CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

David Potěšil
Kristína Gömöryová

Anna Schneiderová
Pavel Hruška
Ranjani Ganji

2019-11-20

Processing of proteomics data in KNIME

Workshop outline

- morning session – theoretical part

10:00 – 10:15	Opening and introduction
10:15 – 10:45	Software container running KNIME
10:45 – 11:00	Coffee break
11:00 – 11:30	Introduction to KNIME
11:30 – 11:45	Coffee break
11:45 – 12:30	Practical applications, our KNIME metanodes
12:30 – 13:30	Lunch break, visit of our laboratories for interested people

Workshop outline

- afternoon session – practical part
  - 13:30 – 16:30 KNIMing and coffee breaks
  - 16:30 – 18:00 Discussion

Workshop organization remarks

- morning session
  - presentation slides will be provided
  - keep your questions to the end of each presentation
- afternoon session
  - virtual workspaces provided to you by us will be kept for you till 31st January 2020
    - changes in the network settings will be necessary though (different IP::port settings)
    - let us know if you would need prolongment
  - data and workflows presented will be provided

Workshop organization remarks

- everyone has working WiFi connection?

- who does not have working environment accessible and wants to have one?

Workshop outline

- morning session – theoretical part
  - 10:00 – 10:15 Opening and introduction
  - **10:15 – 10:45 Software container running KNIME**
  - 10:45 – 11:00 Coffee break
  - 11:00 – 11:30 Introduction to KNIME
  - 11:30 – 11:45 Coffee break
  - 11:45 – 12:30 Practical applications, our KNIME metanodes
  - 12:30 – 13:30 Lunch break, visit of our laboratories for interested people

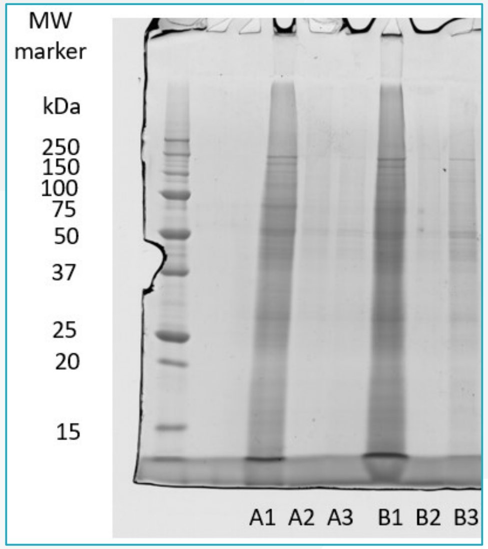
0) General proteomics study

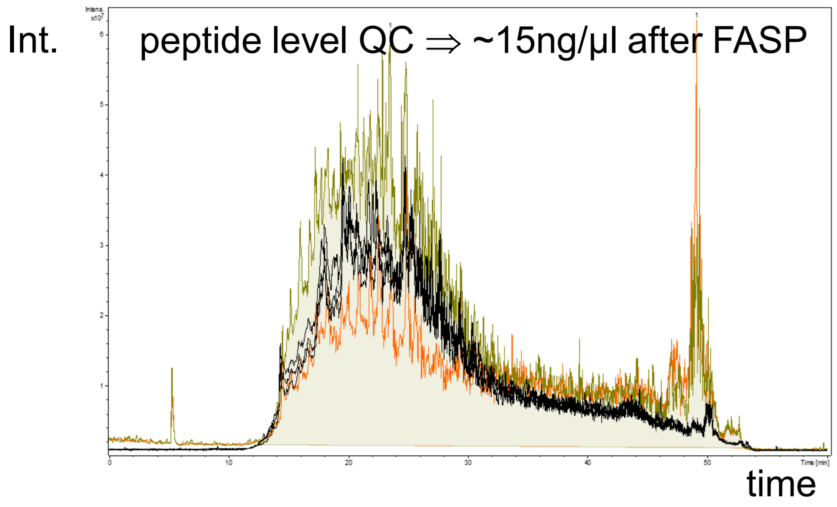
General proteomics study in our core facility

- **email** in our inbox
  - “Hello, we are studying very interesting **protein A** and would love to know **how it will look like if we will get rid of it in our cell line**.
- **discussion** about the study, **experiment design** specification
  - two types of cell line
    - **WT** – wild type
    - **KO** – protein A knocked out
  - **4 replicates**
    - prepared all in parallel
- samples **processing plan** on our side, **instrument booking**
- samples **preparation on the customer side**

General proteomics study in our core facility

- samples **processing in our laboratory**
  - e.g. cells lysis (proteins solubilization)
  - **protein level quality control (QC)** step (1D SDS-PAGE; semi quant.)
  - FASP (proteins ⇒ peptides)
  - **peptide level QC** (LC-UV-MS; semi quant.)
  - peptides transfer into the LC-MS vial prior the LC-MS/MS measurement

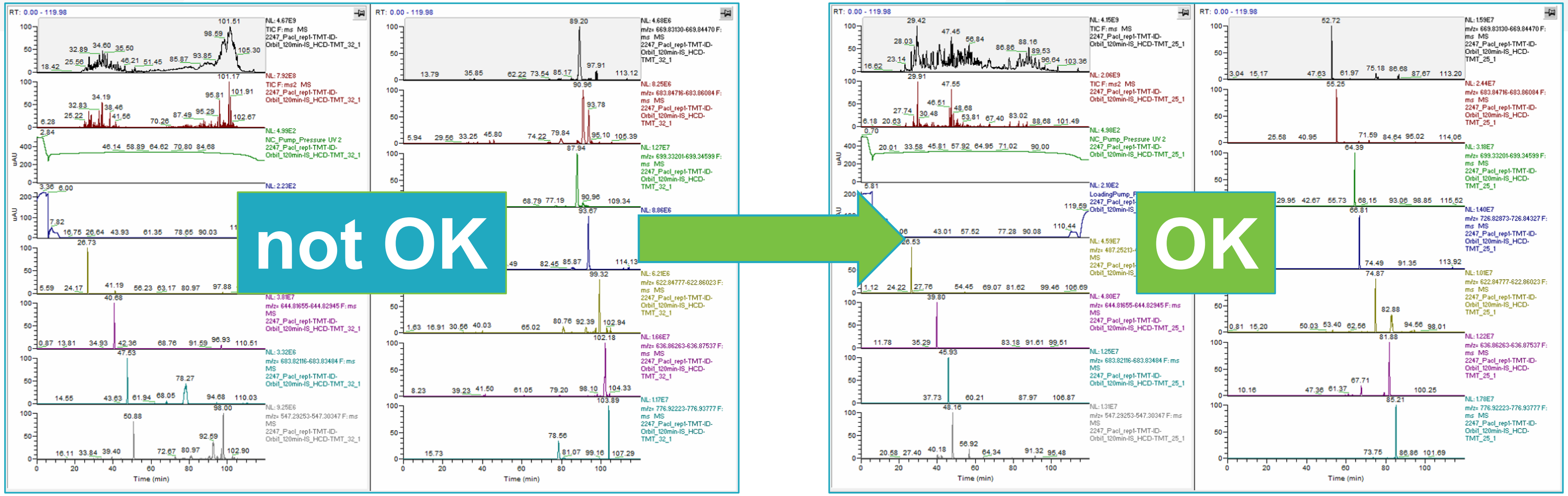



peptide level QC ⇒ ~15ng/µl after FASP

General proteomics study in our core facility

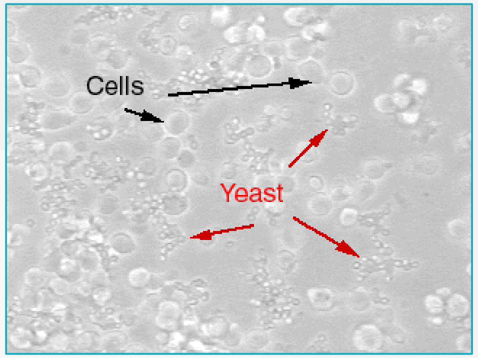
- final **LC-MS analyses** of the resulting peptide mixtures
  - potential issues due to samples matrix and or technical problems with the used instruments
    - e.g. residual detergents affecting the peptides LC separation, “dirty” mass spectrometer
    - iRT injected together with each sample, checked for retention and intensity profiles

  - partial samples reprocessing might be sometimes necessary with another round of the LC-MS analyses



General proteomics study in our core facility

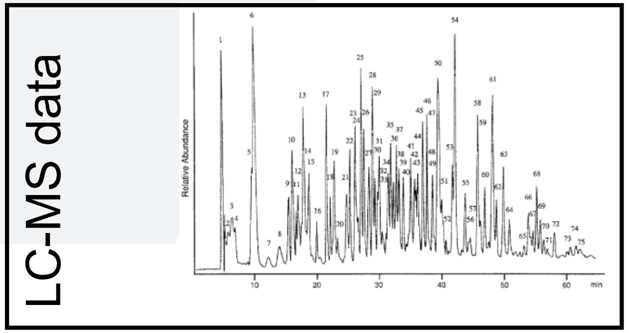
- **QC database searches**
  - potential issues with **protein type contaminations**
    - e.g. FBS, bacterial contamination
    - QC database searches to check on any potential samples contamination
  - potential issues with **not expected peptide or protein level modifications**
    - e.g. partially digested sample (not specific peptides)
    - QC database searches to check presence of not expected modifications

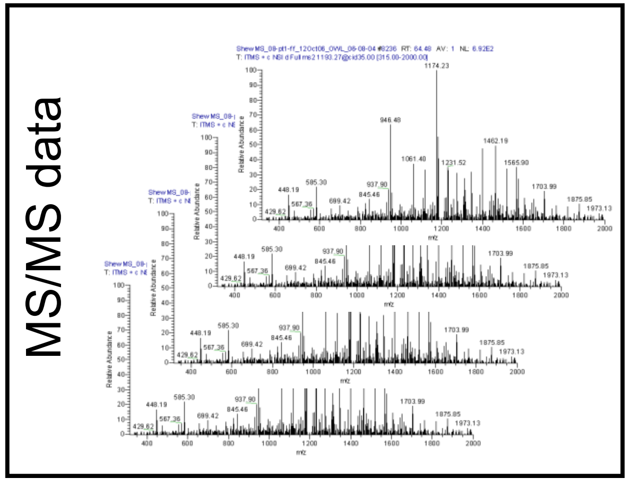


  - adjustment of the final database search conditions if needed
  - **sometimes critical issue** ⇒ new samples preparation

General proteomics study in our core facility

- final **database searches** and **protein quantification**
  - peptides identification, proteins list generation and peptides and proteins quantification
    - including LC-MS signal, not MS/MS based quant., e.g. MBR
    - more and more complex approaches
  - **potential issues** affecting the results worth to check
    - varying or generally too low identification rates
    - separation specific issues (varying peptides LC-MS peaks widths and or retention times)

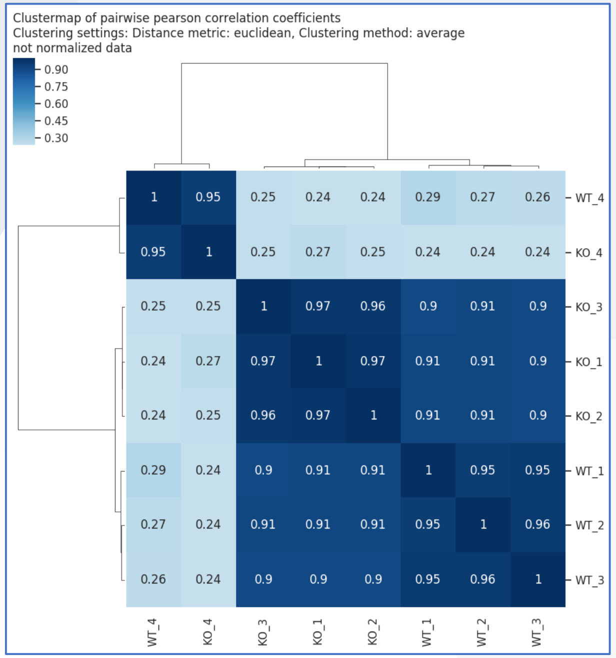




PG	Accession	Protein intensity					
		A1	A2	A3	B1	B2	B3
1	P12345	20	18	19	12	15	13
2	P23456	28	24	23	0	0	0
3	P34567	18	17	15	16	19	18
4	P45678	0	0	0	0	0	0
5	P56789	0	0	0	20	24	20
…	…	…	…	…	…	…	…

General proteomics study in our core facility

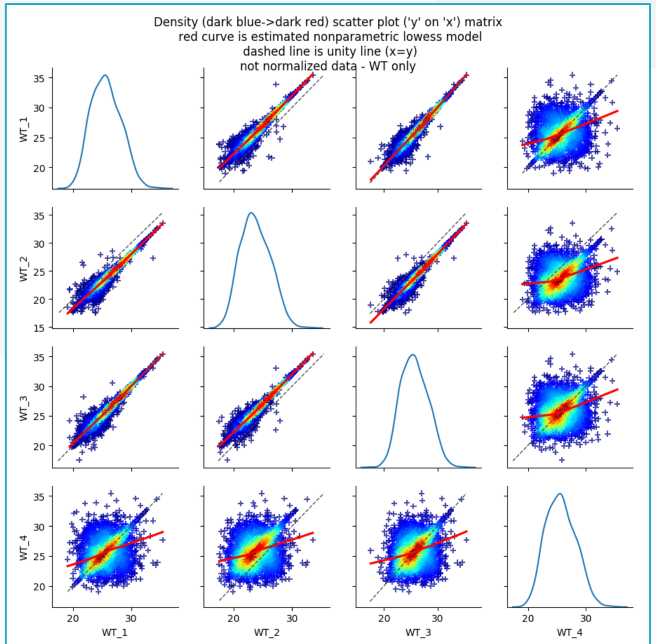
- proteins **quantification data preprocessing**
  - proteins list filtering
  - transformation, normalization
  - missing data imputation
  - + another level of QC steps (cluster analysis, correlation plots inspection, ...)
    - **get to know the data prior statistics...**
  - **potential issues** observed/affecting the results
    - not expected design “features” – **paired design**, **batch effects**
    - **outlying sample replicates**



Clustermap of pairwise pearson correlation coefficients
Clustering settings: Distance metric: euclidean, Clustering method: average
not normalized data

cluster analysis

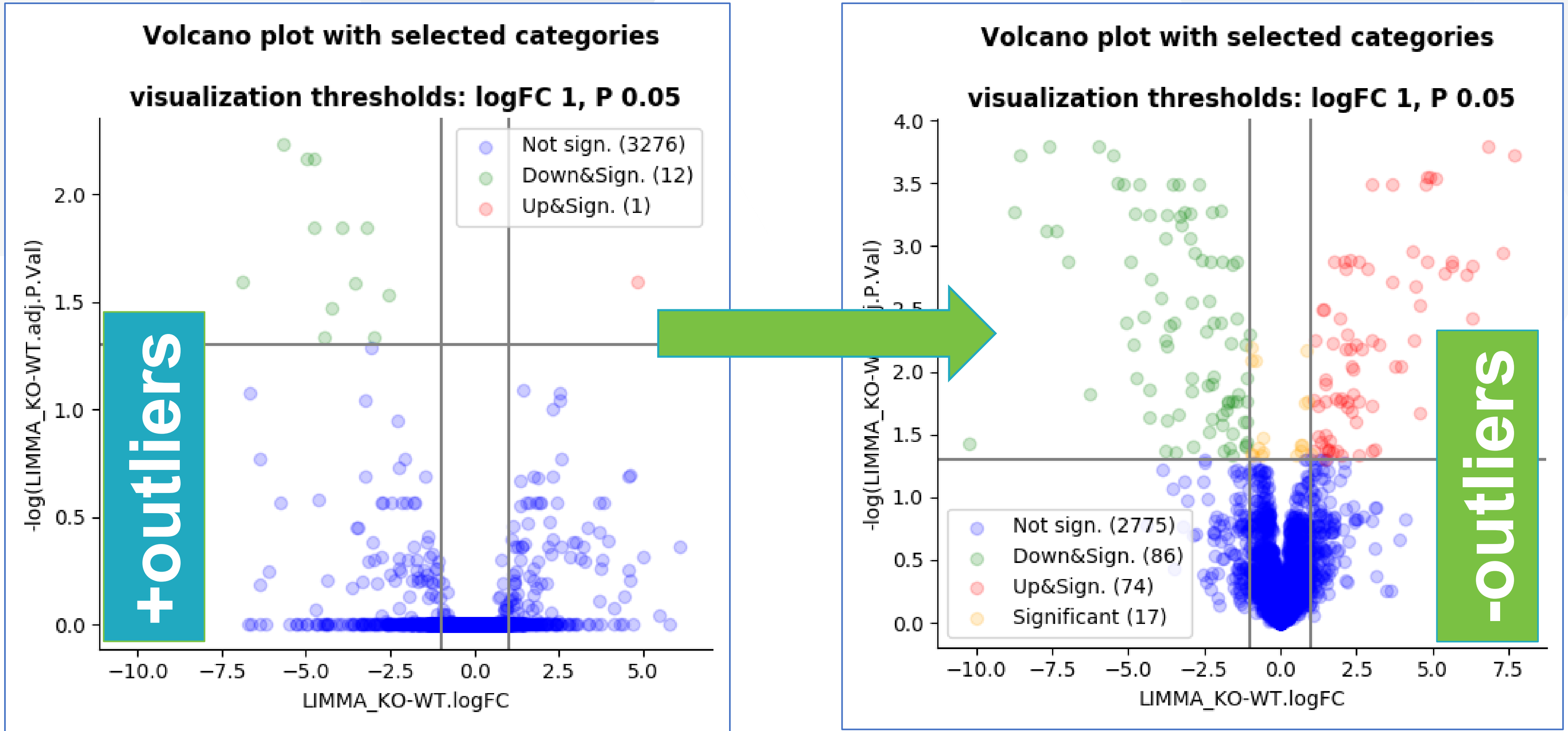
samples protein intensities correlations



Density (dark blue->dark red) scatter plot ('y' on 'x') matrix
red curve is estimated nonparametric lowess model
dashed line is unity line (x=y)
not normalized data - WT only

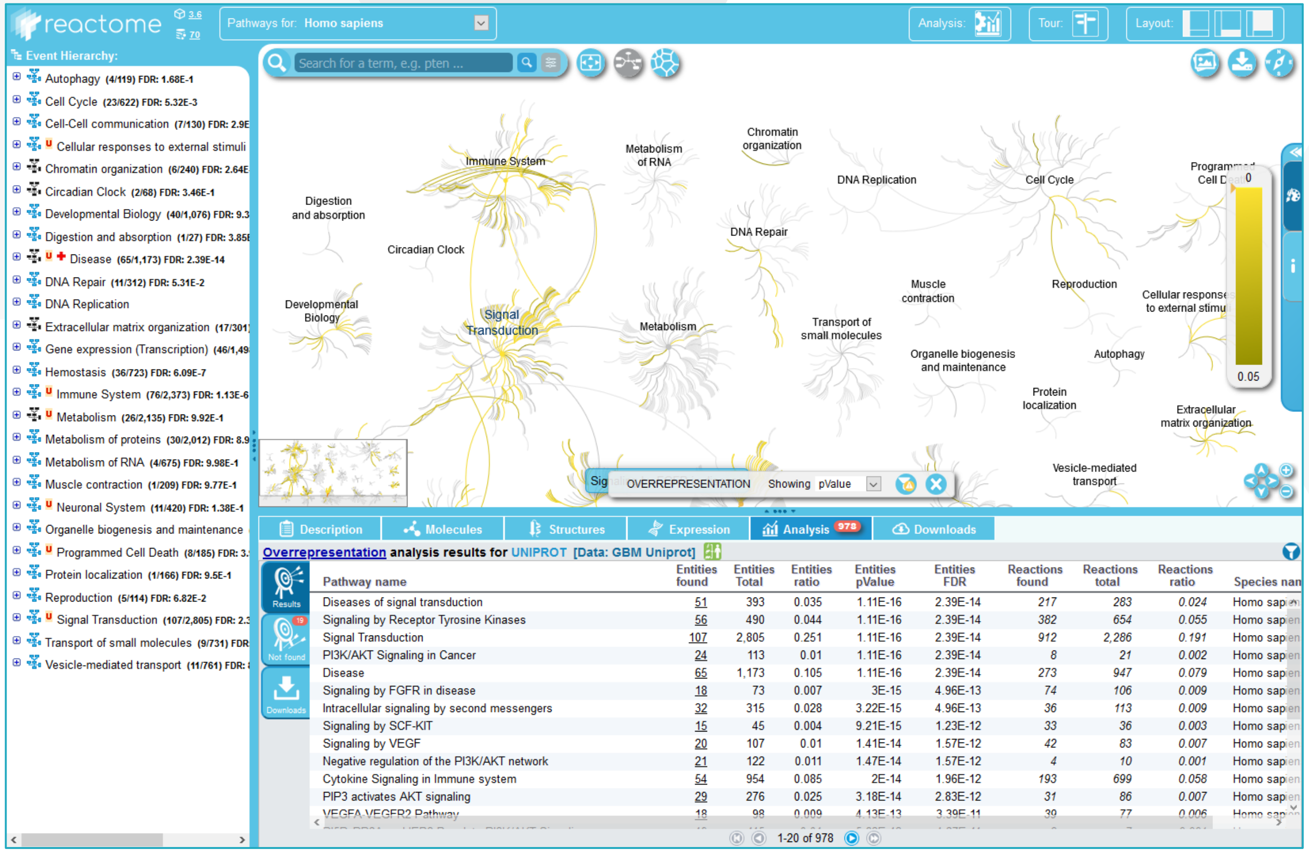
General proteomics study in our core facility

- **statistical data analysis**
  - e.g. LIMMA test (moderated t-test)
  - importance of proper **experimental design**, **outliers exclusion**, ...



General proteomics study in our core facility

- **initial** proteins list **interpretation**
  - using whole protein list or just candidate proteins list
    - e.g. pathway analysis using Reactome



Event Hierarchy:

- Autophagy (4/119) FDR: 1.68E-1
- Cell Cycle (23/622) FDR: 5.32E-3
- Cell-Cell communication (7/130) FDR: 2.9E
- Cellular responses to external stimuli
- Chromatin organization (6/240) FDR: 2.64E
- Circadian Clock (2/68) FDR: 3.46E-1
- Developmental Biology (40/1,076) FDR: 9.3
- Digestion and absorption (1/27) FDR: 3.85E
- Disease (65/1,173) FDR: 2.39E-14
- DNA Repair (11/312) FDR: 5.31E-2
- DNA Replication
- Extracellular matrix organization (17/301)
- Gene expression (Transcription) (46/1,49
- Hemostasis (36/723) FDR: 6.09E-7
- Immune System (76/2,373) FDR: 1.13E-6
- Metabolism (26/2,135) FDR: 9.92E-1
- Metabolism of proteins (30/2,012) FDR: 8.9
- Metabolism of RNA (4/675) FDR: 9.98E-1
- Muscle contraction (1/209) FDR: 9.77E-1
- Neuronal System (11/420) FDR: 1.38E-1
- Organelle biogenesis and maintenance
- Programmed Cell Death (8/185) FDR: 3.
- Protein localization (1/166) FDR: 9.5E-1
- Reproduction (5/114) FDR: 6.82E-2
- Signal Transduction (107/2,805) FDR: 2.3
- Transport of small molecules (9/731) FDR
- Vesicle-mediated transport (11/761) FDR:

Overrepresentation analysis results for UNIPROT [Data: GBM Uniprot]

Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Species nan
Diseases of signal transduction	51	393	0.035	1.11E-16	2.39E-14	217	283	0.024	Homo sapien
Signaling by Receptor Tyrosine Kinases	56	490	0.044	1.11E-16	2.39E-14	382	654	0.055	Homo sapien
Signal Transduction	107	2,805	0.251	1.11E-16	2.39E-14	912	2,286	0.191	Homo sapien
PI3K/AKT Signaling in Cancer	24	113	0.01	1.11E-16	2.39E-14	8	21	0.002	Homo sapien
Disease	65	1,173	0.105	1.11E-16	2.39E-14	273	947	0.079	Homo sapien
Signaling by FGFR in disease	18	73	0.007	3E-15	4.96E-13	74	106	0.009	Homo sapien
Intracellular signaling by second messengers	32	315	0.028	3.22E-15	4.96E-13	36	113	0.009	Homo sapien
Signaling by SCF-KIT	15	45	0.004	9.21E-15	1.23E-12	33	36	0.003	Homo sapien
Signaling by VEGF	20	107	0.01	1.41E-14	1.57E-12	42	83	0.007	Homo sapien
Negative regulation of the PI3K/AKT network	21	122	0.011	1.47E-14	1.57E-12	4	10	0.001	Homo sapien
Cytokine Signaling in Immune system	54	954	0.085	2E-14	1.96E-12	193	699	0.058	Homo sapien
PIP3 activates AKT signaling	29	276	0.025	3.18E-14	2.83E-12	31	86	0.007	Homo sapien
VEGFA-VEGFR2 Pathway	18	98	0.009	4.13E-13	3.39E-11	39	77	0.006	Homo sap

1-20 of 978

General proteomics study in our core facility

- study **results discussion** with the customer

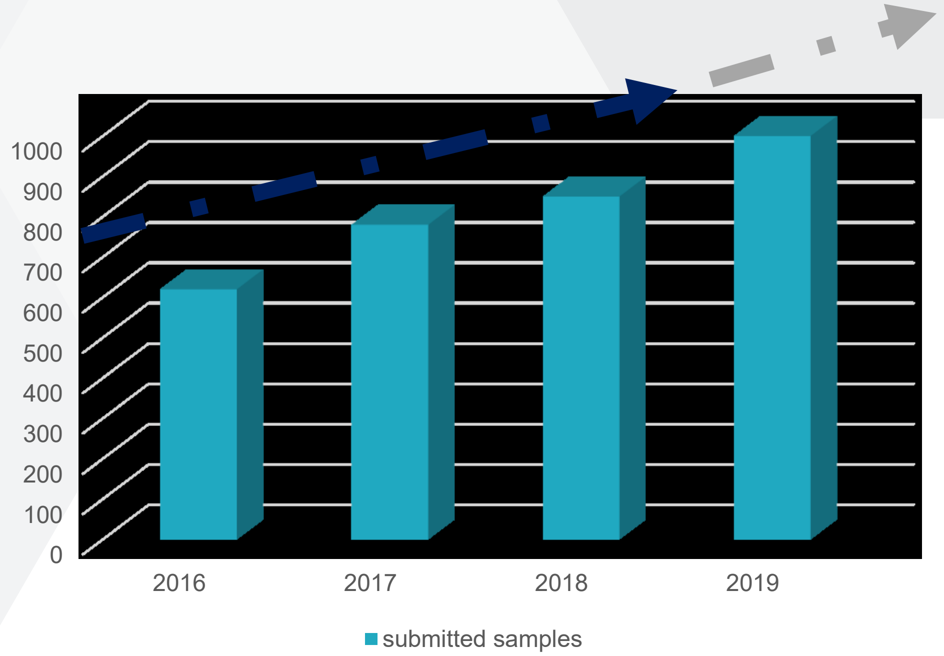
- study **results verification**
  - mainly on the customer side, with our assistance if needed
  - MS data **re-evaluation and finalization** sometimes needed
    - including e.g. MS data database re-searching
    - combination with other omics data
    - combining with another set of samples prepared

  - **even several years long process...**

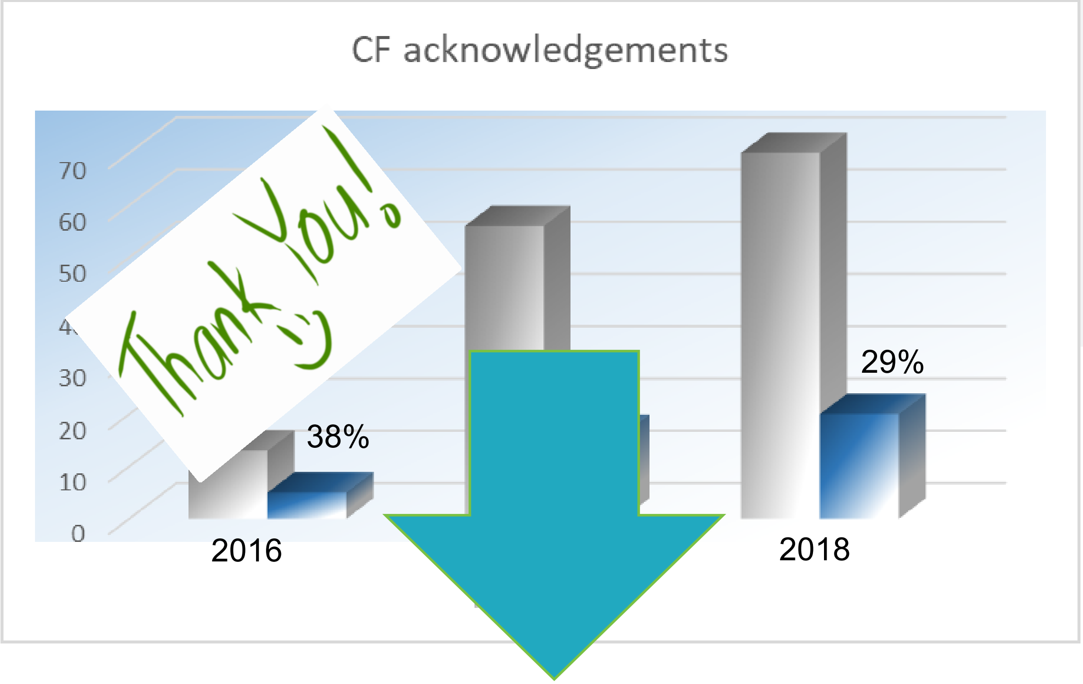
- study **publication**

General proteomics study in our core facility

Number of LC-MS/MS samples (04/2016 – 11/2019)

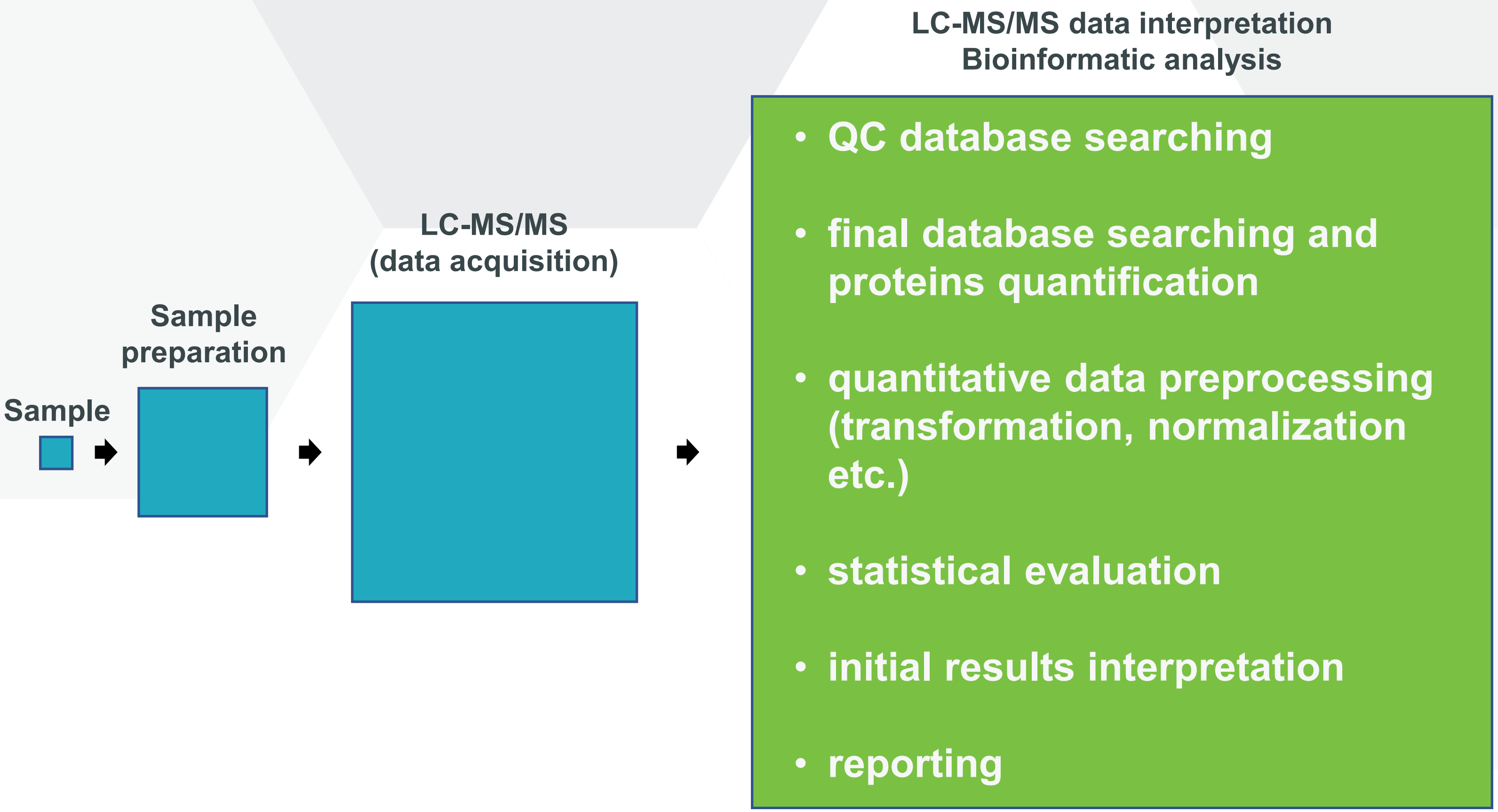


CIISB project - 10 Core facilities



funding of our CF

General proteomics study in our core facility - overview



Sample

Sample preparation

LC-MS/MS (data acquisition)

LC-MS/MS data interpretation
Bioinformatic analysis

- QC database searching
- final database searching and proteins quantification
- quantitative data preprocessing (transformation, normalization etc.)
- statistical evaluation
- initial results interpretation
- reporting

1) Software container running KNIME

Reasons to create such environment?

- more and more **complex** data processing and visualization **steps needed and used**
- **combining individual steps** into **complex pipelines**
  - **flexibility** of the processing ways necessary
    - one approach not generally applicable for all datasets, even though with similar concept
    - multiple settings tested, benchmarked and considered
- to have **easy to be used environment** without any scripting necessary
- to have **reproducible environment**, yet with up to date techniques, i.e. **versioning** needed
  - use identical environment on many installations covering also e.g. R, python specific versions
  - **older versions easily accessible** if needed even after years
- **revisiting the old pipelines** needed (publication, candidates verification, ...)
  - detailed **documentation** of the used tools/steps with **concrete settings**
    - single processing step settings details having major effect on the results!
    - including all the scripts used
- be able to **reproduce/reuse the older pipelines** on other data – **D**on't **R**epeat **Y**ourself
- use, support and build our processing pipelines on **free and open-source** tools

Potential tools/solutions for data processing

- selected tools/solutions
  - Proteome Discoverer – commercial, ready-to-use solution, limited functionality
  - MaxQuant + Perseus – free, but closed source (black boxes)
  - R studio/R console (DEP, MSnbase)
  - Jupyter notebooks – using python, R, or other scripting language
  - Galaxy (https://galaxyproject.org)
    - great community
    - less clear and user-friendly interface
    - harder to get new features in and script on the go

Open for Innovation KNIME

Ubuntu
(incl. desktop environment, browser, text editor, ...)

KNIME Analytics Platform
with additional
ns

python and selected **packages** for use in KNIME

R and selected **packages** for use in KNIME

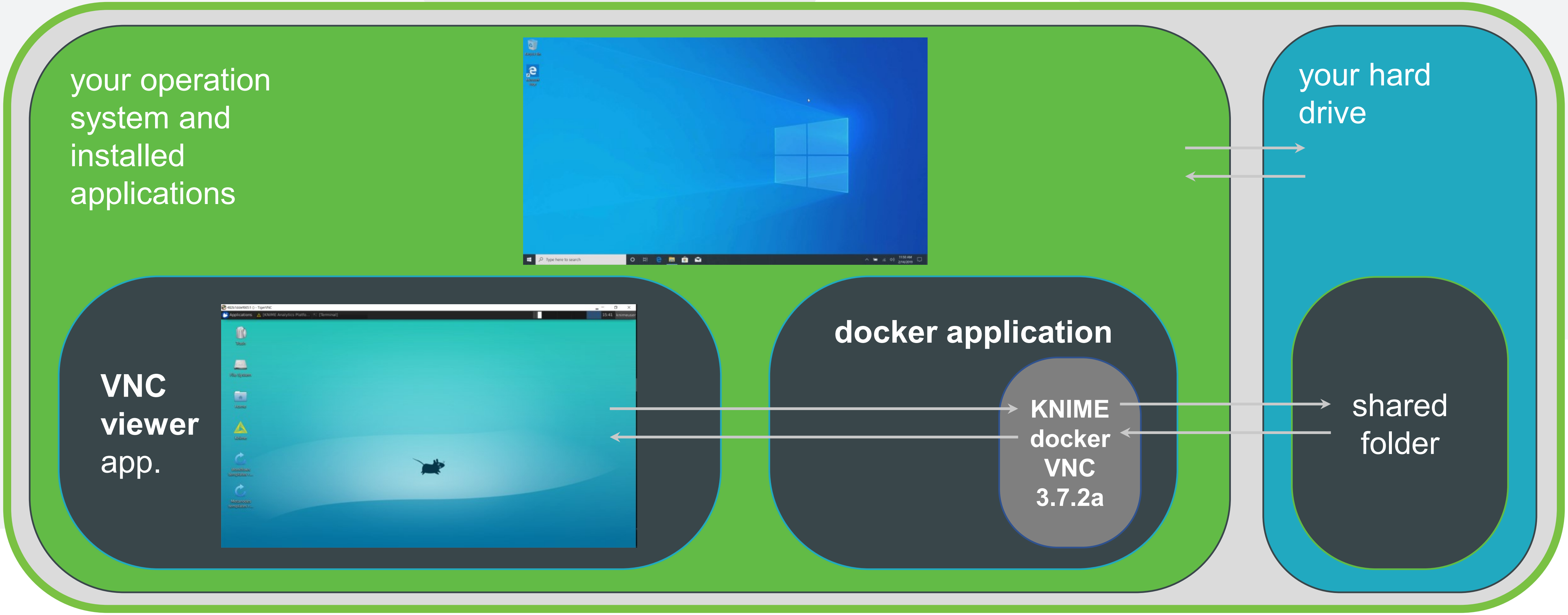
Scripts for selected use cases like git

ote
access via VNC protocol and shared folder

docker

Running software container scheme – local version

your PC (Windows, Linux) or Mac



your operation system and installed applications

VNC viewer app.

docker application

KNIME docker VNC 3.7.2a

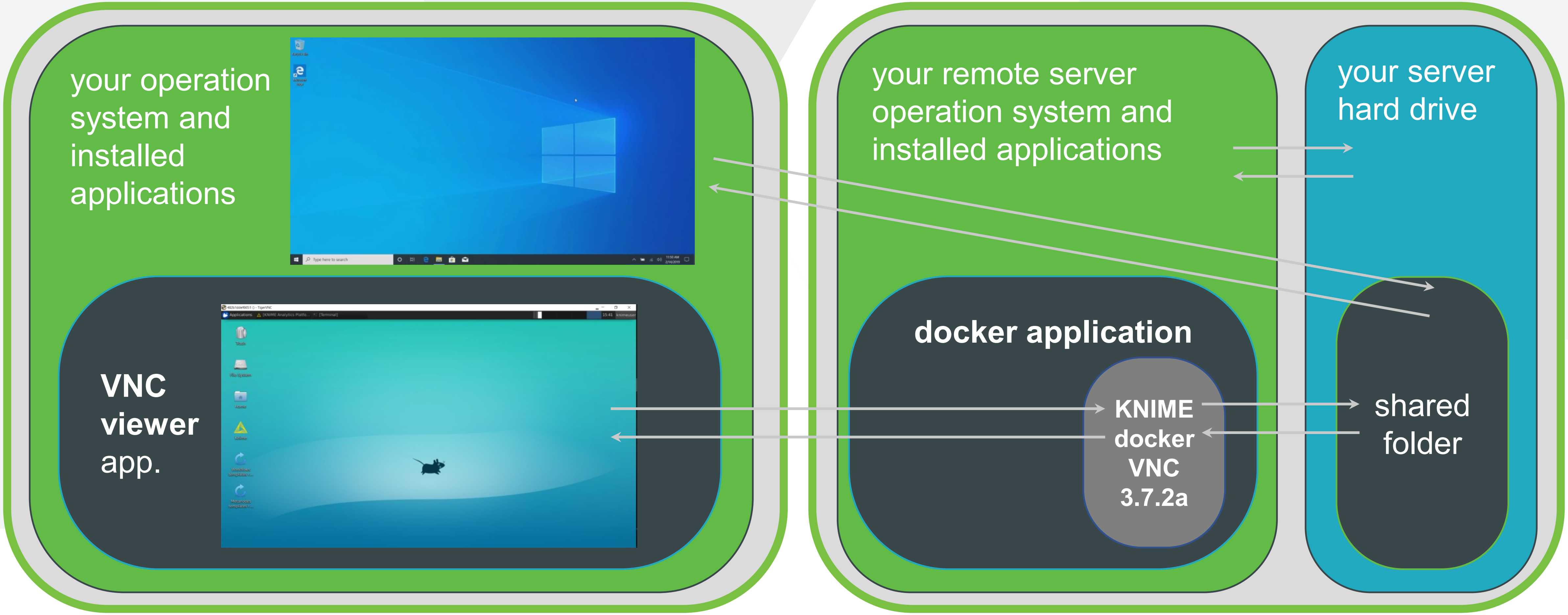
your hard drive

shared folder

Running software container scheme – server version

your PC (Windows, Linux) or Mac

your remote server



your operation system and installed applications

VNC **viewer** app.

your remote server operation system and installed applications

docker application

KNIME docker VNC 3.7.2a

your server hard drive

shared folder

Requirements to run container – run locally

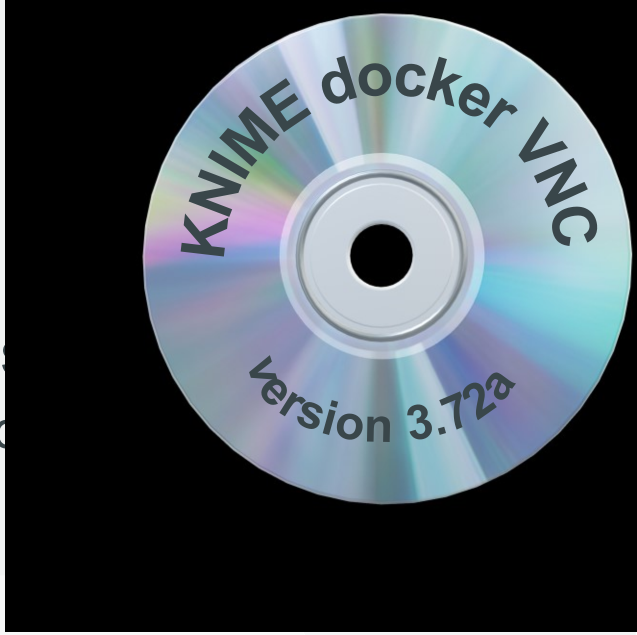
- hardware
  - **supporting virtualization** (you may need to enable it in the bios)
  - desktop or server; Mac hardware 2010 or newer
- operation system
  - **Linux** (64-bit, kernel >=3.10)
  - **Mac** (macOS >=10.12)
  - **Windows** (64-bit, Win 10, Pro, Enterprise or Education, Build 15063 or later)
- software
  - **docker application** installed
  - **VNC viewer** (TigerVNC viewer recommended)
  - **VirtualBox** can **NOT** be installed
  - **folder for sharing** with the environment

Requirements to run container – access to server

- hardware
  - **supporting virtualization** (you may need to enable it in the bios)
  - desktop or server; Mac hardware 2010 or newer
- operation system
  - **Linux** (64-bit, kernel >=3.10)
  - **Mac** (macOS >=10.12)
  - **Windows** (64-bit, Win 10, Pro, Enterprise or Education, Build 15063 or later)
- software
  - **docker application** installed
  - **VNC viewer** (TigerVNC viewer recommended)
  - **VirtualBox** can **NOT** be installed
  - **folder for sharing** with the environment – access only

Software container availability

- GitHub repository with docker files and associated files + script
  - to know how the environment has been build and what components it co
  - https://github.com/OmicsWorkflows/KNIME_docker_vnc



- prebuild, ready to be used docker images on Docker Hub
  - to use the environment directly (info on how to use it is on the GitHub repository)
  - https://hub.docker.com/r/cfprot/knime

Thank you for your attention

Workshop outline

- morning session – theoretical part
  - 10:00 – 10:15 Opening and introduction
  - 10:15 – 10:45 Software container running KNIME
  - **10:45 – 11:00 Coffee break**
  - 11:00 – 11:30 Introduction to KNIME
  - 11:30 – 11:45 Coffee break
  - 11:45 – 12:30 Practical applications, our KNIME metanodes
  - 12:30 – 13:30 Lunch break, visit of our laboratories for interested people