

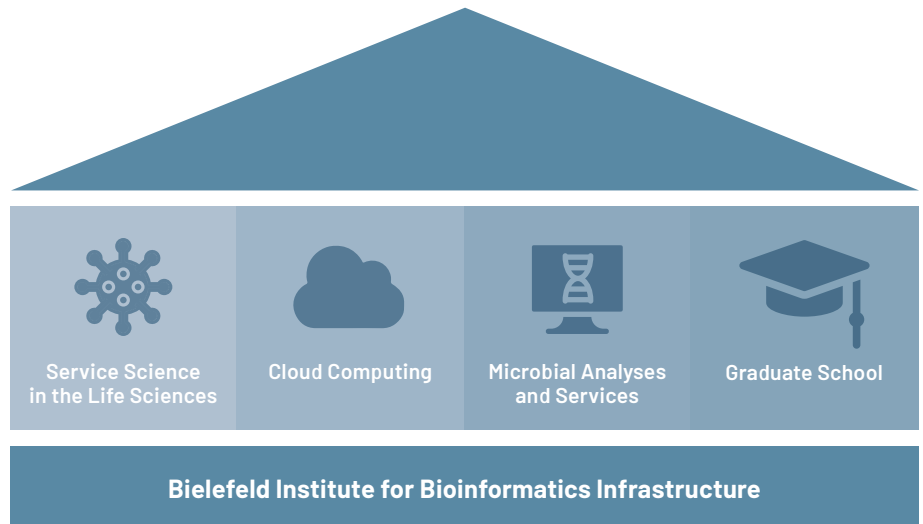


# The Bielefeld Institute for Bioinformatics Infrastructure

# EDITORIAL

*Prof. Dr. Jens Stoye, Head of Institute,  
BIBI, Bielefeld University*





## Dear Reader,

After two years of operation, we would like to use this opportunity to give an overview of the activities that were concentrated and extended at Bielefeld University in the area of bioinformatics services and training within the Bielefeld Institute for Bioinformatics Infrastructure (BIBI).

The institute was founded in summer 2019 and slightly restructured in spring 2021. It pools the bioinformatics service and training activities at Bielefeld University. The goal is to create, together with ZB MED – Information Center for Life Sciences in Köln and Bonn, a powerful institute offering research infrastructure for all areas in the digital life sciences. At the same time original research is carried out – mostly in cooperation with other national and international partners and often with the goal to evaluate and improve existing bioinformatics services. More background

about the history and mission of the institute can be found in the notes by Prof. Dr. Alfred Pühler and Prof. Dr. Dietrich Rebolz-Schuhmann in this booklet.

The institute comprises three scientific units, of which two are currently in operation. Yet to be set into function is the unit for Service Science in the Life Sciences, when the professorship with the same denomination will be filled. The other two scientific units, Cloud Computing headed by Prof. Dr. Alexander Sczyrba and Microbial Analyses and Services headed by Prof. Dr. Alexander Schönhuth, are presented in this brochure by two dedicated sections.

The fourth unit of BIBI is the graduate school Digital Infrastructure for the Life Sciences (DILS), coordinated by Dr. Roland Wittler. It contributes both to the educational aspect, and also to the

research profile of BIBI. By training data scientists with a focus on the development of new bioinformatics methods, the graduate school supports the institute in establishing the young research profile "Service Science in the Life Sciences" at an international level. Selected PhD projects currently ongoing in the DILS graduate school are presented in the third section of this booklet.

The authors of this brochure now wish all readers an interesting reading.

November 2021

Prof. Dr. Jens Stoye



# CONTENT

Editorial	03
Content	04
The process of establishing BIBI	06
ZB MED/BIBI: A strategic alliance	08
<b><i>A FUTURE IN THE CLOUD: high-performance computing for life sciences</i></b>	<b>10</b>
Cloud computing for the life sciences	12
BiBiGrid: Scaling from single virtual machines to high-performance clusters within minutes	16
EU Simba project: Analyzing large scale metagenomics data on the de.NBI Cloud	18
<b><i>UNLOCKING THE GENETIC SECRETS OF MICROORGANISMS: accessing and analysing microbial genome data</i></b>	<b>22</b>
The genetic diversity of viruses on a graphical map	24
Supporting local health authorities in fighting the SARS-CoV-2 pandemic	26

## 10 A FUTURE IN THE CLOUD



# 22

## UNLOCKING THE GENETIC SECRETS OF MICROORGANISMS

# 34

## GRADUATE SCHOOL

**Omics Fusion** 30  
– a web application to analyze and integrate microbial data from multiple omics sources

**GRADUATE SCHOOL** 34  
*“Digital Infrastructure for the Life Sciences” (DILS)*

Large scale detection of regulatory small RNAs in pathogenic bacteria 36

Errors in sequencing data 38  
Quality assessment and bioinformatics solutions

From hidden data and information towards data-driven research 42

**Biogas-GeneMining** 46  
Characterization of the genetic potential of biogas microbiomes by meta-analysis of metagenome datasets

Comparing pangenomes 50

Functional genomics of and bioinformatic analysis tools for seed quality parameters in rapeseed 52

Algorithms for graph-based computational pangenomics 56

Protein-DNA binding specificity is facilitated by DNA shape 62

Computational pangenomics in plants 66

Exploring comparative metagenomic and metatranscriptomic datasets 70

Tool development for comparative gene regulatory network analysis 74

Imprint 78

# The process of establishing BIBI

Alfred Pühler; *Bielefeld University, Bielefeld*



The impulse to found the BIBI institute came from the German Network for Bioinformatics Infrastructure (de.NBI). The de.NBI network is a large-scale BMBF project that was initiated in 2015 and has since been successfully involved in developing a bioinformatics infrastructure. Since BMBF projects are generally time-limited, a sustainable continuation of the de.NBI network was considered from the beginning. A first approach was to make the network permanent in the frame of the Leibniz Association. To this end, plans were developed to combine the Bielefeld parts of the de.NBI network in a "Bielefeld Institute for Bioinformatics Infrastructure" (BIBI) and to follow the plan to integrate the BIBI institute, togeth-

er with the Information Centre for Life Sciences (ZB MED) in Cologne, into the Leibniz Association. This approach was taken from 2018 onwards. First, a cooperation agreement was signed between Bielefeld University and ZB MED. Subsequently, the BIBI institute was founded, whose administrative and user regulations (VBO) came into force in 2019. The BIBI institute is headed by Jens Stoye and consisted of a total of six areas when it was founded. Two of these areas were dedicated to the administration and coordination offices of the de.NBI network and the German ELIXIR node. In addition, areas for cloud computing, microbial bioinformatics and graduate training were also established. The sixth area was planned as

an area for service science in the life sciences for the head of the BIBI institute to be newly appointed. With the establishment of the BIBI institute, all the prerequisites have been provided to strive for the sustainable continuation of the de.NBI network together with ZB MED via the Leibniz track. At the end of 2020, however, this plan was fundamentally changed. At the political level, it was decided that the continuation of the de.NBI network should take place within the framework of the Helmholtz Association. The Forschungszentrum Jülich (FZJ) was given the task and also the financial support to take the de.NBI continuation into its hands. This development obviously had an impact on the BIBI institute. The de.NBI components

**3 May 2021**  
Publication of the  
**modified VBO**  
of the BIBI institute

2020

2021

**13 Aug 2019**

Selection meeting for filling  
**graduate positions** at  
Bielefeld University

of the BIBI institute were removed. However, the modified BIBI Institute, with its areas of service science, cloud computing, microbial bioinformatics and graduate training, was still highly topical and so future-oriented that the plan to join the Leibniz Association together with ZB MED could be continued. Incidentally, the structure of the modified BIBI institute was laid down in an amended VBO, which already came into force on 3 May 2021.

*Prof. Dr. Alfred Pühler*  
Coordinator of the de.NBI network



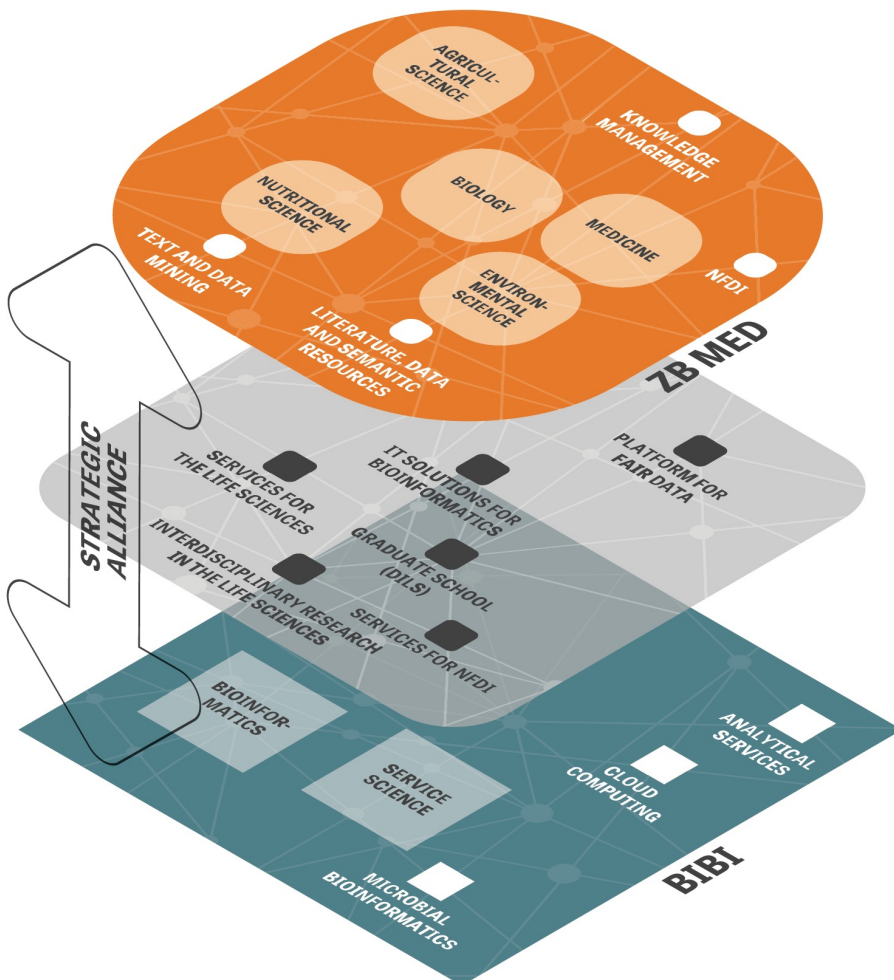


# ZB MED/BIBI:

## A strategic alliance

Dietrich Rebold-Schuhmann; ZB MED, Cologne

**Figure 1:** The alliance of ZB MED (orange level) and BIBI (blue level) introduces numerous new services and extends existing services (grey level) by linking content, data science, cloud computing, and extending the target groups. In this way, ZB MED and BIBI jointly campaign for open science in the life sciences.



The Information Centre for Life Sciences (ZB MED) in Cologne follows a long tradition of licensing and delivering scientific journal articles and scholarly books to researchers in Medicine, agricultural sciences (subsidiary located at the University of Bonn) and life sciences in general. The transformation of the scientific journals towards digital and open access publications and in the same way in recent years for scholarly books have changed the community involvement of ZB MED and is leading to a portfolio of services that support digital delivery of content. These developments are well aligned with the collection and delivery of digital data from the scientific community to the scientific community, which requires an IT Infrastructure and forms the next generation of information delivery.

The Bielefeld Institute for Bioinformatics Infrastructure (BIBI) and ZB MED in Cologne form a strategic alliance, since both institutes deliver IT services into the scientific community in the life sciences and provide complementary solutions that complete each other's portfolio of solutions. Whereas ZB MED is focused on the delivery of content, BIBI advances cloud-based analytical solutions for the life sciences that

## Key Elements of the Strategic Alliance

- October 2018: Cooperation agreement of ZB MED and Bielefeld University on joint development of information services for the life sciences
- August 2019: Foundation of the graduate school "Digital Infrastructure for the Life Sciences" (DILS) at BIBI in cooperation with ZB MED
- Since 2019: Hosting joint scientific workshops
- 2020: Developing a joint strategy: "ZB MED/BIBI 2020-2025: Supporting humans and environment by research and infrastructure"
- Since 2020: Joint supervision of the first doctoral students at the graduate school DILS
- 2021: First DILS retreat with PhD students and faculty members from both institutions

can make efficient use of the content available at ZB MED. Both make use of large-scale IT infrastructure, however ZB MED enriches its content with semantics technologies whereas BIBI analyses large sets of OMICS data, e.g., for biomedical research.

Researchers from BIBI and ZB MED are well established in the life science research community, in bioinformatics as well as medical informatics and agricultural research.

Scientifically they can cover the full range of computer science, bioinformatics, medical informatics, semantics technologies, e.g., generation and use of terminologies and ontologies, and machine learning (including deep learning and AI). Together they provide solu-

tions that cover the complete research life cycle of life science research: a unique combination of literature and information supply including computational analysis of big data. This enables new scientific insights for researchers of all life science disciplines as well as bioinformatics.

BIBI introduces modern infrastructures and additional bioinformatics expertise into the broad repertoire of ZB MED. The strategic alliance of ZB MED and BIBI implements numerous

new services: bioinformatics compute solutions, cloud computing, data science training, and the graduate school "Digital Infrastructure for the Life Sciences" (DILS).

*Prof. Dr. Dietrich Rebholz-Schuhmann,  
Scientific director of ZB MED  
Source: ZB MED*

## Powerful Alliance for the Life Sciences

ZB MED supports the scientific communities with data, information and literature, and BIBI offers cloud-based compute services. Both complement one another in their mission.



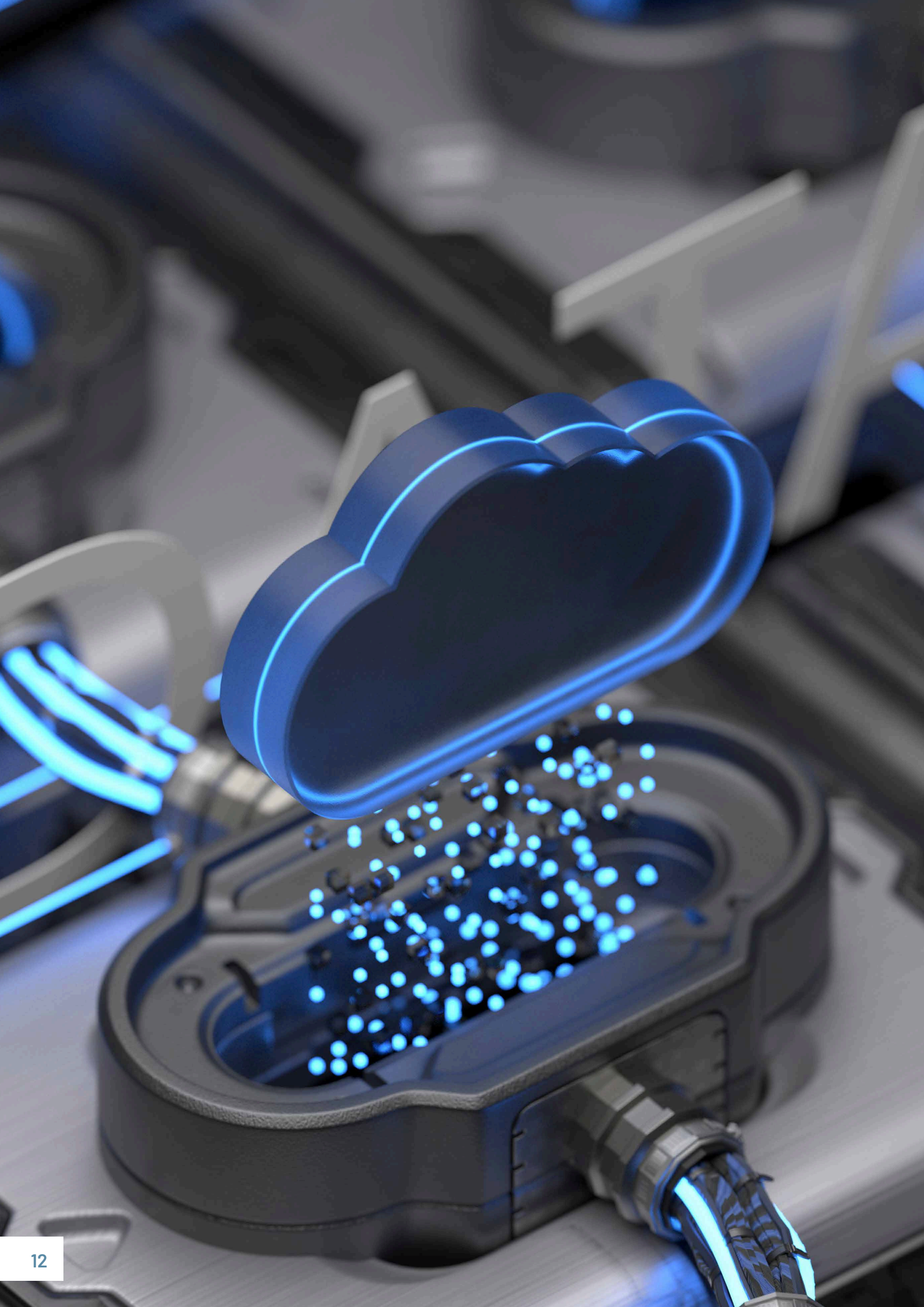
# A FUTURE IN THE CLOUD:

*high-performance  
computing for life  
sciences*

*The life sciences sector places an incredible reliance on data and this all needs to be processed and stored somewhere. With the cloud, this data becomes more accessible and offers a sea of information.*









# Cloud computing for the life sciences

**Alexander Sczyrba; *Bielefeld University, Bielefeld***

The increasingly widespread availability and application of high-throughput technologies in the life sciences, such as (meta-)genomics studies or imaging applications, generate an exponentially increasing amount of experimental data. The number of specialized databases distributed around the world is also growing rapidly. Therefore, the storage, integration and processing of this data becomes the bottleneck of the analysis workflows, as they require infrastructures for data storage as well as services for data processing, analysis and possibly special access approval.

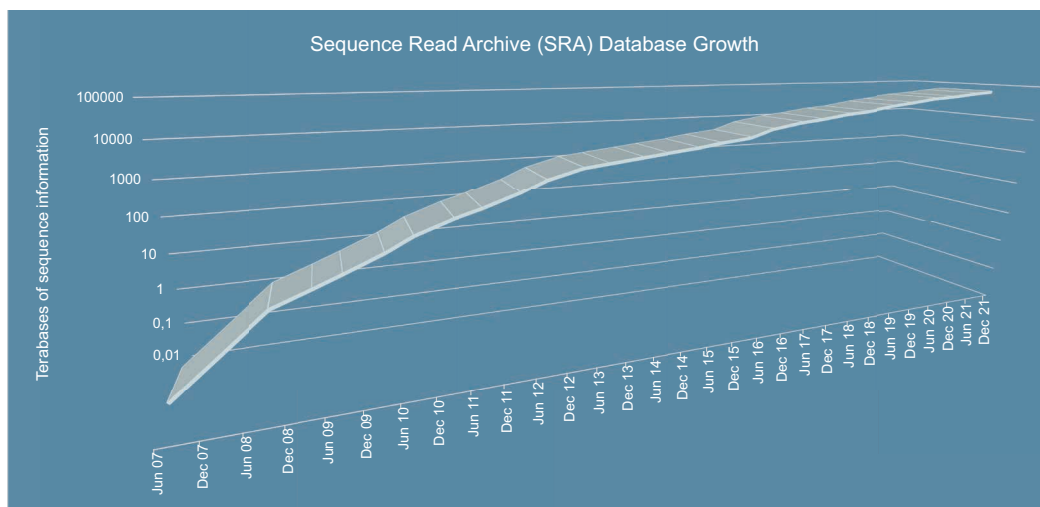
According to the definition of the National Institute of Standards and Technology (NIST), "Cloud computing is a model for enabling ubiquitous, con-

venient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". Cloud computing plays an important role in many modern bioinformatics analysis workflows, from data management and processing to data integration and analysis, including data exploration and visualization. It provides massively scalable computing and storage infrastructures and can therefore represent the key technology for overcoming the aforementioned problems.

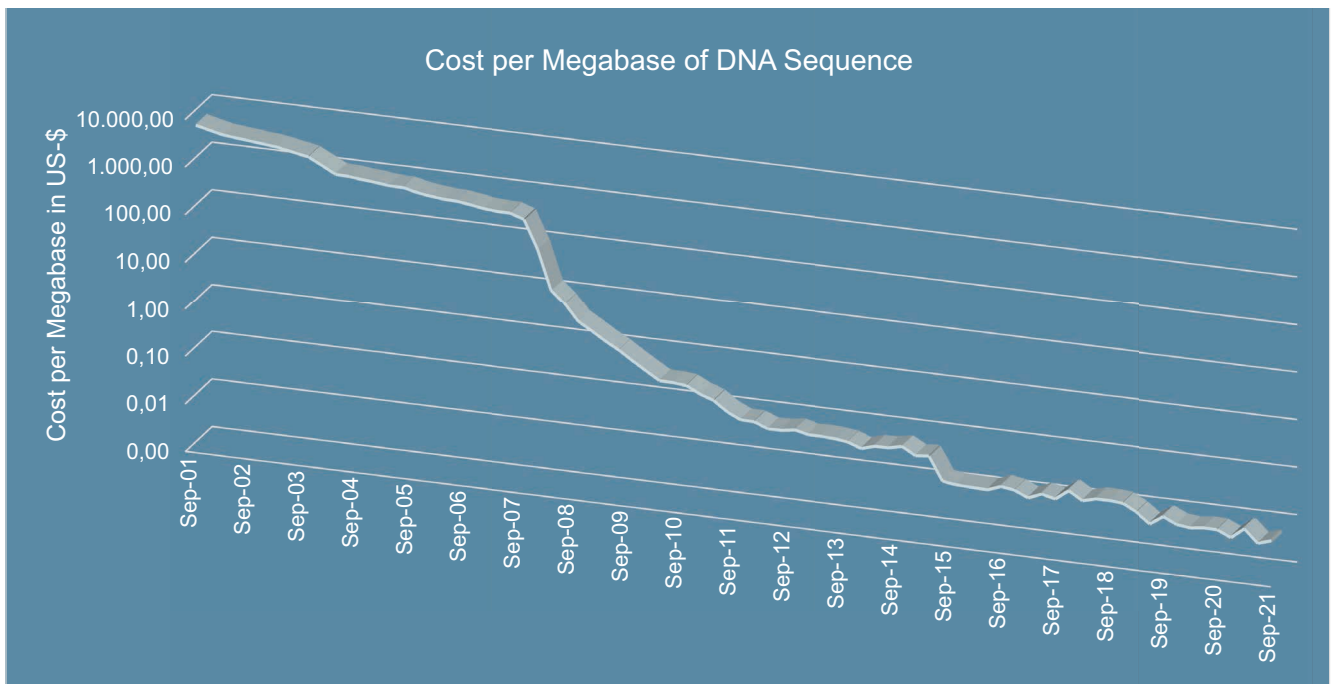


## Cloud computing (bioinformatics) services are often divided into the following areas:

- Data as a Service (DaaS):**  
 provides data storage in a dynamic virtual environment hosted in the cloud, providing data that can be accessed from a variety of connected devices on the Internet. One such example is the National Center for Biotechnology Information (NCBI), which provides the Sequence Read Archive (SRA) data on the Google Cloud Platform (GCP) and Amazon Web Services (AWS) clouds. All publicly-available, unassembled read data and authorized-access human data are available for access and compute through these cloud providers.
- Software as a Service (SaaS):**  
 offers cloud-based tools for performing various bioinformatics tasks, e.g. sequence processing, gene expression analysis, or image analysis.
- Platform as a Service (PaaS):**  
 In contrast to SaaS solutions, PaaS solutions enable users to provide bioinformatics applications and maintain complete control over their instances and the associated data.
- Infrastructure as a Service (IaaS):**  
 This service model is offered in a compute infrastructure that includes servers (usually virtualized) with specific computing capacities and/or storage. The user controls all provided storage resources, operating systems and bioinformatics applications. The German Network for Bioinformatics Infrastructure (de.NBI) Cloud provides such a service free of charge for life scientists in Germany.



**Figure 1:** Growth of Sequence Read Archive (SRA) database hosted at the National Center for Biotechnology Information (NCBI), USA. The data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. [1]



**Figure 2:** Development of costs for sequencing one megabase of genomic information over the last 20 years [2].

Virtual environments such as virtual machines (VMs), Docker or Singularity provide maximal flexibility to the users. In contrast to classical high performance environments they are independent from the installed operating system, software stacks libraries. Special requirements can be fulfilled easily without side effects. Additionally, virtual environments allow easy exchange of analysis workflows and with publication of these environments research becomes reproducible.

The cloud computing department of BIBI develops and provides bioinformatics environments and workflows for bioinformatics analyses, mainly in the field of (meta-)genomics. A mirror of SRA's metagenomics data sets hosted at the de.NBI Cloud site in Bielefeld allows large scale analyses integrating publicly available data. Examples of such projects are described in the following sections.



**References:**

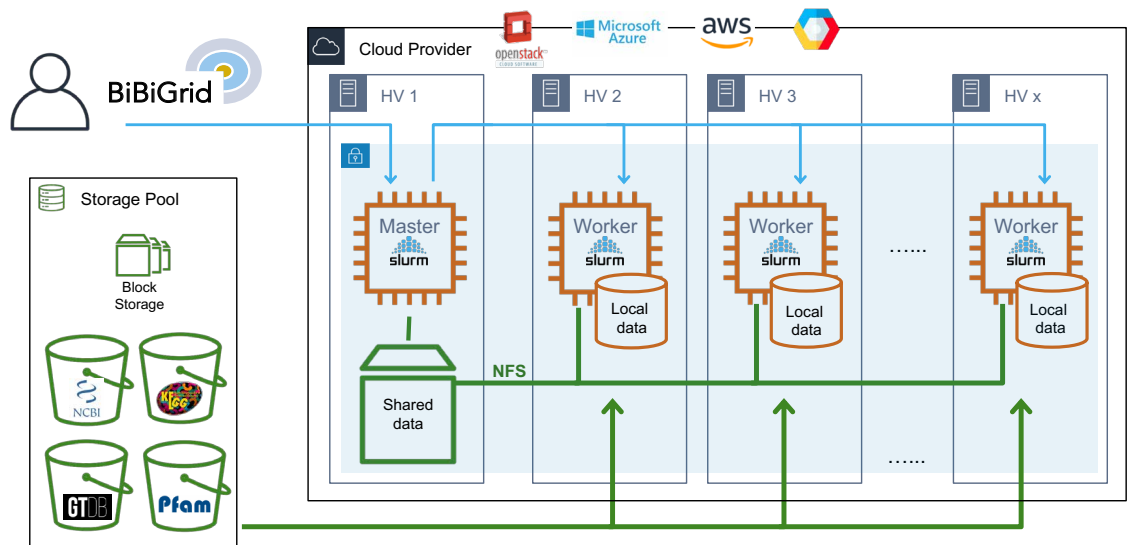
- [1] National Center for Biotechnology Information (NCBI) Website: <https://www.ncbi.nlm.nih.gov/sra/>  
 [2] The Cost of Sequencing a Human Genome. <http://genome.gov/sequencingcosts>

# BiBiGrid:

*Scaling from single virtual machines to high-performance clusters within minutes*

*Jan Krüger; Bielefeld University, Bielefeld*





**Figure 1:** BiBiGrid controls cloud computing environments by launching and configuring virtual machines (VMs) and storage volumes. Software stacks are deployed via Ansible and queuing systems (SLURM) are set up to distribute workloads across the VMs.

Infrastructure-as-a-service (IaaS) is a model of cloud computing in which a virtualized IT infrastructure is made available to users via the Internet. Together with Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS), IaaS is one of three general cloud service models. Within the IaaS model users manage the operating system, middleware, applications and data to take advantage of compute and storage resources. The IaaS provider is responsible for providing virtualization, storage, network, and servers. As a result, users do not need a local data center, avoiding the administrative overhead including maintenance and updating of hardware and software components. The user controls the infrastructure via an application programming interface (API) or a graphical user interface (dashboard). IaaS enables easy scaling and updating as well as the addition of resources as required.

Making use of available IaaS resources turns out to be a challenge for many users not familiar with cloud environments. While launching tens or even hundreds of virtual machines (VMs) is easy using the API or dashboard, a whole software application stack needs to be deployed on these VMs to fully utilize the resources. Many bioinformatics workflows use classical high-performance computing (HPC) environments with scheduling systems to distribute their compute jobs on HPC clusters. These kinds of environments need to be set up in the cloud to easily move the existing workflows to cloud environments.

BiBiGrid [1] is an open source tool for an easy cluster setup inside a cloud environment. BiBiGrid is independent of the operating system and cloud provider. Currently it supports backend implementations for Amazon (AWS), Google (Google Compute), Microsoft (Azure) and OpenStack.

Starting a cluster requires a valid configuration file and credentials of the cloud provider.

The configuration file specifies the composition of the requested cluster. During resource instantiation BiBiGrid configures the network, local and network volumes, (network) file systems and deploys the software for immediate use of the started cluster. When using pre-installed images a fully configured and ready to use cluster is available within a few minutes.

BiBiGrid uses Ansible to configure standard Ubuntu as well as Debian cloud images. Depending on your configuration BiBiGrid can set up an HPC cluster for grid computing (Slurm Workload Manager), a shared file system (NFS on local discs and attached volumes), a cloud IDE for writing, running and debugging (Theia Web IDE) code, and a monitoring system (Zabbix). Custom Ansible scripts can be used to further customize the cluster after first initialization.



#### References:

[1] <https://github.com/BiBiServ/bibigrid>

# EU SIMBA project:

## *Analyzing large scale metagenomics data on the de.NBI Cloud*

Liren Huang; *Bielefeld University, Bielefeld*

SIMBA (Sustainable Innovation of Microbiome Applications in the Food System) is a European innovation project, funded under the EU's Horizon 2020 Funding Programme, which provides a holistic and innovative approach to the development of microbial solutions to increase food and nutrition security. SIMBA focuses in particular on the identification of viable land and aquatic microbiomes that can assist in the sustainability of European agro- and aquaculture. Under the scope of the EU SIMBA project, our research group focuses on exploring microbial communities in large scale publicly available environmental sequencing (metagenomics) data and association studies with plant growth promoting bacteria (PGPB). To that end, our study addresses both computational intensive challenges of searching hundreds of terabytes of public data and sophisticated data mining (e.g. network analysis) on putative PGPB genomes.

We established a scalable bioinformatics workflow for detecting PGPB associated microbes from public data. In particular, we have developed a distributable framework, Sparkhit, that enables screening and mapping [1] terabytes of sequencing data within hours. After preliminary screening of large datasets, EMGB is used as a general purpose bioinformatics workflow for analyzing metagenomics data and visualizing annotation results. We also developed a de-replication tool that can handle large amounts of metagenomics samples and facilitate downstream co-occurrence network analysis. Most tools are containerized (e.g. Docker) and are easily accessible on the cloud.

In the case of the EU SIMBA project, terabytes of public soil metagenome datasets were collected and downloaded on the de.NBI cloud object storage. Associated metadata containing detailed description of the datasets was



categorised. In the first step of our analysis pipeline, Sparkhit is used to map all input sequencing data to the selected PGPB genomes. Sparkhit is an in-house fragment recruitment tool that can be scaled to hundreds of computer nodes. Once high similarity hits are found, corresponding samples are

selected for assemblies or co-assemblies (multiple samples in one bio-project) using the EMGB pipeline. The EMGB pipeline also generates “metagenome assembled genomes” (MAGs) after assembly, representing the microbes present in the samples.

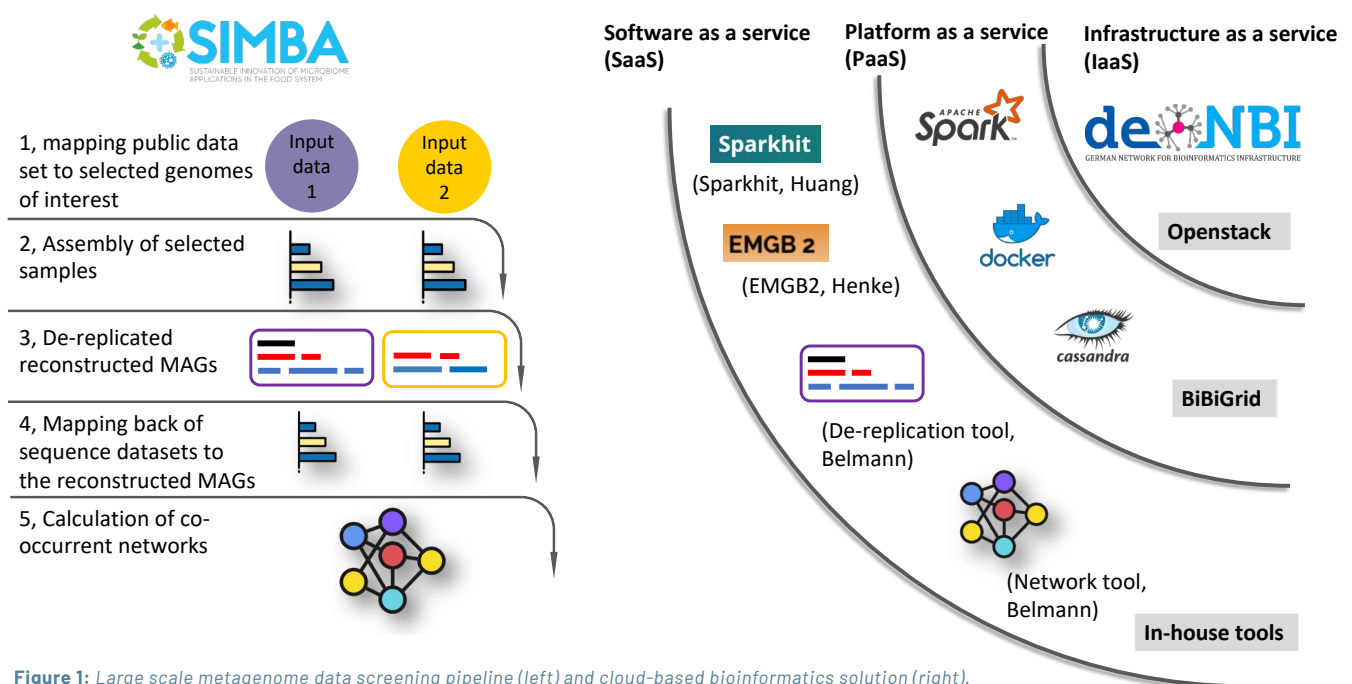


Figure 1: Large scale metagenome data screening pipeline (left) and cloud-based bioinformatics solution (right).



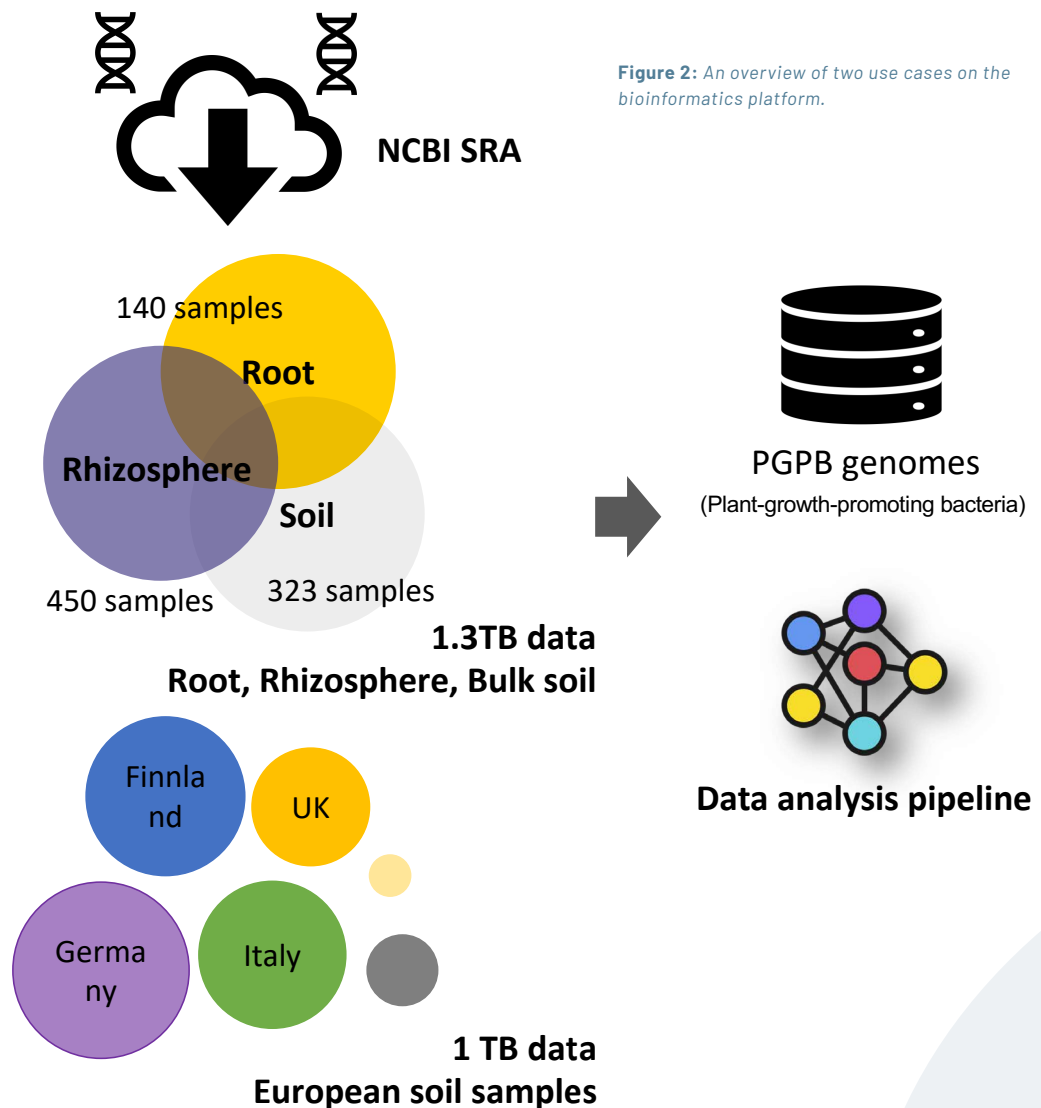


Figure 2: An overview of two use cases on the bioinformatics platform.

To remove redundancy between different samples, the generated MAGs are de-replicated and the representative MAGs are selected for further analysis. To refine our analytical pipeline, we have combined a set of existing tools for the de-replication of MAGs. By comparing and evaluating these tools, we were able to identify the best approach to de-replicate our reconstructed MAGs, and accordingly established a personalized de-replication pipeline.

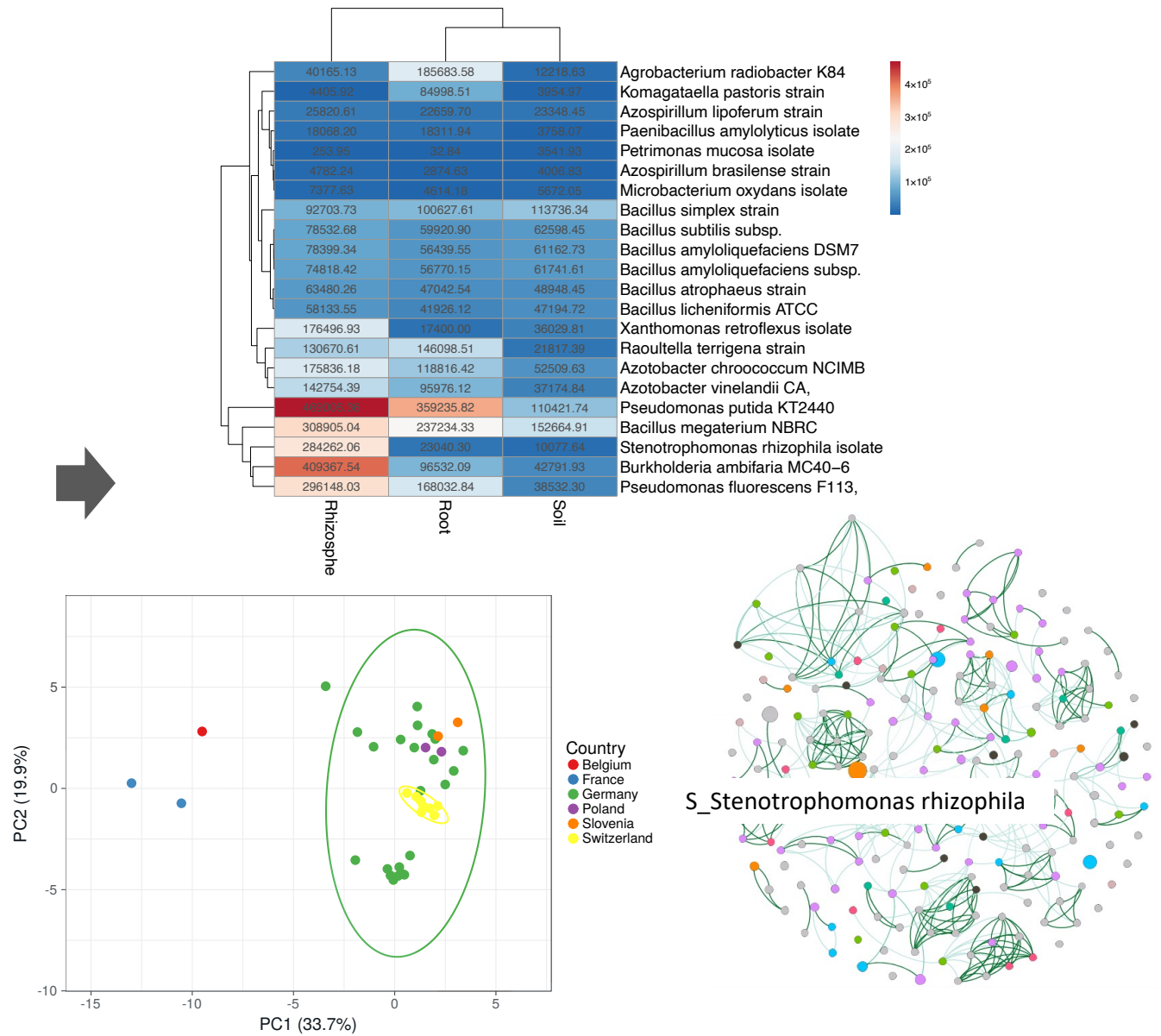
Our de-replication pipeline starts by filtering MAGs with high contaminations and low coverages. After the filtering step, Average Nucleotide Identity (ANI) methods are applied to determine spe-

cies and strain level clusters. Once clusters are formed, representative MAGs are selected based on the ranking algorithm to represent each cluster. Since the core task of MAG dereplication workflows is the estimation of similarity between genomes, which can be done by calculating ANI, we collected and evaluated several ANI-based approaches. Three different datasets (unfiltered, medium, and high MIMAG) from CAMI challenge [2] are used for species and strain level dereplication evaluation.

Once representative MAGs are selected, the pipeline re-maps the sequencing data back to the MAGs and produces MAG-abundance profiles for all sam-

ples. The abundance profiles are used to compare the PGPB diversity between different samples. It can also be used to build co-occurrence networks involving assembled MAGs and known PGPBs.

The intermediate results of the EMGB pipeline are imported into the EMGB browser. In the browser, each individual sample can be selected and its computed results can be explored in a click-button style. Users can also compare different samples by selecting multiple samples in the browser tab. Selected metrics, such as the abundance tables of de-replicated MAGs from all metagenome samples, are also accessible through the web interface.



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 818431 (SIMBA). This output reflects only the author’s view and the Research Executive Agency (REA) cannot be held responsible for any use that may be made of the information contained therein.



**References:**

- [1] Huang et al. Analyzing large scale genomic data on the cloud with Sparkhiti. *Bioinformatics*, 2018, DOI: 10.1093/bioinformatics/btx808.
- [2] Sczyrba et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, 2017, DOI: 10.1038/nmeth.4458.

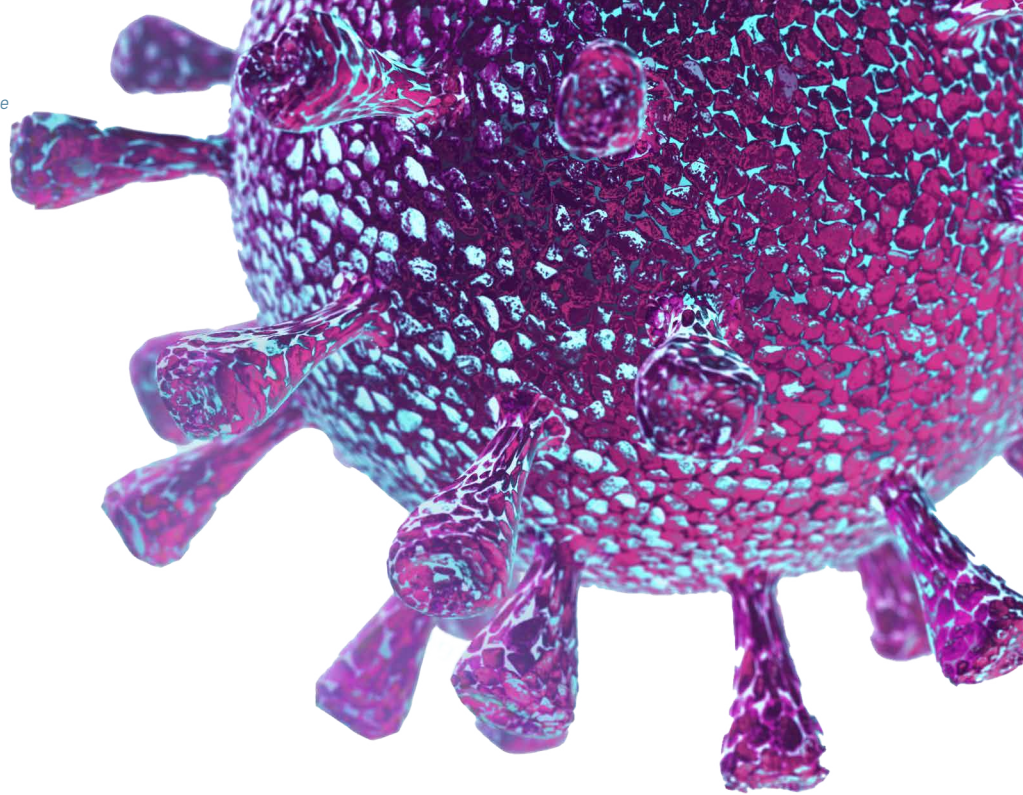
# UNLOCKING THE GENETIC SECRETS OF MICROORGANISMS:

*accessing and analysing  
microbial genome data*

*In the last decade bioinformatics has silently filled in the role of cost effective and target-oriented data analysis. It has enhanced our understandings about the microorganisms' genome structure and the cellular processes in order to treat and control microbial cells as factories.*







# The genetic diversity of viruses on a graphical map

**Alexander Schönhuth; Bielefeld University, Bielefeld**

Various life-threatening viruses mutate insanely fast, thereby protecting the virus from human immune response or medical treatment. Naturally, virus variants form within the infected hosts, when hijacking the host's replication machinery. Therefore, accurate tracking of strains within individual patients or local samples, for example obtained from wastewater, can make a crucial contribution to assessing the evolutionary course of epidemics [1].

We focus on developing methodology for identifying the development of new strains/variants, and to put them into context with existing strains/variants. To do that, all strains and variants are arranged in a "map-like" graphical data structure. This "map of variants" highlights the origin of new variants conveniently, and puts them into context with existing variants.

Further, this enables us to accurately place new infections on this map of variants. As a consequence, new infections can be classified at high resolution. Novel, emerging variants can be spotted quickly and integrated into the map in an evolutionarily consistent manner.

Key to success is to make use of “pangenome graphs” as a relatively new concept to arrange individual, mutually related genomes in an evolutionarily sensible way. Pangenome graphs have recently been emerging, and gradually replacing traditional ways of working with genomes.

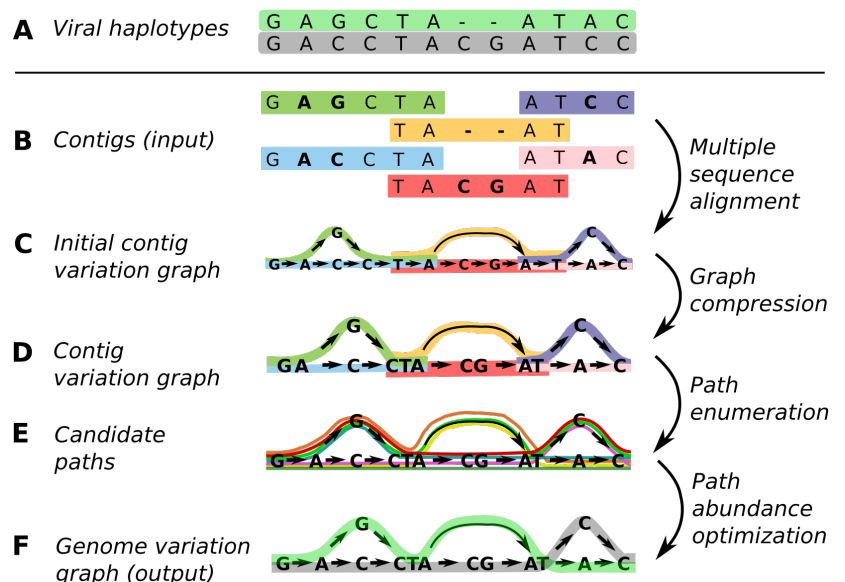
Their advantages are their compactness – which can save petabytes of storage space – their intuitive representation, and their consistency in terms of the evolutionary relationships among the individual genomes.

Recently, we have pointed out ways to make stringent use of pangenome graphs for tracking and analyzing viruses. Therefore, it was important to realize that not only one, but possibly several strains can populate individual hosts. In fact, this is rather common because new strains and variants form within hosts, when the virus hijacks the host’s replication machinery; note that virus particles cannot mutate while circulating between hosts.

The crucial first step is to adapt analysis tools accordingly, and to make it possible to construct pangenome graphs that reflect the within-host diversity of a virus: not considering within-host diversity collapses different mutations, which falsifies one’s view on the evolutionary development of the virus. The challenge however is that considering within-host diversity requires approaches that are essentially novel [2,3].

Once this foundation has been laid, accurate pangenome graphs can be constructed, as we could demonstrate in a corresponding series of papers [4,5]. See Figure 1 for an illustration of the algorithmic steps to be taken towards successful construction of viral pangenome graphs.

**Figure 1:** Pangenome graph construction. (A) Original viral haplotypes, reflecting strain specific genomes. While originally unknown, they are the source of fragments (aka contigs) shown in (B). The task is to reconstruct sequences shown in (A) from fragments shown in (B). This reconstruction proceeds in four further steps: computing a multiple sequence alignment leads to a first graph as shown in (C), compressing the graph by joining letters leads to (D). Eventually, inspecting all possible paths (“candidate paths” in (E)) and evaluating their plausibility relative to the original fragment data leads to the final situation, shown in (F). The final graph enables us to correctly identify the haplotypes that were responsible for generating the data.



#### References:

- [1] S. Posada-Céspedes et al. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus research*, 2017. DOI: 10.1016/j.virusres.2016.09.016.
- [2] J. Baaijens et al. De novo assembly of viral quasispecies using overlap graphs. *Genome Research*, 2017. DOI: 10.1101/gr.215038.116.
- [3] X. Luo et al. Strainline: full-length de novo viral haplotype reconstruction from noisy long reads. In revision at *Genome Biology*. DOI: 10.1101/2021.07.02.450893.
- [4] J. Baaijens et al. Full-length de novo viral quasispecies assembly through variation graph construction. *Bioinformatics*, 2019. DOI: 10.1093/bioinformatics/btz443.
- [5] J. Baaijens et al. Strain-aware assembly of genomes from mixed samples. *RECOMB*, 2020. DOI: 10.1101/645721.





# Supporting local health authorities

*in fighting the  
SARS-CoV-2 pandemic*

Michael Beckstette; *Bielefeld University, Bielefeld*

The global SARS-CoV-2 pandemic poses numerous new challenges and actions on different kinds of administrative and institutional levels ranging from world-wide vaccination initiatives over state-specific changes of laws to new regulations for local authorities. Most notably with the appearance of new emerging virus variants like the more infectious Alpha (B.1.1.7) and Delta (B.1.617.2) variants, local health authorities in Germany were confronted with new tasks such as sequencing of virus genomes and virus lineage detection. The importance of these tasks emerged not only from the necessity to control local outbreaks but also from the obligation to provide data to federal government agencies like the Robert Koch Institute (RKI) for monitoring of the pandemic situation and population risk assessment which is finally used to furnish recommendations to health professionals and governmental decision-makers.

In early 2021, when the Alpha variant of the SARS-CoV-2 virus started to spread over Germany (Figure 1), the Chemical and Veterinary Investigation Office for the region Ostwestfalen-Lippe (CVUA-OWL) contacted researchers of the

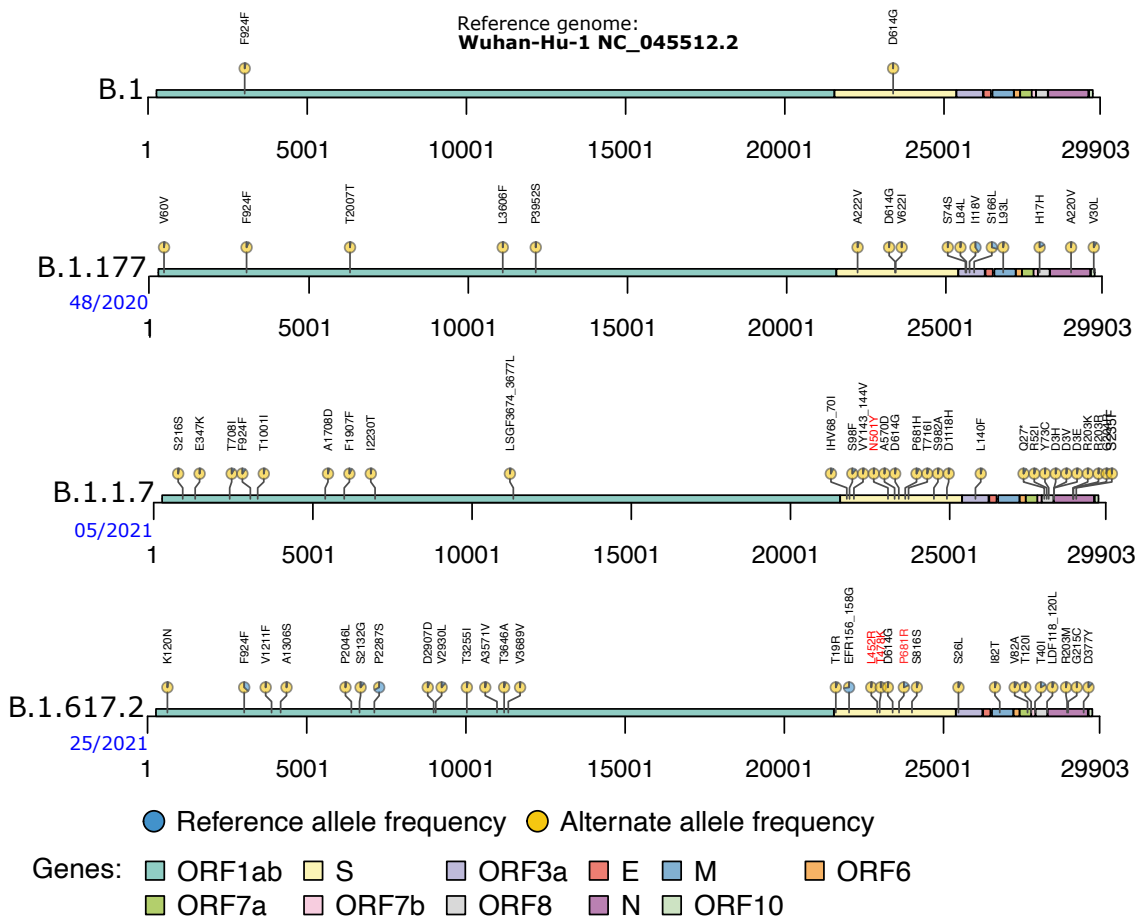
Bielefeld Institute for Bioinformatics Infrastructure (BIBI) and asked for support in the bioinformatics analysis of sequenced virus samples from patients from the OWL region. BIBI's substantiated existing expertise in this field [1, 2] and a pragmatic and efficient collaboration style between researchers from both institutions allowed to

***“The researchers from BIBI were of great help in the bioinformatics analysis of our samples. They provided standardized and easy to use workflows, which have successfully been used to analyze and identify the variants of more than 260 SARS-CoV-2 positive samples.”***

*Dr. Henning Petersen from the CVUA-OWL*

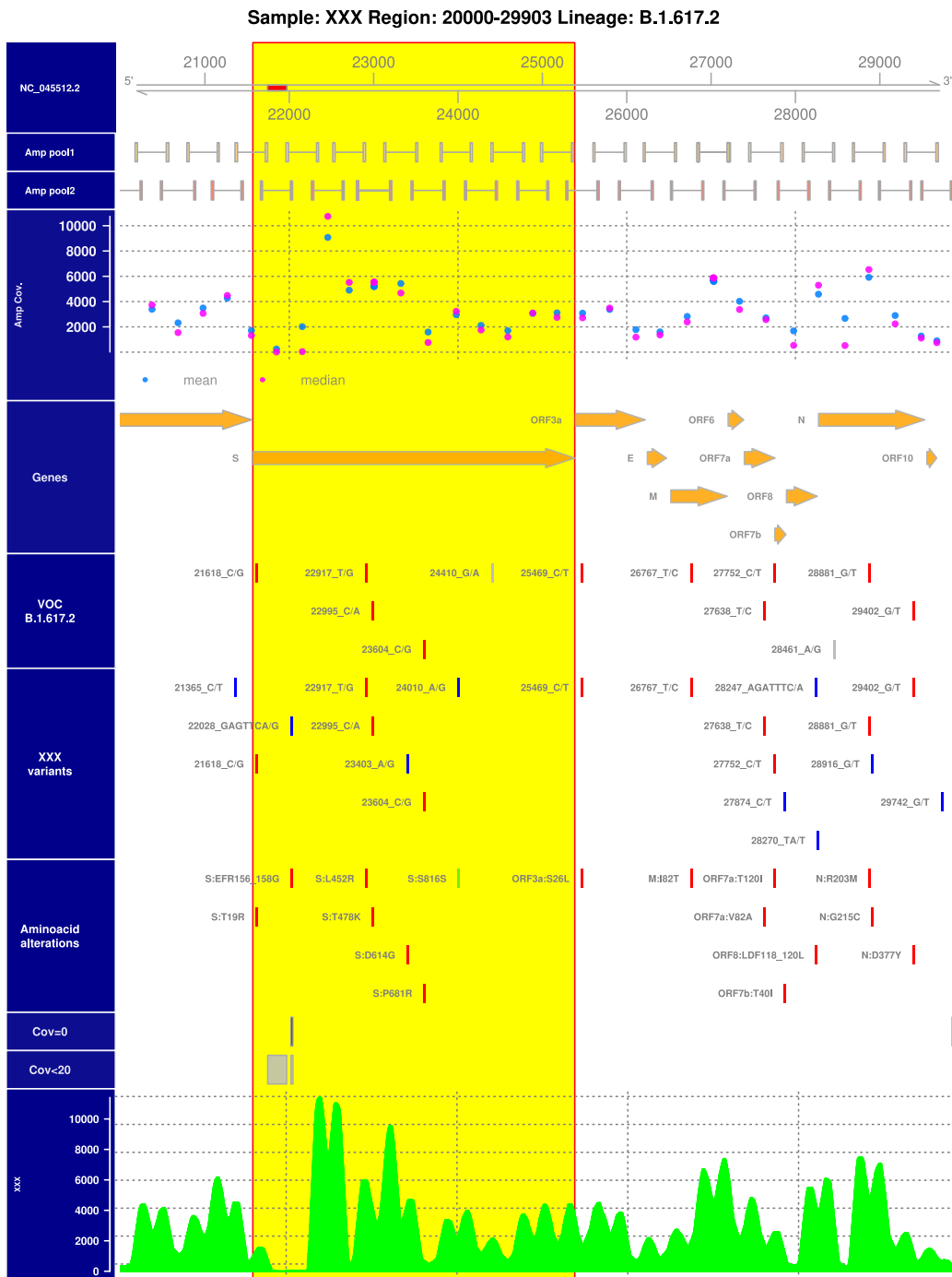
promptly provide a secure and easy to use solution operating on the federated de.NBI cloud computing infrastructure [3]. With the COG-UK [4] and the RKI CovPipe [5] analysis pipelines two widely used, fully automated bioinformatics workflows for the reproducible analysis of SARS-CoV-2 samples could

be offered to CVUA-OWL researchers. The former has widely been used in the COVID-19 Genomics UK Consortium (COG-UK) project which has been a pioneering initiative in the use of large-scale, whole genome sequencing of SARS-CoV-2, with the aim to aid the harmonization of the analysis of sequencing data by providing a standardized analysis workflow. The latter is an alternative bioinformatics pipeline developed by the German Robert Koch Institute (RKI) and widely used in Germany for the analysis of SARS-CoV-2 samples from viral outbreaks. Likewise, to the COG-UK pipeline, the workflow covers and automates all necessary steps from rigorous quality assessment of the input data, read mapping against the SARS-CoV-2 reference genome, over variant calling and generation of a consensus sequence of the virus containing the in the analysis process detected mutations (genomic variants) to lineage assignment. In addition, BIBI researchers enhanced these standard workflows with additional capabilities for comprehensive result visualizations (Figures 1 and 2) allowing to generate epidemiological information that is easily interpretable by public health institutions.



**Figure 1:** The genomic dynamics of SARS-CoV-2. Sketched are genomic variants – with respect to the original SARS-CoV-2 corona virus responsible for the initial outbreak in Wuhan/China – in protein coding regions of different virus strains circulating in Europe over the time of the pandemic. Variants are annotated with their allele frequencies and effect on the corresponding proteins amino acid sequence. If available, the calendar week of first occurrence in Ostwestfalen-Lippe (OWL), based on genome sequencing data of our project partner CVUA-OWL, is given in blue. The B.1 lineage is a large European lineage whose origin roughly corresponds to the Northern Italian outbreak in March 2020. It was the dominant lineage in Germany until summer 2020. The B.1.177 lineage spread mostly over Europa after opening of borders in summer 2020 and appeared for the first time in the OWL data in week 48/2020. B.1.1.7 (Alpha) is the first variant of concern (VOC) that spread over Europe. It was first detected in the United Kingdom in September 2020 and is associated with the N501Y mutation and with evidence for having higher transmissibility than other lineages resulting in rapid growth in the UK and internationally. At the beginning of 2021 (week 05/2021 in OWL) Alpha starts to push away other virus variants and quickly became the dominant variant in spring 2021. The B.1.617.2 (Delta) lineage was first detected in October 2020 in India and classified as a VOC in Mai 2021. It has much increased transmissibility compared to Alpha and is linked to a significantly higher risk of severe COVID-19 disease progression and death. In OWL it appeared in week 25/2021 and since July 2021 it has become the dominant virus lineage accounting today for more than 99 percent of all SARS-CoV-2 infections in Germany. With P681R, L452R and T478K it carries several mutations in the virus' spike protein (S gene) that are linked to higher virus load due to increased replication rates and the ability to partially escape neutralizing antibodies generated by the hosts immune response.





**Figure 2: Analysis report.** Analysis results of a SARS-CoV-2 positive sample from the Ost-westfalen-Lippe region that was classified as belonging to the highly infectious Delta variant (B.1.617.2) of the virus. Shown is an excerpt of the virus genome with coverage information of the used sequencing amplicons, gene annotation, VOC B.1.617.2 defining variants, variants called for this sample with their amino acid alterations and genomic coverage information. The part of the genome coding for the spike protein (S gene) is marked in yellow. Detected, lineage (Delta) specific genomic variants are colored red in the samples variant track.

**References:**

- [1] Schulte-Schrepping et al. Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell*, 2020. DOI: 10.1016/j.cell.2020.08.001.
- [2] Brandt et al. Multiple Occurrences of a 168-Nucleotide Deletion in SARS-CoV-2 ORF8, Unnoticed by Standard Amplicon Sequencing and Variant Calling Pipelines. *Viruses*, 2021. DOI: 10.3390/v13091870.
- [3] German Network for Bioinformatics Infrastructure - de.NBI: <https://denbi.de/cloud>
- [4] COG-UK (ncov2019-artic-nf) pipeline: <https://github.com/connor-lab/ncov2019-artic-nf>
- [5] RKI CovPipe pipeline: [https://gitlab.com/RKIBioinformaticsPipelines/ncov\\_minipipe](https://gitlab.com/RKIBioinformaticsPipelines/ncov_minipipe)





# Omic Fusion

– a web application to analyze and integrate microbial data from multiple omics sources

**Stefan P. Albaum; Nils Kleinbölting; Bielefeld University, Bielefeld**

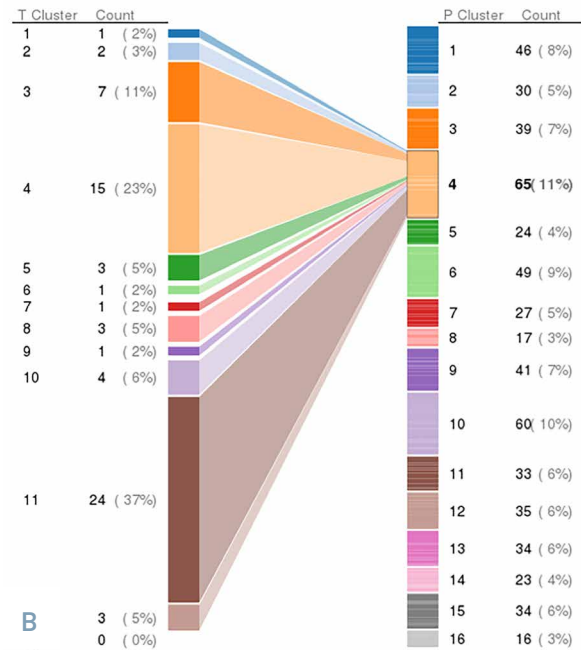
Understanding a system as a whole often requires to consider its different levels and their connections [1]. For a ‚biological organism‘ these levels are represented by: the genome (which genes are encoded on the DNA level), the transcriptome (what and how much is transcribed from the genome to mRNA), the proteome (what is translated into proteins/enzymes) and the metabolome (which metabolites are present – often as products of enzyme processing). The mere presence of a particular gene or allele, respectively,

is not providing insights into a possible number of transcripts of this gene, and even less into the synthesis rates of the corresponding protein or the abundance of specific metabolites synthesized by specific enzymes. Many factors may influence this process. Moreover, synthesized proteins will in turn affect the actual metabolism of an organism. A holistic study of a living organism therefore has to consider all these different “omics” levels including a view into an organism's transcriptome, proteome and metabolome.



**A**

Name	Time: 0min	Time: 10min	Time: 30min	Time: 60min	Time: 120min
aadK <span>T</span>	1.550	-1.141	-0.268	0.255	-0.397
aapA <span>T</span>	-0.059	-0.118	1.607	-0.279	-1.150
abfA <span>T</span>	-1.269	-0.060	1.537	-0.138	-0.071
abh <span>T</span>	1.056	0.541	0.369	-0.471	-1.495
abnA <span>T</span>	0.057	0.936	-1.100	1.030	-0.923
abrB <span>T</span>	1.298	0.615	-0.840	-1.109	0.036
accA <span>P</span>	1.192	-0.898	-1.172	0.353	0.524
accA <span>T</span>	0.647	1.006	-1.362	-0.726	0.435
accB <span>P</span>	0.317	1.418	-1.269	-0.524	0.058
accB <span>T</span>	1.065	-0.152	-1.812	0.347	0.352



Omics Fusion has been developed to support researchers in the analysis of such datasets [2]. The web application provides a comprehensive portfolio of methods to analyze data from different levels of omics in an integrative manner. It is freely available to the public and provided following the software as a service paradigm. Omics Fusion has initially been designed for microbial data, but also has successfully been used for plant and other higher organisms data including human data.

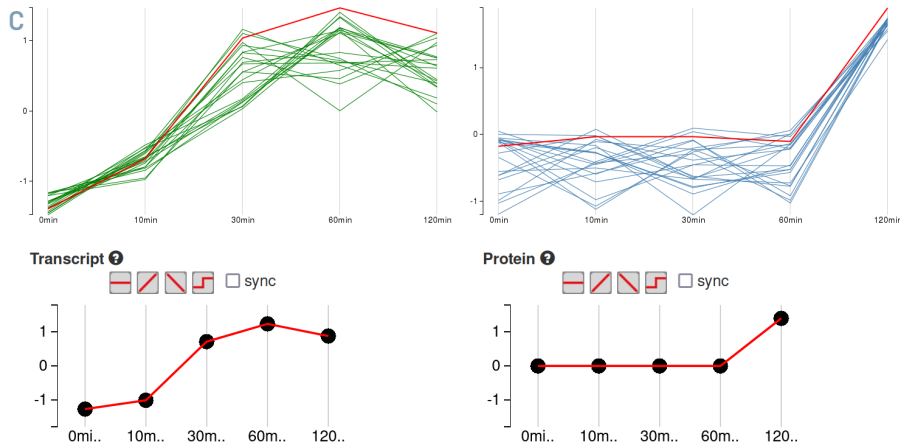
Starting with the upload of data tables containing transcript counts, quantitative protein ratios or metabolome abundance values from high-throughput experiments users can draw on a collection of tools for integrative data analysis and data visualization to gain new insights into a biological system under investigation. This includes functionality to filter, normalize and transpose data and analyses such as variance and regression analysis to determine significantly differentially abundant

transcripts, proteins or metabolites. Methods such as principal component analysis and t-SNE can provide an initial overview on the data by reducing its dimensionality and thereby increasing the interpretability.

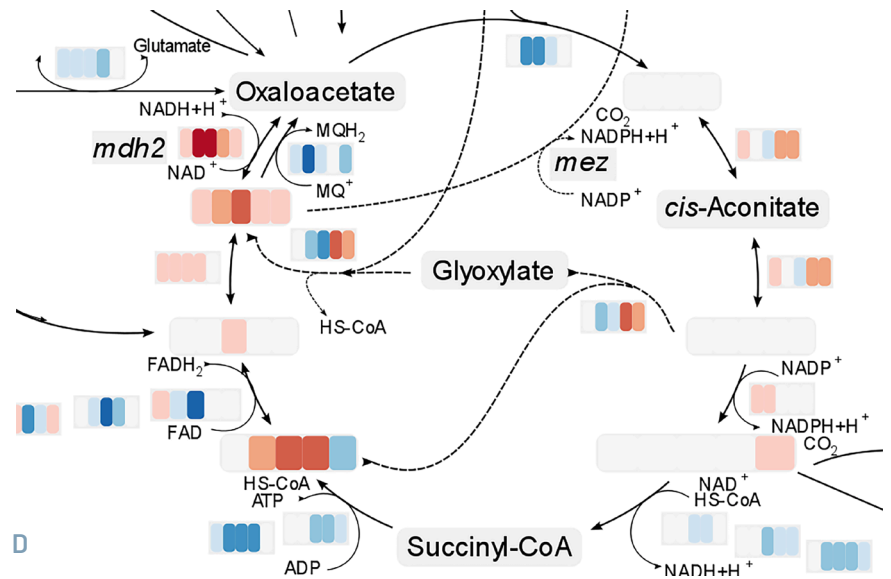
Omics Fusion places a particular emphasis on unsupervised learning methods. Cluster analysis, for example, allows to identify groups of transcripts, proteins and metabolites that show a similar pattern of expression or abundance. A common problem, in this regard, is the determination of an optimal number of clusters fitting to the data. Our software offers various means to apply cluster algorithms on the data and features a fully automated procedure to

determine optimal clustering solutions. Furthermore, specialized cluster methods have been developed and allow, inter alia, the combined detection of transcripts and proteins that show similar patterns of abundance.

Extensive visualization methods enable explorative ways to better understand the data. An intuitive presentation is the combined mapping of transcript, protein and metabolome data on metabolic pathway maps. The tool box for visual and meaningful representation of data, moreover, contains scatter plots, box- and whisker plots and parallel coordinate plots. An important element for understanding biological data is the enrichment of the quantitative infor-



**Figure 1:** Screenshots of the web application Omics Fusion: A) data management, B) cluster analysis, C) cluster profiling visualization, D) pathway mapping.



mation by further descriptions of gene functions or metabolite characteristics and their role in an organism. For this purpose, Omics Fusion provides interfaces to common data repositories as provided by the NCBI [3] and the Uniprot consortium [4] to retrieve annotation data such as enzyme classifications, well-defined function descriptions or metabolic pathway associations. With this information, connections between the data may become visible that beforehand were not obvious.

Omics Fusion is continuously developed. Depending on the needs of the community, new methods are being integrated in the software or existing methods adapted.

#### References:

- [1] Zitnik et al. *Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. An international journal on information fusion*, 2019. DOI: 10.1016/j.inffus.2018.09.012.
- [2] Brink et al. *Omics Fusion - A Platform for Integrative Analysis of Omics Data. Journal of integrative bioinformatics*, 2016. DOI: 10.2390/biecoll-jib-2016-296.
- [3] Benson et al. *GenBank. Nucleic acids research*, 2013. DOI: 10.1093/nar/gks1195.
- [4] Uniprot Consortium. *UniProt: the universal protein knowledgebase in 2021. Nucleic acids research*, 2021. DOI: 10.1093/nar/gkaa1100.



# GRADUATE SCHOOL

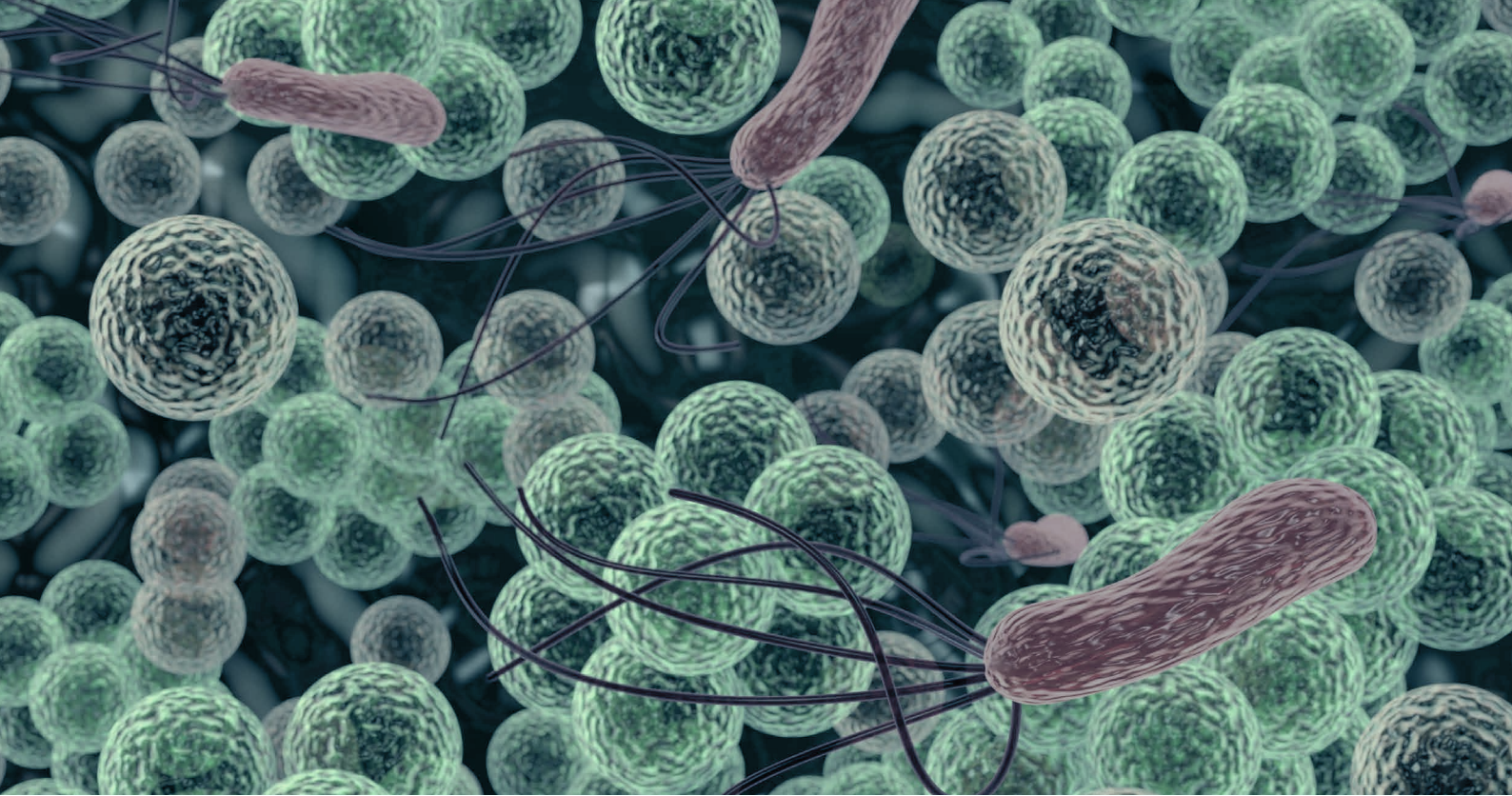
*“Digital Infrastructure for  
the Life Sciences” (DILS)*



***“DILS is a win-win constellation: The PhD students benefit from the academic infrastructure and scientific network in BIBI, and BIBI benefits from active research and method development of the PhD students. Furthermore, it is a platform for launching further bilateral research projects.”***

Dr. Roland Wittler, Coordinator DILS





# Large scale detection of regulatory small RNAs in pathogenic bacteria

Muhammad Elhossary; *ZB MED, Cologne*

Life threatening diseases caused by bacteria combined with a growing drug resistance of these species is a world major health concern. *Gammaproteobacteria* is one important class of bacteria that comprises many critically pathogenic members that are hard to treat. These pathogens adapt to the continuous changes of their surrounding environment. This adaptation involves numerous complex biological processes. The transcription regulation, i.e. the control of gene expression, is one essential process among them. Generally, in bacteria, a class of RNA

known as regulatory small RNAs (sRNA) plays a vital role in regulating gene expression post-transcriptionally. Those regulatory RNAs can influence the activity of messenger RNAs (mRNA) by several mechanisms before they get translated into proteins [1]. The most common mechanism is anti-sense base-pairing between the sRNA and the mRNA which for example can cause translation blockage by binding to ribosome binding sites or within the open reading frame region [2]. Bacteria typically express hundreds of these regulatory RNAs, and each can regulate

several messenger RNAs differently. Up to date, their evolution and biological functions remain largely unknown despite their described importance.

The first step in understanding the regulation of sRNAs is to detect them and then identify their interactions with their target mRNAs. High-throughput sequencing technologies can be utilized to perform global transcriptome and interactome measurements that help to reveal, identify and characterize these regulatory small RNAs [3]. This is followed by a downstream computational analysis using tools that can generate genome-wide high resolution small RNA annotations and characterizing their regulatory networks. For this purpose, a diverse set of twenty species of the *Gammaproteobacteria* class grown in different conditions are studied. Among this set of species, there are model species such as *E. coli* that were extensively studied and can serve as references in comparative studies for poorly studied species.

We will offer queryable interconnected data and have modeled the regulatory networks which afterwards will be integrated into a web platform that provides comparative views of small RNAs and their regulatory networks to efficiently help researchers to address further unanswered biological questions. Furthermore, tools and pipelines developed will also be published, ensuring their easy reproducibility and reuse.

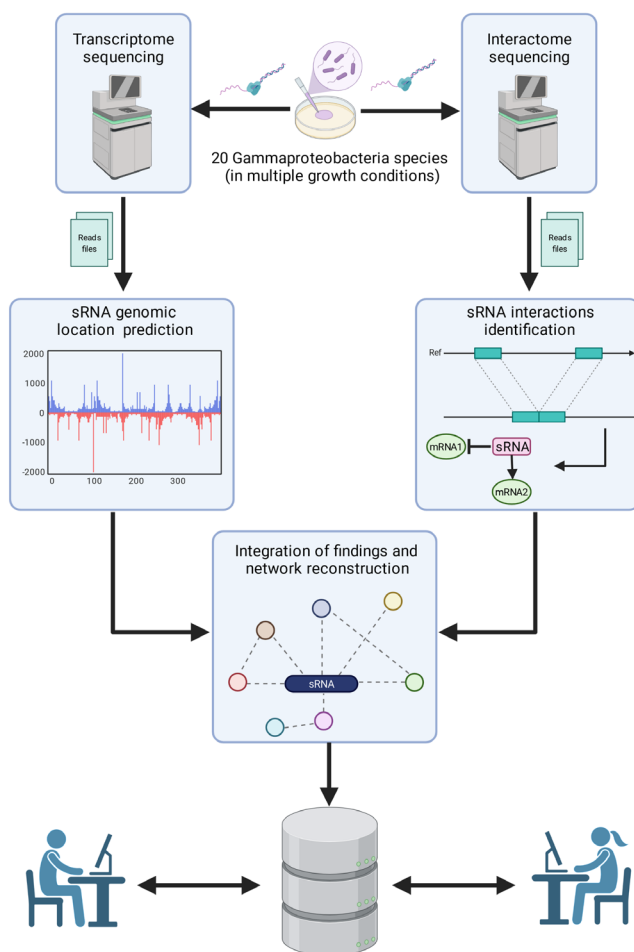


Figure 1: Project plan overview



#### References:

- [1] Holmqvist and Wagner. Impact of bacterial sRNAs in stress responses. *Biochemical Society Transactions*, 2017. DOI: 10.1042/bst20160363.
- [2] Beisel and Storz. Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiology Reviews*, 2010. DOI: 10.1111/j.1574-6976.2010.00241.x.
- [3] Saliba et al. New RNA-seq approaches for the study of bacterial pathogens. *Current Opinion in Microbiology*, 2017. DOI 10.1016/j.mib.2017.01.001.



G C G T **A** A C T G T A G C G **A** C T  
T T C G A C A G T **A** C G T A G T C  
A A C T G C **T** G A C A T G G **T** A C  
G G A C T C G T A C G A **T** T G A C  
C A T **C** G T **C** A G G T C A C G **A** T  
A C A T G A C T G T **G** C A G T G A  
C A T G C C **T** A G T C A G T G A C  
G A G T C G A T C T C G A G C A T  
A A G C T G A C T G T A C T G **C** A  
A C A T G **A** T C G C A T G T C A G  
C C G T A G C A **T** T C G **A** C T A G  
T A C G T **T** G A C C T G C A G T C  
**T** C T A G T A C **G** C T G A C A G T  
G A G C T C **A** G T C G T A G A C T



# Errors in Sequencing Data

## *Quality Assessment and Bioinformatics Solutions*

**Sebastian Jünemann; Bielefeld University, Bielefeld**

DNA sequencing describes the automated process of reconstructing the ordered sequence of nucleotides constituting a DNA molecule. Next-generation sequencing (NGS) instruments, in particular those of the second-generation, are utilizing nano-scale biochemical processes, e.g. sequencing by synthesis, in a highly parallelized manner on a massive scale to sequence single source DNA molecules in multiple copies and repetitions. The advent of NGS technologies had a huge impact on numerous research fields leading to a downright explosion of all sequence based research fields.

Metagenomics (MG) studies microbial communities by the entirety of its genomic content, the metagenome, and addresses, inter alia, the question what is the taxonomic origin of all individual community members, their function within that community and their interaction with other members, the environment or host [1]. Targeted MG, also known as amplicon sequencing, addresses only a subset of whole genome shotgun MG, the taxonomic composition. Here, only a specific marker gene is focused on, e.g. the 16S rRNA gene, which is being selectively extracted and amplified before the sequencing procedure. Metagenomics, and in particular 16S rRNA gene based amplicon surveys, experienced a renaissance with the advent of NGS technologies due to increasingly cheaper and quicker access to raw genomic data allowing to analyze hundreds of samples in parallel.

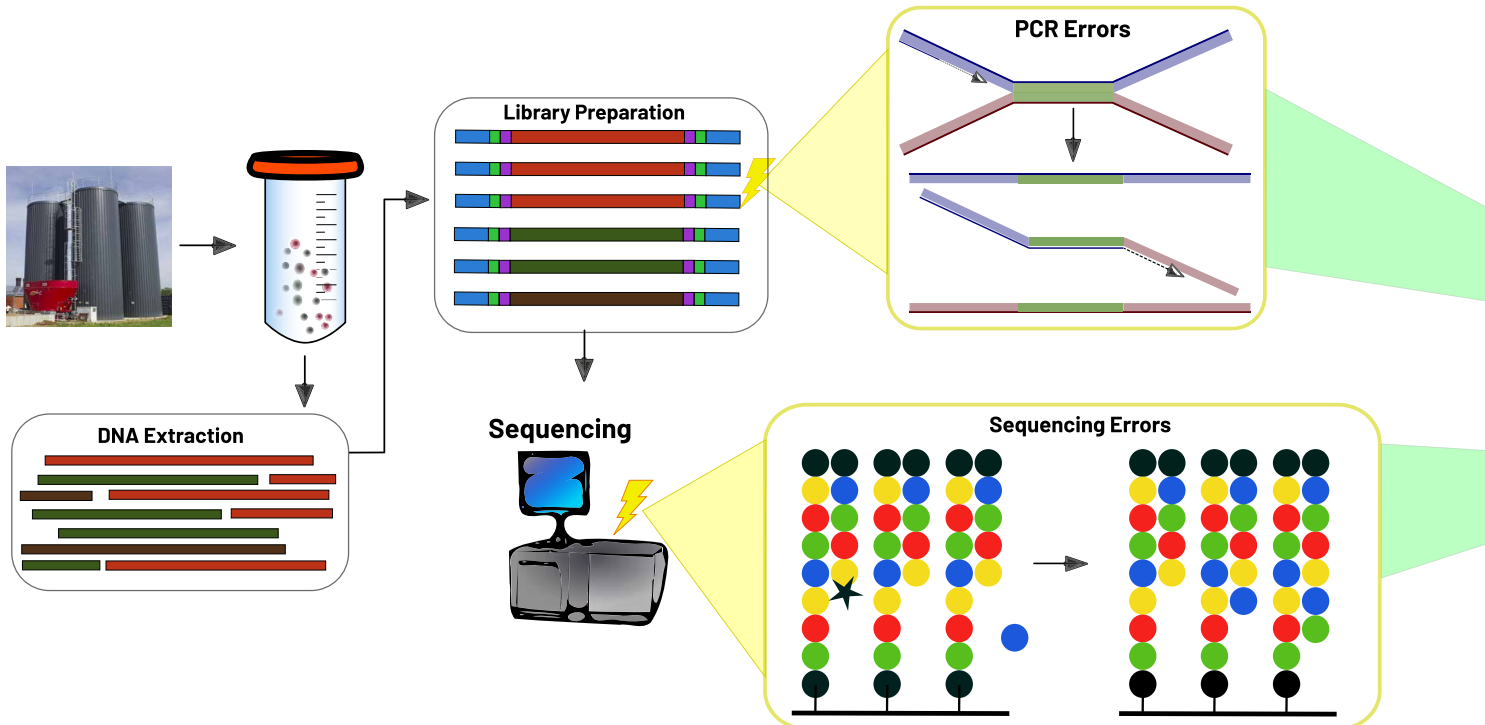
The process of sequencing is remodeling a natural process and is thus naturally affected by different error sources and variation. However, as long as individual errors are random and introduced with a lower probability than their error-free counterparts, errors, in theory, can be compensated by increasing the coverage, i.e. the number each single source DNA molecule is sequenced repetitively. Still, some sequencing errors are methodological and can be impeded only to a certain degree. This means in effect that not all errors are introduced at random but systematically. Thus, each sequencing technology comes with its own error profile [2].

Issues arise also during the library preparation, i.e. the process of treat-

ment from an environmental specimen to a sample ready to be introduced into a sequencing instrument. This usually involves various steps, of which the amplification of DNA material by Polymerase Chain Reaction (PCR) is the most error prone. Errors and artifacts introduced during PCR can have a falsifying impact on the community under study, e.g. artificially inflating species richness or diversity. One of the more difficult error sources to address are so called chimeras, artificial cross hybridized DNA sequences formed from parts of other sequences. Even though their formation can be reduced by adjusting PCR conditions, their formation cannot be prevented completely [3].

Explicit errors, for instance overly long PCR fragments or overlapping sequencing signals, can already be dealt with during library preparation or the sequencing process. For more subtle errors appropriate quality assessment (QA) tools are applied to NGS data to (i) assess the overall data quality, (ii) detect putative erroneous sequencing reads or stretches, (iii) filter the data for errors, or, if possible, (iv) correct the errors [4].

General data QA is done usually by utilizing intrinsic sequence information to generate quality profiles, report about benchmarking properties or to search for known contaminating sequence patterns (e.g. sequencing primers). Such profiles often build the basis for



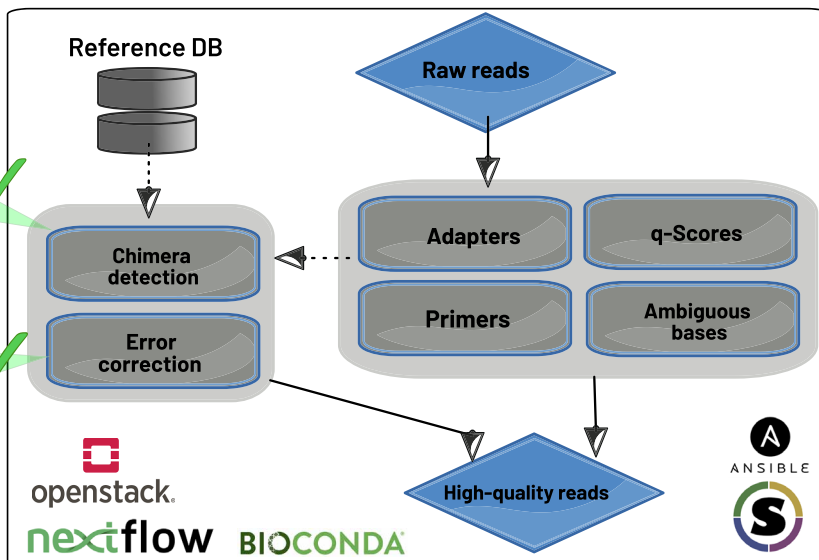
**Figure 1:** Flowchart of targeted Metagenomics from environmental samples to high-quality sequencing data: After sampling and DNA extraction, the DNA library is prepared for and sequenced on a NGS instrument, and errors can occur in both steps (e.g. PCR errors and sequencing errors). These errors are usually addressed by specific algorithms either by filtering erroneous data (e.g. as done by chimeric read detection) or by correcting erroneous data points. In conjunction with other quality assessment (QA) methods, e.g. the removal of adapters and primers and filtering based on low quality scores and ambiguous bases, these algorithms are piped together in a workflow. QA workflows can make use of flexible workflow languages (e.g. using nextflow) and can be deployed as software containers (e.g. using singularity) in cloud environments (e.g. an openstack cloud) for easy access and application.



deducing and applying data specific filtering rules to increase the overall data quality. Detection and filtering of errors can be done extrinsically by comparing sequences of known origin to reference data or intrinsically by matching sequencing reads against each other in order to detect outliers or abnormalities, as done e.g. by our developed chimera checker tool ChimP. Here, a sequence is compared to its potential parents utilizing a specific alignment method and, if the match against different parents is high enough, reported as a chimera. In some occasions, error correction can be applied by grouping reads belonging to the same origin, e.g. based on coverage and nucleotide conformity. Now, random errors can be rectified by calling up a majority based consensus.

Usually, different QA tools cover only one or few of the aforementioned issues. To this end, several tools are usually piped together in a consecutively applied workflow to deal with all potential error sources in a meaningful manner (Figure1). However, NGS is a rapidly evolving field. Sequencing protocols change and new technologies emerge. Thus, old tools need to be adopted or new ones developed tailored to instrument specific properties. In addition, the exponential trend at which sequencing data is being generated is still ongoing. This presents a challenge on the performance of bioinformatics solutions, leaving data processing as the new bottleneck in the field of NGS based research. Moreover, fully standardized QA workflows are, until now,

still no integral part of SOPs of data generators, researchers, or public data archives. This complicates data exchange, reproducibility, and comparability. Therefore, joint and community driven effort is necessary to define and guarantee a minimum data quality standard, and the harmonization of QA processes to be – eventually – integrated into a FAIR life cycle, e.g. by reaching out to the National Research Data Infrastructure (NFDI) consortia.



#### References:

- [1] Jünemann et al. Bioinformatics for NGS-based metagenomics and the application to biogas research. *JBioTec*, 2017. DOI: 10.1016/j.jbiotec.2017.08.012.
- [2] Jünemann et al. Updating benchtop sequencing performance comparison. *Nat Biotechnol*, 2013. DOI: 10.1038/nbt.2522.
- [3] Schloss et al. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS one*, 2011. DOI: 10.1371/journal.pone.0027310.
- [4] Chen et al. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics*, 2017. DOI: 10.1186/s12859-017-1469-3.

# From hidden data and information towards data-driven research

Lisa Langnickel; ZB MED, Cologne

Machine readability and access to data, information and knowledge are core requirements for data-driven research. Furthermore, the enormous growth in freely available, electronic research data increases the need for semantic interoperability as well as computational methods to generate new information and knowledge from the data.

This, however, implies that all published data is stored in a machine-readable format and that data can be accessed. Especially in the medical area, this is hampered by the heterogeneity and missing standardization of the data as well as the restricted access and availability of high quality data.

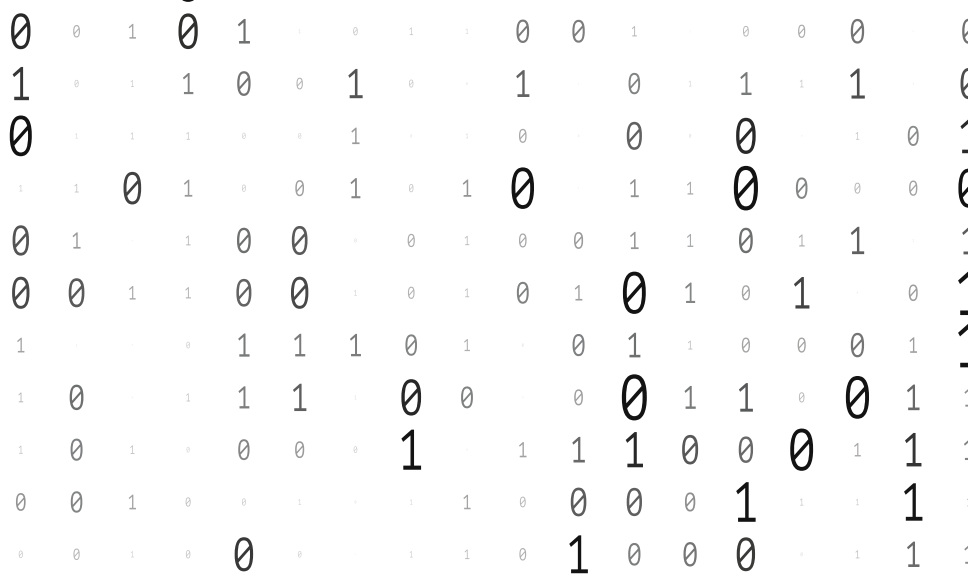
Concerning the literature, there is an increasing need for text mining solutions to make a transfer from unstructured text to machine readable information possible. During the past 20-30 years, intense research has been done in the field of natural language processing (NLP) and also in the specific application field of bio-medical NLP (bioNLP). Currently, Artificial Intelligence (AI) methods seem to be superior to traditional approaches. The success of those methods is, however, dependent on high-quality, labeled data whose availability is strongly limited. In addition, it still remains open whether the results achieved on the specific training/test data are transferable to real world applications.

Another obstacle for data analysis in the medical domain is the access to personalized health data. Despite the growing amount of freely available data, personal data (e.g. clinical or epidemiological data) is usually not publicly available due to data privacy. To circumvent data protection, machine learning (ML) methods will be investigated in order to generate synthetic data.

The overarching aim of this project is to investigate computational methods in order to make biomedical data and information available in a machine-readable format and, thereby, supporting researchers.







## Two examples

### Investigating the robustness of transfer learning-based NER methods

We investigated the robustness of current state-of-the-art text mining methods, such as BioBERT [1], in the area of Named Entity Recognition (NER) of diseases. These machine learning (ML)-based methods are usually trained and evaluated on specific, relatively small corpora and evaluations on corresponding test sets show promising results. For NER of diseases, two different manually labeled data sets are publicly available which consist of training, development and test data. Our first investigation focused on cross-corpora evaluation: training on one dataset and evaluation on the test set of the other data set. We could show that the model achieves an F1-score of only 68% – a drop of about 20% compared to the original test set [2].

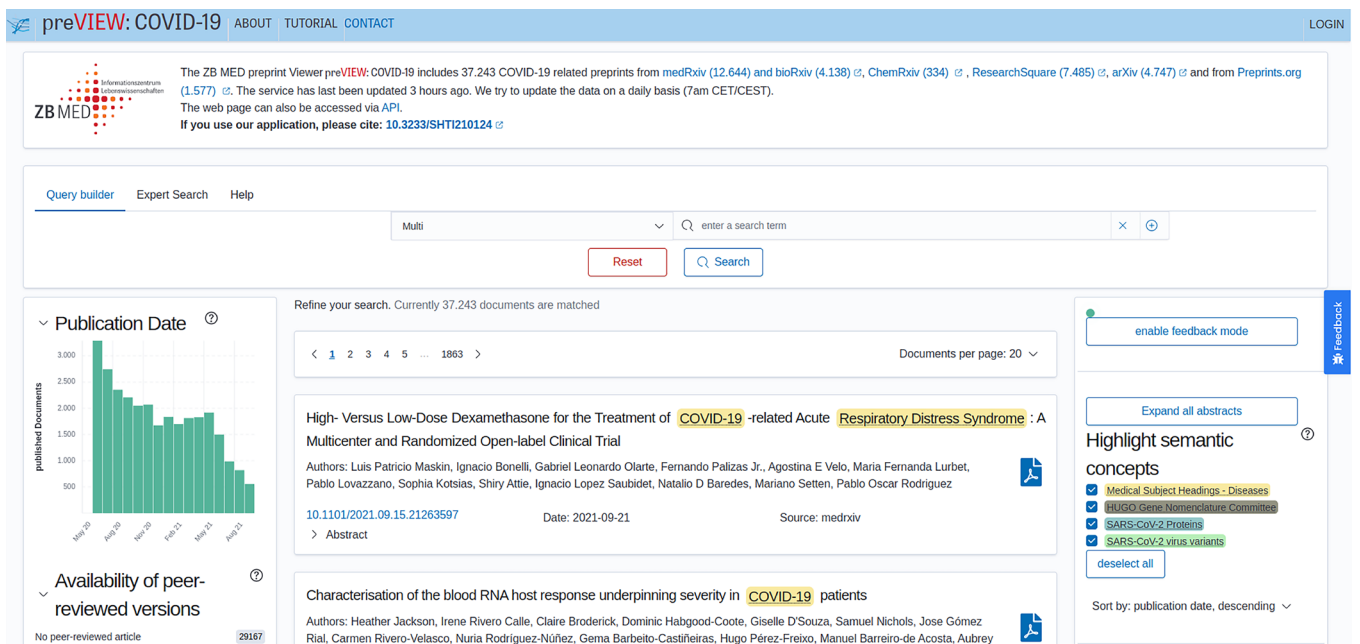
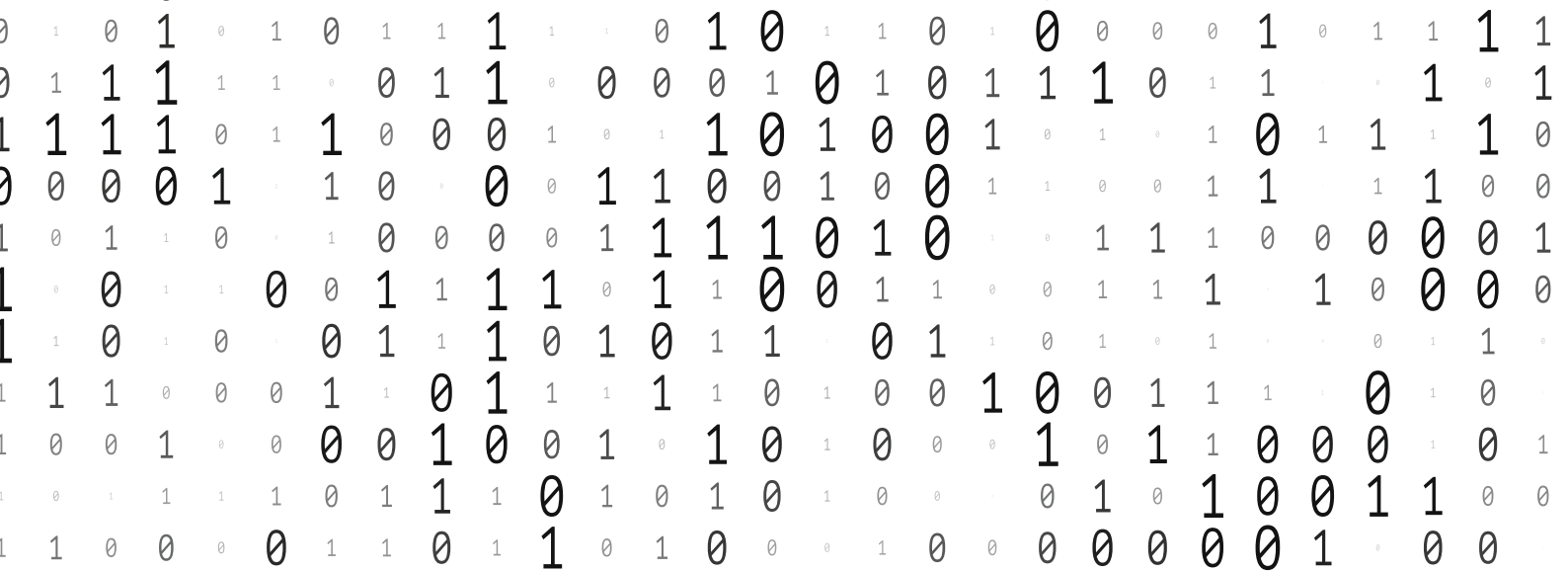
Provided ML-based models are able to generalize, comparable results would be expected from data sets following the same annotation guidelines. An analysis of the two different data sets revealed that the training and corresponding test set (belonging to the same data set) are similar in wording and topics while the data sets as a whole do not. This leads to the assumption that a model trained on one available corpus is not applicable to real world cases and needs to be continuously retrained (called continual learning). Currently, we are investigating compute- and resource-efficient methods.

### Service Science – COVID-19 underlines the need for text mining-based solutions

The current COVID-19 pandemic underlines the need for text mining methods as more than 100 papers – mostly in form of preprints – are currently published per day which makes it infeasible for a human to read all of them. In order to support researchers to cope with this huge amount of information, we set up a text mining-based semantic search engine, called preVIEW, that currently contains more than 37,000 preprints from seven different preprint servers, such as bioRxiv and medRxiv [3]. In accordance to our previous research, we found out that the current machine learning-based state-of-the-art methods are not applicable to services/real world cases because they do not generalize well and are not consistently able to recognize new terms. For

Background: In India, a large number of patients with coronavirus disease - 2019 (COVID-19) presented with common symptoms including fever, dyspnea, cough, musculoskeletal symptoms (fatigue, myalgia, joint pain) and gastrointestinal symptoms. However, information is

Figure 1: Excerpt of an abstract (doi: 10.1101/2021.07.06.21260115), annotated with disease mentions (screenshot taken from <https://preview.zbmed.de>).



example, for the recognition of diseases, the ML-based algorithm TaggerOne [4] missed new terms like COVID-19. As text mining is nevertheless needed to index these high amounts of preprints and thereby find relevant literature, we extended the text mining workflow with additional rule based components and re-evaluated the resulting annotations. Moreover, for new entity classes – i.e. SARS-CoV-2 specific virus proteins and variants of interest – a dictionary-based approach was implemented due to the

lack of training data for supervised learning algorithms.

Whereas preVIEW was developed as a fast prototype together with the user community in the beginning of the crisis, it has been continuously improved towards a sustainable system [5]. In addition, it is currently undergoing evaluation by BioCreative Interactive text mining track [6] in order to evaluate the system usability by a variety of end users.

**Figure 2:** Screenshot of our semantic search engine preVIEW, freely accessible under <https://preview.zbmed.de>.

#### References:

- [1] Lee J et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020. DOI:10.1093/bioinformatics/btz682.
- [2] Langnickel L et al. We are not ready yet: limitations of transfer learning for disease named entity recognition. *bioRxiv*, 2021. DOI:10.1101/2021.07.11.451939.
- [3] Langnickel L et al. COVID-19 preVIEW: semantic search to explore COVID-19 research preprints. *Public Health and Informatics*, 2021. DOI: 10.3233/SHTI210124.
- [4] Leaman R et al. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 2016. DOI:10.1093/bioinformatics/btw343.
- [5] Langnickel et al. preVIEW: from a fast prototype towards a sustainable semantic search system for central access to COVID-19 preprints. *Journal of European Association for Health Information and Libraries*, 2021. DOI:10.32384/jeahil17484.
- [6] BioCreative – Track 4 – COVID-19 text mining tool interactive demo. Accessed October 5, 2021. <https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-4/>









# Biogas- GeneMining

*Metagenomics survey unravels the potential of biogas microbiomes*

**Benedikt Osterholz; Bielefeld University, Bielefeld**

## **The importance of biogas in the portfolio of green fuels**

In times of environmental pollution and global warming, it is important to replace fossil fuels by renewable forms of energy. Solar energy and wind power significantly contribute to these forms of energy. However, also recovery of energy from digestion of biomass has its place among the renewable energies. Biogas produced through decomposition of organic substrates can be converted to electricity and heat and is also transportable using the existing natural gas infrastructure. Storage of biogas for later combustion is also feasible.

Biogas is mostly generated in agricultural biogas plants from energy crops, residual material from agriculture and manure as input substrates. The corresponding anaerobic digestion process is considered to represent the most efficient bioenergy production pathway known [1].

## **Potential for improvement**

In the anaerobic digestion of biomass, a huge number of microbial species is involved. These possess a great variety of metabolic properties, that are exploit-

ed to generate the desired methane within biogas [2,3].

However, the majority of these species that can be detected in biogas reactors have not been adequately characterized, either in terms of their biomass conversion properties or in terms of their respective ecological roles within the microbiological system. Accordingly, the trophic network responsible for the degradation of crop biomass in biogas reactors is only partially understood [2, 4, 5].

Deeper knowledge about interactions between different microbial species considering their metabolic properties is expected to enhance the overall performance of the biogas process. An improved monitoring, management and engineering of biogas microbiomes is envisioned.



### **Analysis of biogas microbiomes applying methods of metagenome research**

To study the function of biogas microbiomes, data on these microbial communities are needed. The overall goal is to access and use publicly available metagenome datasets originating from biogas microbiomes. Sequencing of metagenomic DNA allows to portray the microbial community of a sample and not only the cultivable species, which is why metagenome analyses are preferred when the whole picture is to be viewed [3, 6, 7, 8].

Currently, there are 386 biogas metagenomes that are used for this project. These large data volumes, typical for the life sciences, present another huge problem to tackle.

A lot of different steps are necessary to implement established bioinformatics solutions and concepts for comparative analyses of metagenome datasets representing biogas microbiomes. Certainly, it is not feasible to manually

perform all data processing steps for a large number of samples. Therefore, an automated bioinformatics workflow, the **Metagenomic-Toolkit**, was developed and implemented. Apart from the realization of the general feasibility, our workflow also offers additional benefits like guaranteed reproducibility of all steps, a high rate of portability and support of key cloud computing based technologies.

### **Parallelization using cloud computing**

After the implementation of a suitable workflow, nothing is gained if it does not finish in an adequate time frame, which is a real problem when hundreds of datasets have to be processed. The performance of single computers is not sufficient, for what reason the **Metagenomic-Toolkit** was developed to be executed in a cloud computing environment (Figure 1). All the different steps were parallelized by the workflow through distribution of jobs over hundreds of different compute nodes at the same time and compilation of obtained

# Metagenomic-Toolkit: **nextflow**

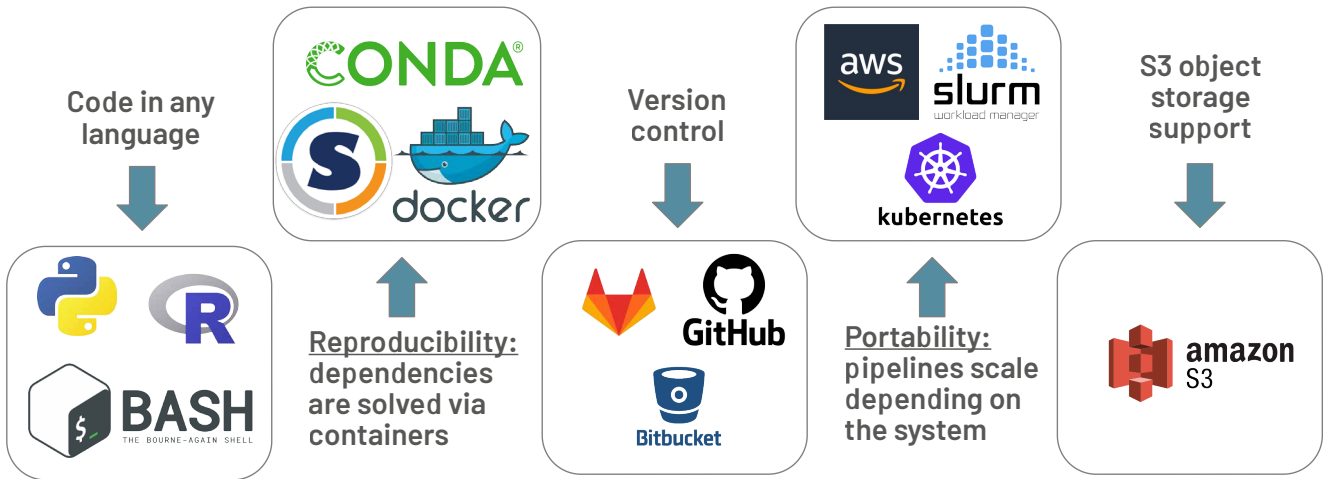


Figure 1: Overview of special features within the Metagenomic-Toolkit.

results at the end of the computation process. In this way, the work time is reduced to an acceptable minimum.

## The goal

Implementation of the Metagenomics-Toolkit is expected to enable representation of detailed microbial networks, identify the core microbiome of biogas communities, determine unique taxa for specific sub-communities and to elucidate relationships between taxonomic units by means of co-occurrence and network analyses. The overall aim of the project is to further upgrade and optimize the biogas process as a whole.

01	Quality control	<ul style="list-style-type: none"> <li>fastp</li> <li>CheckM</li> </ul>
02	Assembly	<ul style="list-style-type: none"> <li>Megahit</li> </ul>
03	Read mapping	<ul style="list-style-type: none"> <li>bwa</li> </ul>
04	Binnig	<ul style="list-style-type: none"> <li>Bowtie</li> <li>Metabat</li> <li>MaxBin</li> </ul>
05	Annotation	<ul style="list-style-type: none"> <li>Diamond</li> <li>Blast</li> </ul>
06	Dereplication	<ul style="list-style-type: none"> <li>Pasolli (ANI distance based)</li> <li>Almeida (dRep based)</li> </ul>
07	Co-occurrence	<ul style="list-style-type: none"> <li>CarveMe</li> <li>Memote</li> <li>Smetana</li> </ul>

Figure 2: The Metagenomics-Toolkit: All available modules and processes that can be combined by the user.

## References:

- [1] Antoni et al. Biofuels from microbes. *Appl Microbiol Biotechnol*, 2007. DOI: 10.1007/s00253-007-1163-x.
- [2] Hassa et al. Indicative marker microbiome structures deduced from the taxonomic inventory of 67 full-scale anaerobic digesters of 49 agricultural biogas plants. *Microorganisms*, 2021. DOI: 10.3390/microorganisms9071457.
- [3] Maus et al. Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates. *Biotechnol Biofuels*, 2016. DOI: 10.1186/s13068-016-0581-3.
- [4] Batstone et al. A review of ADM1 extensions, applications, and analysis: 2002–2005. *Water Sci Technol*, 2006. DOI: 10.2166/wst.2006.520.
- [5] Bremges et al. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. *GigaScience*, 2015. DOI: 10.1186/s13742-015-0073-6.
- [6] Lang et al. Novel florfenicol and chloramphenicol resistance gene discovered in Alaskan soil by using functional metagenomics. *Applied and Environmental Microbiology*, 2010. DOI: 10.1128/AEM.00323-10.
- [7] Hassa et al. Metagenome, metatranscriptome, and metaproteome approaches unraveled compositions and functional relationships of microbial communities residing in biogas plants. *Applied Microbiology and Biotechnology*, 2018. DOI: 10.1007/s00253-018-8976-7.
- [8] Maus et al. Genomics and prevalence of bacterial and archaeal isolates from biogas-producing microbiomes. *Biotechnol Biofuels*, 2017. DOI: 10.1186/s13068-017-0947-1.





# Comparing pangenomes

Luca Parmigiani; *Bielefeld University, Bielefeld*

With the advent of Next-Generation Sequencing technologies, the number of genomes we have at our disposal for each different species is increasing – opening new prospects previously not feasible.

The practice of sequencing and genome analysis, for what regards most of the already sequenced species,

usually involves at some point comparing the nucleotide sequences, previously extracted from the sample, to some reference genome. This reference genome fails to account for all the variability present in nature and can not truly represent a whole species.

In 2005 the term *pangenome* was used by Tettelin et al. [1] to describe the set of all distinct genes present in a species: either present in all genomes, defined as core genes, or present in just

some, called dispensable genes. Their goal was to know how many genomes should be sequenced to fully describe a bacterial species. While this concept of pangenome was later extended beyond bacteria, to plants and animals, one of the most outstanding discoveries at the time was that some species possess

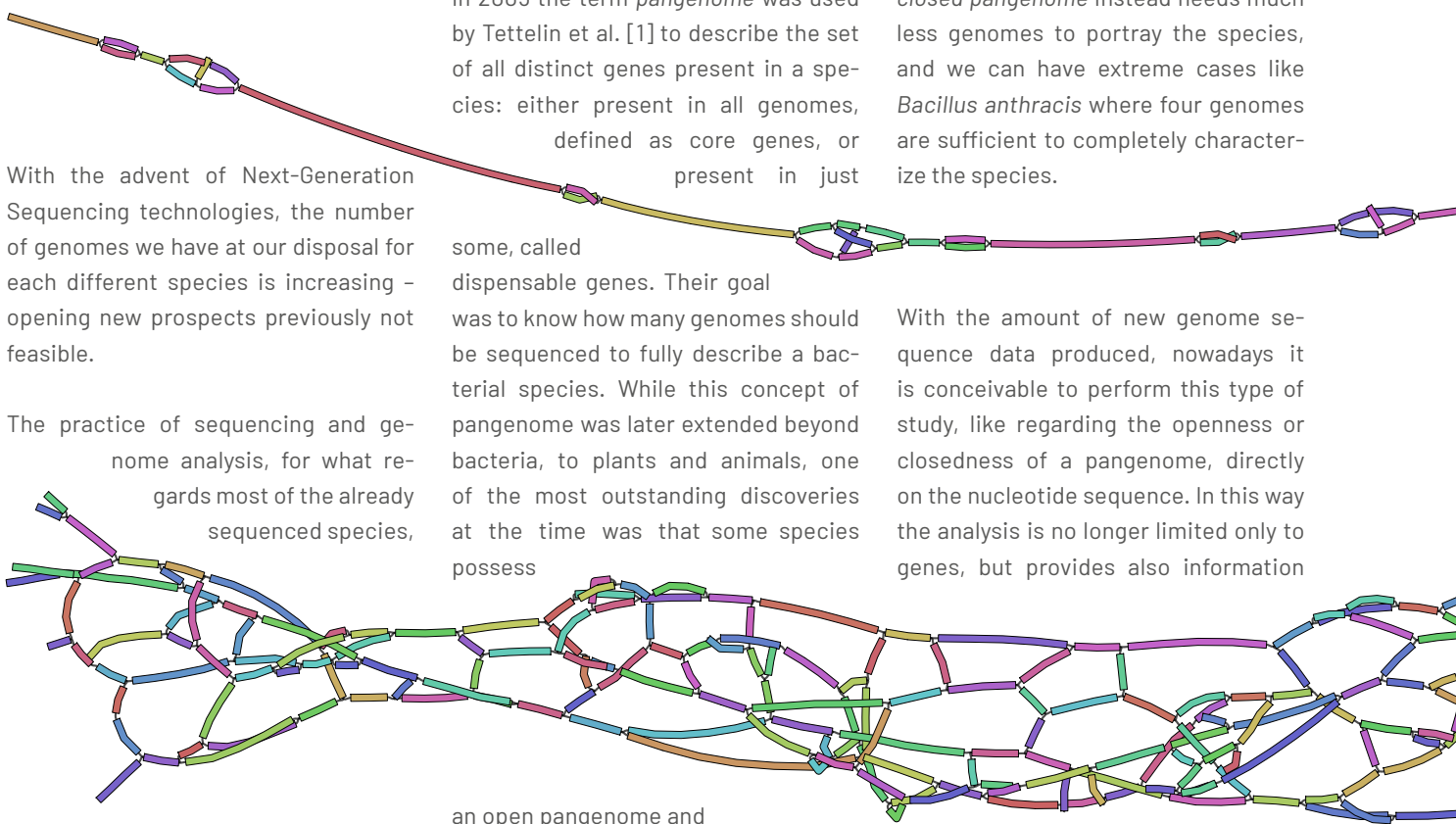
an open pangenome and others a closed pangenome.

For an open pangenome the number of genomes that has to be sequenced in order to get a full picture of the species is large, since new distinct genes are found each time a new sequenced genome is inserted in the pangenome. A

*closed pangenome* instead needs much less genomes to portray the species, and we can have extreme cases like *Bacillus anthracis* where four genomes are sufficient to completely characterize the species.

With the amount of new genome sequence data produced, nowadays it is conceivable to perform this type of study, like regarding the openness or closedness of a pangenome, directly on the nucleotide sequence. In this way the analysis is no longer limited only to genes, but provides also information

about noncoding sequences, small RNAs or other repeated structures.



Here we show early results of the comparison between two different pangenomes, one derived from 58 genomes of *Escherichia coli*, and one from 58 genomes of *Yersinia pestis*.

Even though the length of the genomes of *E. coli* and *Y. pestis* are comparable – 5 million versus 4 million base pairs, respectively – these

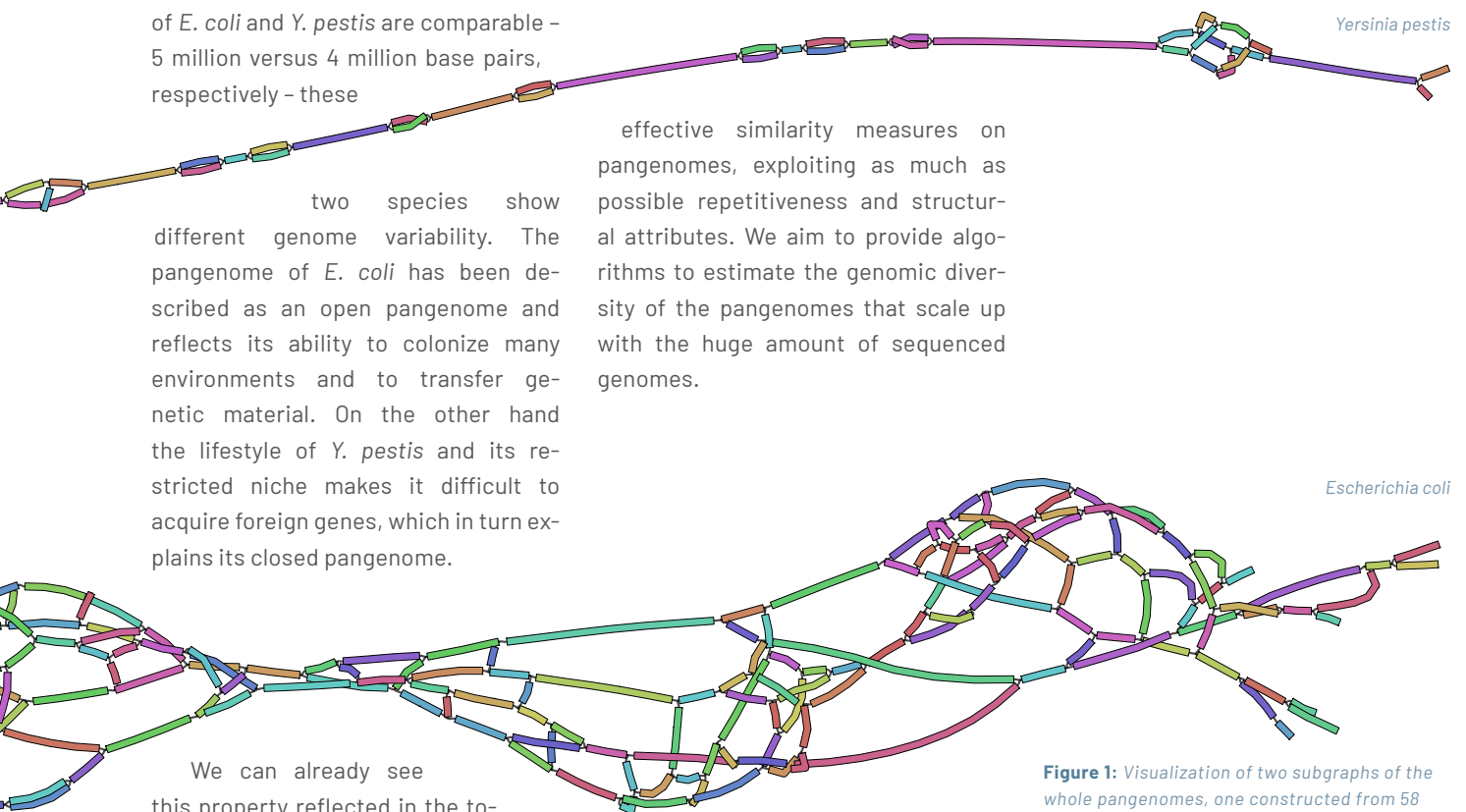
two species show different genome variability. The pangenome of *E. coli* has been described as an open pangenome and reflects its ability to colonize many environments and to transfer genetic material. On the other hand the lifestyle of *Y. pestis* and its restricted niche makes it difficult to acquire foreign genes, which in turn explains its closed pangenome.

We can already see this property reflected in the topology of the graphs shown in Fig. 1. In the two pictures we present a subgraph of the total pangenome for the two

species. Every time a position in the genome contains multiple variations among the species, the graph reflects it by branching and creating multiple paths.

These differences are crucial to define

effective similarity measures on pangenomes, exploiting as much as possible repetitiveness and structural attributes. We aim to provide algorithms to estimate the genomic diversity of the pangenomes that scale up with the huge amount of sequenced genomes.



**Figure 1:** Visualization of two subgraphs of the whole pangenomes, one constructed from 58 genomes of *Y. pestis* and one from 58 genomes of *E. coli*. The visible difference in the amount of variations can be further characterized mathematically, classifying species with an open or a closed pangenome on the basis of their genome sequences and their respective graph, without the necessity of annotating them.



**References:**

[1] Tettelin et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA*, 2005. DOI: 10.1073/pnas.0506758102.



# Functional genomics of and bioinformatic analysis tools for seed quality parameters in rapeseed

Hanna Marie Schilbert; *Bielefeld University, Bielefeld*

The increasing demand in high quality plant based food products requires breeding of improved crop plants. Rapeseed (*Brassica napus L.*) is one of the most important oil crops worldwide. Beside its high-quality fatty acid balance, also the excellent amino acid composition of its protein is of high nutritional value. However, the presence of anti-nutritional components renders rapeseed protein unusable for human consumption. As part of the BMBF-funded project RaPEQ, our aim is to reduce or even remove these

anti-nutritional components. Therefore, we study the molecular basis of relevant seed quality parameters. To achieve this, (i) dedicated tools for the transfer of functional annotation data are developed, (ii) high-throughput sequencing data is harnessed for e.g. mapping-by-sequencing, and (iii) in-depth characterization of involved key genes have been performed.

The development of dedicated tools facilitates the automatic analyses of the genes and encoded enzymes involved,

and provides predictions for their functionalities. All tools created in this project will be made freely available on github, e.g. KIPEs (Knowledge-Based Identification of Pathway Enzymes) (Figure 1)[1].

The results of (i) were incorporated into (ii), namely the analysis of large genomic and transcriptomic data sets to identify loci and genes associated with seed oil, seed protein-, and antinutrients content via mapping-by-sequencing (MBS). MBS combines bulk-segregate-anal-





ysis and next-generation-sequencing, which enables the identification of causal mutations associated with a phenotype of interest. For this task, the development and application of automatic scripts (written in python) is necessary. Candidate genes and sequence variants associated with each seed quality trait were identified. Consequently, causal sequence variants have been validated to be associated with seed glucosinolate content using the BnASSYST diversity panel.

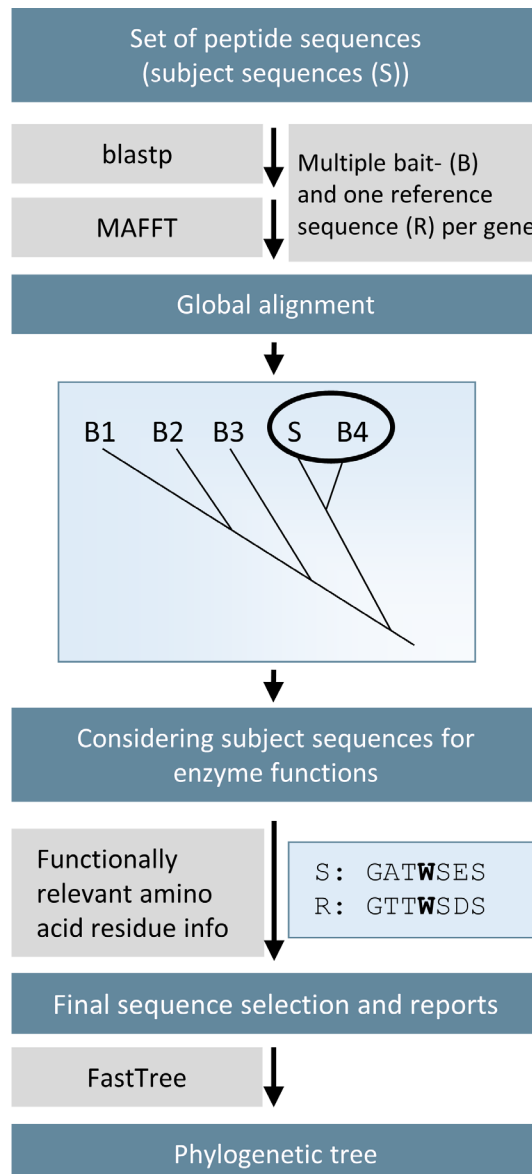
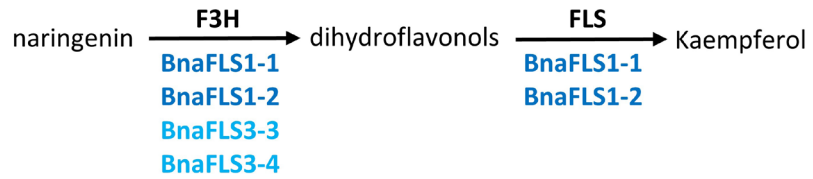


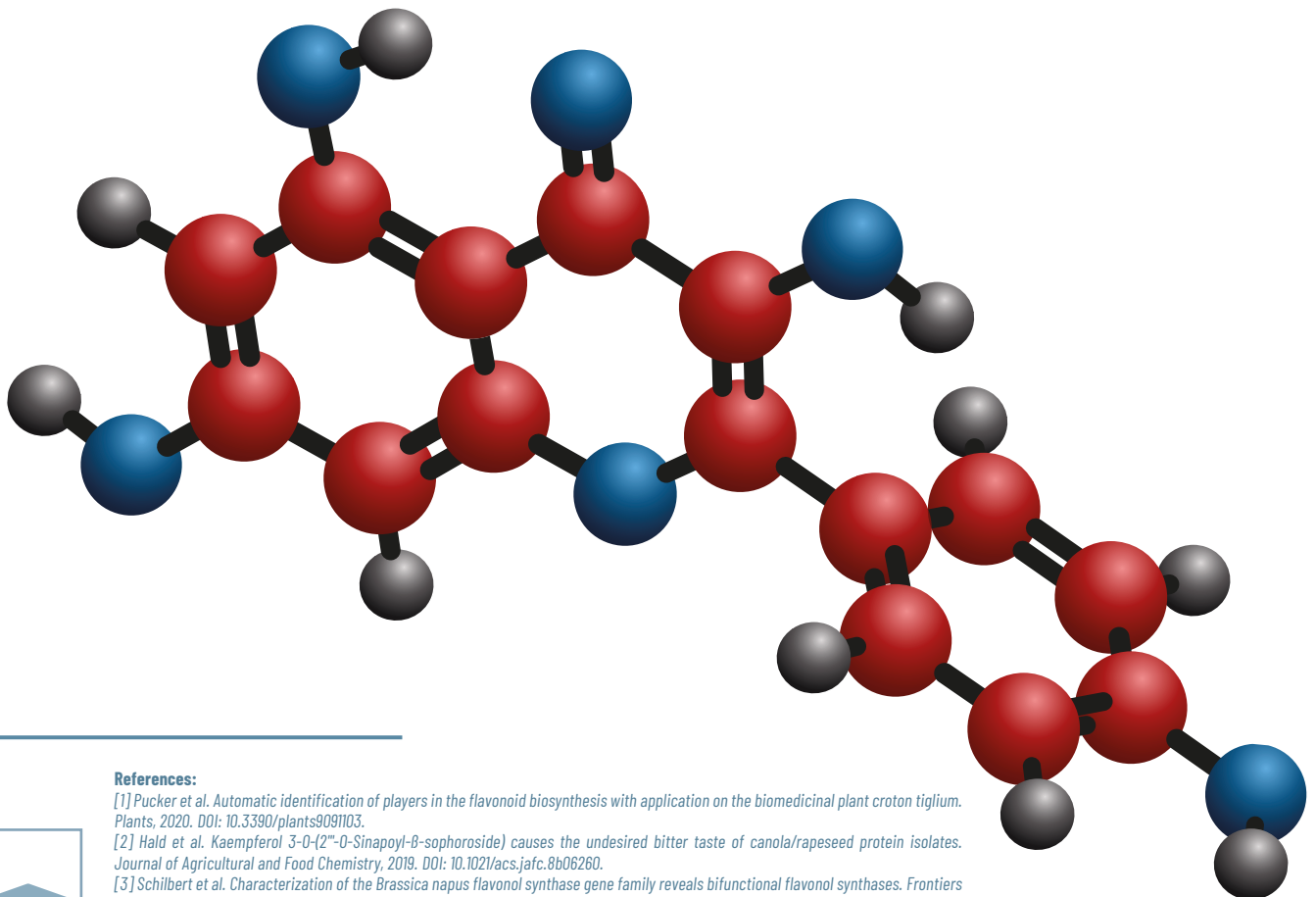
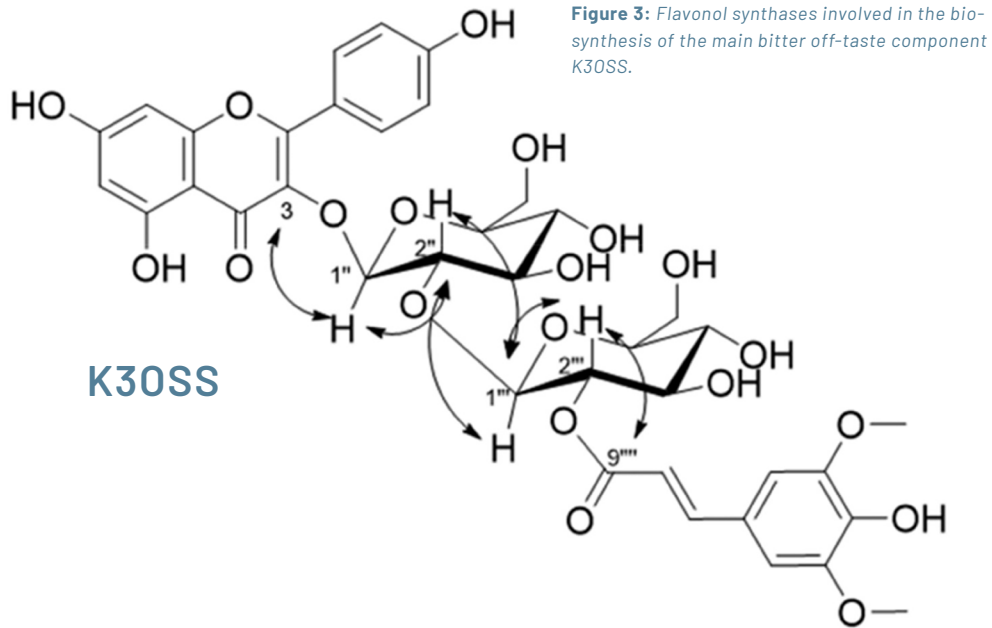
Figure 1: Workflow of KIPES.



Recently, the main bitter off-taste component in rapeseed protein isolates was identified: Kaempferol 3-O-(2"-O-Sinapoyl- $\beta$ -sophoroside) (K3OSS) [2]. The key enzyme of flavonol biosynthesis is flavonol synthase (FLS), which converts dihydroflavonols to flavonols. We identified the members of the FLS gene family via KIPEs (see (i)) based on the genome sequence of the polyploid *B. napus* cultivar Express 617 [4]. In order to identify which FLS genes contribute to flavonol production in seeds, we analyzed several transcriptome data sets (Figure 2). The FLS gene family revealed organ- and development-specific gene expression. In seeds, five FLS genes are expressed: BnaFLS1-1, BnaFLS1-2, BnaFLS2-1, BnaFLS3-3, and BnaFLS3-4. The corresponding gene products were functionally characterized (Figure 3) [3]. Our results provide novel insights into the molecular basis of seed protein, oil, glucosinolate and flavonol biosynthesis in *B. napus*, which can be used for targeted engineering and breeding to support the use of rapeseed protein in human consumption.

	<i>BnaFLS1-1</i>	<i>BnaFLS1-2</i>	<i>BnaFLS2-1</i>	<i>BnaFLS2-2</i>	<i>BnaFLS3-1</i>	<i>BnaFLS3-2</i>	<i>BnaFLS3-3</i>	<i>BnaFLS3-4</i>	<i>BnaFLS3-5</i>	<i>BnaFLS4-1</i>	<i>BnaFLS4-2</i>	<i>BnaFLS4-3</i>	<i>BnaFLS4-4</i>
<b>anther flowering (n=4)</b>	148	131	0	0	3	0	13	4	0	0	0	4	1
<b>stamen (n=1)</b>	6	7	0	0	0	12	12	0	0	0	0	0	0
<b>ovule (n=1)</b>	7	1	1	0	0	0	53	3	0	0	1	1	0
<b>pistil (n=3)</b>	20	33	1	1	0	2	16	0	0	0	0	1	0
<b>sepal (n=1)</b>	12	10	0	0	0	0	5	0	0	0	3	8	5
<b>petal (n=2)</b>	132	150	0	0	0	1	8	2	0	0	0	0	0
<b>seed 23DAF (n=3)</b>	15	4	1	0	0	0	55	76	0	0	0	1	1
<b>seed 35DAF (n=3)</b>	59	29	0	0	0	0	14	29	0	0	0	0	0
<b>seed coat 14DAF (n=7)</b>	14	20	4	0	0	0	82	8	0	1	0	1	0
<b>seed coat 21DAF (n=6)</b>	39	32	12	0	0	0	53	165	1	0	0	3	1
<b>seed coat 28DAF (n=6)</b>	25	20	10	0	0	0	11	146	0	1	0	2	1
<b>seed coat 35DAF (n=6)</b>	15	10	11	0	0	0	14	210	0	3	0	3	1
<b>seed coat 42DAF (n=6)</b>	10	3	16	0	1	0	24	134	0	4	0	2	1
<b>embryo (n=6)</b>	7	60	0	0	0	0	12	2	0	0	0	0	0
<b>endosperm (n=8)</b>	7	3	1	0	0	1	8	66	0	3	2	0	0
<b>seedling (n=9)</b>	5	6	1	0	1	9	25	1	0	0	1	11	2
<b>leaf 35DAF (n=3)</b>	15	9	0	0	0	0	38	1	0	0	1	0	0
<b>stem (n=19)</b>	9	10	4	4	1	1	40	7	0	0	0	6	2
<b>shoot (n=2)</b>	6	10	0	0	0	0	23	0	0	0	0	0	0
<b>root 30DAP (n=20)</b>	0	0	7	5	7	38	22	8	4	1	1	34	31
<b>root 60DAP (n=2)</b>	0	0	21	10	22	50	107	45	19	1	1	90	59

**Figure 2:** Transcriptomic analysis revealed the major FLS genes expressed in *B. napus* seeds.



**References:**

- [1] Pucker et al. Automatic identification of players in the flavonoid biosynthesis with application on the biomedical plant croton tiglium. *Plants*, 2020. DOI: 10.3390/plants9091103.
- [2] Hald et al. Kaempferol 3-O-(2''-O-Sinapoyl-beta-sophoroside) causes the undesired bitter taste of canola/rapeseed protein isolates. *Journal of Agricultural and Food Chemistry*, 2019. DOI: 10.1021/acs.jafc.8b06260.
- [3] Schilbert et al. Characterization of the Brassica napus flavonol synthase gene family reveals bifunctional flavonol synthases. *Frontiers in Plant Science* 2021. DOI: 10.3389/fpls.2021.733762.
- [4] Lee et al. Chromosome-scale assembly of winter oilseed rape Brassica napus. *Frontiers in Plant Science*, 2020. DOI: 10.3389/fpls.2020.00496.





# Algorithms for graph-based computational pangenomics

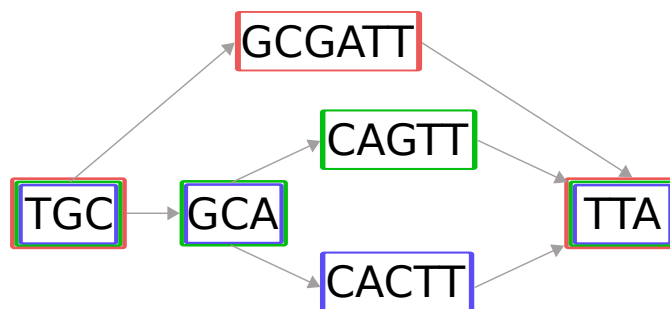
Tizian Schulz; *Bielefeld University, Bielefeld*

Ongoing technological improvements during the last decades have made DNA sequencing become both affordable and fast. Nowadays, sequencing of DNA isolates is a standard approach even in small laboratories, and genetic information is decoded from DNA molecules at a speed comparable to computation times needed for subsequent downstream analyses. Consequently, more and more individual genome sequences are getting abundant for many species across all parts of the tree of life.

The massive increase of genomic data means a great chance for science. New large-scale data sets may allow us to gain many new biological insights. However, it also puts new challenges on algorithms in computational biology. Efficient storage concepts are vital to keep large data sets processible on state-of-the-art hardware. At the same time, processing times need to stay reasonable. Traditional methods often fail to fulfill these demands since they have not been developed for such large-scale data.

The pangenomic approach is one way to handle large data sets while facing the above-described challenges. A pangenome is defined as the set of all genomic information of a species. It can be stored efficiently by using graphical data structures.

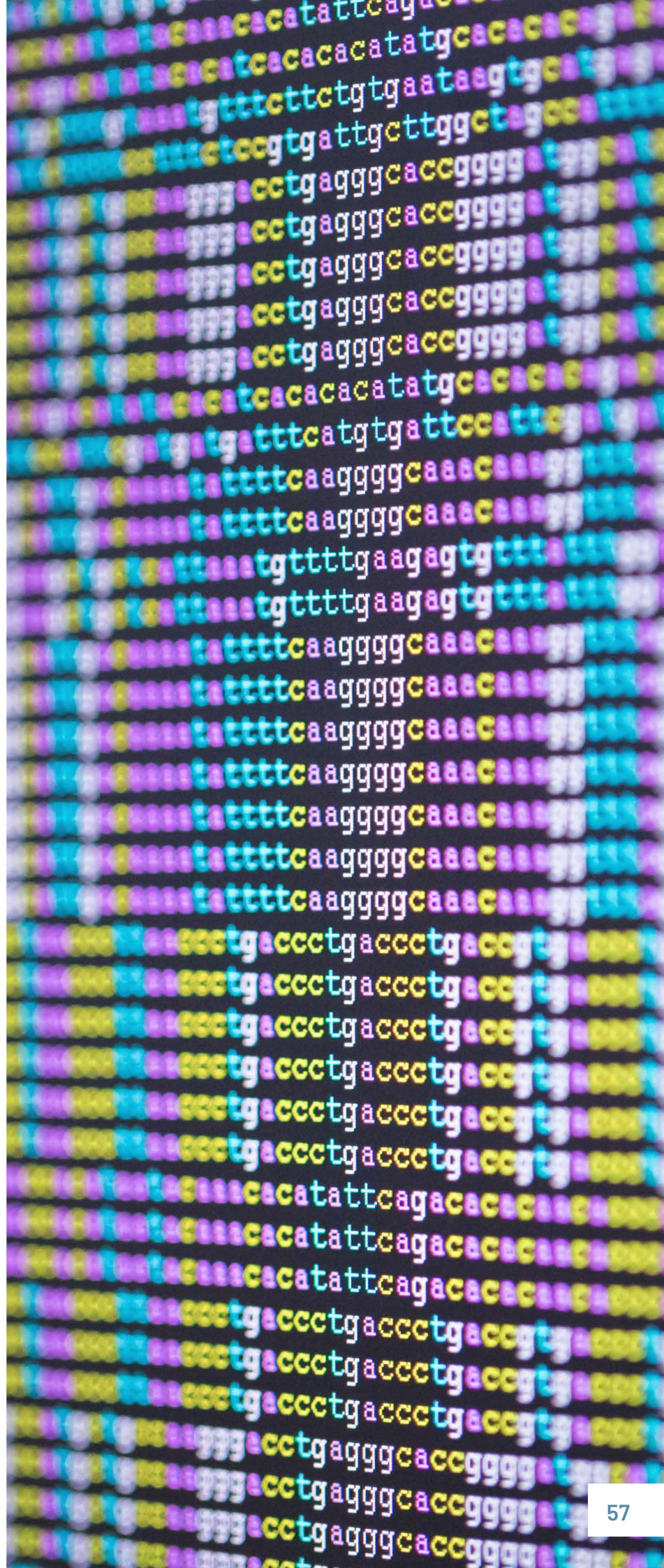
Genome A: TGCACCTTA  
Genome B: TGCAGTTA  
Genome C: TGCGATTA



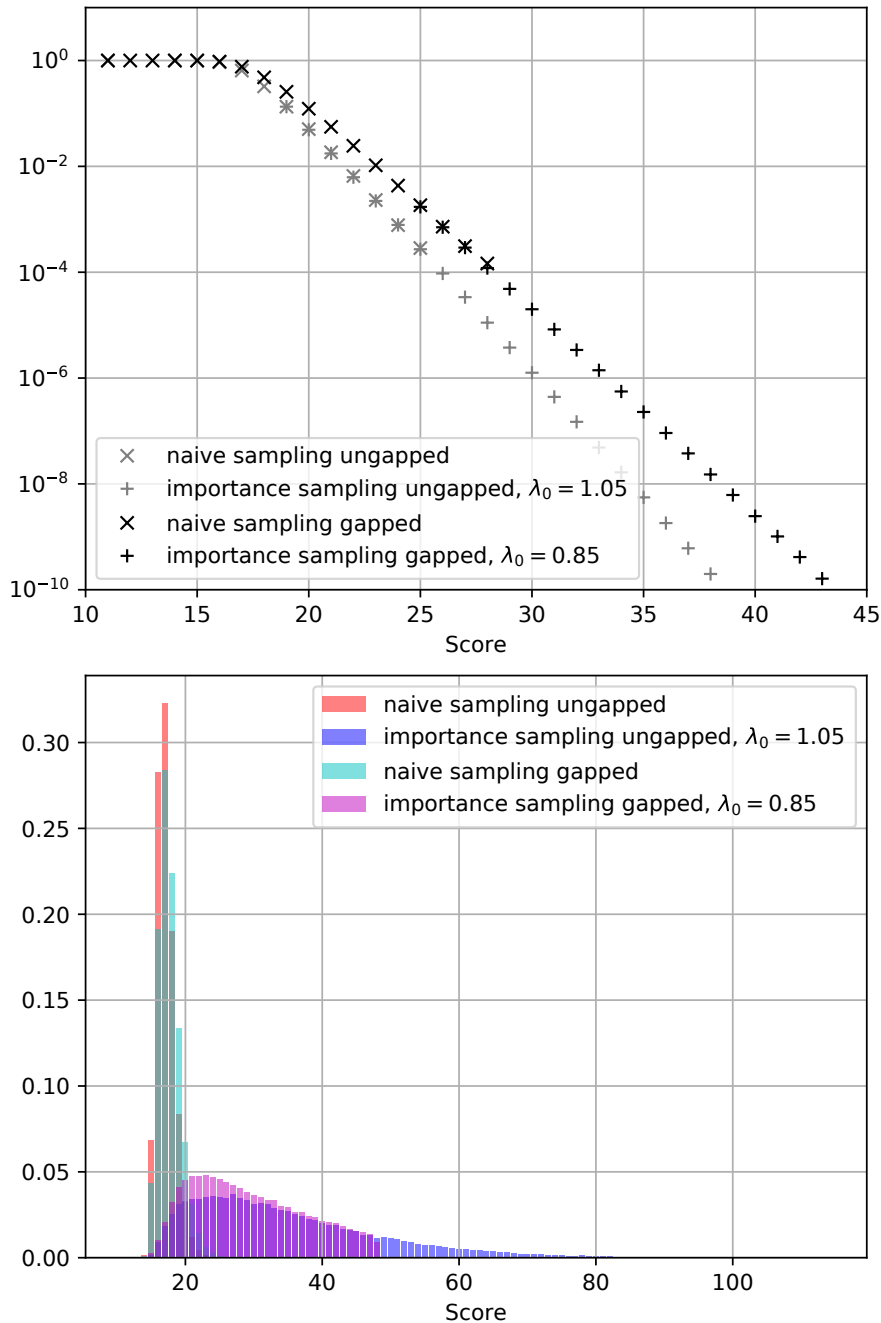
**Figure 1:** Redundant sequence information that appears in several members of a pangenome is stored only once in a sequence graph which reduces memory requirements and facilitates sequence comparisons.

A sequence graph takes perfect advantage of the high amounts of sequence redundancy inside a pangenome of closely related organisms, because sequence parts shared by several individuals are stored in the graph only once (Figure 1). Furthermore, they can facilitate sequence comparisons during analysis. Only looking at the graph's topology already reveals the degree of genetic diversity within the pangenome and allows to distinguish areas of strong sequence conservation from those of high variability. Some graph data structures even allow to omit certain computationally expensive preprocessing steps such as genome assembly that are usually needed after DNA sequencing.

Unfortunately, the processing of a sequence graph is often more complex than the processing of plain sequences, and a modification of algorithms to work on graphs is often not possible in a straightforward way. Therefore, at BIBI and in collaboration with several other groups we have developed algorithms for an efficient analysis of graphical pangenomes.







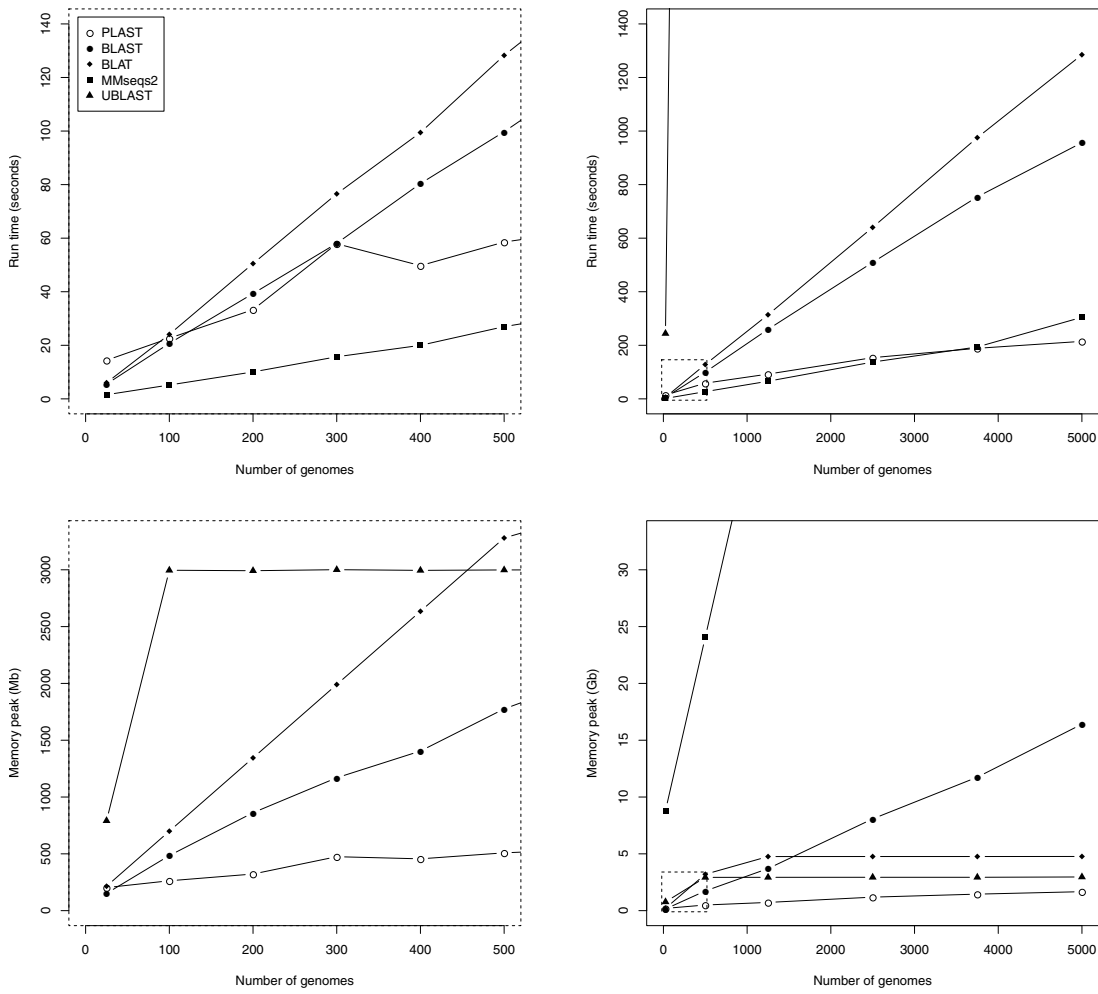
**Figure 2:** Alignment statistics for a pangenome graph differ compared to conventional alignment statistics, because the central assumption of unrelatdness between sequences is violated. Our simulations could show an affine, linear dependency for larger score values (top). A priority sampling method allowed us to gain sufficiently many samples from the distribution's rare event tail with only a moderate sampling in total. Results are shown as histograms (bottom).

The first problem studied focuses on the detection of sequence homology. This is one of the most basic tasks in DNA sequence analysis. Being confronted with a new DNA sequence of unknown biological function, a common way to estimate this function is to search for homologous sequences whose function is already known. Classically, this is done by scanning large sequence databases, calculating alignments between database sequences and the query to measure their similarity, and estimating homology based on an alignment statistic using the gained alignment score. As alignment calculation for large amounts of sequences is costly, heuristic approaches are usually preferred over exact methods.

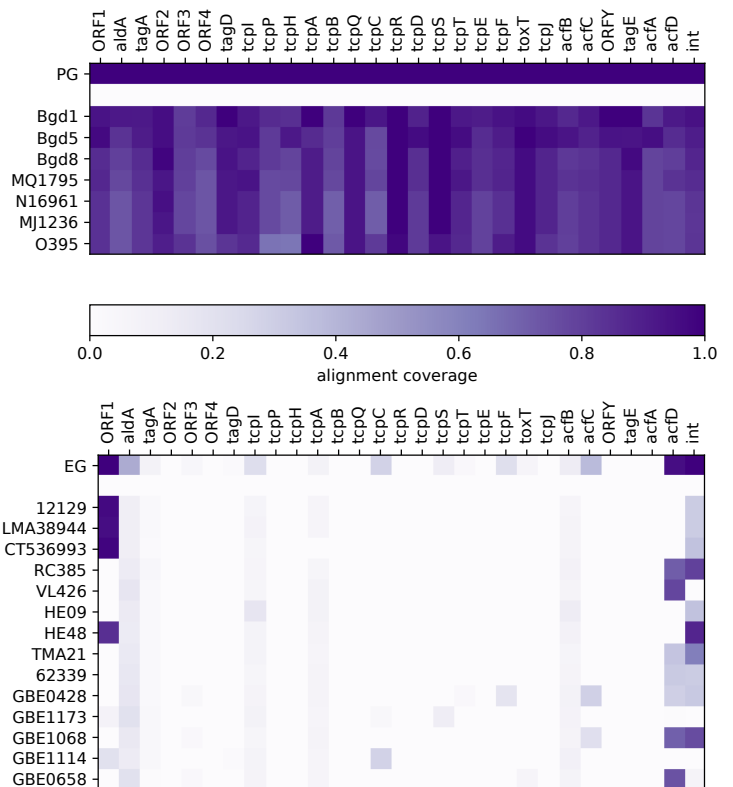
The most popular homology detection method BLAST [1] gains speed by the underlying assumption that in a large database of many unrelated sequences only a small number of sequences matches well to the query and needs to be compared thoroughly while most sequences can be discarded quickly without missing any meaningful results. However, this assumption fails when considering a pangenome where all sequences are closely related. As a consequence, BLAST's run time increases linearly with the number of genomes in a pangenome.

In order to cope with this challenge, we introduced PLAST [2], a method to detect sequence homology between a DNA query sequence and a pangenome represented as a sequence graph. Unlike BLAST, our algorithm makes use of the fact that shared sequence parts are collapsed inside the graph and can calculate alignments for several genomes simultaneously.

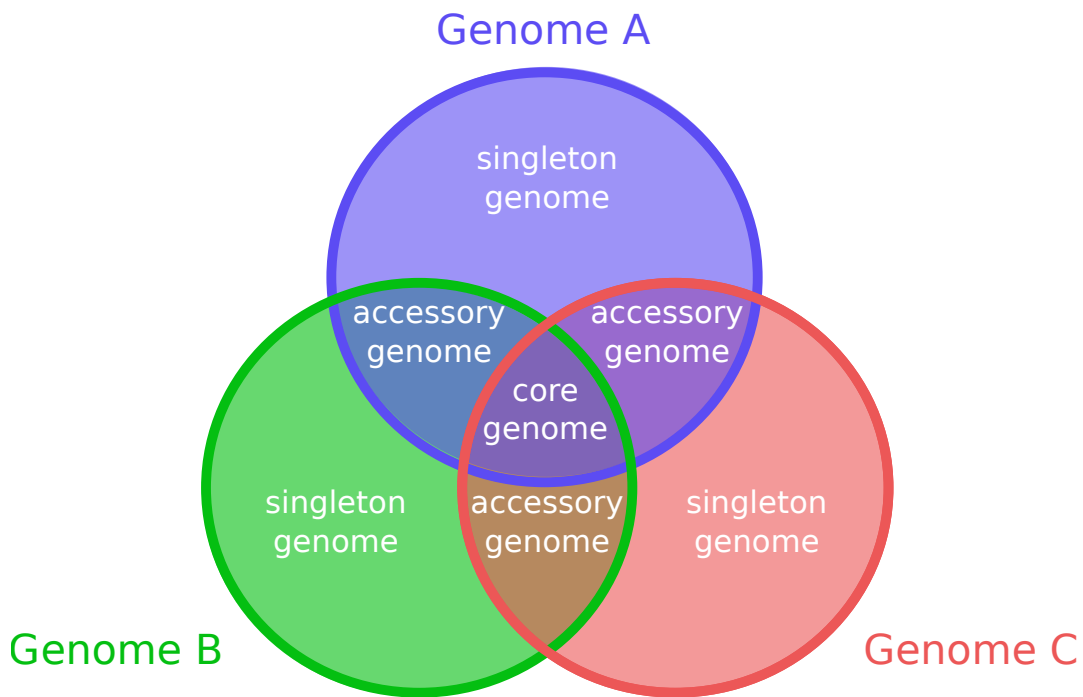




**Figure 3:** Comparisons of our method to classical homology detection tools on pangenomes of different sizes show superior behavior in terms of speed and memory consumption.



**Figure 4:** In a classical use case, PLAST was taken to search for virulence causing genes in a pangenome of *Vibrio cholerae* strains. Viral traits could clearly be found in strains isolated from hospital samples (left) while they were widely absent in strains from environmental samples (right).



**Figure 5:** Due to the high degree of sequence similarity, pangenomes can be partitioned into parts shared by all (core genome) and only several (accessory genome) members. Each member also has some unique parts (singleton genome) usually.

The fact that PLAST operates on closely related sequences within a pangenome made it necessary to introduce an alignment statistic for sequence-to-graph alignments on pangenome graphs – a use case that has never been studied before. Our simulations show that such an alignment statistic follows the same basic rules as a conventional alignment statistic. However, its exact statistical parameters are additionally influenced by the degree of biological diversity inside the pangenome (Figure 2). Comparisons of our tool to BLAST and other state-of-the-art homology detection tools show a superior behavior of PLAST in terms of run time and memory usage (Figure 3). It could also convince in several use case scenarios (see e.g, Figure 4).

The second project focuses on the determination of a pangenomic core. The

core of a pangenome is defined as the set of genetic material that is shared between all members of the pangenome (Figure 5).

Pangenomic core detection is a common and widely used method which has many different applications. Among others, it can be used for studying genetic diversity, has relevance for drug development or vaccine design and is applied in crop plant breeding.

Traditionally, a core genome is determined at the gene level. However, this approach has several shortcomings. Firstly, it does not allow to detect any core features below the genes. Secondly, it is dependent on multiple preprocessing steps like genome assembly and gene annotation. Both can be highly error-prone, which leads to biases and has direct influence on the quality of a

subsequent core detection. Furthermore, gene-based core detection involves many alignment calculations that are resource demanding and time consuming.

In order to extend core detection beyond the gene level and to avoid additional error sources during preprocessing, we introduced a method that works directly on sequences instead of whole genes. It makes use of a sequence graph to allow a fast processing of large data sets. The graph makes it also possible to deal with small variations that appear between different individuals on the sequence level.

Comparisons to traditional methods show that our approach is faster for large data sets and leads to similar results.



**References:**

- [1] Altschul et al. Basic local alignment search tool. *J. Mol. Biol.*, 1990. DOI: 10.1016/S0022-2836(05)80360-2.
- [2] Schulz et al. Detecting high scoring local alignments in pangenome graphs. *Bioinformatics*, 2021. DOI: 10.1093/bioinformatics/btab077.





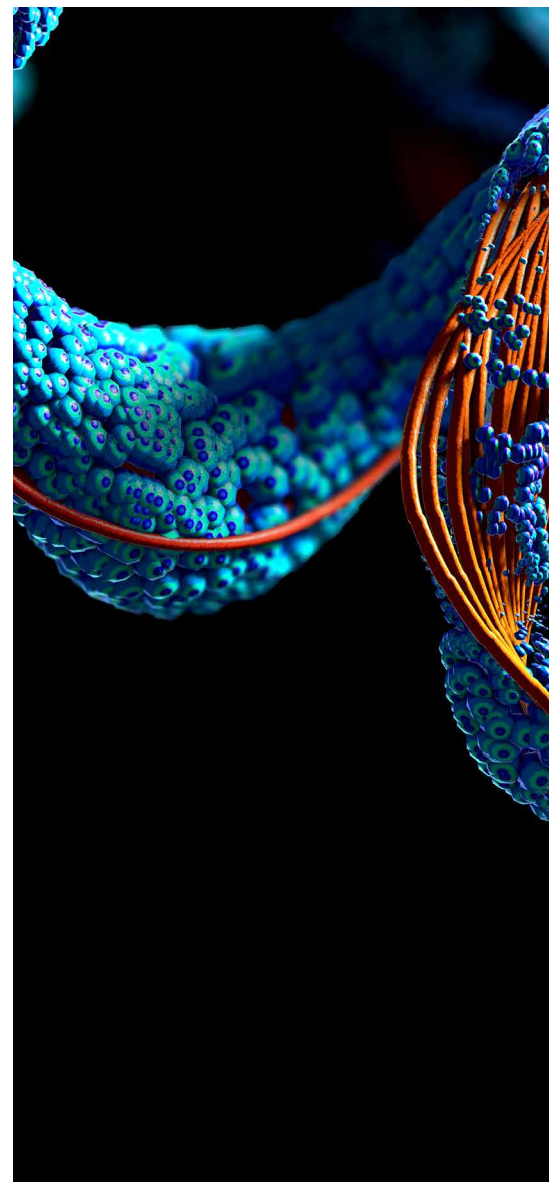
# Protein-DNA binding specificity is facilitated by DNA shape

Janik Sielemann; *Bielefeld University, Bielefeld*

The composition of available proteins for each organism is encoded in its DNA. The regions that encode for the respective proteins are called genes. Gene expression denotes the procedure of an organism to use those genes to eventually build proteins and other gene products. An essential part of this procedure is the promoter sequence, which is located upstream from the protein coding regions and is bound by regulatory proteins. Predicting the expression of a gene from its promoter sequence is one of the key goals of transcriptomics research. The prediction of gene expression will require understanding where exactly regulatory proteins bind genomic DNA. A common approach to visualize those binding locations represents the sequence motif (Figure 1). Even though sequence motifs describe the DNA-binding sites for the

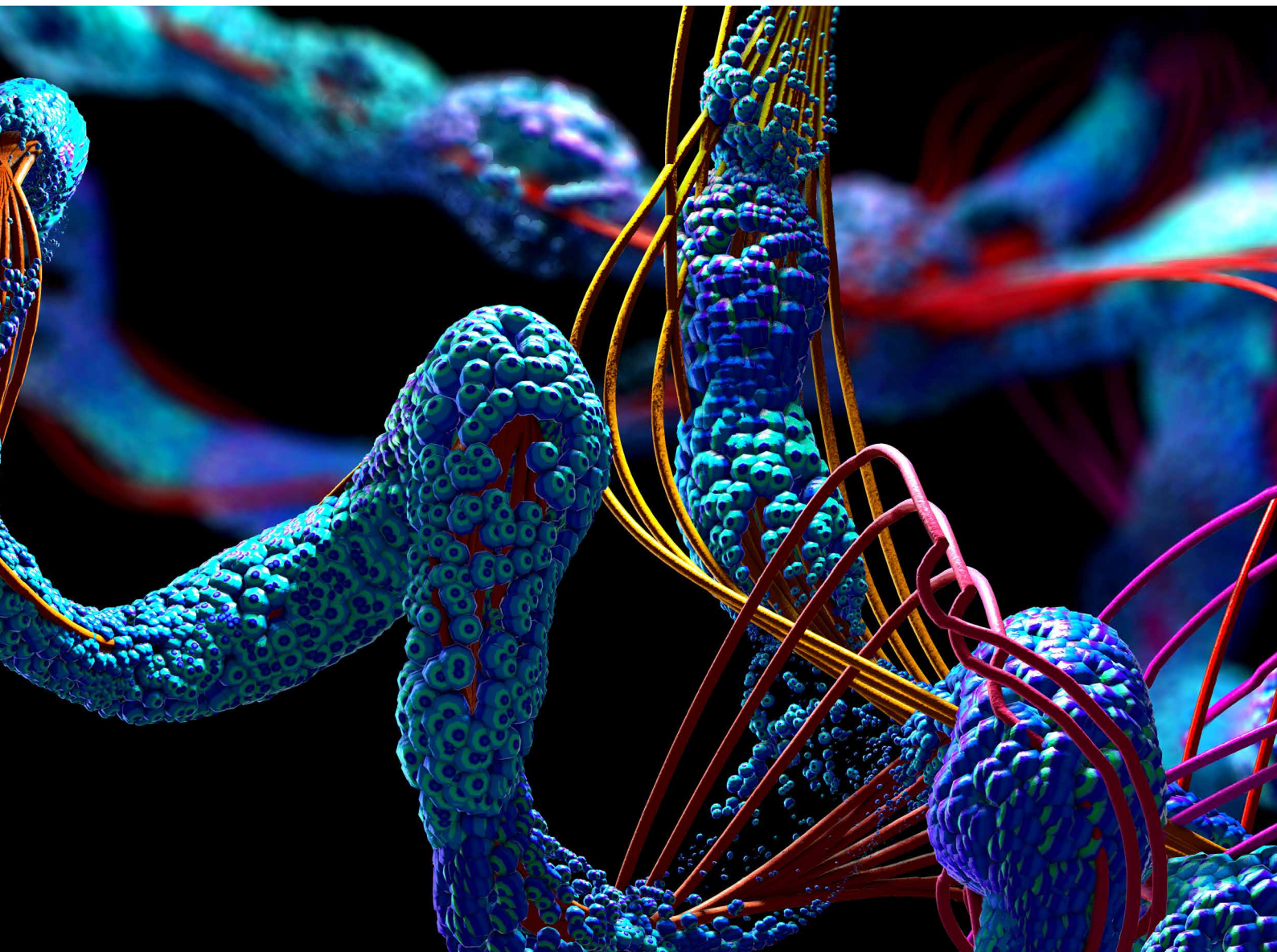
respective protein, they lack predictive power, as they occur more frequently unbound than actually bound by the protein [1].

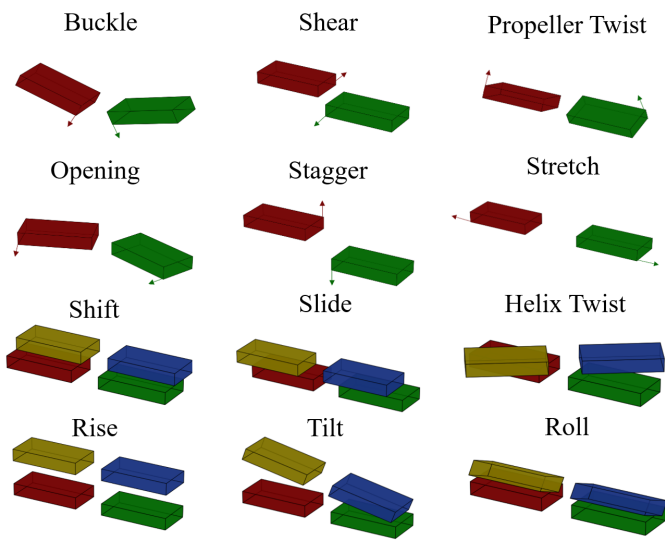
Traditionally, these motifs depict only sequence but neglect DNA shape. DNA is a very constrained molecule since its phosphate sugar backbone runs antiparallel while its bases are paired and arranged in rungs on a helical ladder. However, despite the constraints, the exact position of each base pair and each base in a pair is influenced by its surrounding bases. The pairs can be tilted, shifted, slid, rolled, risen and twisted relative to each other [4] (Figure 2). The bases in a pair can be buckled, sheared, stretched, twisted, opened and staggered [4] (Figure 2). The width of the minor groove is also influenced by the surrounding bases.



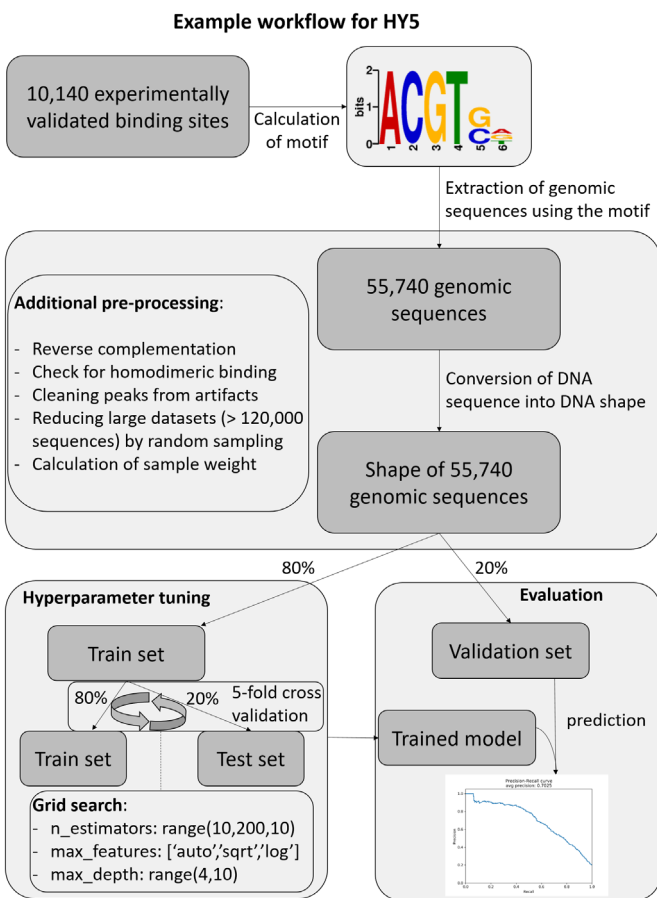


**Figure 1:** Example sequence motif. This sequence motif was generated using experimentally validated DNA binding sites [2] for the protein ANAC070 using MEME-ChIP [3].





**Figure 2:** DNA shape features. A publicly available query table [4] was used to translate DNA sequence into DNA shape features.



**Figure 3:** Example workflow for HY5 in *A. thaliana*. This workflow illustrates the computational steps from publicly available data to trained models capable of predicting Protein-DNA binding affinity.

Since shape may contribute non-linearly and combinationally to binding, machine learning approaches ought to be able to predict binding events using DNA shape features. In the example workflow (Figure 3) a random forest model is built using the DNA shape features of all sequence motif occurrences for the protein HY5 in *Arabidopsis thaliana*.

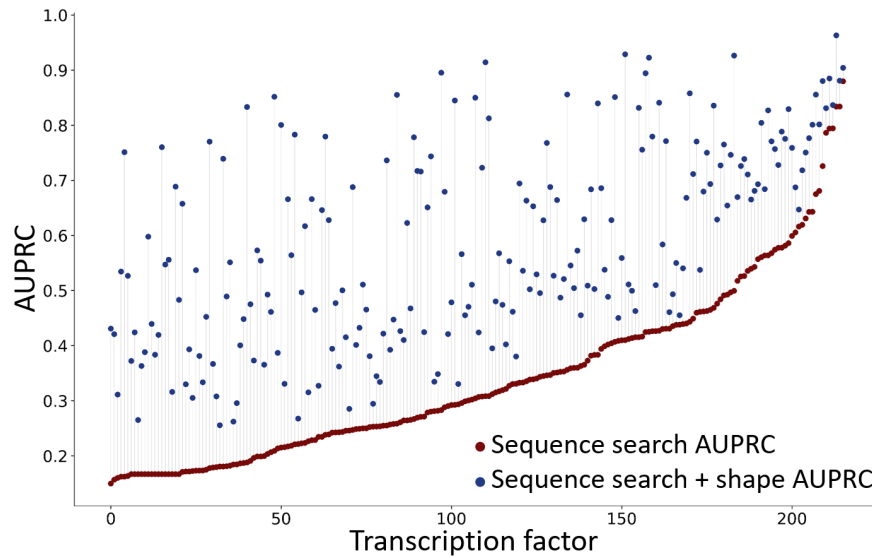
For each of 216 DNA binding proteins, which have publicly available binding data, a model was trained using the workflow above. The performance of those models regarding binding site prediction was compared to a conventional sequence based approach (Figure 4). The area under the precision-recall-curve improved on average by 93.2% using the predictors, which were trained on the DNA shape features in combination with the sequence search.

In addition to evaluating the performance on ground truth data, it was tested whether the models are capa-

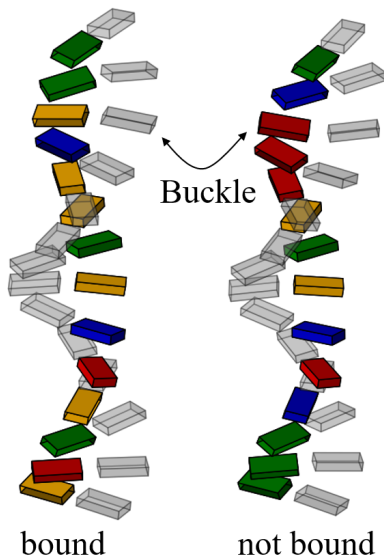


ble of predicting binding affinity to sequences, which are not present in the *A. thaliana* genome. For this, three high scoring and three low scoring random sequences were used for a competition EMSA experiment (Figure 5). Five out of six sequences show the predicted binding behaviour in the experiment, which is within the expected error rate of the model.


In conclusion, the analysis shows that a combination of motif sequence and motif shape enables improved prediction of TF binding on the genomic sequence. This knowledge can now be leveraged to transfer likely binding sites determined in one genome to that of a related species, to better understand evolution of regulation and regulatory motifs, and to build predictive models of gene expression.



**Figure 4:** Comparison of performance regarding binding site prediction. In combination, the sequence search and random forest models improved binding prediction for all tested proteins.



HY5 Protein	-	+	+	+	+	+	+	+
Competitor	-	-	predicted sequences					
Seq. prediction			+	+	+	+	+	+
Shape prediction			-	+	-	+	-	+



**Figure 5:** Competition EMSA experiment. All used sequences contain the binding motif. The model assigned different binding affinities, based on the shape features of the sequence.



**References:**

- [1] Sielemann, J. et al. Local DNA shape is a general principle of transcription factor binding specificity in *Arabidopsis thaliana*. *Nature communications*, 2021. DOI: 10.1038/s41467-021-26819-2.
- [2] O'Malley, R. C. et al. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 2016. DOI: 10.1016/j.cell.2016.04.038.
- [3] Machanick, P. & Bailey, T. L. MEME-CHIP: motif analysis of large DNA datasets. *Bioinformatics*, 2011. DOI: 10.1093/bioinformatics/btr189.
- [4] Li, J. et al. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Research*, 2017. DOI: 10.1093/nar/gkx1145.



# Computational pangenomics in plants

**Katharina Sielemann; *Bielefeld University, Bielefeld***

Understanding the complete genomic diversity of larger taxonomic groups requires the effective comparison of genomes of many species or cultivars, or more pragmatically their genome sequences. While the fast development of long read sequencing technologies enables cost-effective data generation, there is a pressing need to develop tools for the parallel analysis of large (plant) genome sequences for generation of results and creation of knowledge.

As an entry point into comparative genomics, we want to determine and eval-

uate the pangenome (the entire set of sequences, including e.g. genes and structural variations) of a selected taxonomic group. Comparative analysis of high-quality genome sequence assemblies from evolutionary related sources will provide power to an improved identification of genes, pseudogenes, transposable elements and structural genome variation at the high kbp scale.

Differences concerning gene copy number and large-scale structural variations, including insertions and deletions, can be assessed automatically through the

comparison of these assemblies and also by integrating the underlying sequence read data. Comparative genomics between species and against a reference genome sequence will allow the identification of domestication tracks like “variant deserts”. Integration of phylogenetic and phenotypic information will allow the characterization of genomic features that confer unique properties (traits, phenotypes, etc.) to particular species and further the investigation of the relation and ancestry of these species. We develop dedicated tools and automated analyses to answer specific questions in this field.

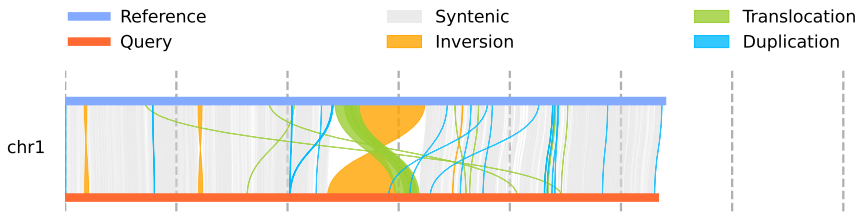


Figure 1: Genomic rearrangements (plotted with SyRI)[1].

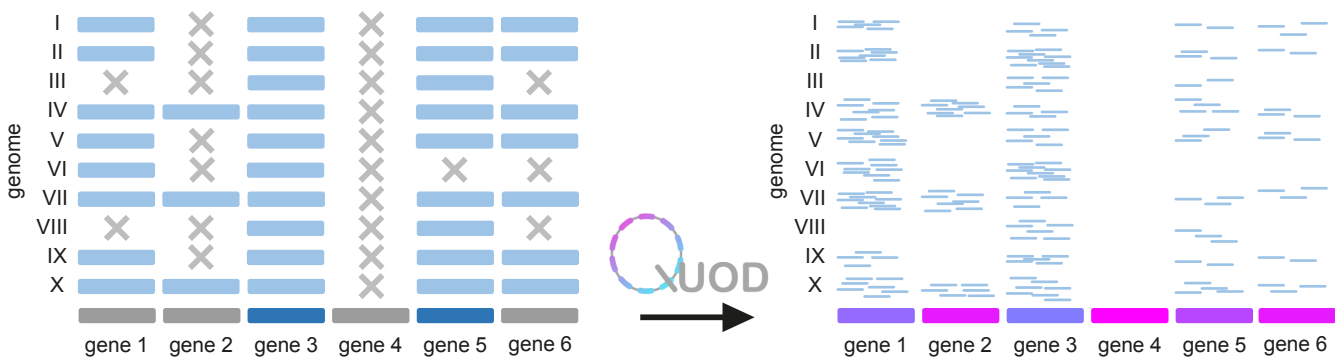


Figure 2: Illustration of the QUOD concept [2].

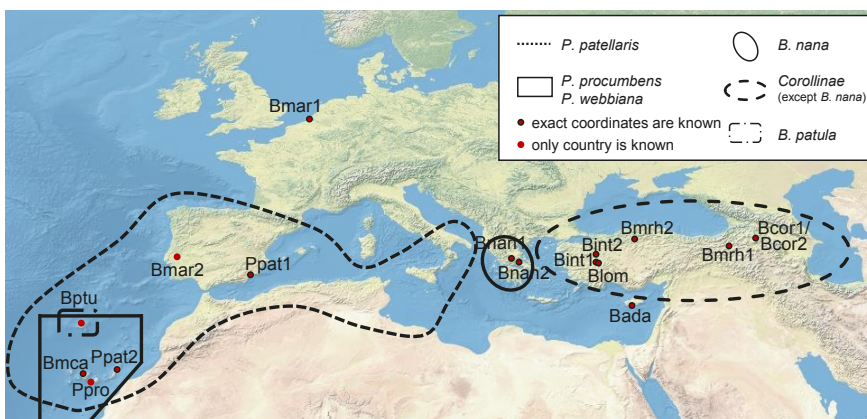


Figure 3: Geographic distribution of the investigated species [3].

As a first part of the project, a method for the quantification of gene dispensability (QUOD) was developed [2]. Dispensability of genes in a phylogenetic lineage, e.g. a species, genus, or higher-level clade, is gaining relevance as most genome sequencing projects move to a pangenome level. Instead of classifying a gene in a binary way as either core (present in all investigated genome sequences) or dispensable (missing in some genome sequences), QUOD assigns a dispensability score to each gene. Hence, QUOD facilitates the identification of candidate dispensable genes which often underlie lineage-specific adaptation to varying environmental conditions.



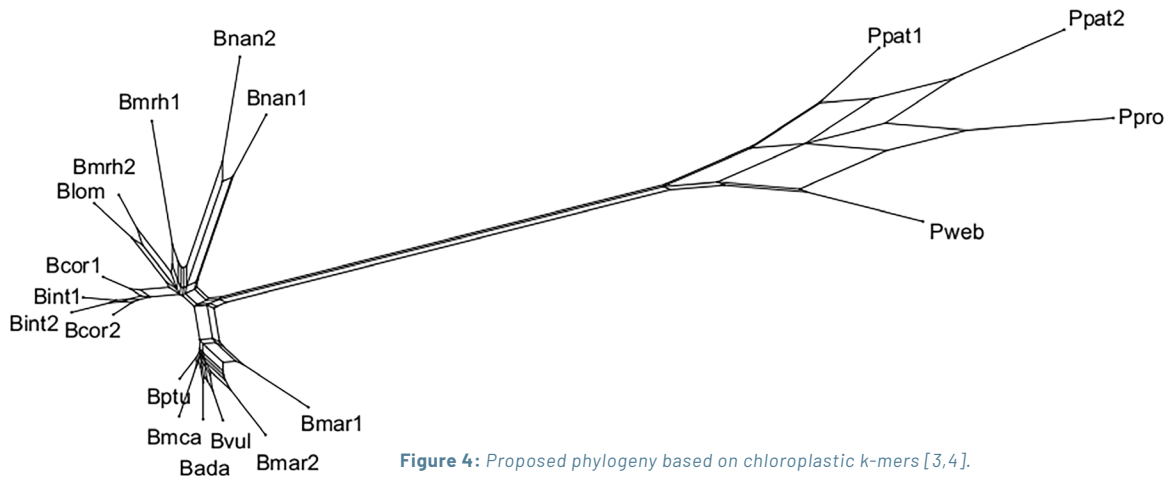


Figure 4: Proposed phylogeny based on chloroplastic k-mers [3,4].

We have selected the Betoideae sub-family including sugar beet (*Beta vulgaris* subsp. *vulgaris*) as evolutionary related target for pangenomic studies. Sugar-producing beets have a high economic value, and crop wild relatives are relevant for breeding. The low genetic diversity within the cultivated beets requires introduction of new traits, for example to increase their tolerance and resistance attributes - traits that often reside in the wild relatives. For this, genetic information of wild beet relatives as well as data on their phylogenetic placements to each other are crucial. To answer this need, in a second part of the PhD project, we sequenced and assembled the complete plastomes sequences from a broad species spectrum of the beet genera *Beta* and *Patellifolia* [3]. This pan-plastome dataset was then used to determine the wild beet phylogeny at high-resolution.

In conclusion, our wild beet plastomes present a new resource to understand the molecular base of the beet germplasm.

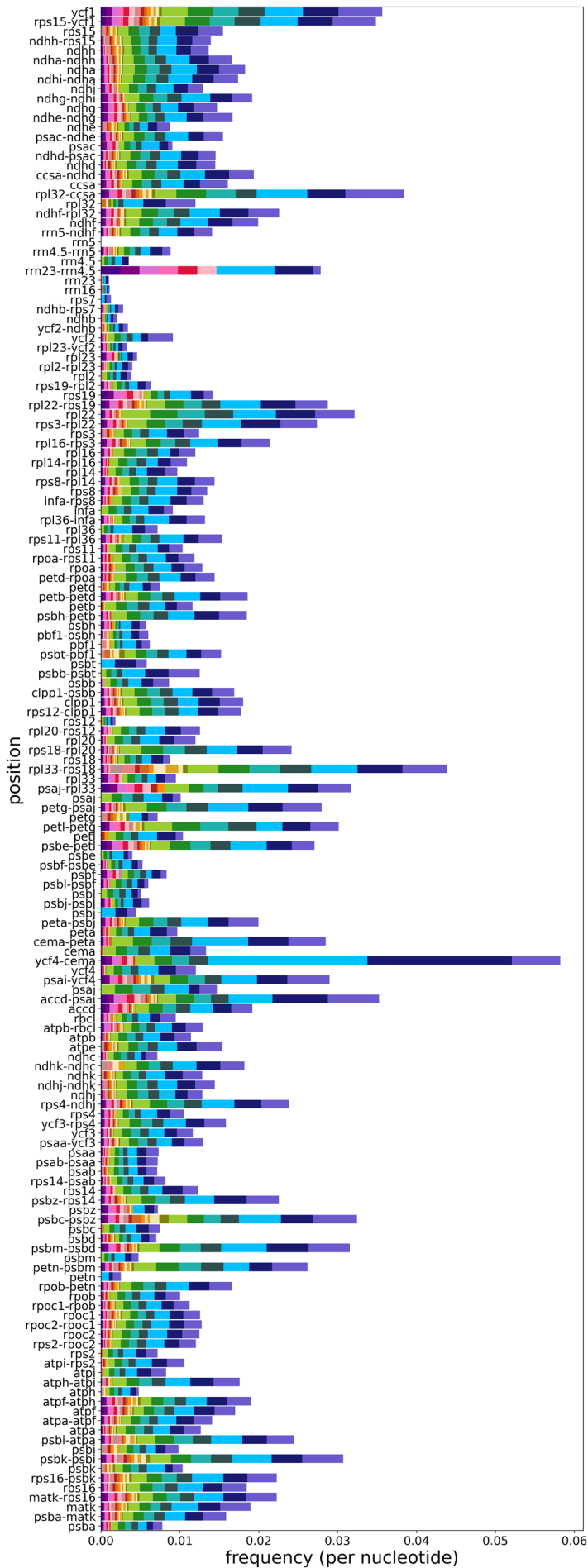


Figure 5: SNP hotspots throughout the plastome assemblies [3].

References:

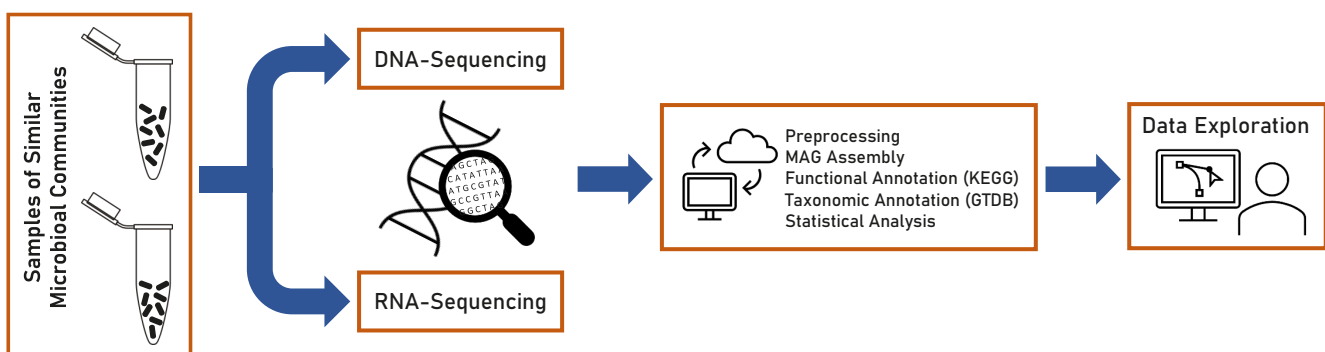
- [1] Goel et al. SyRl: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 2019. DOI: 10.1186/s13059-019-1911-0.
- [2] Sielemann et al. Reference-based quantification of gene dispensability (QUOD). *Plant Methods*, 2021. DOI: 10.1186/s13007-021-00718-5.
- [3] Sielemann et al. Complete pan-plastome sequences enable high resolution phylogenetic classification of sugar beet and closely related crop wild relatives. *bioRxiv*, 2021. DOI: 10.1101/2021.10.08.463637.
- [4] Rempel and Wittler. SANS serif: alignment-free, whole-genome-based phylogenetic reconstruction. *Bioinformatics*, 2021. DOI: 10.1093/bioinformatics/btab444.



# Exploring

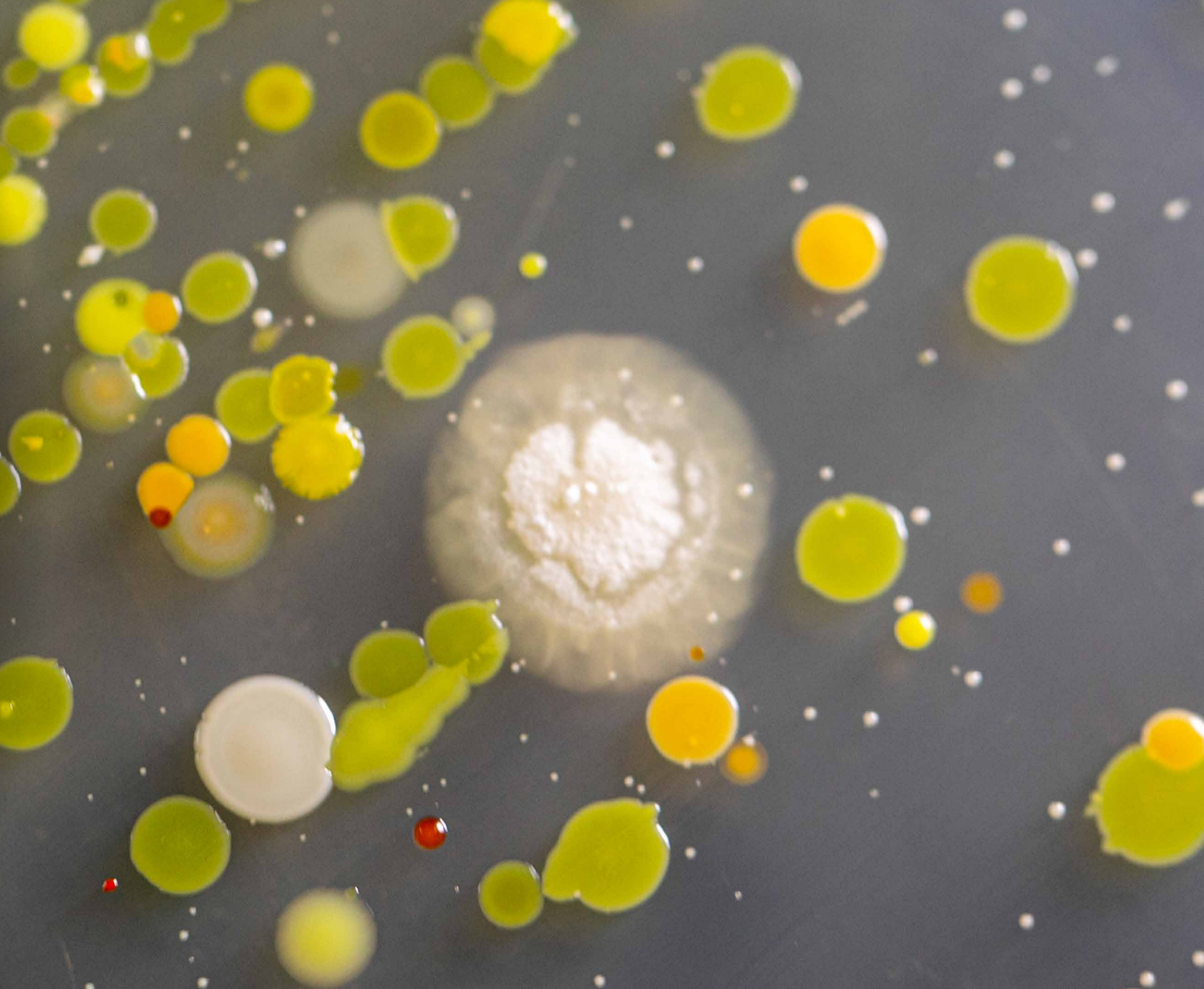
## *comparative metagenomic and metatranscriptomic datasets*

Tom Jonas Tubbesing; *Bielefeld University, Bielefeld*



**Figure 1:** Multiple microbial communities are sampled and genomes as well as mRNA are probed in sequencing experiments. Automated data processing takes place in a cloud environment before a researcher interprets the results of the experiment.



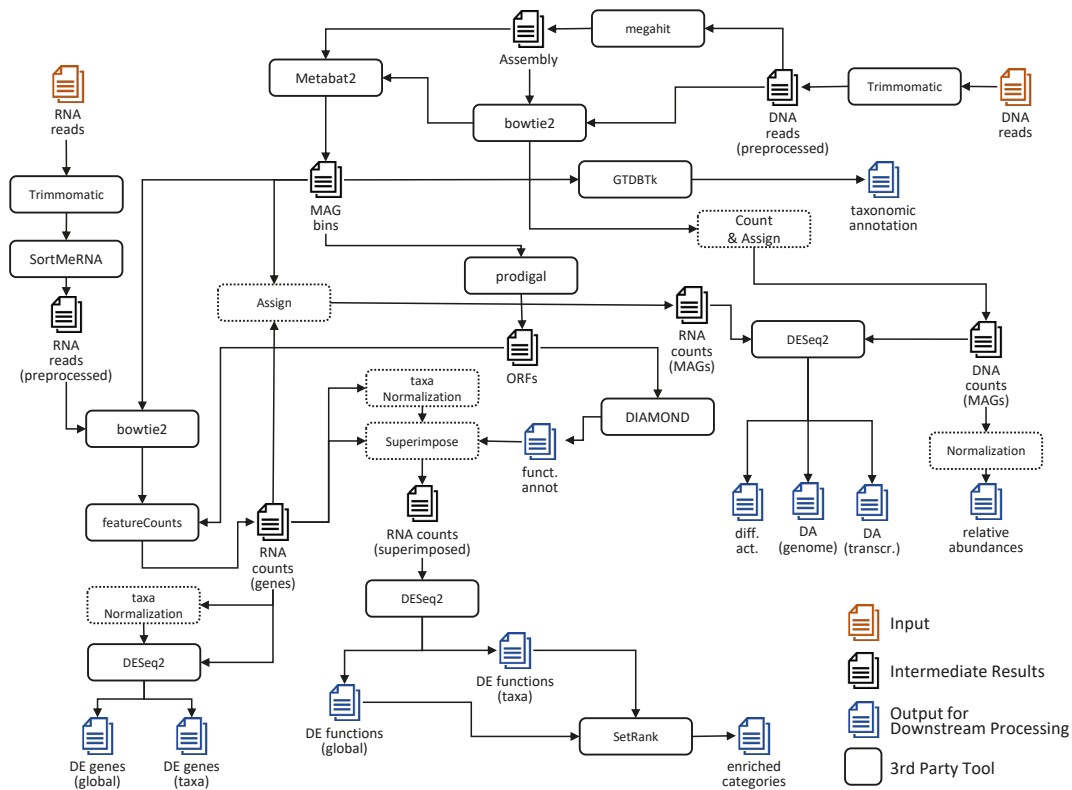


Modern sequencing technologies enable in-depth analyses not only of isolated species, but also of environmental samples containing diverse microbial communities. Learning about these communities is relevant to many different fields like crop cultivation, bio-fuel production and human medicine. The data from such a metagenome sequencing experiment can be used to reconstruct the genomes of bacteria and archaea in the sample. Genomes generated in this way are referred to as metagenome assembled genomes (MAGs). From these, one can learn which species are present in the sample, which functionalities are encoded on their genomes and how abundant

each species in the sample is relative to the others. Sequencing the mRNA of such a sample provides an additional layer of information, and such metatranscriptome data can be ascribed to the various MAGs to gain insight into the transcriptional activities of different species. Comparative experiments can be carried out, where similar microbial communities are sampled from different environments to learn how microbiomes react and adapt (Figure 1).

Processing data from these kinds of sequencing experiments, however, involves the use of many different bioinformatics tools in sequence (Figure 2). Carrying out the necessary steps re-

quires bioinformatics expertise and a significant time investment, which is why it is desirable to automate such tasks. Furthermore, some software tools are too demanding to run on common computer hardware and might still take days to finish on high powered workstations. Thus, this project is concerned with the development of workflows to facilitate an automated data analysis using the resources of the de.NBI cloud [1]. A user has to care only about initial input to and final output of the workflow while a workflow engine takes care of the rest. Containerization of software makes workflows portable and ensures reproducibility.



**Figure 2:** Diagram of a workflow intended for differential gene expression analysis based on a combined metagenomics and -transcriptomics experiment.

If samples from different environmental conditions are to be compared, statistical analyses can be carried out to infer which transcript- or gene-categories are differentially abundant between the two. In a microbiome, two factors drive the change in transcript abundance: Firstly, the strength of gene expression affects how much of a certain transcript is found. Secondly, the relative abundance of the species carrying the gene has a strong influence on transcript counts and can often mask variations in the expression patterns. For this reason, differential expression analysis of microbiomes can answer two different questions depending on the read count normalization approach

that is employed [2]. The first approach elucidates how the abundances of transcripts change between the samples, regardless of what causes the differences. A second approach can be used to measure actual differential expression and thus learn how individual cells respond. In order to learn as much as possible from the samples, we aim to carry out both methods whenever possible. To facilitate interpretation of the results, data pertaining to individual genes and transcripts are aggregated at the level of enzymatic functions.

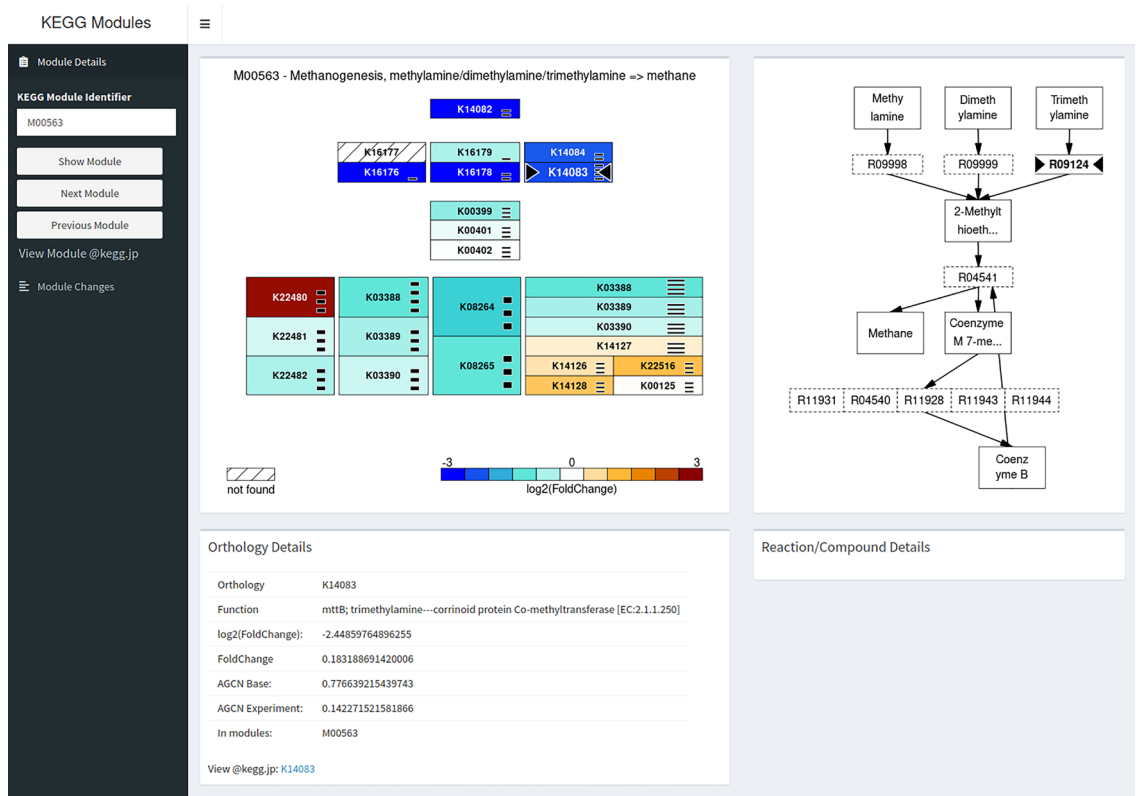
Automated data processing thus yields results from a range of different analyses, mainly in the form of large tables:

Genomes of Species accompanied by taxonomic and functional annotations, relative and differential abundances of genes, and analyses of differential abundance as well as differential expression in the metatranscriptome data. Interpreting these results can be cumbersome: To answer a biological question, it most often becomes necessary to relate information from several different output files, each containing tens of thousands of rows of data. To simplify this work, this project also strives to provide tools to aid in data exploration and interpretation. One of these tools processes results of differential abundance analysis, relates these to biochemical processes defined

in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [3] and is able to provide contextual information regarding the data that a user might want to inspect (Figure 3). Such applications may be hosted in the de.NBI Cloud and accessed through a web browser.

Offering automated workflows enables researchers to carry out analyses even with limited bioinformatics knowledge, while saving time and guaranteeing reproducibility. Aggregating results and relating them in an easy-to-use interface streamlines data exploration and

interpretation. Thus, we will simplify the analysis of metagenome and -transcriptome datasets as well as comparisons between microbial communities.



**Figure 3:** Screenshot of the early version of a web application which relates workflow results to biochemical processes, represented by KEGG modules. The interface provides quick access to the different pieces of information necessary for interpreting results.



**References:**

- [1] de.NBI Cloud – Cloud Computing for Life Sciences. <https://cloud.denbi.de>.
- [2] Klingenberg et al. How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ*, 2017. DOI: 10.7717/peerj.3859.
- [3] Kanehisa et al. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021. DOI: 10.1093/nar/gkaa970.



# Tool development

## *for comparative gene regulatory network analysis*

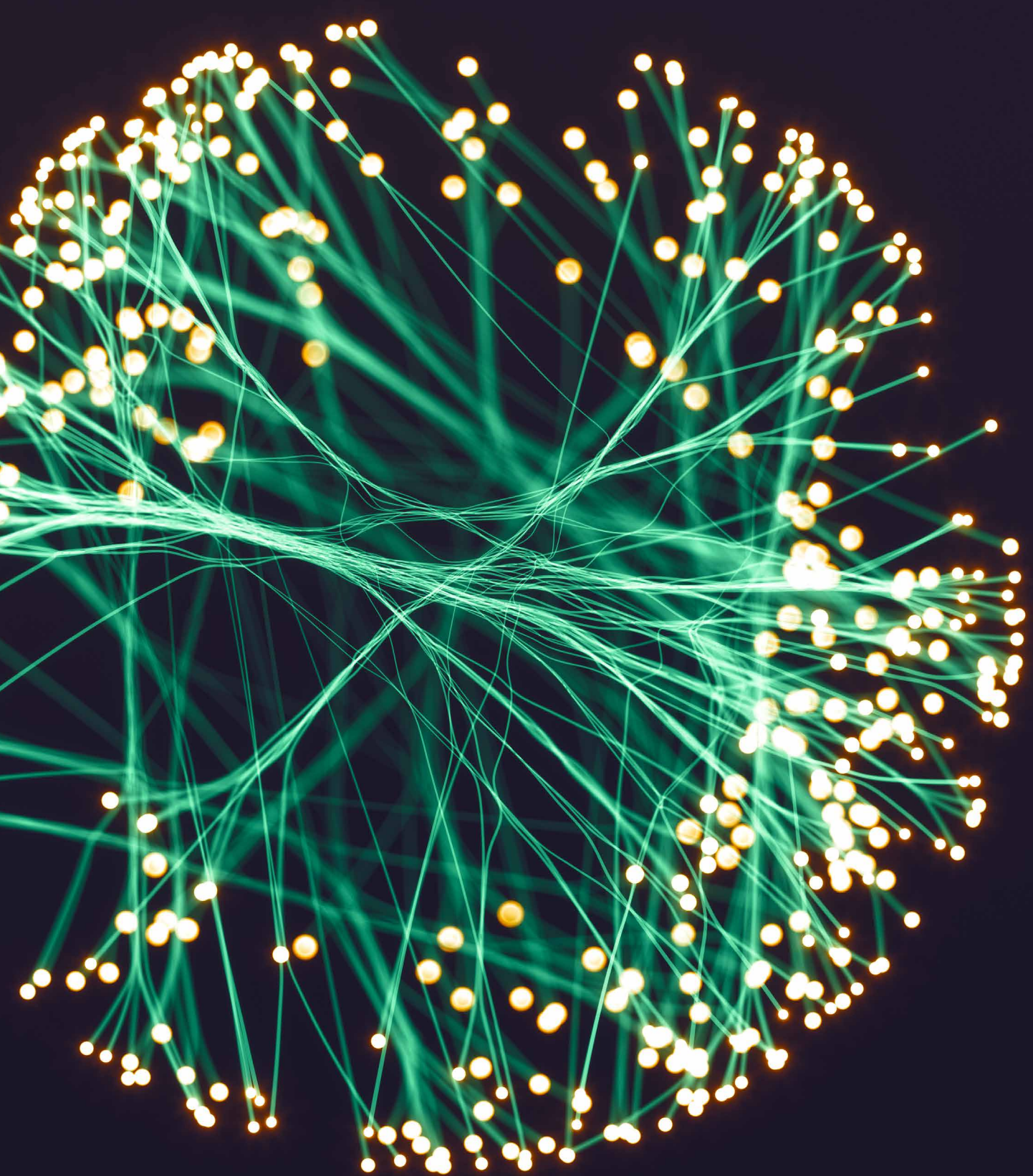
**Donat Wulf; Bielefeld University, Bielefeld**

Plants need to adapt to ever changing environmental conditions. For example: A plant does not get enough water and experiences drought stress. It needs to activate drought genes to better cope with this stressor. These responses are organized in molecular regulatory circuits which consist of transcription factors, histone modifications and DNA methylation. One output from these networks is the transcript abundance. This can be measured by RNA-seq. Methods for gene regulatory network inference from transcript abundance data are available. In the DREAM4 chal-

lenge [1] the objective was to infer the gene regulatory networks from simulated gene expression measurements. In the year afterwards a similar objective in the DREAM5 challenge was set: Infer simulated and in-vivo gene regulatory networks. In the DREAM4 challenge, a training dataset was provided with the solution. The data set was generated with GeneNetWeaver. The random forest-based method GENIE3 has the best performance in this challenge and is followed by the ANOVA and the TIGRESS algorithms. Performance of the remaining contestants is nearly half as good as

GENIE3. The prediction performance of correlation based methods and Bayesian networks is much lower than the performance of GENIE3, ANOVA and TIGRESS [2]. Correlation based network inference is still a commonly used method [3,4].





In previous work, we employed the random forest machine learning algorithm GENIE3 for *Chlamydomonas reinhardtii*. There we were able to infer a gene regulatory network from 1050 publicly available RNA-seq datasets. Its predictions were validated for LRS1 with an average precision of 0.68, with RNA-seq of the mutant [5].

Because of this successful usage of gene regulatory networks, the approach was later employed for the crop plants barley, rice, maize and wheat. With this approach, we confirmed that gene regulatory networks are conserved between species and that closer species show a higher degree of conservation. Candidate transcription factors for the regulation of photosynthesis were identified. Within these candidates, the already known photosynthetic transcription factors GLK1 and GNC were included.

One major limitation is that network inference is computationally resource intensive. The algorithms are not optimized to run on a cloud infrastructure and are not easily accessible. The data acquisition is a critical step for the network inference and needs to be inte-

grated into this procedure. An inferred gene regulatory network contains too much information. Statistical analysis is required to test and generate biological hypotheses. For this purpose, the previously established analysis will be directly integrated into the package. Novel networks will be directly compared with already existing networks. To assign functions to transcription factors, GO-term enrichment for the targets of each transcription factor will directly be performed. If the user has an interest in a special gene list, it will be possible to search for potential regulators and for putative conserved regulators in other species.

A standardized pipeline will be built. Currently the analysis scripts are not optimized to fully utilize high performance computing. For example the GO-term enrichment for each transcription factor is not optimized and takes a lot of time. The number of compared species results in an increase in analyzed transcription factors. This increases the compute time substantially especially for large plant genomes. This leads to the necessity to distribute the workload. The analysis and network inference therefore will be built so that it is

possible to scale across many cores and many machines. This will make comparative network analysis even more accessible, by a reduction of compute time. To achieve this, optimizations will be established either in snakemake [6] or in nextflow [7].

The gene regulatory networks mentioned above were generated with bulk RNA-seq data using already established methods. Predictions by these networks at a high level were successfully validated, but for predictions about single genes the error rate is high. Because of the used data, it is planned to test the possibility to improve the prediction of gene regulatory networks from single cell RNA-seq and ATAC-seq data.

Overall, this project will deliver major advances with the understanding of regulation of photosynthesis and gene regulatory network inference. First steps are taken to decipher the importance of different regulatory levels like transcription factor binding site specificity and open chromatin regions for gene regulation.





---

**References:**

- [1] Greenfield et al. DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS One*, 2010. DOI: 10.1371/journal.pone.0013397.
- [2] Marbach et al. Wisdom of crowds for robust gene network inference. *Nat. Meth.*, 2012. DOI: 10.1038/nmeth.2016.
- [3] Horvath et al. Geometric interpretation of gene coexpression network analysis. *PLoS Comp. Biol.*, 2008. DOI: 10.1371/journal.pcbi.1000117.
- [4] Huang et al. Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. *BMC Plant Biol.*, 2018. DOI: 10.1186/s12870-018-1329-y.
- [5] Lämmermann et al. Ubiquitin ligase component LRS1 and transcription factor CrHy5 act as a light switch for photoprotection in *Chlamydomonas*. *bioRxiv*, 2020. DOI: 10.1101/2020.02.10.942334.
- [6] Mölder et al. Sustainable data analysis with Snakemake. *FI00Res* 10, 33, 2021. DOI: 10.12688/fi00research.29032.2.
- [7] Di Tommaso et al. Nextflow enables reproducible computational workflows. *Nature Biotechnol.*, 2017. DOI: 10.1038/nbt.3820.



# IMPRINT

Prof. Dr. Jens Stoye  
Bielefeld Institute for Bioinformatics Infrastructure (BIBI)  
Faculty of Technology  
Bielefeld University  
Universitätsstraße 25  
33615 Bielefeld

Tel.: +49 (0)521 106 3852  
Fax: +49 (0)521 106 6495  
E-Mail: [bibi@uni-bielefeld.de](mailto:bibi@uni-bielefeld.de)

Editors:  
Editor-in-Chief: Prof. Dr. Jens Stoye (Bielefeld University, BIBI)  
Editorial Team: Dr. Irena Maus (Bielefeld University, CeBiTec)  
Dr. Roland Wittler (Bielefeld University, BIBI)

[www.bibi.uni-bielefeld.de](http://www.bibi.uni-bielefeld.de)

 @BIBIBielefeld

Date: November 2021

Design and Layout:  
MEDIUM Werbeagentur GmbH, Bielefeld

Printing:  
Bruns Druckwelt GmbH & Co. KG, Minden

DOI: <https://doi.org/10.4119/unibi/2959449>

Unless otherwise noted, this publication is licensed under Creative Commons Attribution – Non Commercial – NoDerivatives4.0 International (CC BY NC ND).

For more information see:  
<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>  
<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

 **UNIVERSITÄT  
BIELEFELD**  
 Technische Fakultät

## PHOTO CREDITS:

de.NBI Administration office  
ZB MED  
© [stock.adobe.com/Astibuag](https://www.stock.adobe.com/Astibuag) (Cover)  
© [shutterstock.com/sdecoret](https://www.shutterstock.com/sdecoret) (p. 4,10)  
© [iStock.com/Just\\_Super](https://www.iStock.com/Just_Super) (p. 12, 16, 18)  
© [shutterstock.com/spainter\\_vfx](https://www.shutterstock.com/spainter_vfx) (p. 5, 22)  
© [iStock.com/LeArchitecto](https://www.iStock.com/LeArchitecto) (p. 24)  
© [stock.adobe.com/CROCOTHERY](https://www.stock.adobe.com/CROCOTHERY) (p. 26)  
© [iStock.com/Alkalyne](https://www.iStock.com/Alkalyne) (p. 30)  
© [iStock.com/carloscastilla](https://www.iStock.com/carloscastilla) (p. 5, 34)  
© [stock.adobe.com/Dmitry\\_Knorre](https://www.stock.adobe.com/Dmitry_Knorre) (p. 36)  
© [stock.adobe.com/Gernot\\_Krautberger](https://www.stock.adobe.com/Gernot_Krautberger) (p. 38)  
© [iStock.com/onurdongel](https://www.iStock.com/onurdongel) (p. 42)  
© [shutterstock.com/pingebat](https://www.shutterstock.com/pingebat) (p. 44)  
© [stock.adobe.com/Natasha](https://www.stock.adobe.com/Natasha) (p. 46)  
© [iStock.com/matejmo](https://www.iStock.com/matejmo) (p. 48)  
© [iStock.com/ilyakalinin](https://www.iStock.com/ilyakalinin) (p. 52)  
© [iStock.com/Bacsica](https://www.iStock.com/Bacsica) (p. 54)  
© [iStock.com/jxfzsy](https://www.iStock.com/jxfzsy) (p. 56)  
© [iStock.com/tampatra](https://www.iStock.com/tampatra) (p. 60)  
© [stock.adobe.com/Christoph\\_Burgstedt](https://www.stock.adobe.com/Christoph_Burgstedt) (p. 62)  
© [iStock.com/SandraMatic](https://www.iStock.com/SandraMatic) (p. 66)  
© [iStock.com/Sinhyu](https://www.iStock.com/Sinhyu) (p. 70)  
© [iStock.com/imaginima](https://www.iStock.com/imaginima) (p. 74)  
© [stock.adobe.com/Vjom](https://www.stock.adobe.com/Vjom) (p. 76)  
© [iStock.com/hh5800](https://www.iStock.com/hh5800) (p. 79)

