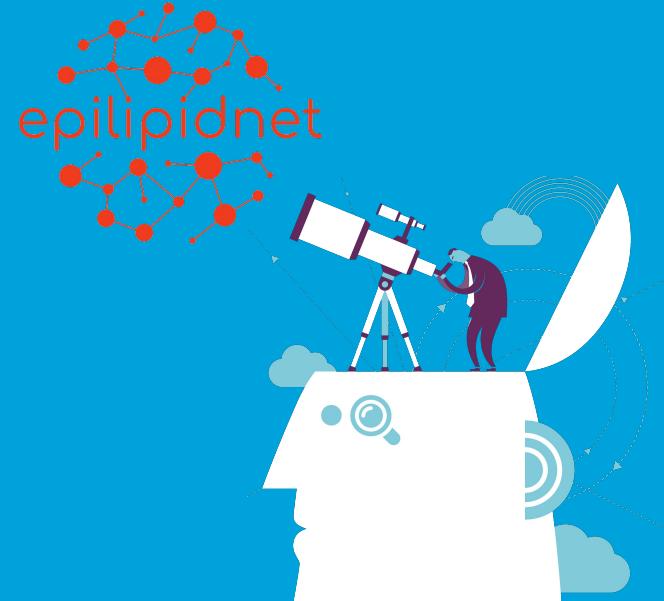


# Metabolomics data analysis

Denise Slenter

ORCID: 0000-0001-8449-1318

Tuesday May 28, 2024



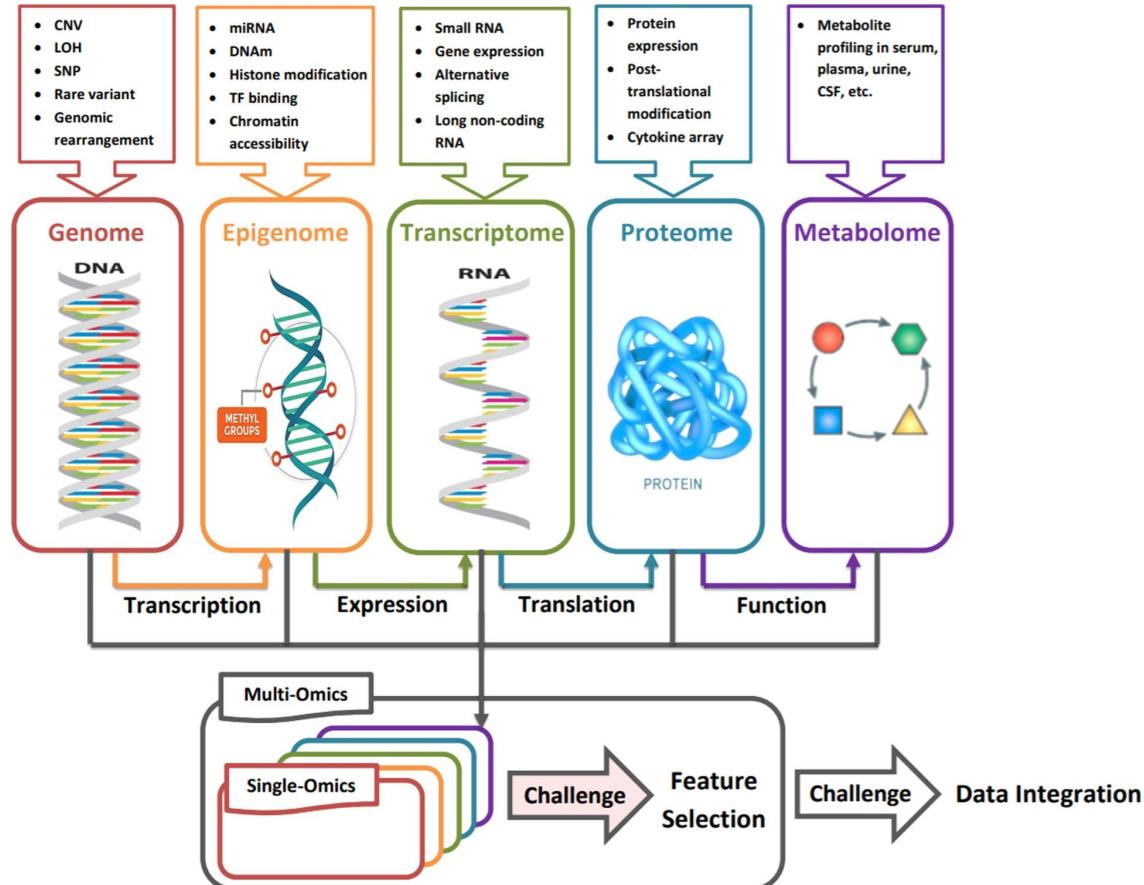
# What will I cover?

- + Existing tools for metabolomics data analysis, pros and cons
- + Steps in processing annotated metabolomics data
- + Working with Biocrates data

# What can I NOT cover?

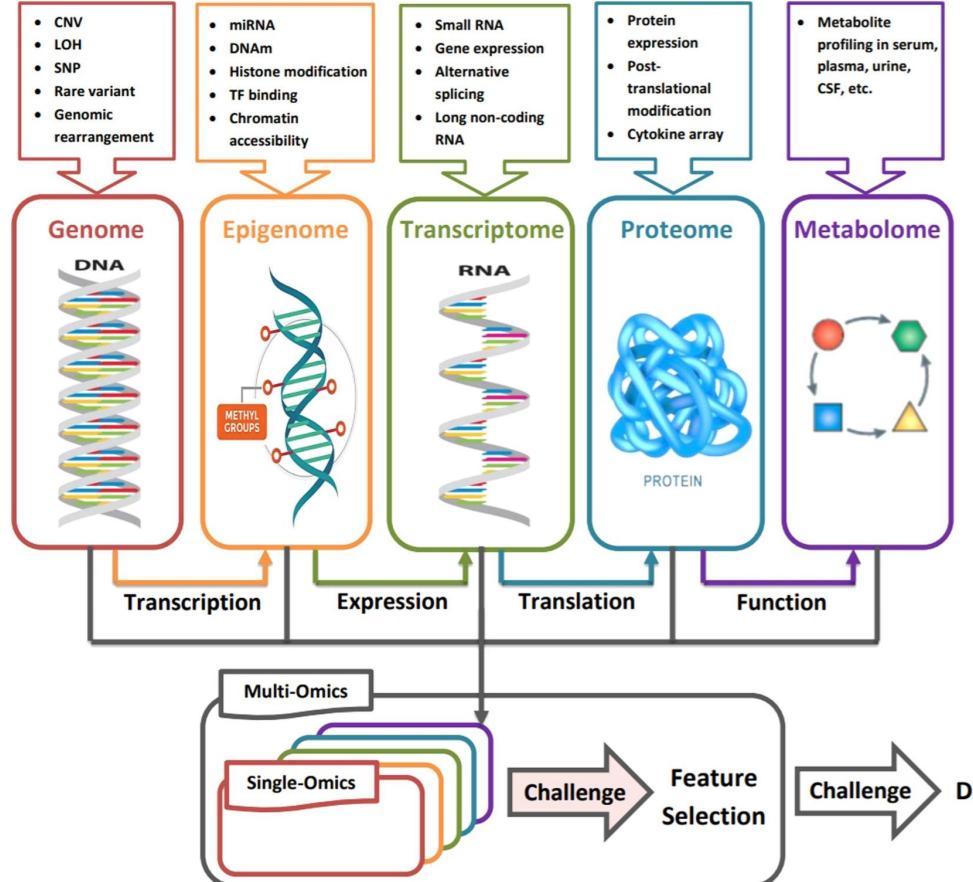
- Annotating raw data
- Overview of the metabolomics research field
- ‘Plug and Play’ data analysis scripts - you will need to write some code yourself :)

# Different data types and analysis techniques



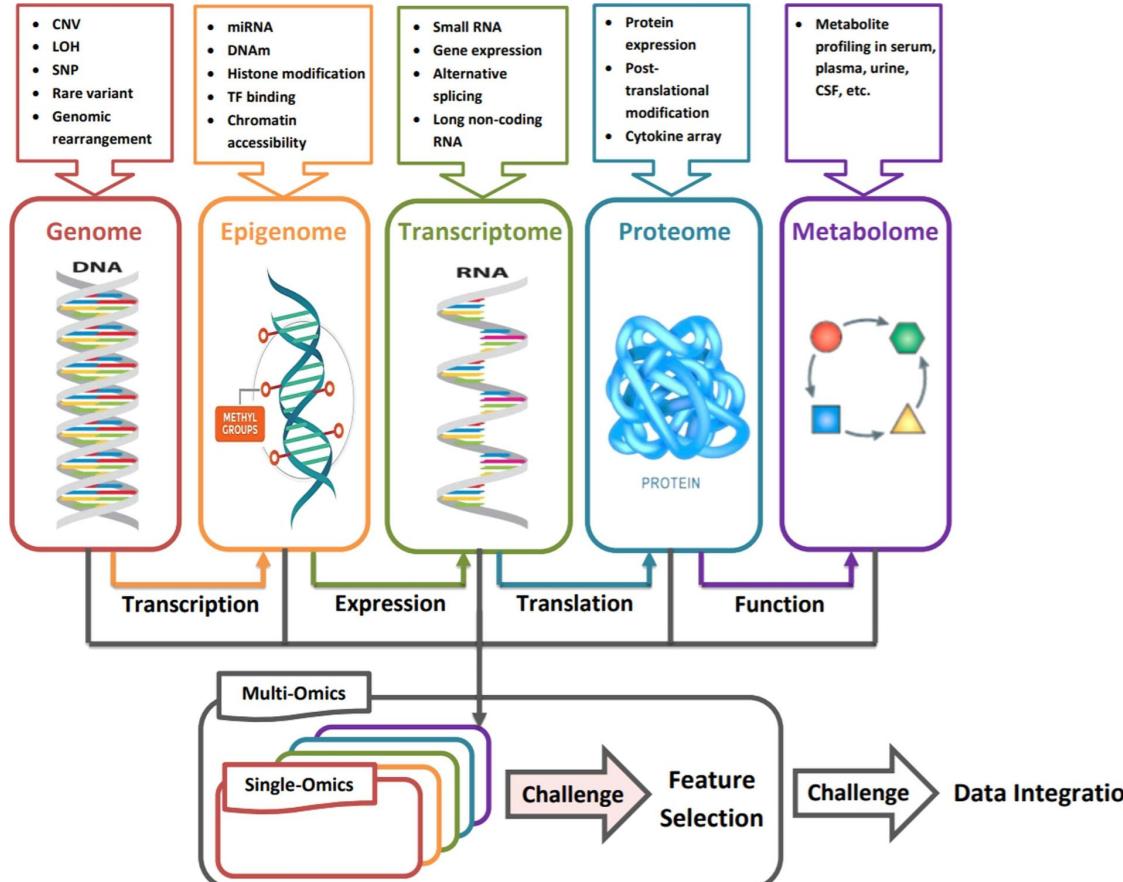
Adapted from: Momeni, Zahra, et al. "A survey on single and multi omics data mining methods in cancer data classification." Journal of Biomedical Informatics 107 (2020): 103466. DOI: [10.1016/j.jbi.2020.103466](https://doi.org/10.1016/j.jbi.2020.103466)

# Different data types and analysis techniques



What data is missing in this overview?

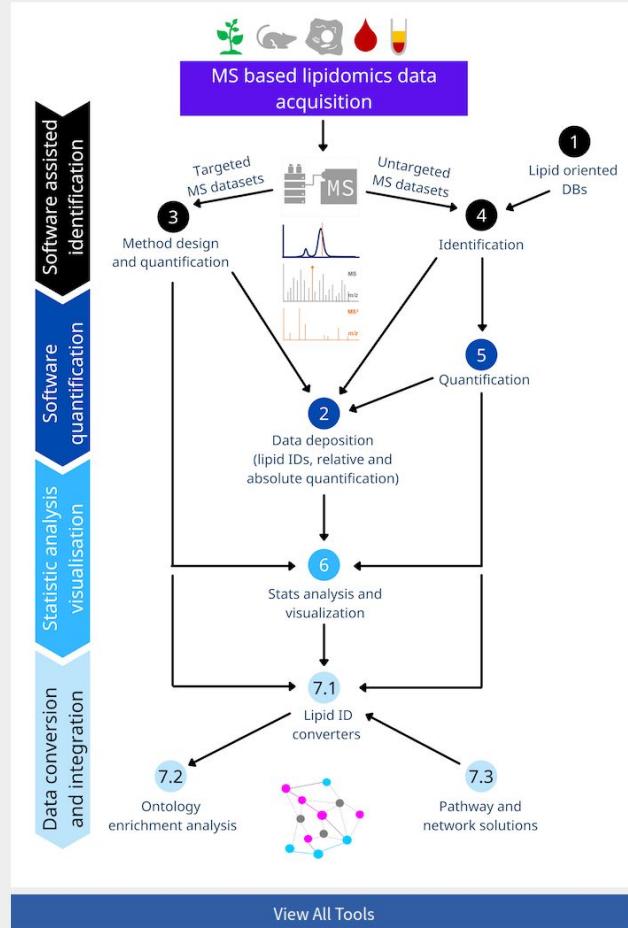
# Different data types and analysis techniques



## Missing:

- Phenotype
- Imaging data
- Fluxomics
- ...

Adapted from: Momeni, Zahra, et al. "A survey on single and multi omics data mining methods in cancer data classification." Journal of Biomedical Informatics 107 (2020): 103466. DOI: [10.1016/j.jbi.2020.103466](https://doi.org/10.1016/j.jbi.2020.103466)

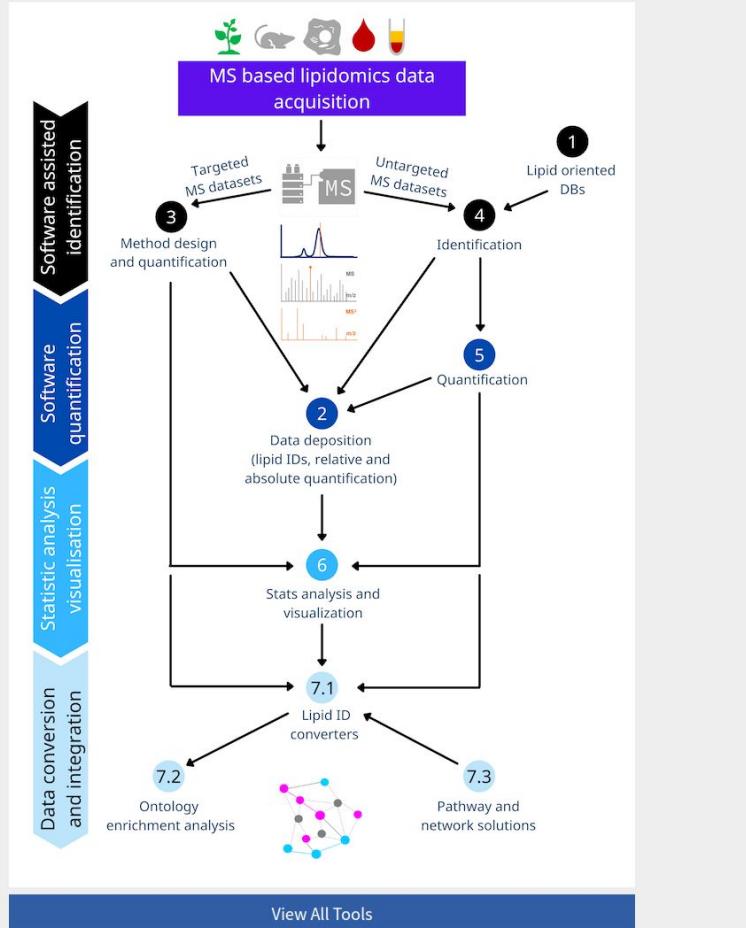


# Overview for Lipidomics analysis tools exists!

## For metabolomics not so much (unfortunately!)

Adapted from: Ni, Zhixu, et al. "Guiding the choice of informatics software and tools for lipidomics research applications." *Nature methods* 20.2 (2023): 193-204. DOI: [10.1038/s41592-022-01710-0](https://doi.org/10.1038/s41592-022-01710-0)

MS Analysis	Structure Drawing	Statistical Analysis
Nomenclature	LIPID MAPS® Software	Lipidomics Tools Guide



# Considerations when comparing tools:

- License & Source Code
- Graphical User Interface (GUI)
- Command Line Interface (CLI)
- Desktop client / web interface
- Input & output formats
- Operating Systems (Windows, Mac, Linux)
- Programming Language (R, Python, Java, Matlab, ...)
- Coverage and IDs used

Adapted from: Ni, Zhixu, et al. "Guiding the choice of informatics software and tools for lipidomics research applications." *Nature methods* 20.2 (2023): 193-204. DOI: [10.1038/s41592-022-01710-0](https://doi.org/10.1038/s41592-022-01710-0)

# Coverage of pathway data (according to RaMP, merging information from 4 pathway databases)

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0

A	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442

B

Total distinct compounds <sup>b</sup>	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022
Chemical properties <sup>c</sup>	256 592	217 776	13 066

- a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).
- b Distinct InChIKeys.
- c Chemical properties are only captured for compounds referenced within RaMP.

# Coverage of pathway data (according to RaMP, merging information from 4 pathway databases)

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0

	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
# Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442

B

Total distinct compounds <sup>b</sup>	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties <sup>c</sup>	256 592	217 776	13 066	44 981

a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

b Distinct InChIKeys.

c Chemical properties are only captured for compounds referenced within RaMP.

# Coverage of pathway data (according to RaMP, merging information from 4 pathway databases)

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0

A	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
# Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442
B					
Total distinct compounds <sup>b</sup>	256 592	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties <sup>c</sup>	217 776	13 066	44 981		

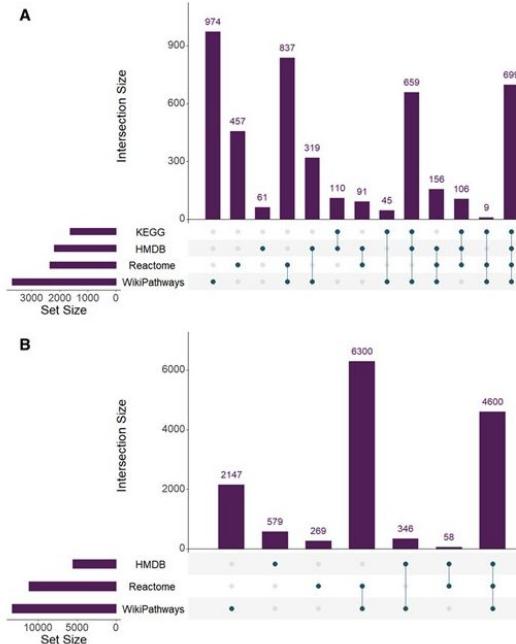
a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

b Distinct InChIKeys.

c Chemical properties are only captured for compounds referenced within RaMP.

Adapted from: Braisted, John, et al. "RaMP-DB 2.0: a renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes." Bioinformatics 39.1 (2023). DOI: [10.1093/bioinformatics/btac726](https://doi.org/10.1093/bioinformatics/btac726)

**Fig 3.**



[Open in new tab](#)

[Download slide](#)

Overlap in content among source databases. Only analytes mapping to pathways are considered, as HMDB contains a large number of metabolites associated only with ontologies, which are not relevant to Reactome and WikiPathways as pathway-centric databases. (A) Overlap in metabolites associated with at least one pathway between source databases in RaMP. (B) Overlap of genes associated with at least one pathway. The filled circle(s) underneath each bar in the plots demonstrate the source databases that the analyte counts are drawn from

# Some existing tools in metabolomics analysis

# Based on pathway databases in RaMP

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0.

A	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
# Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442

B
Total distinct compounds <sup>b</sup>
HMDB v5.0
ChEBI release 212
LIPID MAPS release July 13, 2022

Chemical properties <sup>c</sup>	256 592	217 776	13 066	44 981
----------------------------------	---------	---------	--------	--------

a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

b Distinct InChIKeys.

c Chemical properties are only captured for compounds referenced within RaMP.

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING																				
	<p><b>18 modules, Free to use</b></p> <table border="1"> <tr> <td>Input Data Type</td> <td>LC-MS Spectra (mzXML, mzML, or mzData)</td> <td>MS Peaks (peak list or intensity table)</td> <td>Spectra Processing (LC-MS with R&amp;G)</td> </tr> <tr> <td>Generic Format</td> <td>Peak Annotation (mzML/DEQUAM)</td> <td>Functional Analysis (LC-MS)</td> <td>Functional Meta-analysis (LC-MS)</td> </tr> <tr> <td>Annotated Features (metabolite list or table)</td> <td>Statistical Analysis (metabolite list)</td> <td>Biomarker Analysis</td> <td>Statistical Meta-analysis</td> </tr> <tr> <td>Link to Genomics &amp; Phenotypes (metabolite list)</td> <td>Enrichment Analysis</td> <td>Pathway Analysis</td> <td>Network Analysis</td> </tr> <tr> <td></td> <td></td> <td>Causal Analysis (Metabolite concentration)</td> <td></td> </tr> </table>	Input Data Type	LC-MS Spectra (mzXML, mzML, or mzData)	MS Peaks (peak list or intensity table)	Spectra Processing (LC-MS with R&G)	Generic Format	Peak Annotation (mzML/DEQUAM)	Functional Analysis (LC-MS)	Functional Meta-analysis (LC-MS)	Annotated Features (metabolite list or table)	Statistical Analysis (metabolite list)	Biomarker Analysis	Statistical Meta-analysis	Link to Genomics & Phenotypes (metabolite list)	Enrichment Analysis	Pathway Analysis	Network Analysis			Causal Analysis (Metabolite concentration)		Metabolite ID conversion <a href="https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml">https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml</a>
Input Data Type	LC-MS Spectra (mzXML, mzML, or mzData)	MS Peaks (peak list or intensity table)	Spectra Processing (LC-MS with R&G)																			
Generic Format	Peak Annotation (mzML/DEQUAM)	Functional Analysis (LC-MS)	Functional Meta-analysis (LC-MS)																			
Annotated Features (metabolite list or table)	Statistical Analysis (metabolite list)	Biomarker Analysis	Statistical Meta-analysis																			
Link to Genomics & Phenotypes (metabolite list)	Enrichment Analysis	Pathway Analysis	Network Analysis																			
		Causal Analysis (Metabolite concentration)																				

# Based on pathway databases in RaMP

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0.

**A**

	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
# Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442

**B**

Total distinct compounds <sup>b</sup>	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022
---------------------------------------	-----------	-------------------	----------------------------------

Chemical properties<sup>c</sup>

a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

b Distinct InChIKeys.

c Chemical properties are only captured for compounds referenced within RaMP.

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING															
<p>The Human Metabolome Database (hmdb), PathBank, DRUGBANK, and Small Molecule Pathway Database logos.</p>	<p><b>Input Data Type</b></p> <ul style="list-style-type: none"> <li>LC-MS Spectra (mzXML, mzML, or mzData)</li> <li>MS Peaks (peak list or intensity table)</li> <li>Genetic Format (csv or .txt table files)</li> <li>Annotated Features (metabolite list or table)</li> <li>Link to Genomics &amp; Phenotypes (metabolite list)</li> </ul> <p><b>18 modules, Free to use</b></p> <table border="1"> <tr> <td></td> <td>Peak Annotation (MS/MS Deconv)</td> <td>Spectra Processing (LC-MS with MS2)</td> <td>Functional Analysis (LC-MS)</td> <td>Functional Meta-analysis (LC-MS)</td> </tr> <tr> <td>Statistical Analysis (gene factors)</td> <td>Statistical Analysis (metabolite table)</td> <td>Biomarker Analysis</td> <td>Statistical Meta-analysis</td> <td>Dose Response Analysis</td> </tr> <tr> <td></td> <td>Enrichment Analysis</td> <td>Pathway Analysis</td> <td>Causal Analysis (Metabolite concentration)</td> <td>Network Analysis</td> </tr> </table>		Peak Annotation (MS/MS Deconv)	Spectra Processing (LC-MS with MS2)	Functional Analysis (LC-MS)	Functional Meta-analysis (LC-MS)	Statistical Analysis (gene factors)	Statistical Analysis (metabolite table)	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis		Enrichment Analysis	Pathway Analysis	Causal Analysis (Metabolite concentration)	Network Analysis	<p>Metabolite ID conversion</p> <p><a href="https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml">https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml</a></p>
	Peak Annotation (MS/MS Deconv)	Spectra Processing (LC-MS with MS2)	Functional Analysis (LC-MS)	Functional Meta-analysis (LC-MS)													
Statistical Analysis (gene factors)	Statistical Analysis (metabolite table)	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis													
	Enrichment Analysis	Pathway Analysis	Causal Analysis (Metabolite concentration)	Network Analysis													
<p>Kyoto Encyclopedia of Genes and Genomes (KEGG) logo.</p> <p>Reaction scheme showing L-threonine conversion to N-acetyl-L-threonine:</p> <p>L-threonine + O<sub>2</sub> → N-acetyl-L-threonine + H<sub>2</sub>O</p> <p>Legend: Reaction center (blue circle), Difference arrow (yellow circle), Matched arrow (orange circle), C-C bond (green line).</p>	<p>Licence required; Academic usage Of website for free</p>	<p>KEGG Mapper</p> <p><a href="https://www.genome.jp/kegg/mapper/">https://www.genome.jp/kegg/mapper/</a></p>															

# Based on pathway databases in RaMP

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0.

## A

	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
# Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442

## B

Total distinct compounds <sup>b</sup>	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022
Chemical properties <sup>c</sup>	256 592	217 776	13 066

a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

b Distinct InChIKeys.

c Chemical properties are only captured for compounds referenced within RaMP.

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING															
	<p><b>Input Data Type</b></p> <p>LC-MS Spectra (mzXML, mzML, or mzData)</p> <p>MS Peaks (peak list or intensity table)</p> <p>Genomic Format (csv or .txt table files)</p> <p>Annotated Features (metabolite list or table)</p> <p>Link to Genomics &amp; Phenotypes (metabolite list)</p> 	<p><b>18 modules, Free to use</b></p> <table border="1"> <tr> <td></td> <td>Peak Annotation (MS/MS Deconv)</td> <td>Spectra Processing (LC-MS with MS2)</td> <td>Functional Analysis (LC-MS)</td> <td>Metabolite ID conversion</td> </tr> <tr> <td>Statistical Analysis (gene factors)</td> <td>Statistical Analysis (metabolite table)</td> <td>Biomarker Analysis</td> <td>Statistical Meta-analysis</td> <td>Dose Response Analysis</td> </tr> <tr> <td></td> <td>Enrichment Analysis</td> <td>Pathway Analysis</td> <td>Causal Analysis (Metabolite concentration)</td> <td>Network Analysis</td> </tr> </table>		Peak Annotation (MS/MS Deconv)	Spectra Processing (LC-MS with MS2)	Functional Analysis (LC-MS)	Metabolite ID conversion	Statistical Analysis (gene factors)	Statistical Analysis (metabolite table)	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis		Enrichment Analysis	Pathway Analysis	Causal Analysis (Metabolite concentration)	Network Analysis
	Peak Annotation (MS/MS Deconv)	Spectra Processing (LC-MS with MS2)	Functional Analysis (LC-MS)	Metabolite ID conversion													
Statistical Analysis (gene factors)	Statistical Analysis (metabolite table)	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis													
	Enrichment Analysis	Pathway Analysis	Causal Analysis (Metabolite concentration)	Network Analysis													
	<p>L-threonine OXIDATION</p> <p><chem>CC(=O)N[C@@H](C)C</chem> → <chem>CC(=O)N([O-])[C@@H](C)C</chem></p> <p>Reaction center: N+ - N(O) Difference atom: HO - CH2 Matched atom: C(C) - C(O)</p>	<p><b>Licence required; Academic usage Of website for free</b></p>															
		<p><b>KEGG Mapper</b></p> <p><a href="https://www.genome.jp/kegg/mapper/">https://www.genome.jp/kegg/mapper/</a></p> <p><b>Integrated in analysis tool, downloadable files available</b></p> <p><a href="https://reactome.org/download-data">https://reactome.org/download-data</a></p>															

# Based on pathway databases in RaMP

**Table 2.** Number of analytes and pathways (A) and chemical properties (B) available through RaMP-DB 2.0.

## A

	Total <sup>a</sup>	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	-	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
# Metabolite-pathway mappings	412 775 (+343 120)	367 609	1714	30 804	12 648
# Gene-pathway mappings	401 303 (-695 287)	208 211	8479	125 171	59 442

## B

	Total distinct compounds <sup>b</sup>	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022
Chemical properties <sup>c</sup>	256 592	217 776	13 066	44 981

a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

b Distinct InChIKeys.

c Chemical properties are only captured for compounds referenced within RaMP.

Adapted from: Braisted, John, et al. "RaMP-DB 2.0: a renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes." Bioinformatics 39.1 (2023). DOI: [10.1093/bioinformatics/btac726](https://doi.org/10.1093/bioinformatics/btac726)

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING
	<p><b>Input Data Type</b></p> <ul style="list-style-type: none"> <li>LC-MS Spectra (mzXML, mzML, or mzData)</li> <li>MS Peaks (peak list or intensity table)</li> <li>Genomic Format (csv or .txt table files)</li> <li>Annotated Features (metabolite list or table)</li> <li>Link to Genomics &amp; Phenotypes (metabolite list)</li> </ul> <p><b>18 modules, Free to use</b></p>	Metabolite ID conversion <a href="https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml">https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml</a>
	<p>L-ornithine ODES</p> <p><math>\text{NH}_2-\text{CH}_2-\text{CH}(\text{NH}_2)-\text{COOH} \rightarrow \text{NH}_2-\text{CH}(\text{NH}_2)-\text{COO}^+</math></p> <p>Reaction center: <math>\text{N}=\text{N}^+</math>; Difference atom: <math>\text{HO}-\text{CH}_2</math>; Matched atom: <math>\text{C}=\text{O}</math></p> <p><b>Licence required; Academic usage Of website for free</b></p>	KEGG Mapper <a href="https://www.genome.jp/kegg/mapper/">https://www.genome.jp/kegg/mapper/</a>
	<p><b>Analysis tools</b></p> <ul style="list-style-type: none"> <li>Analyze gene list</li> <li>GO</li> <li>Protein Interaction</li> <li>Species Comparison</li> <li>Protein Domains</li> <li>Protein Evolution</li> </ul>	Integrated in analysis tool, downloadable files available <a href="https://reactome.org/download-data">https://reactome.org/download-data</a>
		<a href="https://www.bridgedb.org/">https://www.bridgedb.org/</a>

# Join our quiz!

Join at [menti.com](https://menti.com) | use code 75 51 85 5

Mentimeter

## Instructions

Go to

**www.menti.com**

Enter the code

**75 51 85 5**



Or use QR code

# Okay, so there are many tools, great!

But wait...

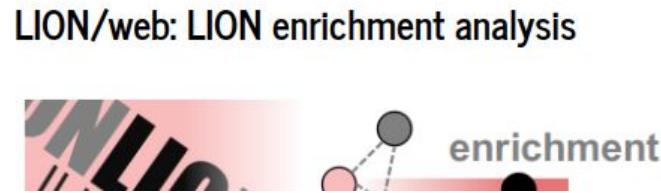
# Okay, so there are many tools, great!

But wait...

Tools can require data to be added in different formats,  
using different IDs, etc.

Samples	Group	CAR(16:0)	CAR(18:0)	CAR(18:1)	CAR(18:2)
S001_2	Affected/Male	32592	7400	25164	16371 39797 461580 342255
S002_27	Affected/Male	37821	13552	40988	26845 51799 526923 409751
S007_51	Affected/Male	9201	6037	6219	10361 18848 461700 168391
S008_59	Affected/Male	132519	15845	245076	159627 24173 437630 326360
S009_39	Affected/Male	24407	9146	51668	32965 42774 337701 362332
S013_29	Affected/Male	30813	7299	35485	25603 58491 386359 385114
S014_22	Affected/Male	33082	8830	36894	21874 49050 542047 420069
S015_5	Affected/Male	29115	7472	38326	23507 35022 230142 298691

A	B	C	D	E	F	
	Nuclei_Control	Nuclei_Control	Nuclei_Control	Nuclei_KLA	Nuclei_KLA	Nucle
#1	#2	#3	#4	#5	#6	
3 GPA(30:1)	0.81091749	1.08513601	1.533164135	1.399345645	1.489582453	1.27
4 GPA(30:0)	8.26E-05	8.07E-05	7.67E-05	7.7E-05	7.45E-05	
5 GPA(32:4)	8.26E-05	8.07E-05	7.67E-05	1.82	7.45E-05	
5 GPA(32:1)	0.375	0.44	7.67E-05	7.7E-05	7.45E-05	
7 GPA(32:0)	2.1	3.52	3.62	4.2	3.39	
3 GPA(34:2)	0.195220877	0.278616003	0.334508539	0.321849498	0.270833173	0.25
9 GPA(34:1)	1.1713253	1.61304	1.6725427	2.3788876	1.8958322	1.
0 GPA(34:0)	1.35	2.64	2.37	2.66	1.76	



# Okay, so there are many tools, great!

But wait...

Tools can require data to be added in different formats,  
using different IDs, etc.

Tip: check the example/tutorial data to find out what is  
expected

# Data processing steps: scaling



Scaling: various techniques for transforming the range of data values. Includes normalization, standardization (Z-score scaling), Min-Max scaling, and robust scaling.

Normalization	Standardization
Scales the data using minimum and maximum values.	Scales the data using the mean and standard deviation.
Values between [0, 1] and [-1, 1].	No specific range
Easily compare findings within and across several data sets	Enables reliable data transmission across various systems
Outliers can affect the range of the data, however these may not skew the entire range as significantly as they would in Z-score scaling.	Outliers can potentially skew the mean and standard deviation, affecting the scaling process.

# Data processing steps: scaling



Scaling: various techniques for transforming the range of data values. Includes normalization, standardization (Z-score scaling), Min-Max scaling, and robust scaling.

Normalization	Standardization
Scales the data using minimum and maximum values.	Scales the data using the mean and standard deviation.
Values between [0, 1] and [-1, 1].	No specific range
Easily compare findings within and across several data sets	Enables reliable data transmission across various systems
<b>Outliers</b> can affect the range of the data, however these may not skew the entire range as significantly as they would in Z-score scaling.	<b>Outliers</b> can potentially skew the mean and standard deviation, affecting the scaling process.

# Data processing steps: Outliers

Data point that significantly differs from other observations in a dataset, outside the overall pattern of the data. Reasons for this: measurement errors, experimental errors, natural variability, genuine extreme values in the data (e.g. IEMs/IMDs).

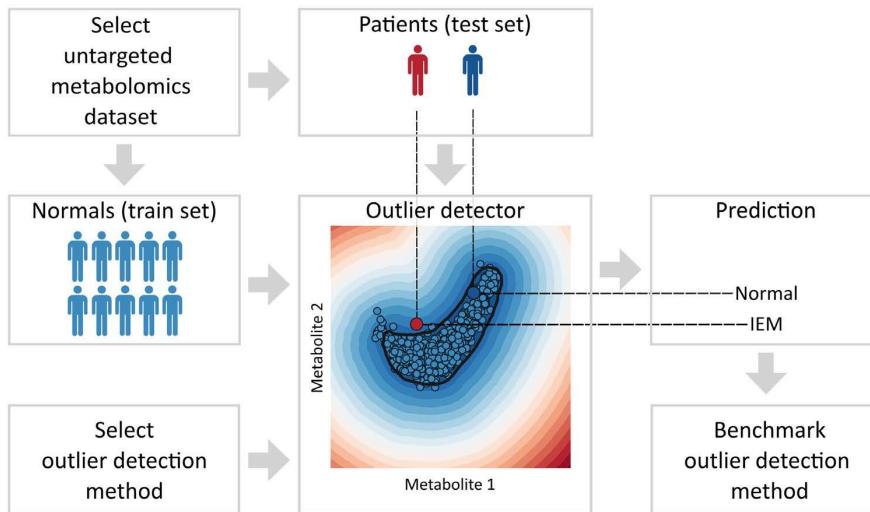


Figure obtained from:  
Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." *Metabolites* 13.1 (2023): 97.  
<https://doi.org/10.3390/metabo13010097>

# Data processing steps: Outliers

Data point that significantly differs from other observations in a dataset, outside the overall pattern of the data. Reasons for this: measurement errors, experimental errors, natural variability, genuine extreme values in the data.

Which one to pick?

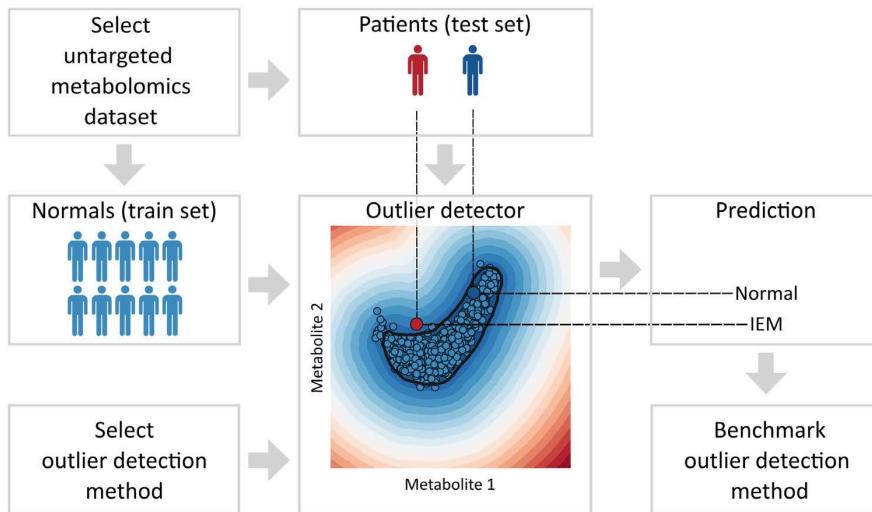
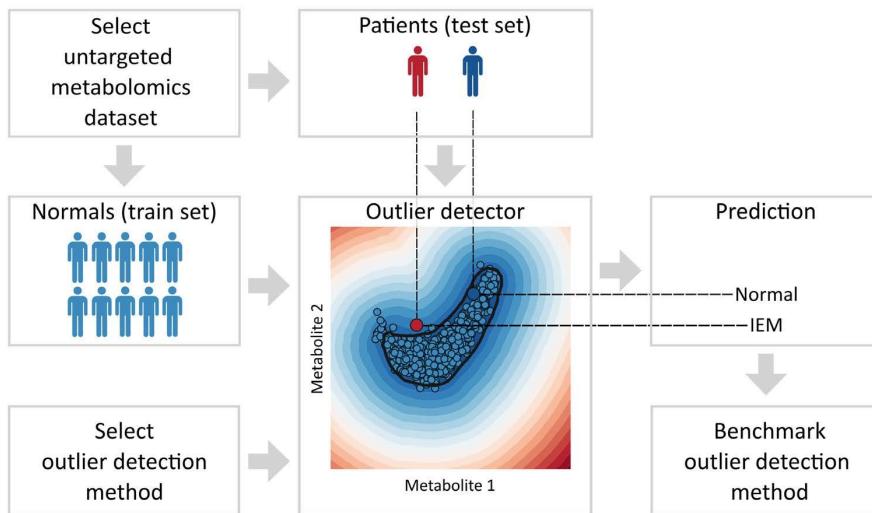


Figure obtained from:  
Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." *Metabolites* 13.1 (2023): 97.  
<https://doi.org/10.3390/metabo13010097>

# Data processing steps: Outliers

Data point that significantly differs from other observations in a dataset, outside the overall pattern of the data. Reasons for this: measurement errors, experimental errors, natural variability, genuine extreme values in the data.



## Which one to pick?

Depends on:

- Your Data
- Your Research Question
- “Standard(s)” in the field
- General Data Distribution

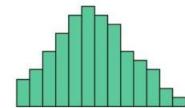
Figure obtained from:

Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." *Metabolites* 13.1 (2023): 97.  
<https://doi.org/10.3390/metabo13010097>

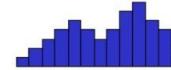
# Data processing steps: distribution

Many data distributions exist:

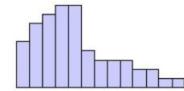
- Normal Distribution (Gaussian Distribution)
- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Log-Normal Distribution
- ...



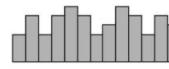
The shape has a bell shape.  
It is *symmetric*.



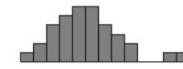
The shape has two humps.  
It is *bimodal*.



The shape has a long tail.  
It is *not symmetric*.



The shape is *flat*.

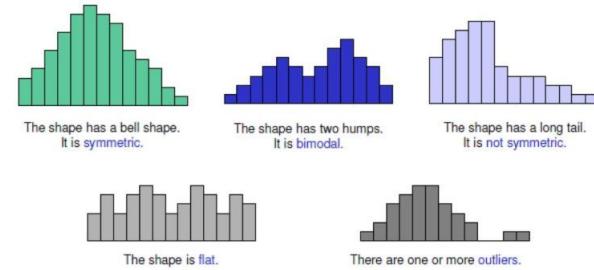


There are one or more *outliers*.

# Data processing steps: distribution

Many data distributions exist:

- Normal Distribution (Gaussian Distribution)
- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Log-Normal Distribution
- ...



## But, which one is it?

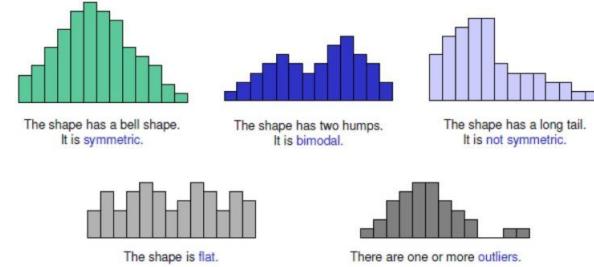
Graphical and statistical methods exist to check:

- Histogram, Boxplot, Scatterplots, Density plots
- Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling

# Data processing steps: distribution

Many data distributions exist:

- Normal Distribution (Gaussian Distribution)
- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Log-Normal Distribution
- ...



## But, which one is it?

Graphical and statistical methods exist to check:

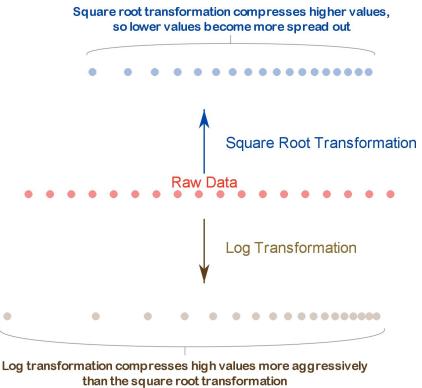
- Histogram, Boxplot, Scatterplots, Density plots
- Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling

Can you change the distribution of your data?

# Data processing steps: transformation

Again, many data transformation techniques exist:

- Logarithmic ( $\log_2$ ,  $\log_{10}$ )
- Square Root/ Cube Root
- Exponential
- Rank
- Box-Cox
- ...



# Data processing steps: transformation

Again, many data transformation techniques exist:

- Logarithmic (log2, log10)
- Square Root/ Cube Root
- Exponential
- Rank
- Box-Cox
- ...

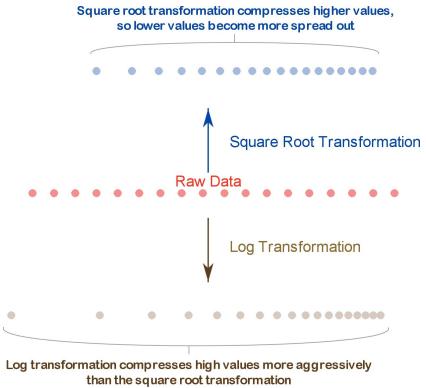
## Which one to use?

Depends on:

- characteristics of the data,
- goal(s) of the analysis,
- assumptions of (later used) statistical method
- 

General rule of thumb:

Experiment with different transformations and evaluate their effects on the data distribution

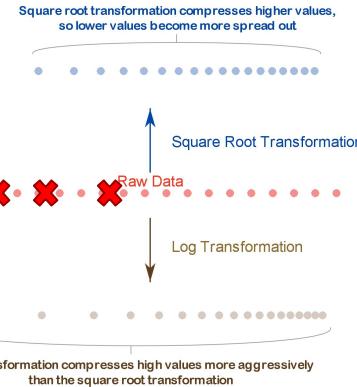


# Data processing steps: transformation

Again, many data transformation techniques exist:

- Logarithmic (log2, log10)
- Square Root/ Cube Root
- Exponential
- Rank
- Box-Cox
- ...

But, what if you have (many) missing data points?



# Data processing steps: missing data

- Identify missing values and their annotations (e.g. "NaN" (Not a Number), "NA" (Not Available), blank cells, software specific error codes)
- Quantify Missing Values: how many are there?
- Why are there missing values in your dataset?

Student_Id	Math	English
1	70	60
2		55
3	45	NaN
4	75	50
5	999999	75
6	90	X
7	95	80
8	??	57
9	80	na
10	n/a	64

Highlighted all the missing values in the dataset

Student_Id	Math	English
1	70	60
2		55
3	45	NaN
4	75	50
5	999999	75
6	90	X
7	95	80
8	??	57
9	80	na
10	n/a	64

Highlighted only the standard missing values in the dataset

# Data processing steps: missing data

- Identify missing values and their annotations (e.g. "NaN" (Not a Number), "NA" (Not Available), blank cells, software specific error codes)
- Quantify missing values: how many are there?
- Why are there missing values in your dataset?
- Replace “wrong” data if possible (check decimal separator!!); keep copy of original data
- Delete specific Rows or Columns:
  - Only for few and randomly distributed missing data are found && when sufficient data points remain.
- Replacing missing values with estimated (e.g. mean, median) or imputed values (predict missing values, e.g. using regression, K-nearest neighbors)

Student_Id	Math	English
1	70	60
2		55
3	45	NaN
4	75	50
5	999999	75
6	90	X
7	95	80
8	??	57
9	80	na
10	n/a	64

Highlighted all the missing values in the dataset

Student_Id	Math	English
1	70	60
2		55
3	45	NaN
4	75	50
5	999999	75
6	90	X
7	95	80
8	??	57
9	80	na
10	n/a	64

Highlighted only the standard missing values in the dataset

# Join our quiz!

Join at [menti.com](https://menti.com) | use code 75 51 85 5

Mentimeter

## Instructions

Go to

**www.menti.com**

Enter the code

**75 51 85 5**



Or use QR code

# Data processing steps (after identification): duplicates?!?!

How to deal with duplicate identifications of MS (/MS) peaks:

- Identify if there are duplicates Names/ IDs
- Review how these duplicates came part of the dataset
- Remove redundant rows
- Merge duplicates for multiple instances of same entity (average of values)
- Retain unique duplicates from entry errors or inconsistencies (label them)
- Don't forget about secondary IDs...

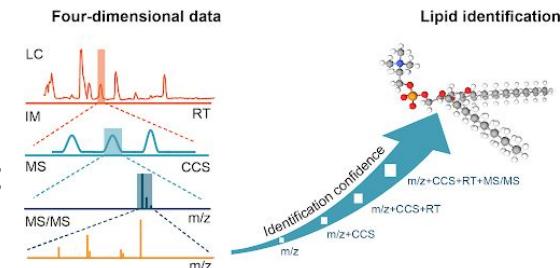
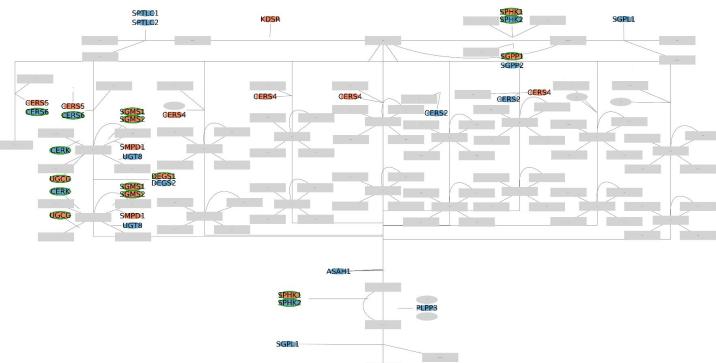
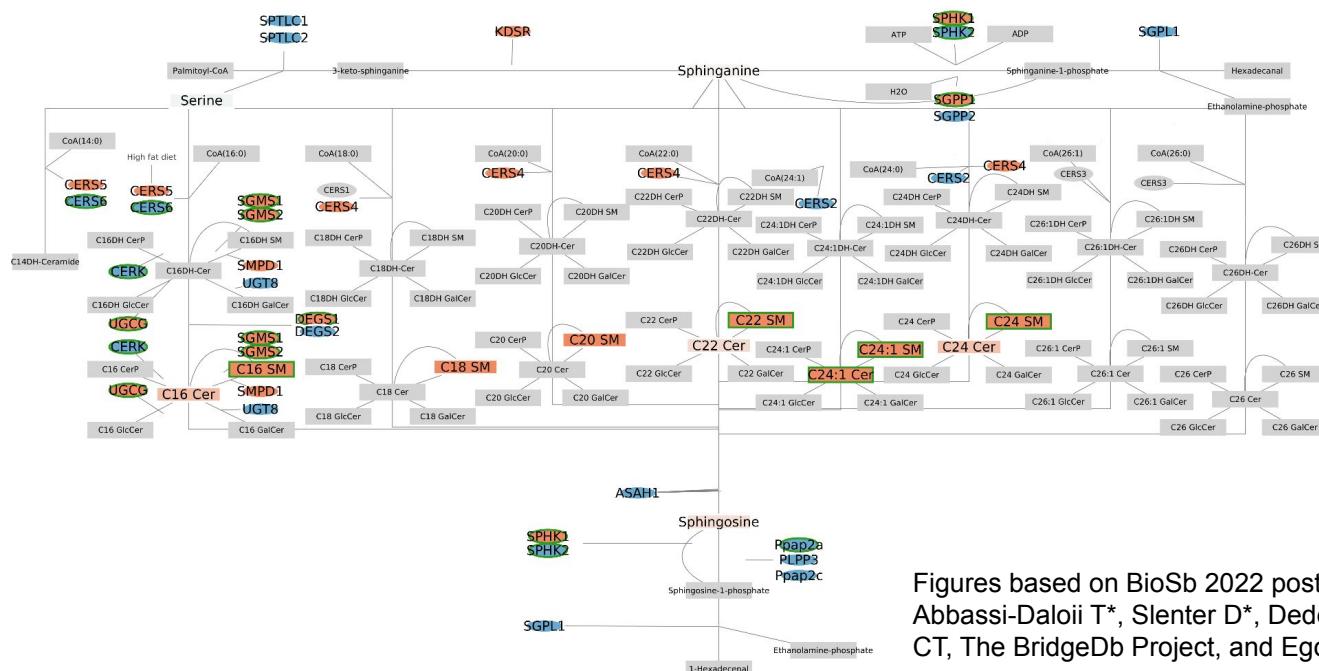


Figure obtained from:  
Chen, Xi, Yandong Yin, and Zheng-Jiang Zhu. "Lipid4DAnalyzer Tutorial." (2022).  
<http://lipid4danalyzer.zhulab.cn/>

# Secondary IDs?

Growing issue in (meta) analysis of biological data

- (1) entries withdrawn/deleted from a database,
- (2) entries split/merged in a database,
- (3) entries referring to the same entity ('unknown' duplicates)



Enhanced multi-omics visualization with BridgeDb identifier mapping adding data to:

2 proteins and 13 metabolites

Figures based on BioSb 2022 poster, authors:  
Abbassi-Daloii T\*, Slenter D\*, Dede Sener D, Basaric H, Kutmon M, Evelo CT, The BridgeDb Project, and Egon Willighagen

# Variance; when to check and what to do?

Data processing: Before performing any statistical analysis or modeling. Identify features or variables with low variability, which do not contribute to your model

Feature Selection: Identify the most informative features for predicting the target variable. High-variance features are often preferred, contain more information may be relevant predictors.

Modeling Assumptions: Some statistical models assume that the variance of the residuals (i.e., the difference between observed and predicted values) is constant across the range of predictors.

Method	Description	Pros	Cons
Descriptive Statistics	Calculate basic statistics such as mean, median, standard deviation, and range for each variable.	- Provides summary measures of variability.	- Does not provide visual representation of data distribution. - May not capture the shape of the distribution or identify outliers.
Box Plots	Visualize the distribution of values for each variable using box plots.	- Provides visual representation of data distribution, including median, quartiles, and outliers.	- Limited to univariate analysis and may not capture multivariate relationships.
Histograms	Plot histograms to visualize the frequency distribution of values within each variable.	- Shows the shape and spread of the data distribution.	- May vary in appearance depending on binning choices. - Limited to univariate analysis.
Coefficient of Variation (CV)	Compute the ratio of the standard deviation to the mean, expressed as a percentage.	- Provides a measure of relative variability, allowing comparison of variability across variables.	- May not be meaningful for variables with zero mean. - Sensitive to the scale of measurement.
Interquartile Range (IQR)	Calculate the difference between the third quartile (Q3) and the first quartile (Q1) in the data distribution.	- Robust to outliers and extreme values. - Provides a measure of variability within the middle 50% of the data.	- Does not capture variability in the tails of the distribution.
Variance Inflation Factor (VIF)	Assess multicollinearity between predictor variables in regression analysis.	- Helps identify high collinearity between variables, which can affect the reliability of regression coefficients.	- Applies specifically to regression analysis and may not be relevant for other types of data analysis. - Does not provide information about variability in the data distribution.

# Correlated data; when to check and what to do?

Hypothesis generation: Exploratory data analysis, understand the relationships between variables

Feature selection for predictive modeling: highly correlated variable lead to multicollinearity issues, affecting model performance and interpretability

Identify redundant information: Removing highly correlated variables to simplify models and improve interpretability without losing predictive performance.

Method	Description	Pros	Cons
Correlation Matrix	Compute pairwise correlations between all pairs of variables in the dataset, represented in a matrix format.	- Provides a comprehensive overview of all pairwise correlations in the dataset.	- Can be computationally intensive for large datasets.
Scatter Plots	Create visual representations of the relationship between pairs of variables using scatter plots.	- Provides a direct visualization of relationships between variables. - Suitable for identifying linear and nonlinear associations.	- May be less effective for datasets with many variables, as it requires examining multiple scatter plots.
Correlation Coefficients	Calculate correlation coefficients (e.g., Pearson, Spearman, Kendall) to quantify the strength and direction of relationships between variables.	- Quantifies the strength and direction of relationships numerically. - Offers flexibility with different correlation coefficients suitable for various types of data and relationships.	- Pearson correlation may not capture nonlinear relationships. - Spearman and Kendall correlations may be less sensitive to outliers and non-normality but may be less powerful for detecting linear relationships.
Heatmap	Visualize correlations in a heatmap format, using color gradients to highlight patterns of correlation between variables.	- Offers an intuitive visual representation of correlation patterns. - Suitable for identifying clusters of correlated variables.	- Heatmaps may be challenging to interpret with a large number of variables. - Color scales can influence interpretation, requiring careful selection.
Statistical Tests	Conduct hypothesis tests (e.g., Pearson correlation test, Spearman correlation test) to assess the significance of observed correlations.	- Allows for formal assessment of whether observed correlations are statistically significant.	- Requires assumptions about data distribution and independence. - May be less informative about the strength and direction of relationships compared to correlation coefficients.
Correlation Thresholding	Set a threshold to identify variables with strong correlations, typically based on the absolute value of correlation coefficients.	- Provides a straightforward way to identify highly correlated variables.	- Arbitrary choice of threshold may influence results. - May overlook relationships with moderate but meaningful correlations.
Partial Correlation	Compute partial correlations to assess the relationship between two variables while controlling for the effects of other variables in the dataset.	- Helps uncover direct relationships between variables by removing the effects of confounding variables.	- Requires assumptions about the absence of direct relationships between control variables and the variable of interest. - May not capture complex relationships involving interactions between multiple variables.

# Break time!

# **Hands-on session - setting up laptop**

11:15-12:00 (so only 45 minutes!)

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides

# **Hands-on session - setting up laptop**

11:15-12:00 (so only 45 minutes!)

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides

**First; does everyone have a GitHub Account?**

<https://docs.github.com/en/desktop/installing-and-authenticating-to-github-desktop/setting-up-github-desktop>

# Hands-on session - setting up laptop

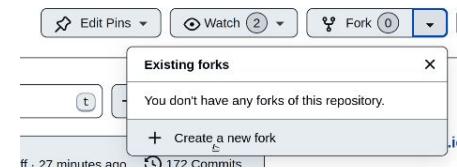
11:15-12:00 (so only 45 minutes!)

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides

Second, find the repository and make a FORK

<https://github.com/NUTRIOME/Workshop1>





# Hands-on session - setting up laptop

Second, find the repository and make a FORK

<https://github.com/NUTRIOME/Workshop1>

## Create a new fork

A fork is a copy of a repository. Forking a repository allows you to freely experiment with changes without affecting the original project.

Required fields are marked with an asterisk (\*).

Owner \*



Repository name \*

/ Workshop1

Workshop1 is available.

By default, forks are named the same as their upstream repository. You can customize the name to distinguish it further.

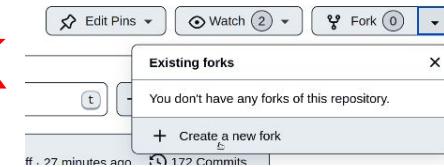
Description (optional)

Copy the main branch only

Contribute back to NUTRIOME/Workshop1 by adding your own branch. [Learn more.](#)

i You are creating a fork in your personal account.

Create fork



Now you are  
allowed to make  
changes to the  
available content!

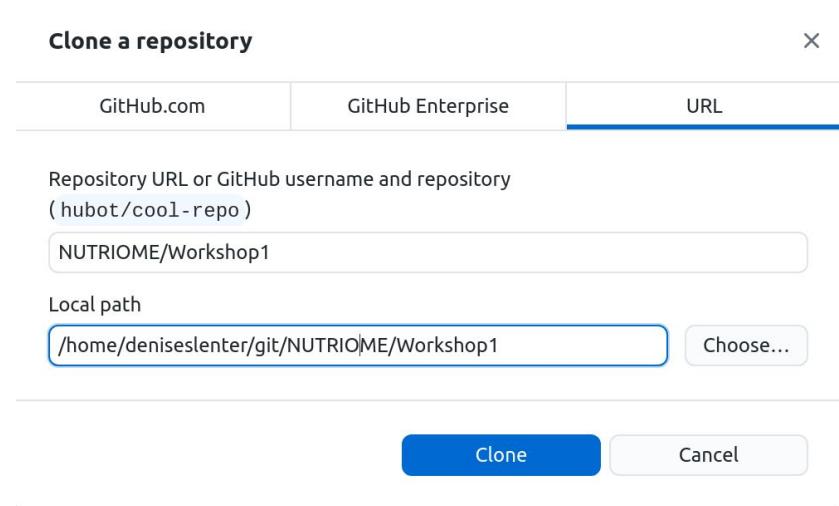
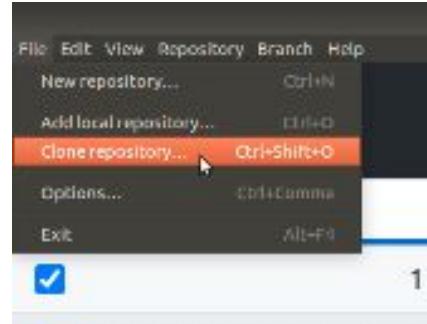
# Hands-on session - GitHub Desktop

- Select File/Clone repository
- In the pop-up menu, select URL (third option from the top), and add:

NUTRIOME/Workshop1

or

\*username/Workshop1



# Hands-on session - GitHub Desktop

The screenshot shows the GitHub Desktop application interface. At the top, there's a menu bar with options: File, Edit, View, Repository, Branch, Help. Below the menu is a header bar with three items: "Current repository Workshop1", "Current branch main", and "Fetch origin Never fetched". The main area is titled "No local changes". It contains a message: "There are no uncommitted changes in this repository. Here are some friendly suggestions for what to do next." Below this message are three suggestions:

- Open the repository in your external editor**  
Select your editor in [Options](#)  
Repository menu or **Ctrl + Shift + A**  
[Open in GNOME Text Editor](#)
- View the files of your repository in your File Manager**  
Repository menu or **Ctrl + Shift + F**  
[Show in your File Manager](#)
- Open the repository page on GitHub in your browser**  
Repository menu or **Ctrl + Shift + G**  
[View on GitHub](#)

At the bottom left, there's a "Summary (required)" field with a user icon, a "Description" field, and a "Commit to main" button.

You should have something that looks like this now!

# Hands-on session - GitHub Desktop

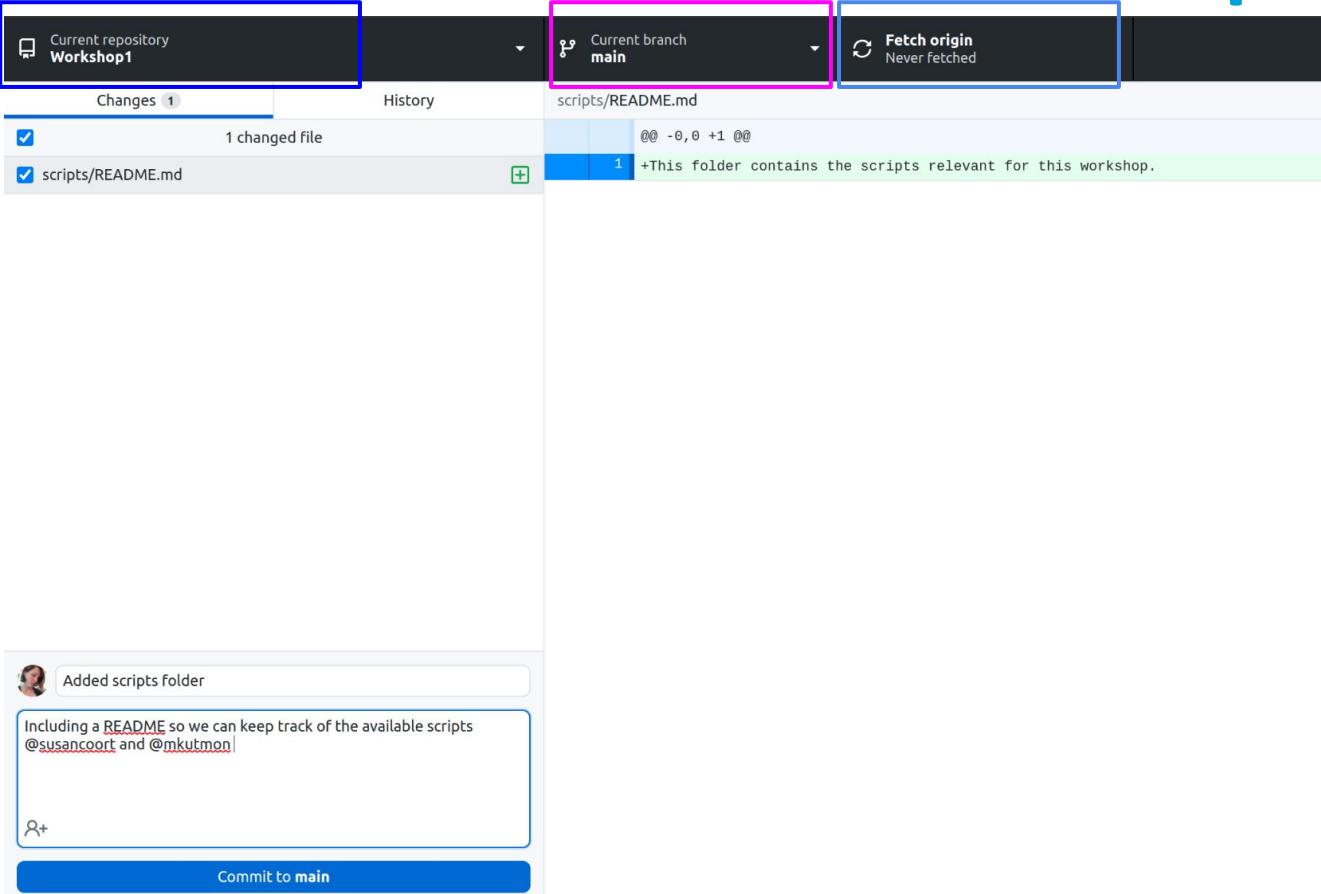
The screenshot shows the GitHub Desktop application interface. At the top, there are three dropdown menus: 'Current repository' set to 'Workshop1', 'Current branch' set to 'main', and 'Fetch origin' status 'Never fetched'. Below these are two tabs: 'Changes' (selected) and 'History'. The 'Changes' tab shows one changed file, 'scripts/README.md'. The commit message in the main pane reads:

```
scripts/README.md
@@ -0,0 +1 @@
1 +This folder contains the scripts relevant for this workshop.
```

In the bottom left, a commit history card shows a previous commit by a user named 'Added scripts folder'. The commit message is: "Including a ~~README~~ so we can keep track of the available scripts @susancoort and @mkutmon". There is a placeholder for adding more contributors with 'R+' and a large blue 'Commit to main' button at the bottom.

# Hands-on session - GitHub Desktop

Folder we  
are working  
in



Branch  
("version")  
we are in

Button to  
update local  
and online

# Hands-on session - GitHub Desktop

The screenshot shows the GitHub Desktop application interface. At the top, there are three main status indicators: 'Current repository Workshop1' (highlighted by a blue box), 'Current branch main' (highlighted by a pink box), and 'Fetch origin Never fetched' (highlighted by a blue box). Below these, the main window displays a commit history with one change: 'scripts/README.md'. A green box highlights the file name 'scripts/README.md'. To the right of the commit history, a preview of the file content is shown, enclosed in a red box. The preview text reads: '@@ -0,0 +1 @@' and '1 +This folder contains the scripts relevant for this workshop.' On the left side of the main window, there are two purple boxes: one for 'Folder we are working in' (containing 'Workshop1') and another for 'Number of changes that occurred' (containing 'Changes 1'). Below the main window, a commit message is shown: 'Added scripts folder' with a note: 'Including a README so we can keep track of the available scripts @susancourt and @mkutmon'. A blue box highlights the 'Commit to main' button at the bottom.

Folder we are working in

Number of changes that occurred

Name of changed file(s)

Branch ("version") we are in

Preview of change(s) in selected file

Button to update local and online

# Hands-on session - GitHub Desktop

The screenshot shows the GitHub Desktop application interface with several callout boxes highlighting specific features:

- Folder we are working in**: Points to the "Current repository" dropdown showing "Workshop1".
- Number of changes that occurred**: Points to the "Changes 1" indicator.
- Name of changed file(s)**: Points to the "scripts/README.md" file listed under changes.
- Your GitHub account icon :D**: Points to the user icon in the commit history.
- Short name for the change (commit message)**: Points to the commit message "Added scripts folder".
- Longer description for change**: Points to the detailed commit message "Including a README so we can keep track of the available scripts @susancoort and @mkutmon".
- Branch ("version") we are in**: Points to the "Current branch main" dropdown.
- Preview of change(s) in selected file**: Points to the preview pane showing the file content and diff.
- Button to save changes in local git**: Points to the "Commit to main" button.
- Button to update local and online**: Points to the "Fetch origin" button.

# Hands-on session - GitHub Desktop

The screenshot shows the GitHub Desktop application interface. At the top, there are two dropdown menus: "Current repository" set to "Workshop1" and "Current branch" set to "main". To the right of the branch dropdown is a blue-bordered box containing the text "Push origin" and "Last fetched 15 minutes ago". Below these are two tabs: "Changes" (selected) and "History", with a sub-section showing "0 changed files". In the bottom left corner, a cyan-bordered box contains the text "Button to save changes in local git: Successful!". The main area displays a commit summary for a commit just made:

- Summary (required): [empty input field]
- Description: [empty input field]
- Commit to main: [blue button]
- Committed just now: [text] Added scripts folder
- Undo: [button]

In the center-right, a large blue-bordered box contains the message "No local changes". It includes a small icon of a document with arrows. Below this, there are three suggestions:

- Push commits to the origin remote**: You have 1 local commit waiting to be pushed to GitHub. Always available in the toolbar when there are local commits waiting to be pushed or `Ctrl + P`. Push origin
- Open the repository in your external editor**: Select your editor in [Options](#). Repository menu or `Ctrl + Shift + A`. Open in GNOME Text Editor
- View the files of your repository in your File Manager**: Repository menu or `Ctrl + Shift + F`. Show in your File Manager
- Open the repository page on GitHub in your browser**: Repository menu or `Ctrl + Shift + G`. View on GitHub

On the far right, another blue-bordered box contains the text "Button to update local and online:" followed by the definitions for Pull and Push.

**Button to update local and online:**

Pull = Download from online to local files

Push = Upload from local to online files

# Steps so far in Terminal:

The screenshot shows a GitHub repository named 'Workshop1' (Public). The repository has 1 branch and 0 tags. The file list includes 'Create tutorials.md', '\_layouts', 'assets/css', 'images', 'lectures', and 'pages'. A context menu is open over the 'Create tutorials.md' file, showing options like 'Clone', 'HTTPS', 'SSH', and 'GitHub CLI'. The 'SSH' option is highlighted with a blue box. The URL 'git@github.com:NUTRIOME/Workshop1.git' is also highlighted with a blue box.

Click this button to copy  
(select SSH for secure transfer  
of data)

```
(base) deniseslenter@deniseslenter-HP-EliteBook-840-G2:~/git/NUTRIOME$ git clone
git@github.com:NUTRIOME/Workshop1.git
(base) deniseslenter@deniseslenter-HP-EliteBook-840-G2:~/git/NUTRIOME/Workshop1/
(base) deniseslenter@deniseslenter-HP-EliteBook-840-G2:~/git/NUTRIOME/Workshop1$ git status
on branch main
Your branch is ahead of 'origin/main' by 1 commit.
  (use "git push" to publish your local commits)
nothing to commit, working tree clean
```

Code to check changes in local  
git: Successful!

```
(base) deniseslenter@deniseslenter-HP-EliteBook-840-G2:~/git/NUTRIOME/Workshop1$ git remote set-url origin git@github.com:NUTRIOME/Workshop1.git
(base) deniseslenter@deniseslenter-HP-EliteBook-840-G2:~/git/NUTRIOME/Workshop1$ git push
Warning: the ECDSA host key for 'github.com' differs from the key for the IP address
Offending key for IP in /home/deniseslenter/
Matching host key in /home/deniseslenter/.ssh/
Are you sure you want to continue connecting (yes/no)? yes
Counting objects: 4, done.
Delta compression using up to 4 threads.
Compressing objects: 100% (3/3), done.
Writing objects: 100% (4/4), 448 bytes | 448.00 KiB/s, done.
Total 4 (delta 1), reused 0 (delta 0)
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To github.com:NUTRIOME/Workshop1.git
  9bie3ea..f1bb9ff main -> main
```

In case you failed previously setting up with SSH like me ;)

<https://docs.github.com/en/get-started/getting-started-with-git/managing-remote-repositories>

Code to update local and online:

git pull = Download from online to local files

git push = Upload from local to online files

# Hands-on session - GitHub Desktop

 **Workshop1** Public

 Edit Pins ▾

 Watch (2) ▾

 main ▾

 1 Branch

 0 Tags

 Go to file



Add file ▾

 Code ▾



**DeniseSI22** Added scripts folder



f1bb9ff · 25 minutes ago

 172 Commits

# Hands-on session - GitHub Desktop

Workshop1 Public

Edit Pins

Watch (2)

main

1 Branch 0 Tags

Go to file

t

Add file

Code

DeniseSI22 Added scripts folder · f1bb9ff · 25 minutes ago · 172 Commits

Note: if you get this message in GitHub Desktop, something in your setup is not correct, ask us for help!

## Error



Authentication failed. Some common reasons include:

- You are not logged in to your account: see File > Options.
- You may need to log out and log back in to refresh your token.
- You do not have permission to access this repository.
- The repository is archived on GitHub. Check the repository settings to confirm you are still permitted to push commits.
- If you use SSH authentication, check that your key is added to the ssh-agent and associated with your account.
- If you used username / password authentication, you might need to use a Personal Access Token instead of your account password. Check the documentation of your repository hosting service.

Close

Open options

# Hands-on session - GitHub Desktop

The screenshot shows the GitHub Desktop application interface. At the top, it displays the current repository as "desktop", the current branch as "the-end-of-it-all" (PR #15640), and the status as "Pull origin" (last fetched 5 minutes ago). The main area shows a diff for the file "app/src/ui/diff/seamless-diff-switcher.tsx". The left sidebar lists three changes: "3 changed files" (with checkboxes checked) and "Stashed Changes" (empty). The right pane displays the code diff with line numbers and highlights for added and deleted code.

```
.... @@ -19,6 +19,7 @@ import {  
    import { Loading } from '../lib/loading'  
    import { getFileContents, IFileContents  
} from './syntax-highlighting'  
    import { getTextDiffWithBottomDummyHunk  
} from './text-diff-expansion'  
    + import { textDiffEquals } from './diff-h  
    elpers'  
    ...  
    /**  
     * The time (in milliseconds) we allow w  
     hen loading a diff before  
    .... @@ -127,7 +128,7 @@ function isSameDiff(prevDiff: IDiff, newDiff: IDiff) {  
        prevDiff === newDiff ||  
        (isTextDiff(prevDiff) &&  
        isTextDiff(newDiff) &&  
        - prevDiff.text === newDiff.text)  
    )  
    }  
    ....  
    127 128  prevDiff === newDiff ||  
    128 129  (isTextDiff(prevDiff) &&  
    129 130  isTextDiff(newDiff) &&  
    130 131  + textDiffEquals(prevDiff, newDiff))  
    131 132  )  
    132 133  }  
    133 134  ....
```

At the bottom, there is a "Commit to the-end-of-it-all" button.

# Join our quiz!

Join at [menti.com](https://menti.com) | use code 75 51 85 5

Mentimeter

## Instructions

Go to

**www.menti.com**

Enter the code

**75 51 85 5**



Or use QR code

# Hands-on session - GitHub Desktop



DeniseSI22 / Workshop1

Type ⌘ to search



Button to  
save changes  
in online git

Code Pull requests Actions

Projects Wiki Security Insights Settings

Files

main + Q

Go to file t

- > \_layouts
- > assets
- > images
- > lectures
- > pages
- > scripts

Folder we  
are working  
in

Workshop1 / scripts / README.md in main

Cancel changes

Commit changes...

Edit Preview

Spaces 2 Soft wrap

```
1 This folder contains the scripts relevant for this workshop.  
2  
3 Contents:  
4 - metabolomics processing script (DIY)  
5 - metabolomics processing script (with answers)  
6
```

# Hands-on



DeniseSI22 / Workshop1

Code Pull requests Actions

## Files

main + Q

Go to file t

- > \_layouts
- > assets
- > images
- > lectures
- > pages
- > scripts

README.md

.gitignore

LICENSE

README.md

\_config.yml

index.md

Folder we are working in

## Commit changes

X

### Commit message

Added contents list

### Extended description

Currently has two examples, to be linked to actual content.

Commit directly to the main branch

Create a **new branch** for this commit and start a pull request [Learn more about pull requests](#)

Button to save changes in online git

| > | + | - | 0 | n | e | profile

Cancel changes

Commit changes...

Spaces 2 Soft wrap

Branch ("version") we are in

Cancel

Commit changes

# Hands-on session - GitHub Desktop



Workshop1 Public

forked from [NUTRIOME/Workshop1](#)

Pin

Watch 0

To request a merge from our content to the original one (Pull Request or PR)

main ▾

1 Branch 0 Tags

Go to file

t

Add file ▾

Code ▾

This branch is 1 commit ahead of

NUTRIOME/Workshop1:main .

Contribute ▾

Sync fork ▾



DeniseSI22 Added contents list



dec57b7 · now



173 Commits

To check for updates from our version to the original one

# Hands-on session - GitHub Desktop



Workshop1

Public

forked from [NUTRIOME/Workshop1](#)

Pin

Watch 0

To request a merge from our content to the original one (Pull Request or PR)

main

1 Branch 0 Tags

Go to file

t

Add file

Code

This branch is 1 commit ahead of

NUTRIOME/Workshop1:main .

Contribute

Sync fork



DeniseSI22 Added contents list

dec57b7 · now 173 Commits

To check for updates from our version to the original one

## No ‘upstream’ changes

This branch is 1 commit ahead of NUTRIOME/Workshop1:main .

Contribute Sync fork

This branch is not behind the upstream NUTRIOME/Workshop1:main .

DeniseSI22 Added contents list

_layouts	Update default.html
assets/css	Update style.scss
images	Add files via upload
lectures	Add files via upload
pages	Create tutorials.md
scripts	Added contents list
.gitignore	Update .gitignore

No new commits to fetch. Enjoy your day!

## ONE ‘upstream’ change

This branch is 1 commit ahead of, 1 commit behind NUTRIOME/Workshop1:main .

Contribute Sync fork

This branch is out-of-date

Update branch to merge the latest changes from the upstream repository into this branch.

Discard 1 commit to make this branch match the upstream repository. 1 commit will be removed from this branch.

Learn more about syncing a fork

DeniseSI22 Added contents list

_layouts	Update default.html
assets/css	Update style.scss
images	Add files via upload
lectures	Add files via upload
pages	Create tutorials.md
scripts	Added contents list
.gitignore	Update .gitignore

Discard 1 commit

Update branch

This branch is 2 commits ahead of NUTRIOME/Workshop1:main .

## ‘upstream’ change integrated in fork

# Hands-on session - GitHub Desktop

## Comparing changes

Choose two branches to see what's changed or to start a new pull request. If you need to, you can also compare across forks or learn more about diff comparisons.

base repository: NUTRIOME/Workshop1 ▾ base: main ▾ ... head repository: DeniseSI22/Workshop1 ▾ compare: main ▾ Able to merge. These branches can be automatically merged.

Discuss and review the changes in this comparison with others. [Learn about pull requests](#)

Create pull request

2 commits 1 file changed 1 contributor

Commits on May 27, 2024

Added contents list ···  
DeniseSI22 committed 11 minutes ago  
Merge branch 'NUTRIOME:main' into main  
DeniseSI22 committed 5 minutes ago

Showing 1 changed file with 4 additions and 0 deletions.

scripts/README.md

00 -1 +1,5 00

This folder contains the scripts relevant for this workshop.

```
1 This folder contains the scripts relevant for this workshop.
```

DeniseSI22 added 2 commits 14 minutes ago

Added contents list ···  
Merge branch 'NUTRIOME:main' into main

DeniseSI22 merged commit 2808dcf into NUTRIOME:main now

Pull request closed

If you wish, you can delete this fork of **NUTRIOME/Workshop1** in the [settings](#).

DeniseSI22 added 2 commits 13 minutes ago

Added contents list ···  
Merge branch 'NUTRIOME:main' into main

Add more commits by pushing to the [main](#) branch on **DeniseSI22/Workshop1**.

This branch has not been deployed  
No deployments

This branch has no conflicts with the base branch  
Merging can be performed automatically.

Merge pull request ▾ or view [command line instructions](#).

# Hands-on session - GitHub Desktop

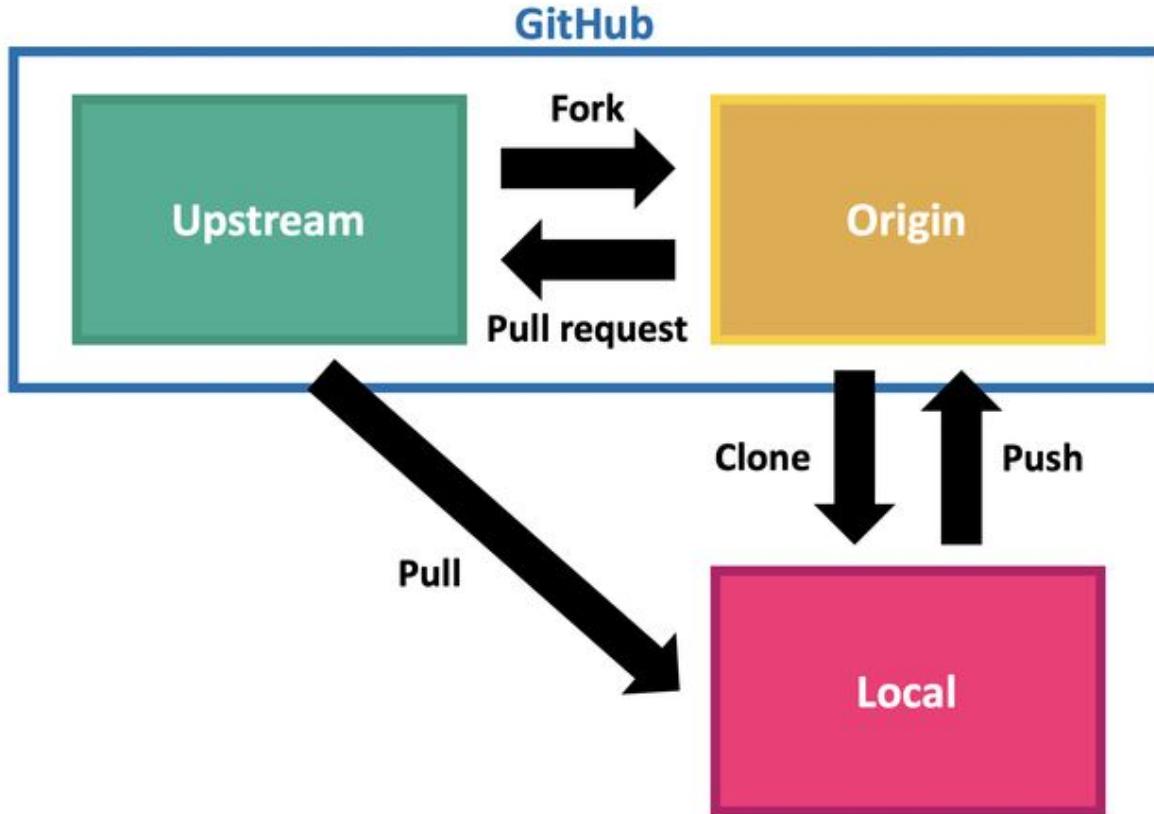


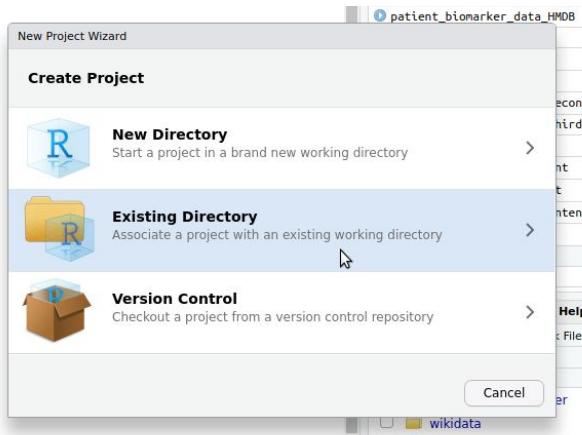
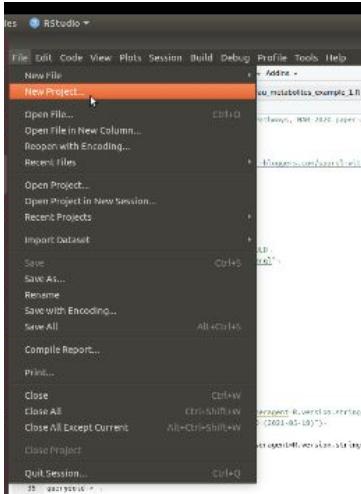
Image from: <https://anujf0510.medium.com/the-guide-to-git-terminology-bbae2a3d8af5> and <https://medium.com/@vishwasacharya/github-desktop-vs-command-line-choose-the-right-tool-for-you-feb58c3f0e30>

# Hands-on session - Adding GitHub to Rstudio

Step 1: Have Rstudio installed

Step 2: Open Rstudio

Step 3: Select File/New Project (top left menu) ;



- Existing Directory
- Browse
- Find location of folders from GitHub NUTRIOME

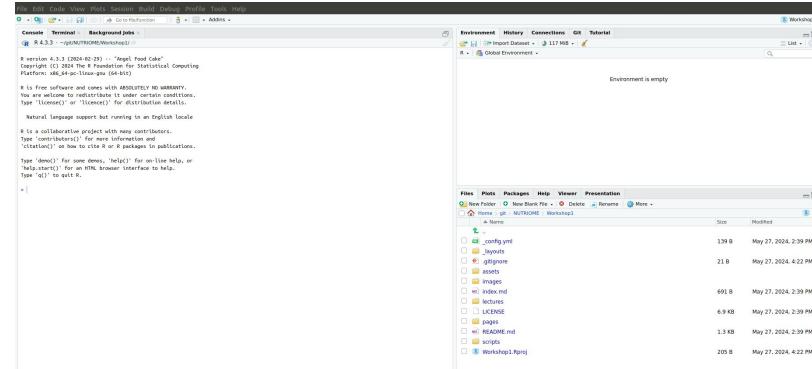
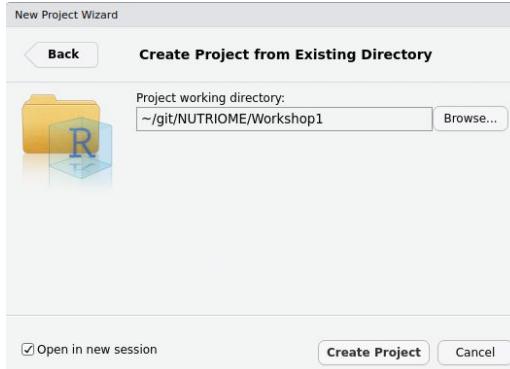
# Break time!?

# Hands-on session - Adding GitHub to Rstudio

Step 4: Select the top folder (Workshop1)

Step 5: Check if ‘Project working Directory’ is similar to printscreen below!

Step 6: Click → Create Project

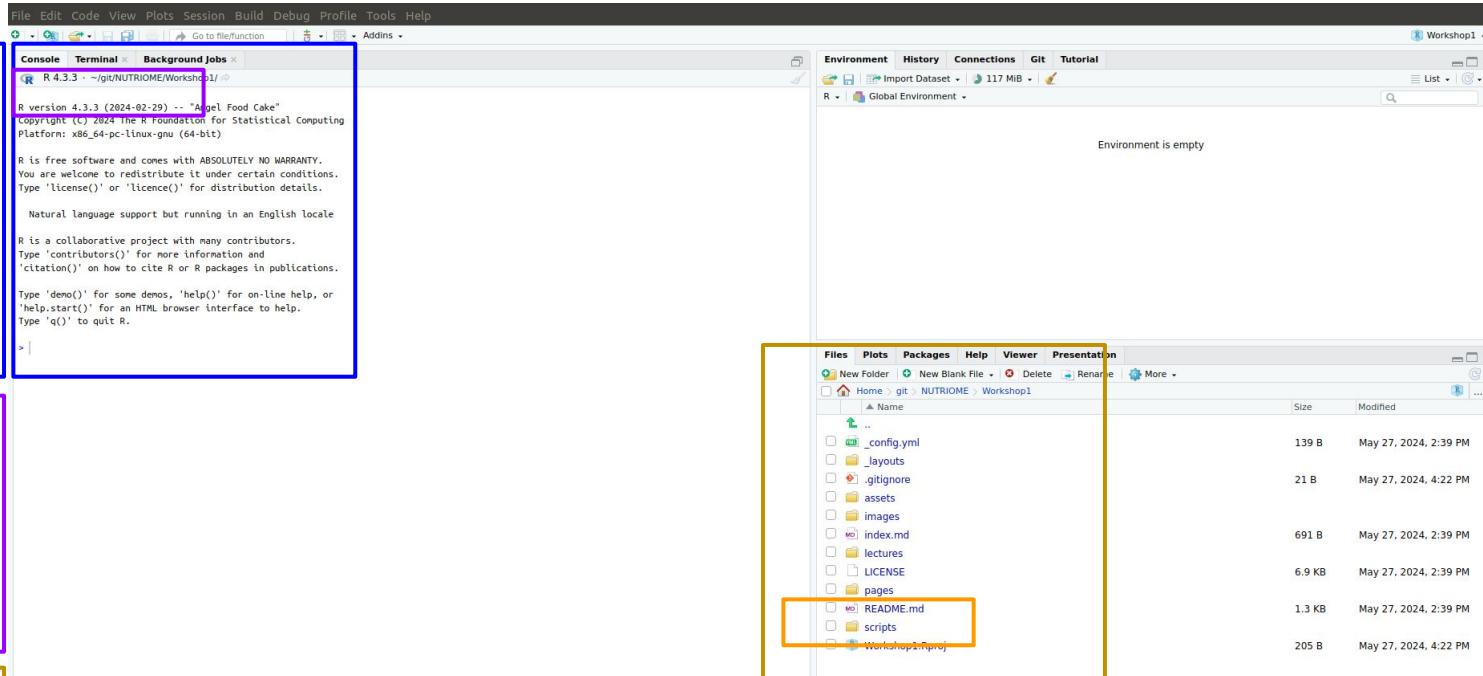


# Hands-on session - Adding GitHub to Rstudio

Section where code is executed (after the >) in console (also note 'Terminal' as second option)

Currently loaded version of R (check this for potential compatibility issues!)

Folder structure (according to GitHub Repository)



Folder with the Scripts  
(double click to unfold and show content)

# Hands-on session - Adding GitHub to Rstudio

The screenshot shows the RStudio interface. On the left is the 'Files' browser, which lists various project files and options like 'New Blank File', 'R Markdown...', and 'Quarto Doc...'. The main area is the 'Source Editor' showing an RMarkdown document named 'nb-demo.Rmd'. The code includes an R chunk that displays summary statistics for the 'iris' dataset:

```
9 Sepal.Length Sepal.Width Petal.Length Petal.Width Species
10 Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa :50
11 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor :50
12 Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica :50
13 Mean :5.843 Mean :3.057 Mean :4.358 Mean :1.500
14 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
15 Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
```

Below this, another R chunk uses ggplot2 to create a scatter plot of Petal.Length vs Petal.Width, colored by Species (setosa, versicolor, virginica) and sized by Petal.Width.

**Petal.Length**

**Petal.Width**

- 0.5
- 1.0
- 1.5
- 2.0
- 2.5

**Species**

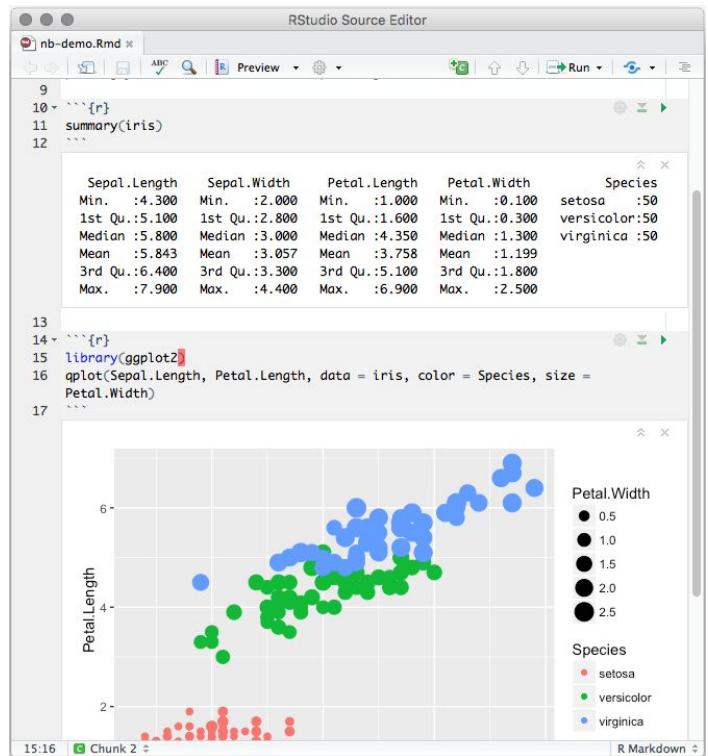
- setosa
- versicolor
- virginica

15:16 | Chunk 2 | R Markdown

# Hands-on session - Adding GitHub to Rstudio

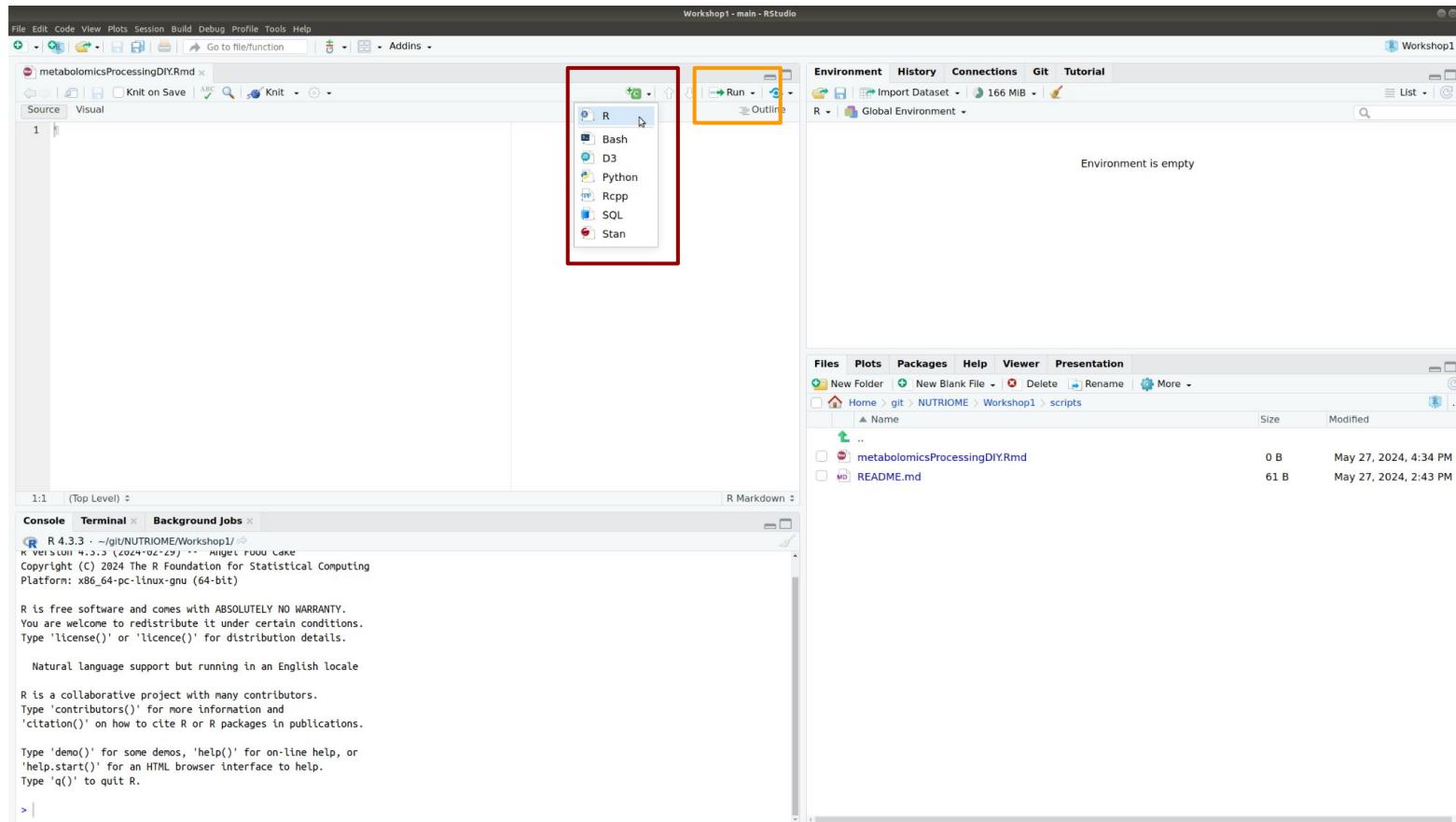
The screenshot shows the RStudio interface with GitHub integration. A green box highlights the top-left menu bar where 'GitHub' is listed under 'File'. Another green box highlights the 'Source' tab in the bottom-left panel, which displays the file 'metabolomicsProcessingDIY.Rmd'.

New 'Markdown' file create in File Explorer and in top left menu



<https://bookdown.org/yihui/rmarkdown/notebook.html>

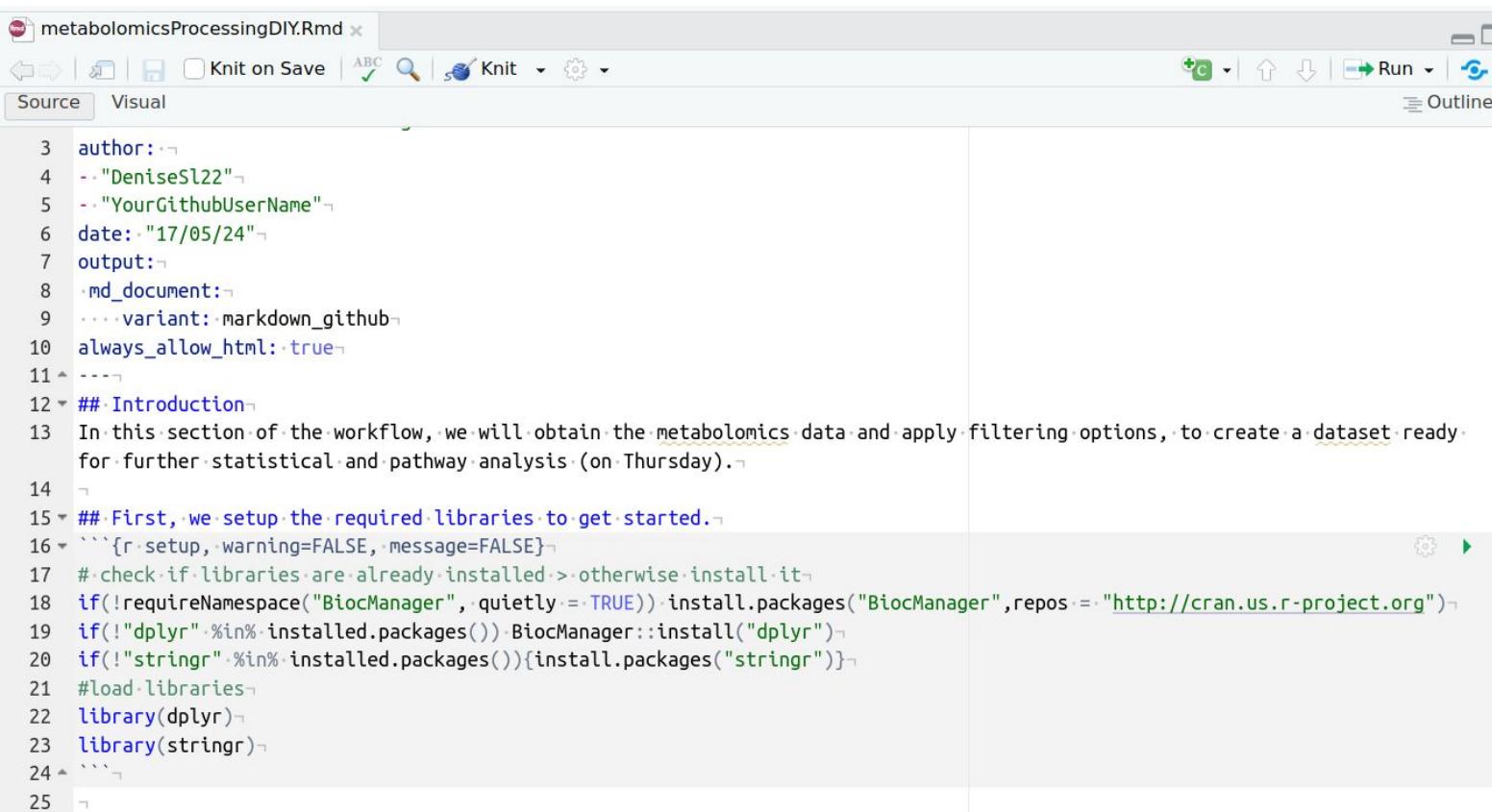
# Hands-on session - Adding GitHub to Rstudio



Button to add a new section of code (code 'chunk')

Button to run your code (or section of it, click small arrow)

# Hands-on session - First programming tasks



```
metabolomicsProcessingDIY.Rmd x
Source Visual
3 author:-
4   - "DeniseSL22"
5   - "YourGitHubUserName"
6 date: "17/05/24"
7 output:-
8   md_document:-
9     variant: markdown_github
10    always_allow_html: true
11  -----
12 ## Introduction
13 In this section of the workflow, we will obtain the metabolomics data and apply filtering options, to create a dataset ready for further statistical and pathway analysis (on Thursday).
14
15 ## First, we setup the required libraries to get started.
16 ``{r setup, warning=FALSE, message=FALSE}
17 # check if libraries are already installed > otherwise install it
18 if(!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager", repos = "http://cran.us.r-project.org")
19 if(!"dplyr" %in% installed.packages()) BiocManager::install("dplyr")
20 if(!"stringr" %in% installed.packages()) install.packages("stringr")
21 # load libraries
22 library(dplyr)
23 library(stringr)
24 ``
```

Make sure to have the following content in your Code chunk (select the correct file) and click 'Run' (or the green arrow in the chunk)!

Add your own GitHub User name to the metadata at the top iso 'YourGitHubUserName'

Save the file (locally) and check your GitHub Desktop

# Hands-on session - First programming tasks

Console Terminal × Background Jobs ×

R 4.3.3 · ~/git/NUTRIOME/Workshop1/ ↗

Natural language support but running in an English locale

R is a collaborative project with many contributors.

Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

```
> # check if libraries are already installed > otherwise install it
> if(!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager", repos = "http://cran.us.r-project.org")
> if(!"dplyr" %in% installed.packages()) BiocManager::install("dplyr")
> if(!"stringr" %in% installed.packages()){install.packages("stringr")}
> #load libraries
> library(dplyr)
> library(stringr)
> |
```

Check the console; if everything looks okay (see Figure on the left, text should be blue and end with >).

In case of errors (red text), ask for help!

Commit your username  
change to your own fork, and Push!

# Hands-on session - First programming tasks

1. Download the metabolomics data (from Surfdrive)
2. Add this file in the 'Data' folder (subfolder of scripts).
3. Check in GitHub Desktop if this file change is not visible (data should not be shared online!)
4. Add the following steps to the script (see Figure on the right). Answers are available on Github (in a separate script), but please try yourself first; you will learn the most from that!

## Read the data:

Download  
'readxl' package

Load the 'readxl'  
package

Add data  
location as  
variable

Read the first  
tab of the data  
file

# Hands-on session - First programming tasks

1. Download the metabolomics data (from Surfdrive)

2. Add this file in the 'Data' folder (subfolder of scripts).

3. Check in GitHub Desktop if this file change is not visible (data should not be shared online!)

4. Add the following steps to the script (see Figure on the right). Answers are available on Github (in a separate script), but please try yourself first; you will learn the most from that!

## Read the data:

Download 'readxl' package

```
install.packages("readxl")
```

Load the 'readxl' package

```
library("readxl")
```

Add data location as variable

Read the first tab of the data file

# Hands-on session - First programming tasks

1. Download the metabolomics data (from Surfdrive)

2. Add this file in the 'Data' folder (subfolder of scripts).

3. Check in GitHub Desktop if this file change is not visible (data should not be shared online!)

4. Add the following steps to the script (see Figure on the right). Answers are available on Github (in a separate script), but please try yourself first; you will learn the most from that!

## Read the data:

Download 'readxl' package

```
install.packages("readxl")
```

Load the 'readxl' package

```
library("readxl")
```

Add data location as variable

```
dataLocation <- paste0(getwd(),  
                      '/Data/NoMa_NMR_intervention_GEOcodes_Sent  
                      Susan.xlsx')
```

Read the first tab of the data file

```
metabolomicsData <- read_excel(dataLocation, 1)
```

# Hands-on session - First programming tasks

The screenshot shows an RStudio interface. On the left is a code editor with an R script titled 'metabolomicsProcessingDIY.Rmd'. The script contains several lines of R code, including library imports and data loading commands. On the right is the 'Global Environment' pane, which lists objects in memory. A pink box highlights the 'metabolomicsData' object, which is described as a '294 obs. of 227 variables' data frame. Below it, another object 'dataLocation' is listed with its value. Two callout boxes point to these entries: one blue box points to 'metabolomicsData' with the text 'Loaded 'dataframes' including number of rows (obs.) and columns (variables)'; another pink box points to 'dataLocation' with the text 'Loaded 'variables' including their values. Can also be a list of values and other data types!'. At the bottom of the screen is a file browser showing a directory structure under 'git / NUTRIOME / Workshop'.

```
metabolomicsProcessingDIY.Rmd | metabolomicsProcessingAnswers....  
Source Visual  
12 ## Introduction:  
13 In this section of the workflow, we will obtain the metabolomics data and apply filtering options, to create a dataset ready for further statistical and pathway analysis (on Thursday).  
14  
15 ## First, we setup the required libraries to get started.  
16 ````{r setup, warning=FALSE, message=FALSE}  
17 # check if libraries are already installed > otherwise install it  
18 if(!requireNamespace("BiocManager", quietly=TRUE)) install.packages("BiocManager", repos = "http://cran.us.r-project.org")  
19 if(!"dplyr" %in% installed.packages()) BiocManager::install("dplyr")  
20 if(!"stringr" %in% installed.packages()) install.packages("stringr")  
21 #load libraries  
22 library(dplyr)  
23 library(stringr)  
24 ````  
25  
26 ##Locate the data (in the file explorer) and load it in R  
27 ````{r}  
28 #Download 'readxl' package  
29 #Load the 'readxl' package  
30  
31 #Add data.location as variable  
32 #Read the first.tab of the data file  
33 ````  
34
```

Environment History Connections Git Tutorial  
Import Dataset 233 MB Global Environment  
Data metabolomicsData 294 obs. of 227 variables  
Values dataLocation "/home/deniseslenter/git/NUTRIOME/Workshop1/scripts/Data/NoMa\_NMR\_intervention...  
Data  
Files Plots Packages Help Viewer Presentation  
New Folder New Blank File Delete Rename More  
Home git > NUTRIOME > Workshop  
Name Size Modified  
.. 139 B May 27, 2024, 2:39 PM  
\_config.yml

Again, ask for help in case of errors/issues/question :)

Click on the dataframe name you used to load the metabolomics data

# Hands-on session - First programming tasks

Column names	Sample name	time	XXL-VLDL-P	XXL-VLDL-L	XXL-VLDL-PL	XXL-VLDL-C	XXL-VLDL-CE
1 1_A3	base	NA		NA	NA	NA	NA
2 1_A4	end	4.792000000000001E-11	1.021E-2	8.524000000000001E-4	2.229000000000001E-3	1.758E-3	
3 NA	delta	NA		NA	NA	NA	NA
4 1_A11	base	NA		NA	NA	NA	NA
5 1_A12	end	NA		NA	NA	NA	NA
6 NA	delta	NA		NA	NA	NA	NA
7 1_F3	base	7.563999999999997E-11	1.575E-2	1.489000000000001E-3	2.269000000000002E-3	1.65599999999999	
8 1_F4	end	1.358000000000001E-10	2.844E-2	3.07999999999998E-3	3.908E-3	2.33899999999999	
9 NA	delta	6.016000000000002E-11	1.269E-2	1.59099999999999E-3	1.638999999999901E-3	6.83000000000000	
10 1_B3	base	1.931000000000001E-10	4.132000000000003E-2	4.50799999999999E-3	7.727000000000004E-3	4.71099999999999	
11 1_B4	end	9.73999999999995E-11	2.076000000000001E-2	2.027000000000002E-3	3.705E-3	2.11999999999999	
12 NA	delta	-9.570000000000003E-11	-2.055999999999998E-2	-2.480999999999902E-3	-4.022000000000004E-3	-2.591E-3	
13 1_B1	base	1.031E-10	2.164E-2	2.532E-3	3.202E-3	2.20000000000000	
14 1_B2	end	NA	NA	NA	NA	NA	NA
15 NA	delta	NA	NA	NA	NA	NA	NA
16 1_B5	base	4.527000000000002E-11	8.95499999999994E-3	5.03199999999998E-4	2.423000000000001E-4	1.752E-4	
17 1_B6	end	NA	NA	NA	NA	NA	NA

# Hands-on session - Second tasks

1. Use the search box to find sample '2\_A5'

2. Convert the textual 'NA' labels with the real NA (not available)

3. Check how many NAs are part of sample '2\_A5'.

4. Visualize the NA information in a heatmap

5. Check GitHub Desktop and update your fork

Credits to:  
<https://thomasadventure.blog/posts/r-count-na/>

## Check missing data:

Convert text  
'NA' to real NA

Fill in the correct sample name

Install and load the required package

Use the  
'heatmaply'  
package

# Hands-on session - Second tasks

1. Use the search box to find sample '2\_A5'

2. Convert the textual 'NA' labels with the real NA (not available)

3. Check how many NAs are part of sample '2\_A5'.

4. Visualize the NA information in a heatmap

5. Check GitHub Desktop and update your fork

Credits to:  
<https://thomasadventure.blog/posts/r-count-na/>

## Check missing data:

Convert text 'NA' to real NA

```
metabolomicsData[metabolomicsData == "NA"] <- NA
```

Fill in the correct sample name

```
sum(is.na(metabolomicsData$`Sample name` ==  
"2_A5")) ##Answer is 98
```

Install and load the required package

Use the 'heatmaply' package

# Hands-on session - Second tasks

1. Use the search box to find sample '2\_A5'

2. Convert the textual 'NA' labels with the real NA (not available)

3. Check how many NAs are part of sample '2\_A5'.

4. Visualize the NA information in a heatmap

5. Check GitHub Desktop and update your fork

Credits to:  
<https://thomasadventure.blog/posts/r-count-na/>

## Check missing data:

Convert text 'NA' to real NA

```
metabolomicsData[metabolomicsData == "NA"] <- NA
```

Fill in the correct sample name

```
sum(is.na(metabolomicsData$`Sample name` == "2_A5")) ##Answer is 98
```

Install and load the required package

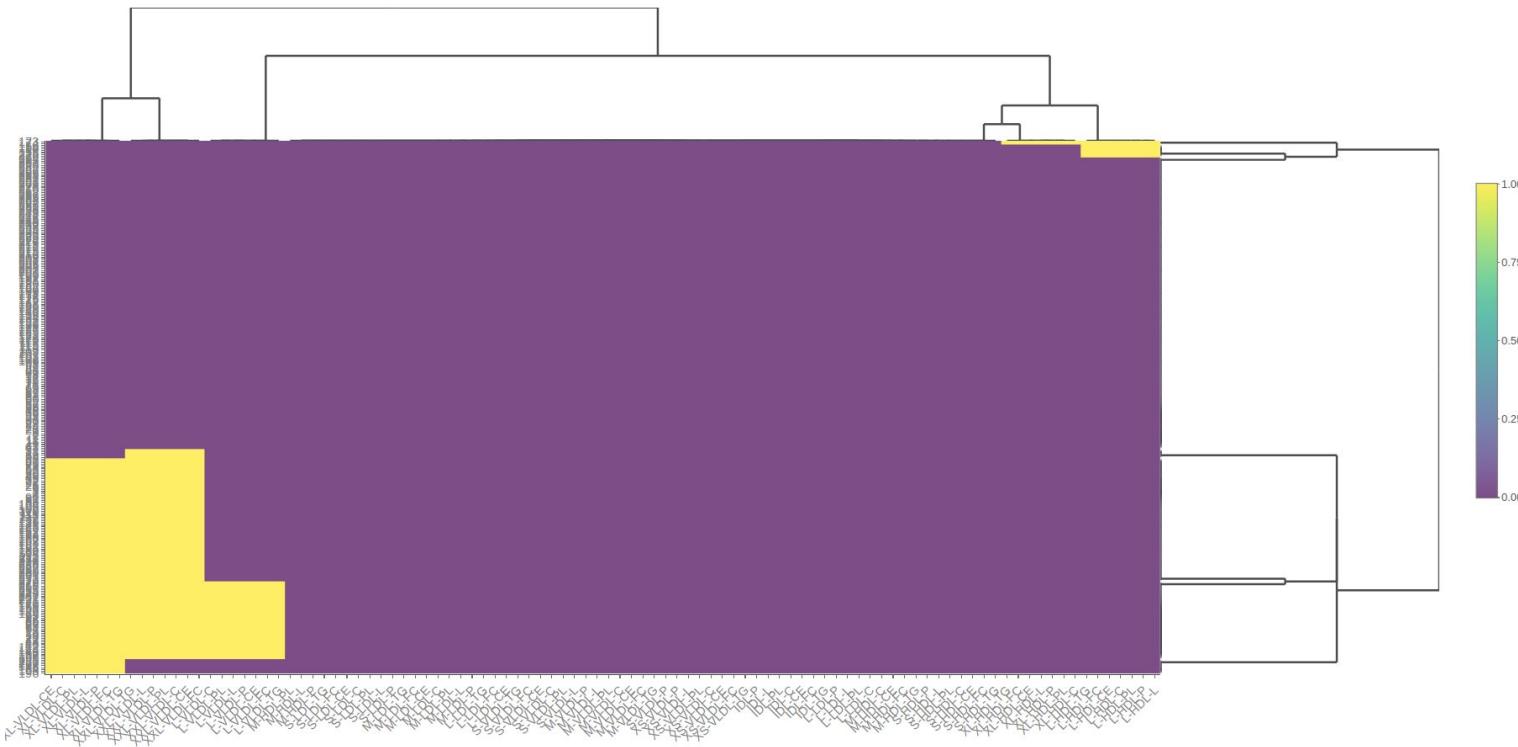
```
install.packages("heatmaply")  
library("heatmaply")
```

Use the 'heatmaply' package

```
heatmaply::heatmaply(is.na10(metabolomicsData[,3:100], grid_gap = 1, colors = heat.colors(200), showticklabels = c(T, F), margins = c(80, 10)))
```

# Hands-on session - Second tasks

Do you see  
anything  
important  
pop-up  
here?



Why did we select: `metabolomicsData[,3:100]` ?

# Hands-on session - Third tasks

1. Make sure all data is captured as numbers (numeric!)  
Otherwise the data is not read correctly by the heatmap functions
2. Remove rows with too many NAs  
Otherwise the clustering distance cannot be calculated
3. Visualize all information in a heatmap  
To be added by yourself

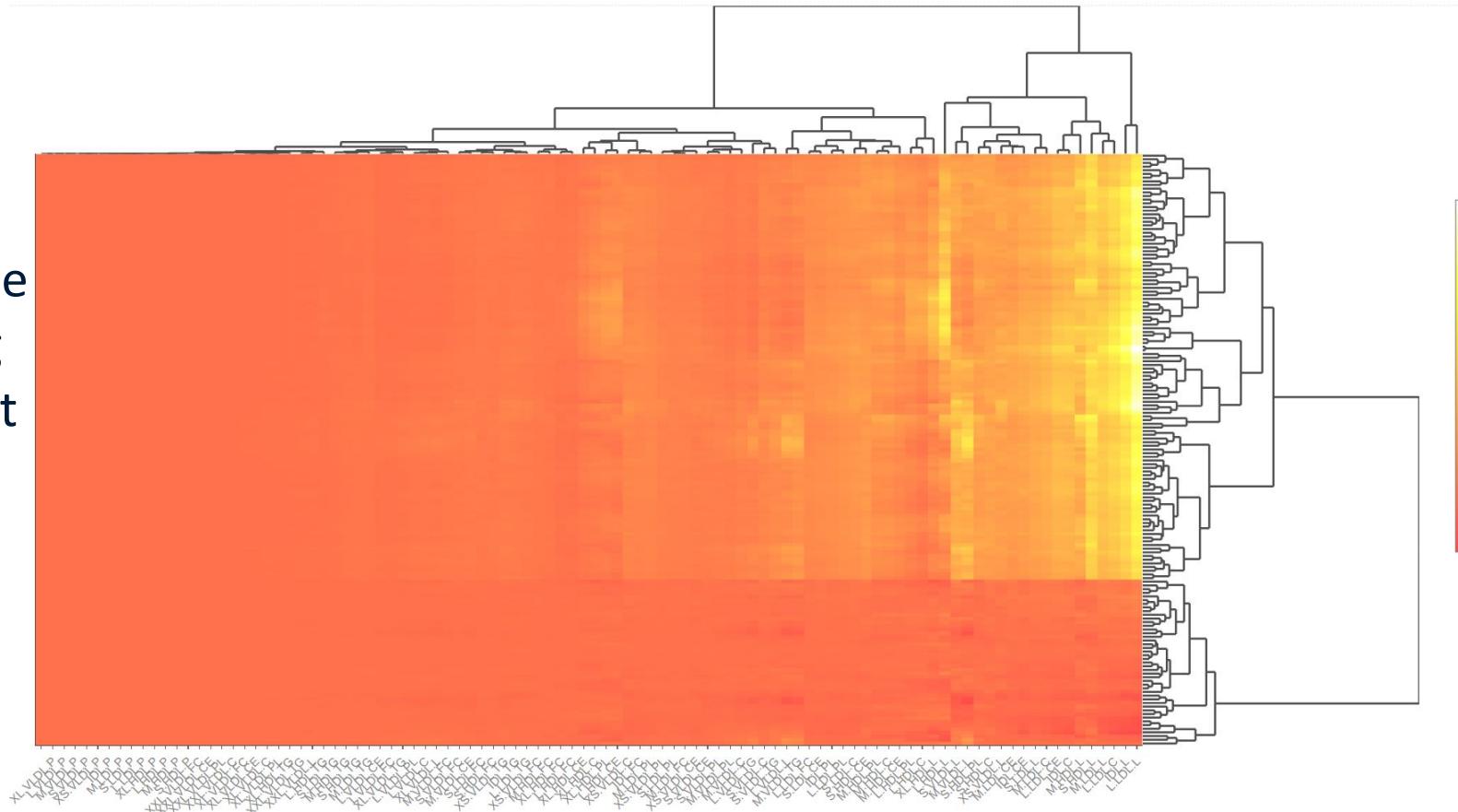
# Hands-on session - Third tasks

1. Make sure all data is captured as numbers (numeric!)  
Otherwise the data is not read correctly by the heatmap functions
2. Remove rows with too many NAs  
Otherwise the clustering distance cannot be calculated
3. Visualize all information in a heatmap  
To be added by yourself

```
# Visualize all information in a heatmap
heatmaply::heatmaply(metabolomicsData[,3:100], grid_gap = 0, colors =
heat.colors(200), showticklabels = c(T, F), margins = c(80, 10))
```

# Hands-on session - Third tasks

Do you see  
anything  
important  
pop-up  
here?

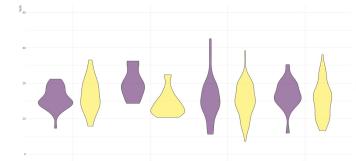


# Hands-on session - Fourth tasks: data distribution and transformation

1. Connect the data to the treatment groups
  - To compare effect of treatment
2. Apply different data transformations and see the effect on the data distribution
  - Four examples given: log2, log10, cube root, square root
3. Calculate some statistics on effect of data transformations
  - Shapiro–Wilk test for normality
4. Visualize transformed data in heatmap
  - To be added by yourself

# Hands-on session - Fifth tasks: check for outliers (can also be done before transformation)

1. Install required packages (can take some time)  
Check if example visualization works!
2. Visualise our remaining data with violin plots (for first 20 variables)
3. Check if any outliers are present; if yes investigate where this could come from.
4. Repeat for other sets of variables



# Hands-on session - Last tasks: check the ratios and compare the groups

1. Up to now we looked at the lipid classes intensity data; what about the other columns (ratios etc.)?

Use for example the heatmap function we used before to visualise

2. Compare the two group

Control versus experimental (check the Diet column)

# Join our quiz!

Join at [menti.com](https://menti.com) | use code 75 51 85 5

Mentimeter

## Instructions

Go to

**www.menti.com**

Enter the code

**75 51 85 5**



Or use QR code