RUB

RUHR-UNIVERSITÄT BOCHUM

USING PROTEIN-GRAPHS TO GENERATE FASTA-DATABASES WITH VARIATIONALLY PEPTIDES

IOW: HOW TO ENABLE SEARCH-ENGINES TO SEARCH FOR VARIATIONAL PEPTIDES?

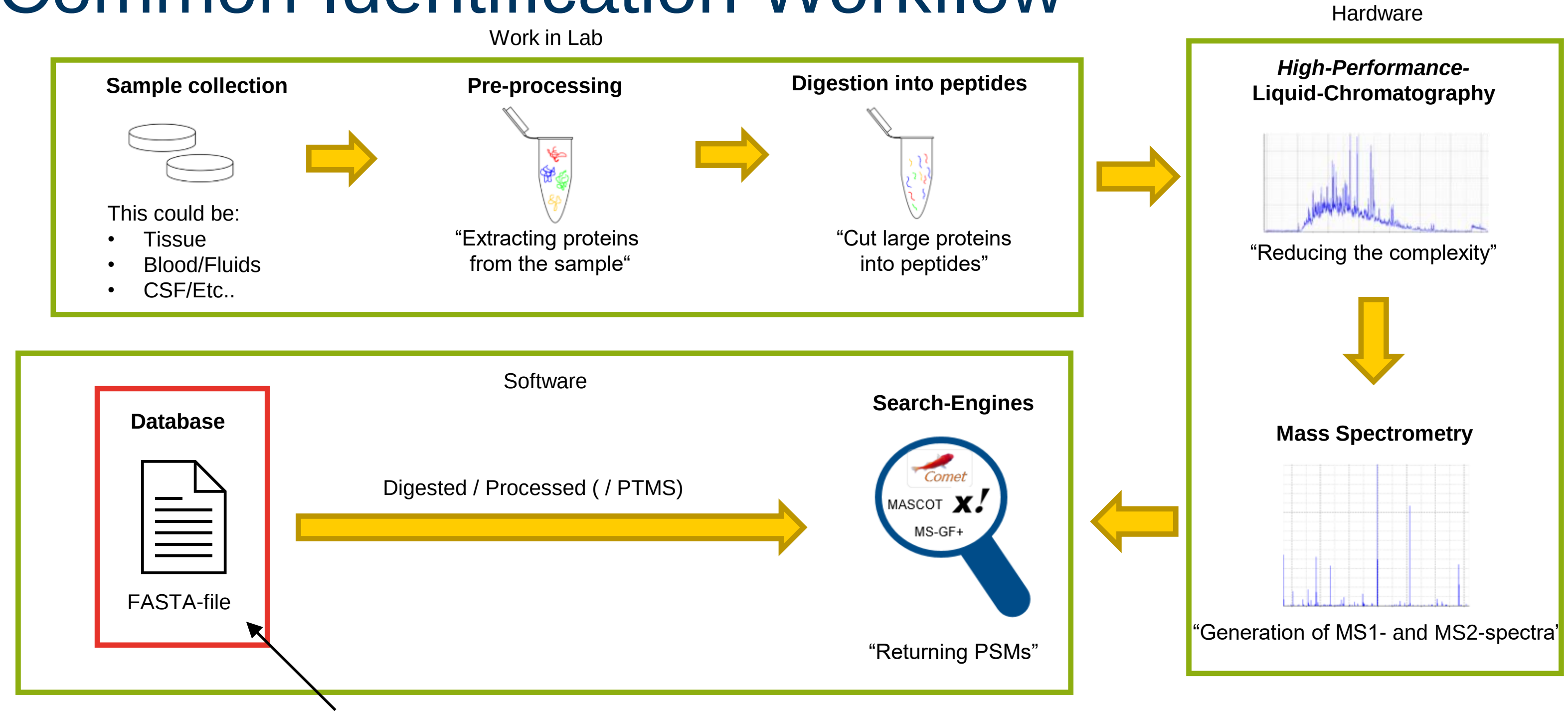
Dominik Lux

This is a graph!

↓

PRODI

Common Identification Workflow



Work in Lab

Sample collection

This could be:

- Tissue
- Blood/Fluids
- CSF/Etc..

Pre-processing

“Extracting proteins from the sample“

Digestion into peptides

“Cut large proteins into peptides”

Hardware

High-Performance- **Liquid-Chromatography**

“Reducing the complexity”

Mass Spectrometry

“Generation of MS1- and MS2-spectra”

Software

Database

FASTA-file

Digested / Processed (/ PTMS)

Search-Engines

“Returning PSMs”

A deeper look into FASTA!

FASTA-files and their potential!

Database

FASTA-file

- Usually contains only canonical sequences (sometimes isoforms, cRAP)
- “Used unprocessed” by search engines

```
…
>sp|ACCESSION|GENE_NAME Protein XYZ Some protein functions…
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
…
…
…
```

Example-Entry in FASTA

→ Large potential:

→ Include additional information (e.g., variants) → UniProt

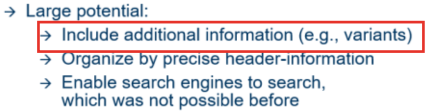
→ Organize by precise header-information → Unique/Shared-entries (by proteins)

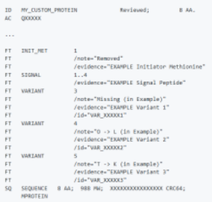
→ Enable search engines to search, which was not possible before →

- Search with “infinite many” miscleavages
- 2 (3 or more) digestion enzymes at once
- ….

→ Sophisticated FASTA-generator would be interesting!

Parsing and Encoding feature Information


→ Large potential:
→ Include additional information (e.g., variants)
→ Organize by precise header-information
→ Enable search engines to search, which was not possible before


```
ID   MY_CUSTOM_PROTEIN          Reviewed;        8 AA.
AC   QXXXXX

...

FT   INIT_MET        1
FT                   /note="Removed"
FT                   /evidence="EXAMPLE Initiator Methionine"
FT   SIGNAL          1..4
FT                   /evidence="EXAMPLE Signal Peptide"
FT   VARIANT         3
FT                   /note="Missing (in Example)"
FT                   /evidence="EXAMPLE Variant 1"
FT                   /id="VAR_XXXXX1"
FT   VARIANT         4
FT                   /note="O -> L (in Example)"
FT                   /evidence="EXAMPLE Variant 2"
FT                   /id="VAR_XXXXX2"
FT   VARIANT         5
FT                   /note="T -> K (in Example)"
FT                   /evidence="EXAMPLE Variant 3"
FT                   /id="VAR_XXXXX3"
SQ   SEQUENCE   8 AA;  988 MW;  XXXXXXXXXXXXXXXX CRC64;
     MPROTEIN
```

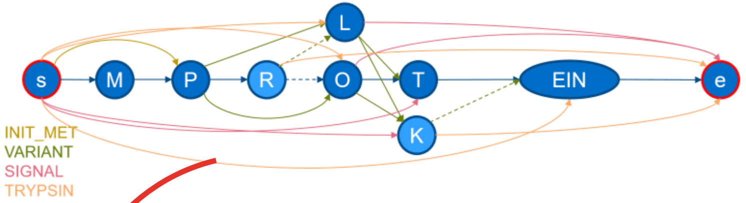
SP-EMBL (UniProt, species)

Contains:
- Canonical- (Isoform-) sequence
- Variants (/Mutagens/Conflicts)
- Peptides (Pro- Signalpeptides)
- …

ProtGraph
Available on:
- BioConda*
- PyPI*
- GitHub*

Python (CLI)

Produces many protein-graphs:
- Encodes feature information
- Allows (multiple) digestion(s)
- Allows to add PTMs
- …



1 protein-graph per entry
(this example contains 46 peptides)

?

Protein-Graph
- Feature-Information represented in graph-format
- Only contains valid paths from s to e
- Very compact representation

Database

FASTA-file

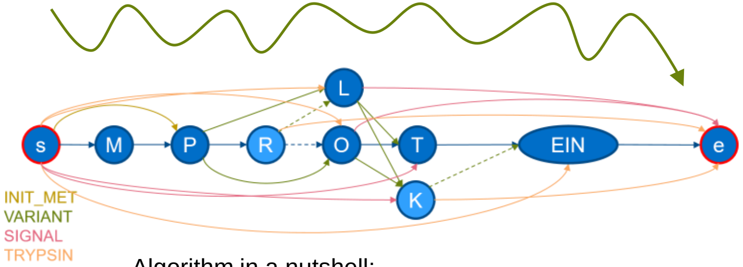
Next Step:

→ Convert Protein-Graphs into FASTA-entries

* conda install -c bioconda protgraph
* pip install protgraph
* https://github.com/mpc-bioinformatics/ProtGraph

Naïve Approach: Depth-First-Search

→ Large potential:
- → Include additional information (e.g., variants)
- → Organize by precise header-information
- → Enable search engines to search, which was not possible before



Algorithm in a nutshell:

- Begin at s
- Report every path to e
- Constrain by specific thresholds
  - Peptide-/Path-Mass
  - #Variants
  - #Miscleavages
  - …

(peptide-) Database

FASTA-file

```
...
>pg|ACCESSION|Protein ...
RPILTIITLEDSSGNLLGR
>pg| ACCESSION |Protein ...
TCPVQLWVDSTPPPGTR
>pg| ACCESSION |Protein ...
CSDSDGLAPPQHLIR
>pg| ACCESSION |Protein ...
KPLDGEYFTLQIR
>pg| ACCESSION |Protein ...
GEPHHELPPGSTK
...
```

A shared peptide between Protein A and B!

...
>pg|ACCESSION|**ProteinA(100:119, mssclvg:1), ProteinB(90:109, mssclvg:1)**
RPILTIITLEDSSGNLLGR

>pg|ACCESSION|**ProteinA(55:72, mssclvg:0, VARIANT[56:56, L->C]**
TCPVQLWVDSTPPPGTR
...

A unique peptide of Protein A with a variant

→ Organize headers by traversed path (concatenate same FASTA-entries)

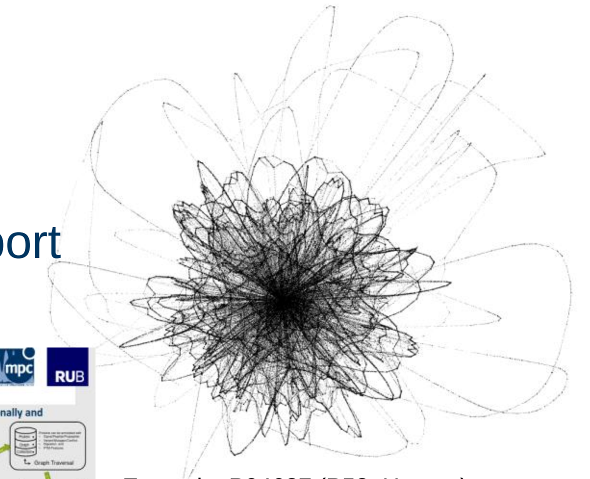
→ Search-Engines would need to search peptide-FASTAs as is (peptidomics)

→ IOW: “Digestion turned off”

Drawbacks of the Naïve Approach

→ “Some“ protein-graphs are too complex, for a naïve export

→ Too complex to save on disk (even in large scale!)



Example: P04637 (P53_Human)

- 1363 Variants
- 36 Mutagens
- 9 Isoforms
- 13988 Nodes, 93924 Edges

- Encodes 1.7E+224 peptides

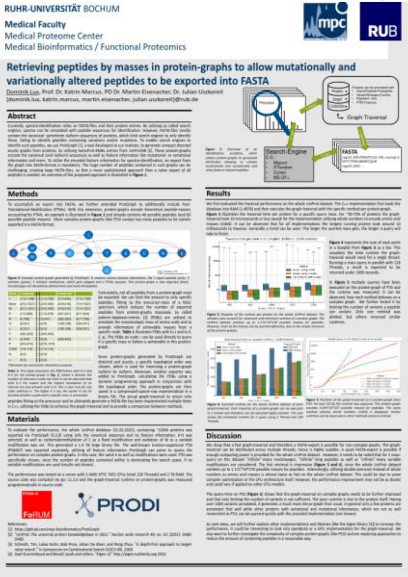
(computed via ProtGraph in a few seconds)

Solution:

- Search Engines do not need all entries
  - Export only peptides, fitting to the MS2-precursor

No details*,

just an overview how it was “solved”



RUHR-UNIVERSITÄT BOCHUM

Medical Faculty
Medical Proteome Center
Medical Bioinformatics / Functional Proteomics

Retrieving peptides by masses in protein-graphs to allow mutationally and variationally altered peptides to be exported into FASTA

Dominik Lux, Prof. Dr. Katrin Marcus, PD Dr. Martin Eisenacher, Dr. Julian Uszkoreit
(dominik.lux, katrin.marcus, martin.eisenacher, julian.uszkoreit)@rub.de

Abstract

[illegible]

Methods

[illegible]

Results

[illegible]

Materials

[illegible]

Discussion

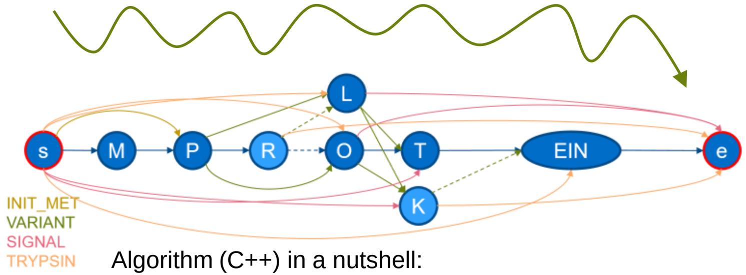
[illegible]

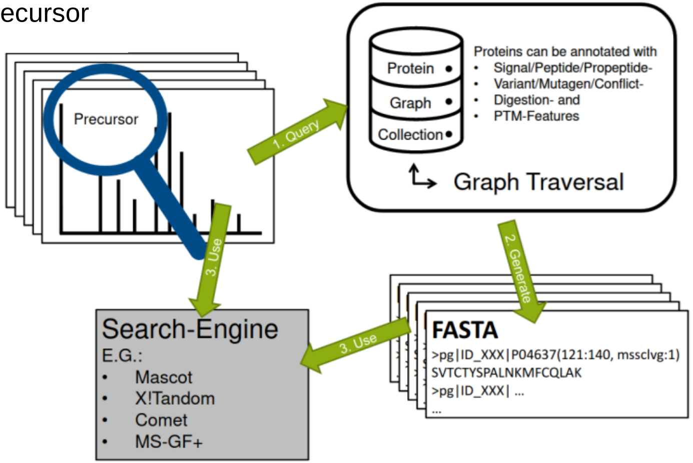
Poster presented at
Proteomic Forum 2022

* This is not a secret! If you are interested how it was “solved”, please ask!

Sophisticated Approach: Target-Value-Search

Problem: Return all Paths, where the mass of the peptide equals to the MS2-precursor
Target-Value: MS2-Precursor (+- Xppm)





Algorithm (C++) in a nutshell:

- Begin at s
- Use so-called PDBs* to traverse only through possible solution paths**
- Reach target e and report a peptide fitting the MS2-Precursor
- Repeat until all solutions are reported
- Constrain by:
  - #Variants

→ Executable on all proteins! (whole UniProtKB, ~230M Proteins)

→ Only a few Proteins, have long running time during traversal***

* Pattern databases, Schmidt et al, A depth first search approach to target value search

** Early expansion-prevention in branches with no solutions

*** Guess: ~200 proteins in the whole UniProtKB

Generation of peptide-FASTA-Databases

ISA*-RAW-files containing 28123 distinct MS2-Precursors (+/-5ppm, Oxidation M (variable), Carbamidomethylation C (fixed))

Organism:	E.Coli	Mus Musculus	~1000 Species (excluding homo sapiens)
# Proteins:	4448	55319	11 973 189
Restrictions:	None	None	None
(peptide-)FASTA Generation**			
(peptide-) Database # Entries	66 680 808 (peptides) (21 GB)	85 702 533 (peptides) (46 GB)	126 462 579 (peptides) (27 GB)
Generation time:	1 h 30m	3 h 54m	1d 14h 4m

Any other species

(probably feasible)

(if no very complex protein-graphs are present)

→ Generation of a MS2-specific-peptide FASTA with features*** is feasible

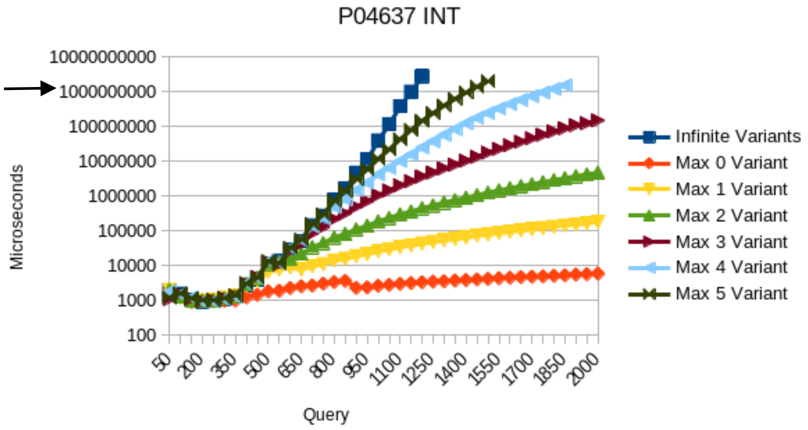
* Internal Standard (for benchmarking)
** Generated on a server with 64 threads
*** All features ProtGraph can parse

What about Homo Sapiens?

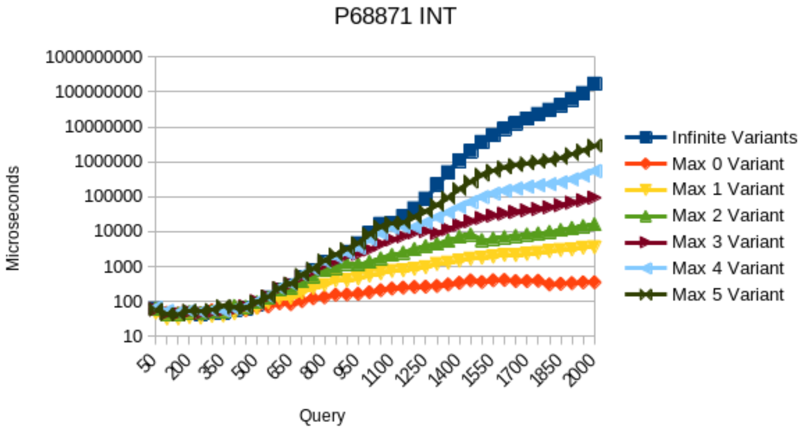
→ Most well researched species (→ most annotated proteins)

Complex protein: P04637 (P53_Human)
Benchmark/Target-Value: Every 50 Dalton
Configuration: All Features, Digested, PTMs:
Carbamidomethylation of C (fixed)
Oxidation of M (variable)

Complex protein: P68871 (HBB_Human)
Benchmark/Target-Value: Every 50 Dalton
Configuration: All Features, Digested, PTMs:
Carbamidomethylation of C (fixed)
Oxidation of M (variable)

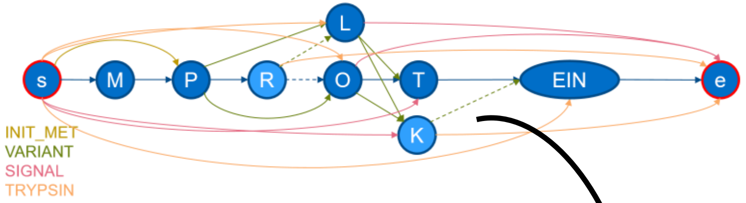


1000 seconds →



Summary

ProtGraph



- Generation of protein-graphs via ProtGraph
- Protein-graphs can contain up to infinite
    - Miscleavages
    - Features (Variants/Peptide/Signal-/Propeptide/Mutagens …)
    - Digestion enzymes

→ Can be exported into a peptide-FASTA-database

→ Search-Engines benefit from it!

→ see above

→ Easier interpretation with the new header-format

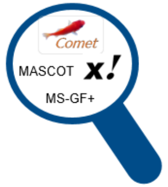
feasible

(peptide-)Database

FASTA-file

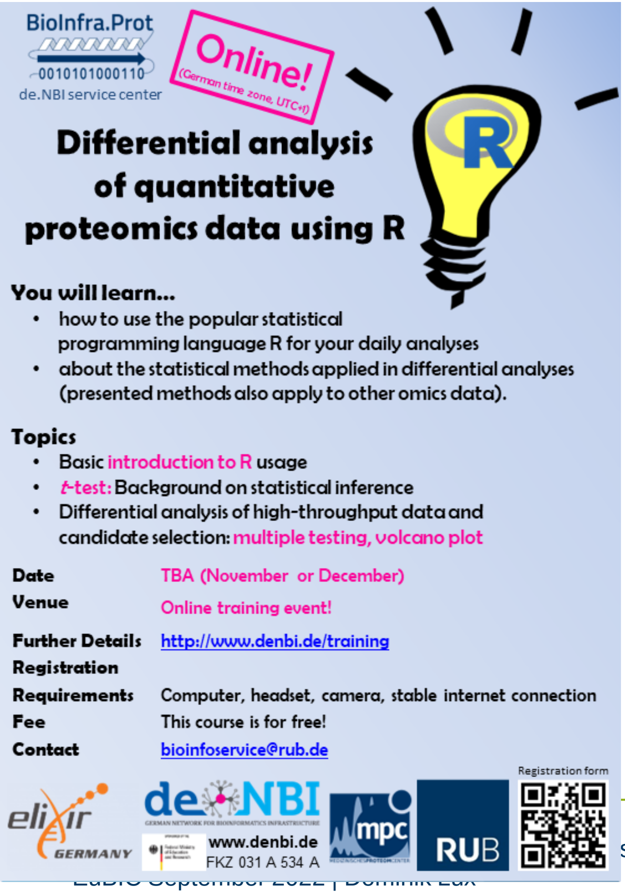
“search as is” (no digestion)

Search-Engines



- Unique/shared PSMs
- (trivial) Inference
- More precise information (how the peptide was generated)

Acknowledgement



BioInfra.Prot

0010101000110

de.NBI service center

Online!

(German time zone, UTC+1)

Differential analysis of quantitative proteomics data using R

You will learn...

- how to use the popular statistical programming language R for your daily analyses
- about the statistical methods applied in differential analyses (presented methods also apply to other omics data).

Topics

- Basic introduction to R usage
- *t*-test: Background on statistical inference
- Differential analysis of high-throughput data and candidate selection: multiple testing, volcano plot

Date	TBA (November or December)
Venue	Online training event!
Further Details	http://www.denbi.de/training
Registration	
Requirements	Computer, headset, camera, stable internet connection
Fee	This course is for free!
Contact	bioinfoservice@rub.de

Registration form

elixir GERMANY

de.NBI

www.denbi.de
FKZ 031 A 534 A

mpc

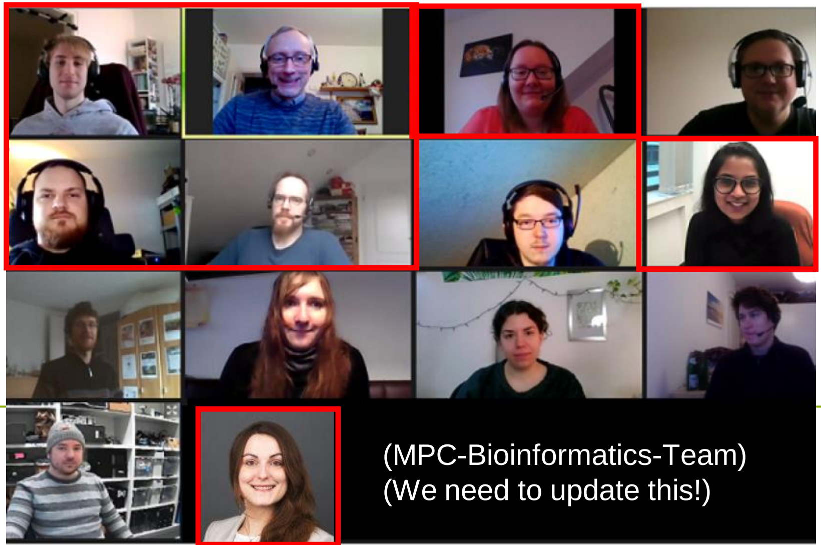
RUB

MPC-Bioinformatics-Team:
PD Dr. Martin Eisenacher
Dr. Julian Uszkoreit
Dirk Winkelhardt
Karin Schork
Sai Spoorti Ramesh
Maike Weber

And Thank you for your attention!

@ProteomCenter
@lululuxii

mpc-bioinformatics
Luxxii



(MPC-Bioinformatics-Team)
(We need to update this!)