de.NBI - Bioinformatics Integrations for KNIME

The Center for Integrative Bioinformatics (CIBI)

Knut Reinert, Freie Universität Berlin, MPI Molgen

What is de.NBI and CIBI?

de.NBI

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
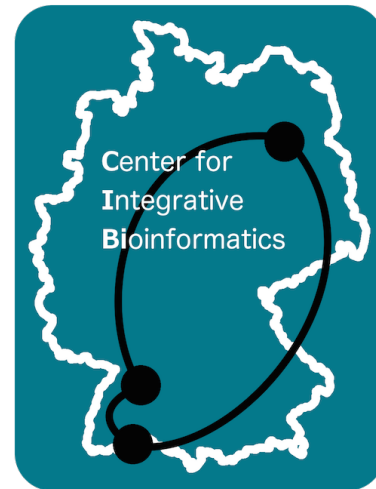
| Home | Mission | Organization | Network | Services | Training | Events | News | Jobs | Help |

Home

Welcome to de.NBI

The 'German Network for Bioinformatics Infrastructure – de.NBI' is a national infrastructure supported by the Federal Ministry of Education and Research

Events

Swiss German Galaxy Days
de.NBI Summer School 2016: From Big Data to Big Insights
International de.NBI-Symposium "Bioinformatics for Human Health & Disease"
de.NBI Mini symposium "Bioinformatics for Metagenome Analysis"
Galaxy DevOps Workshop featured by de.NBI and ELIXIR

CIBI 1.0 – SeqAn, OpenMS, KNIME

SeqAn is a **generic** open-source C++ library of **efficient** algorithms and data structures for the analysis of **NGS data**. **www.seqan.de**
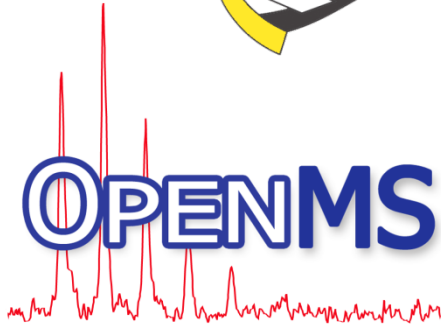
OpenMS is an open-source software library and tool collection for **computational mass spectrometry**. **www.openms.org**

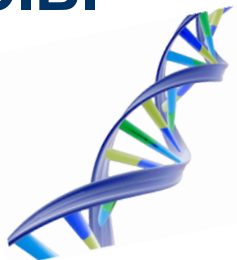**KNIME** is a user-friendly **graphical workbench** for the entire **analysis** process. **www.knime.org**

Address Different Communities

Bioinformaticians
software developers

Experimentalists
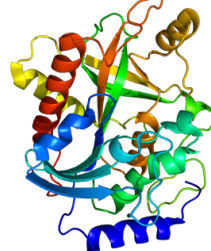*software users*
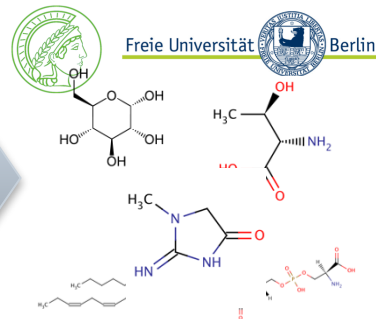
CIBI

Genome

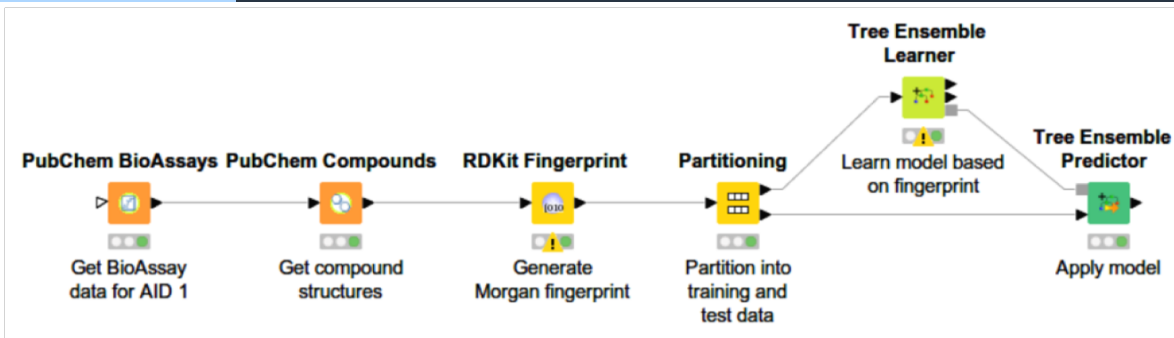Epigenome

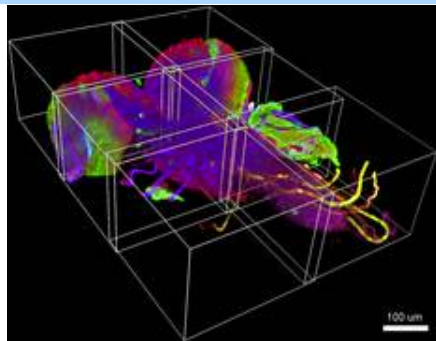Transcriptome

RNOme

Proteome

Interactome

Metabolome

Lipidome

Next-Generation Sequencing

Mass Spectrometry

Imaging

## Workflows

CIBI 2.0 – Partner Projects

- **DAIS – Dresden Analysis-of-Images Suite**
  - Fiji - Leading software platform for image processing
  - Close interactions with the KNIME team already
  - Enables CIBI to move towards integration of imaging and omics
- **MASH – Metabolite Annotation and SHaring**
  - Strengthens CIBI in the area of metabolomics
  - Close interactions with the OpenMS team already
  - Enables CIBI to move towards broader multi-omics integration

Partner Project DAIS

DAIS – Dresden Analysis-of-Images Suite (Myers/Tomancak – MPI CBG)

- Leading software platform for biological image analysis - **Fiji**

- State-of-the-art image processing algorithms

- Integrated with KNIME

20 µm
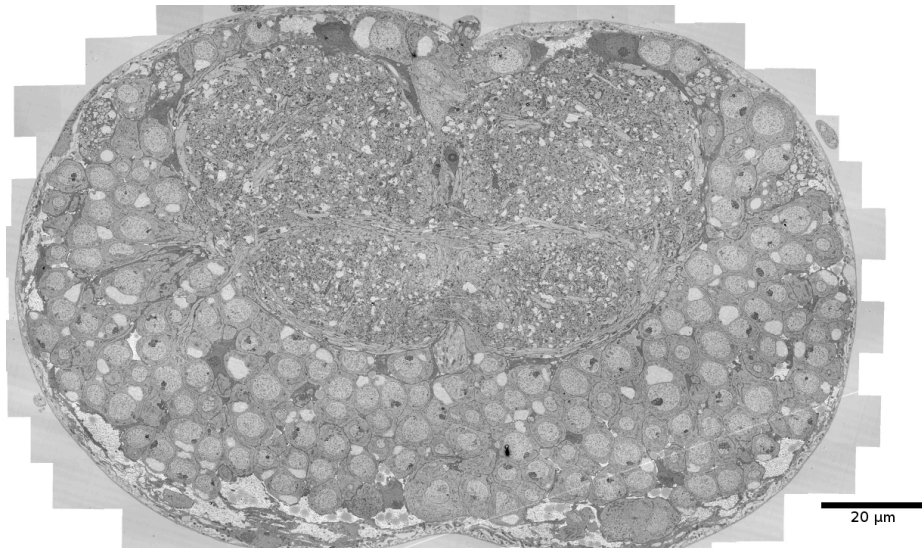
Gene Myers

Pavel Tomancak

Florian Jug

Tobias Pietzsch

KNIME summit 2017 Berlin

Fiji usage map

This page was produced using GeoLite data created by MaxMind, available from MaxMind, downloaded from ipinfodb.com

8

Partner Project MASH

- **MASH – Metabolite Annotation and Sharing**
  (Steffen Neumann, IPB Halle)

  - Strengthens CIBI in the area of metabolomics:

    - Metabolite identification (MassBank, MetFrag)

    - Data sharing (MetaboLights@EMBL-EBI, ISA-Tab)

    - Data clinic first-aid consultations

  - Tightly integrated with European and international metabolomics projects and initiatives

$$S_{C_1} = 1.0$$
$$S_{C_2} = 0.89$$
$$S_{C_3} = 0.63$$
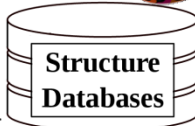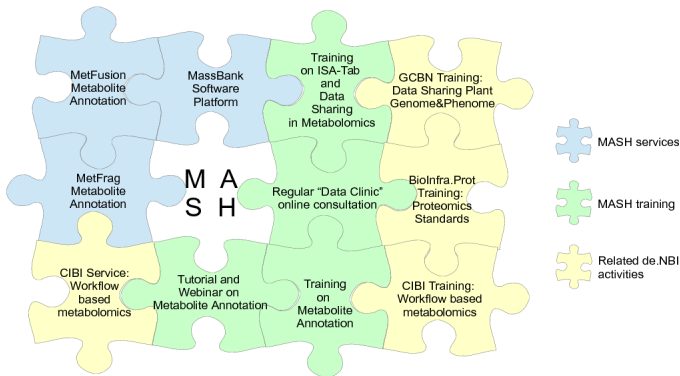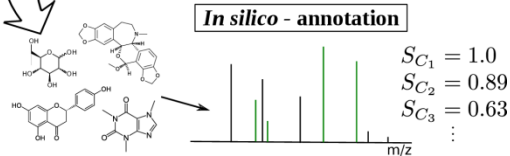
OpenMS and SeqAn integration

Generic KNIME Nodes

CTD Format

Tool description and parameters in XML

```xml
<tool name="MasaiMapper" version="0.7.1 [14053]"
docurl="http://www.seqan.de" category="Read Mapping" >
    <executableName>masai_mapper</executableName>
    <description>Masai Mapper</description>
    <manual>Masai is a fast and accurate read mapper based on approximate seeds and
multiple backtracking.
See http://www.seqan.de/projects/masai for more information.
(c) Copyright 2011-2012 by Enrico Siragusa.
</manual>
    <cli>
        <clielement optionIdentifier="--write-ctd-file-ext" isList="false">
        <mapping referenceName="masai_mapper.write-ctd-file-ext" />
        </clielement>
........
```

GKN plugin with node generator

Generic KNIME Node project can generate node source code and provides a base plugin

https://github.com/genericworkflownodes

GKN plugin:

- generic interfaces/abstract classes (config. dialog, param. IO, tool executors)

- file handling classes/nodes and flow control nodes with fileports (uses&extends KNIME filehandling sources)

Node generator:

- Compatible with both (our) internal and external tools. This means, **ANY** tool can be integrated in KNIME as long as it has a CTD.

Case study 1: Taxonomic profiling

The DNA is loaded into automated sequencers. Celera's automated sequencers run 24-7 and have the ability to decipher more than 100 million letters of genetic code per day - the equivalent of 3 percent of the entire human genetic code every day.

The sequencers create am image of the DNA samples being decoded. The four letters of the genetic code -- A, C, T, G -- each are assigned a color.

Data volume and cost:
In 2000 the 3 billion base pairs of the human genome were sequenced for about 3 billion US$ Dollar

100 million bp per day

Sequencing today...

Illumina HiSeq

100 billion bps per day

Within roughly ten years sequencing has become about 10 million times cheaper
Pangenomics analyses possible

Taxonomic Profiling

Taxonomic profiling is a process of generating qualitative and quantitative information about a composition of a given microbial community.

Challenges
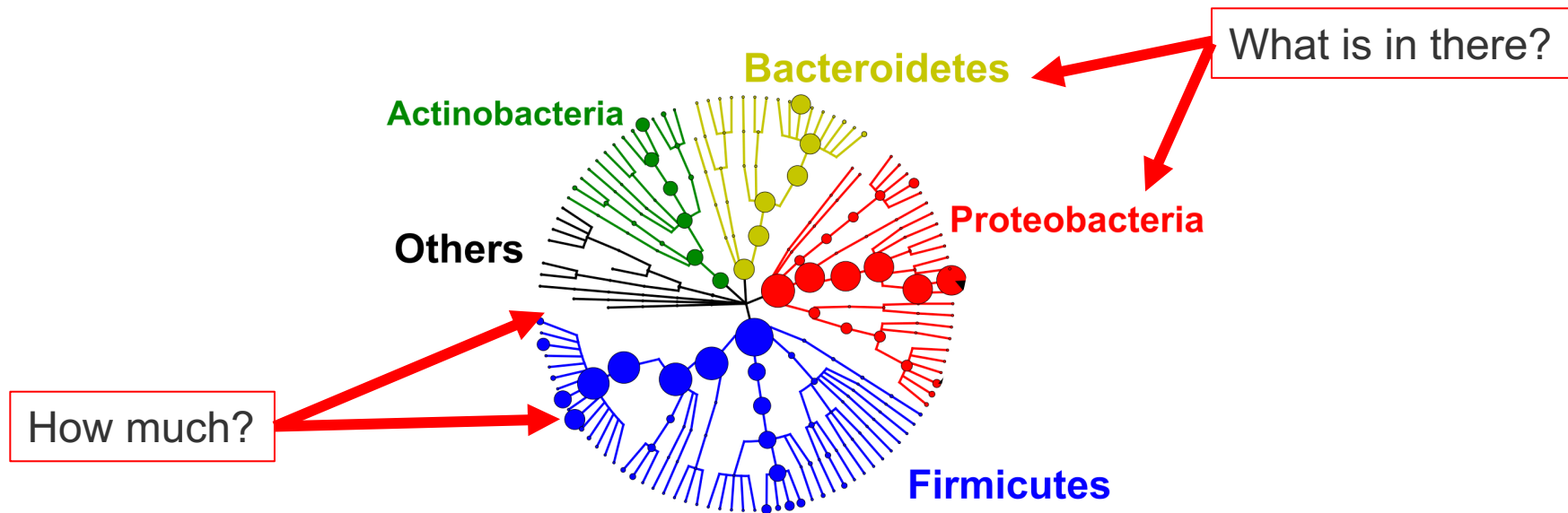
Shared (homologous) regions of genome sequences across multiple microorganisms
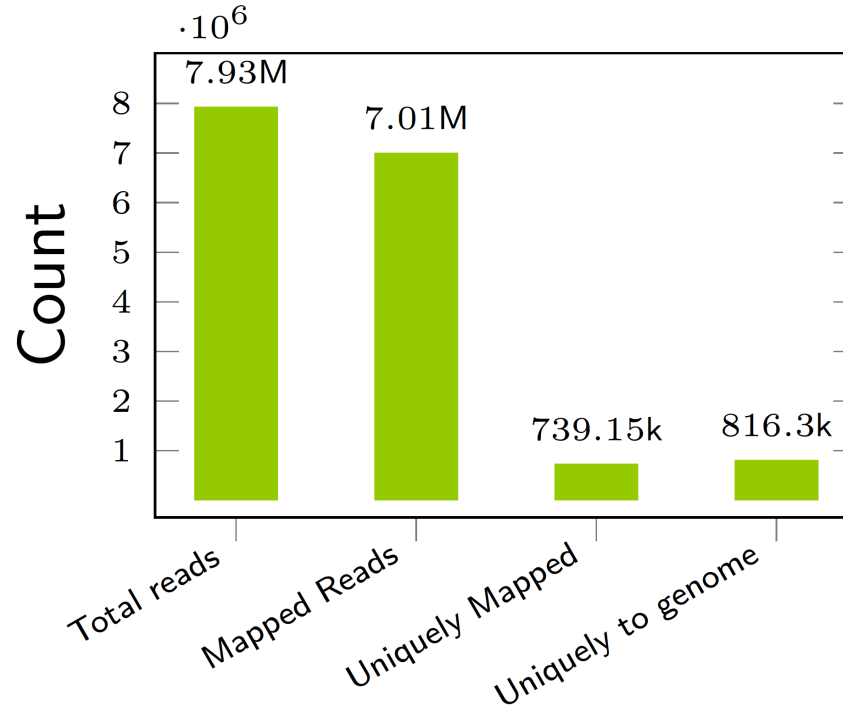
How existing methods try to resolve this ...

Prepare non overlapping reference catalog (MetaPhlAn, GOTTCHA, mOTUs)

- Unable to detect low abundance organisms.

Assign shared reads to their LCA

- Most of the information goes down to the upper levels.

SLIMM - Method
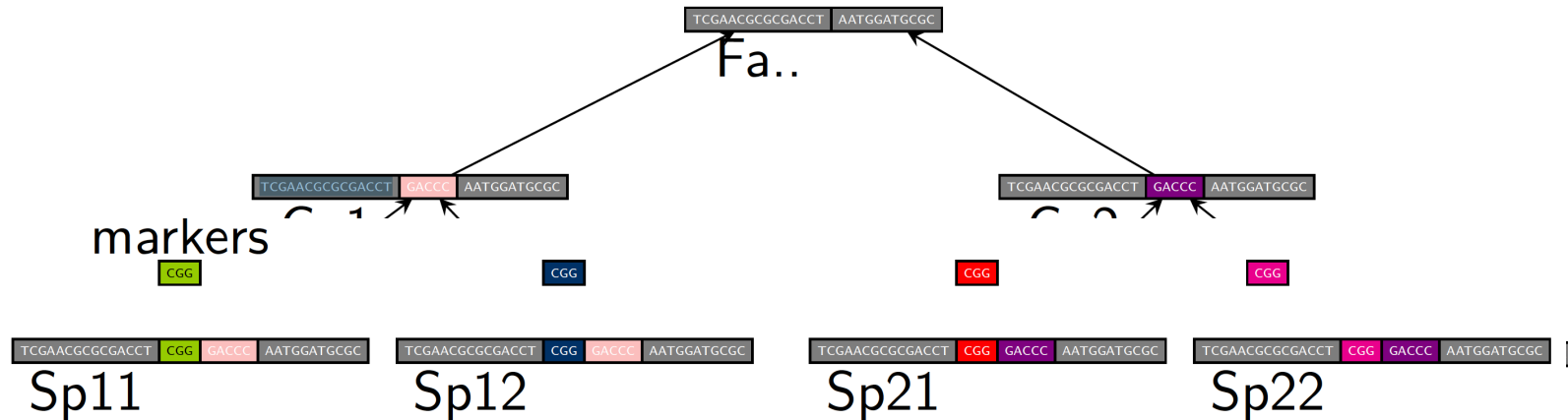
- Collect information about genomes from mapping results

- Bin reads according to their mapping positions

  1. Shared Reads

  2. Uniquely mapping reads

- Discard unlikely genomes based on coverage landscape using quantile based cutoff

- Recalculate reads uniqueness

- Assign reads to their LCA and calculate abundances at a given rank

SLIMM: Species level identification of microorganisms from metagenomes, PeerJ, 2017 (also GCB 2016)
Temesgen Hailemariam Dadi[1,2], Bernhard Renard[3], Lothar H. Wieler[3], Torsten Semmler[3,4], Knut Reinert[1,5]
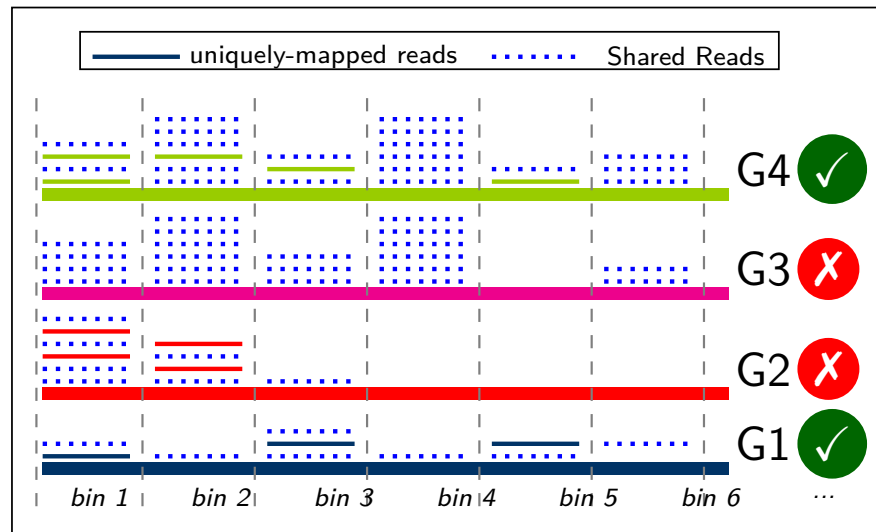
Precision, Recall and F1-Score
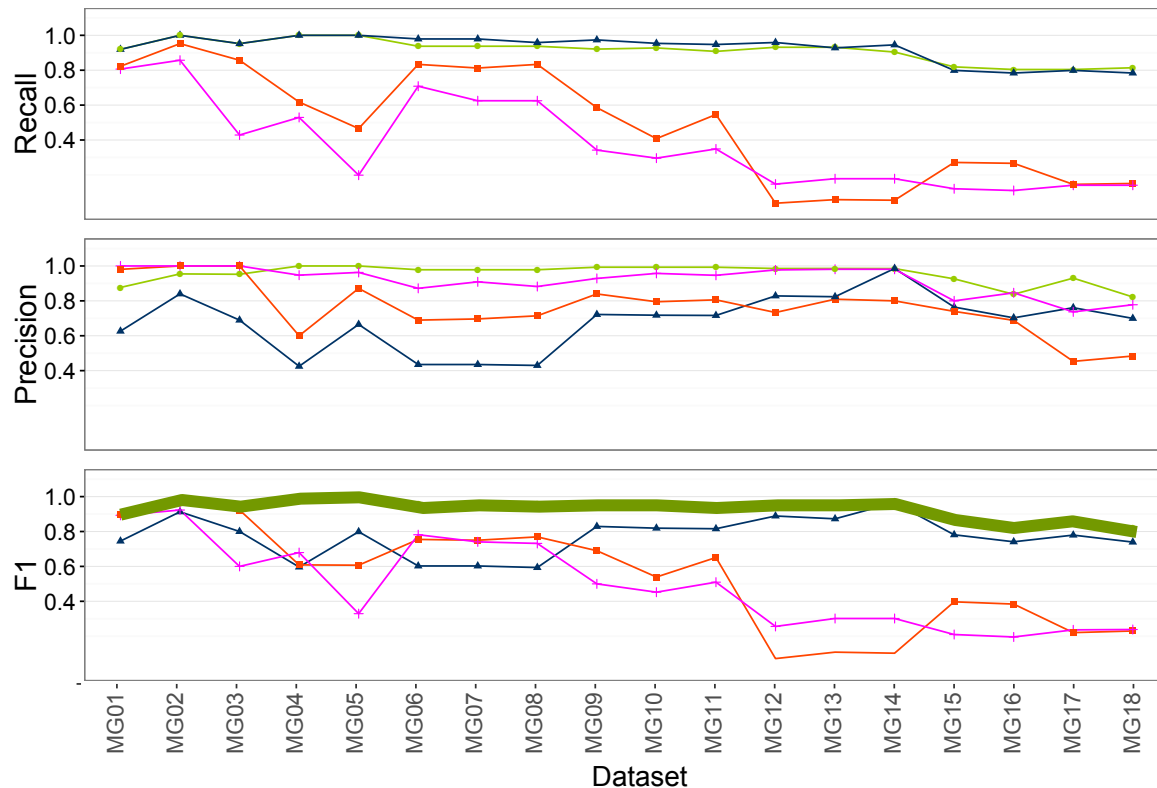
SLIMM: Species level identification of microorganisms from metagenomes, PeerJ, 2017 (also GCB 2016)
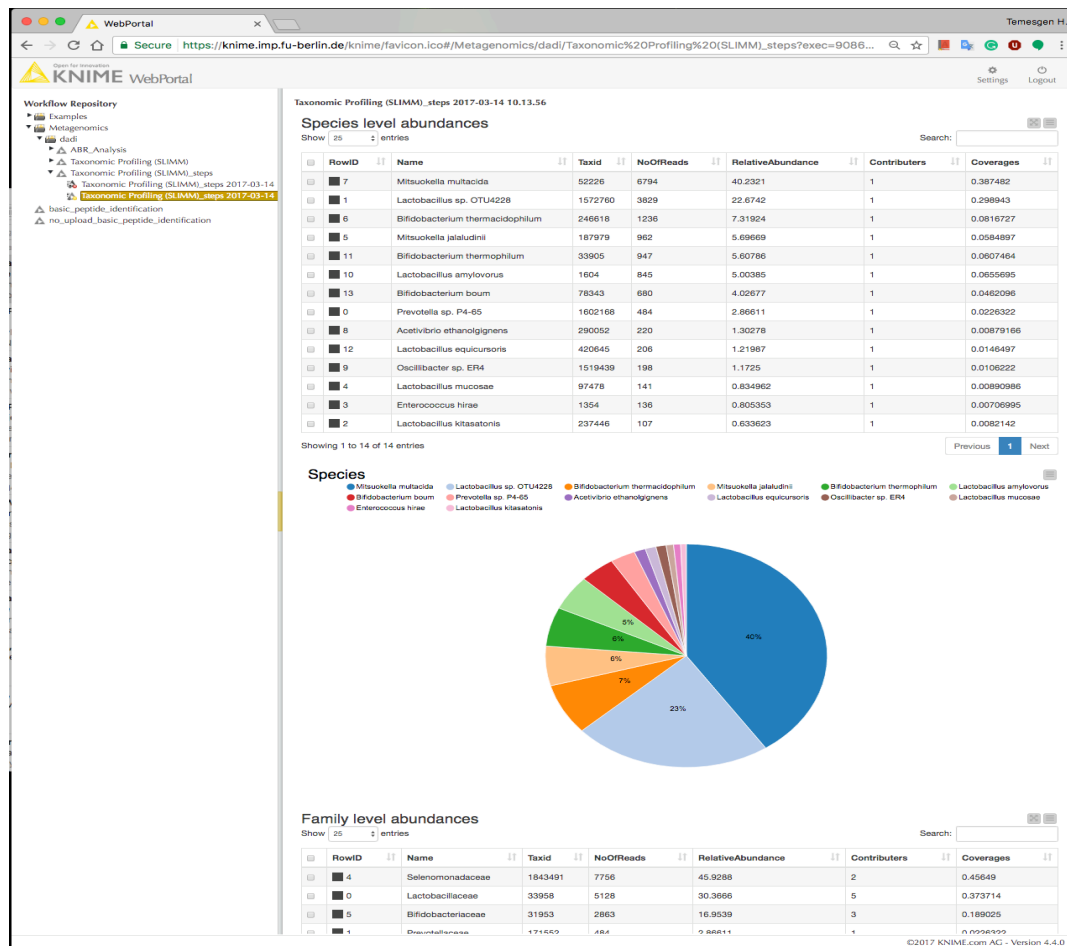Temesgen Hailemariam Dadi[1,2], Bernhard Renard[3], Lothar H. Wieler[3], Torsten Semmler[3,4], Knut Reinert[1,5]
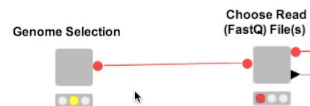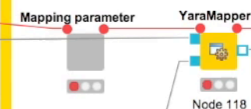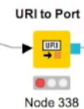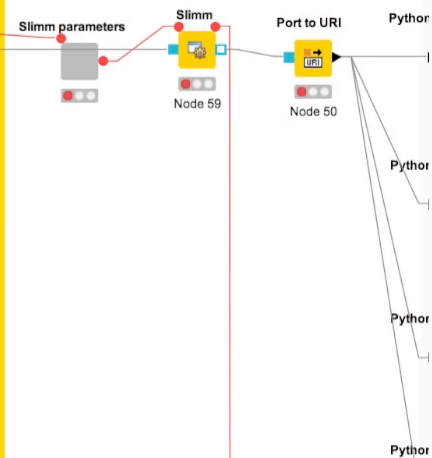
On a server

SLIMM workflow

View on KNIME server

Case study 2: Label free quantitation

Quantitative Data – LC-MS Maps
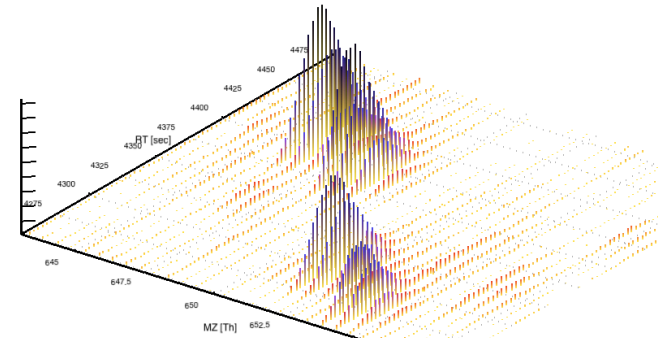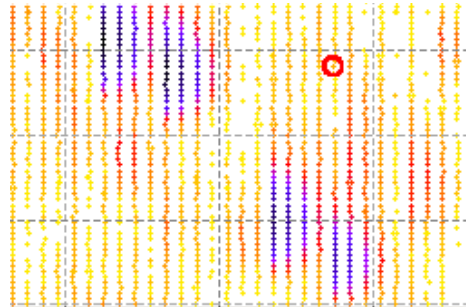
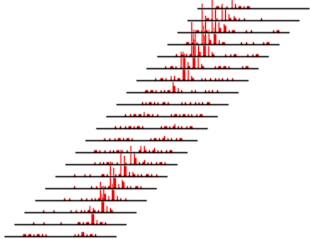Spectra are acquired with rates up to dozens per second

Stacking the spectra yields **maps**

Resolution:

- Up to millions of points per spectrum
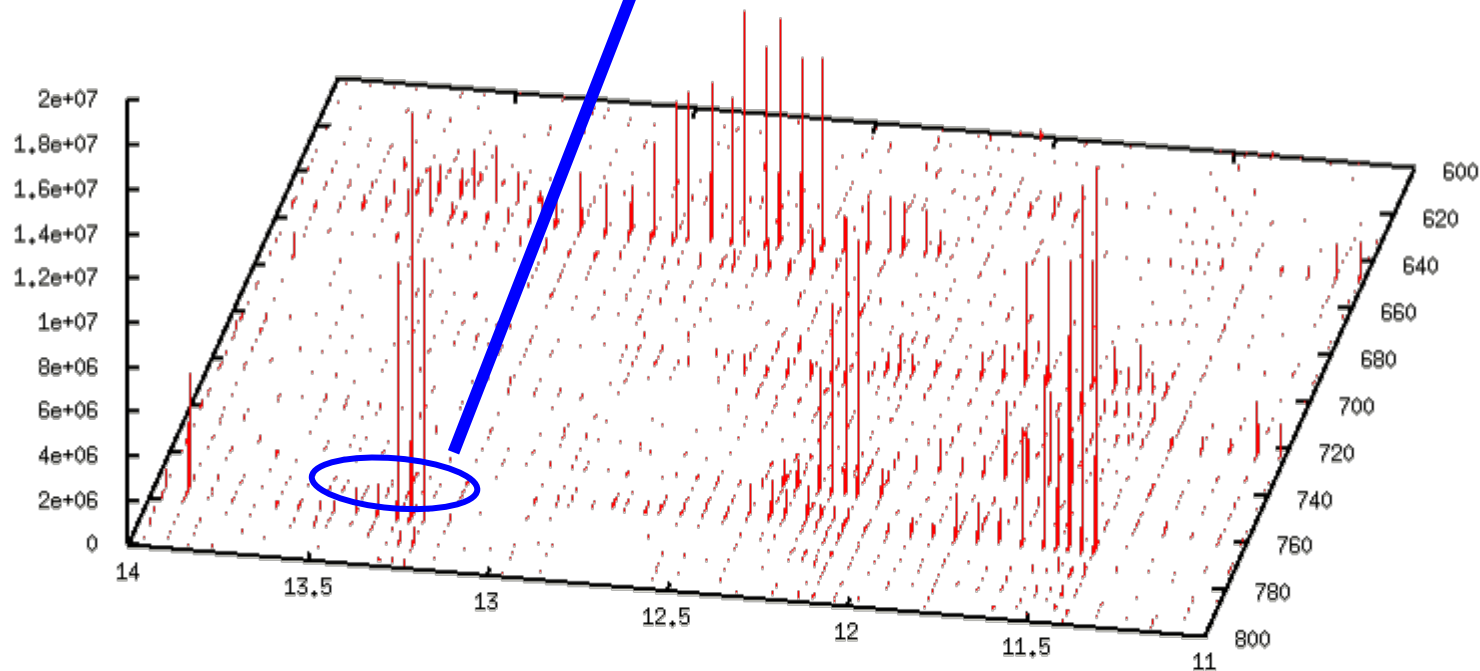
- Tens of thousands of spectra per LC run

Huge 2D datasets of up to hundreds of GB per sample

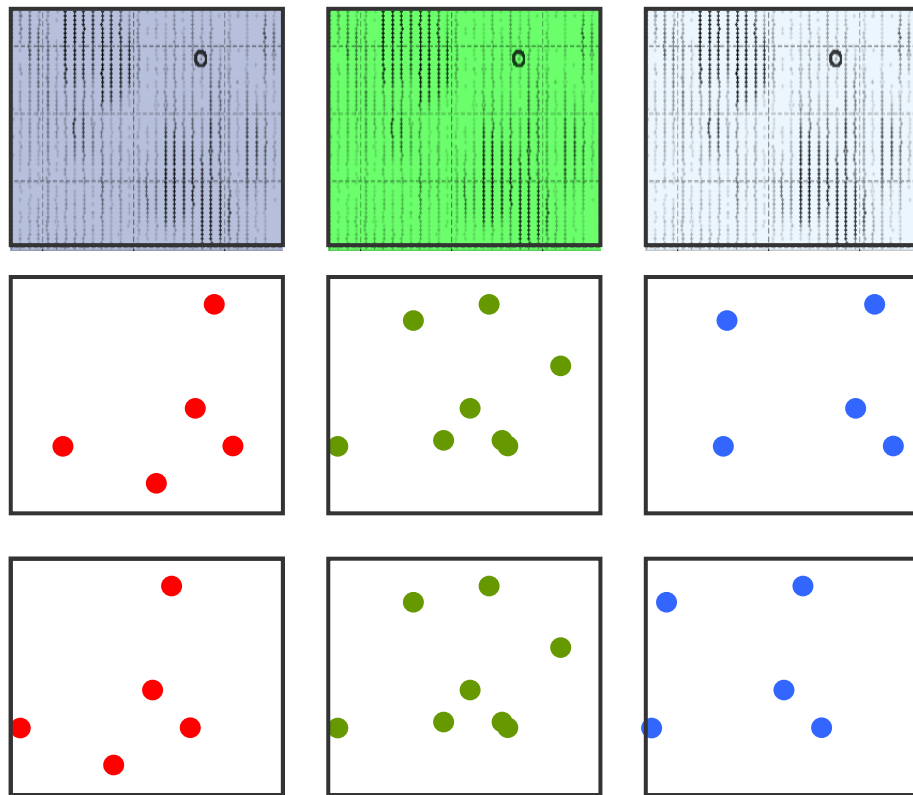MS intensity follows the chromatographic concentration

LC-MS Data (Map)

Quantification via "Features"
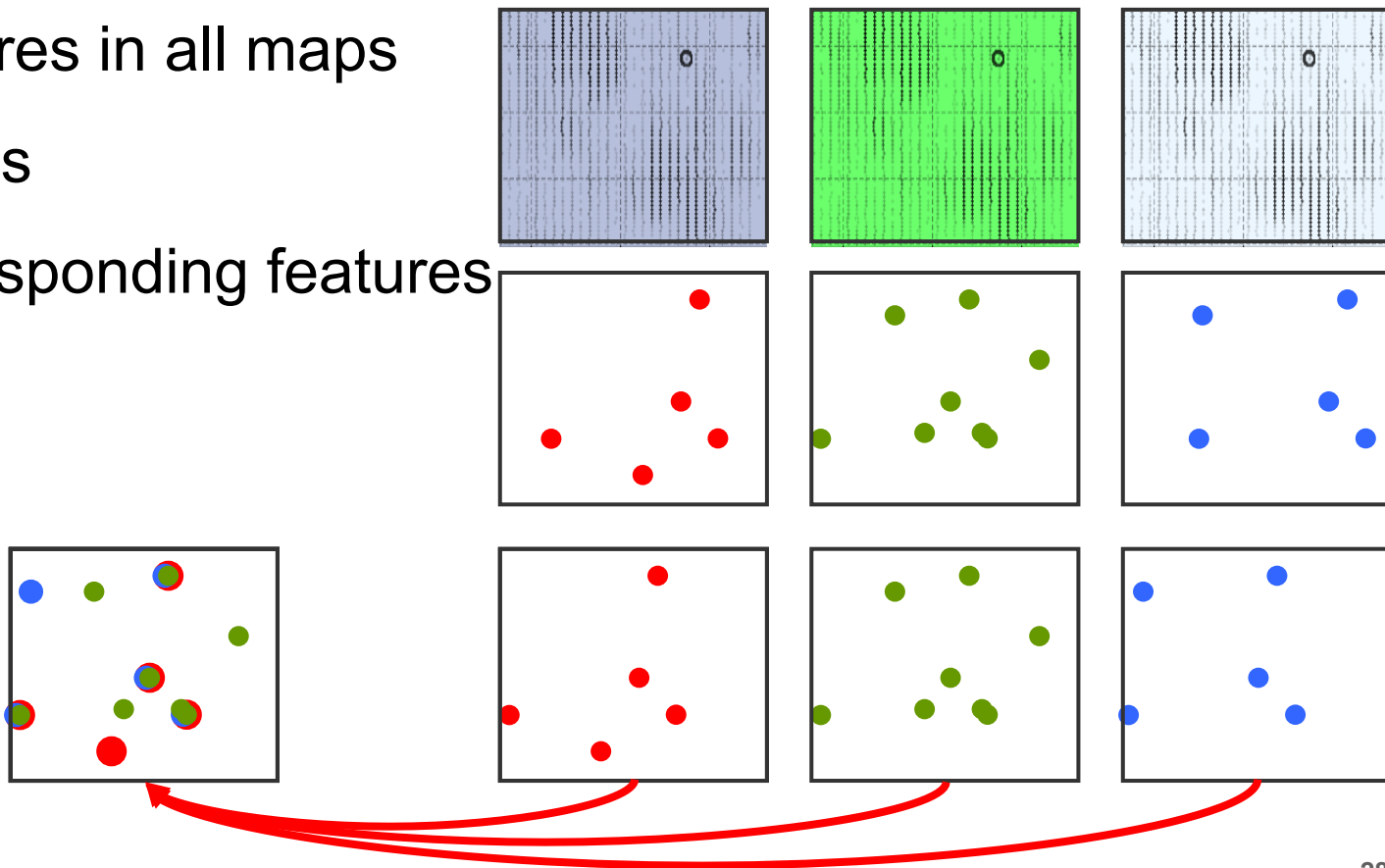(15 nmol/µl, 3x over-expressed, …)

LFQ – Analysis Strategy

1. **Find** features in all maps
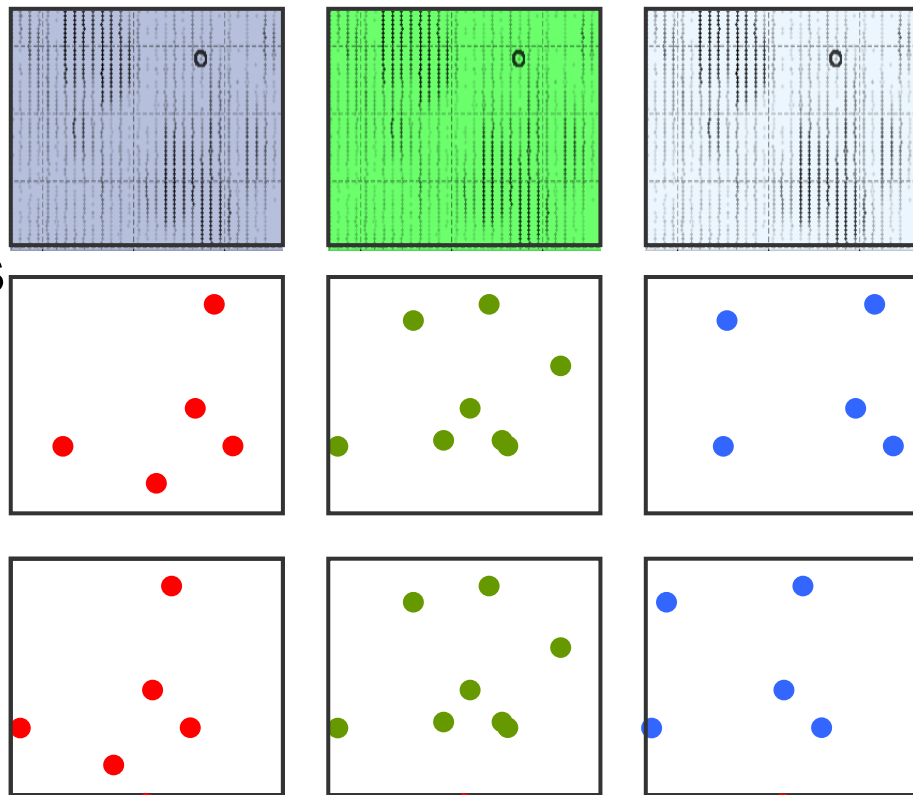
2. **Align** maps

LFQ – Analysis Strategy

1. **Find** features in all maps

2. **Align** maps

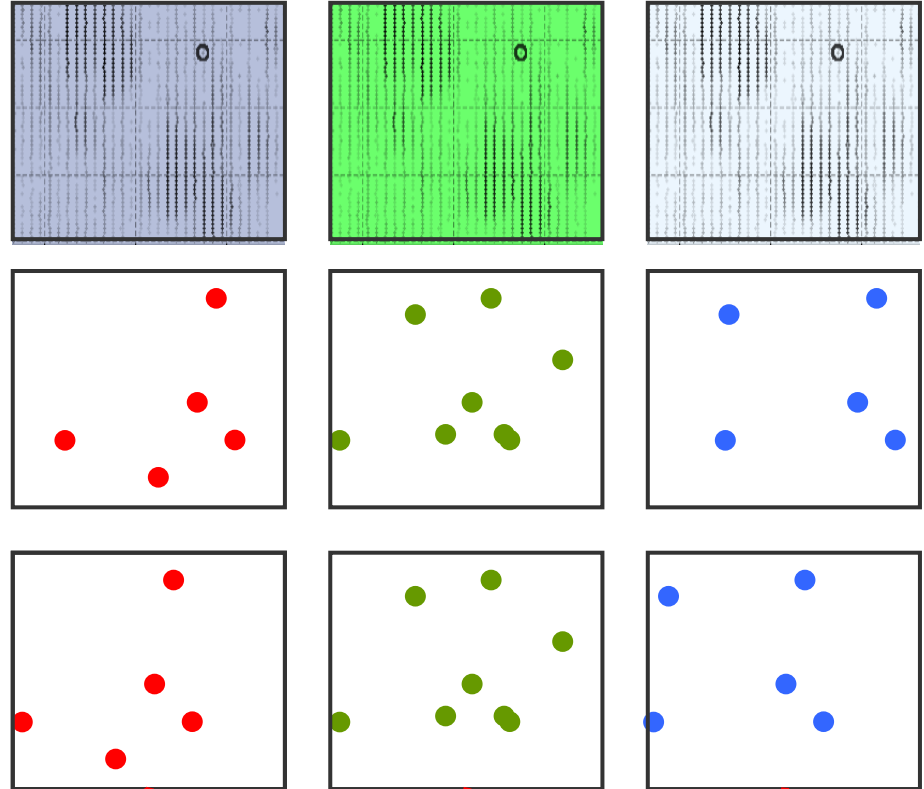3. **Link** corresponding features

LFQ – Analysis Strategy

1. **Find** features in all maps

2. **Align** maps

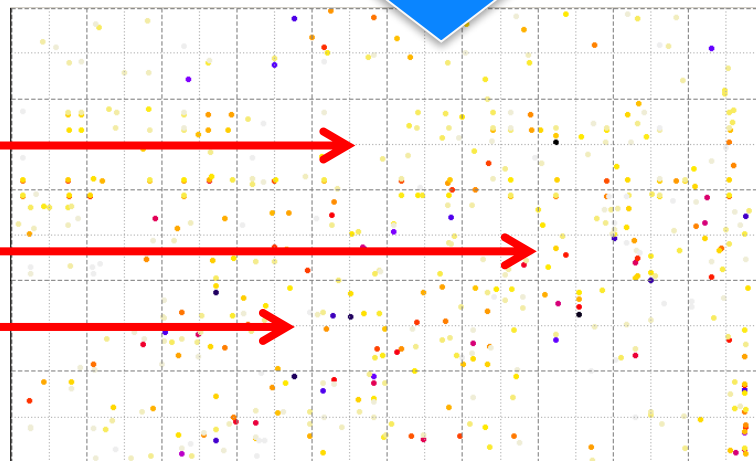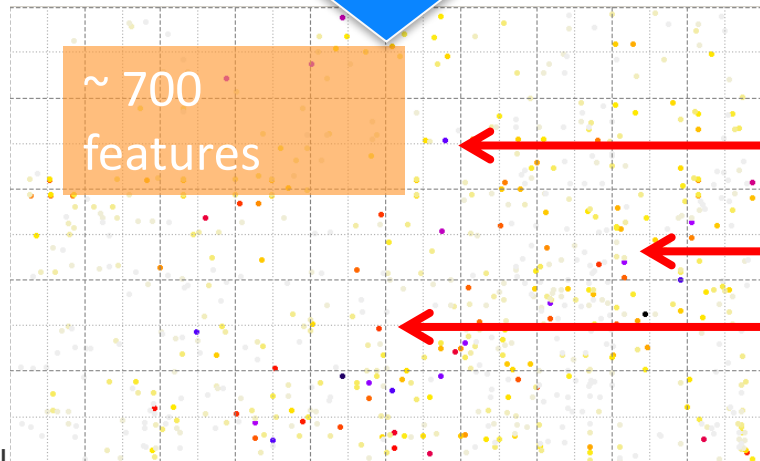3. **Link** corresponding features
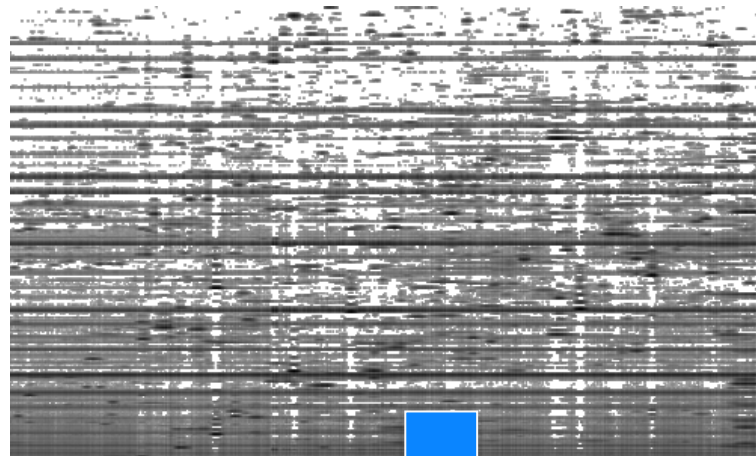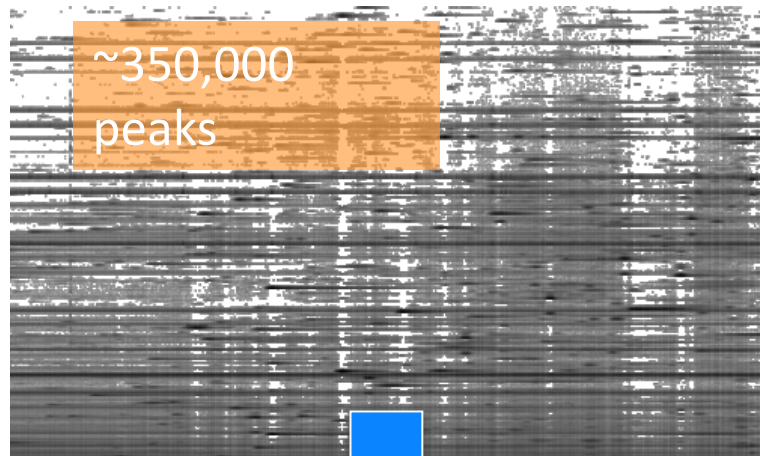
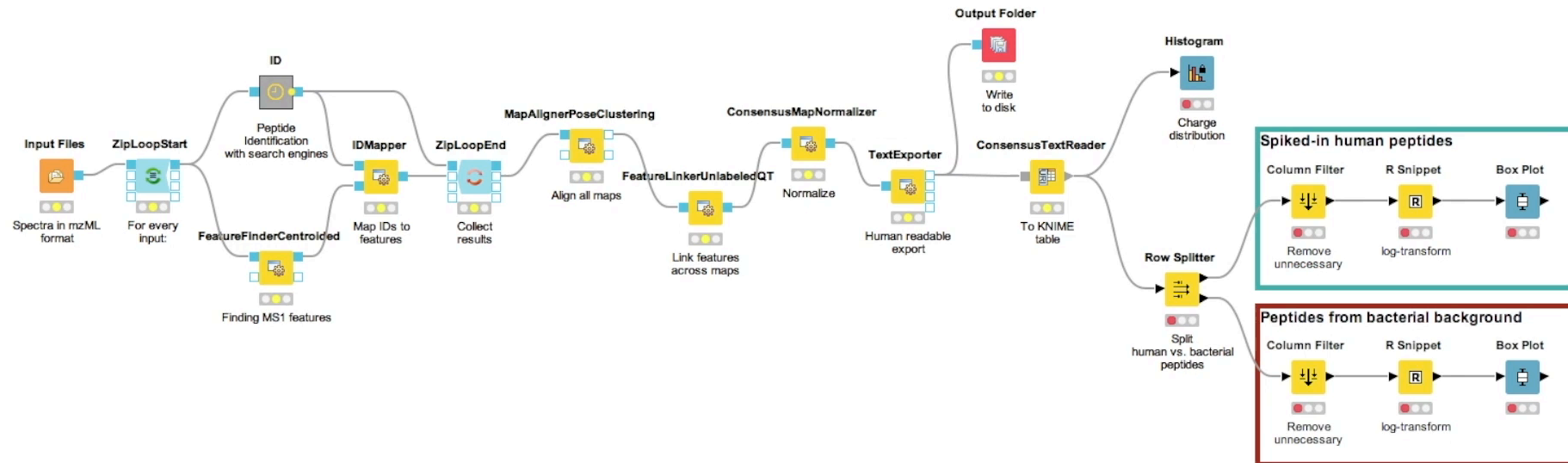4. **Identify** features

GDAFFGMSCK

LFQ – Analysis Strategy

1. **Find** features in all maps

2. **Align** maps

3. **Link** corresponding features

4. **Identify** features

5. **Quantify**

GDAFFGMSCK

1.0 : 1.2 : 0.5

Feature-Based Alignment

~350,000 peaks

~ 700 features

CIBI setup in projects

Ess-B▶A▶R▶
Biologische Gefahren, Analyse und
Resilienz der Lebensmittelwarenketten

BMBF project together
with KNIME, led by BfR

**Bioinformatics
Solution Center**

Permanent consulting positions
@FU Berlin for –omics data analysis

SEQAN   intel

Code modernization in IPCC
Vectorization and multicore support
of the whole SeqAn library

Thank you for your attention!
Tomorrow you can visit
**SeqAn and OpenMS (CIBI/de.nbi)
Integration Workshop
(1-Berlin-Dubai room)**