Data Science in Bioinformatics

WS 22/23

Kick-Off 10.10.2022

The IKIM

- Institute for Artificial Intelligence in Medicine
- Founded 2020
- At the moment 5 groups, but we are growing

Prof. Dr. Folker Meyer

Head of Data Science Research Group

Dr. Ivana Kraiselburd
Head of Junior Research Group

Dr. Dr. Ricarda Schmithausen

Dr. Simon Magin

Researchers

Katharina Block

Ann-Kathrin Brüggemann

Alexander Thomas

Josefa Welling

PhD candidates

Kerstin Bornemann

Hendrik Schmidt

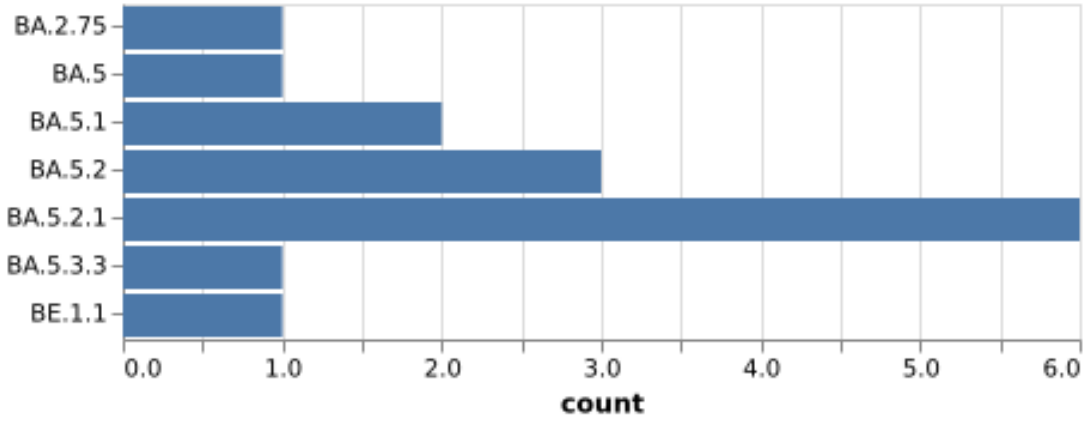
Administration and Tech

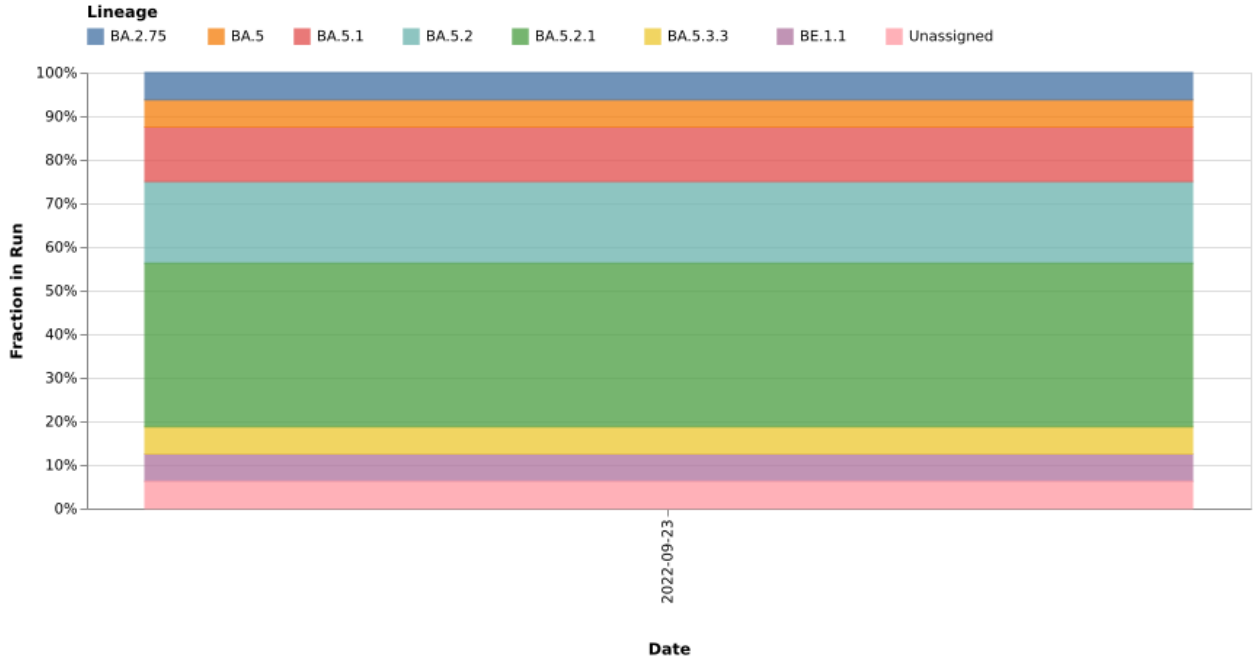
Research topics

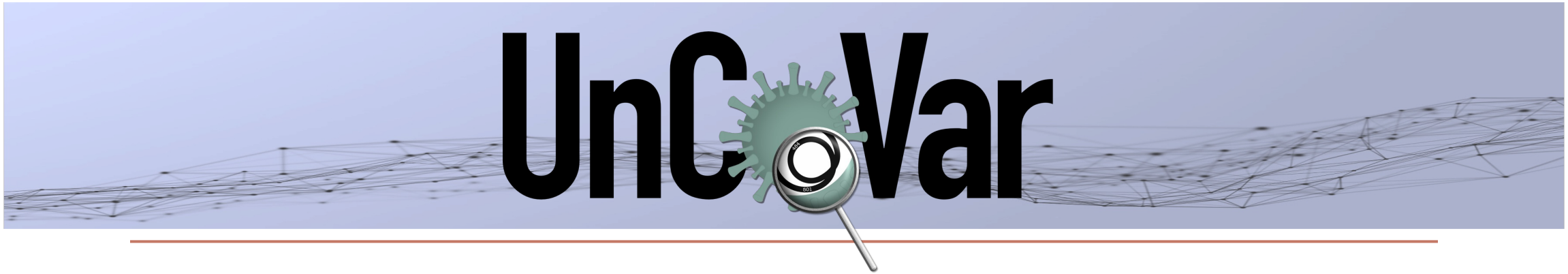
- Analysis of microbial communities (16S rRNA)
- Time series analysis of microbial communities
- Identification of antibiotic resistances
- SARS-CoV-2

SARS-CoV-2 Research

- Started working with SARS-CoV-2 in 2020
- Build laboratory to sequence SARS-CoV-2 → certified since June 2022
- Build pipeline for SARS-CoV-2 variant calling → UnCoVar
- Started SARS-CoV-2 detection in wastewater
- Plan to expand this to other viruses (like norovirus)


BA.2.75
BA.5
BA.5.1
BA.5.2
BA.5.2.1
BA.5.3.3
BE.1.1
0.0
1.0
2.0
3.0
4.0
5.0
6.0
count


Lineage
BA.2.75
BA.5
BA.5.1
BA.5.2
BA.5.2.1
BA.5.3.3
BE.1.1
Unassigned
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Fraction in Run
2022-09-23
Date


UnCoVar

The project – What are we going to do?

- Some biological and bioinformatics basics
  - Exercises to consolidate the knowledge
- Individual work:
  - Presentation of topics and tools important for the pipeline
- Group work:
  - Implementation of a pipeline to call SARS-CoV-2 variants
  - Presentation of your final pipeline

The project – Why are we doing this?

- State-of-the-art research
- Variety of bioinformatic methods
- Applicable for different tasks
- Impression of the research of our group

The timeline so far

Day	Date	Topics	Lecturers
1	10.10.2022, 10:00	Introduction, Questions, Presentation Topics	Folker Meyer, Hendrik Schmidt
2	17.10.2022, 14:00	Biology basics, Viruses	Folker Meyer, Simon Magin
3	24.10.2022, 14:00	Conda, Bioinformatics Basics, Git, Reproducibility	Hendrik Schmidt, Alexander Thomas, Folker Meyer
Dropped	31.10.2022, 14:00		
4	07.11.2022, 14:00	Snakemake	Johannes Köster

Further on in the semester

- 14.11. – 05.12.: individual presentations
- 12.12. – 16.01.: Q&A sessions for problems with your pipeline
- 23.01. & 30.01.: final presentation of your pipeline

Grade composition

- Both presentations
- The pipeline
  - Does it work?
  - Your code
  - Documentation of what you did, when you did it and who participated
- Detailed description on our website in the expectations file

Biological knowledge

- Who knows the three domains of life?
- Do you know the term deoxyribonucleic acid?
- Who has worked with biological topics after school?
- Has anyone worked with sequence data before?

Binding registration

- Only 12 participants: **binding registration until October 16**
- Build teams & allocate presentation topics on October 17

Presentation Topics

Topic	Description
Bioinformatics and Covid	Basics concerning SARS-CoV-2 analysis in bioinformatics
Sequencing Technologies	Illumina, Nanopore etc.
Quality Control (QC)	Checking of data reliability
Trimming & Filtering	Bias and error reduction
Alignment	Sequence analysis - Algorithms
Alignment 2	Sequence analysis - Implementation
Assembly	Reconstructing sequences - Algorithms
Assembly 2	Reconstructing sequences - Implementation
Scaffolding	Linking of genomic sequences
Assembly control	Ranking of assemblies
Variant calling	Identification of represented variants
Lineage Assignment	Identification of represented lineages


Questions so far?

IKIM infrastructure

- Home computers: not enough resources
- Where will we be running the software/pipelines?
  - On our institute internal cloud (~160 servers, 1 PB of storage space)
  - Specifically: c44, c45, c47
  - All of which are Ubuntu server instances

IKIM infrastructure

- How do you access the IKIM cluster?
- Via SSH
  - OpenSSH on MacOS/Linux
  - OpenSSH on Windows
  - PuTTY on Windows
- We use public key auth, so first: generate a public-private key pair
- A good write-up of configurations and steps to take can also be found at:
  - https://ikim-essen.github.io/ClusterDocs/getting-started/

IKIM infrastructure

- When your configuration is correct and your account has been added:
  - „ssh c44“ → host key verification prompt → home directory
- Start moving around and familiarize yourself with the machine
- For those unfamiliar with Unix cmdline:
  - cd
  - ls
  - mkdir
  - cp / mv / rm

IKIM infrastructure

- What‘s next?
  - Please generate your public-private key pair and send the **public** key + your full name to me via email (hendrik.schmidt@uk-essen.de)
    - Hint: Password Manager (keepassxc for example) with ssh-agent integration
  - We will create your user accounts
  - Please set up your ssh client correctly, so you can access the servers
    - If you need help: https://ikim-essen.github.io/ClusterDocs/getting-started/
    - If you cannot find the solution to your problem on the website: contact me via Email with a detailed description of what did not work
  - Session on 24th October: more on development workflows, git, conda, mamba and containers