Welcome
to the DKFZ!

GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

News from PCF at DKFZ

Martin Schneider

Topics from DKFZ

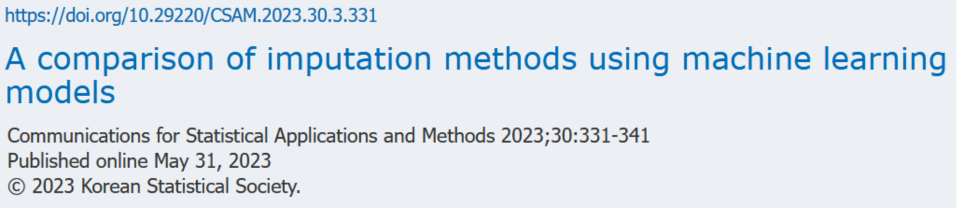
- missRanger
- AlphaPept on Linux
- New Software how do you test?
- The field needs to talk about FDR
- Additional thoughts

missRanger

- missForest

  Stekhoven, D.J. and Buehlmann, P. (2012),
  'MissForest - nonparametric missing value imputation for mixed-type data’,
  Bioinformatics, 28(1) 2012, 112-118, doi: 10.1093/bioinformatics/btr597

- Statistics benchmarks



https://doi.org/10.29220/CSAM.2023.30.3.331

A comparison of imputation methods using machine learning models

Communications for Statistical Applications and Methods 2023;30:331-341
Published online May 31, 2023
© 2023 Korean Statistical Society.

We can see that missForest takes much longer to impute missing values than missRanger, while there is not much difference in performances.

missRanger

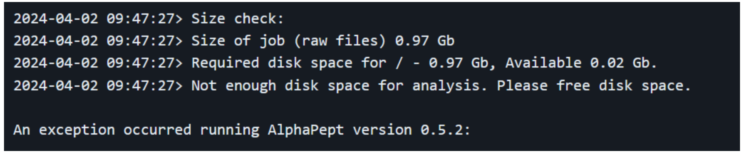
- Results constantly differ in between missRanger and missForest
- missRanger based on ground truth performs worse
- missForest now used with only 20 trees performs equivalent and is much faster

AlphaPept on Linux

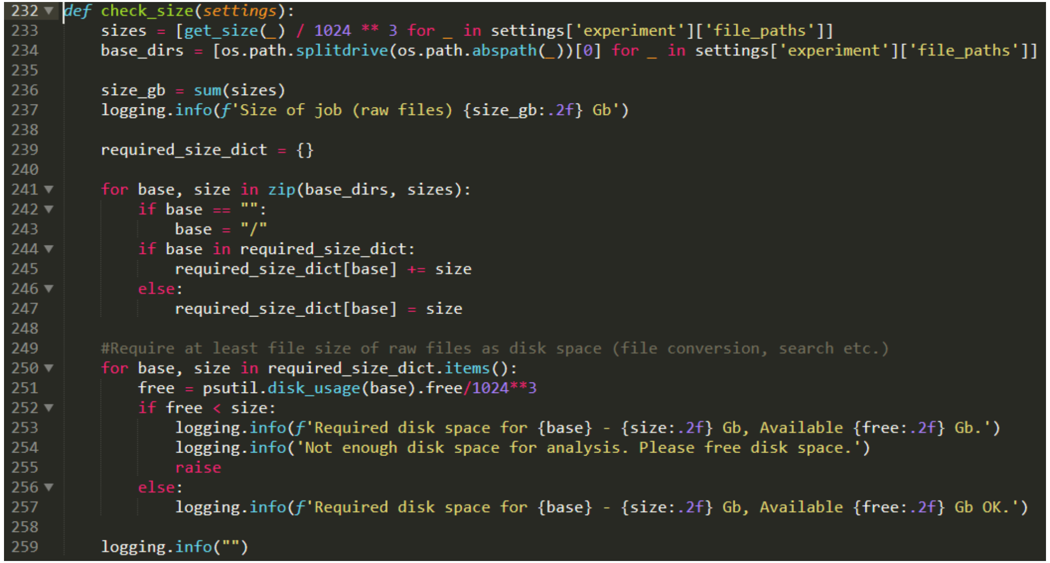
- DKFZ requires local QC solution
- ShinyQC with MaxQuant ~ 60 min
- ShinyQC 2.0 with AlphaPept
  - Fast ~ 8 min
  - Fully open source
  - Fails to run on Linux
  - Some MS level information missing

AlphaPept on Linux

- Fails to run on Linux
- Docker for ARM from scratch for AMD64
- Not enough disk space



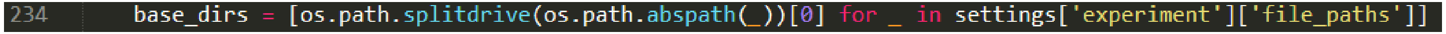
```
2024-04-02 09:47:27> Size check:
2024-04-02 09:47:27> Size of job (raw files) 0.97 Gb
2024-04-02 09:47:27> Required disk space for / - 0.97 Gb, Available 0.02 Gb.
2024-04-02 09:47:27> Not enough disk space for analysis. Please free disk space.

An exception occurred running AlphaPept version 0.5.2:
```



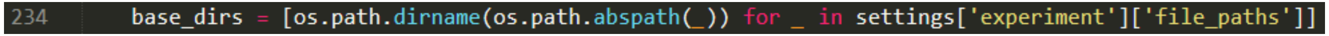
```python
def check_size(settings):
    sizes = [get_size(_) / 1024 ** 3 for _ in settings['experiment']['file_paths']]
    base_dirs = [os.path.splitdrive(os.path.abspath(_))[0] for _ in settings['experiment']['file_paths']]

    size_gb = sum(sizes)
    logging.info(f'Size of job (raw files) {size_gb:.2f} Gb')

    required_size_dict = {}

    for base, size in zip(base_dirs, sizes):
        if base == "":
            base = "/"
        if base in required_size_dict:
            required_size_dict[base] += size
        else:
            required_size_dict[base] = size

    #Require at least file size of raw files as disk space (file conversion, search etc.)
    for base, size in required_size_dict.items():
        free = psutil.disk_usage(base).free/1024**3
        if free < size:
            logging.info(f'Required disk space for {base} - {size:.2f} Gb, Available {free:.2f} Gb.')
            logging.info('Not enough disk space for analysis. Please free disk space.')
            raise
        else:
            logging.info(f'Required disk space for {base} - {size:.2f} Gb, Available {free:.2f} Gb OK.')

    logging.info("")
```

AlphaPept on Linux

- Fix check_size



```
    base_dirs = [os.path.splitdrive(os.path.abspath(_))[0] for _ in settings['experiment']['file_paths']]
```



```
    base_dirs = [os.path.dirname(os.path.abspath(_)) for _ in settings['experiment']['file_paths']]
```

- Additional info on MS level

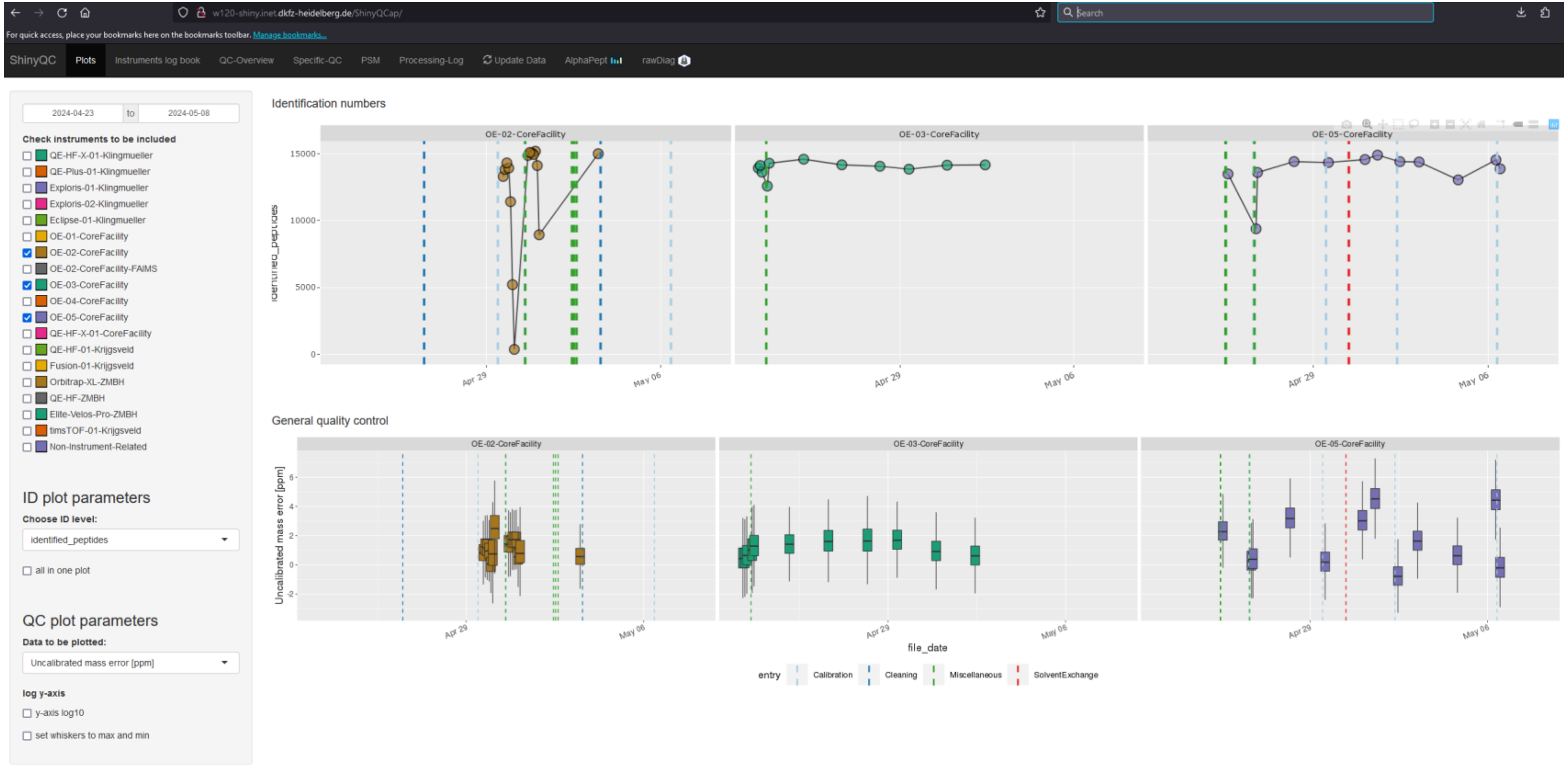
- Info on LC level
  - Thermo Raw File Reader
  - Own EXE for extracting LC info

AlphaPept on Linux

- TIC, Injection Time, und AGC-Fill for MSMS
  - rawrr / rawDiag crashes
  - Table for all individual scans too large

- Thermo Raw File Reader
  - Own implementation to extract
  - Visual Studio 2022 to use required NuGet
  - Visual Studio Code did not get it to work
  - Got experience how to run it with local NuGet?

AlphaPept on Linux



w120-shiny.inet.dkfz-heidelberg.de/ShinyQCap/

For quick access, place your bookmarks here on the bookmarks toolbar. Manage bookmarks...

2024-04-23 to 2024-05-08

Check instruments to be included

- □ QE-HF-X-01-Klingmueller
- □ QE-Plus-01-Klingmueller
- □ Exploris-01-Klingmueller
- □ Exploris-02-Klingmueller
- □ Eclipse-01-Klingmueller
- □ OE-01-CoreFacility
- ☑ OE-02-CoreFacility
- □ OE-02-CoreFacility-FAIMS
- ☑ OE-03-CoreFacility
- □ OE-04-CoreFacility
- ☑ OE-05-CoreFacility
- □ QE-HF-X-01-CoreFacility
- □ QE-HF-01-Krijgsveld
- □ Fusion-01-Krijgsveld
- □ Orbitrap-XL-ZMBH
- □ QE-HF-ZMBH
- □ Elite-Velos-Pro-ZMBH
- □ timsTOF-01-Krijgsveld
- □ Non-Instrument-Related

ID plot parameters

Choose ID level: identified_peptides

□ all in one plot

QC plot parameters

Data to be plotted: Uncalibrated mass error [ppm]

log y-axis

□ y-axis log10

□ set whiskers to max and min

New Software how do you test?

- DIA-NN 1.9.2
- Spectronaut 19.4.241104.62635
- MSAID 1.3.2 (CHIMERYS)
- AlphaDIA 1.8.2
- AlphaPept 0.5.0
- FragPipe 22.0

The field needs to talk about FDR

- DIA-NN 1.9.2 → FDR control
  - …that ensures that each neural network in an ensemble is only used for prediction on samples it has not been trained on.
- Spectronaut default:
  - false discovery rate not in line with reality
- target-decoy; mathematical FDR?; machine learning FDR?

- FDR should match reality
- Which experiment, fading of ratios etc.?

Additional thoughts

- Based on washes tests most abundant peptides are carried over
- Custom internal heavy standard?
  - QC for individual samples (different concentrations of synthetic peptides)
  - Distributed over RT
  - Normalization based on standard

Thank you for your attention!

Further information on www.dkfz.de

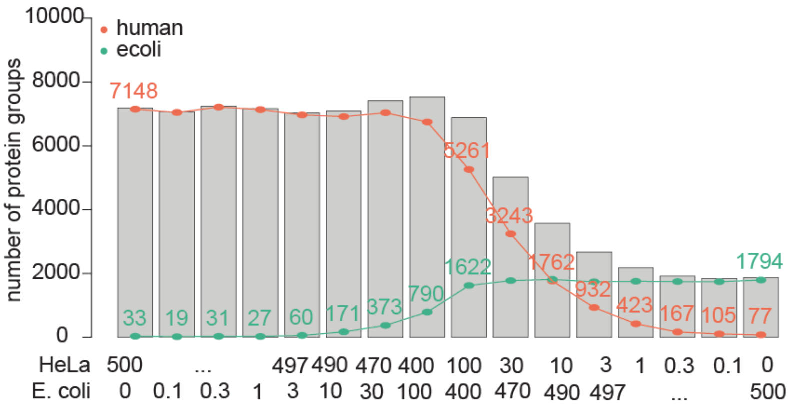
dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

Optimization HYE, fade-in fade-out, thoughts

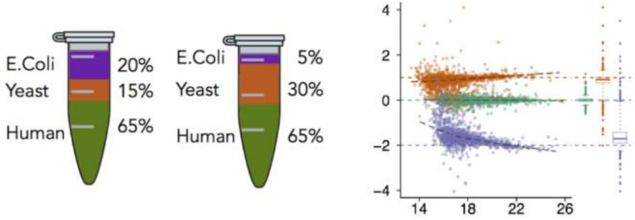
- What is carried over?

- How much is in the washes?

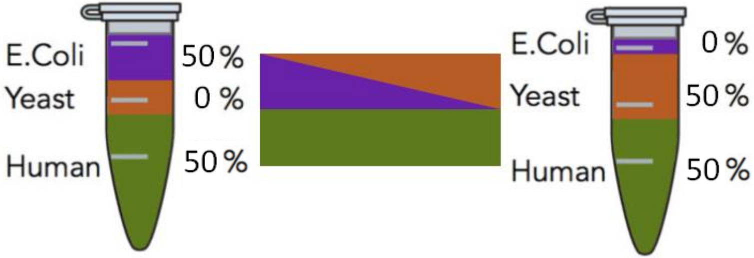


Separate analysis, default settings (Method Evaluation)

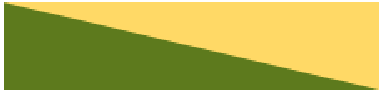
Optimization HYE, fade-in fade-out, thoughts



- HYE with dilution series



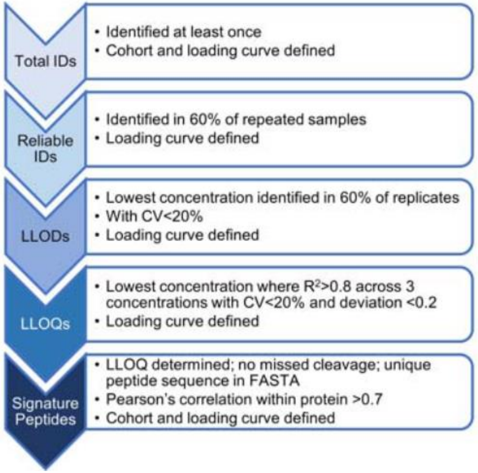
- Human with A. thaliana dilution



Optimization HYE, fade-in fade-out, thoughts

Paradigm shift in biomarker translation: a pipeline to generate clinical grade biomarker candidates from DIA-MS discovery

[1]Qin Fu, [1]Manasa Vegesna, [1]Niveda Sundararaman, [2]Eugen Damoc, [2]Tabiwang N. Arrey, [2]Anna Pashkova, [1]Emebet Mengesha, [1]Philip Debbas, [1]Sandy Joung, [1]Dalin Li, [1]Susan Cheng, [1]Jonathan Braun, [1]Dermot P.B. McGovern, [1*]Christopher Murray, [2*]Yue Xuan, and [1#]Jennifer E. Van Eyk



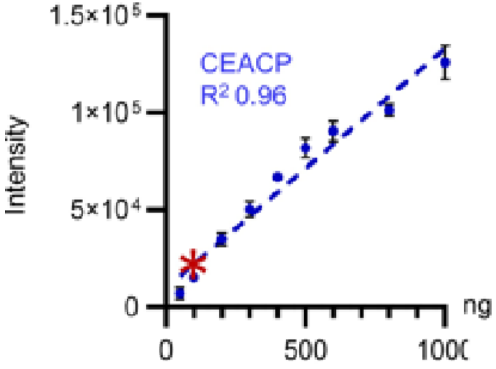
Total IDs
- Identified at least once
- Cohort and loading curve defined

Reliable IDs
- Identified in 60% of repeated samples
- Loading curve defined

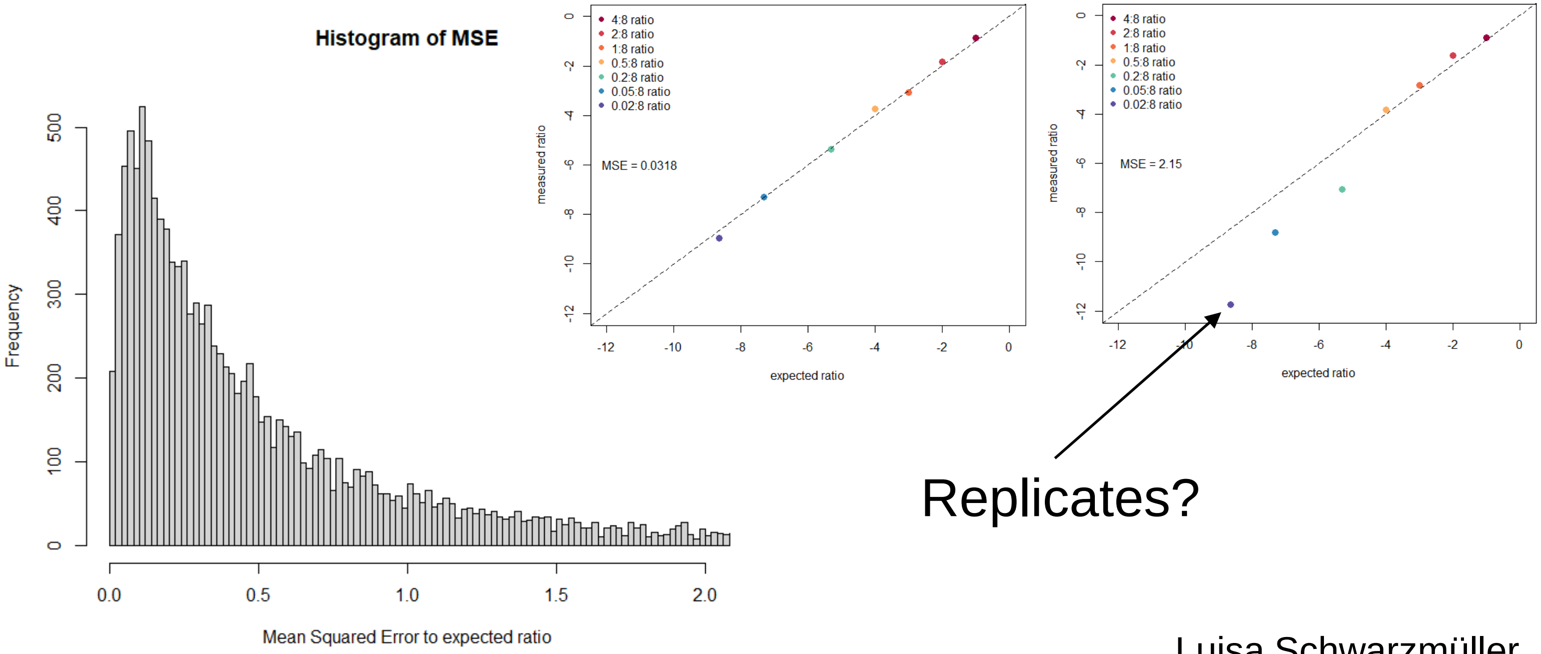
LLODs
- Lowest concentration identified in 60% of replicates
- With CV<20%
- Loading curve defined

LLOQs
- Lowest concentration where R^2>0.8 across 3 concentrations with CV<20% and deviation <0.2
- Loading curve defined

Signature Peptides
- LLOQ determined; no missed cleavage; unique peptide sequence in FASTA
- Pearson's correlation within protein >0.7
- Cohort and loading curve defined



Optimization HYE, fade-in fade-out, thoughts



Replicates?

Luisa Schwarzmüller
luisa.schwarzmueller@dkfz.de

Optimization HYE, fade-in fade-out, thoughts

- Info on all proteins in different abundance
- Pseudo ground truth → they should be on the line
- Feed models HMM, random forest, deep learning
- Which parameter combination good proxy for quality of:
  - Quantification
  - Detection