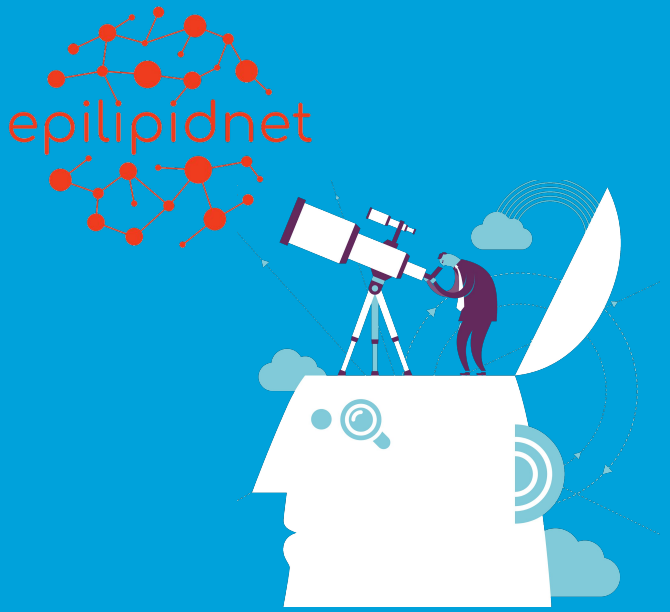
Metabolomics data analysis

Denise Slenter

ORCID: 0000-0001-8449-1318

Tuesday May 28, 2024



Maastricht University

What will I cover?

+ Existing tools for metabolomics data analysis, pros and cons

+ Steps in processing annotated metabolomics data

+ Working with Biocrates data

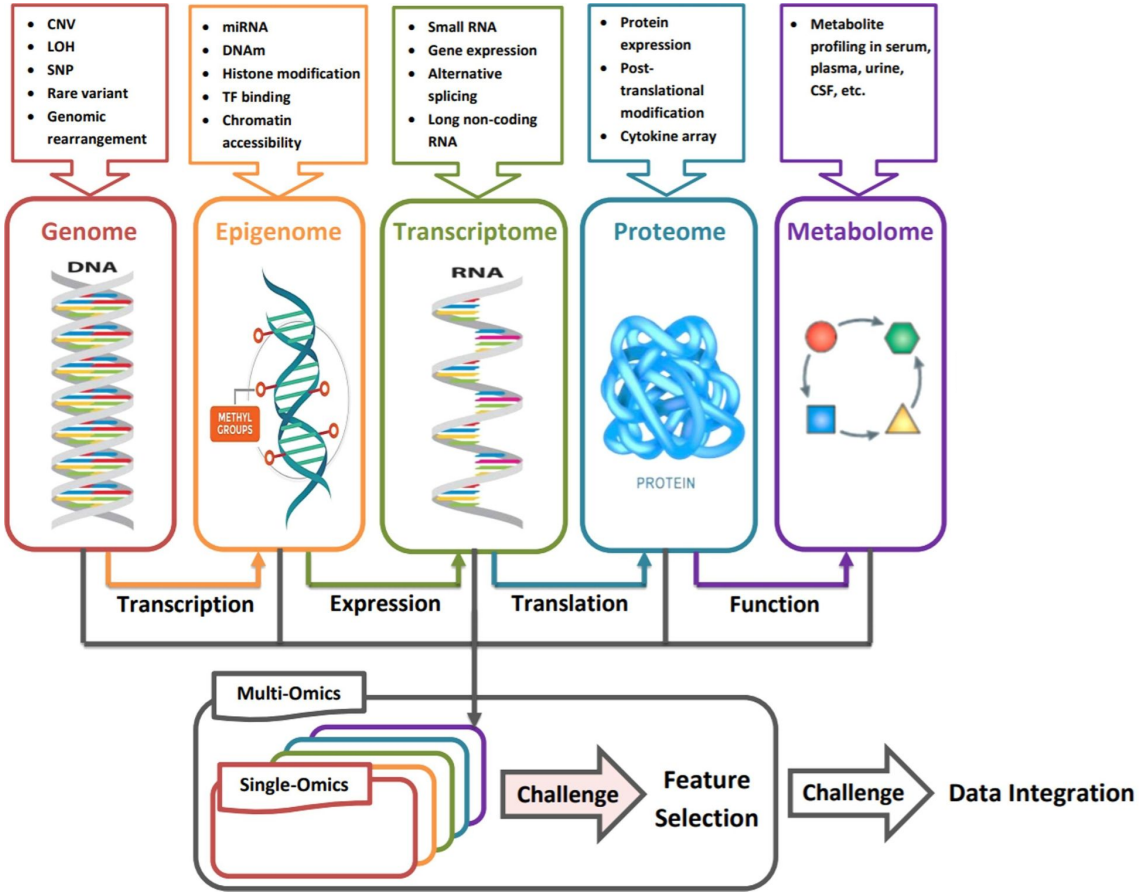
What can I NOT cover?

- Annotating raw data

- Overview of the metabolomics research field

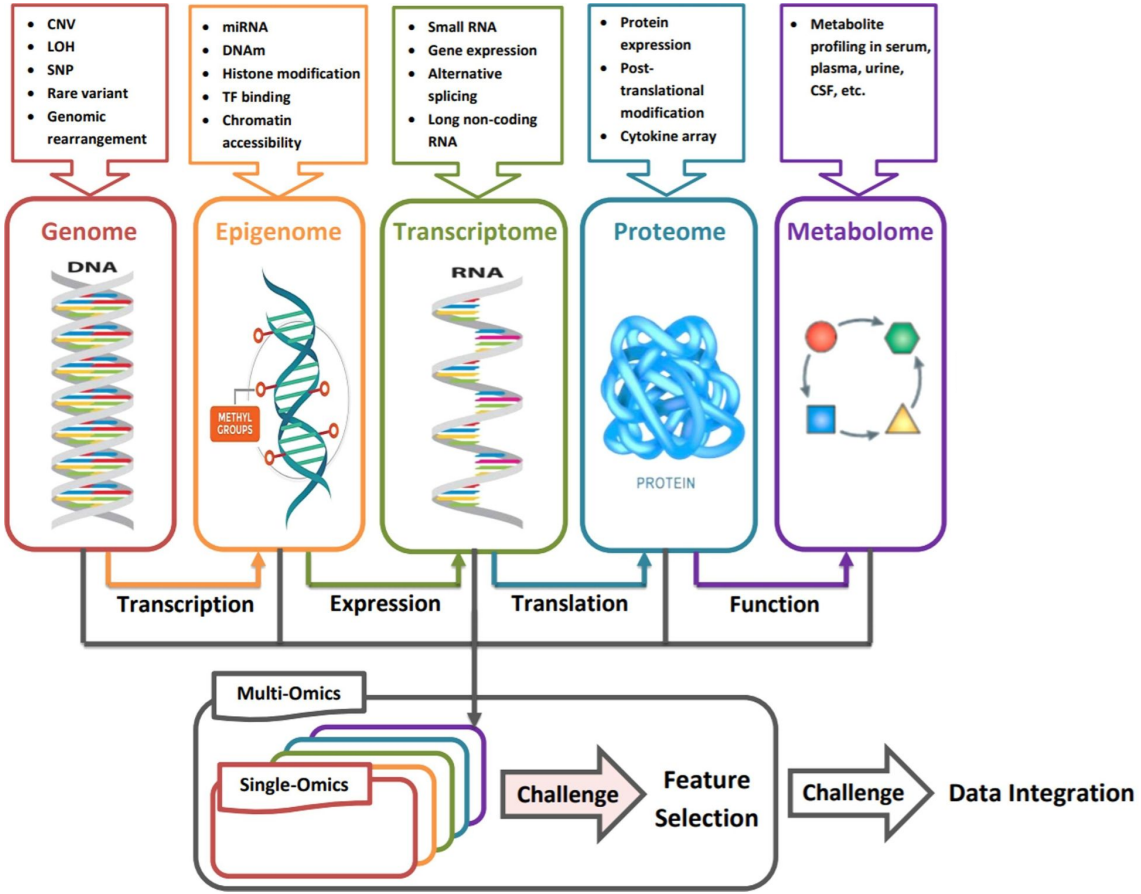
- ‘Plug and Play’ data analysis scripts - you will need to write some code yourself :)

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

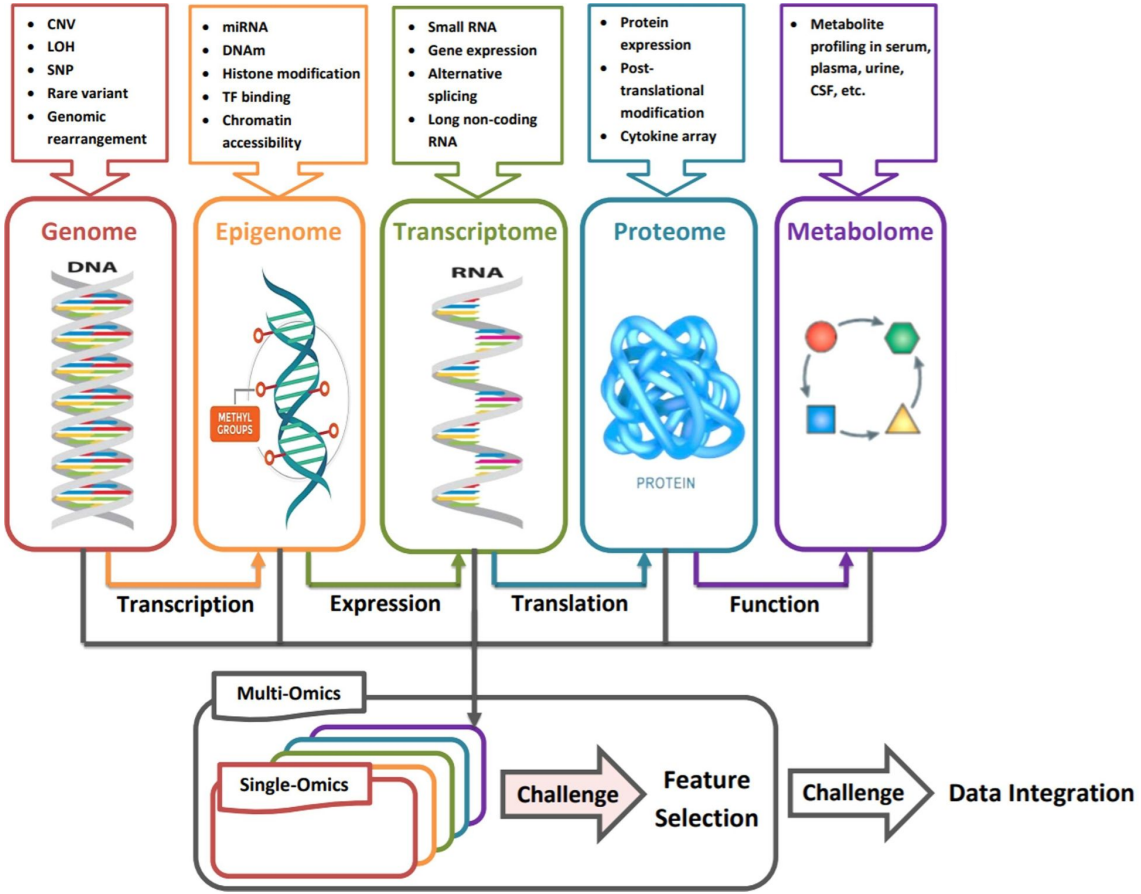
Different data types and analysis techniques



What data is missing in this overview?

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

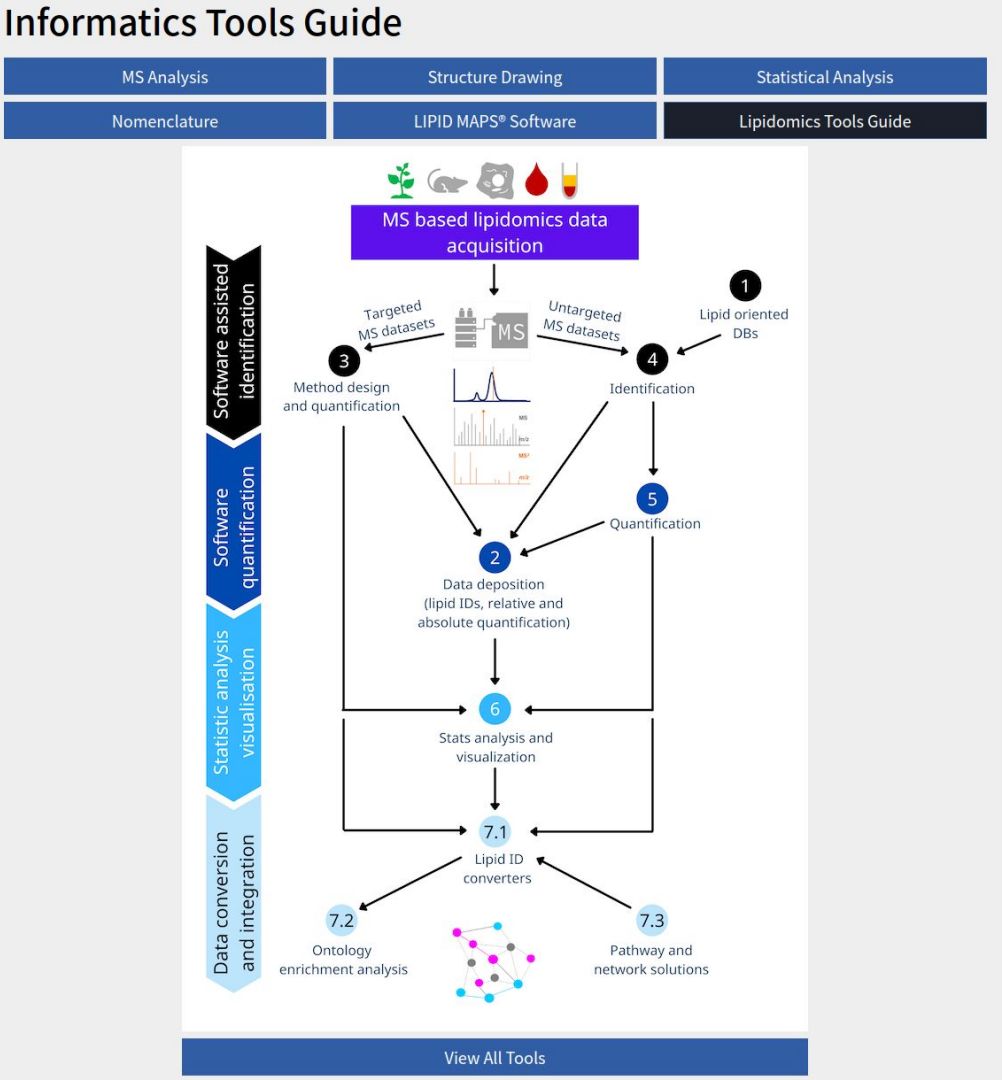
Different data types and analysis techniques



Missing:

- Phenotype
- Imaging data
- Fluxomics
- …

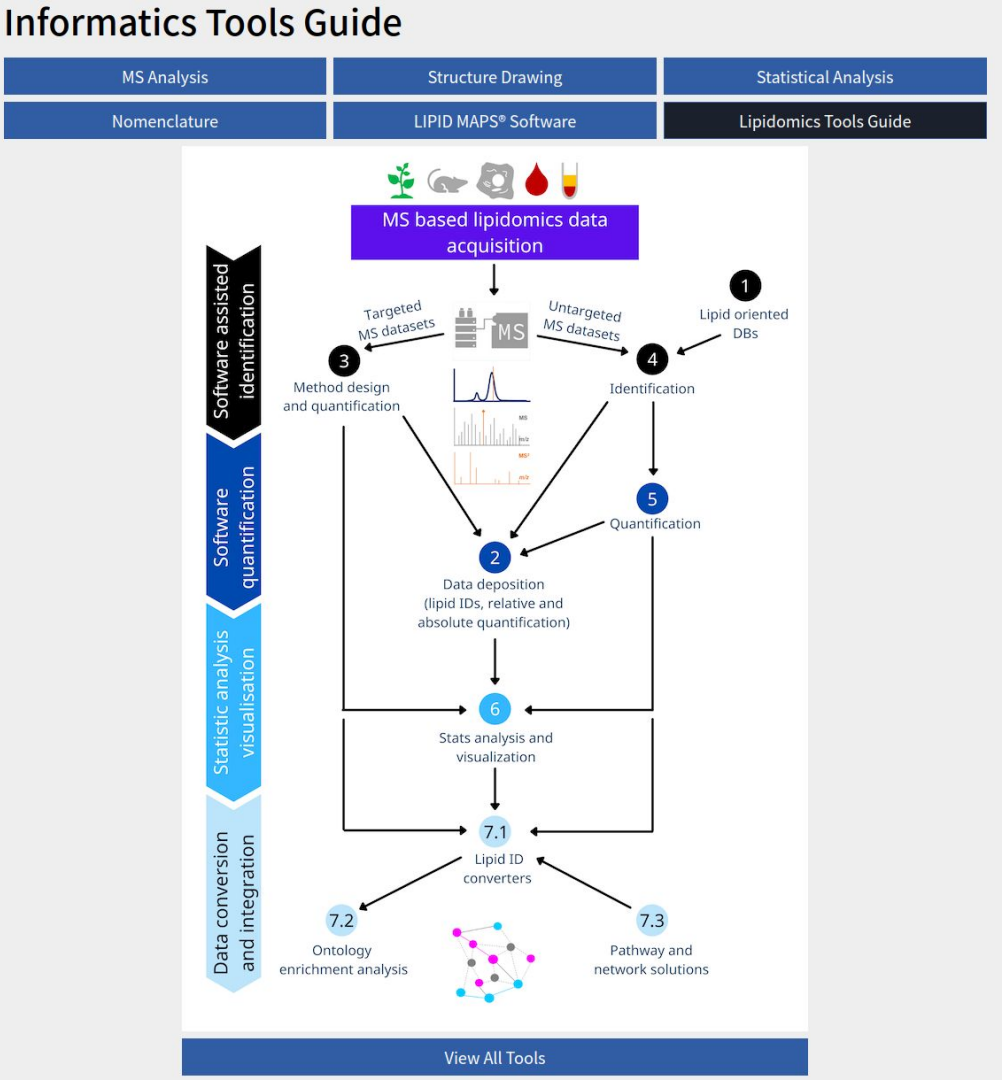
Adapted from: Momeni, Zahra, et al. "A survey on single and multi omics data mining methods in cancer data classification." Journal of Biomedical Informatics 107 (2020): 103466. DOI: 10.1016/j.jbi.2020.103466



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



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

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[a]	**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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022
Chemical properties[c]	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

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[a]	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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022
Chemical properties[c]	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

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[a]	**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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties[c]	256 592	217 776	13 066	44 981	

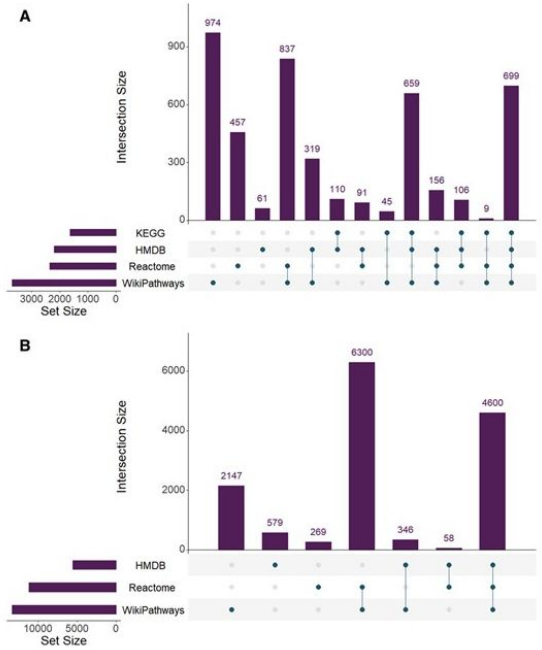
a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

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

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

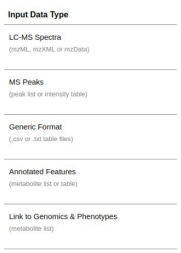
A					
	Total[a]	**HMDB v5.0**	**KEGG (from HMDB 5.0)**	**Reactome v81**	**WikiPathways v20220710**
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway 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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties[c]	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

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING
hmdb The Human Metabolome Database; MetaboAnalyst 6.0; Path Bank; DRUGBANK Small Molecule Pathway Database	 18 modules, Free to use 	Metabolite ID conversion https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml

Based on pathway databases in RaMP

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

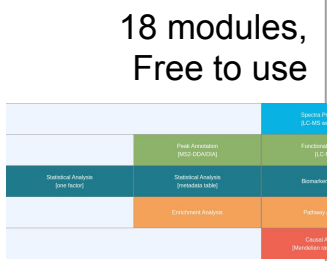
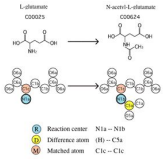
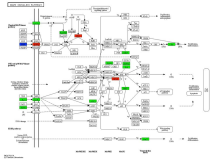
A					
	Total[a]	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775 (+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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties[c]	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

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING
hmdb The Human Metabolome Database; MetaboAnalyst 6.0; Path Bank; DRUGBANK; Small Molecule Pathway Database	 18 modules, Free to use	Metabolite ID conversion https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml
KEGG Kyoto Encyclopedia of Genes and Genomes 	 Licence required; Academic usage Of website for free	KEGG Mapper https://www.genome.jp/kegg/mapper/

Based on pathway databases in RaMP

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

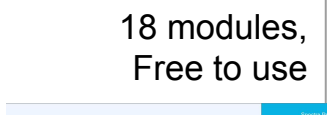
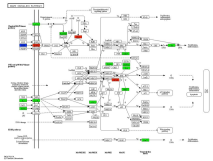
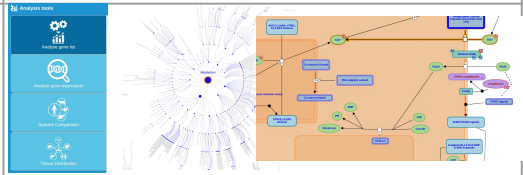
A					
	Total[a]	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775 (+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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties[c]	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

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING
hmdb, MetaboAnalyst 6.0, PathBank, DRUGBANK, Small Molecule Pathway Database	 18 modules, Free to use	Metabolite ID conversion https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml
KEGG Kyoto Encyclopedia of Genes and Genomes	 Licence required; Academic usage Of website for free	KEGG Mapper https://www.genome.jp/kegg/mapper/
reactome, ChEMBL, Ensembl		Integrated in analysis tool, downloadable files available https://reactome.org/download-data

Based on pathway databases in RaMP

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

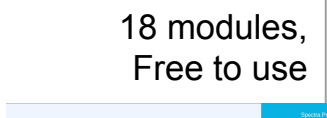
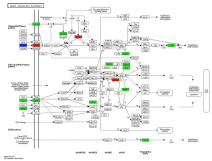
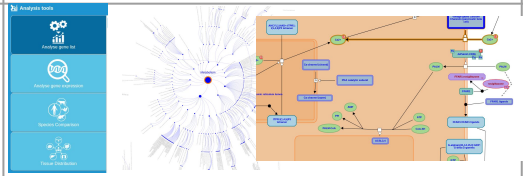
A					
	Total[a]	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775 (+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[b]	HMDB v5.0	ChEBI release 212	LIPID MAPS release July 13, 2022	
Chemical properties[c]	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).

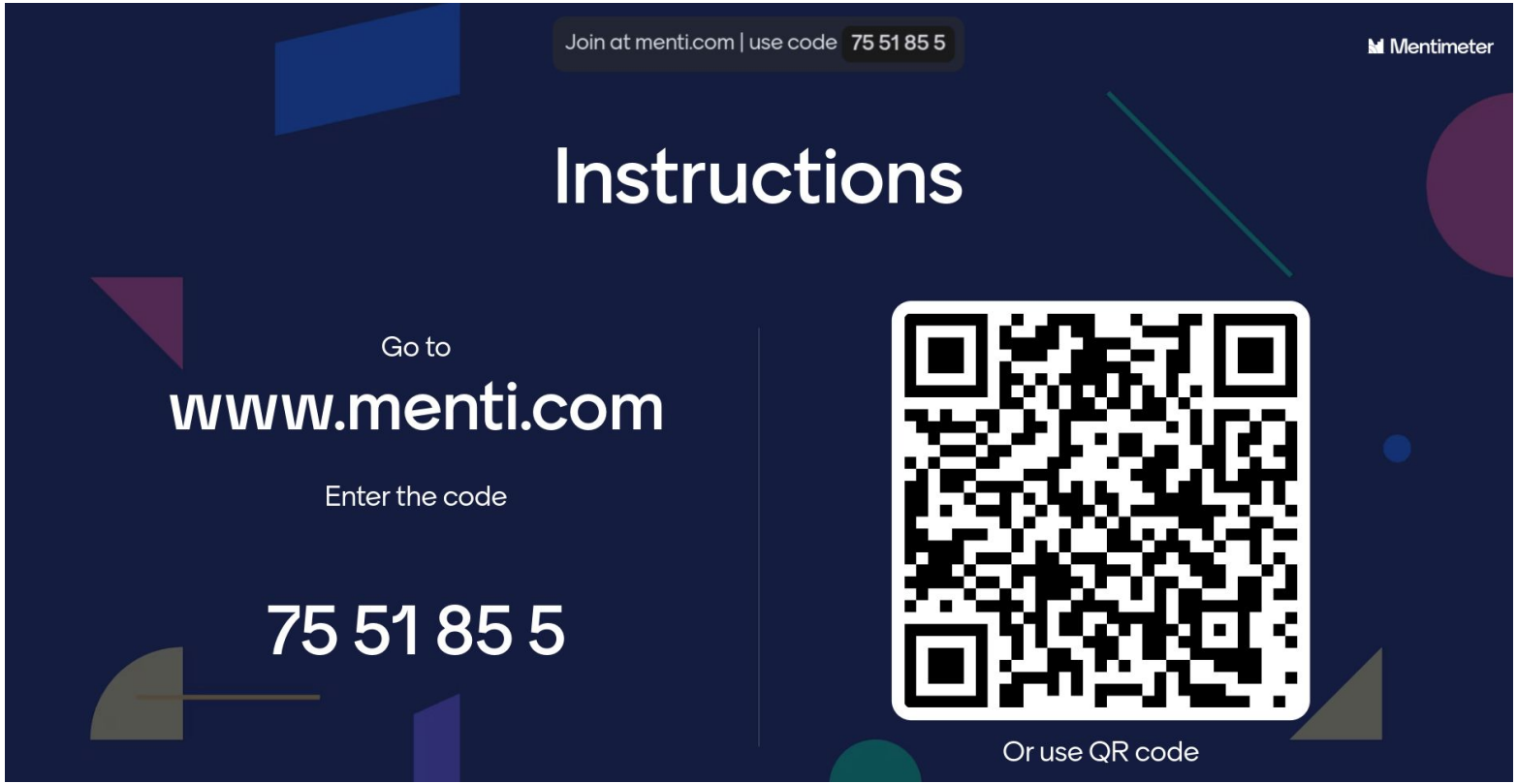
b Distinct InChIKeys.

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

DATABASE AND TOOLS	ANALYSIS OPTIONS	ID MAPPING
hmdb, MetaboAnalyst 6.0, PathBank, DRUGBANK, Small Molecule Pathway Database	 18 modules, Free to use	Metabolite ID conversion https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml
KEGG, Kyoto Encyclopedia of Genes and Genomes	 Licence required; Academic usage Of website for free	KEGG Mapper https://www.genome.jp/kegg/mapper/
reactome, ChEMBL, Ensembl		Integrated in analysis tool, downloadable files available https://reactome.org/download-data
WikiPathways Pathways for the People, CTL, RDF, NDEx	PathVisio, Cytoscape, neo4j	https://www.bridgedb.org/ BridgeDb

Join our quiz!



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	
1		Nuclei_Control	Nuclei_Control	Nuclei_Control	Nuclei_KLA	Nuclei_KLA	Nucle
2		#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	
6	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	
8	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.
10	GPA(34:0)	1.35	2.64	2.37	2.66	1.76	

LIPID MAPS Statistical Analysis Tools

Statistical Analysis Tools for User-Uploaded Data

LION/web: LION enrichment analysis

Maastricht University

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
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: 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).

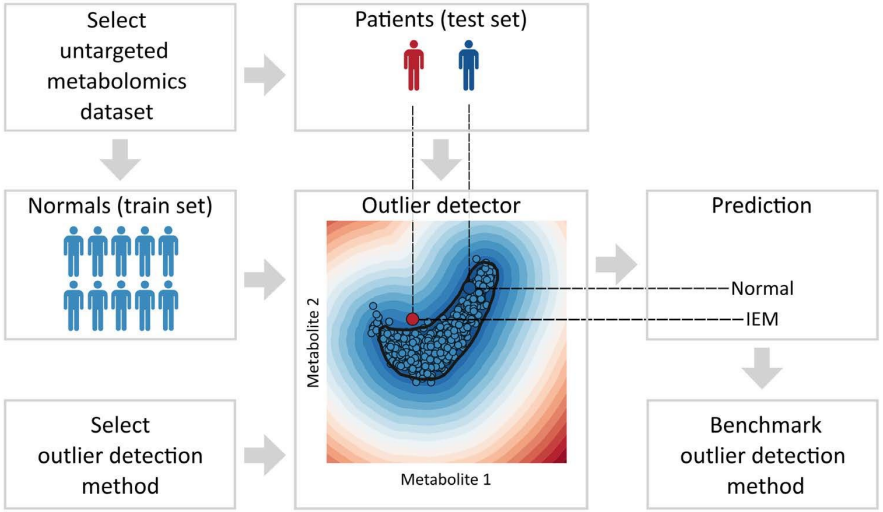


Figure obtained from:
Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." Metabolites 13.1 (2023): 97.
https://doi.org/10.3390/metabo13010097

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?

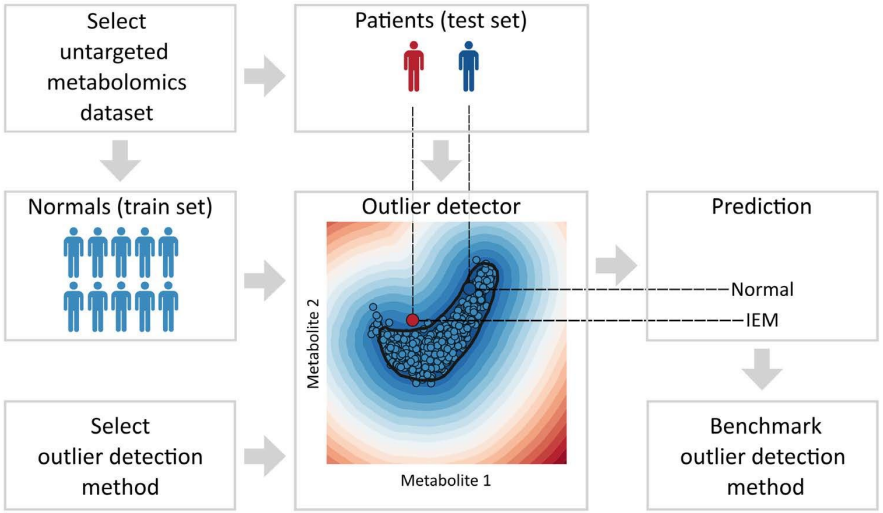
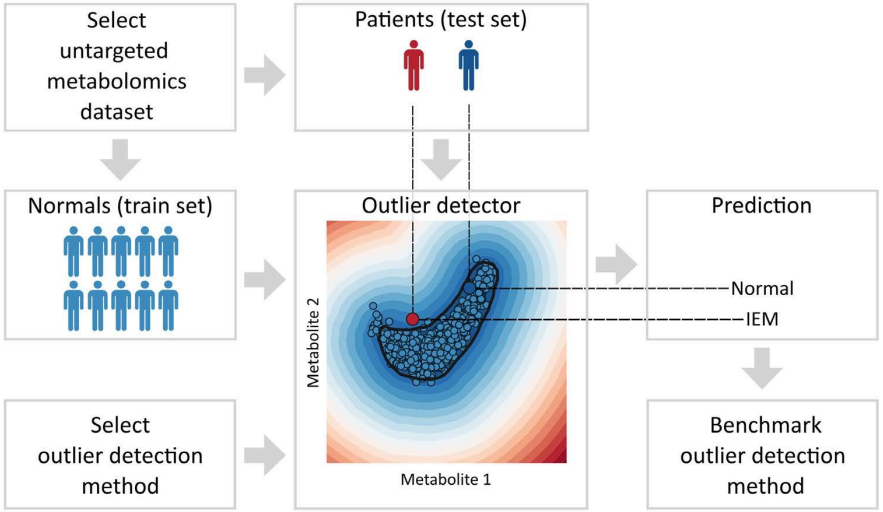


Figure obtained from:
Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." Metabolites 13.1 (2023): 97.
https://doi.org/10.3390/metabo13010097

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.



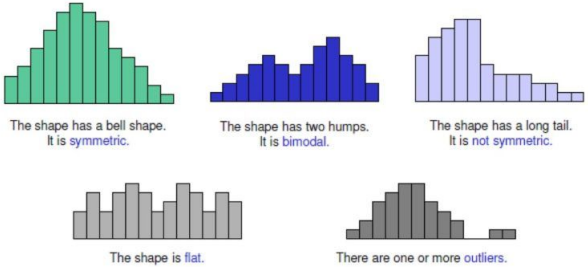
Which one to pick?

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

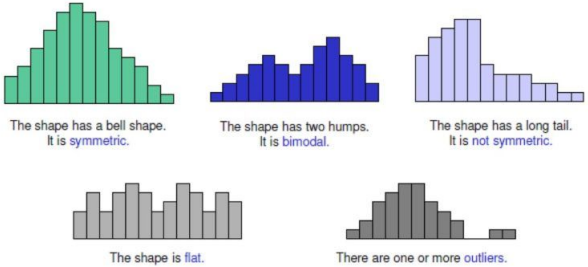
Data processing steps: distribution



Many data distributions exist:

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

Data processing steps: distribution



Many data distributions exist:

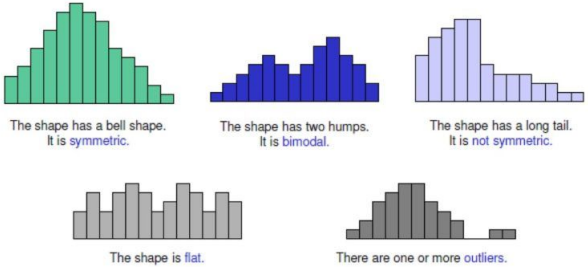
- Normal Distribution (Gaussian Distribution)
- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Log-Normal Distribution
- …

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
- …

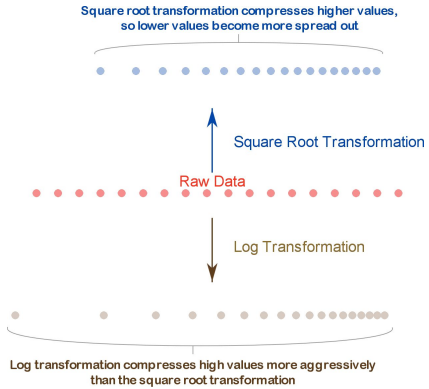
But, which one is it?

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?

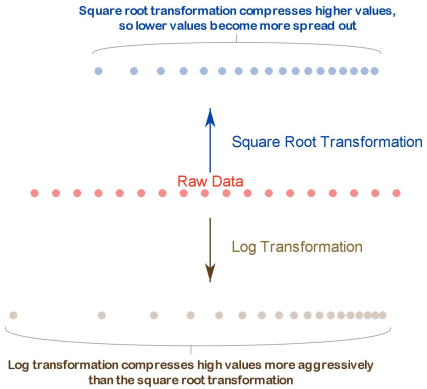
Data processing steps: transformation



Again, many data transformation techniques exist:

- Logarithmic (log2, log10)
- 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
-

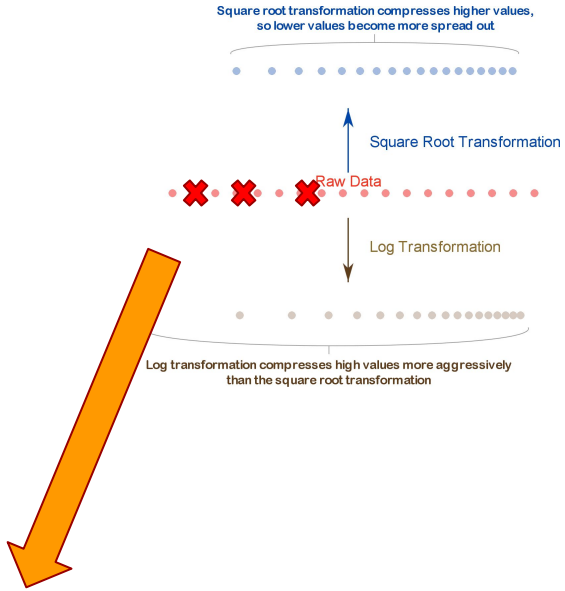
General rule of thumb:

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

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

- 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?

Data processing steps: missing data

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

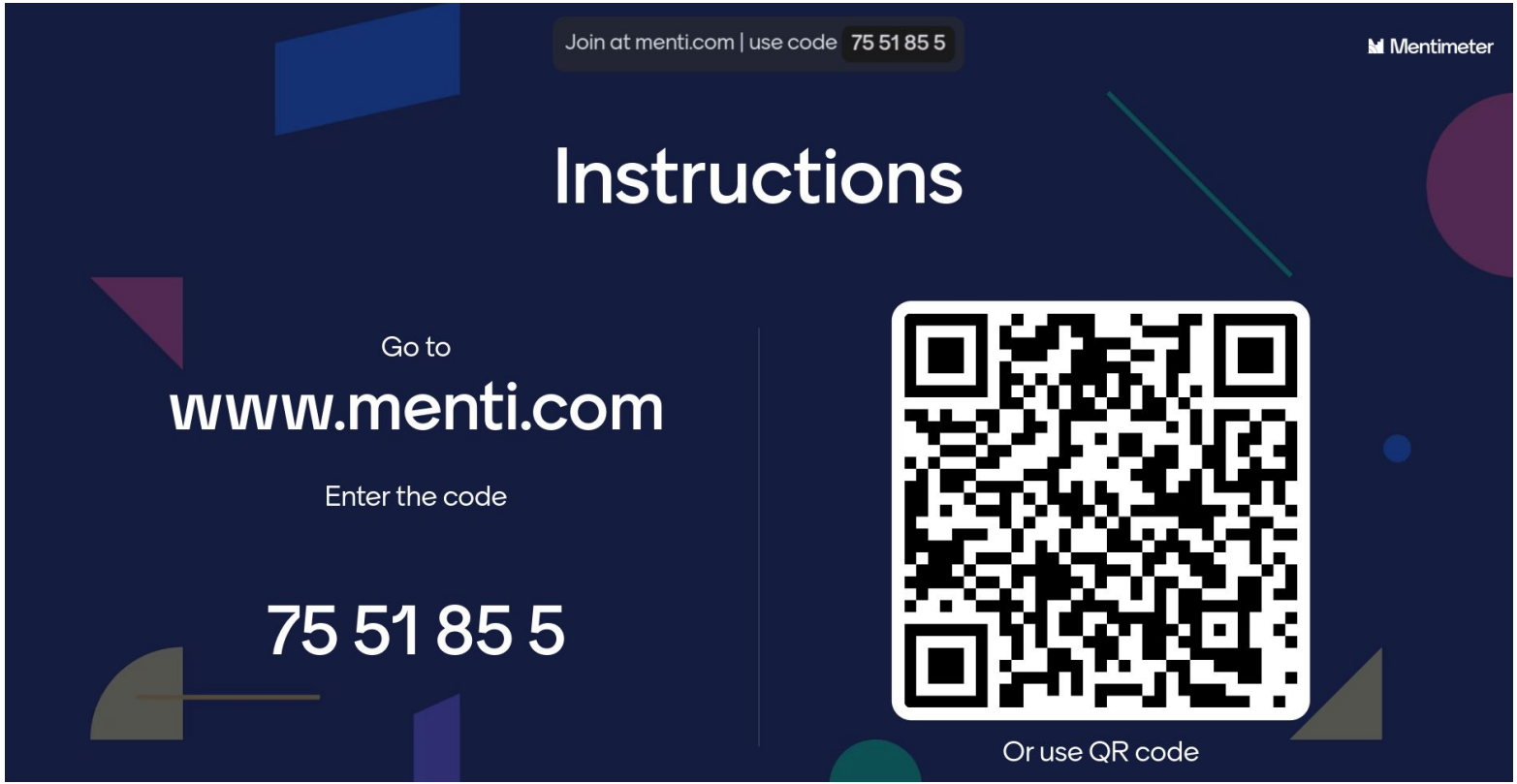
Highlighted all the missing values in the dataset

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

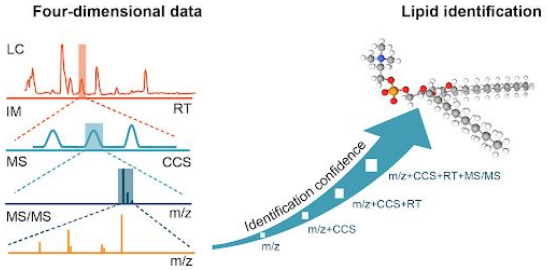
Highlighted only the standard missing values in the dataset

- 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)

Join our quiz!



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



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

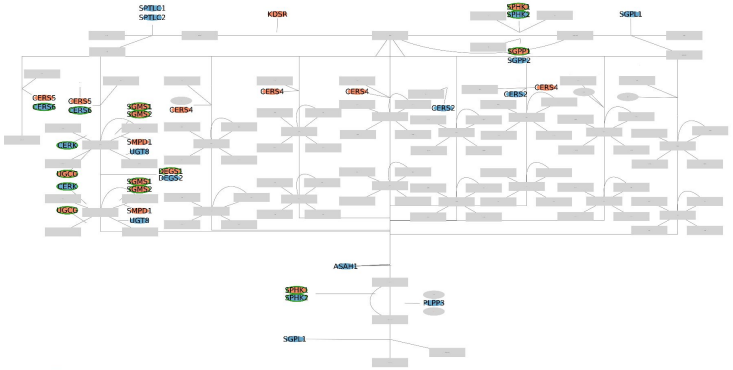
- Identify if there are duplicates Names/ IDs
- Review how these duplicates came part of the dataset
- Remove redundant rows
- Merge duplicates for multiple instances of same entity (average of values)
- Retain unique duplicates from entry errors or inconsistencies (label them)

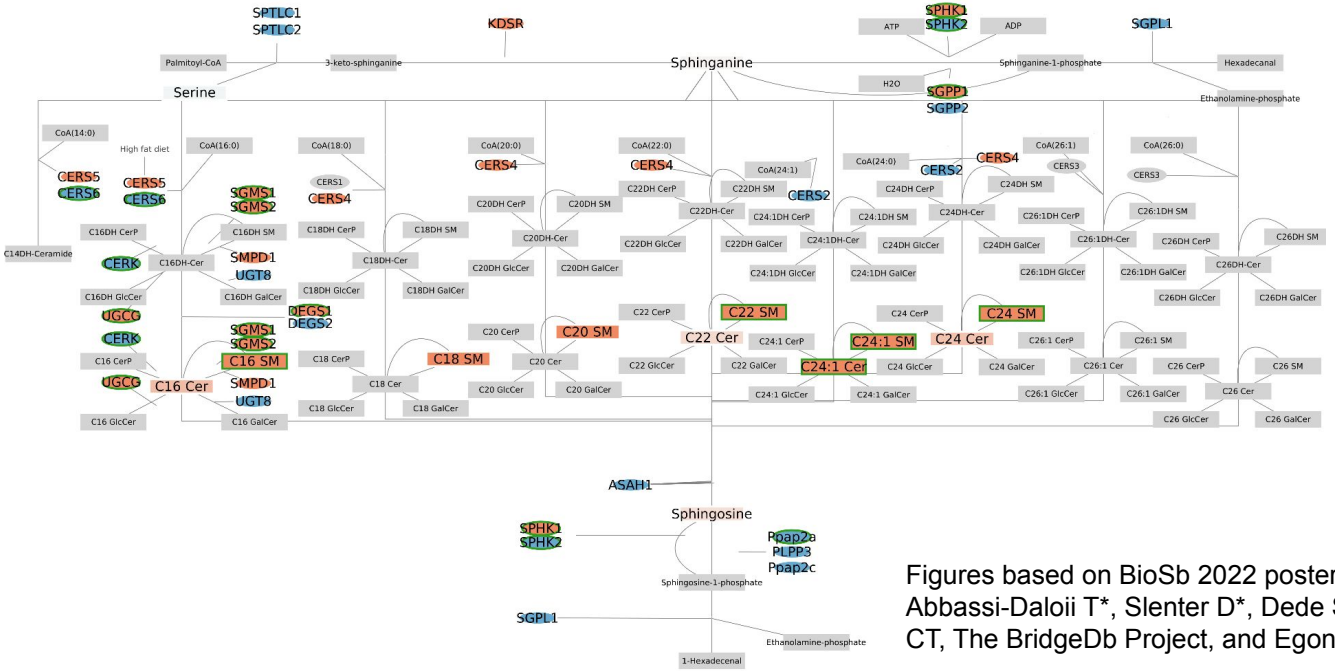
- Don’t forget about secondary IDs…

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

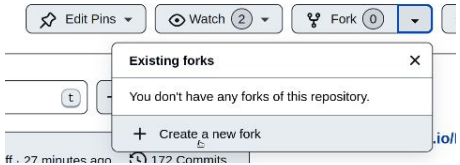
11:15-12:00 (so only 45 minutes!)

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides

Second, find the repository and make a FORK

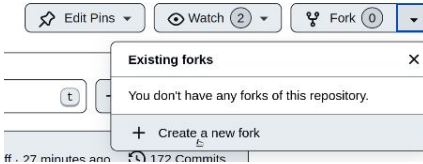
https://github.com/NUTRIOME/Workshop1



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 * DeniseSl22 / **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.

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

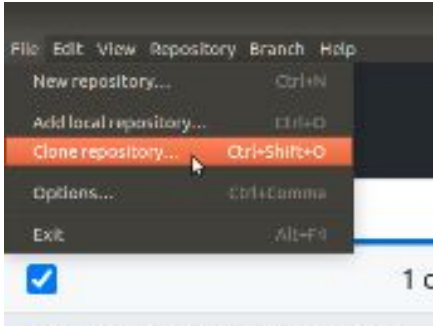
- Select File/Clone repository

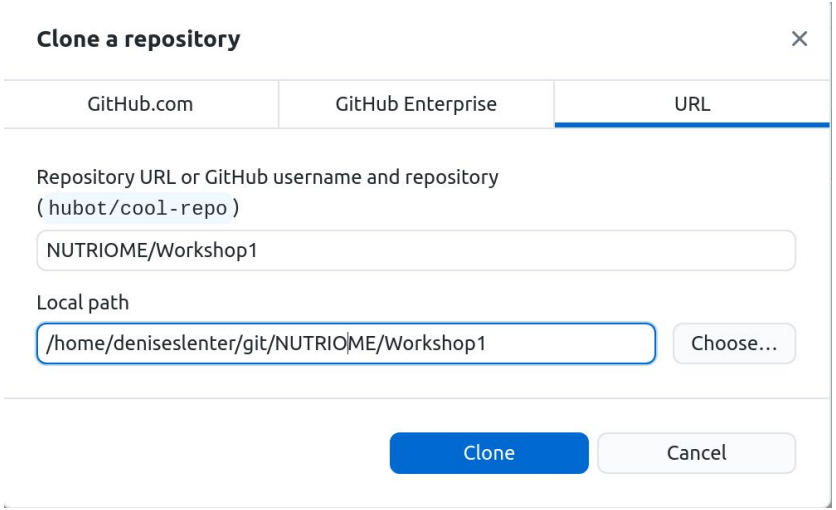
- In the pop-up menu, select URL (third option from the top), and add:

  NUTRIOME/Workshop1

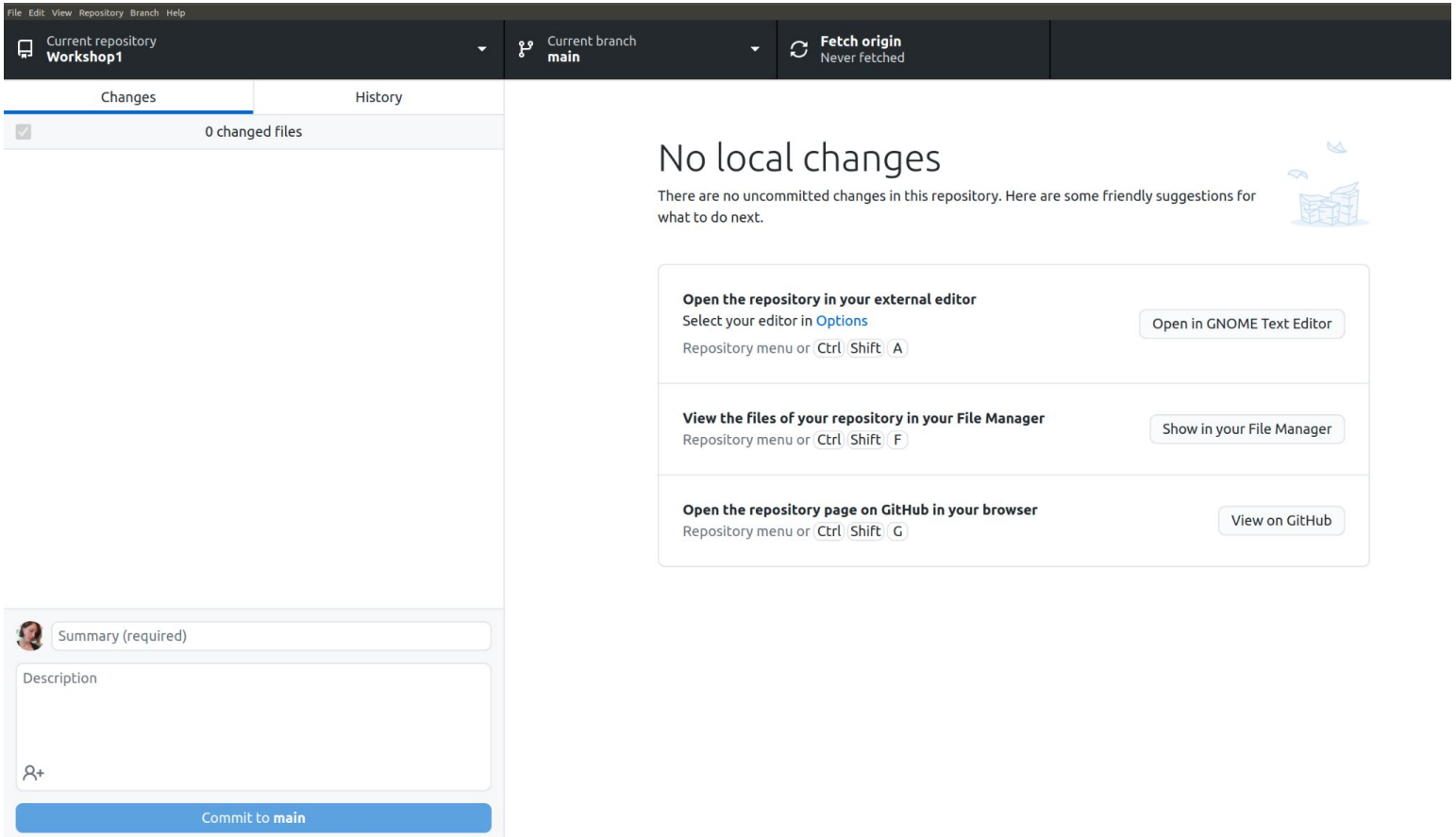
  or

  *username/Workshop1



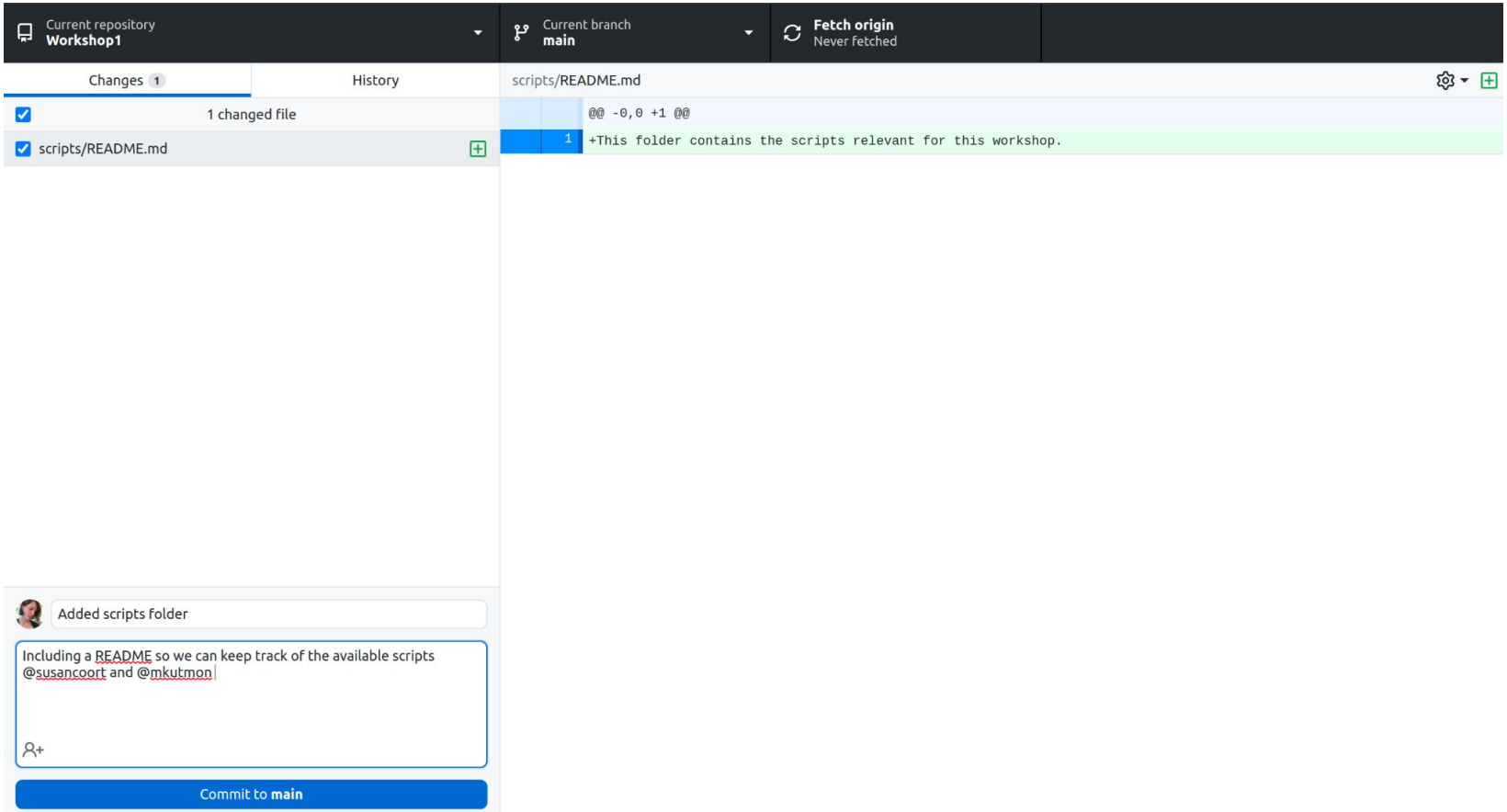


Hands-on session - GitHub Desktop



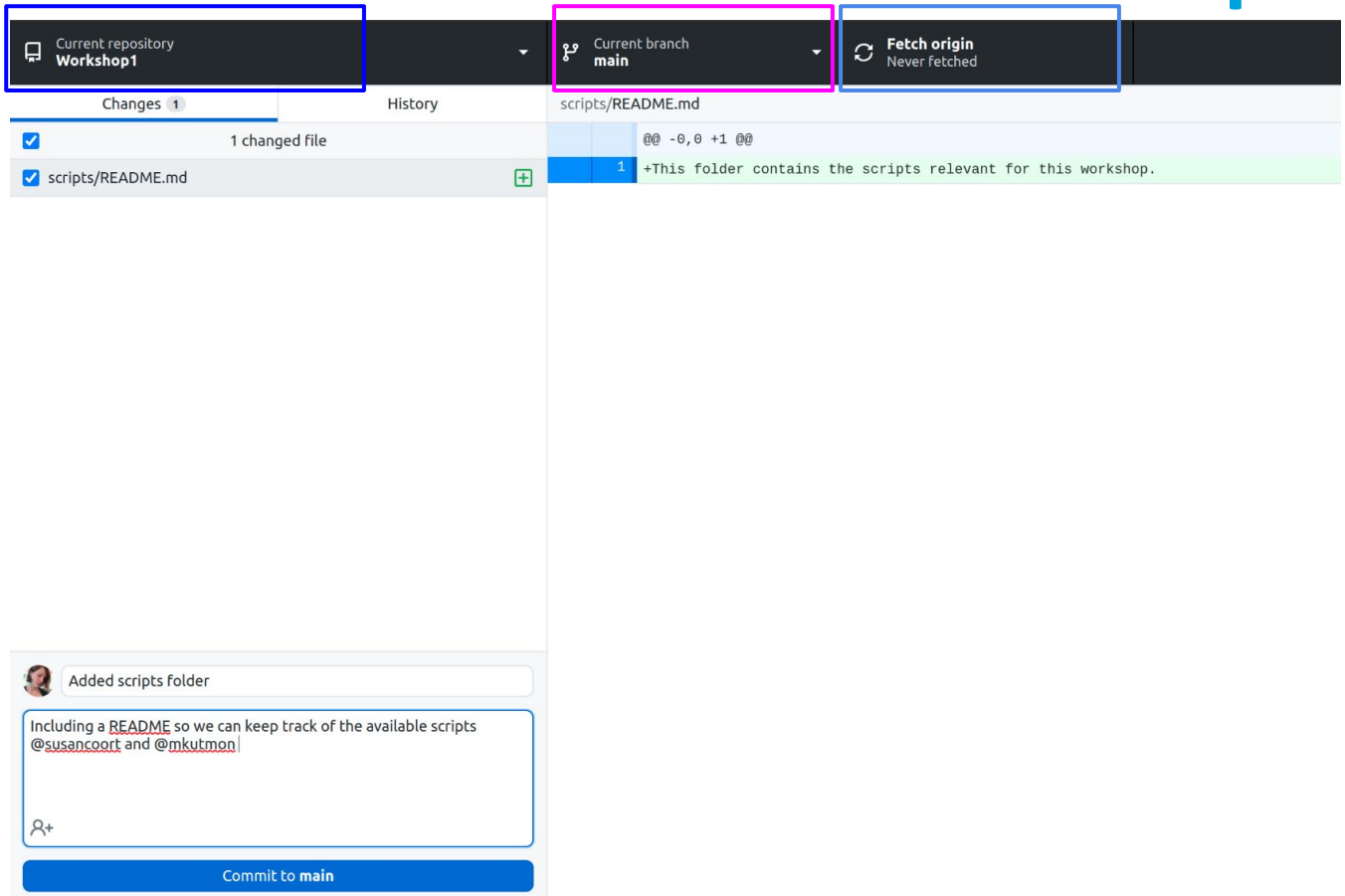
You should have something that looks like this now!

Hands-on session - GitHub Desktop



Hands-on session - GitHub Desktop

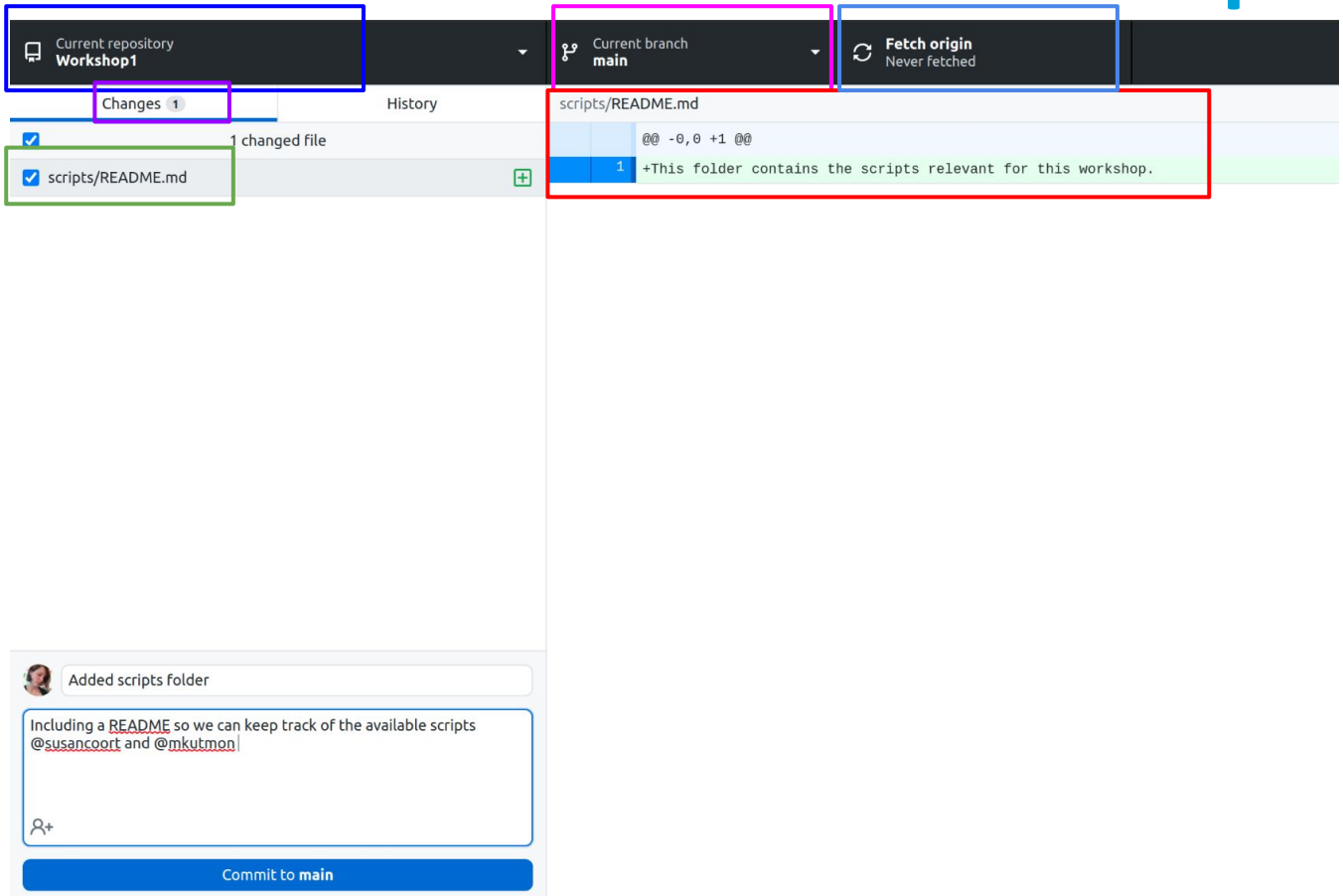
Folder we are working in



Branch ("version") we are in

Button to update local and online

Hands-on session - GitHub Desktop



Folder we are working in

Number of changes that occured

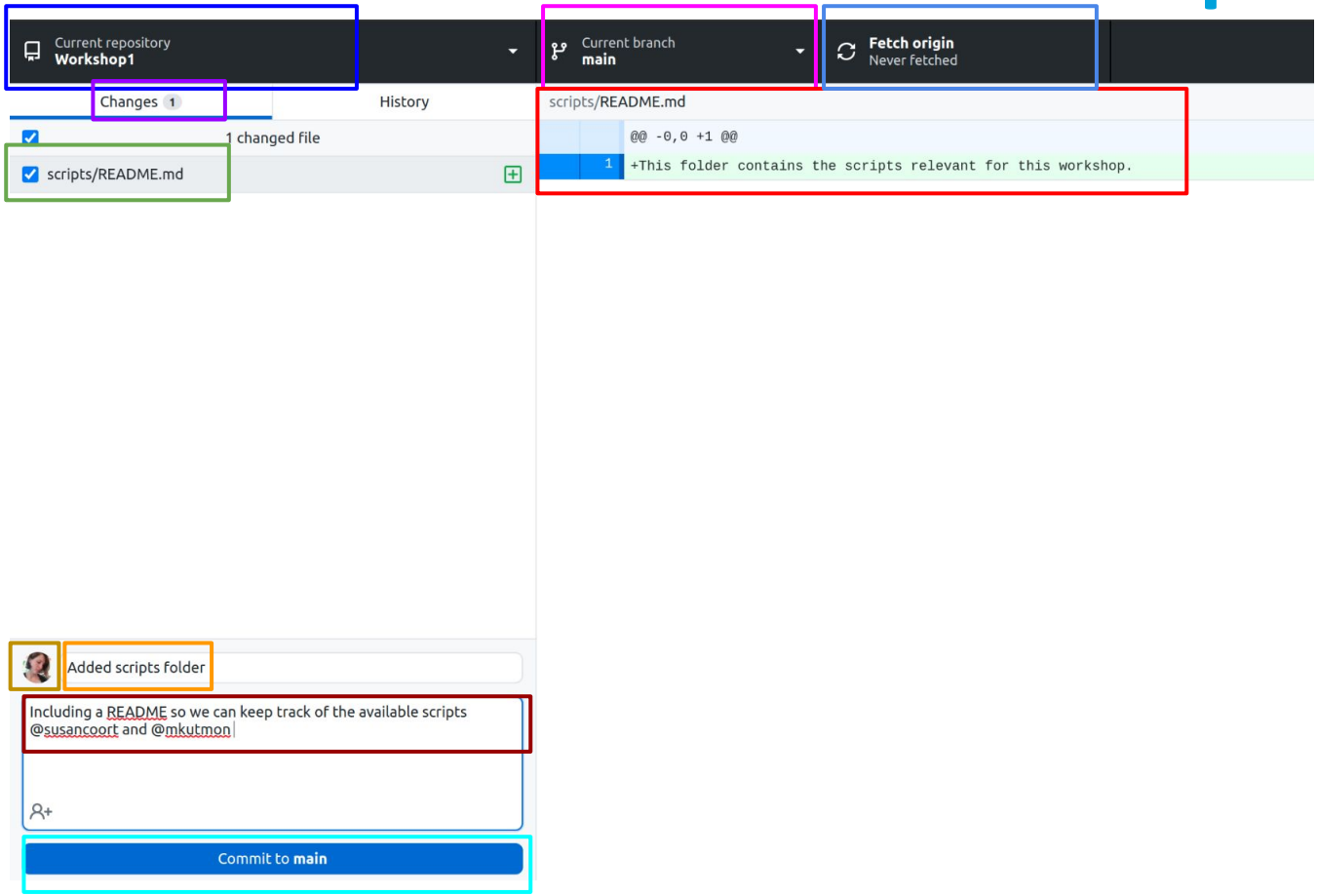
Name of changed file(s)

Branch ("version") we are in

Preview of change(s) in selected file

Button to update local and online

Hands-on session - GitHub Desktop



Folder we are working in

Number of changes that occured

Name of changed file(s)

Your GitHub account icon :D

Short name for the change (commit message)

Longer description for change

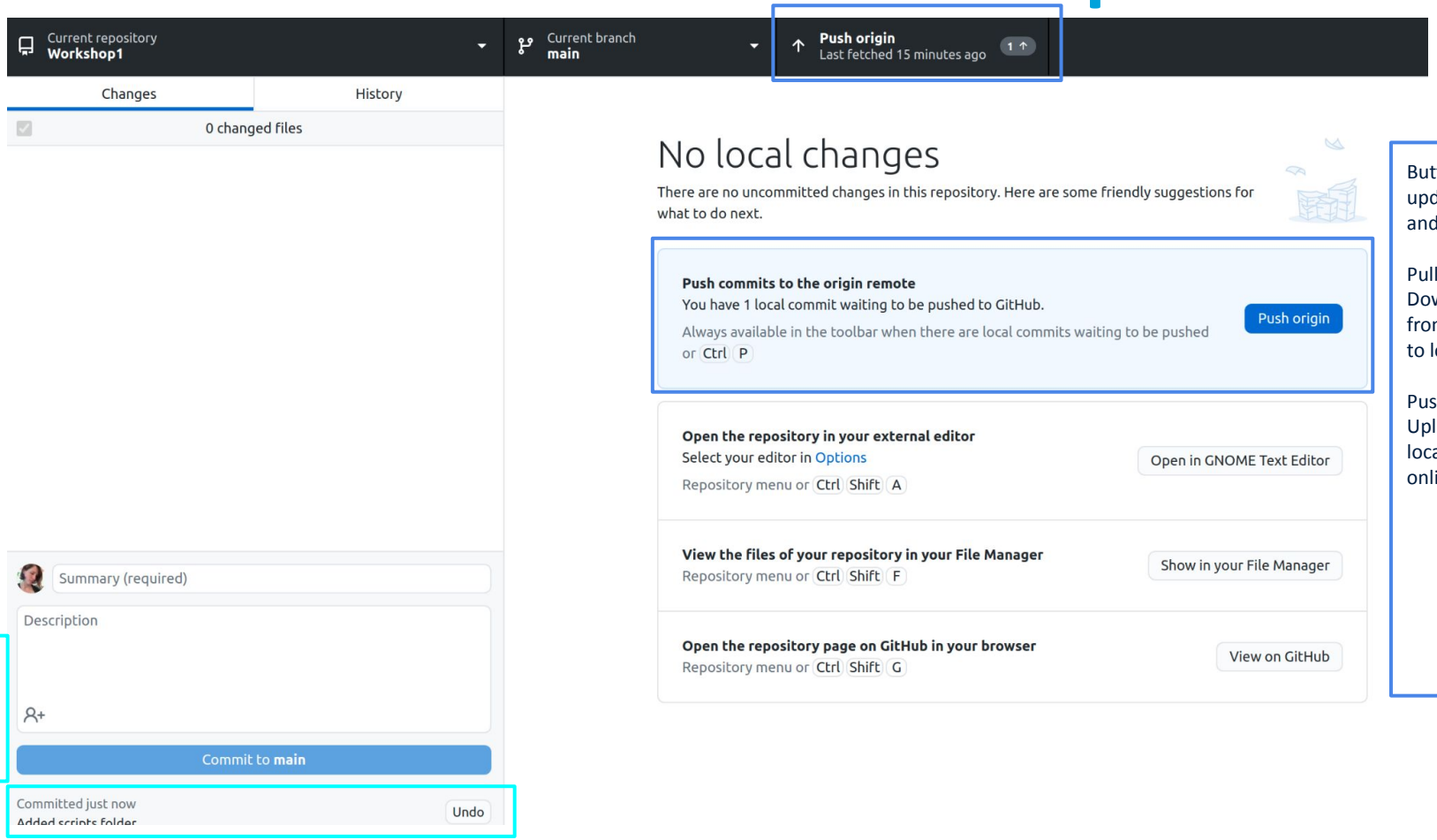
Branch ("version") we are in

Preview of change(s) in selected file

Button to save changes in local git

Button to update local and online

Hands-on session - GitHub Desktop



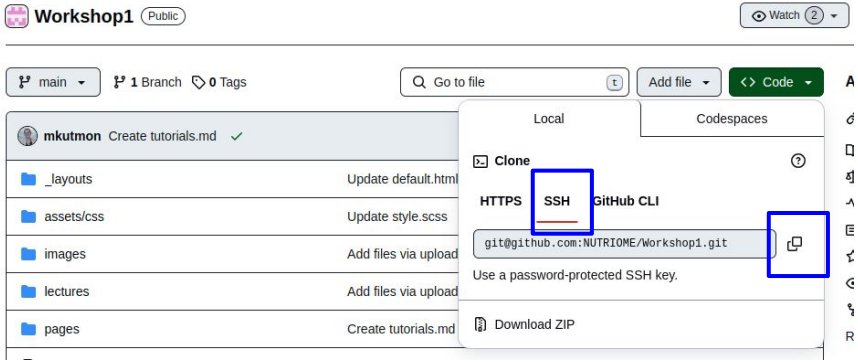
Button to save changes in local git: Successful!

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:



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:~$ cd git/NUTRIOME/Worksh
op1/
(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 add
ress
Offending key for IP in /home/deniseslenter/
Matching host key in /home/deniseslenter/.ss
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
   9b1e3ea..f1bb9ff  main -> main
```

Code to update local and online:

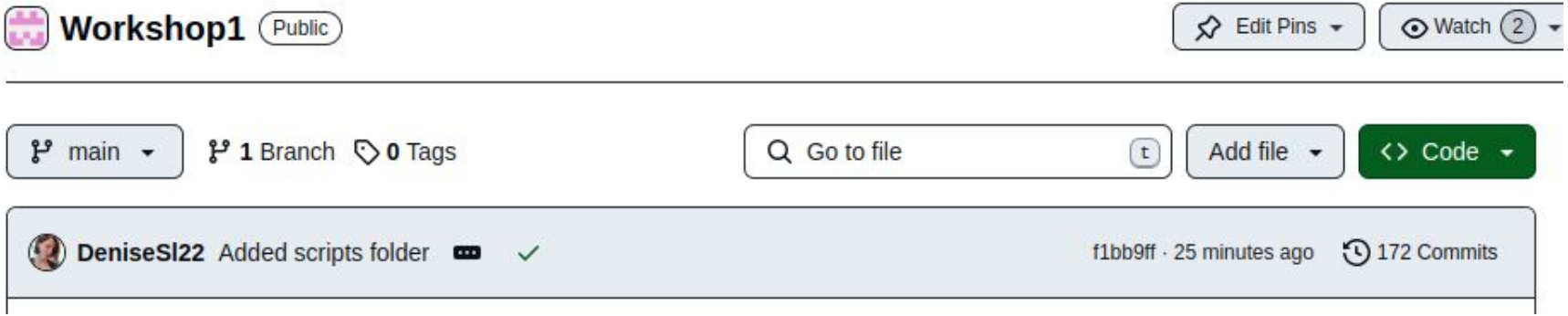
git pull = Download from online to local files

git push = Upload from local to online files

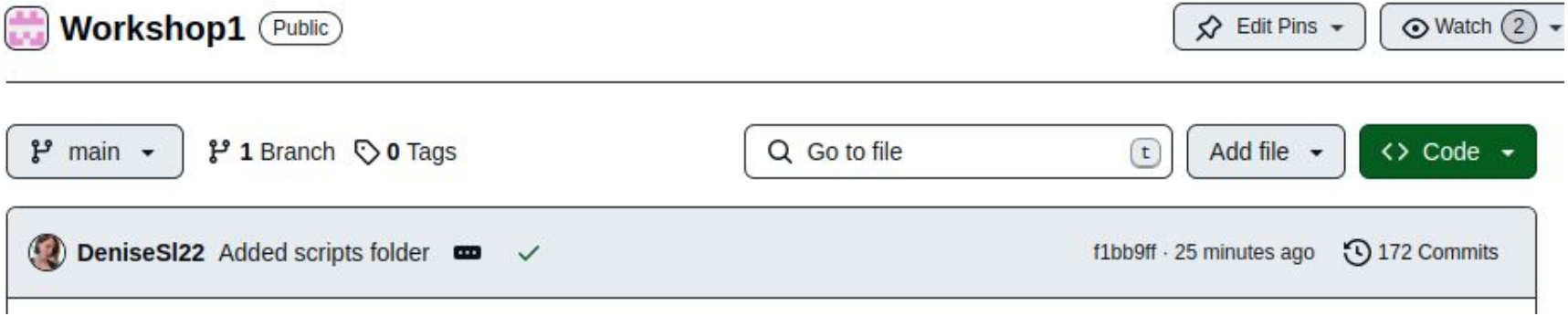
In case you failed previously setting up with SSH like me ;)

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

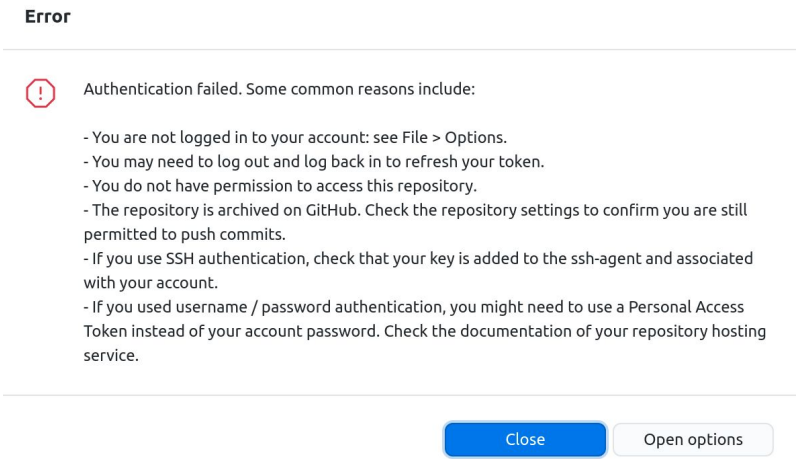
Hands-on session - GitHub Desktop



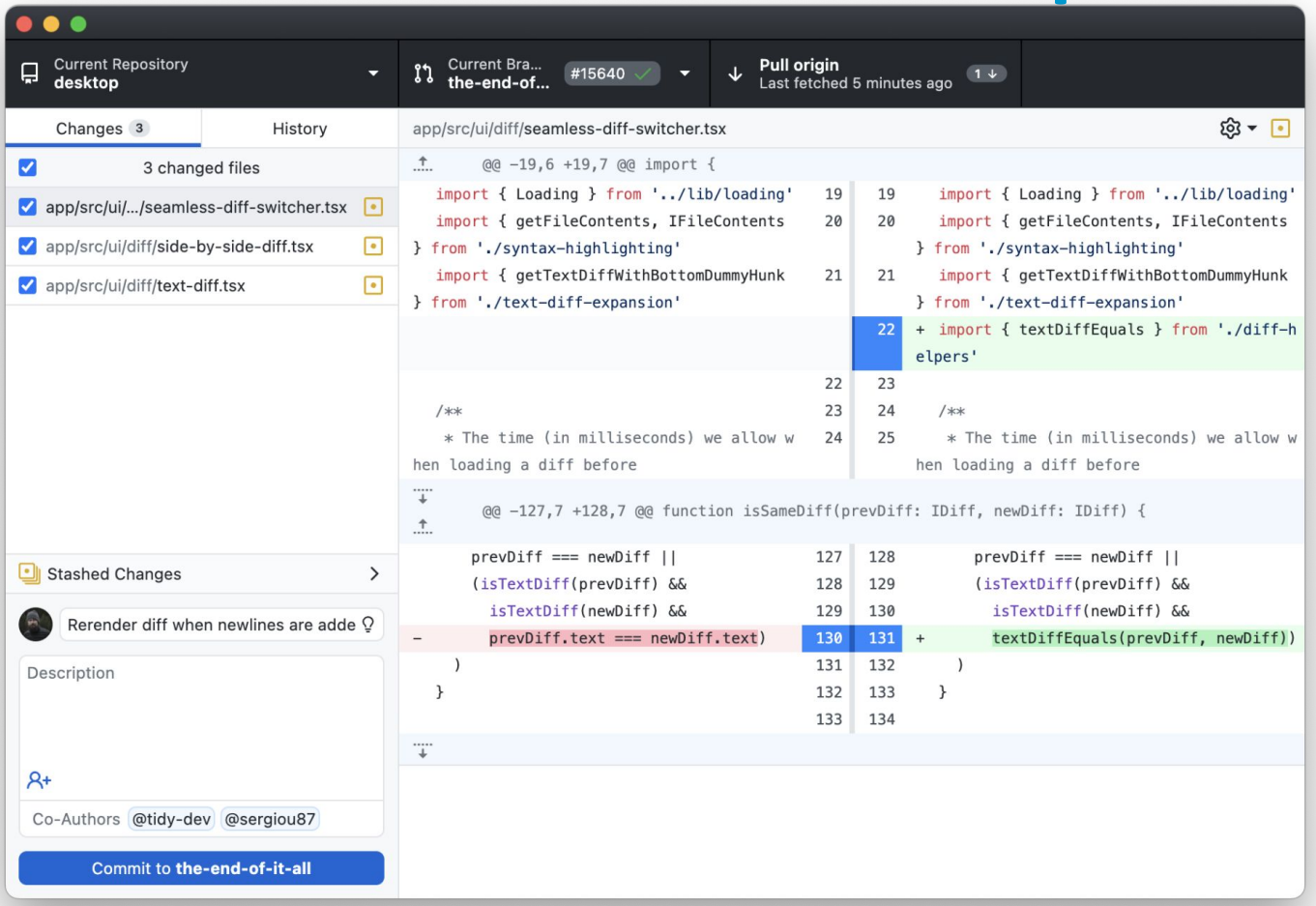
Hands-on session - GitHub Desktop



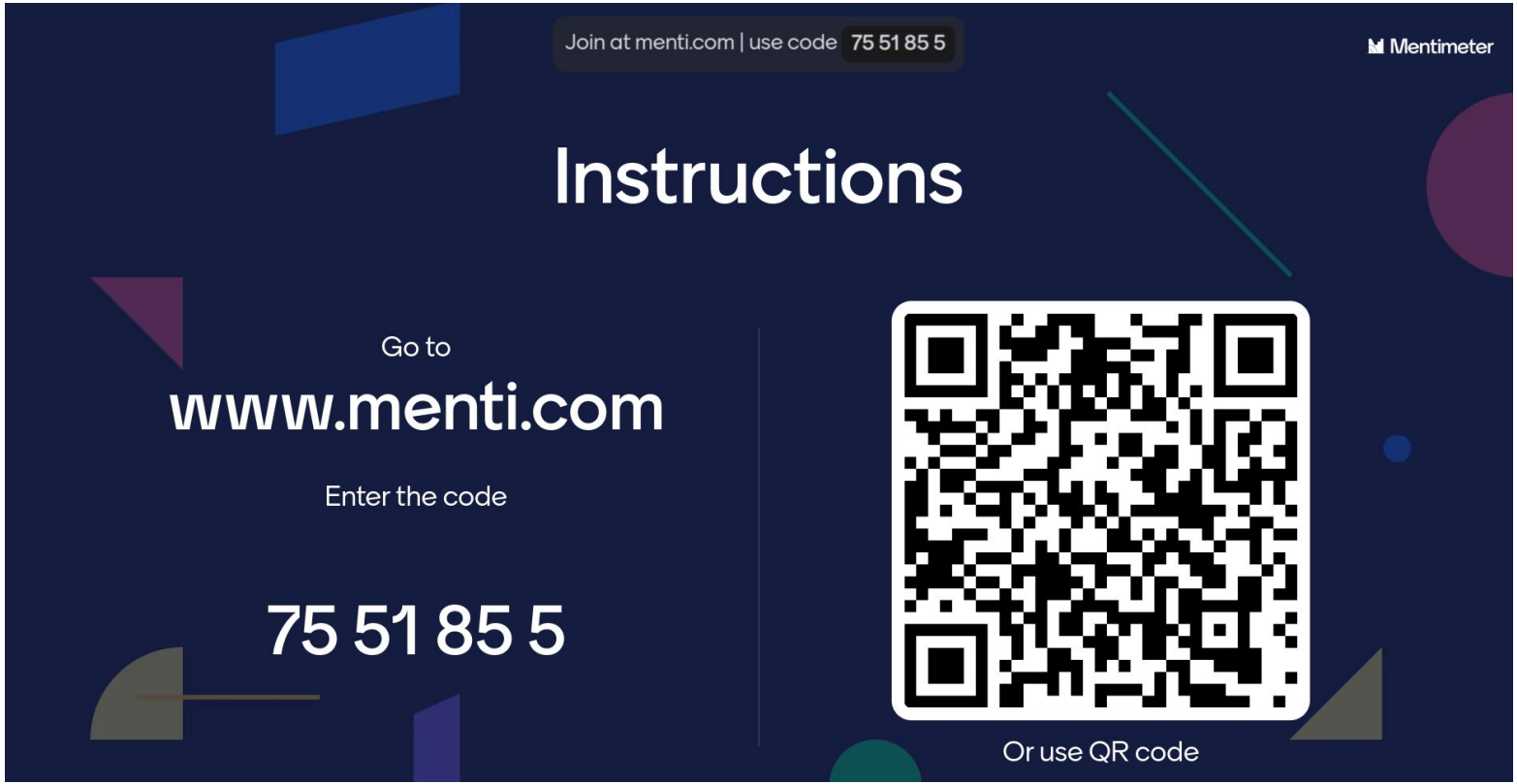
Note: if you get this message in GitHub Desktop, something in your setup is not correct, ask us for help!



Hands-on session - GitHub Desktop



Join our quiz!



Hands-on session - GitHub Desktop

Button to save changes in online git

DeniseSI22 / **Workshop1**

Type / to search

Code | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

Files

main

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

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

DeniseSl22 / **Workshop1**

Code | Pull requests | Actions

Files

main

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

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

Cancel | Commit changes

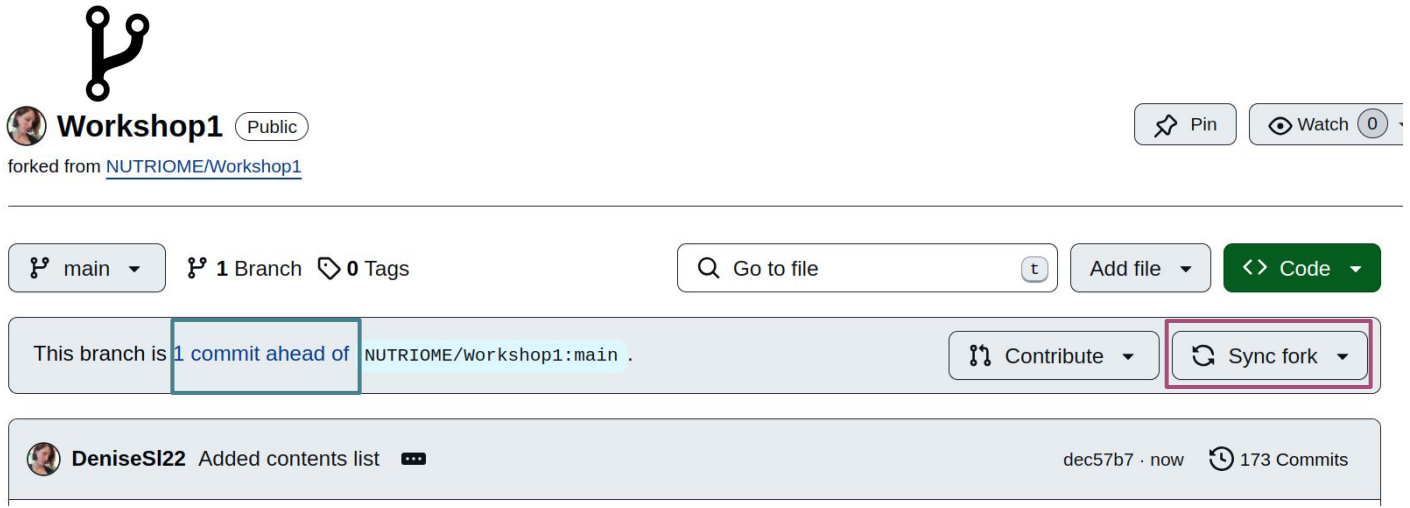
Button to save changes in online git

Cancel changes | Commit changes...

Spaces | 2 | Soft wrap

Branch ("version") we are in

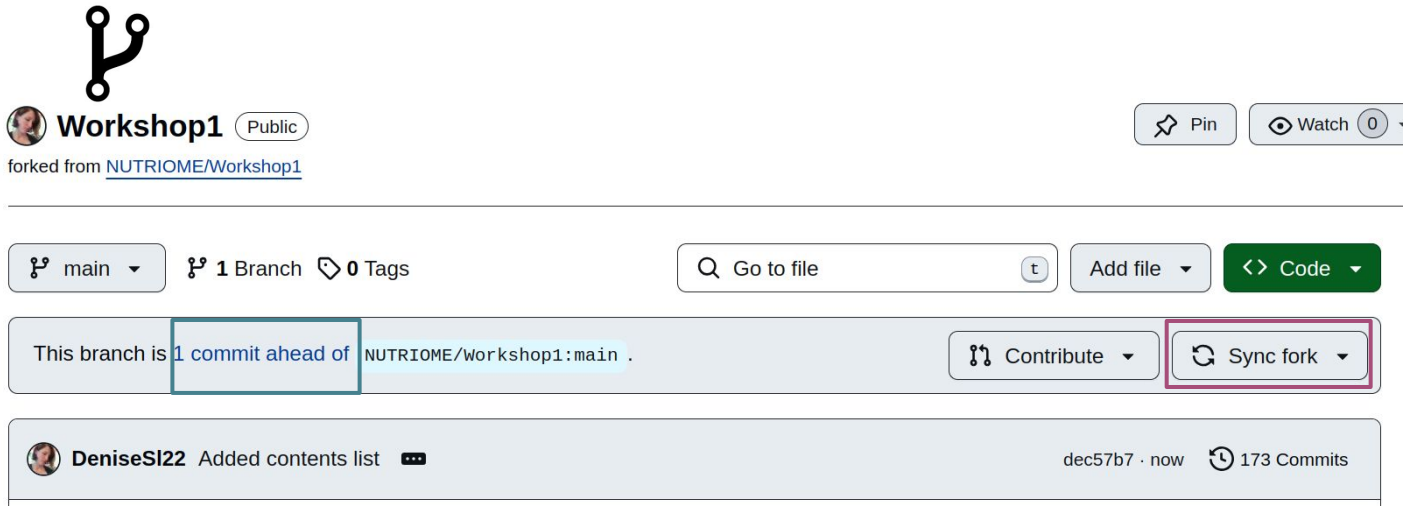
Hands-on session - GitHub Desktop



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

To check for updates from our version to the original one

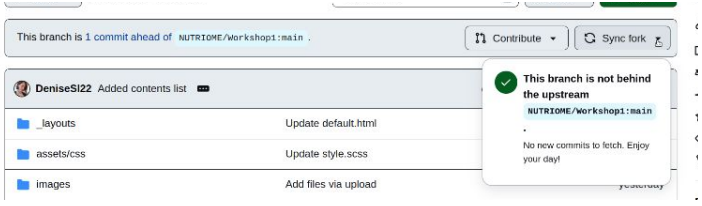
Hands-on session - GitHub Desktop



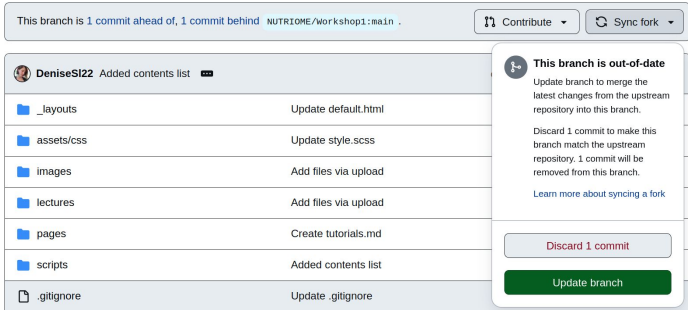
To request a merge from our content to the original one (Pull Request or PR)

To check for updates from our version to the original one

No ‘upstream’ changes

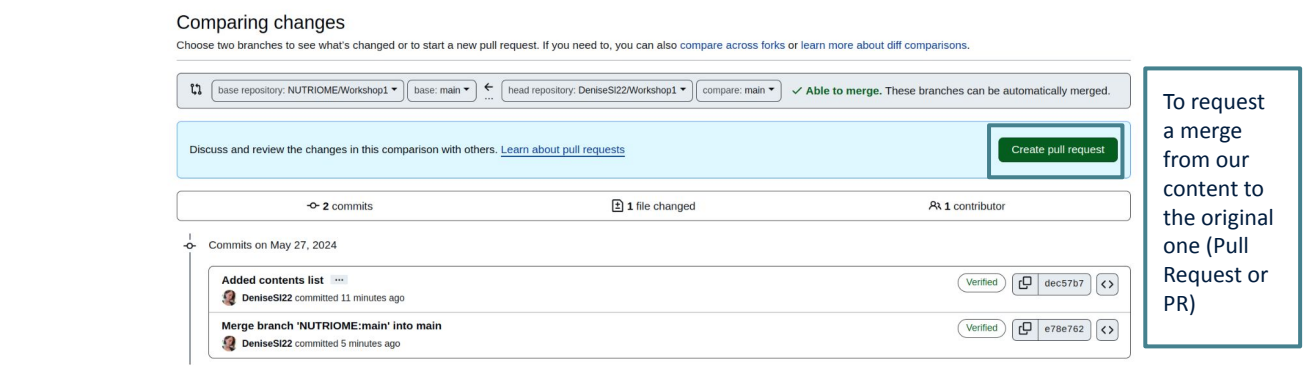


ONE ‘upstream’ change

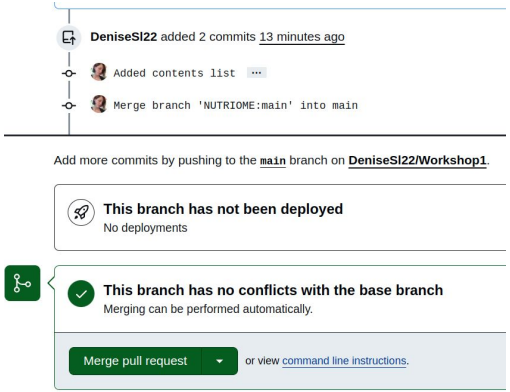


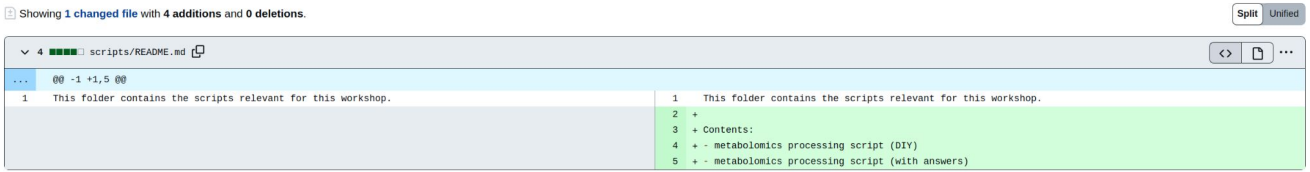
‘upstream’ change integrated in fork

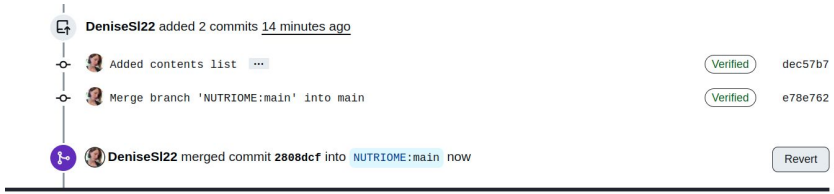
Hands-on session - GitHub Desktop



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







Hands-on session - GitHub Desktop

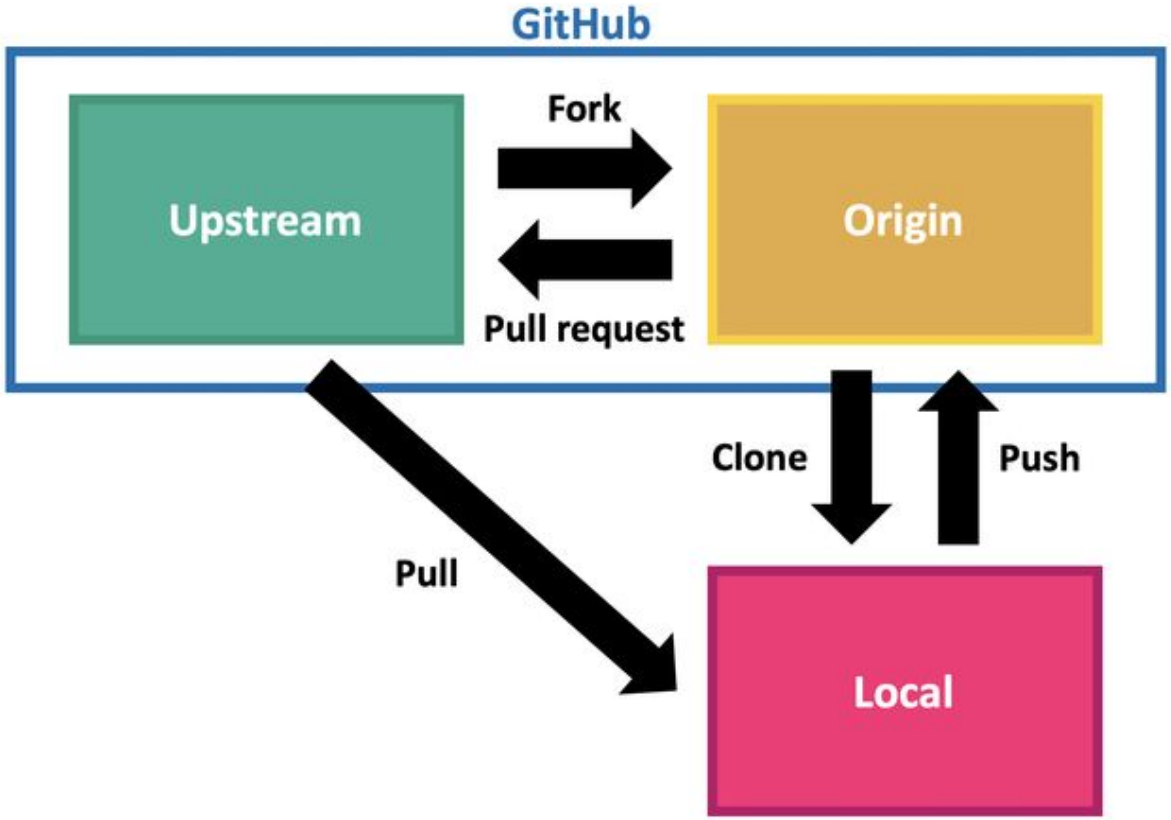


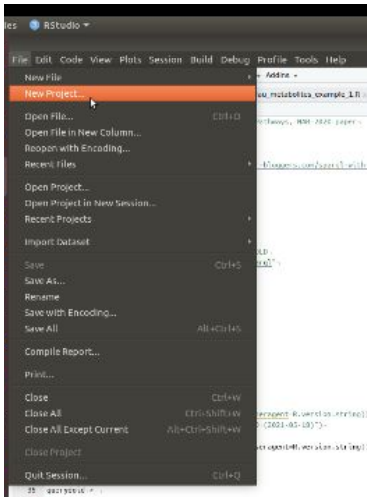
Image from: https://anujf0510.medium.com/the-guide-to-git-terminology-bbae2a3d8af5 and https://medium.com/@vishwasacharya/github-desktop-vs-command-line-choose-the-right-tool-for-you-feb58c3f0e30

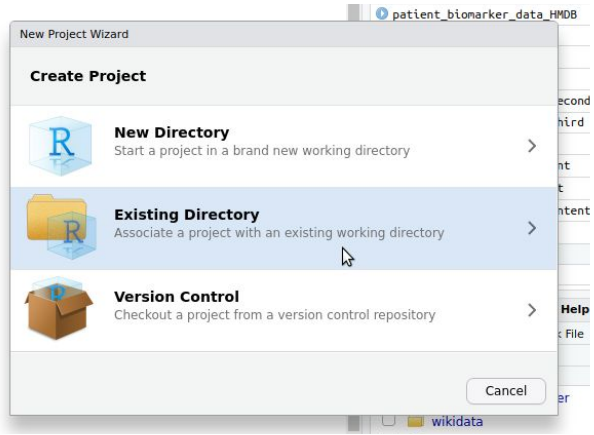
Hands-on session - Adding GitHub to Rstudio

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

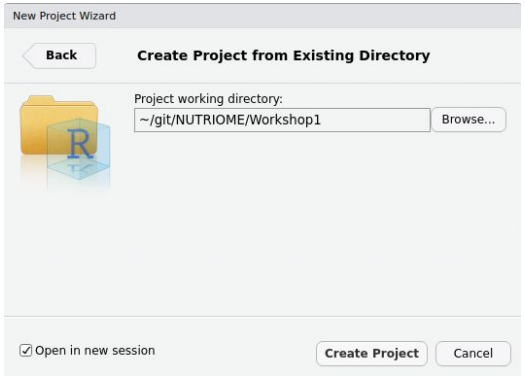
Break time!?

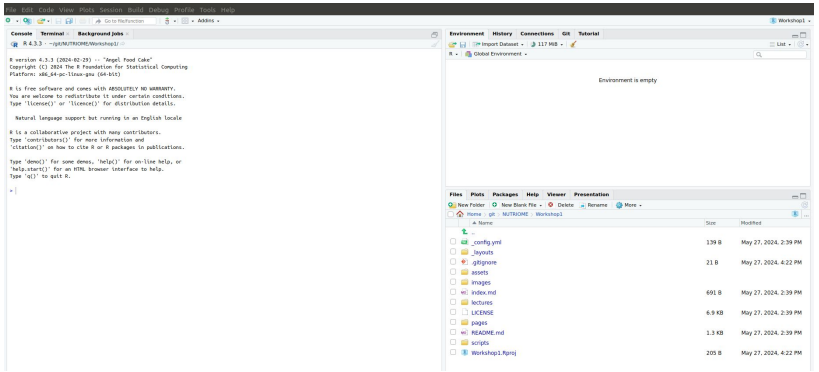
Hands-on session - Adding GitHub to Rstudio

Step 4: Select the top folder (Workshop1)

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

Step 6: Click → Create Project



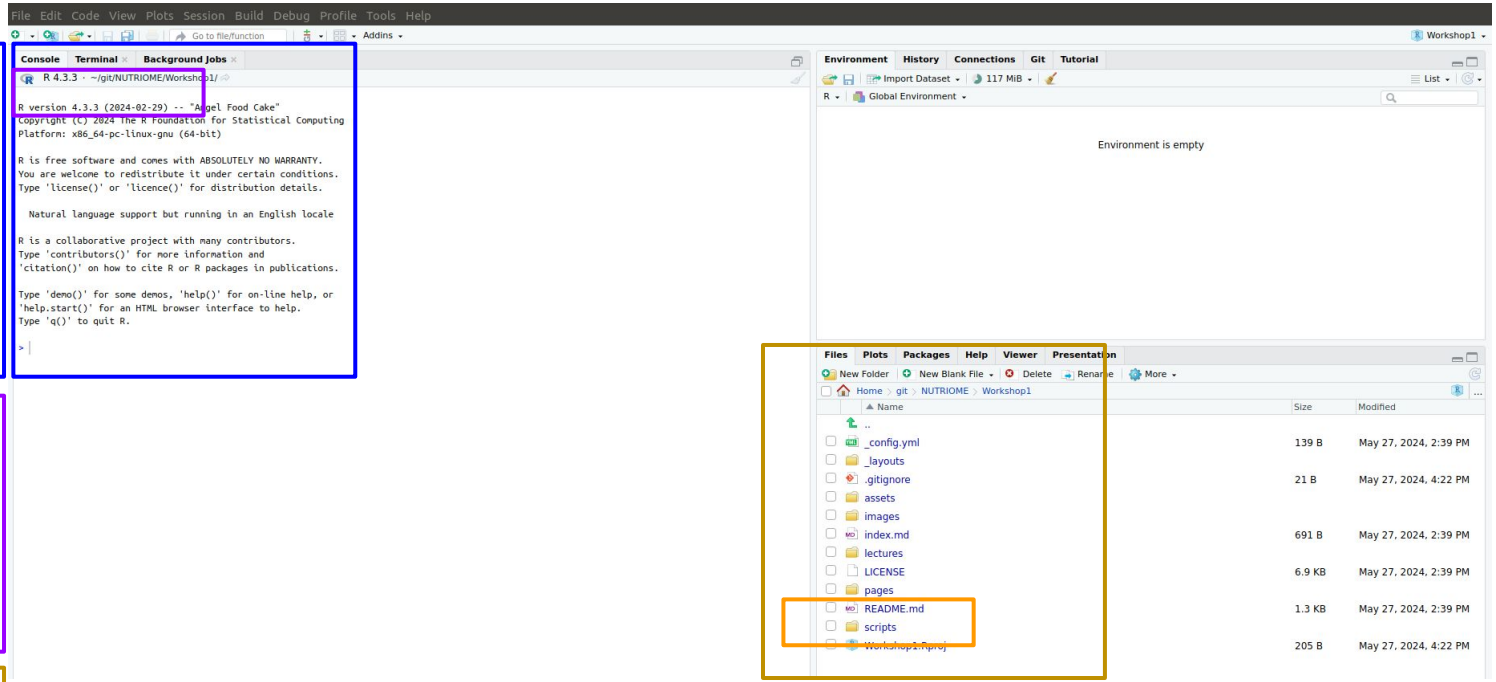


Hands-on session - Adding GitHub to Rstudio

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

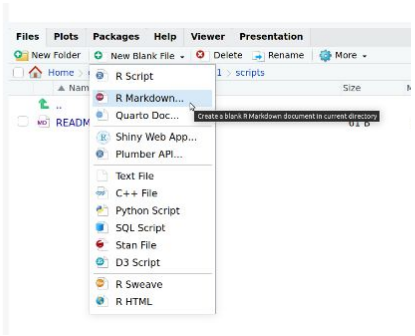
Currently loaded version of R (check this for potential compatibility issues!)

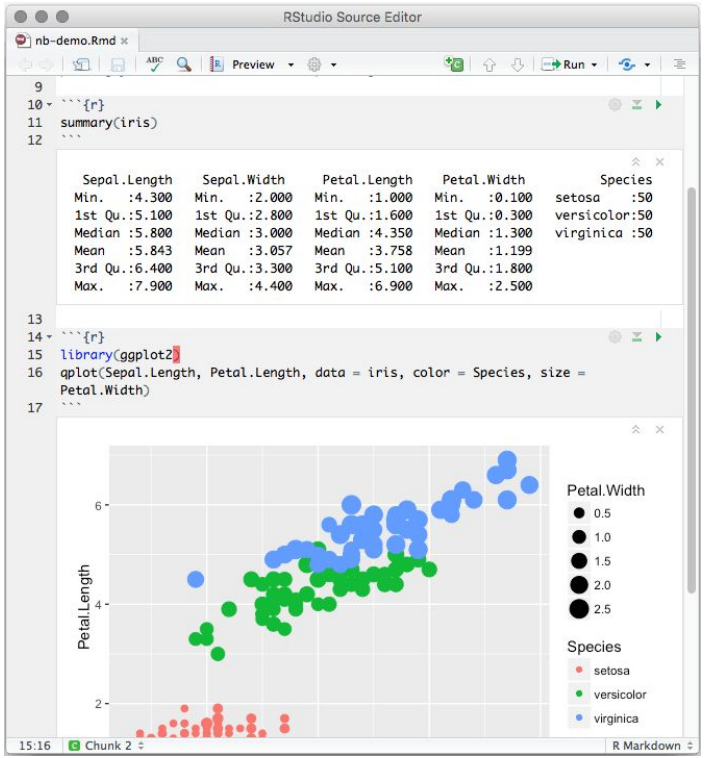
Folder structure (according to GitHub Repository)



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

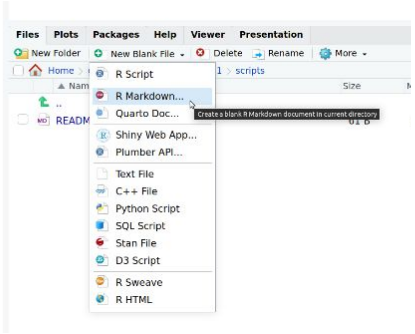
Hands-on session - Adding GitHub to Rstudio

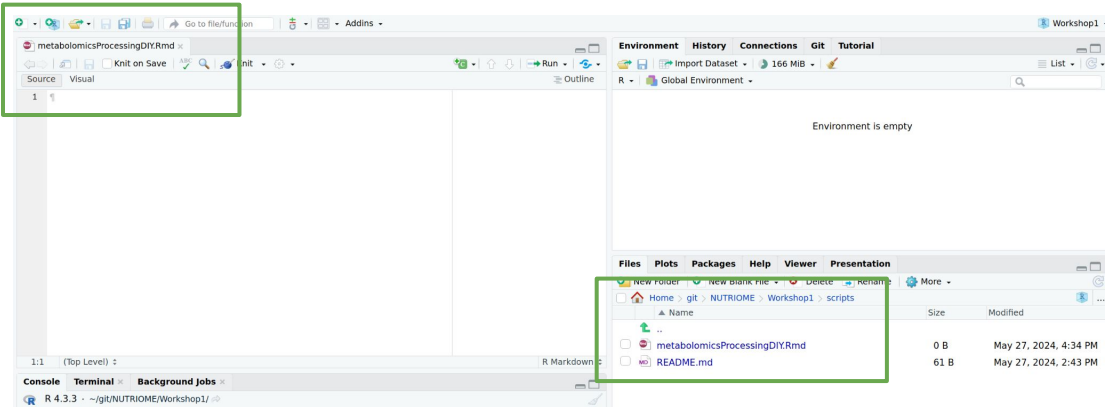


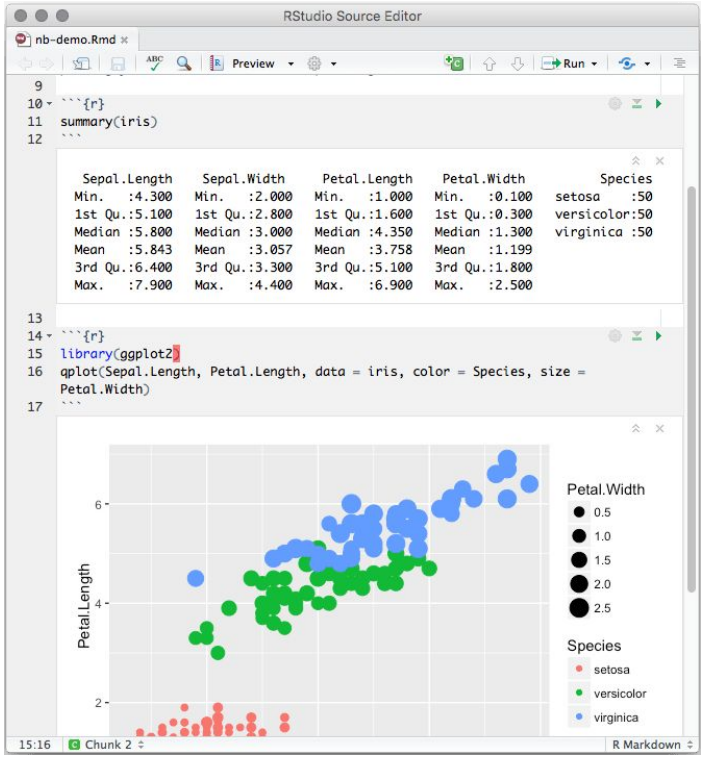


https://bookdown.org/yihui/rmarkdown/notebook.html

Hands-on session - Adding GitHub to Rstudio



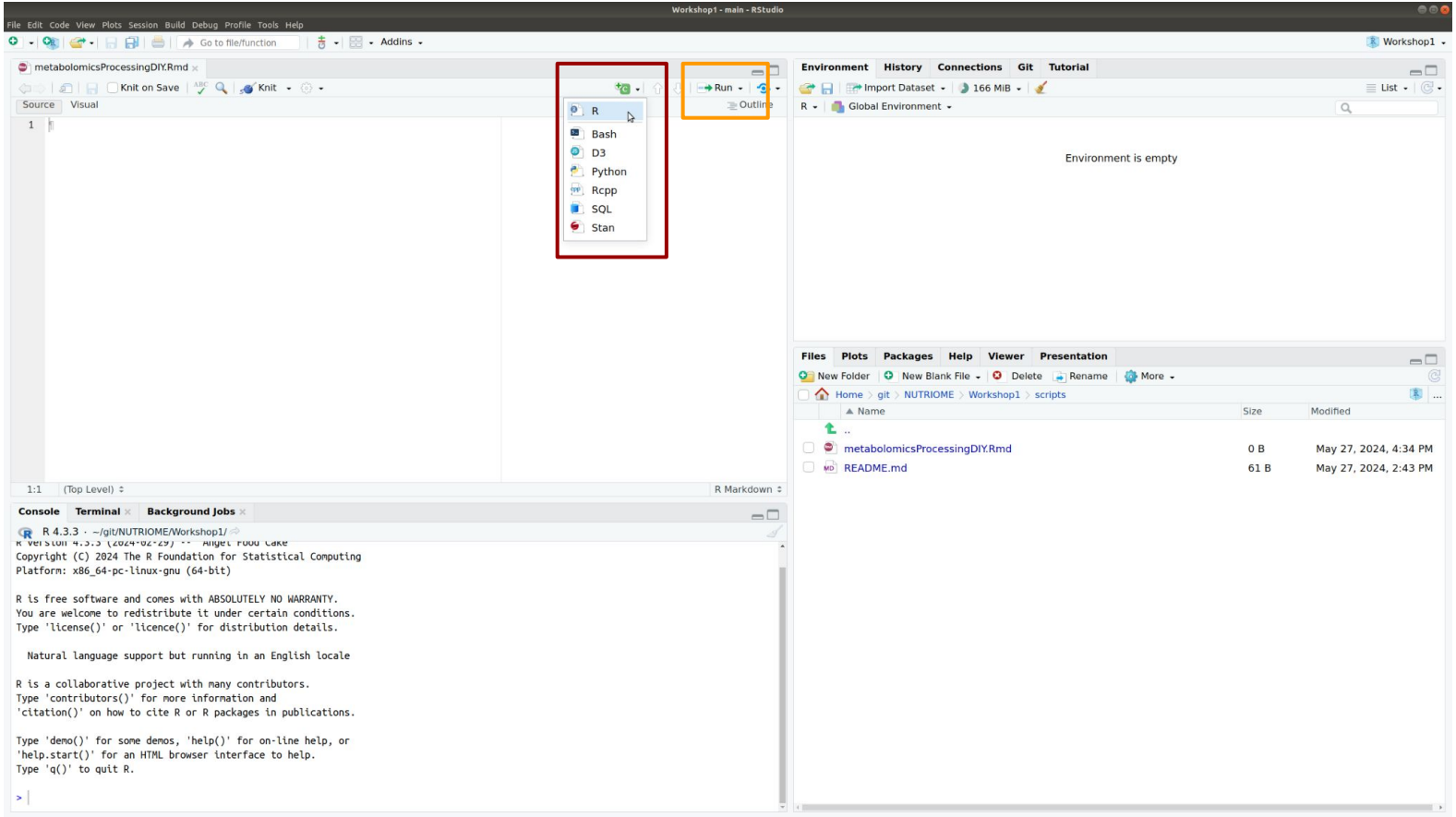




New ‘Markdown’ file create in File Explorer and in top left menu

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

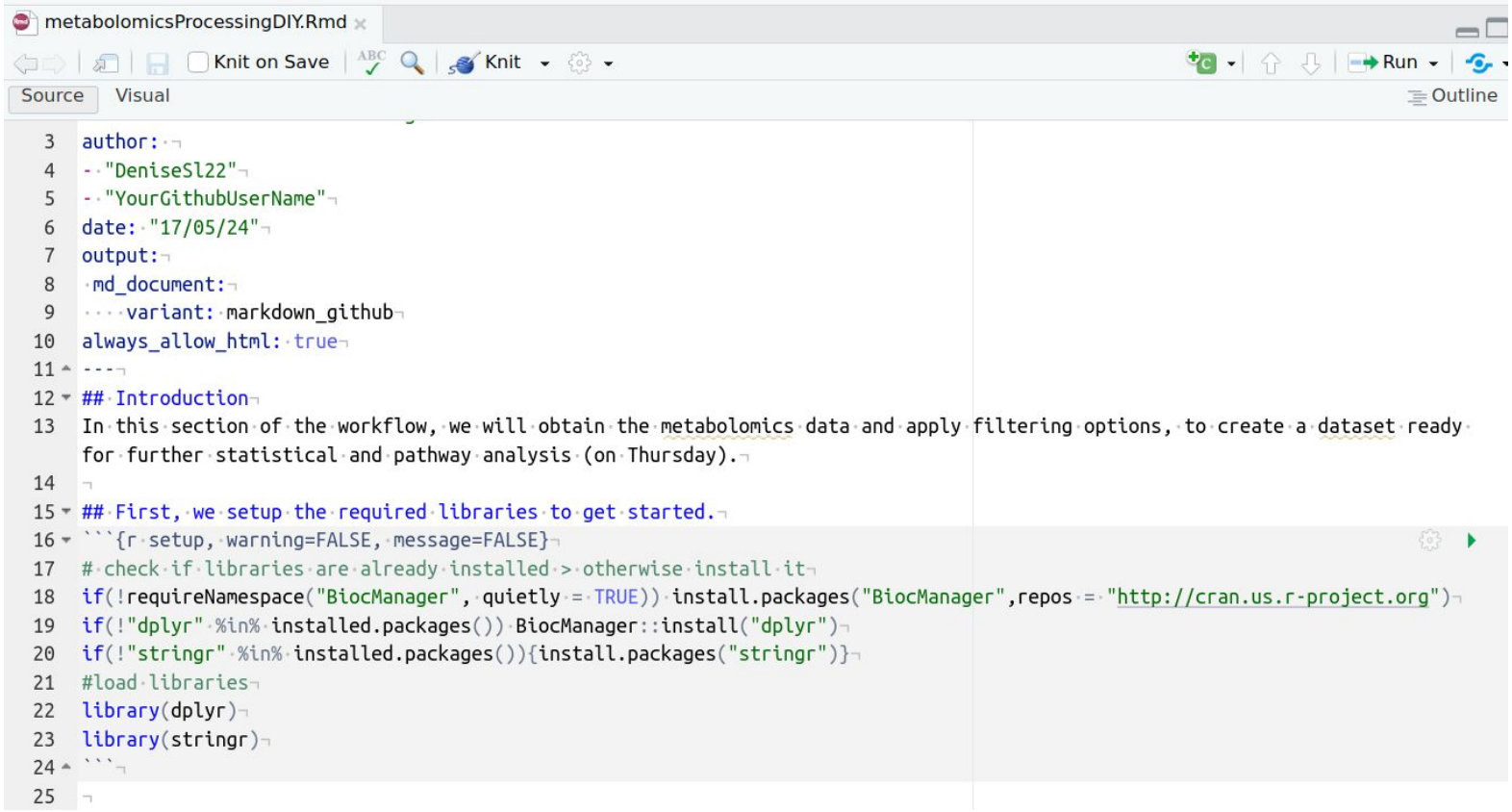
Hands-on session - Adding GitHub to Rstudio



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

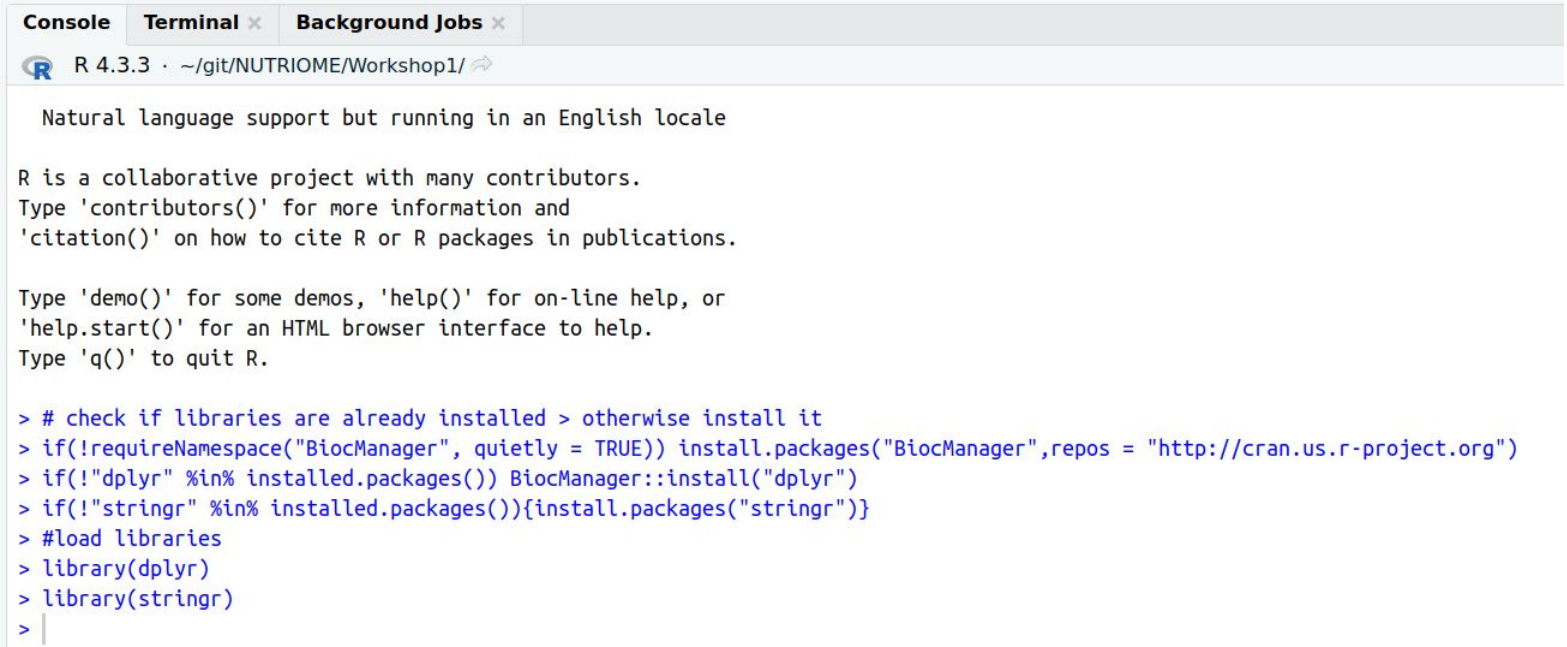


Make sure to have the following content in your Code chunk (select the correct file) and click 'Run' (or the green arrow in the chunk)!

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

Save the file (locally) and check your GitHub Desktop

Hands-on session - First programming tasks



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!

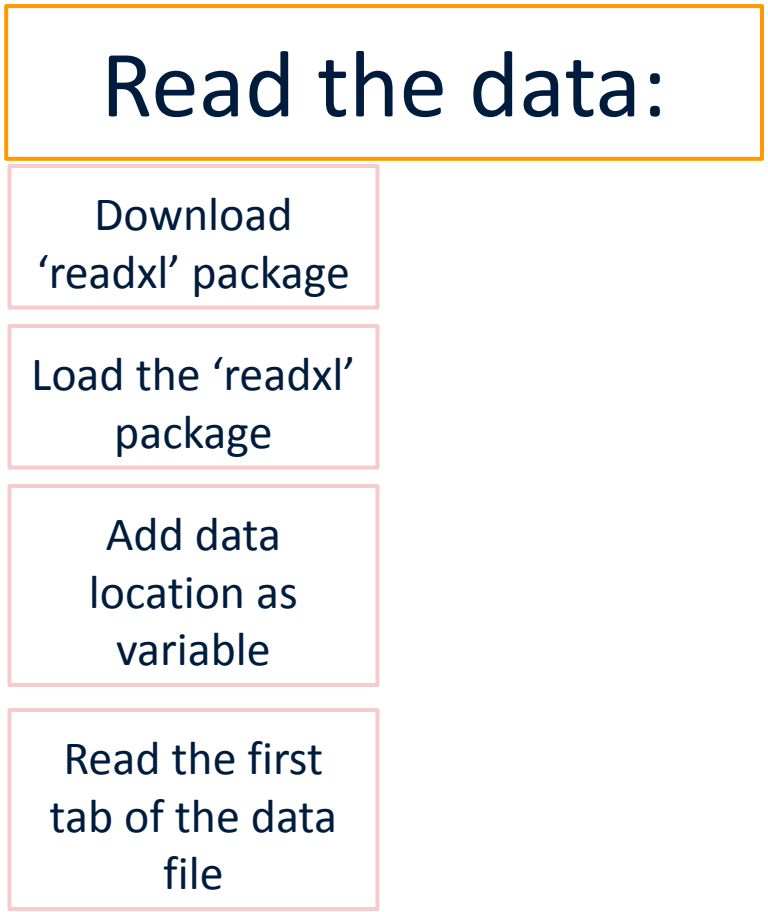
Hands-on session - First programming tasks

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!)



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).

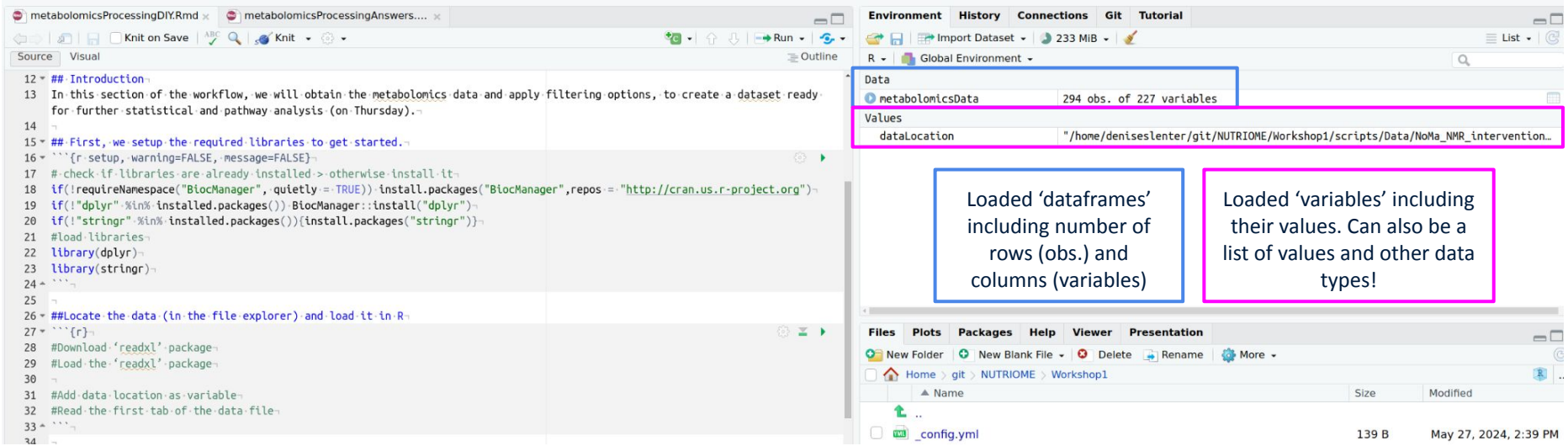
3. Check in GitHub Desktop if this file change is not visible (data should not be shared online!)

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



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

Dataframe name

Search Box to explore data

Sample number in Dataframe

Sample name in dataset

No data measured (NA or NaN) → Not zero!

metabolomicsProcessingDIY.Rmd | metabolomicsData | metabolomicsProcessingAnswers....

Filter | Cols: « < 1 50 > »

	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	4.7920000000000001E-11	1.021E-2	8.5240000000000001E-4	2.2290000000000001E-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	7.5639999999999997E-11	1.575E-2	1.4890000000000001E-3	2.2690000000000002E-3	1.6559999999999
8	1_F4	end	1.3580000000000001E-10	2.844E-2	3.0799999999999998E-3	3.908E-3	2.3389999999999
9	NA	delta	6.0160000000000002E-11	1.269E-2	1.59099999999999E-3	1.6389999999999901E-3	6.8300000000000
10	1_B3	base	1.9310000000000001E-10	4.1320000000000003E-2	4.5079999999999999E-3	7.7270000000000004E-3	4.7109999999999
11	1_B4	end	9.7399999999999995E-11	2.0760000000000001E-2	2.0270000000000002E-3	3.705E-3	2.1199999999999
12	NA	delta	-9.5700000000000003E-11	-2.0559999999999998E-2	-2.4809999999999902E-3	-4.0220000000000004E-3	-2.591E-3
13	1_B1	base	1.031E-10	2.164E-2	2.532E-3	3.202E-3	2.2000000000000
14	1_B2	end	NA	NA	NA	NA	NA
15	NA	delta	NA	NA	NA	NA	NA
16	1_B5	base	4.5270000000000002E-11	8.9549999999999994E-3	5.0319999999999998E-4	2.4230000000000001E-4	1.752E-4
17	1_B6	end	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

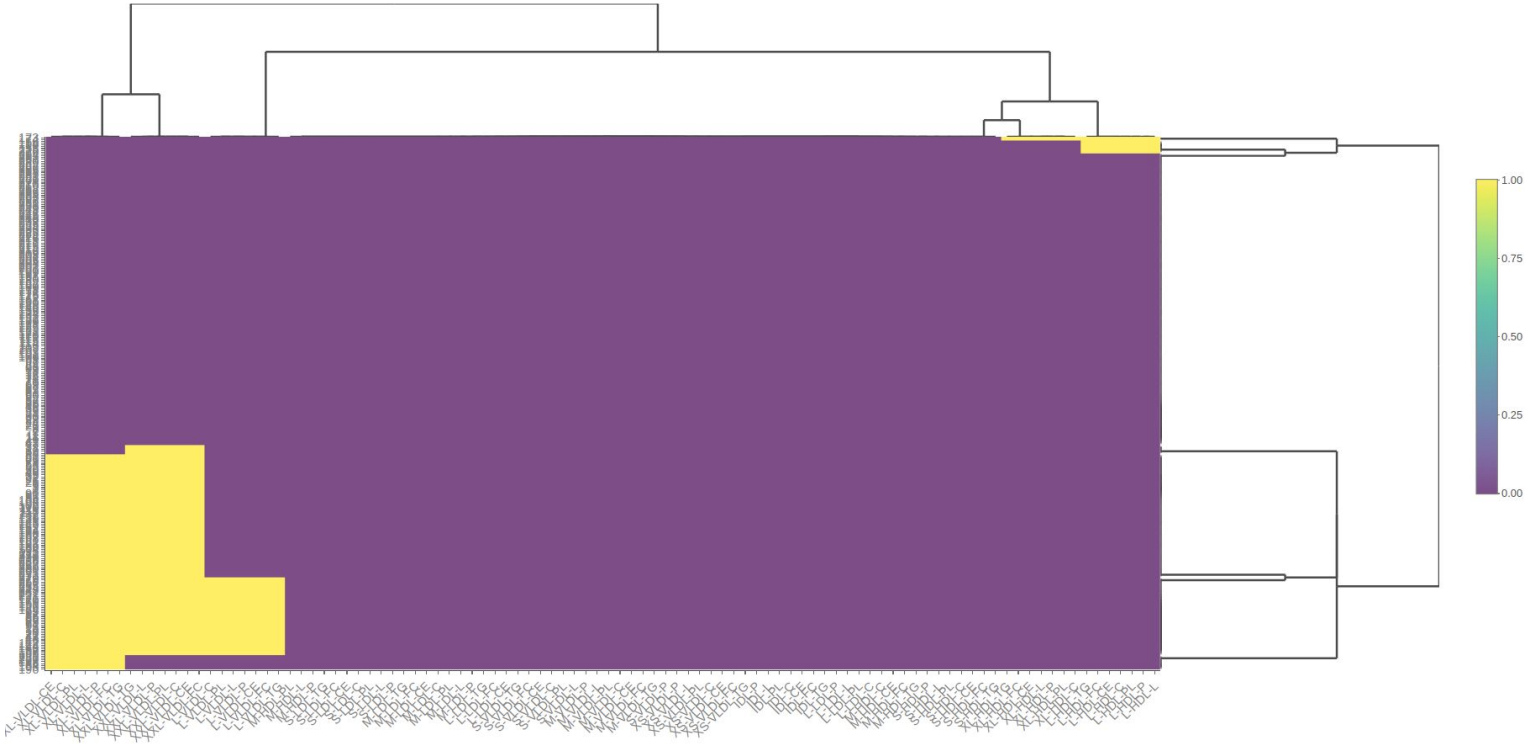
Credits to: https://thomasadventure.blog/posts/r-count-na/

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

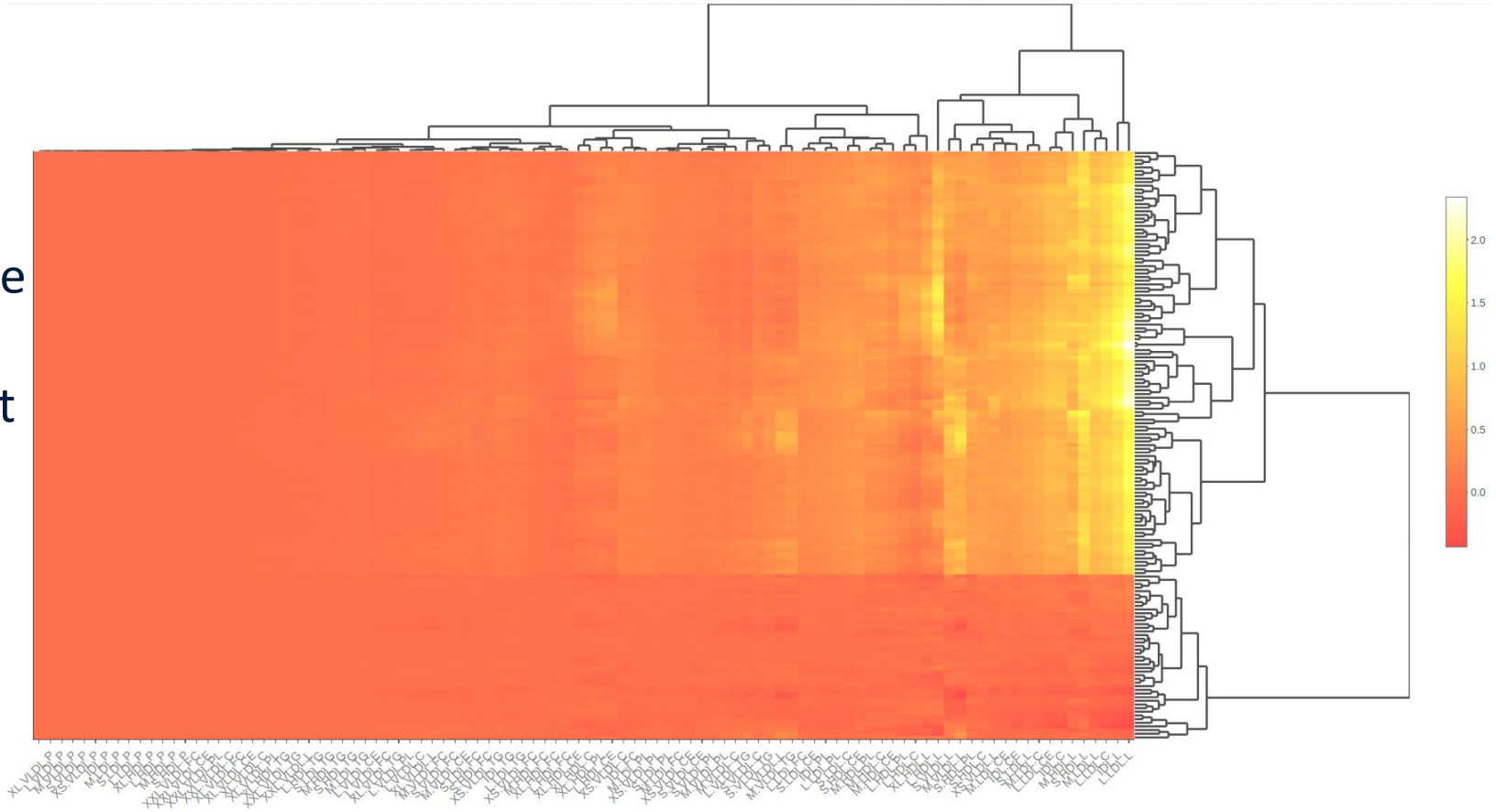
3. Visualize all information in a heatmap

   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

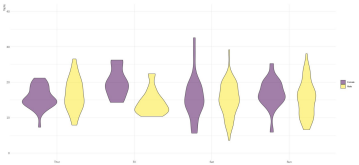
4. Visualize transformed data in heatmap

   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

Credits to: https://statsandr.com/blog/outliers-detection-in-r/ ; https://r-graph-gallery.com/violin_grouped_ggplot2.html

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!