



Central European Institute of Technology
BRNO | CZECH REPUBLIC

Softwarový kontejner a nové KNIME komponenty pro procesování proteomických dat

2019-11-29 Olomouc

MUNI

David Potěšil
Kristína Gömöryová
Michal Bednařík
Michal Cupák
Vítězslav Bryja
Zbyněk Zdráhal

Proč jsme vytvořili a používáme softwarový kontejner?

- mít možnost použít stále **komplikovanějších kroků** při procesování a vizualizaci dat
 - ⇒ použití **aktuálních technik a postupů**
- pro snadnou **kombinaci jednotlivých kroků** do **workflow**
 - ⇒ **flexibilní** zpracování dat (jeden postup nestačí, testování a hodnocení jednotlivých kroků)
- abychom měli **reprodukovatelné a zpětně dostupné prostředí včetně** jeho **aktualizace** (verze)
 - ⇒ **použití** identického prostředí **na více místech**
 - ⇒ **starší verze prostředí jednoduše dostupné** i za několik let
- pro **zpětné projití** dříve použitých workflow (publikace, ověření kandidátních proteinů, ...)
 - ⇒ **přístup k dokumentaci** jednotlivých kroků a použitých nástrojů a **konkrétním nastavením**
 - nastavení jednoho kroku může mít hlavní vliv na konečné výsledky!
 - včetně všech použitých skriptů
- být schopni **znovupoužít starší workflow** na nová data ⇒ *Don't Repeat Yourself*
- i přesto mít **jednoduše použitelné prostředí** bez nutnosti skriptovat ⇒ **ne** jen pro **geeky**
- používat, podporovat a vytvářet **svobodné a open-source** nástroje ⇒ **otevřené a zdarma**

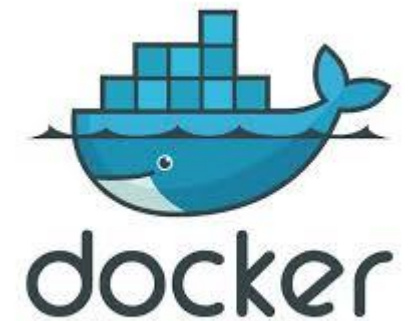
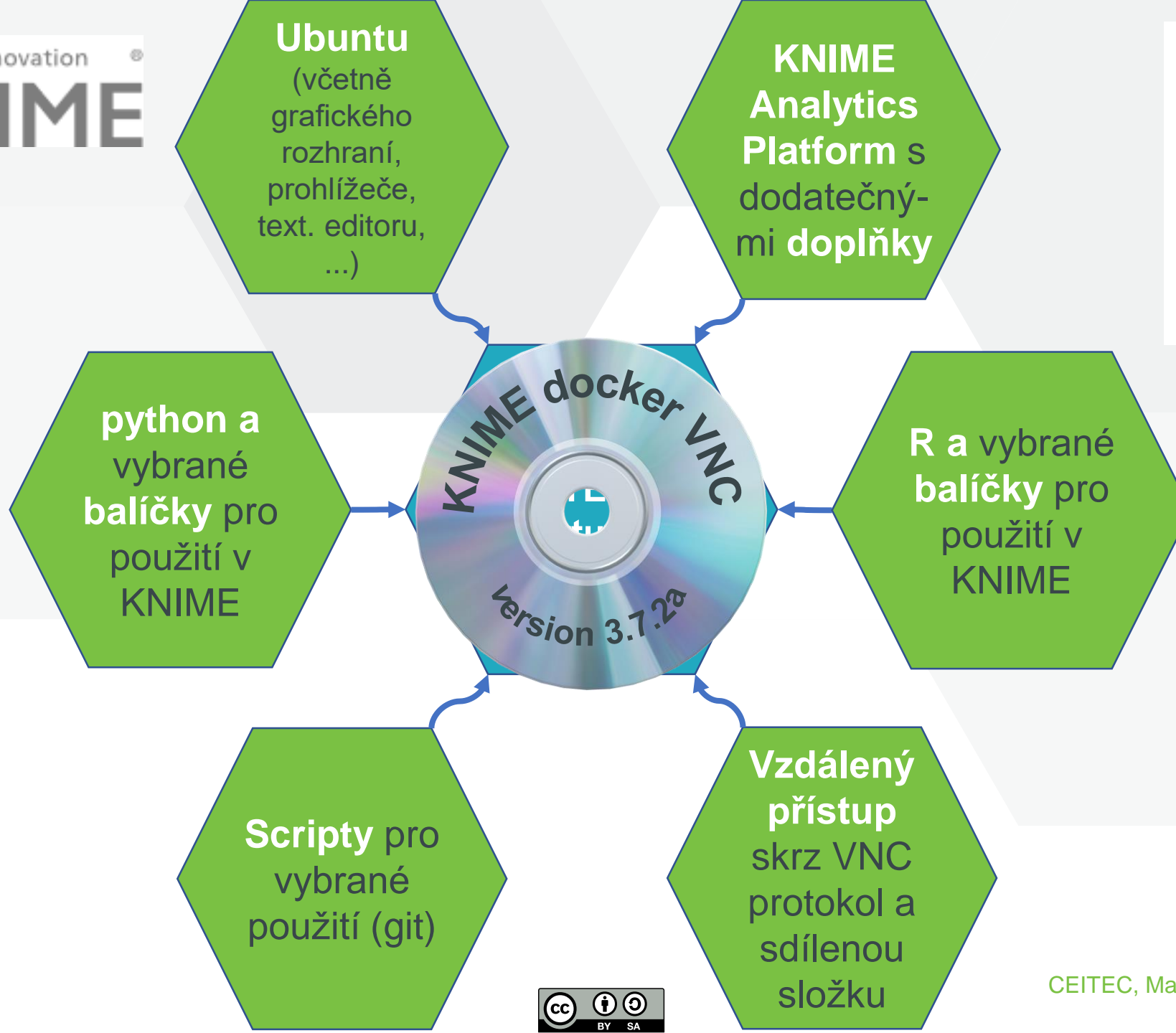
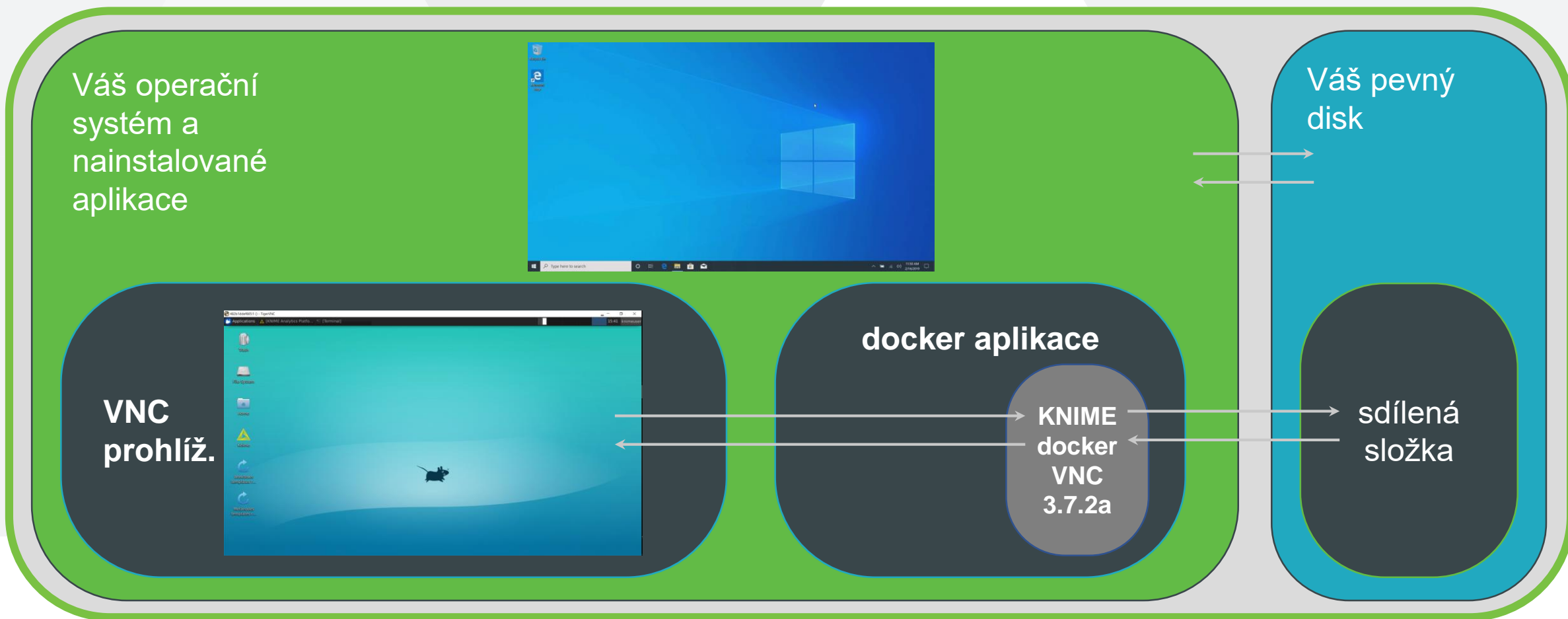


Schéma běhu softwarového kontejneru – lokální verze

Váš PC (Windows, Linux) nebo Mac





Trash



File System



Home



Knime

KNIME aplikace



Workflows
templates r...



Metanodes
templates r...



File Edit View Node Help

75%

KNIME Explorer

Workflow_templates

MQ_PGs_LFQ_gen

graph1_files

graph_files

heatmap_files

Workflow Coach

Recommended Nodes

Node Repository

ungroup

Ungroup - /Manipulation/Row/

GroupBy - /Manipulation/Row/

Group Loop Start - /Workflow C

Database GroupBy - /Databas

Round Double - /Manipulation/

MRMTransitionGroupPicker - /K

Grouped ScatterPlot - /Scriptin

Independent groups t-test - /Au

Unpivoting - /Manipulation/Row

*0: MQ_PGs_LFQ_general_0.6a

File Reader

proteinGroups.txt file from __inputs__ folder

MQ PGs filtering (e.g. cRAP)

filters e.g. cRAP

Column Resorter

resorts columns to get them in the most meaningful way

Math Formula (Multi Column)

log2 calculation for all intensity columns

Math Formula (Multi Column)

log2 calculation for all LFQ intensity columns

Normalization tests - MQ protein groups

test of different norm. approaches (median, quantile, loessF, vsn, MaxLFQ)

LIMMA tests

computes LIMMA test

sorts PGs

Example dataset

MQ_PGs_LFQ_general workflow

Description of the workflow and general remarks

MQ_PGs_LFQ_general is a KNIME workflow developed for the general processing of label-free bottom-up mass spectrometry data.

Please note, that you should understand e.g. data structure and experimental design used within the study to apply the correct processing approach! The pr

Input data

Node Description

File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.

When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Outline

Console

KNIME Console

WARN Column Rename (Regex) 0:844:0:0:850 Pattern replace resulted in duplicate column names; resolved c

WARN CASE Switch Data (Start) 0:844:0:0:73 Errors overwriting node settings with flow variables: Unknow

WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unknow

WARN Column Rename (Regex) 0:844:239:0:850 Pattern replace resulted in duplicate column names; resolved

WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unkn

WARN Column Rename (Regex) 0:844:300:277:0:428 Pattern did not match any column name, leaving input unc

WARN Column Rename (Regex) 0:844:300:290:0:428 Pattern did not match any column name, leaving input unc

330M of 408M

KNIME Analytics Platform

File Edit View Node Help

75%

Quick Access

KNIME Explorer

Workflow_templates

- MQ_PGs_LFQ_gen
- graph1_files
- graph_files
- heatmap_files

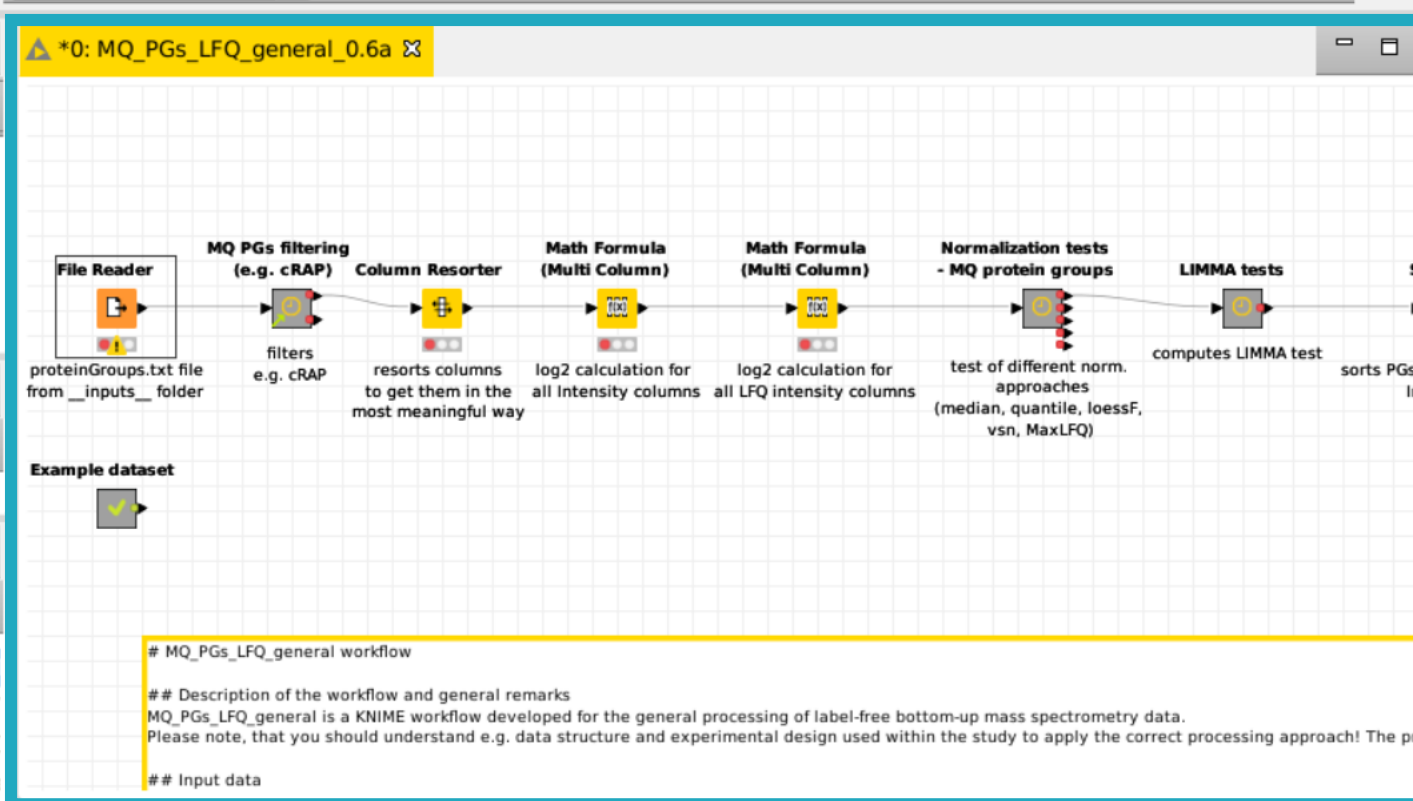
Workflow Coach

Recommended Nodes

Node Repository

ungroup

- Ungroup - /Manipulation/Row/
- GroupBy - /Manipulation/Row/
- Group Loop Start - /Workflow C
- Database GroupBy - /Database
- Round Double - /Manipulation/
- MRMTransitionGroupPicker - /K
- Grouped ScatterPlot - /Scriptin
- Independent groups t-test - /Au
- Unpivoting - /Manipulation/Row



Workflow editor

File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.

When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Outline

Console

KNIME Console

```
WARN Column Rename (Regex) 0:844:0:0:850 Pattern replace resulted in duplicate column names; resolved c
WARN CASE Switch Data (Start) 0:844:0:0:73 Errors overwriting node settings with flow variables: Unknow
WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unknow
WARN Column Rename (Regex) 0:844:239:0:850 Pattern replace resulted in duplicate column names; resolved
WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unkn
WARN Column Rename (Regex) 0:844:300:277:0:428 Pattern did not match any column name, leaving input unc
WARN Column Rename (Regex) 0:844:300:290:0:428 Pattern did not match any column name, leaving input unc
```

330M of 408M

KNIME Explorer

Workflow_templates

MQ_PGs_LFQ_gen

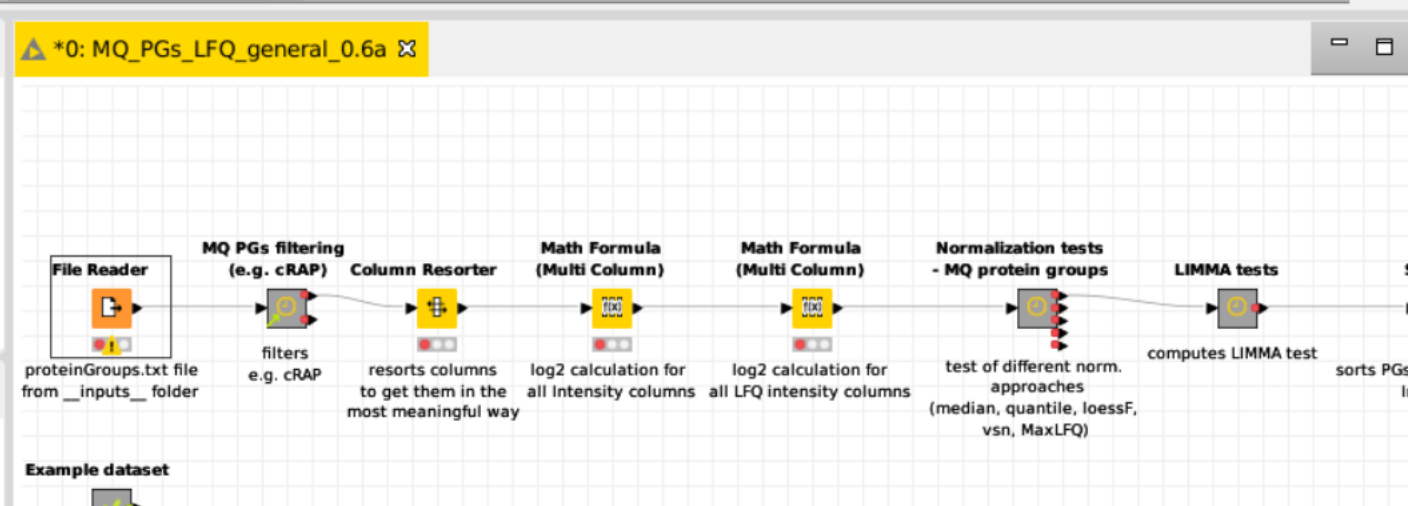
graph1_files

graph_files

heatmap_files

Workflow Coach

Recommended Nodes



Node Description

File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.

When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Node Repository

ungroup

Ungroup - /Manipulation/Row/

GroupBy - /Manipulation/Row/

Group Loop Start - /Workflow C

Database GroupBy - /Database

Round Double - /Manipulation/

MRMTransitionGroupPicker - /C

Grouped ScatterPlot - /Scriptin

Independent groups t-test - /A

Unpivoting - /Manipulation/Row

Repositář KNIME nodů

```
# MQ_PGs_LFQ_general workflow

## Description of the workflow and general remarks
MQ_PGs_LFQ_general is a KNIME workflow developed for the general processing of label-free bottom-up mass spectrometry data.
Please note, that you should understand e.g. data structure and experimental design used within the study to apply the correct processing approach! The pr

## Input data
```

Outline

Console

KNIME Console

```
WARN Column Rename (Regex) 0:844:0:0:850 Pattern replace resulted in duplicate column names; resolved c
WARN CASE Switch Data (Start) 0:844:0:0:73 Errors overwriting node settings with flow variables: Unknow
WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unkn
WARN Column Rename (Regex) 0:844:239:0:850 Pattern replace resulted in duplicate column names; resolved
WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unkn
WARN Column Rename (Regex) 0:844:300:277:0:428 Pattern did not match any column name, leaving input unc
WARN Column Rename (Regex) 0:844:300:290:0:428 Pattern did not match any column name, leaving input unc
```


KNIME Analytics Platform

File Edit View Node Help

75%

Quick Access

KNIME Explorer

Workflow_templates

MQ_PGs_LFQ_gen

graph1_files

graph_files

heatmap_files

Workflow Coach

Recommended Nodes

Node Repository

ungroup

Ungroup - /Manipulation/Row/

GroupBy - /Manipulation/Row/

Group Loop Start - /Workflow C

Database GroupBy - /Databas

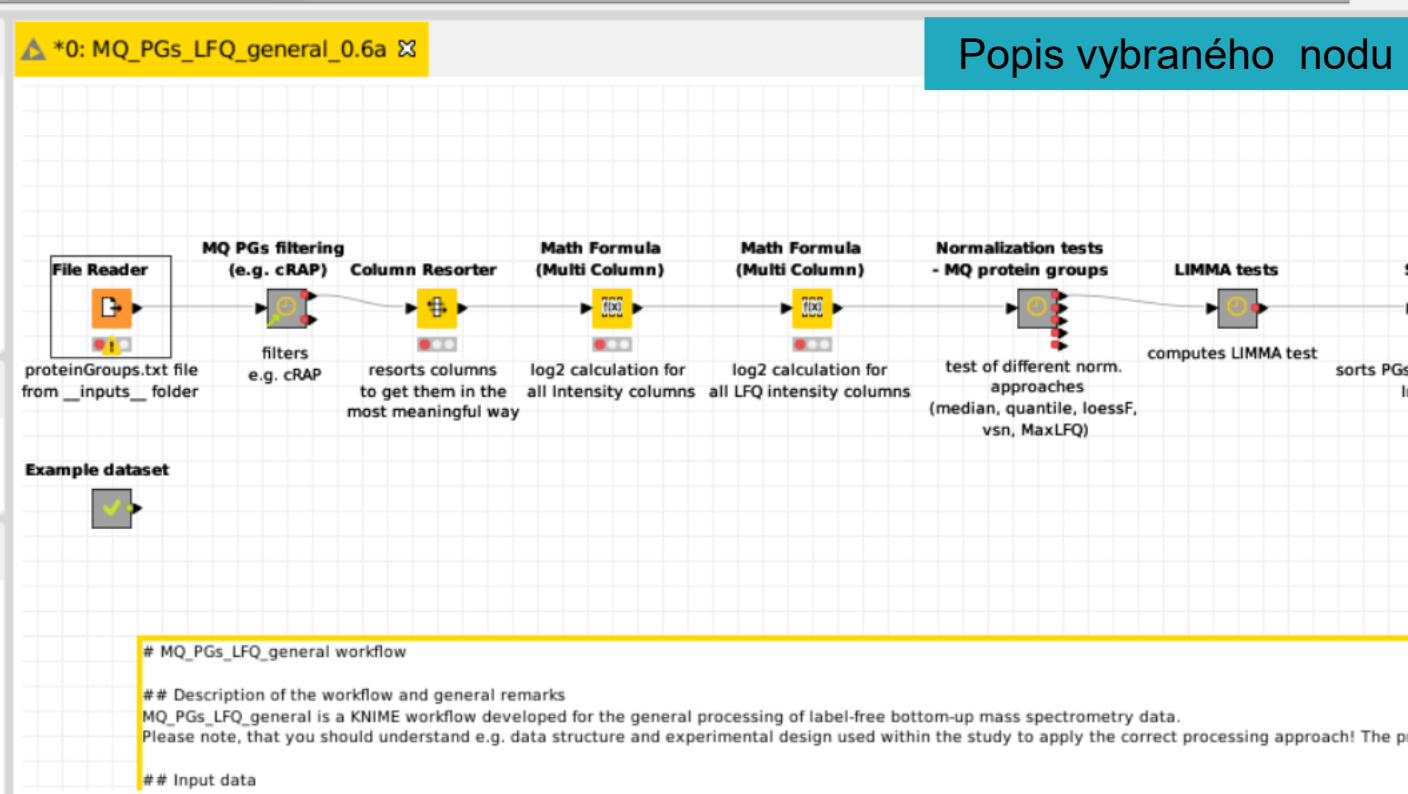
Round Double - /Manipulation/

MRMTransitionGroupPicker - /K

Grouped ScatterPlot - /Scriptin

Independent groups t-test - /Ai

Unpivoting - /Manipulation/Row



Popis vybraného nodu

Node Description

File Reader

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.

When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Outline

Console

KNIME Console

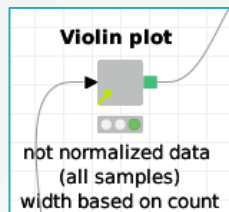
```
WARN Column Rename (Regex) 0:844:0:0:850 Pattern replace resulted in duplicate column names; resolved c
WARN CASE Switch Data (Start) 0:844:0:0:73 Errors overwriting node settings with flow variables: Unknown
WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unknown
WARN Column Rename (Regex) 0:844:239:0:850 Pattern replace resulted in duplicate column names; resolved
WARN CASE Switch Data (Start) 0:844:239:0:73 Errors overwriting node settings with flow variables: Unknown
WARN Column Rename (Regex) 0:844:300:277:0:428 Pattern did not match any column name, leaving input unchanged
WARN Column Rename (Regex) 0:844:300:290:0:428 Pattern did not match any column name, leaving input unchanged
```

330M of 408M

Co když KNIME nody nestačí?

- stovky nodů pro zpracování a vizualizaci dat již přímo v KNIME
- ale co když...
 - se objevil nový statistický nástroj (např. jako R balíček), který prostě musíte použít ☺
 - by se mi líbil typ grafu co jsem viděl tuhle v publikaci použitý na má data
 - často používám skript pro specifické procesování dat s menší/větší obměnou proměnných
 - bych chtěl kolegům umožnit používat podobné nástroje jako já, i když neumí skriptovat
 - ...
- udělejte si vlastní nod!
 - koncept vizuálního programování v KNIME – „naklikej si svůj nod“

Příklad našeho nodu – *Violin plot* (Houslový graf)



pohled se strany **uživatel**e nodu

Dialog - 0:844:291:0 - Violin plot (not normalized data)

File

QuickForms Flow Variables Memory Policy Job Manager Selection

Columns to process ☒ Change

☐ Manual Selection ☒ Wildcard/Regex Selection ☐ Type Selection

Pattern:

☒ Wildcard ☐ Regular expression ☒ Case Sensitive

Mismatch (Exclude)

- ☒ Protein IDs
- ☒ Majority protein IDs
- ☒ Peptide counts (all)
- ☒ Peptide counts (razor+unique)
- ☒ Peptide counts (unique)
- ☒ Protein names
- ☒ Gene names
- ☒ Fasta headers
- ☒ Number of proteins

Match (Include)

- ☒ Intensity WT_1_log2
- ☒ Intensity WT_2_log2
- ☒ Intensity WT_3_log2
- ☒ Intensity WT_4_log2
- ☒ Intensity KO_1_log2
- ☒ Intensity KO_2_log2
- ☒ Intensity KO_3_log2
- ☒ Intensity KO_4_log2

x axis label ☐ Change

y axis label ☒ Change

Graphs subtitle ☒ Change

Violin plot scale (width) setting ☒ Change

☒ count ☐ width

Manual y axis limits? ☐ Change

Manual y axis limits ☐ Change

Prefix to remove ☒ Change

Suffix to remove ☒ Change

OK Apply Cancel

Violin plot

Metanode to create Violin plot from selected columns of input table.

Note: any data preprocessing (like transformation, normalization) should be done prior the metanode usage!

Used programs and tools and their respective licenses at the time of the metanode creation. Version numbers and the licenses might differ based on your local installation. Please inspect your local installation and contact us if you can not locate your local version and or license terms.

KNIME nodes (The KNIME nodes consists of the following GNU GPL 3.0 License. Licence terms are available here: <https://www.gnu.org/licenses/gpl.html>)

Python 3 (The Python consists of the following Python 3.6 License. Licence terms are available here: <https://docs.python.org/3.6/license.html>)

Python package Seaborn (The Seaborn consists of the following BSD License. Licence terms are available here: <https://opensource.org/licenses/BSD-3-Clause>)

Python package Matplotlib (The Matplotlib consists of the following Python Software Foundation License (BSD compatible). Licence terms are available here: <https://matplotlib.org/users/license.html>)

Python package Pandas (The Pandas consists of the following BSD License. Licence terms are available here: <https://opensource.org/licenses/BSD-3-Clause>)

The metanode was created in KNIME 3.7.1 running inside the docker image (<https://hub.docker.com/r/cfprot/knime/>), tag 3.7.1a.

This version of metanode is available under the GNU GPL 3.0 License, unless stated otherwise. The full version of the license terms is available at <https://www.gnu.org/licenses/gpl.html>.

Version: 0.4.3 from 2019-03-20

Contact person: David Potesil (david.potesil@ceitec.muni.cz)

More information can be found at https://github.com/OmicsWorkflows/KNIME_metanodes

Dialog Options

Columns to process

select columns to be processed

x axis label

how the graph x axis should be titled

y axis label

how the graph y axis should be titled

Graphs subtitle

additional information that should be present as the graphs subtitle

Violin plot scale (width) setting

sets the violin plot scale settings

- count - violin plot scale (width) will reflect the number of values

- width - all violin graphs will have the same width irrespective the number of values

Manual y axis limits?

whether to use manually set y axis limits (checked) or use automatic limits (unchecked)

manual y axis limits

limits of y axis in the form of two numbers separated by semicolon; use point (.) as decimal separator

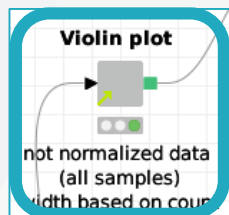
Prefix to remove

common data columns prefix to be removed prior plotting

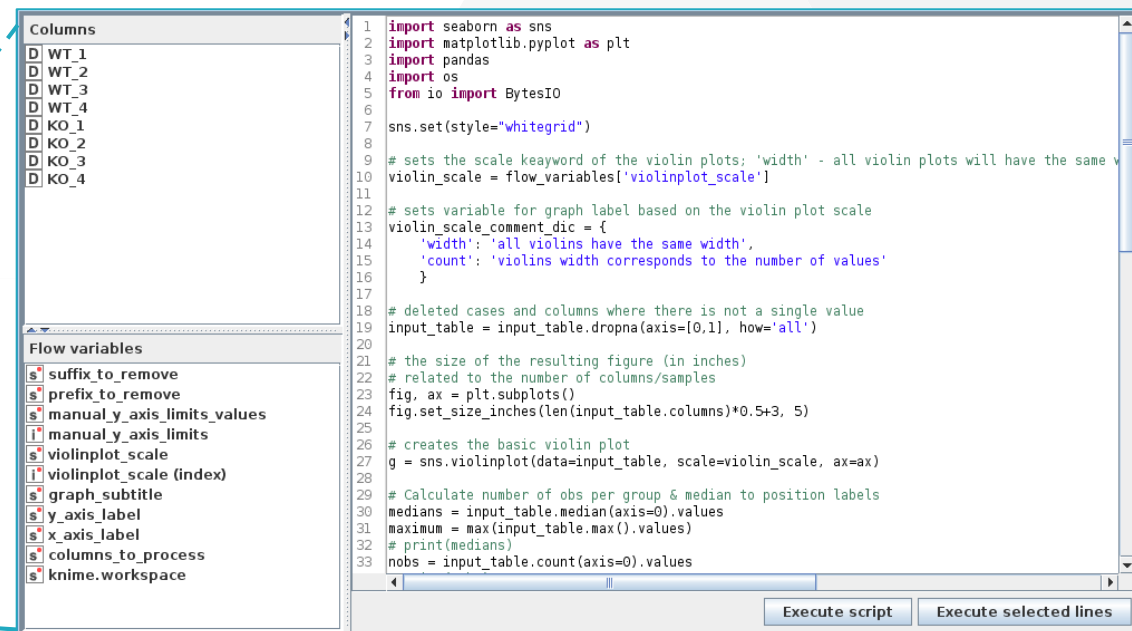
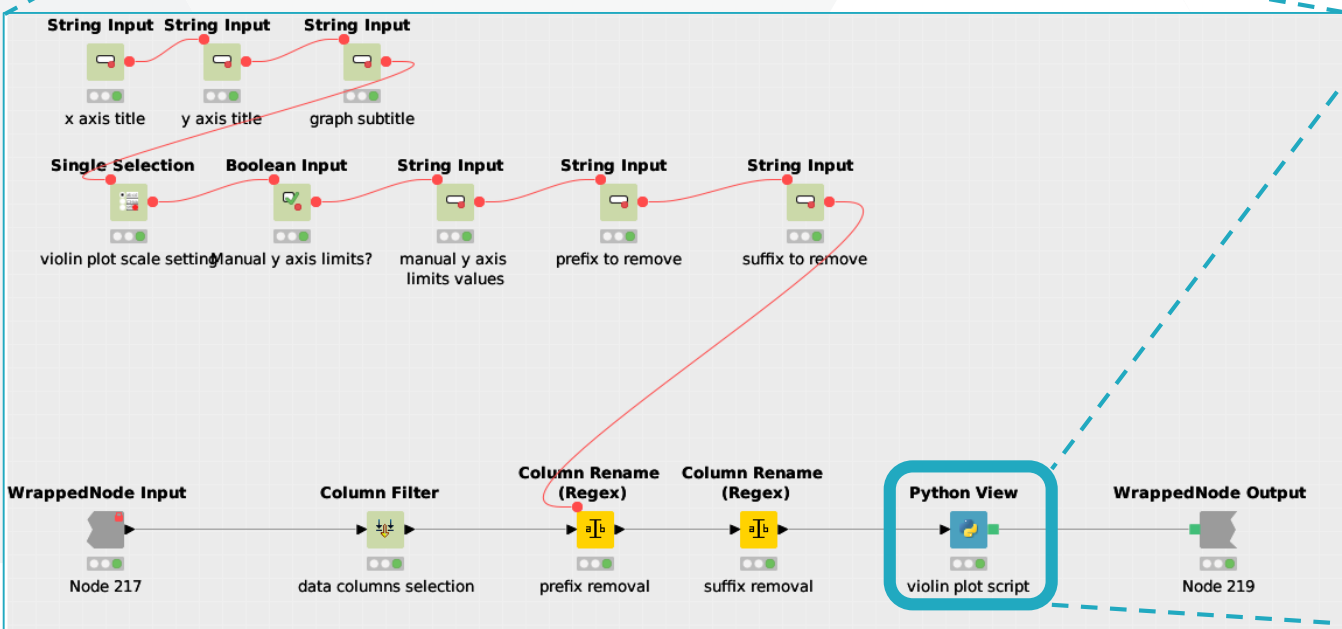
Suffix to remove

common data columns suffix to be removed prior plotting

Příklad našeho nodu – *Violin plot* (Houslový graf)



pohled ze strany tvůrce nodu



```
1 import seaborn as sns
2 import matplotlib.pyplot as plt
3 import pandas
4 import os
5 from io import BytesIO
6
7 sns.set(style="whitegrid")
8
9 # sets the scale keyword of the violin plots; 'width' - all violin plots will have the same width
10 violin_scale = flow_variables['violinplot_scale']
11
12 # sets variable for graph label based on the violin plot scale
13 violin_scale_comment_dic = {
14     'width': 'all violins have the same width',
15     'count': 'violins width corresponds to the number of values'
16 }
17
18 # deleted cases and columns where there is not a single value
19 input_table = input_table.dropna(axis=[0,1], how='all')
20
21 # the size of the resulting figure (in inches)
22 # related to the number of columns/samples
23 fig, ax = plt.subplots()
24 fig.set_size_inches(len(input_table.columns)*0.5+3, 5)
25
26 # creates the basic violin plot
27 g = sns.violinplot(data=input_table, scale=violin_scale, ax=ax)
28
29 # Calculate number of obs per group & median to position labels
30 medians = input_table.median(axis=0).values
31 maximum = max(input_table.max().values)
32 # print(medians)
33 nobs = input_table.count(axis=0).values
```


Dostupnost softwarového kontejneru a našich KNIME metanodů

- GitHub repositář se soubory docker, pomocnými soubory a skripty
 - včetně návodu jak použít předpřipravené docker obrazy (*images*)
 - víte jak bylo prostředí vytvořeno a jaké komponenty obsahuje
 - https://github.com/OmicsWorkflows/KNIME_docker_vnc
- GitHub repositář s KNIME nody
 - https://github.com/OmicsWorkflows/KNIME_metanodes
 - použitelné optimálně v rámci softwarového kontejneru
 - měly by pracovat i v samostatné KNIME instalaci
 - použité KNIME doplňky a python/R balíčky bude potřeba doinstalovat před jejich použitím



Dostupnost softwarového kontejneru a našich KNIME metanodů

- GitHub repositář se soubory a skripty
 - včetně návodu jak použít (přes)
 - víte jak bylo prostředí vytvořeno
 - <https://github.com/Omicron>
- GitHub repositář s KNIME metanodami
 - <https://github.com/Omicron>
 - použitelné optimálně v rámci
 - měly by pracovat i v samostatném kontejneru
 - použité KNIME doplňky a
 - použití

Děkuji za
pozornost!

