usually accompanied by an additional need to adapt existing or implement new bioinformatics solutions for data analysis with help of tailored user-friendly software.

As part of the core unit (Z1) 'Genome signatures and integrated systems biology of pathogen-host interaction and cellular networks in the infected and injured lung' of the clinical research unit KF0309, scientists from the bioinformatics department at Justus Liebig University Giessen try to understand cell-specific gene regulatory events and their impact on cellular networks in infected lungs. They make use of high-throughput scRNA-Seq data from Drop-Seq-based protocols to analyze the transcriptional states in murine lung organoids and to identify specific



**FIGURE 1:** Schematic overview of a scRNA-Seq data analysis conducted with WASP. As a first step, users start the Snakemake-based workflow on a Linux-based system, providing a FASTQ file with the reads and a reference genome with corresponding annotation. The results (quality metrics) of

the workflow are then visualized in an R Shiny web application which allows users to select valid barcodes and subsequently generates a gene expression matrix file (CSV) containing counted UMIs per gene and cell barcode. Subsequently, the generated gene expression matrix or a similar externally generated

file is uploaded to the post-processing Shiny web application for further analysis. WASP enables users to perform the Post-processing in an automated or manual mode by using a dynamic web page similar to the pre-processing.

cell types. Murine lung organoids serve as a suitable model for respiratory viral disease modelling. In this context scientists from the Goesmann Lab in Giessen provide ready-to-use bioinformatics software tools, databases and tailored workflows. Furthermore, a user-friendly web-accessible software (WASP) is being developed that conducts Drop-Seq-based scRNA-Seq data analysis. The first steps (pre-processing) implemented in WASP facilitate the initial processing of raw reads generated with the ddSEQ protocol. These steps include an automated mapping to a reference genome, read quantification, processing of unique molecular identifiers (UMI), demultiplexing of the barcodes, quality control as interactive visualization and the generation of gene expression matrices based on filtered and selected barcodes. The software also includes more detailed biological analysis steps (post-processing) such as normalization, clustering, detection of differentially expressed genes and finally interactive visualizations of the results using machine learning algorithms like t-SNE and UMAP. The pre-processing pipeline of WASP is implemented as Snakemake workflow and the visualization steps are carried out with R Shiny applications. WASP is suitable for gene expression analysis, including detection of differentially expressed genes, clustering of cellular populations and interactive graphical visualizations. The R Shiny application (post-processing) can be used with gene expression matrices generated by the WASP pipeline, as well as with externally provided data from other sources.

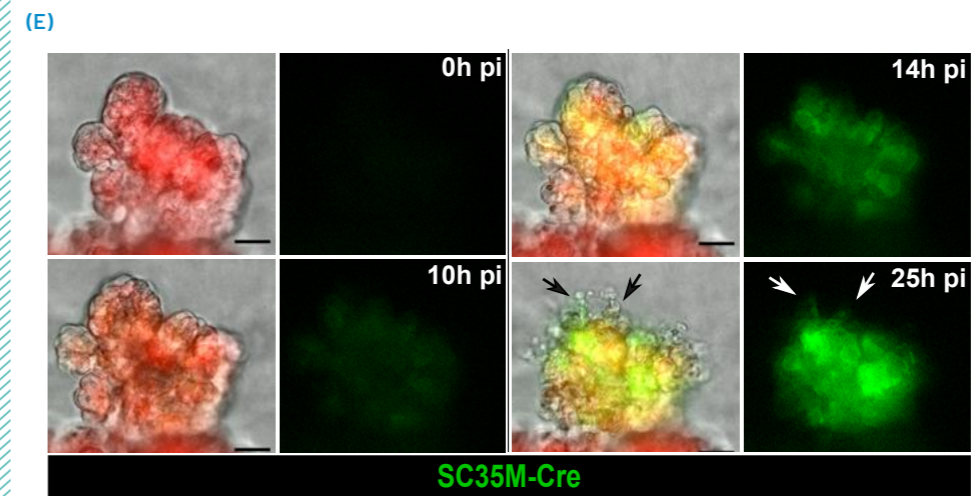
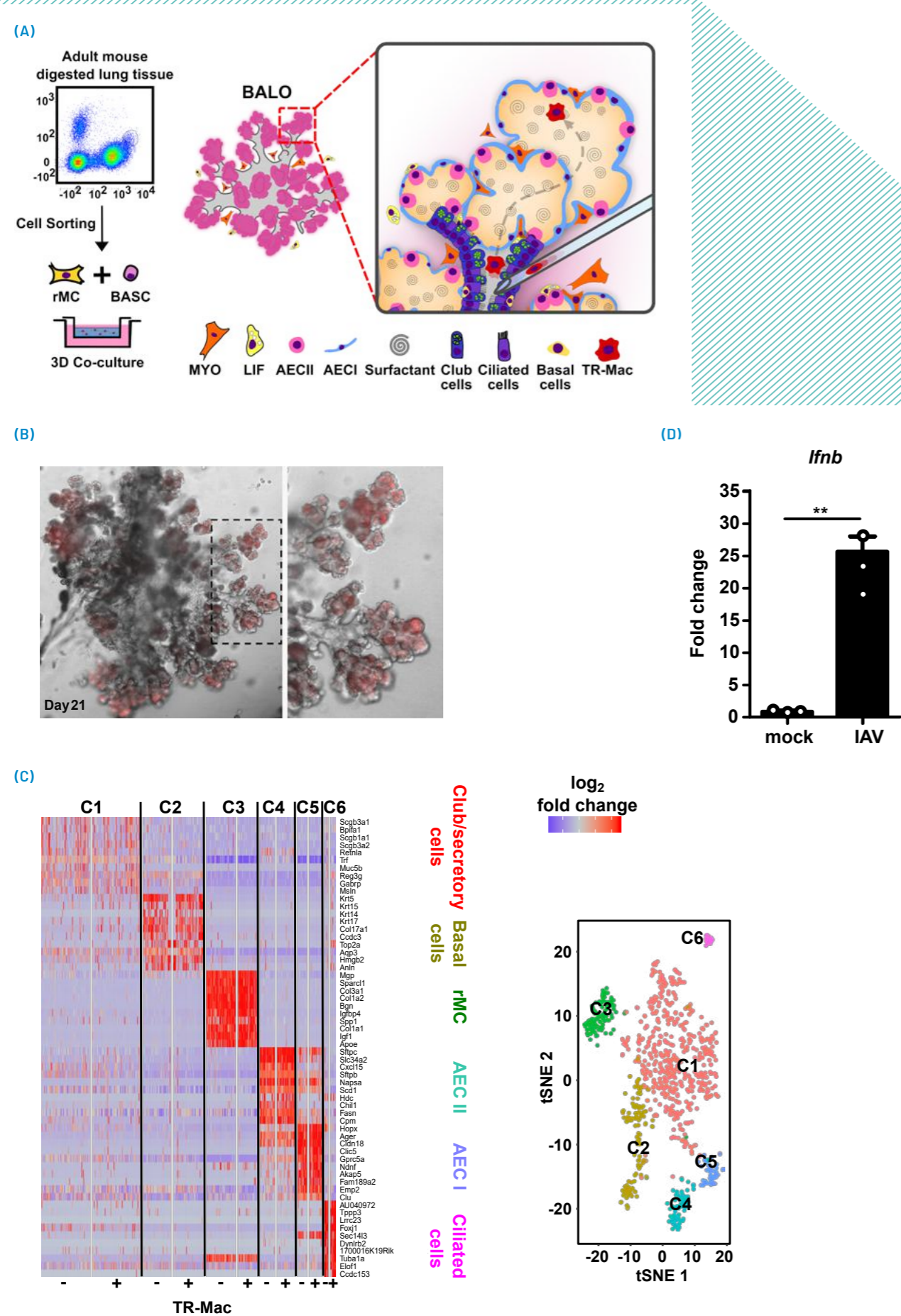
WASP will on the one hand be partially provided as a standalone desktop version to enable use on a standard laptop and on the other hand as a software container using Docker. The latter one allows to benefit from state-of-the-art virtualization technologies and enables a deploy-

ment inside cloud environments such as the de.NBI Cloud, the cloud computing platform of the German Network for Bioinformatics Infrastructure (de.NBI). This comprehensive portfolio of bioinformatics software solutions as well as a structured acquisition and storage of experimental data is offered to all CRU members to process and analyze the above-mentioned data sets in a standardized and highly automated fashion.

#### DISSECTION OF THE CELLULAR COMPOSITION OF MURINE BRONCHIO-ALVEOLAR LUNG ORGANIDS USING SINGLE-CELL RNA SEQUENCING AND WASP

Organoids derived from mouse and human stem cells have rapidly become a powerful tool to study diseases such as cystic fibrosis and lung cancer [1]. In accordance, we have established a novel three-dimensional (3D) murine bronchioalveolar lung organoid (BALO) model that allows clonal expansion and self-organization of FACS-sorted bronchioalveolar stem cells (BASC) upon co-culture with lung-resident mesenchymal cells (rMC) [2]. BALO yield a highly branched 3D structure within 21 days of culture, mimicking the cellular and structural composition of the bronchioalveolar compartment (Figure 2A-B). In this regard, scRNA-Seq and the newly developed WASP software were utilized to dissect the cellular composition of BALOs. Day 21 BALO analysis showed the presence of four distinct clusters including two epithelial clusters with alveolar- and airway-associated genes and two mesenchymal clusters expressing myofibroblast- and lipofibroblast-associated genes. To add another level of complexity to our model, BALOs were complemented by direct microinjection of tissue-resident macrophages (TR-Mac). ScRNA-Seq and WASP comparative analysis of day 23 BALO cultures with and without micro-

injected TR-Mac showed the presence of six distinct clusters defined as club/secretory cells (Scgb3a2, Muc5b), basal cells (Krt5, Trp63), rMC (Col1a2, Apoe), alveolar epithelial cells (AEC) type II (Sftpc, Cxcl15), AEC type I (Hopx, Ager), and ciliated cells (Foxj1, Tppp3) (Figure 1C). Moreover, data revealed that genes associated with cell proliferation (Fos, Fosb, Areg) and inflammatory processes (Erg1, Atf3) were downregulated in BALO with TR-Mac, whereas genes related to cell differentiation (Neat1) and maturation (Cyp2f2, Ces1d) were upregulated suggesting that the presence of TR-Mac in BALO drives epithelial differentiation while reducing cell proliferation and stress signals. Ultimately, we aimed to model disease, therefore, BALOs were infected with different strains of Influenza A virus. Our data showed that BALO supports viral replication and spread, and mounts an antiviral immune response (Figure 1D-E). Notably, the significance of our *ex vivo* model is now highlighted by the SARS-CoV-2 pandemic where 3D models such as BALO can be used in combination with scRNA-Seq and WASP data analysis to uncover molecular processes occurring during lung infection, injury, and repair that are currently difficult to study *in vivo*. This experimental approach is supported by compute sources, data management, technical support and years of experience in handling sequencing data provided by the de.NBI BIGI Service Center at Justus Liebig University Giessen.



**FIGURE 2:** Development of a lung organoid model from single BASC reflecting the distal lung structure including bronchial and alveolar injury and regeneration processes after IAV infection. (A) Schematic representation of BALO generation from flow-sorted BASC and rMC, structure and cellular composition. (B) Staining of AECII in BALO with RFP LysoTracker. (C) Heat map (left)

and tSNE plot (right) of the comparative analysis of digested day 23 BALO cultures with (+) and without (-) microinjected TR Mac depicting six distinct clusters (C1, club/secretory cells, red; C2, basal cells, yellow; C3, rMC, green; C4, AEC II, blue; C5, AEC I, purple; and C6, ciliated cells, pink). (D) mRNA expression of *Ifnb* in mock and PR8 IAV-infected BALO at 48 h pi (n=3 biological replicates with pooled cells from four

cultures per replicate). Bar charts presented as the mean ± S.E.M. and probability determined using unpaired t-test (\*\*P < 0.01) (E) Representative fluorescence images of a day 21 distal region in BALO generated from mTmG reporter mouse and infected by SC35M-Cre IAV after 0, 10, 14 and 25 h. Arrows indicate cell death. Scale bars represent 25 μm and 10 μm in the insert [2].

**REFERENCES:** [1] Stem Cells International 2020; Article ID:5847876. DOI: <https://doi.org/10.1155/2020/5847876>.  
[2] The EMBO Journal 2020;39:e103476. DOI: <https://doi.org/10.15252/emj.2019103476>.

**AUTHORS:** Susanne Herold<sup>1,4</sup>, Andreas Hoek<sup>5</sup>, Ana Ivonne Vazquez-Armendariz<sup>1,4</sup>, Karina Brinkrolf<sup>6</sup>, Jan Philipp Mengel<sup>6</sup>, Torsten Hain<sup>6,7</sup> and Alexander Goesmann<sup>5</sup>

<sup>1</sup> Internal Medicine II and Cardio-Pulmonary Institute (CPI), Justus Liebig University Giessen, Aulweg 130, 35392 Giessen

<sup>2</sup> Universities of Giessen and Marburg Lung Center (UGMLC), Klinikstraße 33, 35392 Giessen

<sup>3</sup> German Center for Lung Research (DZL), Aulweg 130, 35392 Giessen

<sup>4</sup> Institute for Lung Health (ILH), Aulweg 130, 35392 Giessen

<sup>5</sup> Bioinformatics & Systems Biology, Justus Liebig University Giessen, Heinrich-Buff-Ring 58, 35392 Giessen

<sup>6</sup> Medical Microbiology, Justus Liebig University Giessen, Schubertstraße 81, 35392 Giessen

<sup>7</sup> German Center for Infection Research (DZIF), Partner site Giessen-Marburg-Langen, Schubertstraße 81, 35392 Giessen

# OPEN DATA, SOFTWARE AND ANALYTICS AS A RESPONSE TO EMERGING PATHOGEN THREATS

Our society is undergoing a major transformation due to the COVID-19 pandemic. The global scientific efforts have highlighted the need for open data, analysis tools and computational resources. The Galaxy community provides access to novel and publicly available data, to computational resources, maintains tools, training materials and develops projects to contribute to SARS-CoV-2 research.

The COVID-19 pandemic has transformed the way science is perceived by society. Many scientists from all over the world have joined the effort to tackle the public health emergency. However, the coordination between experts of different fields is, more than ever, crucial to address the current global crisis. The openness of the data, analysis tools, and computational resources is another fundamental aspect to pave the way towards a common global solution.

In this article, we describe how our community maintains an ecosystem of open software tools; presenting several data analysis projects that have contributed to the SARS-CoV-2 research. Furthermore, we show the way in which education is shifting to an online paradigm and how to make use of the available resources to adapt to the new model. The development of all these projects has been possible thanks to the public infrastructure

created and maintained in the last years. Galaxy is, therefore, used to address the current worldwide challenges by providing access to novel and publicly available SARS-CoV-2 data, and enabling the processing of complex and computationally heavy datasets.

## GALAXY AS A GATEWAY FOR DATA ANALYSIS

The European Galaxy platform (<https://usegalaxy.eu>) serves as a gateway for **data analysis and access to databases** for scientists of all disciplines. Galaxy is an **open-source**, highly modular, flexible and extensible system that focuses on easily accessible and reproducible research. To use Galaxy, users just need a web browser, avoiding the drawbacks of downloading and installing software.

Galaxy Europe offers more than **2,500 command-line bioinformatics tools** in

one single user-friendly graphical interface. The functionality varies from text manipulation or data visualization to more complex analysis tools serving different communities: Genomics, Proteomics, Metagenomics, Plant Science, Imaging, Microbial Science, Climate Science, NGS, Metabolomics, and many more (Figure 1). The tools are constantly maintained, by updating them regularly (keeping track of the versions) and integrating new ones upon user demand or when changes in the underlying technology demand it.

One of the most powerful features in Galaxy is the creation of automated pipelines, the Galaxy **workflows**. A workflow is a collection of tools that need to be run in a specific order and can be easily extracted either from existing histories or by dragging and dropping tools from the tools panel and connecting them to each other.

The **reproducibility** of the data analysis is ensured by saving the metadata (this means, the parameters selected) associated with the analyses. Anyone can reload the tools and run them again with the same parameters. This functionality is very similar to the concept of an electronic lab notebook but applied here to data analysis. Furthermore, all datasets, histories, and workflows within Galaxy can be easily shared with other users, and outside of Galaxy.

The European Galaxy server offers more than 2,500 bioinformatics tools and the team behind it is among the main contributors to the Galaxy training material (<https://training.galaxyproject.org>). The registration for users of the European Galaxy server is free of charge as well as the attendance to the workshops. Initially, 250 GB of storage are provided to the users upon their registration, though this can be easily extended for a

specific project or for a defined time frame on request.

The main European Galaxy server (<https://usegalaxy.eu>) has more than 23,000 users from 100 different countries and more than 11 million jobs performed. In 2020, the average number of active users per month (users that have submitted at least one job) is above 1,700. Around 500,000 jobs are run per month, meaning more than 16,000 per day and around 12 jobs per minute.

## DATASETS & DATA ACCESS

Public databases, like the European Nucleotide Archive (ENA) or the Sequence Read Archive (SRA), are key to ensure that the SARS-CoV-2 sequence datasets are stored long-term in public archives and adhering to the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Within de.NBI, special connectors have

been integrated to those public archives to easily provide scientists with those data in Galaxy and the different clouds. Although keeping track of the latest data is time-consuming and not straightforward due to the rapid data growth, all publicly available SARS-CoV-2 sequences have been made available. Moreover, downloading thousands of datasets can take days. To overcome these issues, as well as to provide a quick turn-around in analyzing the latest sequences, a collection of identifiers of relevant sequences has been created and is updated daily in an automated way. Those sequences are deposited in the European Galaxy server<sup>1</sup> – one of the biggest public Galaxy instances – so that they can be used immediately by everyone. To facilitate COVID-19 research, the recent SARS-CoV-2 reference genome is continuously integrated and optimized indices are created to access it from all related tools.

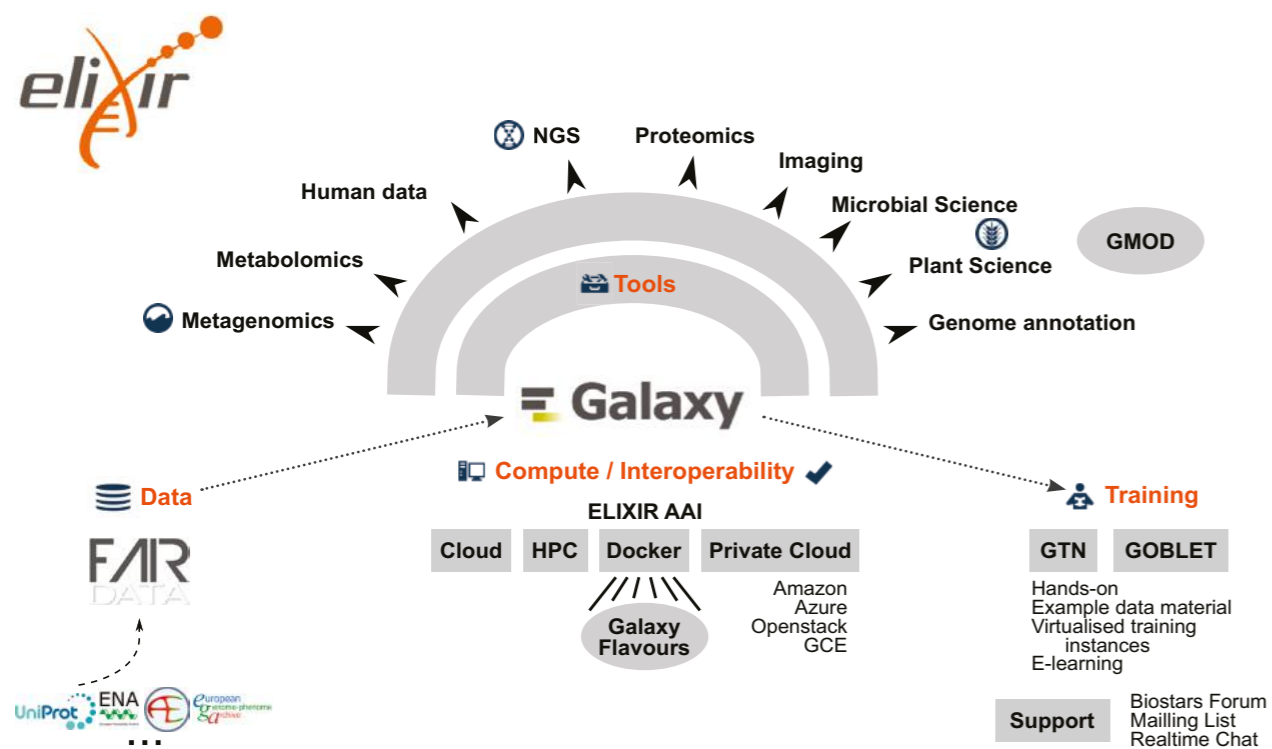


FIGURE 1: Galaxy communities and software stack.

## DATA ANALYSIS PROJECTS

There are a number of COVID-19 related projects that are currently running on the European Galaxy server. All data, provenance information, and ready-to-use workflows are available in <https://usegalaxy.eu> and the WorkflowHub<sup>2</sup>. A more detailed description of the projects can be found at <https://covid19.galaxyproject.org>. In total, **16 COVID-19 workflows** have been developed, are supported and constantly improved since the beginning of March 2020. This extraordinary effort was only possible thanks to the infrastructure that has been built over the last year with the help of regional, national, and international funding.

### Initial analysis of COVID-19 data using Galaxy, Bioconda, and public research infrastructure

The COVID-19 pandemic has revealed the strong need for reproducible analy-

tics and access to compute power. Now more than ever, the publication of results whose analytical procedures are not fully reproducible and transparent is no longer acceptable. Galaxy has already been used to re-analyze and to assess the reproducibility of the first COVID-19 genome papers and the results are published in the article 'No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics' [1]. The project is focused on **underscoring the importance of access to raw data** and to demonstrate that existing **community efforts in curation and deployment of biomedical software can reliably support rapid reproducible research during global crises**. All the workflow descriptions and versions of the software used are detailed in the article (Figure 2). These workflows can be executed by any researcher all over the world in any of the five global Galaxy instances in the US<sup>3</sup>, in Europe<sup>4</sup> (Germany, France, Belgium), and in Australia<sup>5</sup>.

### Reproducible and scalable workflows for SARS-CoV-2 variation and transcript expression analysis

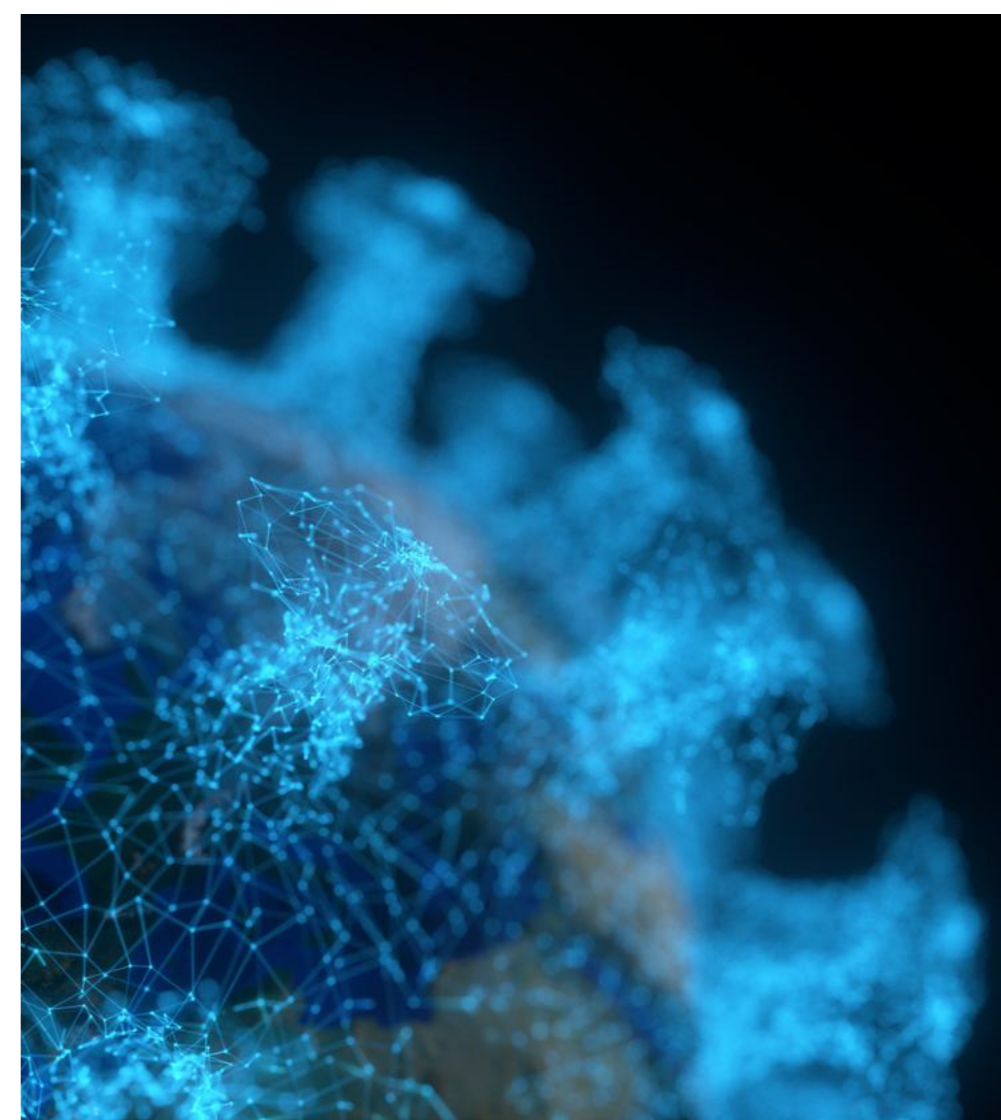
One area of COVID-19 research, in which reproducible data analysis is of particular importance, is the study of virus evolution. As part of its adaptation to humans as its host, in response to host factors in specific human populations, and, in the future, as a mechanism to evade vaccines and/or drug effects, SARS-CoV-2 could acquire nucleotide sequence changes in its 30 kb single-stranded RNA genome and/or change the relative expression of its at least nine structural proteins. The ability to track such changes in near real time and to detect the emergence of novel epidemiologically relevant strains rapidly could be of enormous practical value. However, any claims of occurrence of isolates and strains with altered genomic sequence or protein composition are worth little if the underlying methods leading to the detection of these changes are not fully documented and reproducible.

Through the European Galaxy Server, we are providing a selection of well-tested workflows for variation detection from viral sequencing data obtained using a variety of different techniques and from different sources. Specifically, we offer nucleotide variation detection workflows optimized for Illumina paired-end and single-end sequenced whole genomic DNA<sup>6</sup> and for PCR-amplified viral DNA obtained via protocols released by the ARTIC network<sup>7</sup>. For the detection of the expression levels of viral proteins and their isoforms from viral whole-genome sequencing data, we provide a workflow<sup>8</sup> that classifies Nanopore- or Illumina-sequenced reads according to the subgenomic transcripts they are derived from and counts the number of observations of each known or suspected transcript (Figure 3).

All of these workflows have been validated against publicly available SARS-CoV-2 sequencing data and are scalable from analysis of just a handful of in-house samples to large-scale reanalysis of thousands of published datasets. Several of these workflows have also been used as accessory workflows in our experimental study described in the following section.

### Analyzing the landscape of SARS-CoV-2 RNA modifications with the first european Direct RNA Sequencing (DRS) datasets

Gaining knowledge on how SARS-CoV-2 interacts with its host cells and causes COVID-19 is crucial for the intervention of novel therapeutic strategies. SARS-CoV-2, like other coronaviruses, is a positive-strand RNA virus. The viral RNA is modified by RNA-modifying enzymes provided by the host cell. Direct RNA sequencing using nanopores enables unbiased sensing of canonical and modified RNA bases of the viral transcripts. In this work, we used DRS to precisely



annotate the open reading frames and the landscape of SARS-CoV-2 RNA modifications. We provide the first DRS data of SARS-CoV-2 in infected human lung epithelial cells. From sequencing three isolates, we derive a robust identification of SARS-CoV-2 modification sites within a physiologically relevant host cell type. A comparison of our data with the DRS data from a previous SARS-CoV-2 isolate raised in monkey renal cells, reveals consistent RNA modifications across the viral genome [2]. Conservation of the

RNA modification pattern during the progression of the current pandemic suggests that this pattern is likely to be of relevance for the life cycle of SARS-CoV-2 and represents a possible target for drug interventions (Figure 4).

### Drug design and screening against SARS-CoV-2 Mpro

Non-Structural Proteins (nsps) are vital for the life-cycle of SARS-CoV-2 and can be cleaved from a large precursor (encoded by ORF1ab) by enzymes such as the

<sup>2</sup> <https://workflowhub.eu/workflows?filter%5Btag%5D=covid-19>  
<sup>1</sup> <http://usegalaxy.org>

<sup>3</sup> <http://usegalaxy.eu>  
<sup>4</sup> <https://usegalaxy.org.au>

<sup>6</sup> <https://covid19.galaxyproject.org/genomics/4-Variation/>  
<sup>7</sup> <https://covid19.galaxyproject.org/artic/>

<sup>8</sup> <https://usegalaxy.eu/u/wolfgang-maier/w/sars-cov-2-classify-ont-reads-by-transcript-junctions>  
<sup>9</sup> <https://covid19.galaxyproject.org/cheminformatics/Histories/>



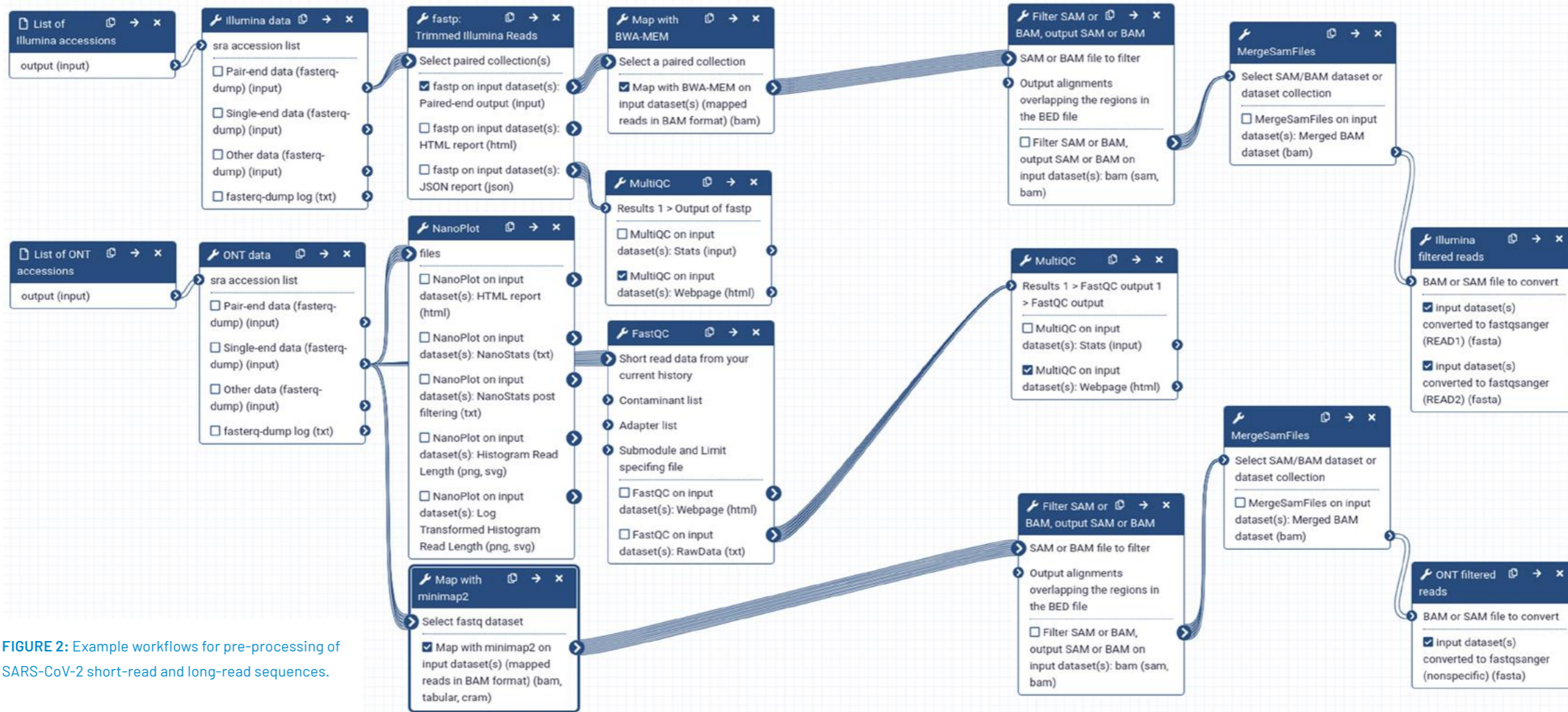


FIGURE 2: Example workflows for pre-processing of SARS-CoV-2 short-read and long-read sequences.

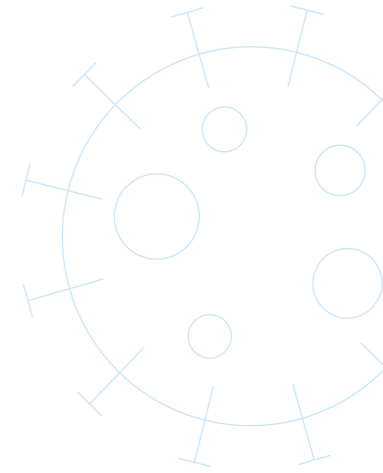


FIGURE 3: Spike protein (S) gene variants in published sequence datasets from the early phase of the pandemic.

main protease Mpro. We performed computational analyses (using protein-ligand docking) to identify potentially inhibitory compounds that can bind to Mpro and can be used to control viral proliferation. This work analyzed around 41,000 compounds considered to be likely to bind to one of the SARS-CoV-2 proteases. These proteases were chosen based on recently published X-ray crystal structures, and identified 500 high-scoring compounds. All the workflows used for this analysis as

well as individual compound lists can be accessed, for each fragment structure, in the corresponding Galaxy histories<sup>9</sup>. **Single-cell transcriptome analysis with the human cell atlas** The Human Cell Atlas project has made available a large amount of SARS-CoV-2 data under <https://www.covid19cellatlas.org>. The de.NBI is working closely with the European Bioinformatics Institute and the Human Cell Atlas project to

standardize the analysis of SARS-CoV-2 single-cell RNA data. This involves the development of new tools, data importers and exporters and visualisations. The European Galaxy server is hosting tools and workflows that support this initiative under <https://humancellatlas.usegalaxy.eu>. General training material about single-cell transcriptomics is available as part of the Galaxy Training Network (GTN)<sup>10</sup> under the tag 'single-cell'.

Sample	CHROM	POS	REF	ALT	DP	AF	SB	DP4	IMPACT	FUNCLASS	EFFECT	GENE	CODON	type	
20590	SRR11140744	NC_045512	17373	C	T	5249	0.996571	0	3,2,2782,2462	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gc/C/gcT	S
29046	SRR11140746	NC_045512	17373	C	T	5045	0.996432	0	3,3,2428,2610	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gc/C/gcT	S
31304	SRR11140748	NC_045512	17373	C	T	6125	0.994776	4	3,7,2948,3163	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gc/C/gcT	S
32645	SRR11247077	NC_045512	28144	T	C	140	0.992857	0	1,0,139,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	ORF8	tTa/tCa	S
32640	SRR11241255	NC_045512	17858	A	G	122	0.991803	0	1,0,121,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	tA/tGt	S
32218	SRR11140750	NC_045512	17373	C	T	141	0.985816	0	1,1,68,71	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gc/C/gtT	S
32648	SRR11247078	NC_045512	18060	C	T	140	0.985714	0	2,0,138,0	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	ctC/ctT	S
32641	SRR11247076	NC_045512	3406	A	C	112	0.982143	0	1,0,110,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	gaA/gaC	S
32643	SRR11247077	NC_045512	18060	C	T	141	0.978723	0	1,0,139,0	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	ctG/ctT	S
32646	SRR11247078	NC_045512	17747	C	T	122	0.97541	0	2,0,120,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	cCt/ctT	S
32647	SRR11247078	NC_045512	17858	A	G	132	0.969697	0	2,0,128,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	tA/tGt	S
32635	SRR11241254	NC_045512	5572	G	T	127	0.968504	0	1,0,125,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	atG/atT	S
32644	SRR11247077	NC_045512	20281	T	C	147	0.911565	0	12,0,134,0	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	Ttc/Ctc	S
19956	SRR11177792	NC_045512	24034	C	T	16123	0.7431	2147483647	1166,2801,8168,3979	LOW	SILENT	SYNONYMOUS_CODING	S	aaC/aaT	S
32312	SRR10903401.fastq	NC_045512	24323	A	C	316	0.379747	3	100,93,58,64	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	S	Aaa/Caa	S
25829	SRR11140744	NC_045512	11335	G	T	8376	0.370941	4	2495,2761,1511,1605	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gtG/gtT	S
32180	SRR11140750	NC_045512	11335	G	T	433	0.21709	1	152,186,45,50	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gtG/gtT	S
28405	SRR11140746	NC_045512	11335	G	T	7928	0.197275	4	3040,3310,773,801	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gtG/gtT	S
32299	SRR10903401.fastq	NC_045512	19164	C	T	88	0.181818	0	34,38,8,8	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gaC/gaT	S
30645	SRR11140748	NC_045512	11335	G	T	8174	0.178982	9	3393,3312,777,692	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	gtG/gtT	S
32384	SRR10903402.fastq	NC_045512	11563	C	T	385	0.145455	2	173,155,33,24	LOW	SILENT	SYNONYMOUS_CODING	orf1ab	tgC/tgT	S
32554	SRR10971381.fastq	NC_045512	9464	T	C	76	0.105263	8	30,37,1,7	MODERATE	MISSENSE	NON_SYNONYMOUS_CODING	orf1ab	Ttt/Ctt	S

## TRAINING ACTIVITIES

Dedicated training material for the analysis of the SARS-CoV-2 genome<sup>11</sup> and the investigation of potential drug candidates have been developed already since February 2020 to make it more convenient for researchers to dive into data analysis by provisioning easy access of the corresponding data, tools, workflows, and related tutorials through Galaxy.

In this section, we briefly introduce how the Galaxy training materials are being used for University-level teaching during the pandemic and highlight some Galaxy tutorials developed for SARS-CoV-2 data analysis and published on <https://training.galaxyproject.org>.

### Teaching during the pandemic

The education paradigm is shifting towards an online scenario, which requires an adaption of the traditional classroom pedagogical approach [4]. During the pandemic period, there have been many courses taught using the European Galaxy server, thanks to the convenient and scalable nature of e-learning in Galaxy [5]. Galaxy offers Training Infrastructure as a Service (TlaaS)[3] for the community, providing with the computed resources together with more than 170 e-learning materials – covering a variety of different scientific topics – developed as part of the Galaxy Training Network. A statistical overview about the TlaaS training can be found here: <https://usegalaxy.eu/tiaas/stats>. The transparent sharing of data analysis histories makes it easy for teachers to assist remotely, track down mistakes or enable assessments. According to the users of the European Galaxy server, it seems to be a common use-case to provide an online webinar introducing the topic, leaving the students one week to analyse one dataset (using the GTN-material) and finally requesting

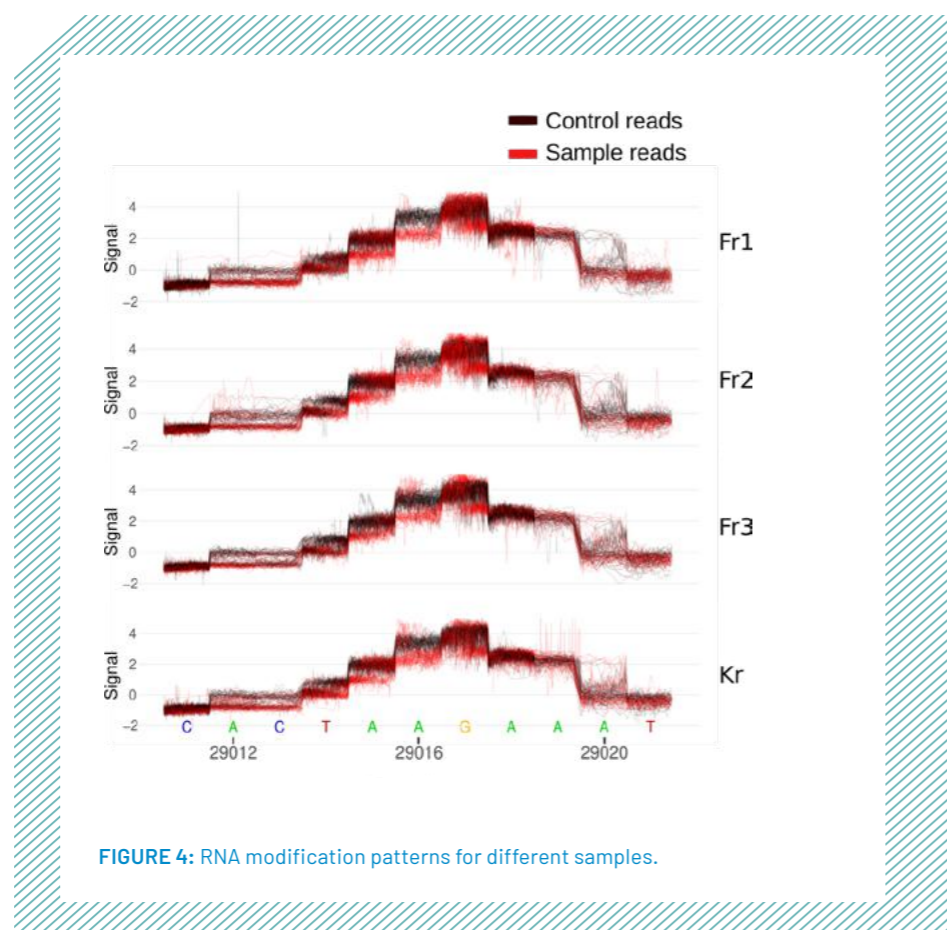


FIGURE 4: RNA modification patterns for different samples.

a shared-history as proof that students have done their homework.

### New tutorial: virtual screening of the SARS-CoV-2 main protease with rDock and pose scoring

This tutorial<sup>12</sup> supports the work performed in March 2020 by InformaticsMatters<sup>13</sup>, the Diamond Light Source<sup>14</sup>, and the European Galaxy Team to perform virtual screening on candidate ligands for the SARS-CoV-2 main protease (Mpro).

The original drug design study required almost 20 years of CPU time, excluding the GPU resources consumed. Since repeating this analysis is not feasible as a tutorial, we reproduced the workflow with a small library of just 100 molecules, on a single Mpro fragment structure. The links to the original Galaxy histories are in the tutorial, with notes to explain which parts of the actual analysis had to be accomplished differently to the tutorial.

### New tutorial: assembly of SARS-CoV-2 gGenome with preprocessing to remove human genome reads

In some research or clinical contexts, it is very challenging to purify DNA/RNA for sequencing from just the specimen of interest. Instead, isolated DNA is contaminated, sometimes heavily, with DNA/RNA of a different origin. This is the case for example with microbiome samples, which typically display considerable contamination with host DNA, or with samples of body fluids for pathogen detection. The contamination can pose an issue with certain types of analyses, in particular with genome assembly.

This tutorial<sup>15</sup> guides users through the preprocessing of sequencing data of BronchoAlveolar Lavage Fluid (BALF) samples obtained from early COVID-19 patients in China. Since those samples are expected to be contaminated significantly with human sequenced reads, the

goal is to enrich the data for SARS-CoV-2 reads by identifying and discarding reads of human origin before trying to assemble the viral genome sequence.

### Galaxy-ELIXIR webinar series: FAIR data and open infrastructures to tackle the COVID-19 pandemic

The Galaxy Community and ELIXIR organised a webinar series<sup>16</sup> to demonstrate how open access and open science are fundamental for fast and efficient response to public health crises, by analysing and publishing SARS-CoV2 data. The main goal of the series was to show workflows for SARS-CoV-2 data analyses in publicly accessible infrastructure.

In a series of five webinar sessions, experts from ELIXIR and the Galaxy community in the US and Europe used exclusively open source tools and the Galaxy platform with a focus on reproducible and transparent research. The trainers guided participants step-by-step through setting up and executing the SARS-CoV-2 data analyses workflows developed by the global Galaxy community<sup>17</sup>. After completing the series, participants were able to fully reproduce the workflows and conduct their own analyses of SARS-CoV-2 data.

The webinar series started on the 30th of April 2020 with the first introductory session and subsequent sessions took place in weekly intervals, addressing the following topics:

- Introduction to Galaxy and the Galaxy workflows for SARS-CoV-2 data analysis
- Genomics/Variant calling
- Cheminformatics: Screening of the main protease
- Evolution of the virus
- Behind the scenes: Global Open Infrastructures at work

### Hackathons and Co-Fests

The RBC has participated in the multiple Hackathons and Co-Fest during the outbreak with different roles: mentor, expert, attendee and as an infrastruc-

ture provider. During the #WeVsVirus Hackathon organised by the German government, the RBC offered compute resources, interactive environments (like Jupyter Notebooks and RStudio instances) and taught people about drug design and virtual screening techniques using the Galaxy training material. The Virtual BioHackathon has been used to integrate tools specifically for variation graphs into Galaxy and to create Conda packages as well as BioContainers for the COVID-19 research community. The de.NBI cloud was used for the assembly of SARS-CoV-2 and to run Jupyter Notebooks. Moreover, an international team of researchers published all Galaxy related COVID-19 workflows into the preview version of the <https://WorkflowHub.eu>.



**REFERENCES:** [1] PLoSPathog 2013;16(8):e1008643. DOI: <https://doi.org/10.1371/journal.ppat.1008643>. [2] bioRxiv [Preprint] 2020. DOI: <https://doi.org/10.1101/2020.07.18.204362>. [3] bioRxiv [Preprint] 2020. DOI: <https://doi.org/10.1101/2020.08.23.263509>. [4] Preprints 2020:2020080532. DOI: 10.20944/preprints202008.0532.v1. [5] Preprints 2020:2020090457. DOI: 10.20944/preprints202009.0457.v1.

**AUTHORS:** Beatriz Serrano-Solano<sup>1</sup>, Wolfgang Maier<sup>1</sup>, Simon Bray<sup>1</sup>, Gianmauro Cuccuru<sup>1</sup>, Anika Erxleben<sup>1</sup>, Bérénice Batut<sup>1</sup>, Mehmet Tekman<sup>1</sup>, Rolf Backofen<sup>1</sup>, Björn Grüning<sup>1</sup>

<sup>1</sup> University of Freiburg, Department of Computer Science, Georges-Köhler-Allee 106, 79110 Freiburg

# BioInfra.Prot SUPPORTS PUBLICATION OF PROTEOMICS DATASETS as well as study design and data analysis for COVID-19 research projects



Proteomics is an important field of research in the course of the COVID-19 pandemic. The de.NBI service center BioInfra.Prot provides a comprehensive portfolio of services over the entire process of the study. This includes literature research on biomarkers, study planning, statistical and bioinformatical consulting as well as data publication in the data repository PRIDE.

## COVID-19 PANDEMIC – A CHALLENGE FOR SCIENTISTS

The pandemic of Coronavirus disease 2019 (COVID-19) caused by the virus SARS-CoV-2 has an enormous negative impact on all affected health systems, economies and societies. Thus, to alleviate and overcome the negative consequences as fast as possible it is crucial to focus extensive research efforts on the development of effective diagnostic test methods, drugs and vaccines. Worldwide collaboration of interdisciplinary researchers and experts involved in COVID-19-related research and efficient sharing of high-quality data and results are needed to accelerate this endeavor. Moreover, the support of local and nationwide research regarding study design and data analysis by expert statisticians and bioinformaticians is an important contribution to COVID-19-related research efforts. The German Network for Bioinformatics Infrastructure (de.NBI) supports biologists, medical doctors and all other scientists regarding bioinformatical and statistical questions. The service center Bioinformatics for Proteomics (BioInfra.Prot) [1] located in Bochum and Dortmund offers support specialized in data processing and data analysis of proteomics data.

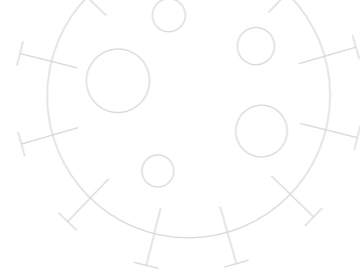
## ROLE OF PROTEOMICS IN COVID-19 RESEARCH

Proteomics is the science of studying proteins, the organism's executive molecules, that perform important molecular tasks, e.g. as enzymes. In the context of COVID-19, proteins are important research targets, e.g., for their role as receptor proteins or virus surface proteins. These proteins are potential drug targets and may serve as biomarkers for the diagnosis or prognosis of disease progression.

Due to the urgent situation caused by the COVID-19 pandemic, a lot of research has been focused on this topic. Nevertheless, there is still a great need for research, especially in the field of proteomics, as the first studies mostly focused on genetic aspects of SARS-CoV-2.

## SEARCH FOR PREVIOUS RESEARCH FINDINGS

Due to the large number of publications on biomarkers (Figure 1), it is impossible for a scientist to read all papers in detail and extract the information relevant to them. At this point, the web-based biomarker database BIONDA [2], which is provided by BioInfra.Prot (Figure 2), can support the scientific community.



BIONDA uses text-mining methods to extract potential disease and biomarker relationships from literature and present them to the user in a structured tabular form. The use of this database allows a quick overview of already published proteins, genes and miRNA connected to various diseases. At the beginning of the pandemic, the coronavirus proteins provided by UniProt [3] have been added to the database as well as a quick search function on the BIONDA homepage. Currently, 70 COVID-19 related publications and 182 Biomarker-COVID-pairs are integrated in BIONDA. This helps scientists to perform an efficient literature search and get an overview of the current state of research regarding COVID-19 biomarkers.

Moreover, in the proteomics field, it is strongly recommended to upload raw data to a publicly accessible repository. This ensures that the data is securely stored, treated according to FAIR principles [4] and made available to the public for re-analysis. The ProteomeXChange Consortium [5] unites different worldwide operating repositories for proteomics data and is an extremely important infrastructure, which enables the provi-

sion of SARS-CoV-2-related proteomics datasets to all researchers worldwide. The largest of these repositories is the PRoteomics IDentifications Database (PRIDE) [6]. In close collaboration with the PRIDE team at EMBL-EBI headed by Juan Antonio Vizcaino, BioInfra.Prot staff members are involved in the curation of PRIDE datasets and corresponding user consulting (Figure 2). PRIDE currently stores 36 public datasets related to COVID-19, of which 27 were processed and set online by our team in Bochum on the PRIDE web portal. To simplify the research within PRIDE, it also offers a comfortable quick search for COVID-19 datasets on its website.

#### REQUEST FOR STUDY DESIGN AND SAMPLE SIZE ESTIMATION

BioInfra.Prot offers consulting regarding bioinformatical and statistical analysis of proteomics data, as well as the planning of studies and experiments. Especially the estimation of an appropriate sample size is a crucial step and is required by ethics committees for approval of the study. For this step information from comparable prior experiments is

required such as expected variances or effect sizes. This information is usually taken from pilot studies or similar studies from literature. In case of the COVID-19 studies, many laboratories without prior experience in this kind of diseases start researching on this topic and may not have the needed information available.

In this scenario, PRIDE is a valuable repository, that allows to reuse public datasets for this task. Additionally the European Bioinformatics Institute (EMBL-EBI) offers a portal with more COVID-19 related data, also beyond proteomics [7]. Thus, public data from COVID-19 patients or infected cell lines are available and can be used to estimate variances of the protein measurements and get an idea of possible effect sizes. The fast publication of COVID-19 datasets allows a well-founded sample size calculation and planning of intended studies. At the moment, many calls for grants and funding regarding COVID-19 are open and a fast and meaningful study planning is of utmost importance. Without PRIDE, this would be hindered by the time-consuming execution of pilot studies.

#### EXPERIMENT, DATA ANALYSIS AND PUBLICATION OF DATASETS

Thanks to BioInfra.Prot's assistance in study design and sample size planning (Figure 2), a research proposal has a good chance of being accepted. In this case, the scientists conduct their study and collect data, e.g. from COVID-19 patients or infected cell lines.

For the (pre-)processing and analysis of the proteomics data the services of BioInfra.Prot can be used again. The service 'Bioinformatical consulting and analysis of proteomics data' helps with the complex (pre-)processing of raw and results data. We also offer a variety of software tools to

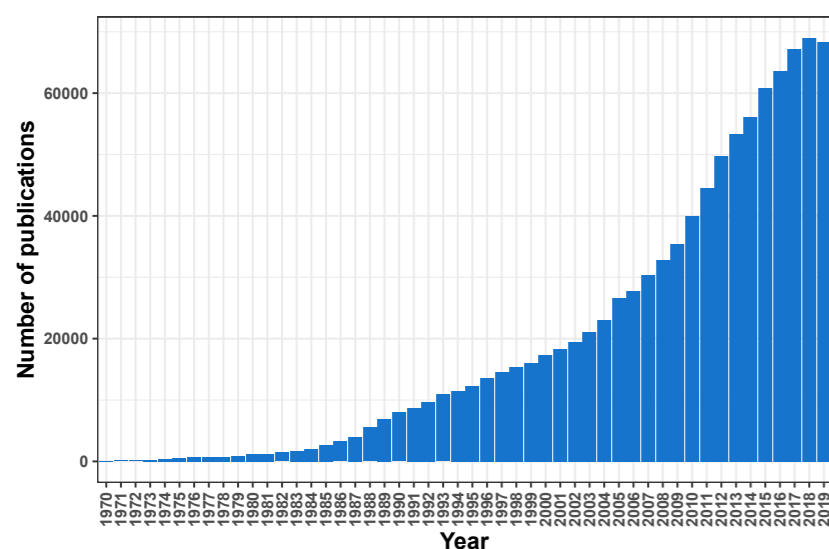


Figure 1: Number of biomarker-related publications in PubMed [11] per year since 1970.

perform different steps of the workflow. For example, the CalibraCurve [8] tool supports the targeted analysis of individual proteins, which may be used for the identification of SARS-CoV-2 proteins for diagnosis. PIA [9] (Figure 2) takes over the task of protein inference in high-throughput experiments that for example compare the whole proteome of COVID-19 patients with healthy controls. BioInfra.Prot also offers a bioinformatics consulting service, which supports the selection of suitable tools and their operation.

After the pre-processing various statistical analyses can be carried out in order to answer the research question in an optimal way. For example, a statistical test can find out proteins that are up- or downregulated in COVID-19 patients. A single protein may not be sufficient to predict the severity of the disease, so instead a biomarker panel consisting of a combination of several peptides is required. Some machine learning methods required for the calculation are implemented in the tool PAA [10]. Furthermore, assistance is provided regarding graphical representations of the results and preparation of a manuscript. As described above, it is essential to upload the raw data into a repository like PRIDE. The upload service of BioInfra.

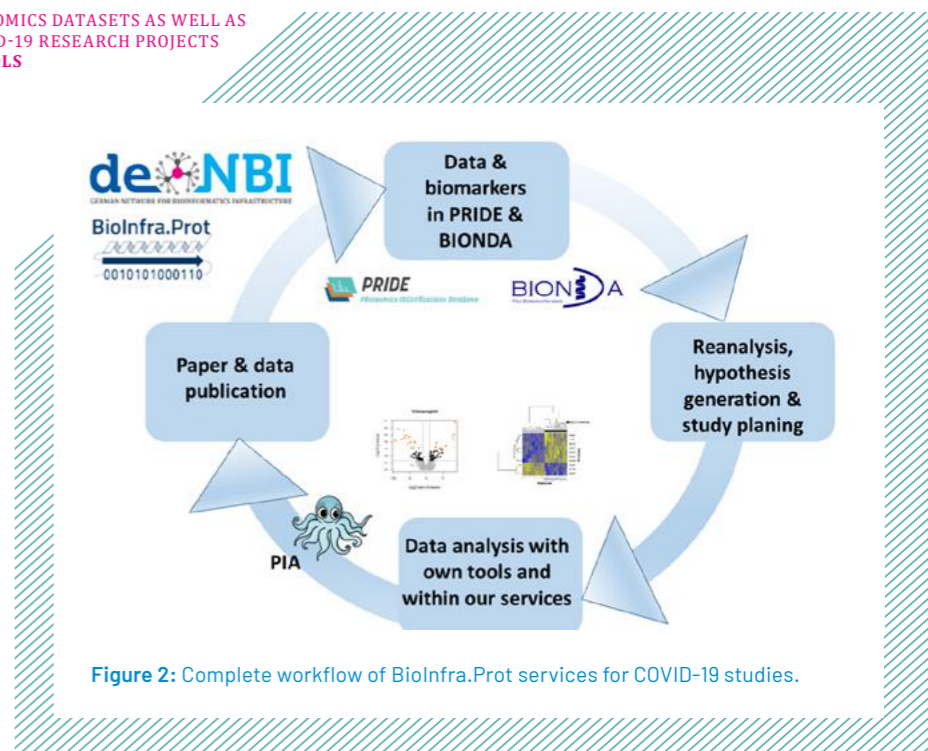


Figure 2: Complete workflow of BioInfra.Prot services for COVID-19 studies.

#### CONCLUSION

We are involved in the COVID-19-related proteomics research at multiple positions. As shown in Figure 2 our research workflow forms a circle: We use COVID-19-related proteomics datasets and biomarkers from past studies that are available in PRIDE and BIONDA for data re-analysis, hy-

pothesis generation and the planning/design of new studies. We analyze data of ongoing studies partly with our own software tools and in the course of our consulting and data analysis services. Subsequently, we support also paper and data publication. Finally, the published datasets and biomarkers are available in PRIDE and BIONDA so that our workflow has come full circle.

#### REFERENCES:

- [1] J Biotechnol 2017;261:116-125. DOI: 10.1016/j.jbiotec.2017.06.005.
- [2] <http://bionda.mpc.ruhr-uni-bochum.de/>
- [3] <https://covid-19.uniprot.org/> [4] Scientific data 2016;3:160018. DOI: <https://doi.org/10.1038/sdata.2016.18>.
- [5] Nucleic Acids Research 2020;48:D1145-D1152. DOI: <https://doi.org/10.1093/nar/gkz984>.
- [6] Nucleic Acids Res 2020;47(D1):D442-D450. DOI: <https://doi.org/10.1093/nar/gkz984>.
- [7] <https://www.covid19dataportal.org/> [8] Proteomics 2020;20:1900143. DOI: <https://doi.org/10.1002/pmic.201900143>.
- [9] Journal of Proteome Research 2015;14(7):2988-2997. DOI: <https://doi.org/10.1021/acs.jproteome.5b00121>.
- [10] Bioinformatics 2016;32(10):1557-1579. DOI: <https://doi.org/10.1093/bioinformatics/btw037>.
- [11] PubMed. Bethesda (MD): National Library of Medicine (US), NCBI; 2004 [cited 2020.09.11]. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/>.

**AUTHORS:** Karin Schork<sup>1,2</sup>, Anika Frericks-Zipper<sup>1,2</sup>, Martin Eisenacher<sup>1,2</sup>, Michael Turewicz<sup>1,2</sup>

<sup>1</sup> Medizinisches Proteom-Center, Ruhr-University Bochum, Universitätsstraße 150, 44801 Bochum

<sup>2</sup> Center for Protein Diagnostics (ProDi), Medical Proteome Analysis, Ruhr-University Bochum, Universitätsstraße 150, 44801 Bochum



# FAST, ADAPTED, CURATED FAIR DATA FOR COVID-19 RESEARCH

Research projects such as the international COVID-19 Disease Map initiative and the German COVID-19 study hub of NFDI are supported by de.NBI-SysBio tools and services in organizing and sharing research data 'FAIRly'. This is done *via* the data management platform FAIRDOMHub/SEEK which is quickly adapted to the users' needs. COVID-19 related literature is manually curated and used for basic research about the curation process of SABIO-RK to provide the research community with high quality kinetics data.

A key message of the work described below is that it is important to have prepared, established, flexible tools that can be adapted to unpredictable challenges in research.

## DATA MANAGEMENT FOR COVID-19 RESEARCH

### Disease maps in the FAIRDOMHub

More than 230 researchers around the world have joined forces to reconstruct the molecular processes of the virus-host interactions of COVID-19 and visualize them in the form of 'disease maps'. The COVID-19 Disease Map<sup>1</sup> [1] is an effort to build a comprehensive, standardized knowledge repository of SARS-CoV-2 virus-host interaction mechanisms, guided by input from domain experts and based on published work. The resulting map is a platform for visual exploration and computational analyses of molecular processes involved in SARS-CoV-2 entry, replication, and host-pathogen interactions, as well as immune response, host cell recovery and repair mechanisms. It supports the research community and improves the understanding of this disease to facilitate the development of efficient diagnostics and therapies.

To ensure a transparent view of the contributors and community resources, the COVID-19 Disease Map community uses the de.NBI service FAIRDOMHub [2], our public research data management system running on HITS servers, that is based on the SEEK system [3]. For more than 10 years FAIRDOMHub/SEEK has been developed by the SDBV group at HITS in collaboration with the group of Carole Goble from the University of Manchester (UK) and other partners

COVID-19 is a disease that was previously unknown, with a wide variety of symptoms and properties. Many of these properties got associated with COVID only in the first few months of the ongoing pandemic. This triggers a huge demand for data on the COVID-19 disease, as well as on molecular mechanisms and activities of its causing SARS CoV-2 virus, this translates into extensive interdisciplinary research efforts. All these data needs to be bundled to render it findable, make it comparable and interoperable by the use of consistently harmonized data formats and metadata standards, and finally make it accessible for its reuse in research and clinical practice. In other words, the COVID-19 research data needs to be 'FAIR' (Findable, Accessible, Interoperable and Reusable) for making it useful. The Scientific Databases and Visualization (SDBV) group of the Heidelberg Institute for Theoretical Studies (HITS) that coordinates the de.NBI Systems Biology node, participates in various national, as well as international initiatives around COVID-19 data infrastructure and develops services for the corresponding data management.

within the FAIRDOM initiative. Through this FAIRDOM service we support the COVID-19 Disease Map community with the FAIRDOMHub project space (Figure 1) that lists and bundles contributors, data, computational models and literature, as well as all data and curation guidelines for this community (<https://fairdomhub.org/projects/190>). The underlying SEEK system is adapted by us and others to the needs of this community, for example by implementing communication channels that allow the researchers to directly get in contact with their collaboration partners. Thus, with the FAIRDOMHub project space we provide the central data and model exchange platform, as well as a communication hub for the COVID-19 Disease Map community supporting this international initiative with a crucial data infrastructure.

### COVID-19 Study Hub based on SEEK for clinical and epidemiological studies

Due to the pressing need in the progressing pandemic, a quickly growing number of clinical and epidemiological COVID-19 studies are planned or already ongoing but there is a lack of coordination among these efforts to secure common standards, comparable results and, most importantly, unified access to them. Several partners in the upcoming German National Research Data Infrastructure (NFDI) for Public Health Data (nfdi4health), including the SDBV group of HITS, are developing a COVID-19 study hub for Germany, funded by DFG as COVID-19 task force (<https://www.nfdi4health.de/index.php/task-force-covid-19-2/>). This meta platform, that will be publicly accessible, bundles information on relevant clinical and epidemiological studies in Germany, their publicly available study documents and results, e.g. measured parameters, as well as other information about the studies.

<sup>1</sup><https://doi.org/10.17881/covid19-disease-map>

FAIRDOM HUB

Home / Projects Index / COVID-19 Disease Map

## COVID-19 Disease Map

Here we share resources and best practices to develop a disease map for COVID-19. The project is progressing as a broad community-driven effort. We aim to establish a knowledge repository on virus-host interaction mechanisms specific to the SARS-CoV-2. The COVID-19 Disease Map is an assembly of molecular interaction diagrams established based on literature evidence.

Programme: Disease Maps  
FAIRDOM PAL: No PALs for this Project  
SEEK ID: <https://fairdomhub.org/projects/190>  
Project created: 27th Mar 2020  
Public web page: <http://doi.org/10.17881/covid19-disease-map>  
Organisms: Severe acute respiratory syndrome coronavirus 2, Homo sapiens

Related items

People (267) Institutions (128) Data files (1) Models (25) Publications (87) Documents (7)

Reactome pathways for SARS-CoV infections

The pathways focused on SARS-CoV infections curated in Reactome. These pathways are work-in-progress.

Creators: Marc Gillespie, Robin Haw, Peter D'Eustachio  
Submitter: Marek Ostaszewski  
Model type: Graphical model  
Model format: SBGN-ML PD  
Environment: Not specified

Organism: Homo sapiens  
Investigations: No Investigations  
Studies: No Studies  
Modelling analyses: No Modelling analyses

Created: 4th May 2020 at 11:05. Last updated: 13th Oct 2020 at 12:49

Nsp4 and Nsp6 interactions (COVID-19 Disease Maps)

Interactions of Nsp4 and Nsp6 proteins of SARS-CoV-2.

Creators: Armau Montagud, Miguel Ponce-de-Leon  
Submitter: Marek Ostaszewski  
Model type: Graphical model  
Model format: SBML  
Environment: Not specified

Organism: Severe acute respiratory syndrome coronavirus 2  
Investigations: No Investigations  
Studies: No Studies  
Modelling analyses: No Modelling analyses

Created: 20th Apr 2020 at 15:16. Last updated: 21st Aug 2020 at 14:59

Orf3a interactions (COVID-19 Disease Map)

**FIGURE 1:** Screenshot of the COVID-19 Disease Map project space on the SEEK-based FAIRDOMHub (<https://fairdomhub.org/projects/190>).

This COVID-19 study hub is partially based on the SEEK software [3], our de.NBI service software platform for collaborative projects, to support the data storage and exchange. SEEK will be used to store and make accessible study documents, such as study protocol templates or data dictionaries, as well as information on study metadata structures like data models to describe subjects and their measured values, their clinical parameters and treatment outcomes and similar information. Additionally, direct links to primary resources and websites of the studies will be included. Using the SEEK web services this information also will directly feed a COVID-19 study search portal, implemented by our partners at ZB MED in Cologne, Germany. Standardization of metadata

describing the studies, as well as their subjects and results are also part of the HITS activities in the nfdi4health COVID-19 task force, contributing to the aim of making the studies and their content 'FAIR'.

#### LITERATURE CURATION OF COVID-19 DATA

##### COVID-19 kinetics data extraction for SABIO-RK

Since 2006, the relational database SABIO-RK (<http://sabiork.h-its.org/>) [4] offers structured and annotated data about biochemical reactions and their reaction kinetics to support modellers in simulating complex metabolic networks and experimentalists looking, e.g. for experimental conditions or alternative re-

actions of an enzyme. All the data is manually curated, originating mainly from scientific literature by being read twice: firstly by students extracting and entering the data to a pre-database and secondly by biocurators controlling and further annotating the data before inserting them to the final database. This manual curation process guarantees a high degree of correctness and completeness of the data which then becomes freely available via a web-based search interface enabling data export in diverse formats as well as by web services for programmatic access also being implemented in systems biology tools.

FAIRDOM HUB

Home / Investigations Index / Consortium for Clinical Characterization of COVID-19 by EHR (4CE)

## Consortium for Clinical Characterization of COVID-19 by EHR (4CE)

Consortium website: <https://covidclinical.net/>  
Slack: <https://c19i2b2.slack.com/>  
Owner: Nils Gehlenborg (nils@hms.harvard.edu)  
i2b2 tranSMART Foundation Call to Action: <https://transmartfoundation.org/covid-19-call-to-action/>

SEEK ID: <https://fairdomhub.org/investigations/376>  
Projects: COVID-19 related studies and tools in Germany

Selected: Consortium for Clinical Characterization of COVID-19 by EHR (4CE) (Investigation)  
Description: Consortium website: <https://covidclinical.net/> Slack: <https://c19i2b2.slack.com/> Owner: Nils Gehlenborg...  
SEEK ID: <https://fairdomhub.org/investigations/376>

Consortium for Clinical Characterization of COVID-19 by EHR (4CE)

- General Information
- Websites
- Chats
- Phase 1
  - Instructions
    - GitHub
    - LOINC Codes
    - Lab list
  - COVID19\_Data\_Files\_Description
    - covid19i2b2/COVID19\_Data\_Files\_Description.txt at master · GriffinWeber/covid19i2b2 · GitHub
    - BIDMC\_IRB\_Application.pdf
      - covid19i2b2/BIDMC\_IRB\_Application.pdf at master · GriffinWeber/covid19i2b2 · GitHub
  - Phase 1.1
    - Analysis of hospitalized COVID-19 patients in the Mount Sinai Health System using electronic medical records (EMR) reveals important prog
  - Data Dictionary
    - Lab List
    - File Descriptions
- Phase 2

**FIGURE 2:** First pilot prototype for structuring publicly available documents and information from COVID-19 studies in the nfdi4health COVID-19 study hub based on the SEEK installation FAIRDOMHub (work in progress).

Since the manual curation process is very time consuming (in comparison to data acquisition by text mining) we're not trying to keep up with the data overload accumulating in the last decades but instead are displaying exemplarily the band width of kinetic data in the scientific literature including all organisms and pathways, as well as metabolic and signaling data. Moreover, we offer data collection on demand and have therefore, in the course of the de.NBI funding, set up the freely available data curation service.

In order to support scientists in the search for drugs against the SARS-CoV-2 virus causing COVID-19 but also other coronaviruses like SARS- and MERS-CoV we've put a focus on collecting kinetic data (especially the inhibition values IC50 and Ki) of papers describing promising antiviral agents like e.g. inhibitors of the main protease 3CLpro or the papain-like protease (PLpro) by using the CORD-19 dataset [5].

Within the BMBF-funded sister project 'SABIO-VIS', different ways of visualization of query results in SABIO-RK are implemented. The focus of the project is to improve the search functionalities and the navigation through the database content as well as to get a fast and flexible overview of available data sets. Using a broad variety of visualization concepts,

an overview about for example data for RNA viruses in SABIO-RK are presented in Figure 3.

#### Comparative studies of literature triage for SABIO-RK

Within the BMBF-funded 'DeepCurate' project to support and improve the literature curation process in SABIO-RK there is one focus on analyzing and optimizing the paper triage system using eye tracker methods. Due to the increasing interest in coronavirus-related data in SABIO-RK the project was extended to analyze the paper triage process for COVID-19 relevant publications by comparing the two domains of metabolic and virus-related literature. In database curation, information is extracted from unstructured scientific documents, normalized, and integrated into databases for structured



FIGURE 3: Screenshot of different visualization concepts of RNA virus data in SABIO-RK.

access. Document triage is the first step in database curation, which deals with how human domain experts choose curatable source documents based on criteria required for a specific database. We conduct an empirical study to fulfill our ergonomic purpose sending effective feedback to the curators while using two different sets of source documents: one consists of literature from the core metabolic reaction domain of the SABIO-RK database, while the new domain contains documents on coronavirus. In this case,

the document triage is a screen-based process, and we use an eye-tracker to collect low-level behavioral data like gaze coordinates, and pupillometric features. We use these features to interpret how the expert curators use visual search and reading skills in a manual triage process. We also control for the curators' cognitive load within the domains of curation. We find that the curators have a significantly higher cognitive load for the triage task of the new COVID-19 virus literature. We measure the cognitive load using

pupil diameter measurement. A representative graph from a curator is given in Figure 4A. In Figure 4B, we also present a representative figure to show how an expert curator uses the reading and searching strategy for the process of a COVID-19 document triage. We compute this graph using the number of fixations (represents the reading activity) and the number of saccades (represents the visual search activity) at one point of time in the course of the process of triage.

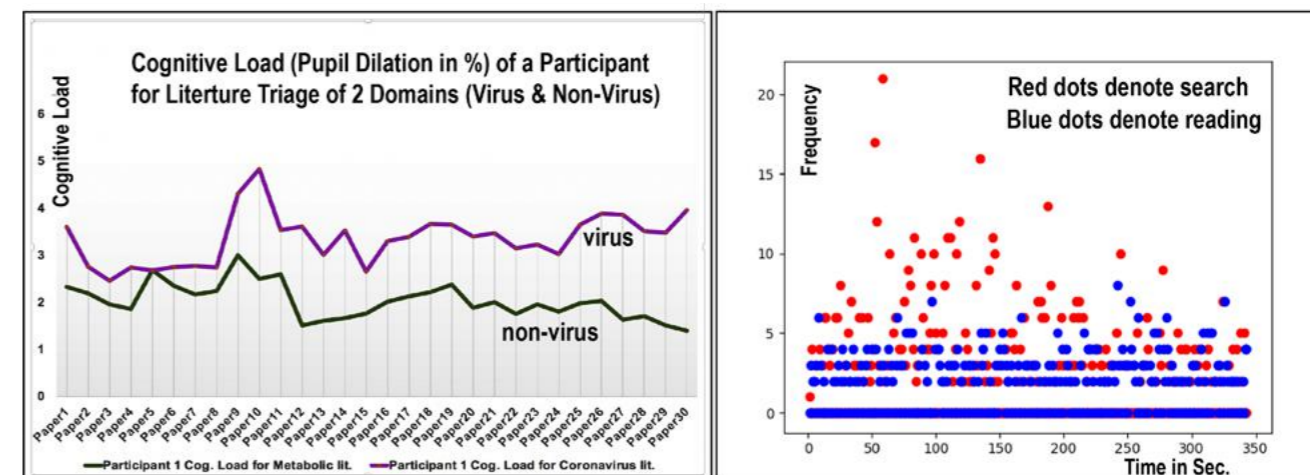


FIGURE 4: (A) left: Cognitive load (B) right: Reading and searching of a participant for triage activity for two domains for 30 papers; participant for one paper.

## CONCLUSION

We have described the data management platform FAIRDOMHub/SEEK (ELIXIR UK/ELIXIR DE) as a flexible platform for storing and sharing research data within national and international COVID-19 research communities and for supporting the FAIRness of their data. The ability to react fast to new scientific challenges by adapting the data management platform and its underlying SEEK software is demonstrated especially for newly es-

tablished and fast developing projects such as the international COVID-19 Disease Map initiative and national German efforts within NFDI. Together with standardization activities for data and metadata [6] this creates data infrastructures and 'FAIRification' services for COVID-19 related research.

Furthermore, we have described how literature curation of SARS-CoV-2 and other virus-related literature benefits SABIO-RK's users and increases our gen-

eral understanding of the curation process. Additionally, with the help of artificial intelligence and sensor technologies, some initiatives are also taken to speed up the curation process in the COVID-19 scenario, without compromising on the curation quality.

Here, we highlighted the focus of the de.NBI Systems Biology Service Center (de.NBI-SysBio) on highly flexible tools and services as well as on high quality and FAIR data.

## REFERENCES:

- [1] Sci Data 2020;7(1):136. DOI: 10.1038/s41597-020-0477-8. [2] Nucleic Acids Research 2017;45(D1):D404-D407. DOI: 10.1093/nar/gkw1032. [3] BMC Systems Biology 2015;9:33. DOI: 10.1186/s12918-015-0174-y. [4] Nucleic Acids Res 2018;46(D1):D656-D660. DOI: 10.1093/nar/gkx1065. [5] ArXiv [Preprint] 2020;arXiv:2004.10706v4. [6] Encyclopedia of Bioinformatics and Computational Biology, Vol. 2, 2019, Pages 884-893. DOI: 10.1016/B978-0-12-809633-8.20471-8.

**AUTHORS:** Maja Rey<sup>1</sup>, Andreas Weidemann<sup>1</sup>, Ulrike Wittig<sup>1</sup>, Dorotea Dudas<sup>1</sup>, Sucheta Ghosh<sup>1</sup>, Martin Golebiewski<sup>1</sup>, Xiaoming Hu<sup>1</sup>, Wolfgang Müller<sup>1</sup>

<sup>1</sup>Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnengweg 35, 69118 Heidelberg

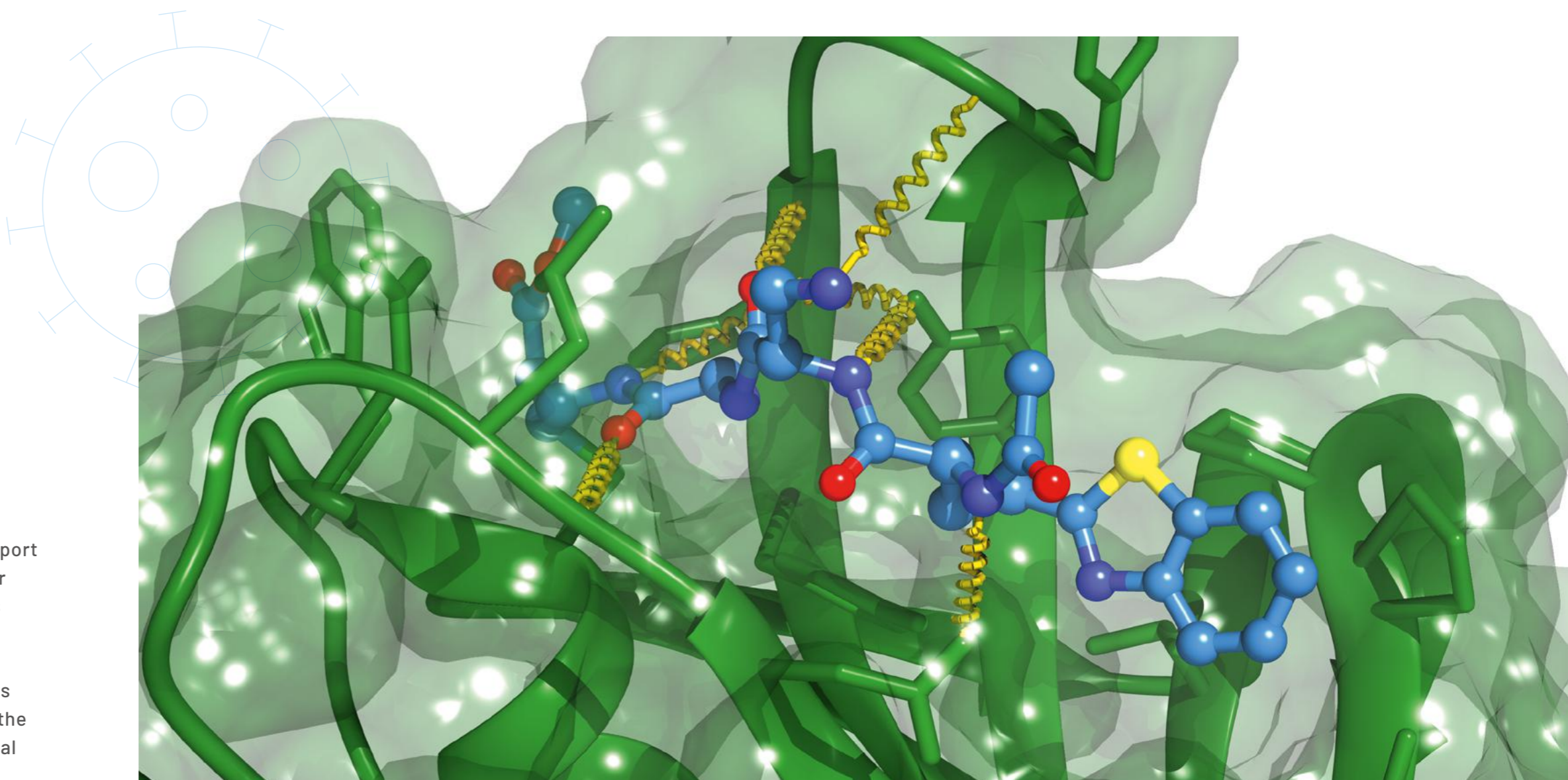
# BIOINFORMATICS ASSISTS DISCOVERY OF DRUGS AGAINST SARS-CoV-2

To concentrate the efforts on drug development against SARS-CoV-2, researchers from all over the world have started collaborating and sharing their data. In this challenging scenario, bioinformatics came out as one of the essential solutions to analyze SARS-CoV-2 data as it provides insights into virus behavior leading to the identification of potential drugs and drug targets against SARS-CoV-2.



# COVID-19 DRUG RESEARCH – GAINING INSIGHTS WITH PROTEINSPLUS

The ProteinsPlus web server provides excellent support for in-depth analyses of protein structures and their ligands. Its innovative structure-based design tools enable navigation through protein pockets, quality assessments, site comparison, interaction visualization etc. in a highly intuitive user interface. In this article, we present their successful application for the repositioning of already known drugs with a potential effect on promising SARS-CoV-2 targets.



The outbreak of the SARS-CoV-2 pandemic poses unprecedented challenges to physicians, medicinal chemists and molecular biologists worldwide. A previously unknown virus threatens the health systems and economy and leads to ever-rising numbers of infections and deaths. Scientists quickly initiated projects coping with the task of discovering reliable drugs to fight this new virus strain.

Drug repurposing (also known as drug repositioning), i.e., the application of approved drugs for new therapeutic purposes,

was adopted as a potential strategy to fight SARS-CoV-2. Cellular, biochemical and structure elucidation workflows in high-throughput settings were set up to explore the possibilities of drug repurposing. In this article, we present how the ProteinsPlus tools support such workflows and lead to important findings which might accelerate the search for known drugs with impact on the therapy of SARS-CoV-2 infections. Identifying similar binding sites of already well-known targets with known drugs might help to prioritize drugs with inhibiting

effect on related SARS-CoV-2 targets. Moreover, the discovery of new binding sites might reveal new starting points for repurposing. Herein, we present the ProteinsPlus tools SIENA, PoseView, GeoMine, HyPPI and DoGSiteScorer (Figure 1) to highlight application domains for the investigation of targets and, subsequently, the establishment of drug repurposing strategies. We first introduce these tools in more detail. Subsequently, we apply them to two SARS-CoV-2 proteases which are crucial for viral replication.

## A SUBSELECTION OF PROTEINSPLUS TOOLS

**ProteinsPlus** [1, 2] is a web portal enabling the analysis of protein structures focusing on protein-ligand interactions.

**SIENA** [3] performs an automated assembly and preprocessing of protein binding site ensembles. Starting with a single binding site, SIENA searches the Protein Data Bank (PDB) for alternative conformations of the same or sequentially closely related binding sites. The

method is based on an indexed database and a new algorithm for the detection of protein binding sites conformations. SIENA provides the user with a sequence alignment of the binding sites as well as superimposed protein structures which are, apart from the transferred coordinates, equal to the original files from the PDB and thus contain all structural details and further information.

**PoseView** [4] automatically creates 2D diagrams of interaction patterns of ligand binding sites according to chem-

ical drawing conventions. Interactions between the molecules in a binding site are estimated based on their 3D arrangement by a built-in interaction model that is based on atom types and simple geometric criteria. Using only the topological information, the interacting partners are arranged in an easily ascertainable layout that enables quick comparisons between binding sites.

**GeoMine** [5, 6] enables textual, numerical and 3D searching with full chemical awareness in protein-ligand interfaces of

the entire PDB. A geometric query may be constructed of binding site elements, i.e., atoms, bonds and interactions, whose relationships are described by distances and angles. All binding site elements can be further specified by different properties. GeoMine uses a database with a tailor-made index that is derived from the PDB and enables a deterministic and precise search. Queries are typically processed in the range of seconds to a few minutes.

HyPPI classifies interaction types of protein-protein complexes into permanent, transient or crystal artifact. Permanent protein-protein complexes are only stable in their complexed state and the subunits would denature upon complex dissociation. Transient protein-protein complexes are stable in the complexed as well as in the monomeric form depending

on the necessary function of the complex. Crystal artifacts have no biological function and are formed during the crystallization process.

DoGSiteScorer [7] is a grid-based method which uses a Difference of Gaussian filter to detect potential binding pockets – solely based on the 3D structure of the protein – and splits them into sub-pockets. Global properties, describing the size, shape and chemical features of the predicted pockets, are calculated and reported with two druggability measures.

#### THE SARS-CoV-2 MAIN PROTEASE

The first crystal structure of one of the currently most interesting targets, the viral main protease  $M^{pro}$  (also known as 3CL $^{pro}$ ) in complex with a binding site-defining ligand was published in February 2020

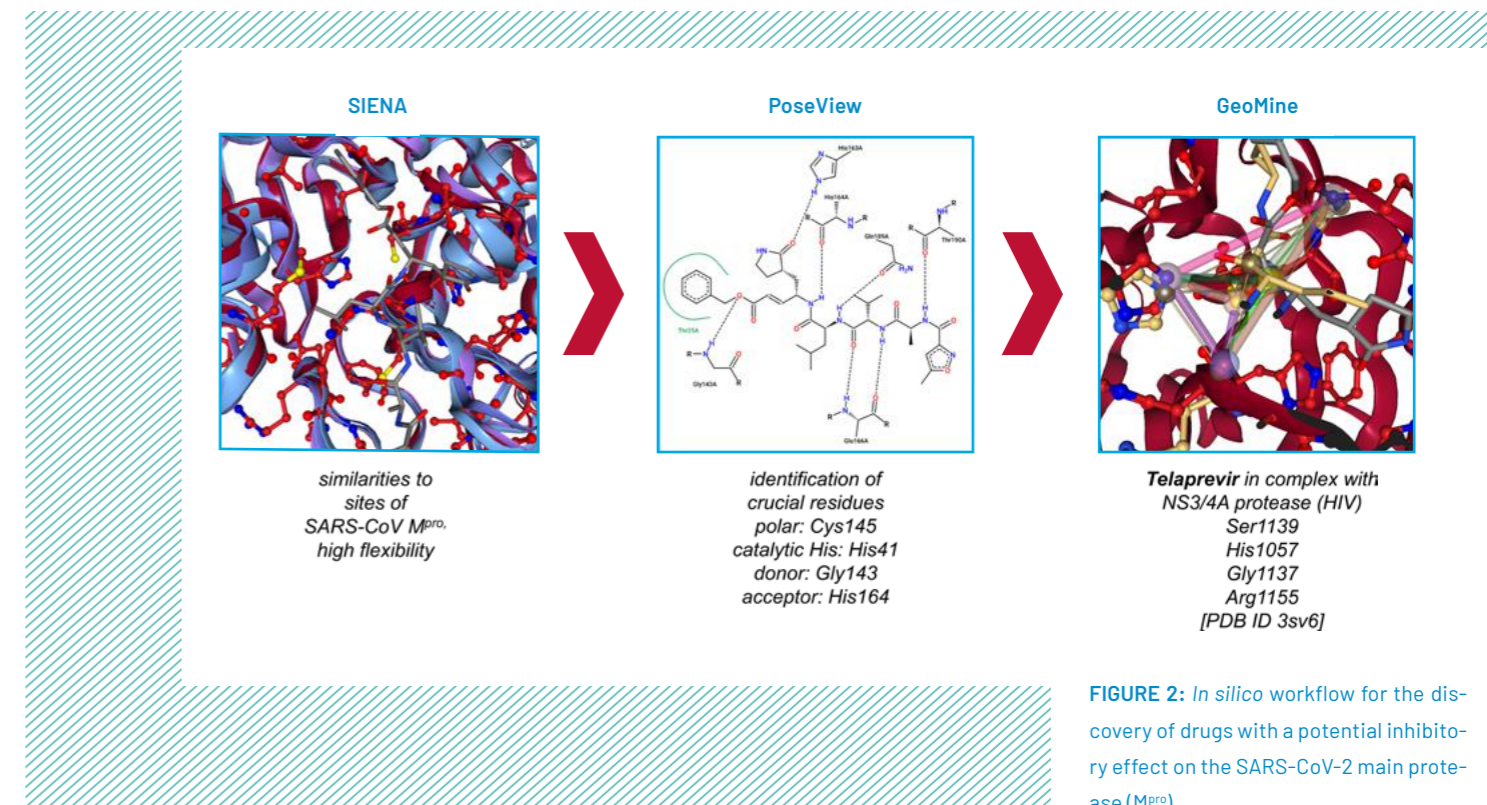
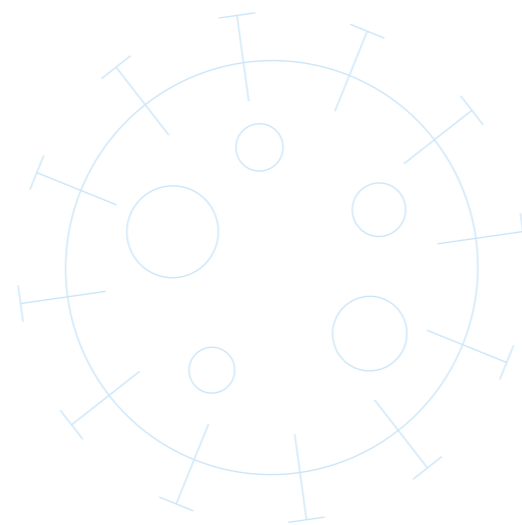


FIGURE 2: *In silico* workflow for the discovery of drugs with a potential inhibitory effect on the SARS-CoV-2 main protease ( $M^{pro}$ ).

(PDB ID 6lu7). We chose this target structure for our first application scenario (Figure 2). Using SIENA to identify related binding sites, we discover similarities to other viral proteases. However, after the removal of sequence duplicates, we only find similarities to SARS-CoV and further SARS-CoV-2  $M^{pro}$  structures. Hence, the knowledge of sequentially closely related protease binding sites is not sufficient to enable the discovery of related targets with known drugs for repurposing. However, we can already derive important conclusions concerning the flexibility of the binding site. With backbone RMSD values of up to 2.5 Å, our binding site of interest is highly flexible which should be considered in further drug design efforts.

Subsequently, we investigated interactions between  $M^{pro}$  and the crystal structure ligand using PoseView. This enabled us to generate a more property-centric comparison of the binding site of interest with GeoMine by the generation of user-defined queries. In the case of  $M^{pro}$ , we modeled the binding site based on the catalytic cysteine and histidine

residues and two hydrogen bond donor/acceptor functionalities in the proximity. Additionally, we used a ligand atom to define the ligand position. All distances between these features were modeled and a distance tolerance of 0.8 Å was used to cope with the high site flexibility. Applying this query for comparison against 377,714 sites, we found matches with the sites of 277 PDB entries. Within the results, we find a match with the HIV NS3/4A protease structure in complex with Telaprevir. Intriguingly enough, in August, a crystal structure of Telaprevir in complex with SARS-CoV-2  $M^{pro}$  was solved (PDB ID 6zrt). The alignment shows a high similarity of the residues involved in ligand binding. Although both proteins share significant binding site similarities, a purely sequence-based approach cannot retrieve the detected similarities. We compared this query to sites of a PDB subselection, only consisting of structures in complex with known drugs, to further decrease the number of relevant matches. This restriction led to similar sites in complex with known drugs of mammalian proteins, but also proteins

from different viral strains (e.g., Hepatitis C, Enterovirus A) which enable further in-depth investigations.

#### THE SARS-CoV-2 PAPAIN-LIKE PROTEASE

Another viral target of interest is the papain-like protease (PL $^{pro}$ ) from SARS-CoV-2 (e.g., PDB ID 6wu0). In this case, a more elaborate strategy had to be pursued (Figure 3). Applying SIENA to compare the target site to related binding sites, only matches with coronavirus proteins were found. Therefore, we followed the same approach as applied for  $M^{pro}$ , i.e., modeling binding site properties and comparing them with GeoMine. Promising similarities were found to the binding sites of the human proteasome. Intriguingly, it was shown before that proteasome inhibitors impair SARS replication. They are commonly used for the management of hematological malignancies. However, they show a considerable cytotoxic effect which hampers a repurposing strategy with the established workflow for the active site.

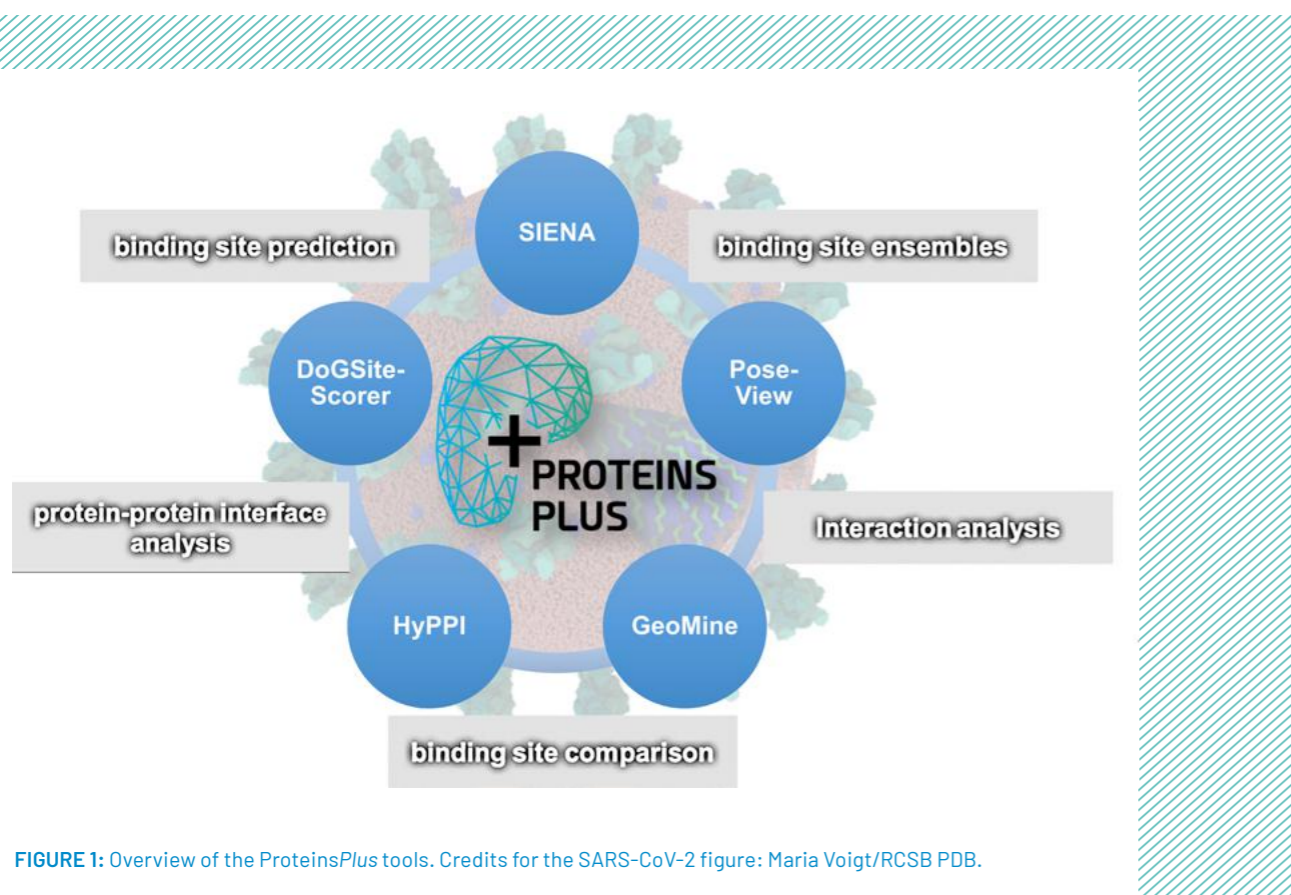
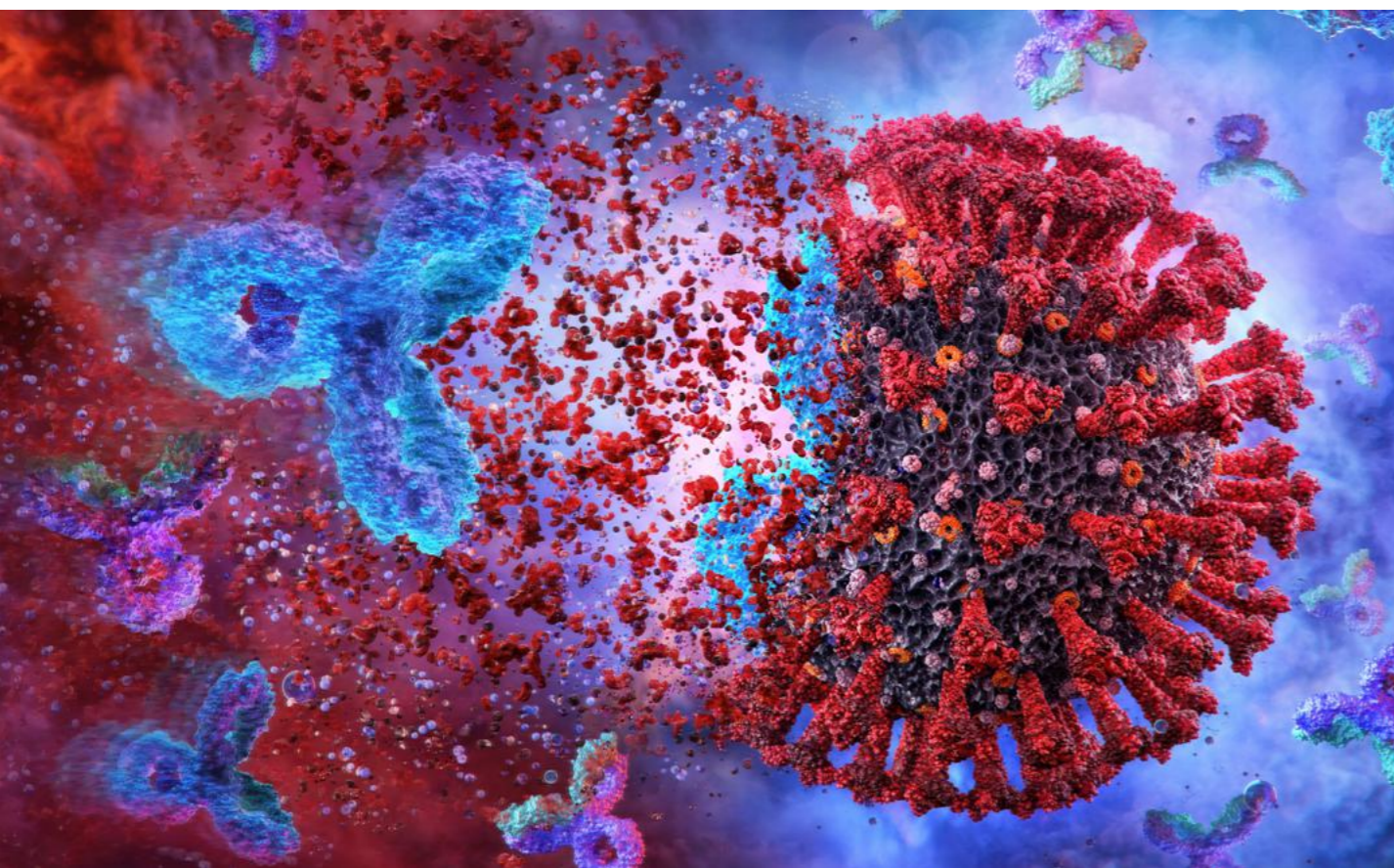


FIGURE 1: Overview of the ProteinsPlus tools. Credits for the SARS-CoV-2 figure: Maria Voigt/RCSB PDB.



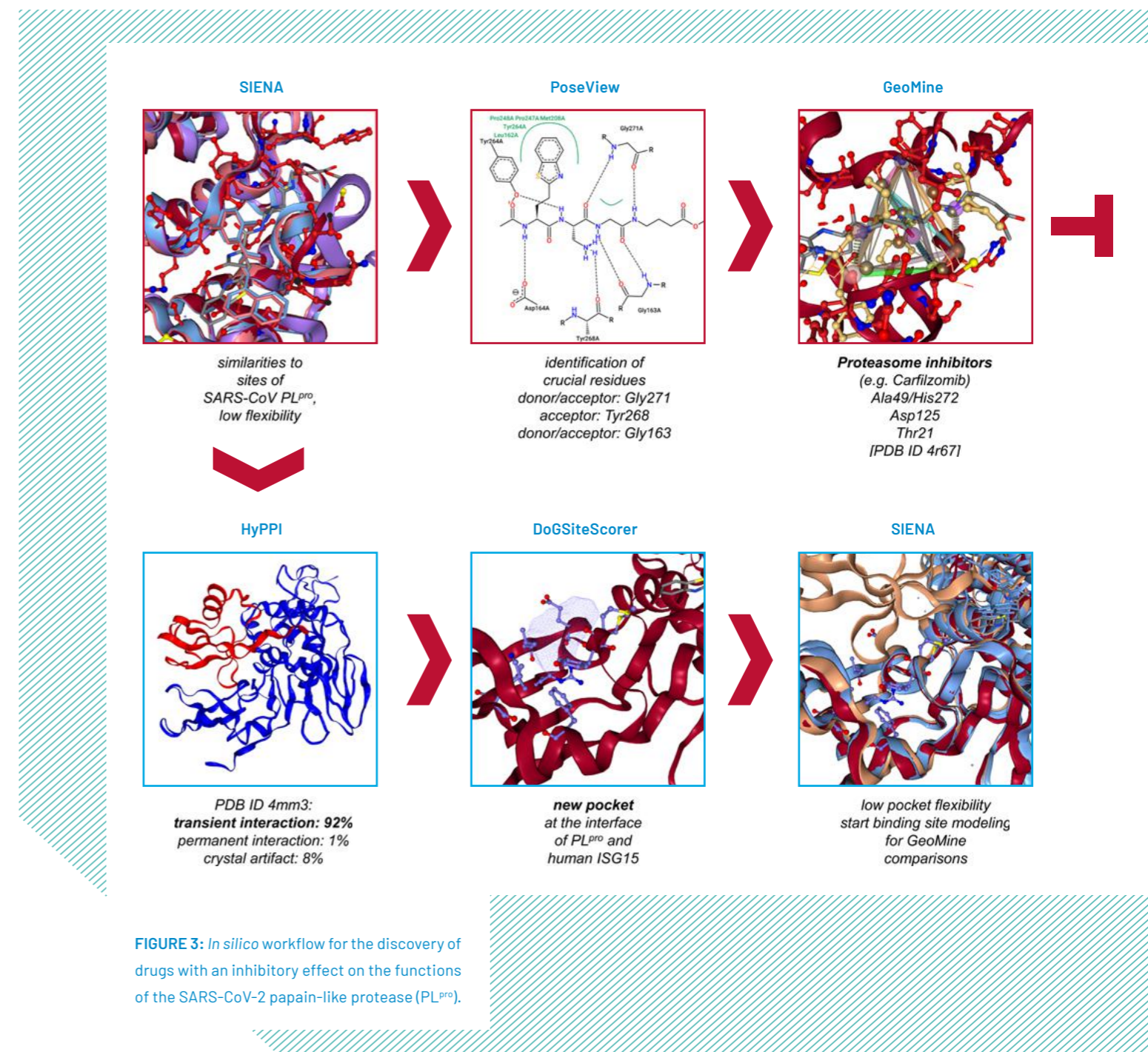
However, papain-like proteases from SARS-CoV are also known to antagonize the host innate immune response by reversing the post-translational modification of proteins conjugated with Ubiquitin-like interferon-stimulated gene product 15 (ISG15). A complex structure of both proteins (PDB ID 4mm3) could be extracted from our SIENA results. Due to the mode of action, we would suspect that this is a transient complex which should be easier to address by small molecules than a permanent one. The ProteinsPlus tool HyPPI also predicts a transient interface for the SARS-CoV structure in complex with human ISG15 with a probability of 92%. The binding of a molecule at the interaction interface of PL<sup>pro</sup> and human ISG15 might diminish the anti-immune effect. We applied DoGSiteScorer to our PL<sup>pro</sup> structure and identified 13 pockets. The pocket predicted most druggable is that of the co-crystal-

lized ligand. We screened the remaining pockets for those with interface residues of the above described transient protein-protein interaction. The residues of the fourth-smallest pocket are part of the interface. Although this pocket has a comparably lower druggability score, we applied SIENA for an initial flexibility analysis. Most of the similar sites showed an overall all-atom RMSD below 1 Å and modeling of the binding site residue properties to generate GeoMine queries is applicable. Thus, this newly identified pocket of PL<sup>pro</sup> might serve as a new starting point for further repurposing analyses.

#### PROTEINSPUS SUPPORTS THE SEARCH FOR DRUGS

During the last six months, the number of visitors to the ProteinsPlus web server increased by 80% to now about

3,600 visitors per month with more than 20,000 page views highlighting the supporting role of ProteinsPlus for COVID-19 research. In joint projects, such as our collaboration in a large consortium of scientists concerned with the identification of novel leads and drugs to cope with COVID-19 [8], ProteinsPlus tools can provide important insights. The two examples nicely illustrate the manifoldness and the complementing character of the ProteinsPlus tools. The web server helps to investigate protein binding sites in a comprehensive way and provides quick access to novel insights into yet unexplored targets. The application of the presented tools will hopefully foster the progress in our current endeavor to repurpose known drugs for application to the inhibition of the described proteases.



**FIGURE 3:** *In silico* workflow for the discovery of drugs with an inhibitory effect on the functions of the SARS-CoV-2 papain-like protease (PL<sup>pro</sup>).

#### REFERENCES:

- [1] Nucleic Acids Res 2017;45:W337-W343. DOI: 10.1093/nar/gkx333. [2] Nucleic Acids Res 2020;48:W48-W53. DOI: 10.1093/nar/gkaa235. [3] J Chem Inf Model 2016;56:248-259. DOI: 10.1021/acs.jcim.5b00588. [4] Bioinformatics 2006;22:1710-1716. DOI: 10.1093/bioinformatics/btl150. [5] Bioinformatics 2020;btaa693. DOI: 10.1093/bioinformatics/btaa693. [6] J Chem Inf Model 2017;57:148-158. DOI: 10.1021/acs.jcim.6b00561. [7] J Chem Inf Model 2012;52(2):360-72. DOI: 10.1021/ci200454v. [8] BioRxiv [Preprint] 2020. DOI: 10.1101/2020.05.02.043554.

**AUTHORS:** Christiane Ehrh<sup>1</sup>, Katrin Schöning-Stierand<sup>1</sup>, and Matthias Rarey<sup>1</sup>

<sup>1</sup> Universität Hamburg, ZBH – Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg



# BIOINFORMATICS TOOLS REVEAL SINGLE-CELL TRANSCRIPTOME DYNAMICS of SARS-CoV-2 infection

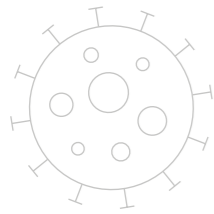
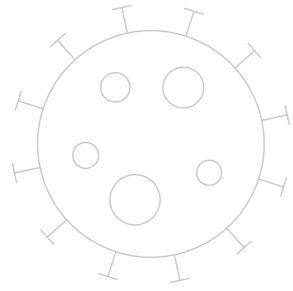
Gene expression profiles of SARS-CoV and SARS-CoV-2 infections in three human cell lines were established using bulk and single-cell transcriptomics in order to follow the transcriptome dynamics during the infection process and to identify potential drug targets. Applying the robust and reproducible PiGx pipeline we were able to identify potential candidate relevant for the infection and as a potential drug target.

The coronavirus disease 2019 (COVID-19) pandemic, caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is an ongoing global health threat with more than thirty million infected people, since its emergence in late 2019. Detailed knowledge of the molecular biology of the infection is indispensable for the understanding of the viral replication, host responses, and disease progression. Ultimately, this knowledge should lead to novel therapeutic interventions, which can ameliorate or suppress the disease progression.

We have teamed up with the labs of Prof. Dr. Markus Landthaler from MDC and Prof. Dr. Christian Drosten from Charité. The ex-

perimental teams infected three human cell lines with SARS-CoV (which caused the outbreak of the severe acute respiratory syndrome from 2002-2004) and the, currently pertinent, SARS-CoV-2. To get the insight about the dynamics of the infection, they have profiled the infection time course using bulk and single-cell RNA sequencing [1].

With Prof. Dr. Markus Landthaler, we already have a long standing collaboration, focused on investigating the cell response to herpes viral infections [2,3]. Through the collaboration we have developed tools and processes, and best practices for the analysis of single cell data of viral infections. We specifically focused on using transcriptomic dynamics to track the host cell response to the infection by inferring the activation of host cell signalling pathways.

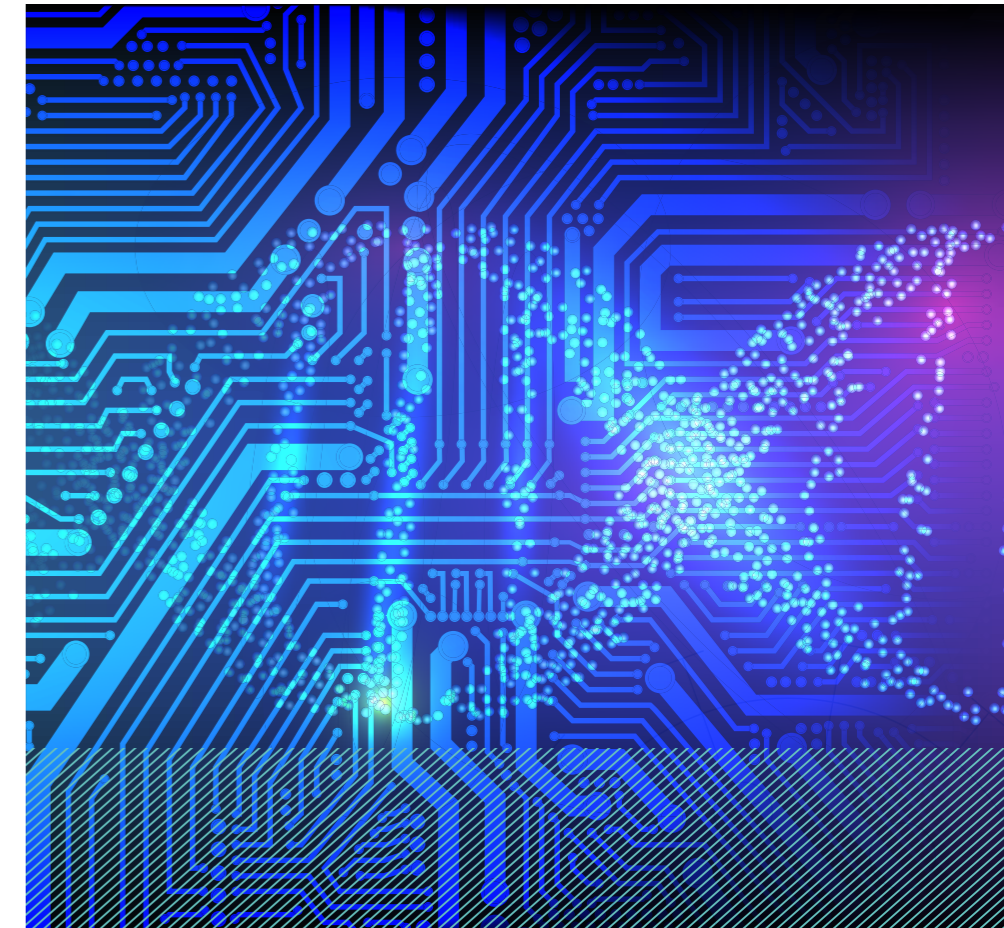


This previous experience in handling infectious single cell data, enabled us to promptly jump to the task at hand. We have analyzed the datasets using our robust and reproducible PiGx pipeline (<https://bioinformatics.mdc-berlin.de/pigx/>). PiGx pipelines envelop the best practices tools for specific genomics analyses (such as single cell RNA sequencing or ChIP-seq) with GUIX - a deterministic software management system [4]. GUIX provides a bit-for-bit reproducibility of the whole dependency tree, therefore enabling completely reproducible workflow execution on any computing environment.

Governed by our previous work, we have elucidated the dynamics of signalling pathway activation during the viral infection; and looked at the differences in the host cell response to SARS-CoV, with respect to SARS-CoV-2. These analyses revealed a broad induction of interferon stimulated genes. Unexpectedly, the induction was much more potent during

the infection with SARS-CoV-2 than SARS-CoV. In addition, the temporal resolution of transcriptional responses suggests that the interferon regulatory factor activities are followed by a strong activation of the nuclear factor- $\kappa$ B (NF- $\kappa$ B) pathway. The activation of the NF- $\kappa$ B pathways causes an induction of pro-inflammatory cytokines (such as interleukin 6, and tumor necrosis factor), which are biomarkers for the severity of the COVID-19 disease [2].

As a negative control, we have infected a cell line model (H1299) which does not express neither the SARS-CoV-2 receptor protein ACE2; nor the membrane bound TMPRSS2 enzyme, needed for activation of the SARS-CoV2 viral particle. To our surprise, both the SARS-CoV, and SARS-CoV-2 managed to infect a subpopulation of cells in the cell culture. The infection was however, extremely slow, giving us a snapshot of early transcriptional changes happening just after the



viral entry. Through the comparison of infected and uninfected subpopulations, we have noticed a higher expression of the heat shock protein 90 (HSP90) in the infected cells. HSP90 is a well known protein chaperone. It is one of the central cellular helper proteins facilitating protein folding, intracellular transport, and degradation of proteins. HSP90 also safeguards the cell from endoplasmic reticulum stress (ER stress). ER stress happens when a large amount of newly synthesized mRNAs start getting translated on the endoplasmic reticulum membrane - like when a virus starts the production of a large amount of its constituent proteins. Coronavirus infections readily cause ER stress. Following this observation, we surmised that by inhibiting HSP90 function might, we ameliorate the SARS-CoV-2 infection. HSP90 is one of the most upregulated proteins in many cancers, and several highly specific compounds have therefore been developed for perturbing its function. We have tested a set of inhibitors on both the cell line models, and primary patient lung epithelial cells infected with SARS-CoV-2. All of the HSP90 inhibitors significantly reduced the progression of viral infection, already in nanomolar concentrations. Furthermore, the compounds did not perturb the activation of the antiviral interferon response, while substantially decreasing the activation of the NF- $\kappa$ B, and the pro-inflammatory cytokines.

This project was made entirely possible through strong scientific collaborations established while doing basic biological research. The well polished interactions, between the three laboratories, enabled a prompt reaction, and adaptation to the new challenge. Our findings underscore the importance of well established laboratory practices and reproducible computational processes, as well as the existence of stable infrastructure, for the swift execution of high quality research.

**REFERENCES:** [1] bioRxiv 2020. DOI: <https://doi.org/10.1101/2020.05.05.079194>. [2] Genome Biol. 2017;18(1):209. DOI: 10.1186/s13059-017-1329-5. [3] GigaScience 2018;7(12):gij123. DOI: <https://doi.org/10.1093/gigascience/gij123>.

**AUTHORS:** Vedran Franke<sup>1</sup>, Kirstin Mösbauer<sup>2</sup>, Emanuel Wyler<sup>1</sup>, Nikolaus Rajewsky<sup>1</sup>, Christian Drosten<sup>2</sup>, Markus Landthaler<sup>1,3</sup>, Altuna Akalin<sup>1</sup>  
<sup>1</sup> Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, Hannoversche Str 28, 10115 Berlin  
<sup>2</sup> IRI Life Sciences, Institut für Biologie, Humboldt Universität zu Berlin, Philippstraße 13, 10115, Berlin  
<sup>3</sup> Institute of Virology, Charité-Universitätsmedizin Berlin and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin

# VIRTUAL SCREENING FOR SARS-CoV-2 DRUG DEVELOPMENT

## using open research and compute infrastructures

The European Galaxy Team, in collaboration with the UK's Diamond Light Source, has developed numerous computational tools and workflows for virtual screening of drug candidates against target proteins. In response to the COVID-19 pandemic, these were run using de.NBI infrastructure to develop a shortlist for chemical compounds with the potential to block the activity of the SARS-CoV-2 main protease.

As a result of the COVID-19 pandemic, many resources have been invested into finding a therapeutic for the disease. Efforts to develop vaccines to confer immunity on the population have in particular received attention. Another potential avenue would be developing an antiviral drug, which would chemically interact with one or more of the components of the virus, blocking its function and rendering the virus inactive. The SARS-CoV-2 virus is made up of 29 different proteins, several of which are considered potentially 'druggable' (i.e. it might be possible to interfere with their function by means of a drug). One which has received much attention is the so-called 'main protease'.

### WHAT IS THE SARS-CoV-2 MAIN PROTEASE?

A protease is an enzyme which has the function of breaking down other proteins into smaller parts. When the SARS-CoV-2 virus is replicated within a human cell,

many of the so-called 'non-structural proteins' are synthesised as part of long polypeptide chains. The main protease (Figure 1) has the essential role of splitting these polypeptide chains into the individual proteins, which have various different roles in the viral life cycle. If it were possible to block the action of the main protease (often referred to as Mpro) with a drug, the synthesis of these non-structural proteins would also be prevented, thus stopping the virus from replicating. The best way to do this is to find a molecule capable of binding to the active site of the protein, therefore preventing the polypeptide chain from entering and being broken down.

Another good reason to focus on the main protease is that the equivalent, very similar protein in the original SARS coronavirus was already the subject of much research, so there is already a knowledge base which helps to understand the SARS-CoV-2 main protease.



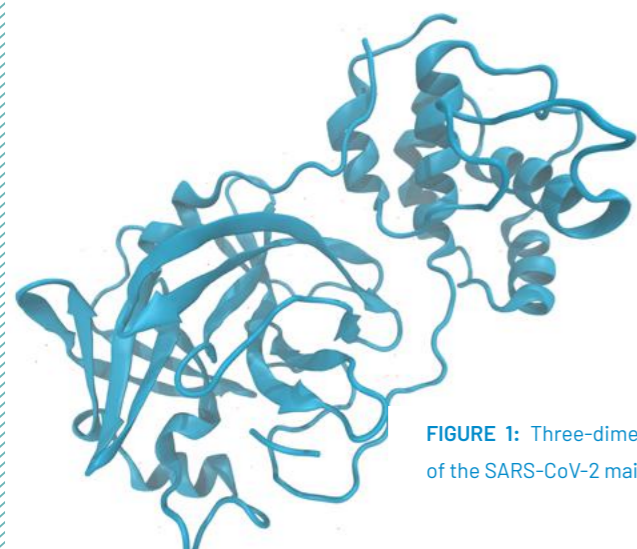


FIGURE 1: Three-dimensional structure of the SARS-CoV-2 main protease.

produce 'crystal structures' – effectively, a series of three-dimensional images of the protein, each with one of the fragments in the protein active site. The Diamond Light Source has developed a system named the Fragalysis fragment network which can be used to expand a list of fragments into a comprehensive list of drug candidates which contain one or more of the fragments within their molecular structure.

#### WHAT IS VIRTUAL SCREENING?

Drug development is a costly and time-intensive process; the number of molecules which could potentially act as drugs is unimaginably large and testing all of them experimentally is not feasible. 'Virtual screening' is a solution to this problem, in which a list of candidates is first tested 'virtually' (*in silico*) via computational methods (e.g. chemical simulations) and then the best-performing candidates can then be purchased from chemical suppliers for laboratory *in vitro* or *in vivo* tests.

#### HOW CAN A LIST OF CANDIDATES BE OBTAINED FOR VIRTUAL SCREENING?

In this project fragment screening was used, in collaboration with the Diamond Light Source (Figure 2) in Oxfordshire, UK, to help suggest molecules that would be likely to bind to the main protease binding site. Fragment screening is an experimental method in which protein crystals are exposed to many samples of small molecules ('fragments') to see which are successfully taken up into the protein structure, usually into the active site. These crystals can then be used to



FIGURE 2: Robot used for automated protein crystallography at the Diamond Light Source. Source: <https://www.diamond.ac.uk/industry/Industry-News/Latest-News/Synchrotron-Industry-News--MXnews4/MX-Sample-changer-upgrade-BART-.html>

The Diamond Light Source had already performed fragment screening of the main protease at extremely rapid speed by the end of March 2020, building on the crystal structure of the main protease, released in January this year by Prof. Dr. Zihe Rao. After generating their own protein crystals, the scientists at Diamond were able to confirm the protein active site was empty and accessible – perfect for fragment screening. Indeed, the first 600-crystal experiment was completed in 72 hours, through growing large numbers of crystals, optimising the soaking conditions, soaking and harvesting all 600 crystals and completing the data collection run. The list of fragment hits (successfully bound fragments) from this initial run and other details were pre-released on March 6<sup>th</sup>. By the 24<sup>th</sup> of March, the entire experiment was complete, and the results made publicly available, with a total number of active site fragments of 71.

An initial list of around 42,000 candidate molecules was assembled by using Fragalysis to elaborate from the initial fragment hits. The fragment network takes a large set of compounds (in this case, the fragment hits), and splits them up into smaller parts. These fragments form the nodes in a graph network. The edges between these nodes describe how the fragments of molecules can be linked together to make new molecules. From this information, we know how we can change a molecule by searching the network for new fragments to add to an initial hit, with transformations described along the edges in the graph network.

After this list of compounds had been produced using Fragalysis, some processing steps were required:

- Compounds can sometimes exist in different 'charge states'; i.e. some of the atoms may be positively or negatively charged, or they may all be neutral. Each of these will have a different

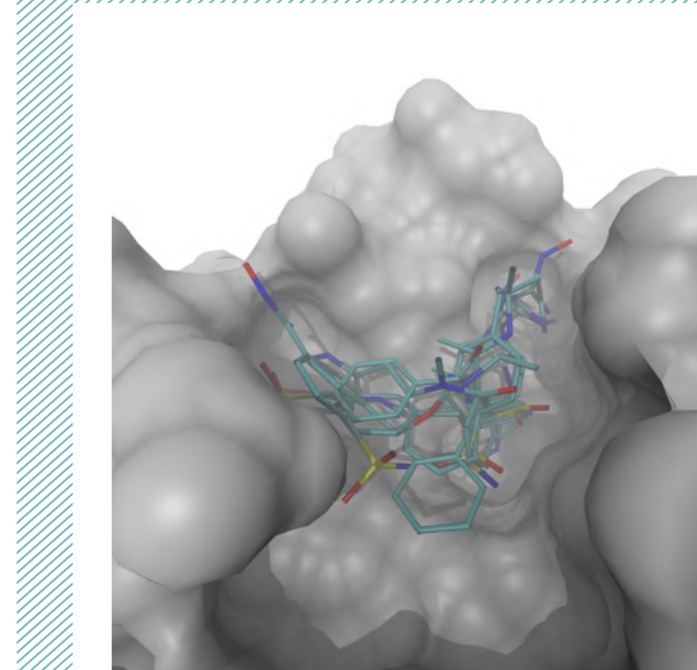


FIGURE 3: Some poses generated by rDock, overlaid on the protein active site.

**The flexibility of open research infrastructures, such as the Galaxy project, makes it possible to complete large projects at a short time-scale, as required to combat a public health emergency like the COVID-19 pandemic.**

chemical behaviour, so it was necessary to generate all charge states that could potentially exist in a typical biological environment.

- Molecules are generally depicted by chemists as 2D drawings on a page (or in a computer database) but in the real world they exist in three-dimensional space, just like the main protease for which the screening was performed. Therefore, 3D conformations (shapes) were also generated for each of the charge states.

The full list of compounds was then passed to the main part of the workflow, protein-ligand docking.

#### WHAT IS PROTEIN-LIGAND DOCKING?

One method that can be used for virtual screening is protein-ligand docking. (Ligand here refers to a small molecule, such as a drug, which binds to a protein.) This involves computationally simulating the optimal position of a molecule in the active site of the protein. The docking procedure can either generate a single 'pose' for the molecule, or multiple different possibilities – in this project, 25 poses were generated for each of the molecules generated by Fragalysis, using a piece of open-source software called rDock (Figure 3).

### HOW ARE THE DOCKING POSES EVALUATED?

One problem is that docking provides the poses without itself evaluating their quality (e.g. how realistic it is that the molecule binds in this way). Therefore, some other method is needed to determine pose quality. We used two different techniques:

- TransFS is a method based on deep learning techniques. There are large databases of protein-ligand binding data and structures available online which can be used to train an algorithm to distinguish 'good' and 'bad' binding poses. The model which is generated by this process can then be used to test new, unknown protein structures (like the docking poses generated) for the quality of the protein-ligand binding. A crucial element of the TransFS training procedure is that a large number of 'junk', randomly generated ligand structures are fed into the model, so that it learns to successfully classify poor binders as well as good ones.
- The SuCOS method compares the poses directly with the fragments. As each of the candidate molecules contains at least one of the fragments, the poses should be similar to the positions of the original fragments from the fragment screen.

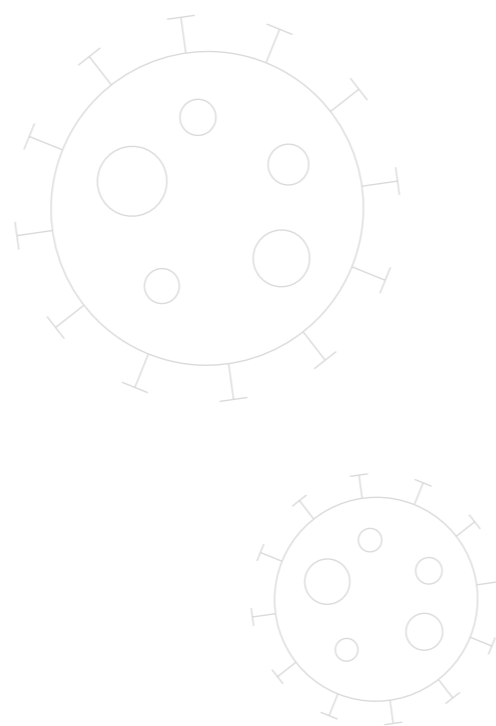
### HOW WAS THE PROJECT MANAGED?

Each of the steps of the analysis were combined into a workflow using Galaxy, a popular workflow management system in computational biology. This enabled the entire virtual screening procedure to be launched at the click of a button – new workflows could be easily started, as soon as new data became available. A schematic of the entire workflow is provided in Figure 4.

Furthermore, performing the screening as part of a Galaxy workflow allowed access to the compute resources provided by the European Galaxy server, <https://usegalaxy.eu>. These resources consist of a network of compute nodes, distributed across multiple European countries, to which individual calculations could be sent (Figure 5). As the nature of this project required a very high level of computational resources, the possibility of spreading the jobs across multiple servers was helpful in reducing the time needed for workflow completion. The project was carried out in close collaboration with the administrators of the European Galaxy server and of local compute centers.

As a result of this close collaboration, and the scale of the compute infrastructure available, initial results from the virtual screening were ready by the 4<sup>th</sup> of March – two days before Diamond publicly released the first results from the fragment screen. This demonstrated the flexibility of open research infrastructure, like the Galaxy project, which makes it possible to assemble and complete large projects at short time-scales, as required to effectively combat a public health emergency like the COVID-19 pandemic.

We have also publicised this project widely through the Galaxy community, including extensive documentation under <https://covid19.galaxyproject.org/cheminformatics> and download links for all data. A tutorial has also been written for the Galaxy Training Network (accessible at <https://training.galaxyproject.org/training-material/topics/computational-chemistry/tutorials/covid19-docking/tutorial.html>) which allows any user with a European Galaxy account and a basic understanding of the Galaxy platform to use and rerun the tools and workflows we have developed.



### WHAT ARE THE NEXT STEPS?

Obtaining SuCOS and TransFS scores for each of the poses allows ranking of all molecules. Our next aim, in collaboration with our experimental partners at the Diamond Light Source, is to purchase some of the top-ranking compounds from chemical suppliers in order to perform *in vitro* experiments, such as ligand binding assays, which will help to assess experimentally which of the compounds are effective at binding to the main protease.

#### Facts and figures:

- 7 crystal structures in initial virtual screening
- 42,000 candidate molecules for testing
- 159,000 candidates (including charge forms)
- 25 poses per candidate
- Up to 10,000 docking jobs running in parallel
- Total poses generated: 17 x 159k x 25 = 67 million
- 10k CPUs used, 20 GPUs, total of over 25 years CPU time

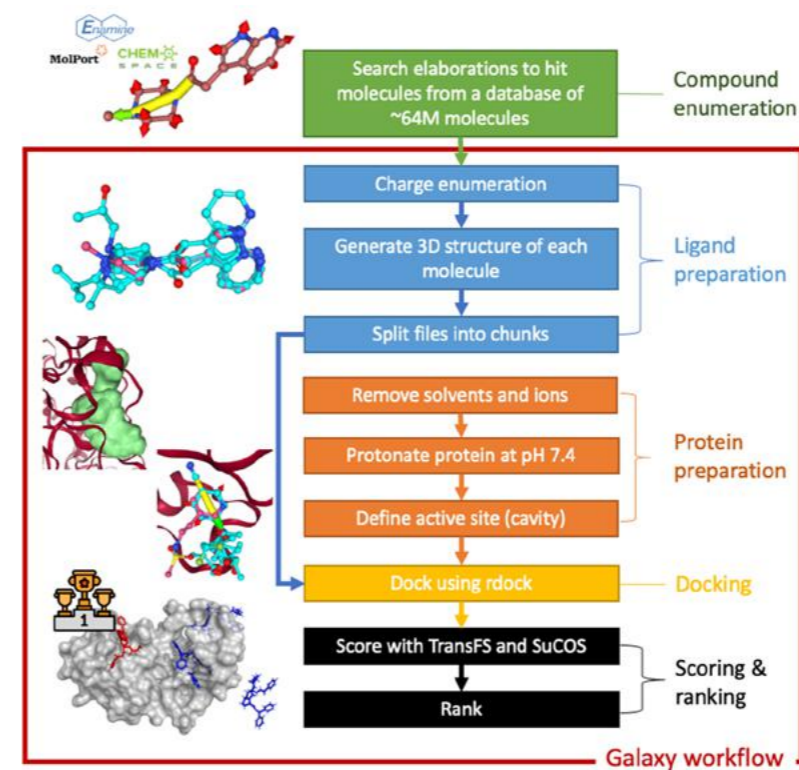


FIGURE 4: Schematic depicting the entire virtual screening workflow. Source: <https://covid19.galaxyproject.org/cheminformatics/>

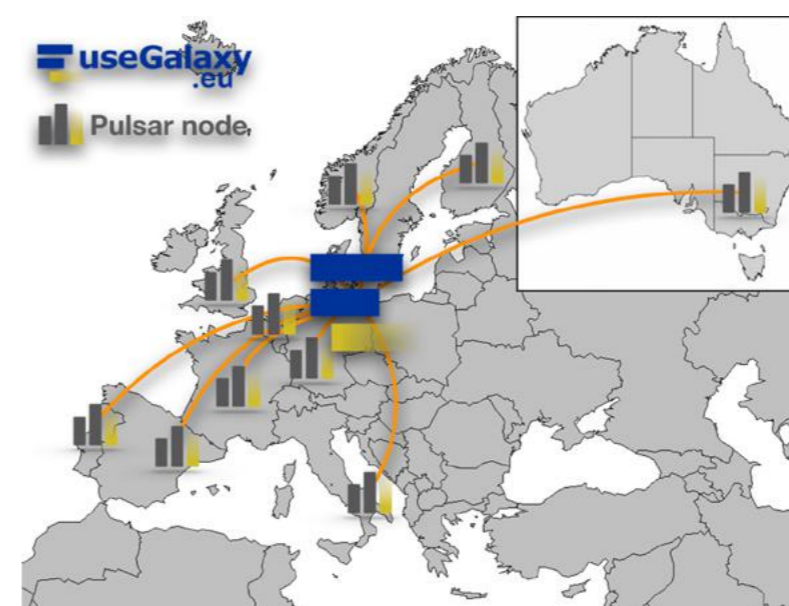


FIGURE 5: usegalaxy.eu compute nodes across Europe. Source: <https://pulsar-network.readthedocs.io/>

**AUTHORS:** Simon Bray<sup>1</sup>, Beatriz Serrano-Solano<sup>1</sup>, Björn Grüning<sup>1</sup>

<sup>1</sup> University of Freiburg, Department of Computer Science, Georges-Köhler-Allee 106, 79110 Freiburg



# IDENTIFY POTENTIAL DRUGS AND DRUG TARGETS AGAINST SARS-CoV-2 BY HOST FACTOR siRNA SCREENING

For the interactive platform for data analytics and image processing, KNIME ([knime.com](http://knime.com)), the Erfle Group at University Heidelberg offers de.NBI workflows for the systematic phenotyping of human cells. These workflows have been used for large scale imaging analysis in siRNA screens. In the current COVID-19 pandemic, we contribute to the search of suitable druggable target genes and proteins. This article gives further details about these KNIME workflows and the scientific approach.

## SEARCH FOR SARS-CoV-2 TREATMENT

At the end of last year, cases of severe pneumonia of unknown cause in Wuhan, China, associated with SARS-like acute lung failure (Acute Respiratory Distress Syndrome, ARDS), were observed on an increasing basis. In January 2020, next generation sequencing was used to identify a new coronavirus (SARS-CoV-2) [1] that causes this disease, which was then called COVID-19 [2]. SARS-CoV-2 is highly contagious, spread quickly and is responsible for the current global pandemic. Further COVID-19 patients with ARDS can deteriorate rapidly and die due to sepsis or multiple organ failure [1, 3, 4]. The overall mortality of hospitalized patients has been reported to be between 2.3% [5] and 11% [4]. As of now, there is no established therapy for COVID-19, clinical trials for vaccine and drug candidates are still ongoing with open outcome.

Treatment is mainly based on supportive and symptomatic measures, such as the treatment of organ failure and secondary infections [3, 5]. It is therefore absolutely essential to quickly develop alternative therapies that prevents SARS-CoV-2 infection and replication. Different attempts have been made in the past months: first, the development of passive vaccinations from plasma of recovered COVID-19 patients is one possibility. However, this only feasible for individual cases of special severity and no solution on a global basis. Second, antibodies have

been discovered that inhibit Ebola virus glycoproteins such as ZMapp, mAb114 [6] and REGN-EB3 [7], prevent the uptake of Ebola virus into host cells, and actually reduced mortality in Ebola patients [8]. However, such monoclonal antibody cocktails for the high demand, as now for COVID-19, cannot be produced in large quantities. Thus, a promising alternative for COVID-19 therapy is to use clinically tested active ingredients that are developed to treat other diseases. Information about such active ingredients and their targets is collected in extensive databases (such as Drugbank, ChEMBL, TTD, PharmGKB, BindingDB, PubChem, etc.). The experimental screening of active substances already led to identification of clinically approved active substances, which application inhibited the pathogenic coronaviruses MERS-CoV and SARS-CoV-1 was demonstrated in experimental model systems [9]. The group of Holger Erfle at University Heidelberg has joined forces with the groups of Rainer König and Sandra Ciesek at the University Hospitals in Jena and Frankfurt to set up the project 'SARSiRNA', in which a screening approach, druggable genome screening, is used to identify genes whose knockdown will potentially increase the survival of infected host cells.

## 'SARSiRNA' - OUR CONTRIBUTION TO FINDING SUITABLE DRUG TARGETS

Like all positive-strand RNA viruses, SARS-CoV-2 is dependent on host factors

for entry, viral gene expression, virion assembly and release [10]. If a part of this machinery is inhibited, the virus cannot complete its replication cycle which in turn prevents the destruction of the cell and virus spread. Using the druggable genome screening approach we aim to identify key host factors that play a role in the virus life cycle. Human cells have evolved in such a way that they can compensate for the failure of individual proteins to maintain cellular fitness. This means that individual host factors can, if necessary, be replaced by alternative pathways. In the SARSiRNA project, we use systems biology approaches to identify host factor dependency modules and to determine drug combinations with which these mechanisms can be blocked synergistically.

For the high throughput screening schematically depicted in Figure 1, we use a library targeting the druggable genome by small interfering RNA (siRNA) fragments [11, 12]. After transfection into the cells, siRNAs bind to the mRNA of a specific gene and inhibit their translation. First, this library is distributed to microwell plates. In each well one mRNA is targeted by specific siRNA sequences. For each gene three different siRNA sequences are applied, each in an individual well. As a host model, human epithelial cells (Heterogeneous human epithelial colorectal adenocarcinoma, Caco-2 cells) is used. This model system allows permissive SARS-CoV-2 replication and was used for proteome analysis and antiviral screening [13, 14]. Caco-2 cells are seeded into the microwell plates to allow the internalization of the siRNA, and subsequently infected with different SARS-CoV-2 isolates [15]. Under normal circumstances permissive cells infected with SARS-CoV-2 display cytotoxicity related morphological changes (CPE) that leads to cell death. However, if a host factor essential for the virus is missing, the cells

FIGURE 1: Schematic of the screen. Individual druggable genes are knocked down by siRNA sequences, the effects are observed microscopically in multiwell plates in high-throughput. The resulting images are analyzed with de.NBI tools. Protein interaction networks are constructed and possible drugs identified.

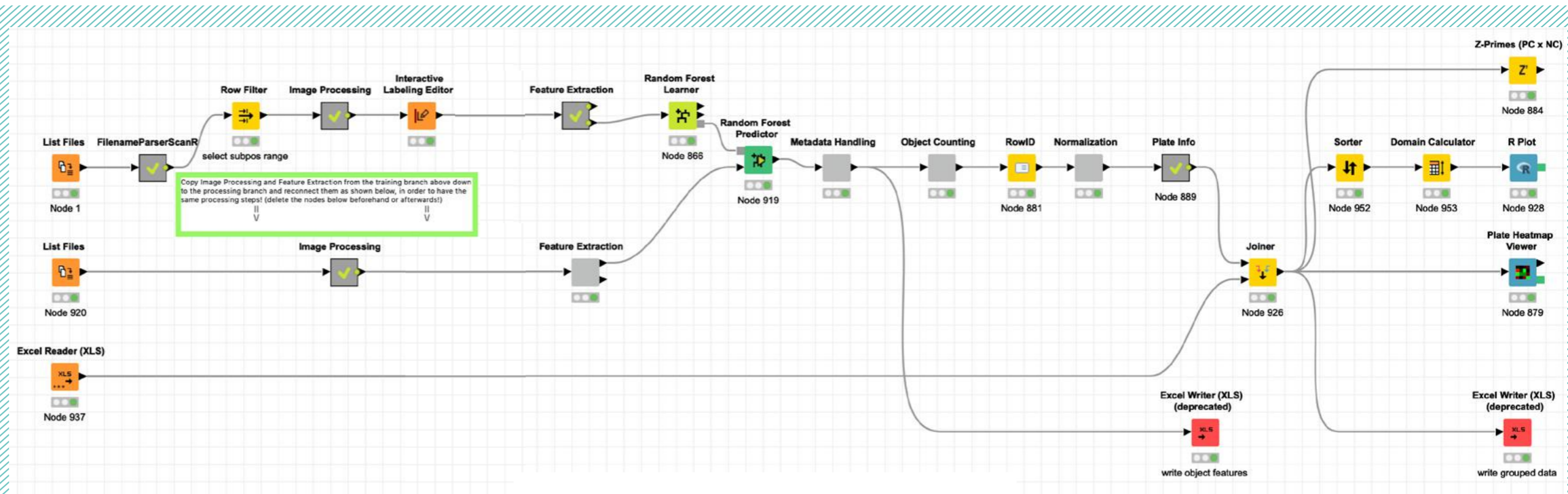
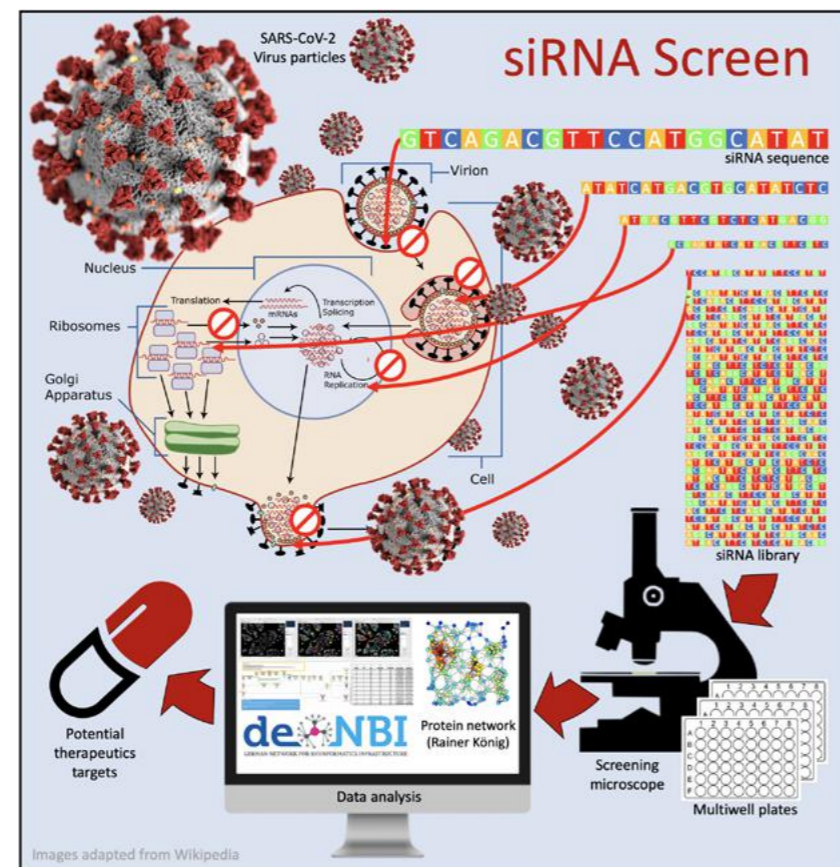


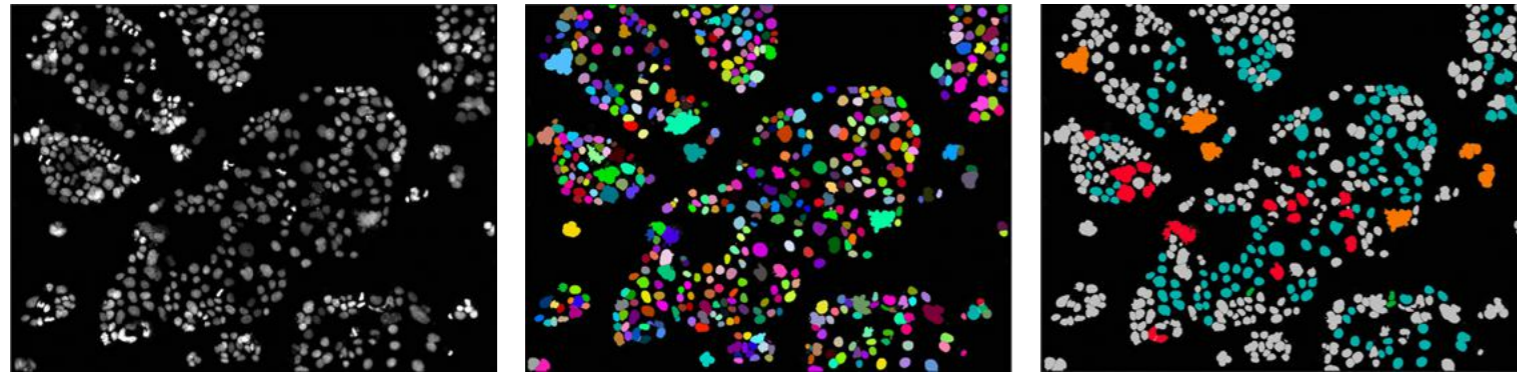
FIGURE 2: KNIME workflow for phenotypic classification. The workflow consists of interconnected nodes, each performing a processing step which can be configured by the user.

are protected and survive. This can be microscopically observed and quantified by high throughput image acquisition and image processing [16]. For the quantification of image data, we implement the systematic phenotyping workflows for KNIME provided by de.NBI (see below). Together with the department of Sandra Ciesek, the screening is conducted for over 9,000 druggable gene targets, each mRNA addressed by three different siRNA sequences independently in order to minimize off-target effects. From the proteomics data and image data a SARS-CoV-2-specific protein interaction network is constructed by the group of Rainer König. In this way, more effective active ingredients and combinations of active ingredients can be determined, the effects of which can be precisely embedded in the relevant pathways.

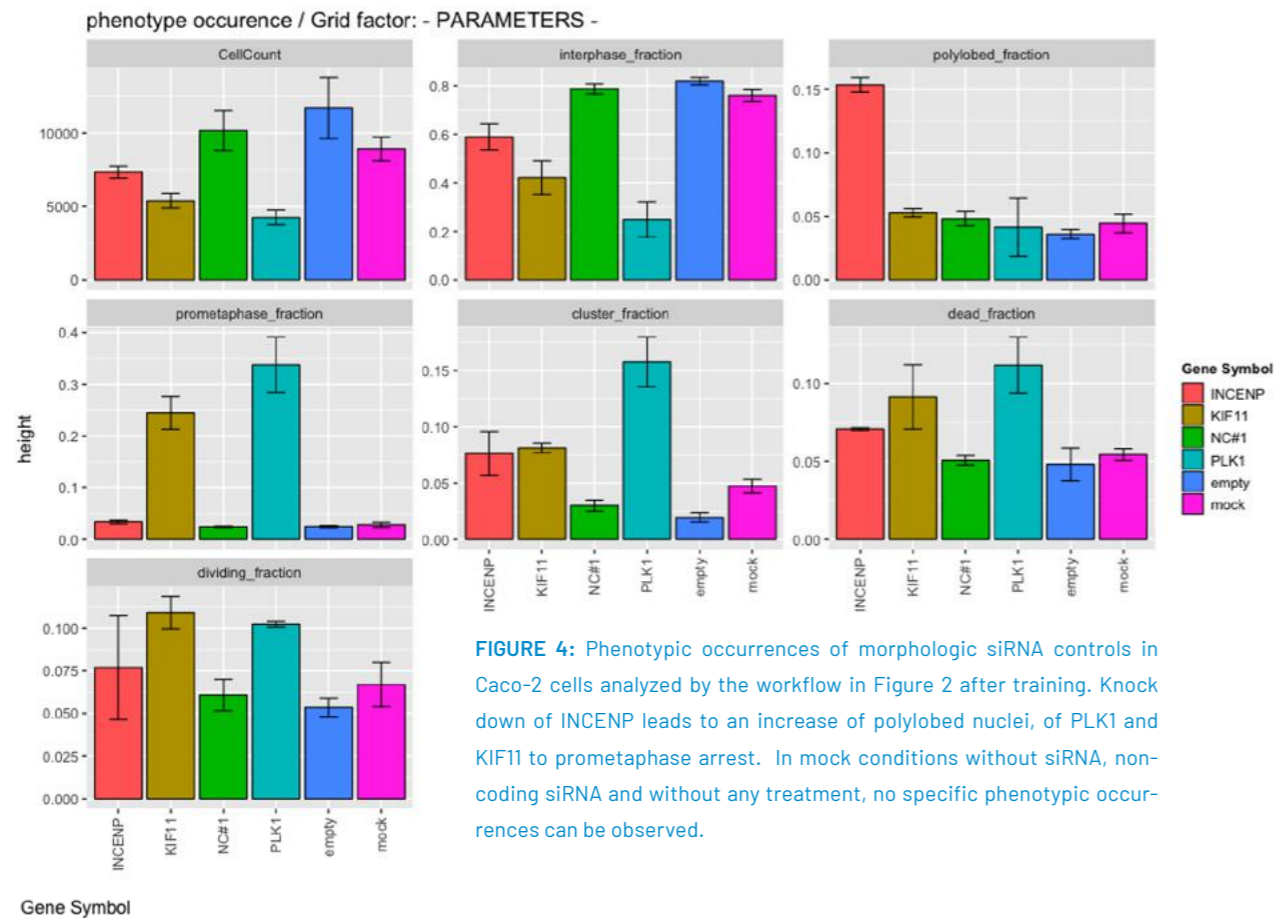
#### de.NBI TOOLS ARE CRUCIAL FOR SARSiRNA

The image processing and quantification workflows for systematic phenotyping used in this project are constructed in KNIME, the Konstanz Information Miner ([www.knime.org](http://www.knime.org)) and implement the KNIME image processing plugins based on ImgLib2 among others. In this interactive data analysis software, individual processing steps, represented as nodes, can be interconnected to complete processing pipelines and workflows. The workflows applied in this project is shown in Figure 2. Images are loaded and segmented for cellular objects, from which a set of features is calculated. These features can be statistical or geometrical features, Tamura features or Haralick features. In exemplary images the user

can define a set of phenotypic classes and assign exemplary objects to each class by marking these objects within the image. This is exemplarily shown in Figure 3. The workflow then includes learning the feature distribution for each class and classifying novel untrained objects accordingly. This results in a characterization of observed phenotypes for each condition. Furthermore, additional information of the screen, metadata, based on the filename and provided in additional spreadsheets is associated with the data. Tables listing the features and classifications for each object along with its associated metadata as well as grouped results can be written to spreadsheets for further analysis. Results can be visualized as shown in Figure 4.



**FIGURE 3:** Exemplary images. Left the original microscopic image is shown, in the middle the resulting segmentation. On the right side, the user can assign phenotypes to exemplary cells in order to train the workflow. All untrained cells are then classified based on this assignment.



**FIGURE 4:** Phenotypic occurrences of morphologic siRNA controls in Caco-2 cells analyzed by the workflow in Figure 2 after training. Knock down of INCENP leads to an increase of polylobed nuclei, of PLK1 and KIF11 to prometaphase arrest. In mock conditions without siRNA, non-coding siRNA and without any treatment, no specific phenotypic occurrences can be observed.

**REFERENCES:** [1] Lancet 2020;395(10223):497-506. DOI: [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5). [2] N Engl J Med 2020;382(8):727-33. DOI: [10.1056/NEJMoa2001017](https://doi.org/10.1056/NEJMoa2001017). [3] JAMA 2020;323(11):1061-1069. DOI: [10.1001/jama.2020.1585](https://doi.org/10.1001/jama.2020.1585). [4] Lancet 2020;395(10223):507-13. DOI: [https://doi.org/10.1016/S0140-6736\(20\)30211-7](https://doi.org/10.1016/S0140-6736(20)30211-7). [5] Clin Transl Med 2020;9(1):19. DOI: <https://doi.org/10.1186/s40169-020-00271-z>. [6] Lancet 2019;393(10174):889-98. DOI: [https://doi.org/10.1016/S0140-6736\(19\)30036-4](https://doi.org/10.1016/S0140-6736(19)30036-4). [7] The lancet infectious diseases 2018;18(8):884-93. DOI: [https://doi.org/10.1016/S1473-3099\(18\)30397-9](https://doi.org/10.1016/S1473-3099(18)30397-9). [8] N Engl J Med 2019;381:2293-2303. DOI: [10.1056/NEJMoa1910993](https://doi.org/10.1056/NEJMoa1910993). [9] Drug Discov Today 2020;25(4):668-688. DOI: <https://doi.org/10.1016/j.drudis.2020.01.015>. [10] J Virol 2003;77(15):8181-6. DOI: [10.1128/JVI.77.15.8181-8186.2003](https://doi.org/10.1128/JVI.77.15.8181-8186.2003). [11] Nature 2010;464(7289):721-7. DOI: <https://doi.org/10.1038/nature08869>. [12] Biotechnol J 2010;5(1):39-49. DOI: [10.1002/biot.200900226](https://doi.org/10.1002/biot.200900226). [13] Nature 2020;583(7816):469-72. DOI: <https://doi.org/10.1038/s41586-020-2332-7>. [14] Nature research [Preprint] 2020. DOI: [10.21203/rs.3.rs-23951/v1](https://doi.org/10.21203/rs.3.rs-23951/v1). [15] Int J Mol Sci 2020;21:4396. DOI: <https://doi.org/10.1101/2020.04.20.052258>. [16] Methods Mol Biol 2015;1251:59-66. DOI: [10.1007/978-1-4939-2080-8\\_4](https://doi.org/10.1007/978-1-4939-2080-8_4).

**AUTHORS:** Authors: Manuel Gunkel<sup>1</sup>, Rainer König<sup>2</sup>, Tuna Toptan<sup>3</sup>, Sandra Ciesek<sup>3</sup>, Holger Erfle<sup>1</sup>

<sup>1</sup> Advanced Biological Screening Facility, BioQuant, Heidelberg University, Im Neuenheimer Feld 267, 69120 Heidelberg

<sup>2</sup> Center for Sepsis Control and Care, Research Group Systems Biology of Sepsis, Jena University Hospital, Am Klinikum 1, 07747 Jena

<sup>3</sup> Institute of Medical Virology, University Hospital, Goethe University Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main

# COVID-19 RESEARCH is supported by de.NBI services

The COVID-19 pandemic has caused immense time pressure to analyse molecular biological data related to the new coronavirus. The German Network for Bioinformatics Infrastructure (de.NBI) helps to cope with this challenge – with bioinformatics tools, online training and a powerful cloud infrastructure. To take a closer look at the role of the de.NBI network in COVID-19 and SARS-CoV-2 research, Irena Maus, who is responsible for public relations at the de.NBI Administration Office, interviewed the de.NBI Coordinator Alfred Pühler and the Head of Node of ELIXIR Germany Andreas Tauch.



Alfred Pühler, de.NBI Coordinator

**IRENA MAUS:** Professor Pühler, in the last five years, the de.NBI network has built up an extensive bioinformatics infrastructure. What are the outstanding features of the de.NBI network?

**ALFRED PÜHLER:** The German Network for Bioinformatics Infrastructure was established by the Federal Ministry of Education and Research in 2015 to provide researchers in the life sciences with an appropriate infrastructure for the analysis of large amounts of data. The network currently counts more than 300 scientists and consists of 40 projects located in eight service centres. The eight service centres are thematically oriented and cover various sub-disciplines in the life sciences, including all aspects required for data-based medicine.

The infrastructure set up in the de.NBI network covers the areas of service, training and compute. In the service area, a variety of analysis programmes are available to evaluate life science data. Closely connected to the service area is the training area, which teaches



Andreas Tauch, Head of Node of ELIXIR Germany

er use of specific analysis programmes. In the training area, training courses for the use of the above-mentioned analysis programmes are provided. In total, the de.NBI network offers more than 70 training courses per year. In the last five years, more than 6,000 de.NBI users have already been trained. In the meantime, the compute area has gained special importance. Here, at six locations in Berlin, Bielefeld, Freiburg, Gießen, Heidelberg and Tübingen, thanks to an extensive BMBF funding, a federated de.NBI Cloud has been established, which is available free of charge to all life science researchers. This de.NBI Cloud enjoys general popularity. More than 300 projects could be calculated on the de.NBI Cloud. In the meantime, the 1000<sup>th</sup> de.NBI Cloud user has already been registered.

**IRENA MAUS:** Triggered by the corona pandemic, COVID-19 research has become very important. What is the role of the de.NBI network in COVID-19 research questions?

**ALFRED PÜHLER:** The current corona pandemic represents a major challenge for our society and therefore requires special attention by established scientific organizations. Molecular biological research on SARS-CoV-2 viruses should be advanced, epidemiological aspects of the infection process needs a better understanding and the course of COVID-19 diseases has to be analyzed. In addition, efforts must be made to develop drugs to treat patients and to produce suitable vaccines. All these research areas generate large amounts of data, which only reveal their value after careful bioinformatics analysis. The evaluation of large amounts of data is one of the main focuses of de.NBI and therefore, the network was prepared to step into the analysis of COVID-19 research data. The de.NBI network has thus passed the acid test. It was able to react at short notice to the current demand for data analysis programmes.

the users of de.NBI services in handling analysis programmes and scientific results achieved. In the compute sector, the de.NBI network plays an important role by establishing the de.NBI Cloud for data analysis of de.NBI customers.

**IRENA MAUS:** Can you characterize the areas of service, training and compute in more detail to generate a better understanding of the de.NBI network?

**ALFRED PÜHLER:** The service area of the de.NBI network provides almost 150 different services, most of them as analysis programmes and workflows that are intended for the interpretation of life science data sets. Potential de.NBI users are referred to the supervisors of the respective services via the de.NBI websites or the de.NBI helpdesk, with the goal to provide information for the prop-



**de.NBI**  
German Network für  
Bioinformatics Infrastructure

is a national, academic and non-profit infrastructure providing bioinformatics service and training to users in life science research and biomedicine in Germany and Europe.

**IRENA MAUS:** The de.NBI network has a large number of members. Which COVID-19 projects are in the hands of these de.NBI members?

**ALFRED PÜHLER:** Initial surveys in the de.NBI network have shown that many of the de.NBI members are involved in COVID-19 research. We could trace a total of 29 COVID-19 research projects, which are presented with short abstracts on the de.NBI website<sup>1</sup>. These research projects can be divided into several categories covering the entire COVID-19 research landscape. This extensive collection of de.NBI-related COVID-19 research projects was the incentive to produce this brochure entitled 'Data Analysis for the COVID-19 Research - Contributions of the German Network for Bioinformatics Infrastructure'. In total, 15 research projects represented by the de.NBI members were collected into the brochure, which could be divided into three categories.

**IRENA MAUS:** Professor Tauch, members of the de.NBI network are also running ELIXIR Germany, the national node of the European ELIXIR infrastructure. How did ELIXIR respond to the COVID-19 pandemic?

**ANDREAS TAUCH:** Since 2016 Germany is a full member of the intergovernmental organisation ELIXIR that brings together life science resources from across Europe. As a national node, ELIXIR Germany is involved in European research activities led by ELIXIR. We rapidly provided on the ELIXIR website a range of services that can be used for studying the SARS-CoV-2 coronavirus and the COVID-19 disease. ELIXIR also made available European computing services to support COVID-19 research projects. In this regard, Germany provided priority access in the de.NBI for projects relating to COVID-19.

**IRENA MAUS:** You are referring to the coordinating role of ELIXIR in Europe. How are data of the SARS-CoV-2

coronavirus and the COVID-19 disease collected in Europe?

**ANDREAS TAUCH:** It is a central goal of ELIXIR to develop life science resources across Europe in such a way that researchers can more easily find, analyse and share scientific data. For instance, all raw and consensus viral sequences should be deposited in the European Nucleotide Archive ENA. ENA is an ELIXIR

Core Data Resource for nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. In addition, the European Commission, the European Bioinformatics Institute in Hinxton and ELIXIR have launched in April 2020 the European COVID-19 Data Portal. This portal brings together relevant datasets for sharing and analysis as an effort to accelerate COVID-19 research.




**ELIXIR Germany and ELIXIR**

is an intergovernmental organisation that brings together life science resources from across Europe. ELIXIR Germany is the national Node of the European ELIXIR infrastructure since 2016.



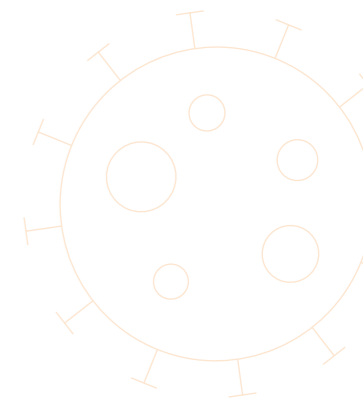
Irena Maus, de.NBI Administration Office

**IRENA MAUS:** How is ELIXIR Germany contributing to the European COVID-19 Data Portal?

**ANDREAS TAUCH:** The COVID-19 Data Portal brings together and continuously updates the relevant COVID-19 datasets and tools. ELIXIR Germany has rapidly established a task force to address the challenges arising from managing human omics data in the COVID-19 context within Germany. The task force is managed by Prof. Dr. Peer Bork from EMBL Heidelberg and brings together expertise from the German COVID-19 OMICS Initiative DeCOI in sequencing and NGS data generation. The Administration Office of ELIXIR Germany participates in regular administrative meetings of the National Coordinating Team of the COVID-19 Data Portal. We are therefore involved in scientific and strategic decisions that will shape the future development of the COVID-19 Data Portal.

**IRENA MAUS:** How was it possible to implement the European COVID-19 Data Portal so quickly?

**ANDREAS TAUCH:** We currently have 23 countries in ELIXIR, and we work together for several years with the ELIXIR Hub in Hinxton using a 'Hub and Node' model. The ELIXIR Hub is like a headquarter, and the national node is generally a network of organisations that works within a member state. The German ELIXIR node was built on the governance of the de.NBI network and it thus uses established decision-making processes. Likewise, the Heads of Nodes meet regularly on the European level to develop ELIXIR's scientific and technical strategy. The scientific and technical knowledge in ELIXIR is thus based on a European network of experts. The European response to the COVID-19 pandemic is therefore an excellent example of the value of established bioinformatics infrastructures that goes beyond the initial goals for the good of the society.



**Prof. Dr. Andreas Tauch**  
Head of Node of ELIXIR Germany

tauch@cebitec.uni-bielefeld.de

**Prof. Dr. Alfred Pühler**  
de.NBI Coordinator

puehler@cebitec.uni-bielefeld.de



# THE GERMAN NODE WITHIN ELIXIR EUROPE

ELIXIR, the European Life Science Infrastructure for Biological Information was founded in 2014 as an intergovernmental organization and brings together life science resources from across Europe. The consortium currently consists of 23 members. ELIXIR Germany is the German Node of ELIXIR since 2016. The node is run by members of the German Network for Bioinformatics Infrastructure (de.NBI). The infrastructure of ELIXIR Germany is represented by eight service units distributed across Germany and the associated EMBL Heidelberg. ELIXIR Germany is coordinated from Bielefeld University and funded by the German government. The National Node is led by the Head of Node Prof. Dr. Andreas Tauch.



- ELIXIR GERMANY ACTIVITIES**
- CONTRIBUTION TO THE CORE DATA RESOURCES**
    - BRENDA
    - SILVA
  - 20 RESEARCH PROJECTS WITH GERMAN CONTRIBUTION**
  - SUPPORT TO EUROPEAN COVID-19 RESEARCH**
    - TASK FORCE FOR COVID-19 HUMAN OMICS DATA MANAGEMENT
    - COLLABORATION WITH GERMAN HUMAN GENOME-PHENOME ARCHIVE (GHGA)
    - INTEGRATION WITH THE EUROPEAN COVID-19 DATA PORTAL

21

PARTICIPATING GERMAN INSTITUTES

65

TOOLS & SERVICES

79

TRAINING COURSES PER YEAR



“ELIXIR demonstrates the European spirit through sharing data and expertise while agreeing on best practices.”

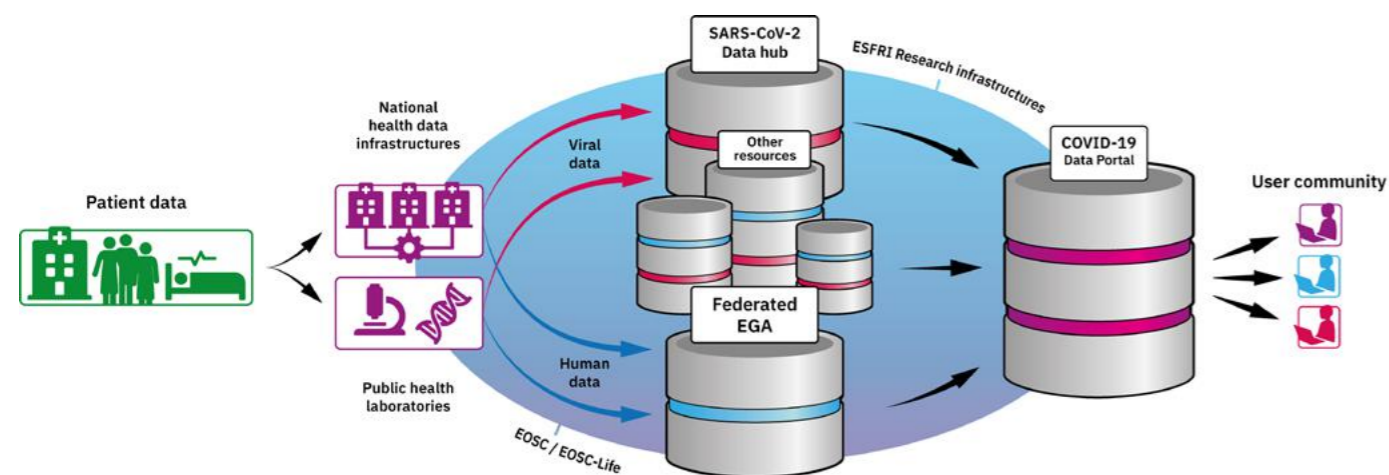
**Vera Ortseifen**  
Node Coordinator  
vera@cebitec.uni-bielefeld.de  
www.denbi.de/elixir-de

# ELIXIR SUPPORT to COVID-19 research

In addition to national activities, ELIXIR Germany supports European COVID-19 initiatives like the European COVID-19 Data Platform. The Portal facilitates open data sharing and analysis in order to accelerate coronavirus research by comprehensive presentation of resources relevant to COVID-19. The Data Platform provides researchers and public health experts not only with sequences resulting from research activities, but also with struc-

tural, expression, clinical and epidemiological data as well as an extensive collection of literature. In order to support the European COVID-19 Data Portal de.NBI/ELIXIR-DE established a task force for COVID-19 human omics data management and international dissemination. This task force is coordinated by Peer Bork (EMBL Heidelberg) and addresses the challenges arising from managing human omics data in the COVID-19 context within Germany:

The task force engages national communities, such as GHGA (led by the de.NBI members Oliver Stegle, Oliver Kohlbacher and Jan Korbel) or DeCOI (German COVID-19 OMICS Initiative), in order to FAIR-ify generated COVID-19 data and improve data exchange in the European context. Ultimately, this supports researchers and public health experts by making this essential research data accessible to the international community.



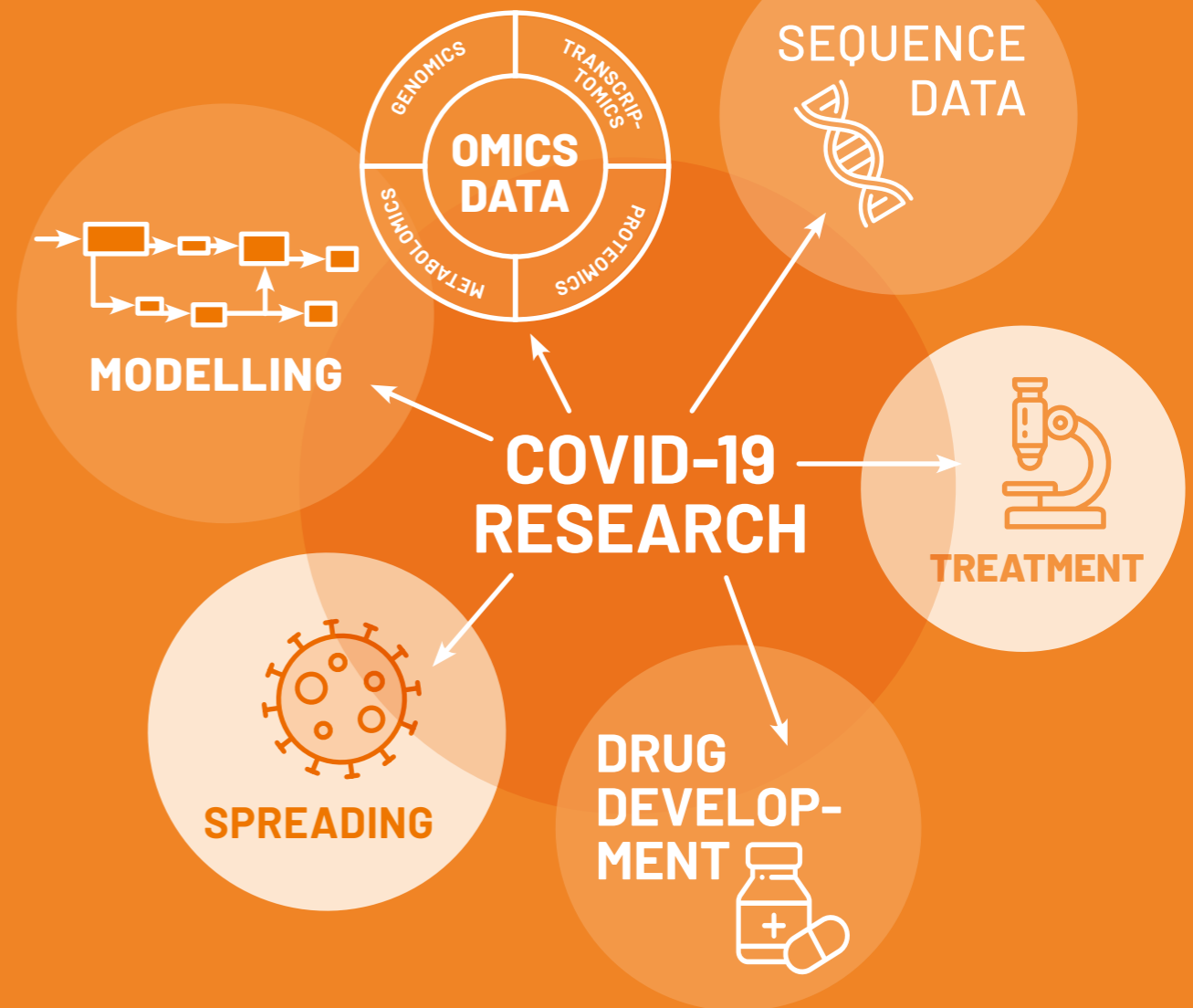
CREDIT: Spencer Phillips/EMBL-EBI



“Both for research and surveillance of SARS-CoV-2, large amounts of data are quickly needed. As a virus does neither accept regional nor national borders, it is essential to openly share data, coordinate activities in Germany and integrate them with international ones, such as the European COVID-19 Data Portal.”

**Peer Bork**  
Director of EMBL Heidelberg  
bork@embl.de  
[www.denbi.de/de-nbi-elixir-de-task-force-for-covid-19-human-omics-data](http://www.denbi.de/de-nbi-elixir-de-task-force-for-covid-19-human-omics-data)

# TO FIGHT PANDEMIC we need better data



For more information please contact  
[contact@denbi.de](mailto:contact@denbi.de) or visit these websites:

 **COVID-19 Data Portal**



[covid19dataportal.org](http://covid19dataportal.org)





[elixir-europe.org/services/covid-19](http://elixir-europe.org/services/covid-19)

# IMPRINT

Prof. Dr. Alfred Pühler  
German Network for Bioinformatics Infrastructure (de.NBI)  
de.NBI Administration Office  
Bielefeld University  
Center for Biotechnology (CeBiTec)  
Universitätsstraße 27  
33615 Bielefeld

Tel: +49 (0)521 106 8750  
Fax: +49 (0)521 106 89046  
E-Mail: [contact@denbi.de](mailto:contact@denbi.de)

Editors:  
Editor-in-Chief: Prof. Dr. Alfred Pühler (Bielefeld University, CeBiTec)  
Editorial Team: Prof. Dr. Andreas Tauch (Bielefeld University, CeBiTec),  
Dr. Tanja Dammann-Kalionski (Bielefeld University, CeBiTec),  
Dr. Doris Jording (Bielefeld University, CeBiTec),  
Dr. Irena Maus (Bielefeld University, CeBiTec).

[www.denbi.de](http://www.denbi.de)  
 @denbiOffice  
 [linkedin.com/company/de-nbi](https://www.linkedin.com/company/de-nbi)

Date: January 2021

Cover:  
Korbinian Stöckle, Rejwan Rasit Toplu: Coronavirus. Glass, blown and  
freely formed. Glashütte Gernheim 2020. Photo: Peter Hübbe

Photo credits:  
Andreas Kühlken  
de.NBI Administration Office  
iStockphoto  
AdobeStock  
unsplash

Design and Layout:  
MEDIUM Werbeagentur GmbH, Bielefeld

Printing:  
Bruns Druckwelt GmbH & Co. KG, Minden

ISBN 978-3-943363-07-4  
DOI: <https://doi.org/10.4119/unibi/2950259>

SPONSORED BY



Fkz 031A532B  
(de.NBI Administration Office)

