GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Data management to support deep learning approaches for microbial genomics

Prof. Dr. Alexander Goesmann

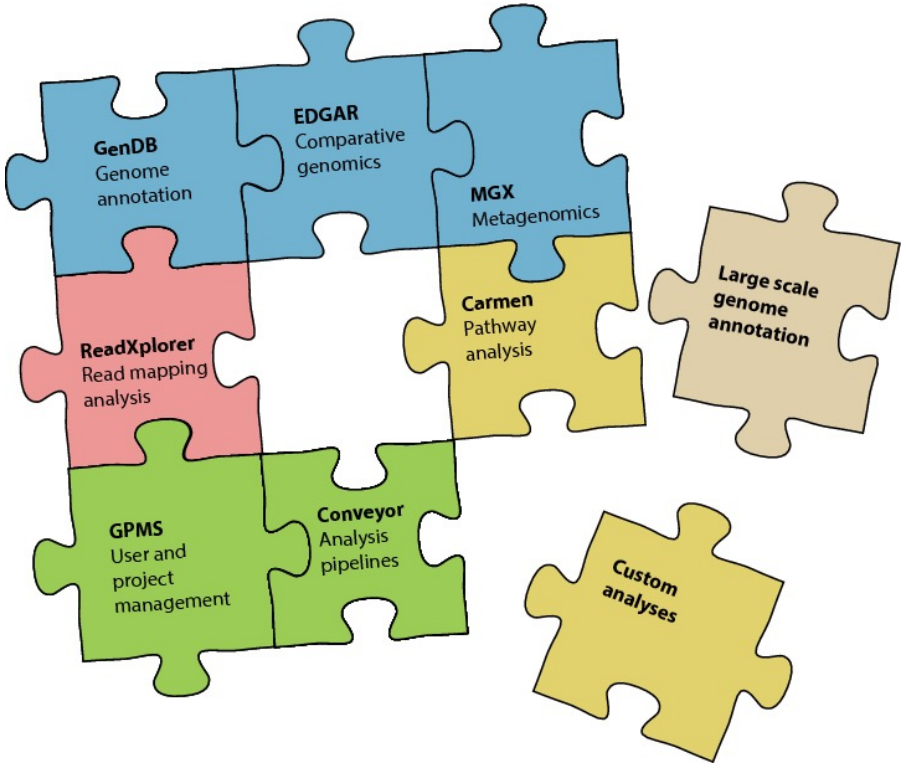
BVMed-Hygieneforum, 8.12.2021

Outline

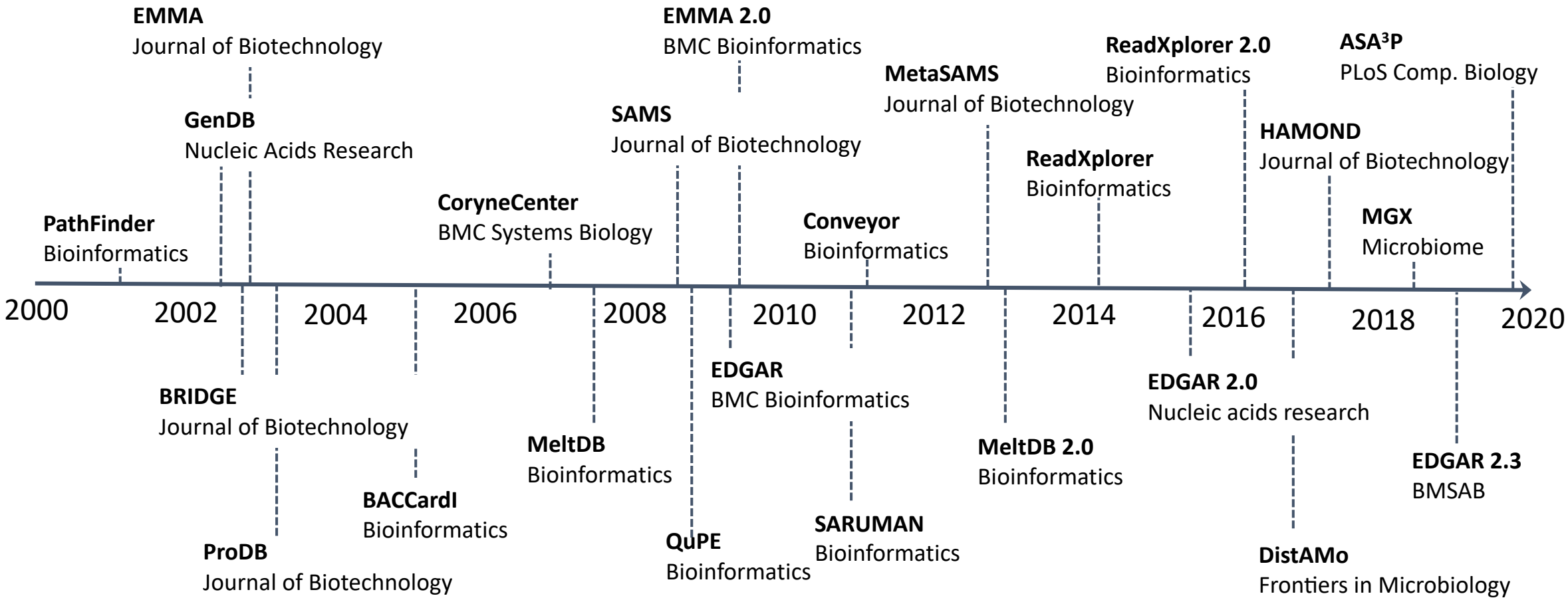
- Two decades of bioinformatics software development
- Current research projects and recent developments
- The Deep-iAMR project
- openBIS

Expertise of our group

- 20 years in bioinformatics software development
- Focus on microbial genomics and transcriptomics with some eukaryotic projects
- High-performance computing & scalable big data analysis
- Large portfolio of in-house developed software platforms
- Experience with various workflow systems such as Conveyor, Galaxy, Nextflow, Snakemake, CWL, …
- Containerization for flexible cloud computing deployment
- de.NBI service center for microbial bioinformatics
- Close collaboration with ELIXIR on international level



Our collection of software tools



Our current research projects

- de.NBI
- NFDI4BioDiversity
- NFDI4Microbiota
- FAIR Data Spaces – Aufbau eines gemeinsamen Cloud-basierten Datenraums für Wirtschaft und Wissenschaft
- KFO 309 – Virus-induced lung injury
- GRK2355 – Regulatory networks in the mRNA life cycle: from coding to non-coding RNAs
- FOR5116 – Communication in host-microbe interaction via exRNA
- LOEWE Schwerpunkt "Diffusible Signals"
- ICIPS – Innovation und Koevolution in der sexuellen Reproduktion von Pflanzen
- Evolution von Gennetzwerken: Die Ranunculales als Modellordnung für evolutionäre Innovationen
- Deep-iAMR
- DeepDomains

The BiGi center for microbial bioinformatics

- Software applications continued as de.NBI services
- Bioinformatics consulting
- Training courses
- Software tools for the field of microbial genome research
- Reusable workflows including Galaxy server
- Storage and compute resources
- Cloud computing environment including Kubernetes

State-of-the-art sequencing technologies

HiSeq X Ten

Feature	Illumina HiSeq X Ten
Read length	150 bp
Reads/Run	6,000 Mio
Yield	**18 Tb**
Time/Run	72 hours
Price/Mb	**0.007 $**
Price/Instrument	10 Mio $

⇒ **1,000 $ per human genome**

MinION™

Feature	MinION
Read length	10,000 bp
Reads/Run	up to 0.6 Mio
Yield	**up to 6 Gb**
Time/Run	up to 48 hours
Price of reagents/Run	99 $
Price/Mb	**0.20 $**
Price/Instrument	900 $

Data management and computational analyses

Project Collaboration & Communication

Nextcloud — File Sharing

Mattermost — Messaging Platform

BigBlueButton — Video Conferencing

COMMON WORKFLOW LANGUAGE, nextflow, Snakemake, Galaxy

Reproducible Bioinformatics Workflows

eLabFTW, openBIS

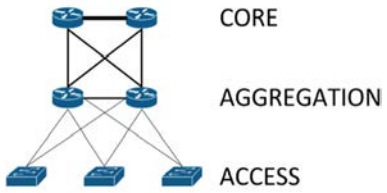
FAIR Data Management

Cloud:
- 16.064 CPU Cores
- 180 TB RAM

Network:
- 10 Gbit
- 40 Gbit

Storage:
- 35 PB
- Volumes / Object Storage

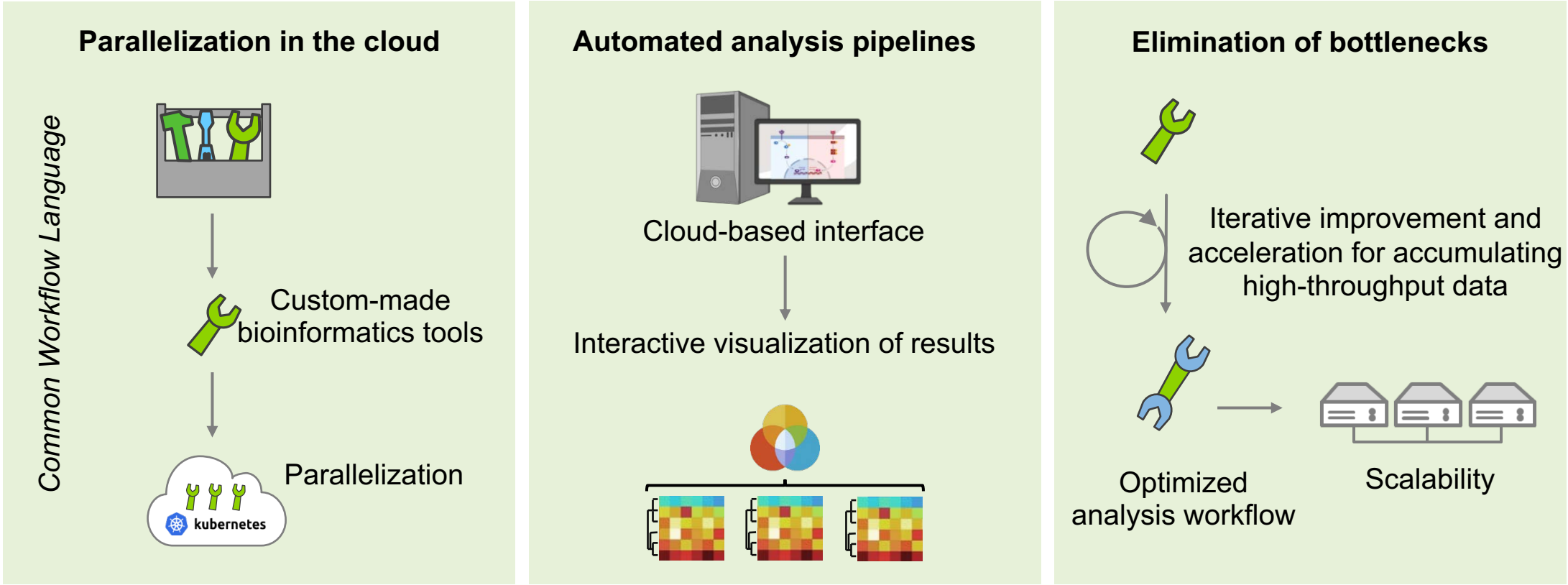
openstack, kubernetes



ceph

All systems are hosted on servers of the Bioinformatics Core Facility at JLU Giessen

Cloud-based analysis workflows



- Focus on next-generation sequencing data analysis workflows
- Continuous improvement of pipelines

ASA³P – Scalable analysis of microbial genomes

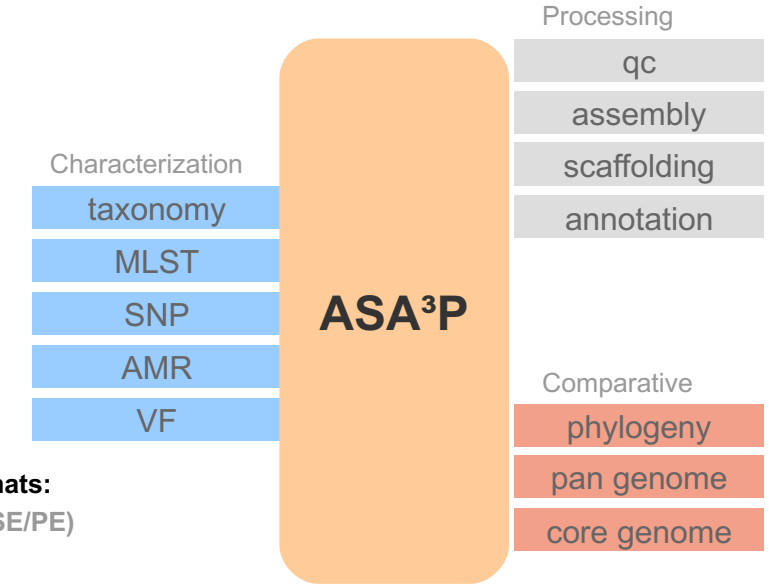
Illumina

PacBio

ONT

Input formats:

- FastQ (SE/PE)
- Bax.h5
- Bam
- FastA
- Genbank
- EMBL
- GFF
- GFF + Fasta



Species	# Genomes
Acinetobacter baumannii	123
Actinobacillus pleuropneumoniae	21
Citrobacter freundii	74
Enterobacter aerogenes	18
Enterobacter cloacae	130
Enterococcus faecalis	193
Enterococcus faecium	356
Escherichia coli	2162
Klebsiella oxytoca	65
Klebsiella pneumoniae	432
Listeria monocytogenes	141
Proteus mirabilis	80
Pseudomonas aeruginosa	114
Serratia marcescens	564
Staphylococcus aureus	64
Total	4,537

- Docker container for local execution
- Ready-to-use single VMs in de.NBI cloud with up to 144 vCPUs
- Full service with password protected web-based access
- Scalable cloud setup based on BiBiGrid to process 1,000+ genomes / day

Collaboration with Med. Microbiology, JLU

ASA³P – Main result page

- Genome Analyses
  - Quality Control
  - Assembly
  - Scaffolds
  - Annotation
- Genome Characterization
  - Taxonomic Classification
  - MLST
  - Antibiotic Resistances
  - Virulence Factors
  - Reference Mapping
  - SNP Detection
- Comparative Analyses
  - Core/Pan Genome
  - Phylogeny

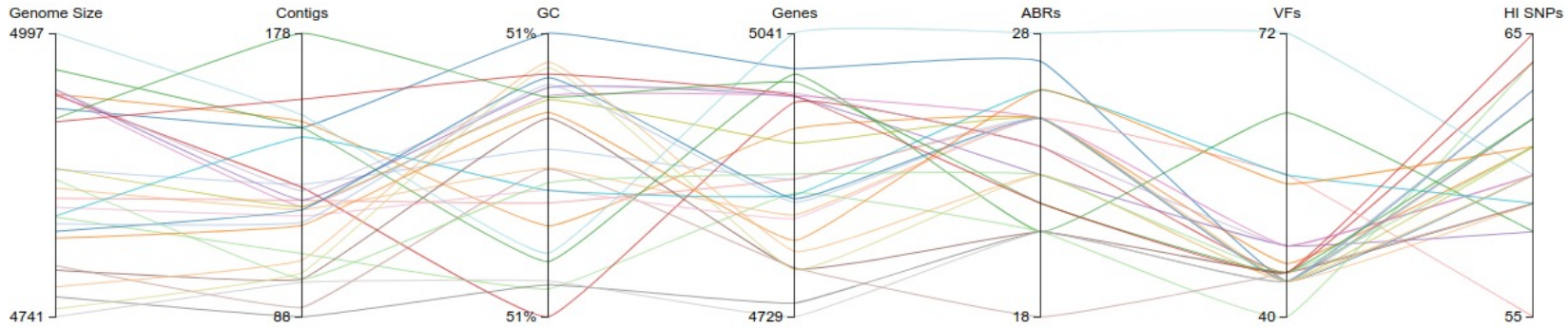
example-ecoli-ST410

This is an example project comparing various Ecoli ST410 strains published by Falgenhauer et al. 2016.

1.0.1 Escherichia

Oliver
Schwengers
oliver.schwengers@computational.bio.uni-giessen.de

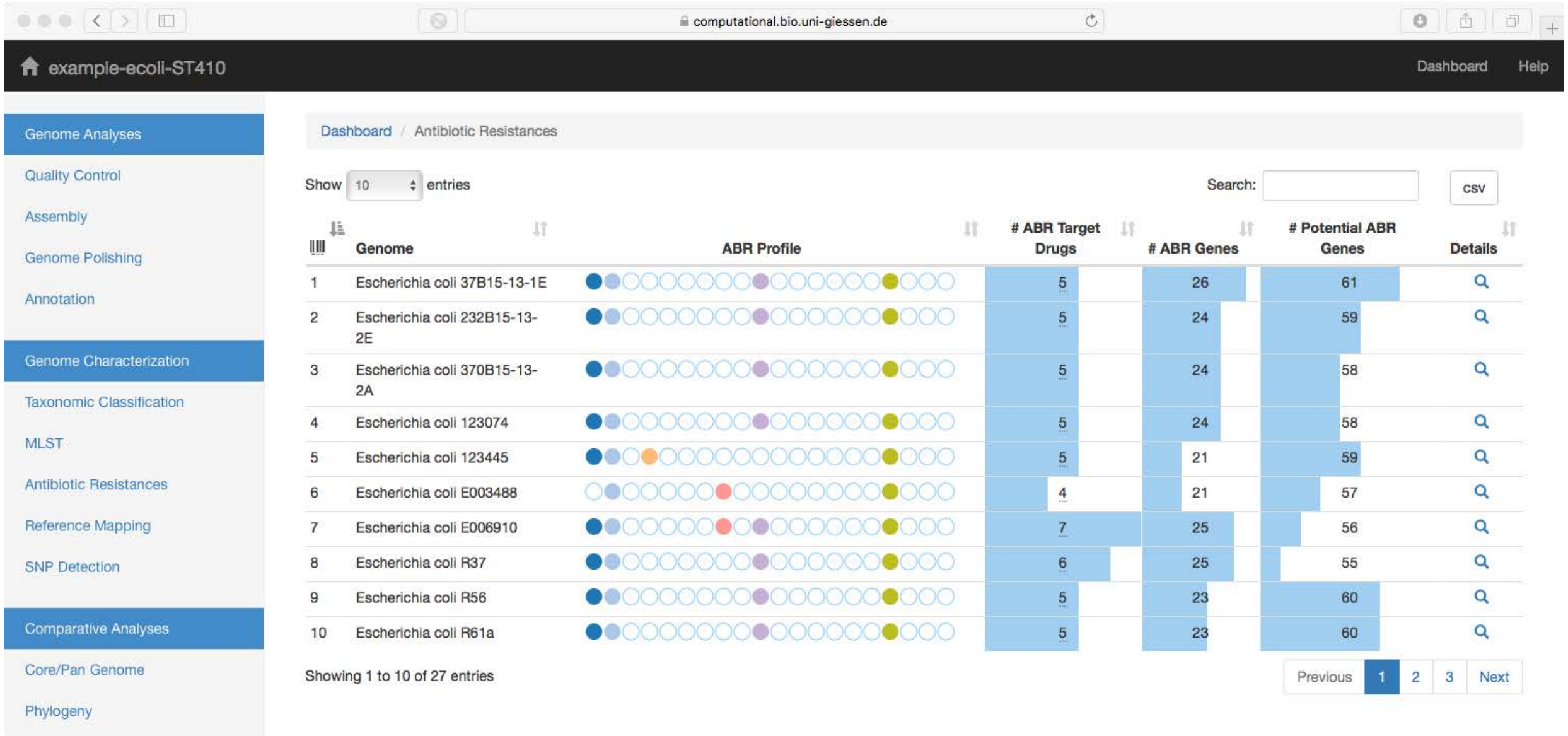
2017-12-05 14:36:45 +01:00
2017-12-05 16:50:21 +01:00
02:13:36.046



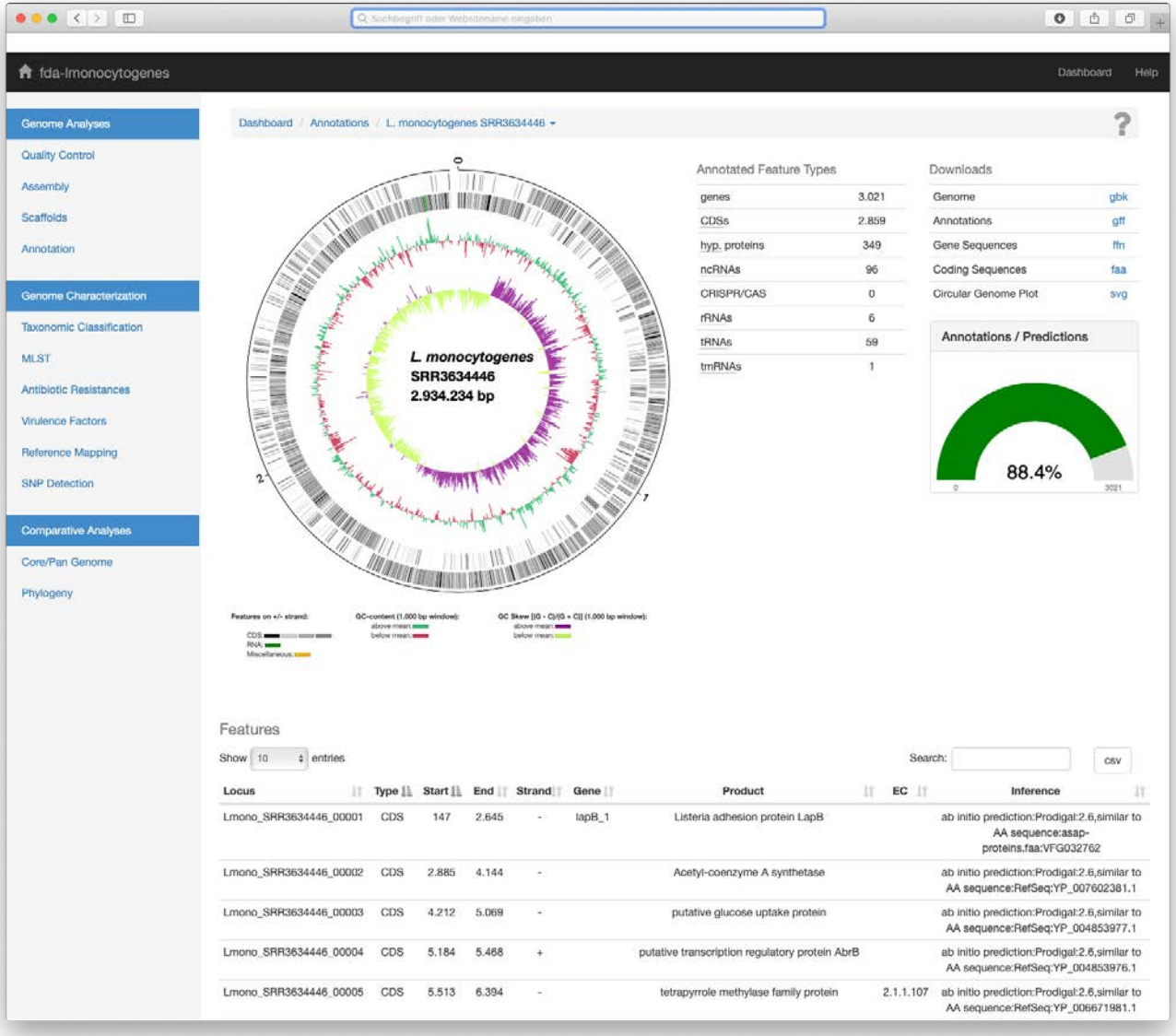
Show 10 entries Search: csv

	Genome	Tax Class	Genome Size	# Contigs	GC	# Genes	# ABR	# VF	# HI SNPs
1	E. coli 37B15-13-1E	Escherichia coli	4.929	148	51	5.002	27	44	60
2	E. coli 232B15-13-2E	Escherichia coli	4.874	130	51	4.880	25	44	62
3	E. coli 370B15-13-2A	Escherichia coli	4.942	150	51	4.936	25	46	61
4	E. coli 123074	Escherichia coli	4.857	122	51	4.841	25	44	59
5	E. coli 123445	Escherichia coli	4.964	148	51	4.996	21	63	58
6	E. coli E003488	Escherichia coli	4.831	108	51	4.865	21	40	64

ASA^{3}P – Antibiotic resistance profiles



ASA3P – Current status and features



Oliver Schwengers

Runtime

- 1,280 *Listeria monocytogenes* strains
- 24 VMs with 768 cores in cloud
- 1 day

Features

- Flexible configuration
- Extensible analysis modules
- Results in standard file formats
- Interactive visualization and browsing

Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, Goesmann A (2020) ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. PLoS Comput Biol. 2020 Mar 5;16(3):e1007134

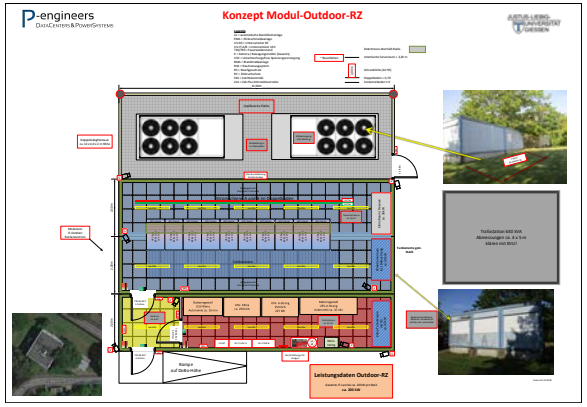
Expansion of IT Infrastructure

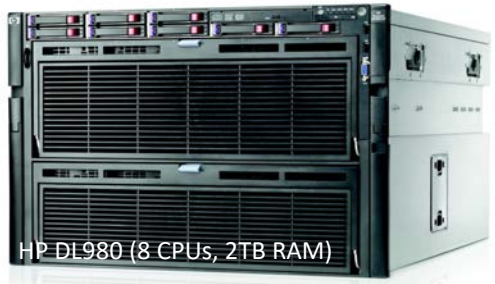
Extension of IT facilities

- Modular data center with 10 racks in container
- 200 kw total
- Investment of 1,5 Mio € (JLU)

Extension of de.NBI cloud

- Additional 16 PB of storage
- Compute nodes, GPU nodes, high-memory servers
- Investment of 2,2 Mio € (BMBF)




HP DL980 (8 CPUs, 2TB RAM)

The Deep-iAMR project

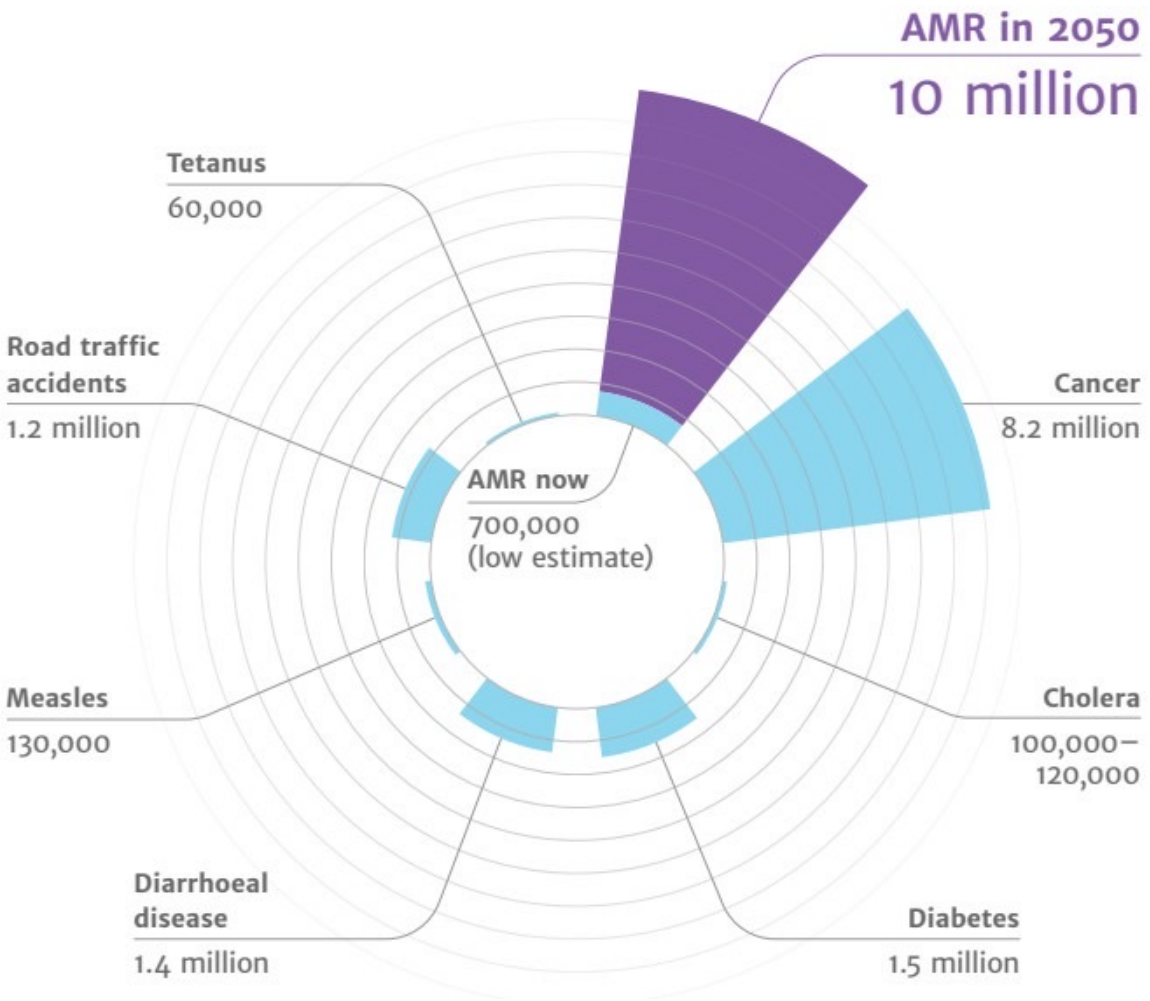
Deep-iAMR – The team

- Prof. Dr. Trinad Chakraborty (Medical Microbiology, JLU Giessen)
  - DNA sequencing, phenotypic genome characterization
  - Experimental validation
- Prof. Dr. Dominik Heider (Data Science in Biomedicine, University of Marburg)
  - Machine learning, deep learning
  - Model development, evaluation, optimization
- Prof. Dr. Alexander Goesmann (Bioinformatics and Systems Biology, JLU Giessen)
  - Genome assembly and annotation
  - Bioinformatic genome characterization
- Project duration: 2020 – 2022
- BMBF funding: 1.1 Mio €
- FKZ 031L0209A, 031L0209B

Bundesministerium für Bildung und Forschung

https://www.gesundheitsforschung-bmbf.de/de/deep-iamr-identifizierung-von-neuen-antimikrobiellen-resistenz-targets-durch-deep-learning-10900.php

Deep-iAMR – Motivation



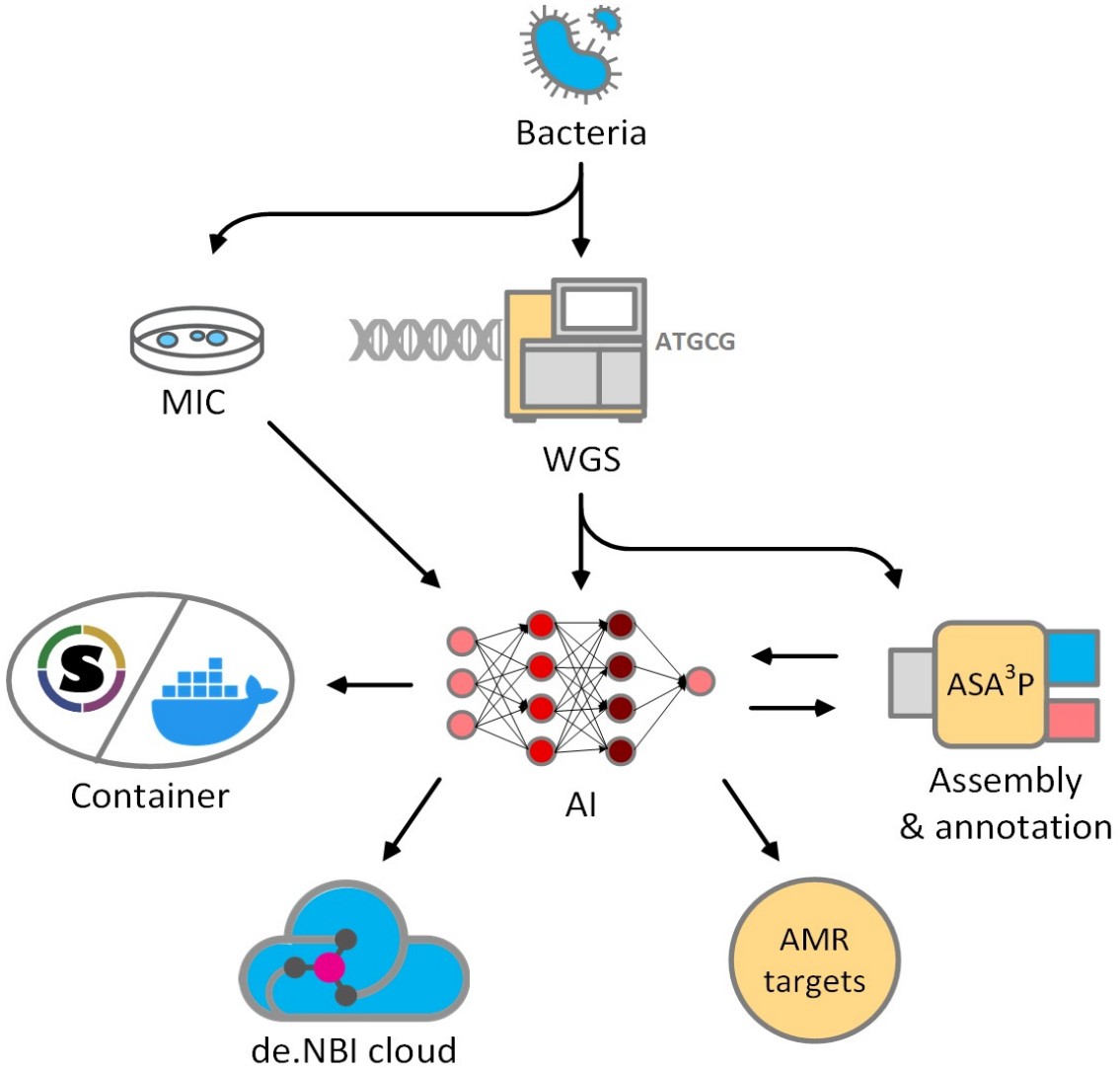
Source: https://amr-review.org

Bacteria ...

- are unicellular
- are omnipresent on earth
- have extremely diverse genomes
- have growth rates down to 9.8 min
- can exchange genetic material, including antibiotic resistance genes

- AMR & MDR bacteria pose a raising threat worldwide
- Even „last resort“ drugs become ineffective
- Global concern for humans, animals, environment → ONE HEALTH approach
- Potentially up to 10 M deaths per year in 2050 without effective measures

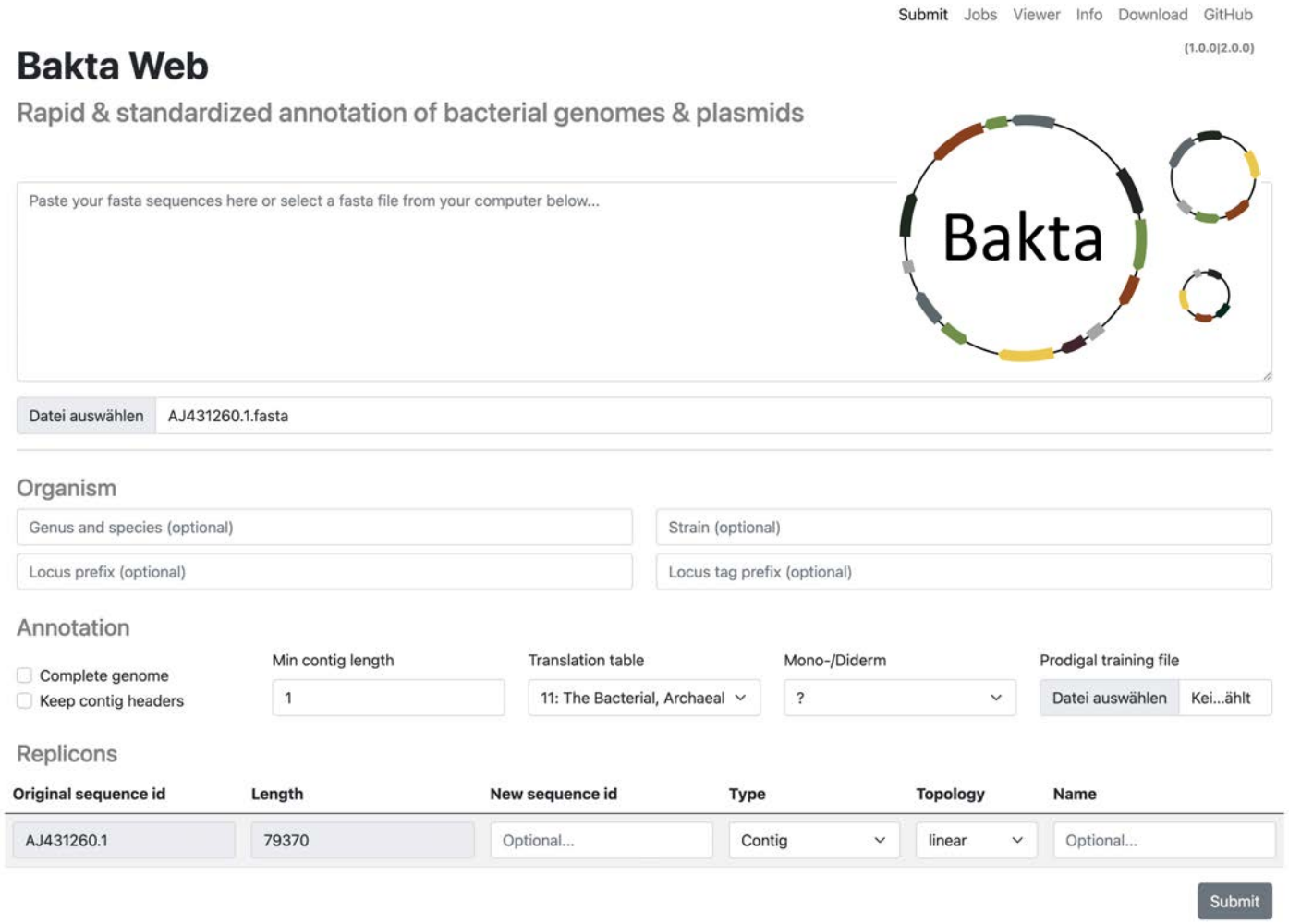
Deep-iAMR



- Sequencing of more than 1,000 *E. coli* genomes and their plasmids
- Analysis of ~1,500 public samples
- Determination of minimal inhibitory concentrations (MIC)
- Evaluation of various feature types (e.g. DNA sequence patterns, SNPs, genes)
- Construction of different machine learning models and performance evaluation
- Experimental validation
- Iterative refinement

Goal: Better characterization of resistance profiles and identification of potential new AMR targets

Bakta

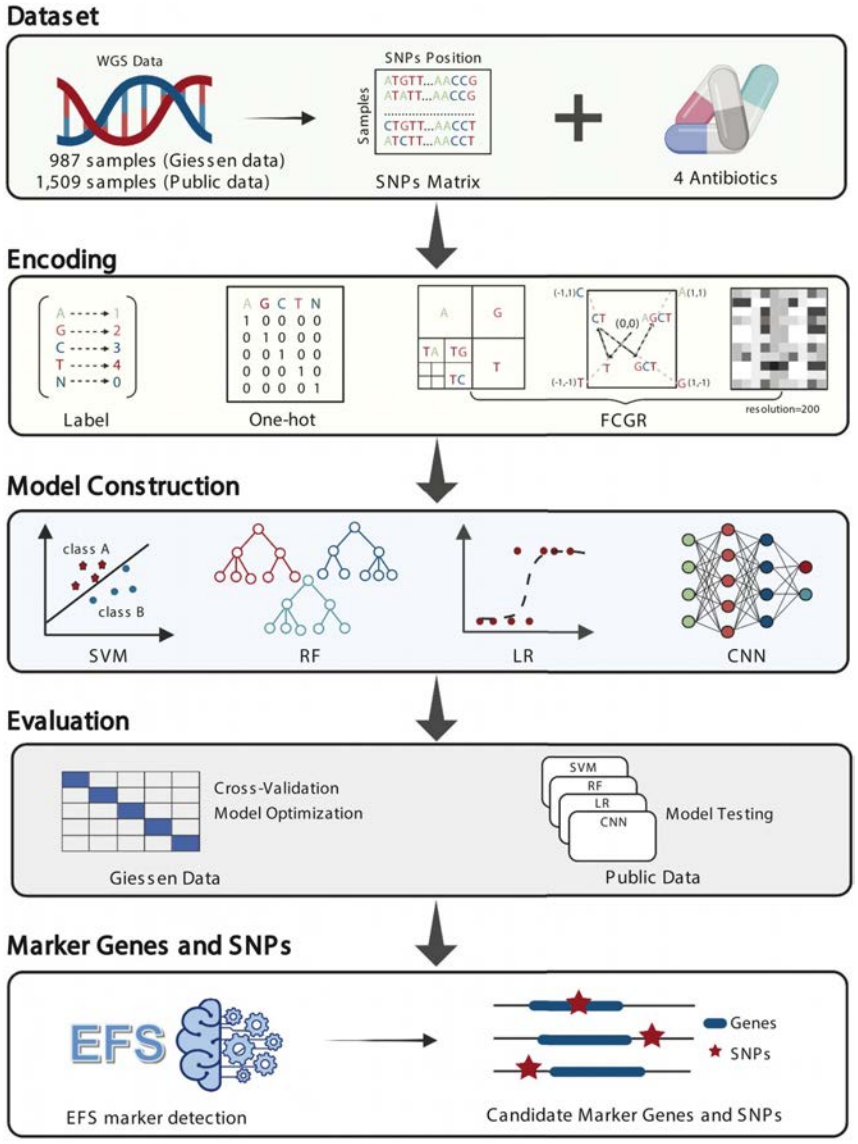


- Rapid and standardized annotation of bacterial genomes and plasmids
- Alignment-free sequence identification
- Command-line application for local installation
- Web-based portal to annotate and browse genomes
- Comprehensive feature annotation including CDS, short ORFs, tRNAs, ncRNAs, CRISPR, …
- Runtime ~10 min for a single genome

⇒ bakta.computational.bio

Schwengers O, Jelonek L, Dieckmann M, Beyvers S, Blom J, Goesmann A (2021) Bakta: rapid & standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics, 7(11). DOI: 10.1099/mgen.0.000685

Prediction of antimicrobial resistance



- Experimental data for 987 local samples
- Focus on four antibiotics: ciprofloxacin, cefotaxime, ceftazidime and gentamicin
- Investigation of four machine learning methods for predicting AMR to four different drugs in *E.coli* from whole-genome sequence data, here mainly SNPs
- Performance evaluation based on cross-validation on our own data and testing of model performance on public data
- Identification of candidate marker genes and SNPs
- In-depth experimental validation and further characterization ongoing

Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, Hauschild AC, Schwengers O, Heider D (2021) Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics, 6:btab681.

Scientific outreach

Oliver Schwengers

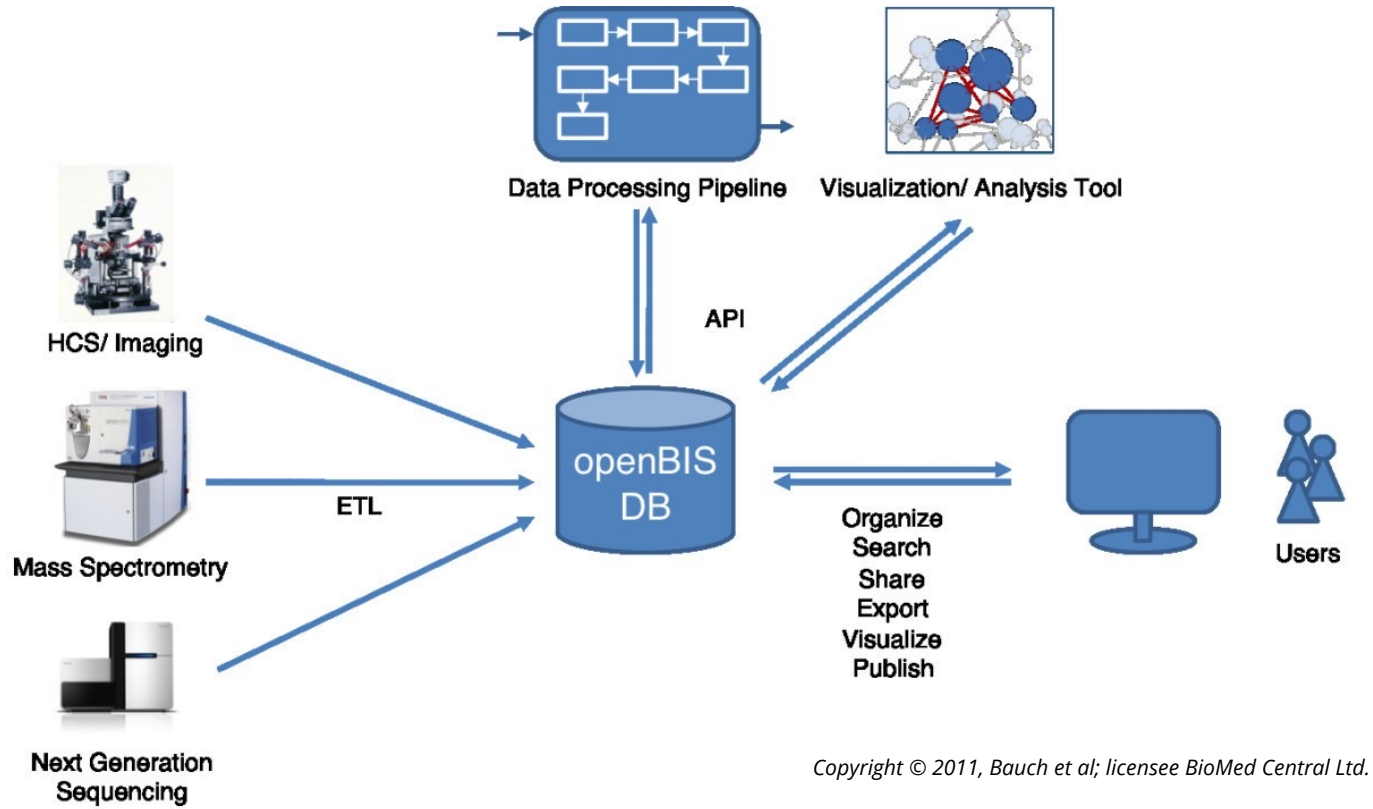
Can Imirzalioglu

https://www.youtube.com/watch?v=UCv8QEKwXts

https://www.denbi.de/images/Downloads/deNBI_KI_brochure.pdf

openBIS

openBIS – System overview

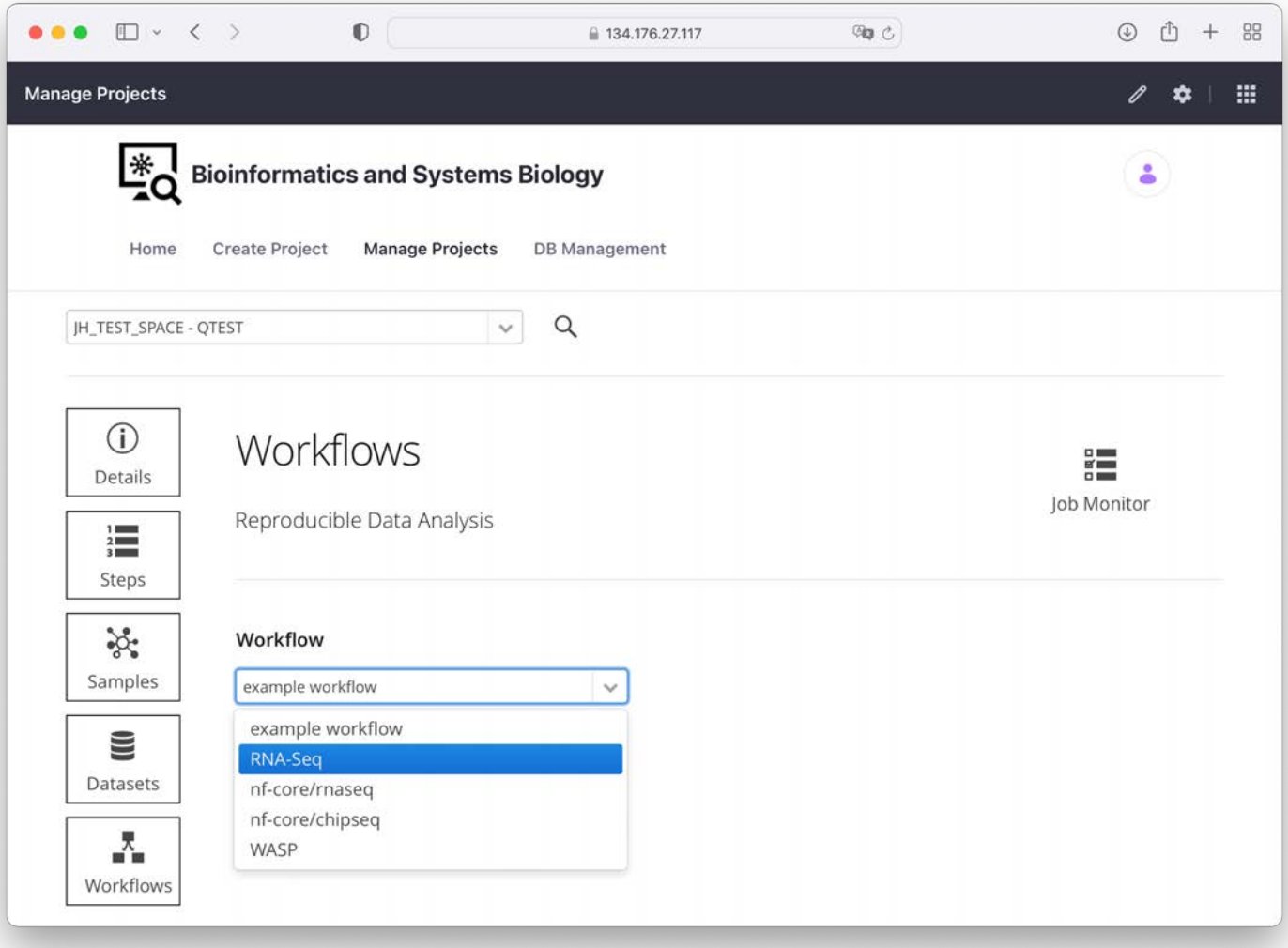


- Digital notebook
- Data management
- Inventory management
- Highly generic data schema
- Modular system design
- Access and rights management
- Support for integrated data analysis
- Excel-compatible import/export

=> openbis.ch

Bauch A, Adamczyk I, Buczek P, Elmer FJ, Enimanev K, Glyzewski P, Kohler M, Pylak T, Quandt A, Ramakrishnan C, Beisel C, Malmström L, Aebersold R, Rinn B (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics, 12:468.

openBIS – Current work



- Local installation & system configuration
- Scalable setup
- Access to HPC & storage
- Data encryption based on Crypt4GH
- Development & integration of workflow registry

Thanks for your attention!

Contact: alexander.goesmann@cb.jlug.de

Homepage: www.computational.bio

Bundesministerium für Bildung und Forschung

de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

ptJ Projektträger Jülich Forschungszentrum Jülich

Federal state of Hesse

DFG Deutsche Forschungsgemeinschaft