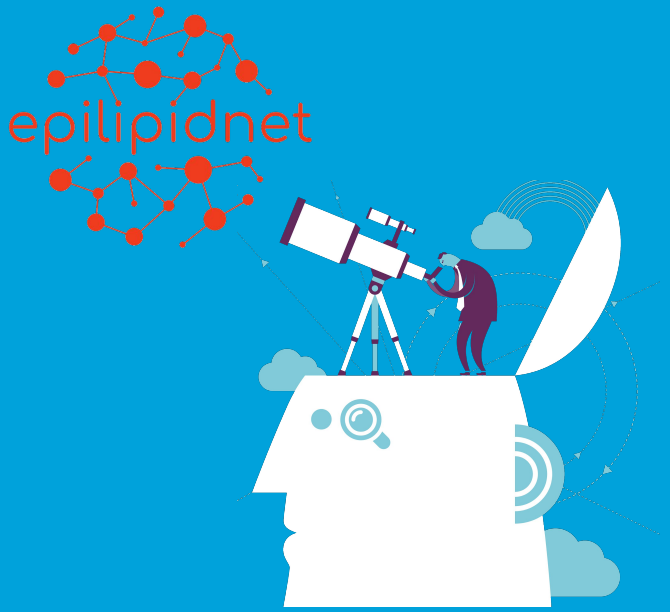
Metabolomics data analysis

Denise Slenter

ORCID: 0000-0001-8449-1318

Tuesday March 11, 2025



Maastricht University

What will I cover?

+ Existing tools for metabolomics data analysis, pros and cons

+ Steps in processing annotated metabolomics data

+ Working with Biostudies data

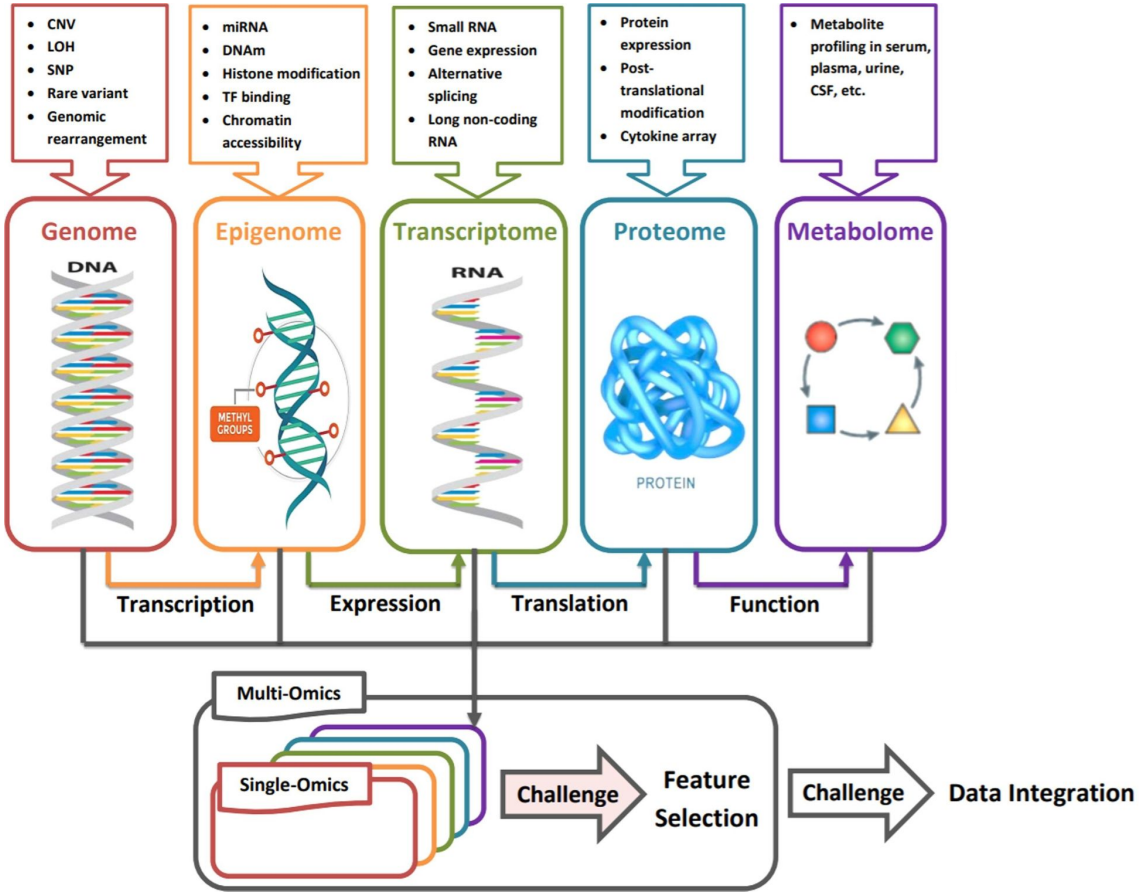
What can I NOT cover?

- Annotating raw data

- Overview of the metabolomics research field

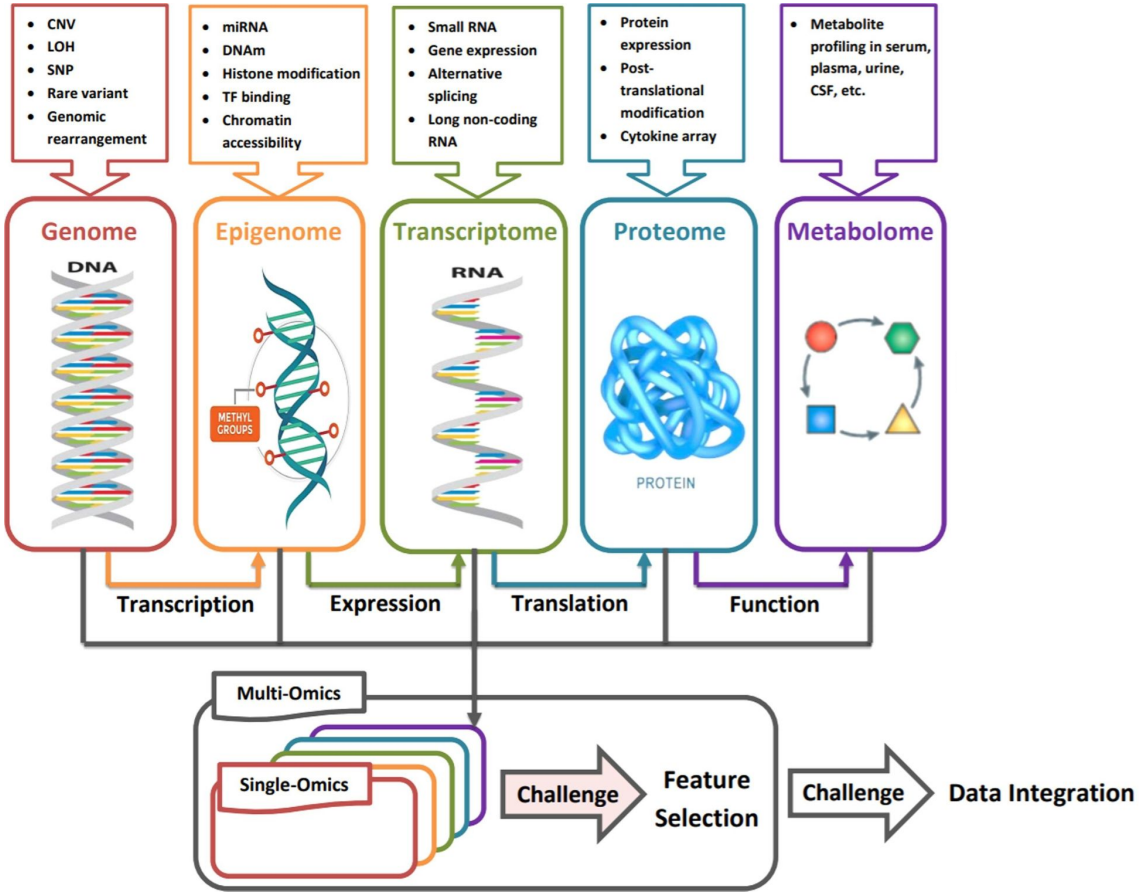
- ‘Plug and Play’ data analysis scripts - you will need to write some code yourself :)

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

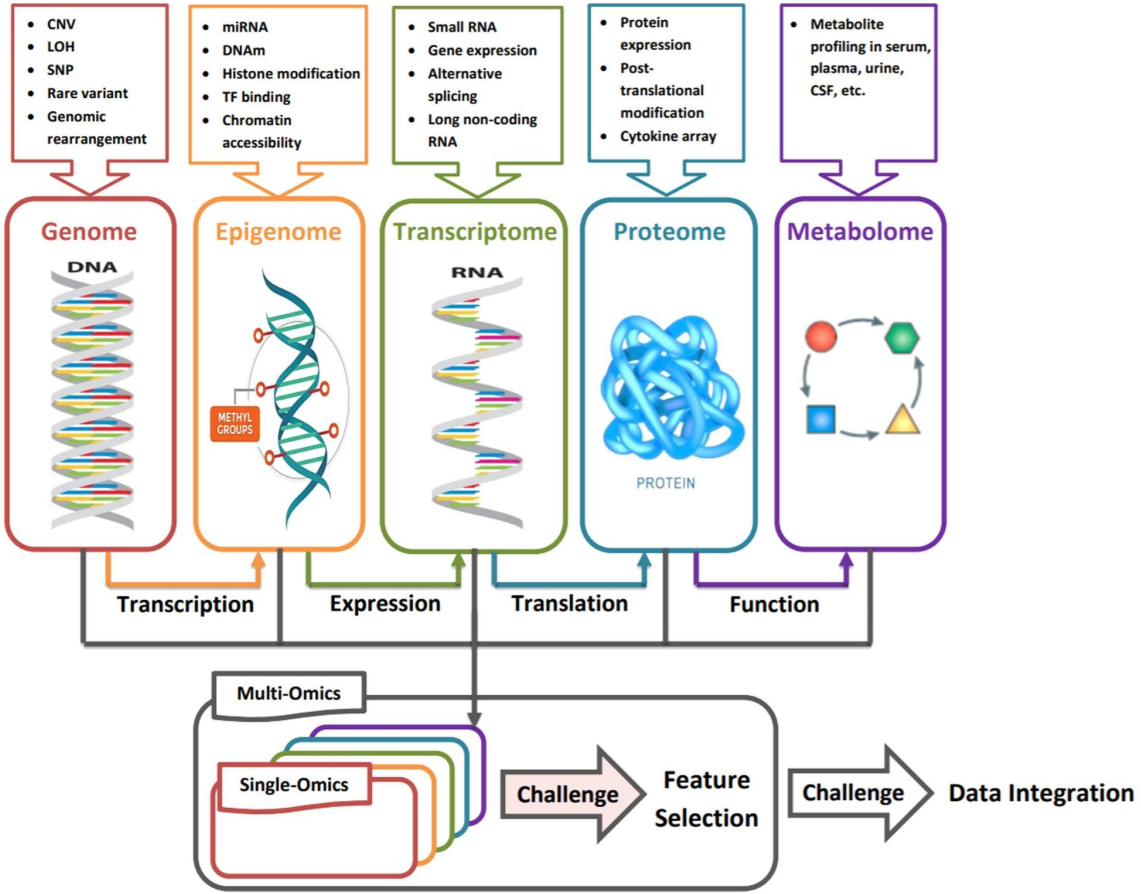
Different data types and analysis techniques



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

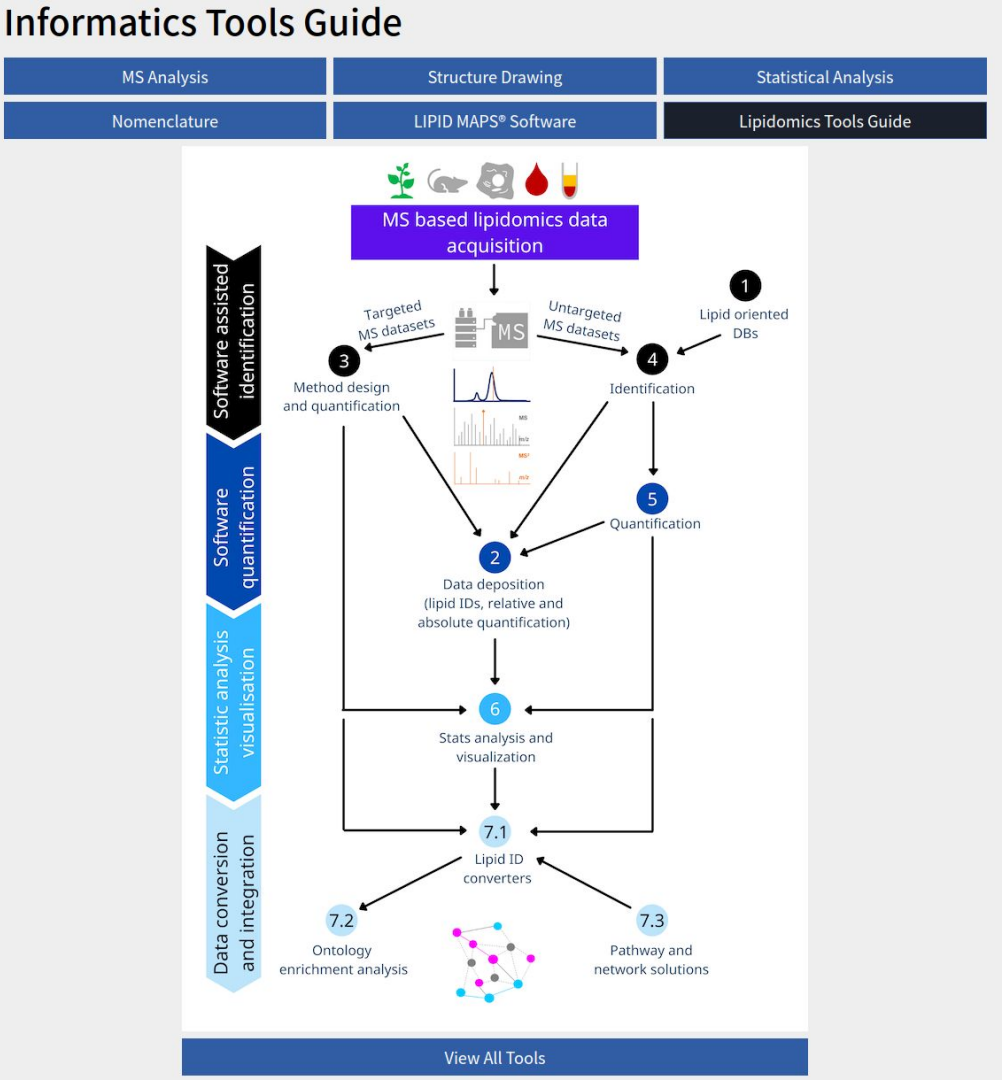
Different data types and analysis techniques



Missing:

- Phenotype
- Imaging data
- Fluxomics
- …

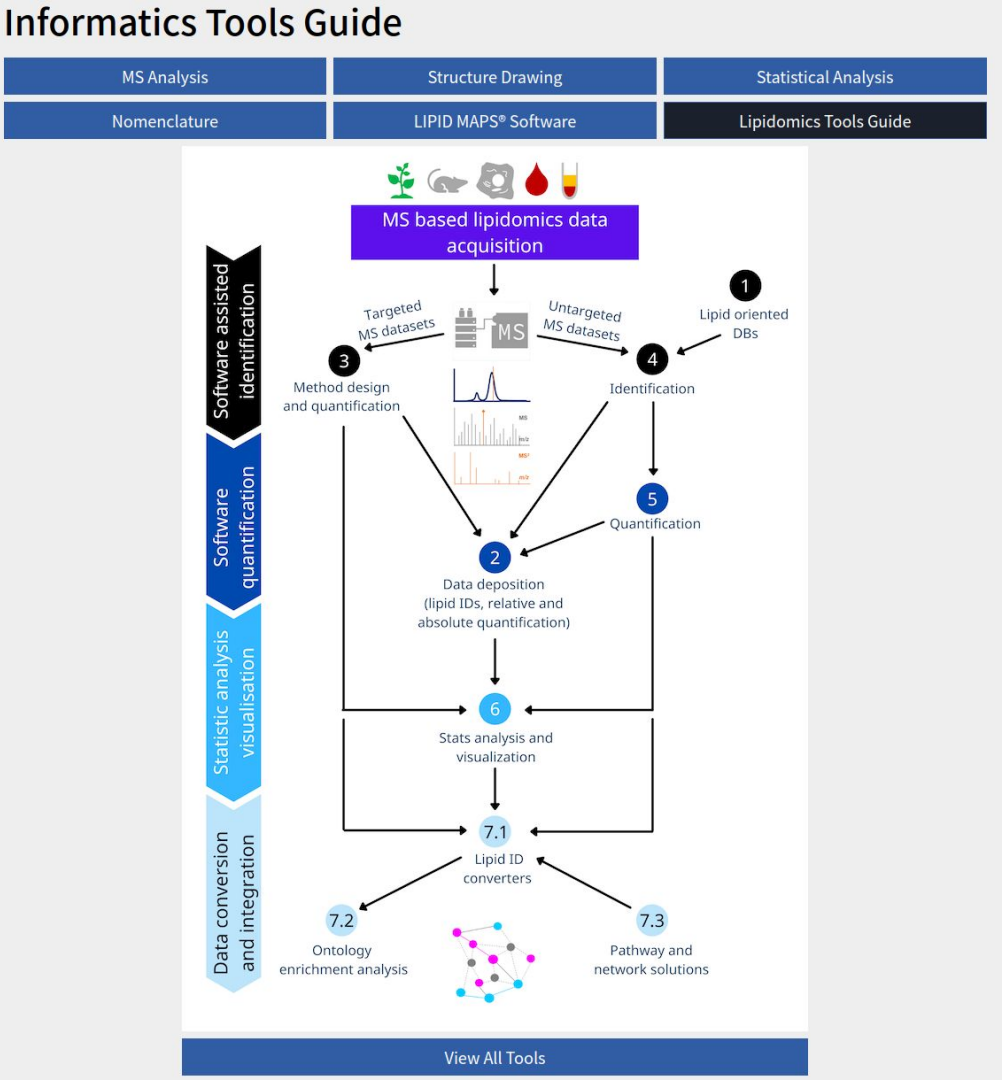
Adapted from: Momeni, Zahra, et al. "A survey on single and multi omics data mining methods in cancer data classification." Journal of Biomedical Informatics 107 (2020): 103466. DOI: 10.1016/j.jbi.2020.103466



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	

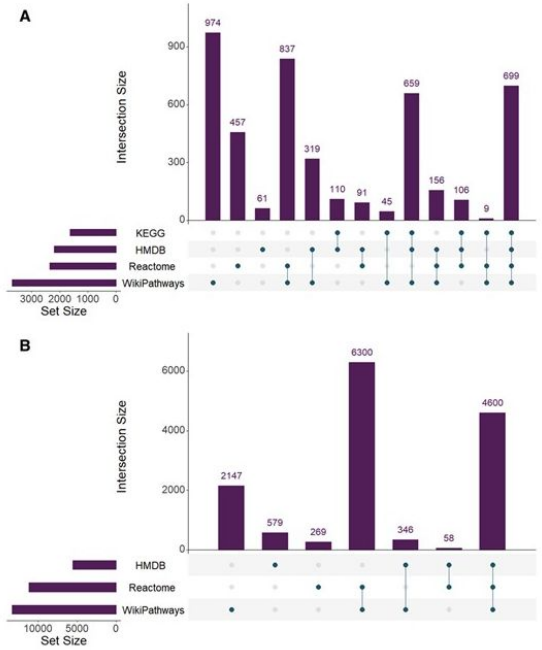
a The number in parentheses represents the difference in numbers compared to the previous RaMP version (1.1.0).

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

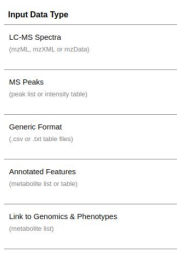
A					
	Total[a]	**HMDB v5.0**	**KEGG (from HMDB 5.0)**	**Reactome v81**	**WikiPathways v20220710**
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway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

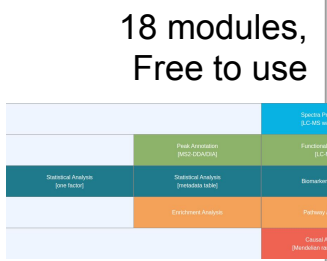
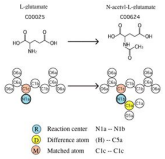
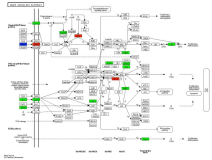
A					
	Total[a]	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775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

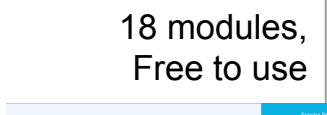
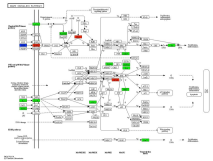
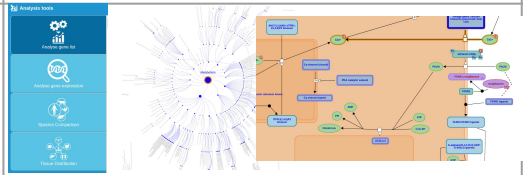
A					
	Total[a]	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775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

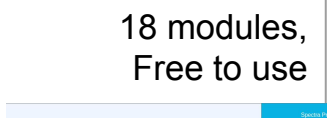
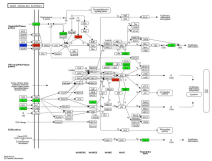
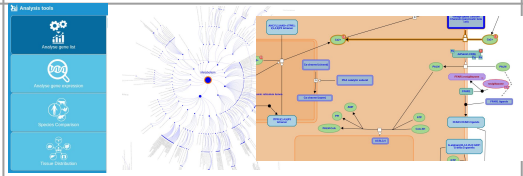
A					
	Total[a]	HMDB v5.0	KEGG (from HMDB 5.0)	Reactome v81	WikiPathways v20220710
# Distinct metabolites	256 086 (+142 361)	216 683	5898	2355	3695
# Distinct genes/enzymes	15 827 (+410)	7111	–	11 227	13 393
# Distinct pathways	53 831 (+2035)	49 613	363	2583	1272
#Metabolite-pathway mappings	412 775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

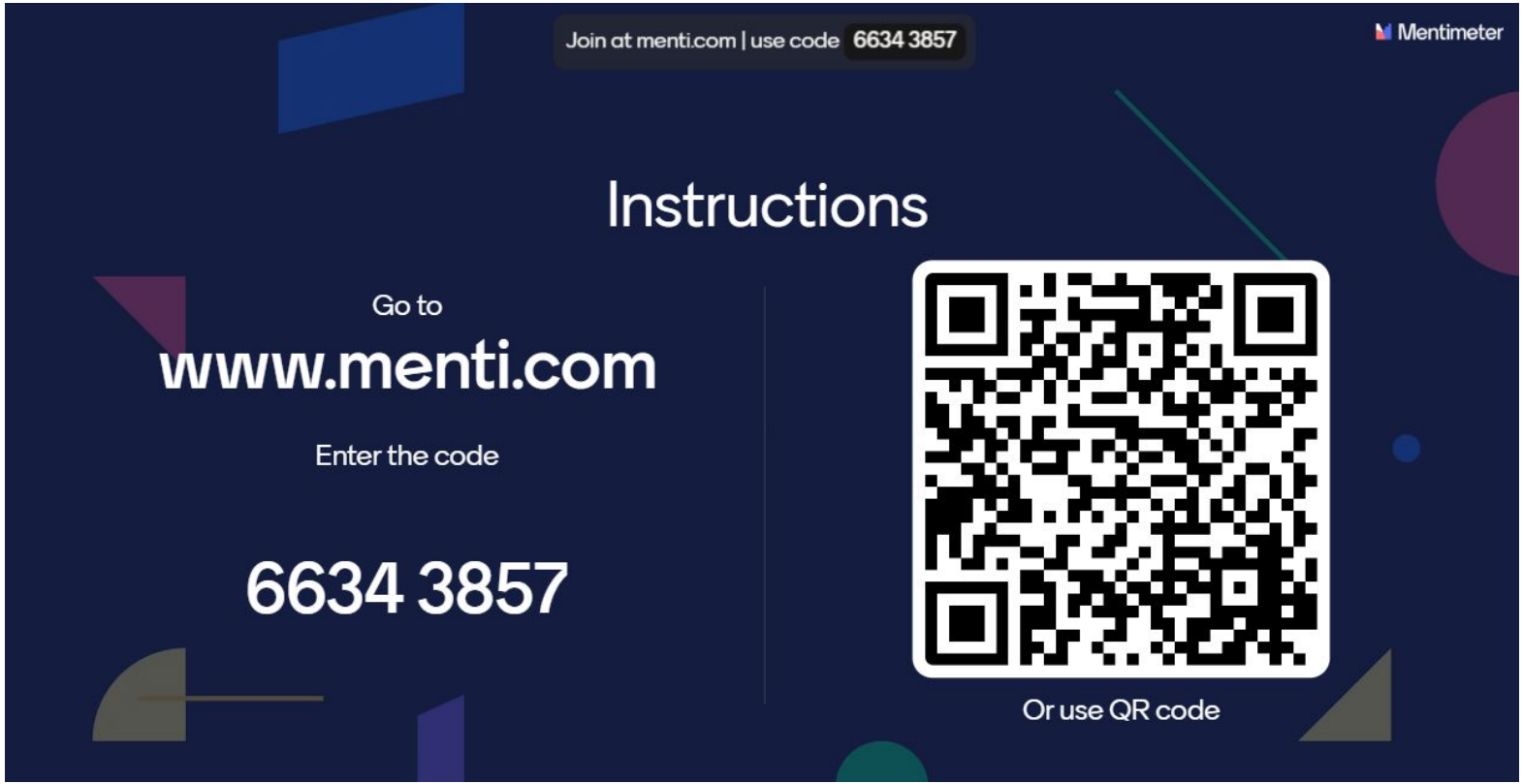
b Distinct InChIKeys.

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).

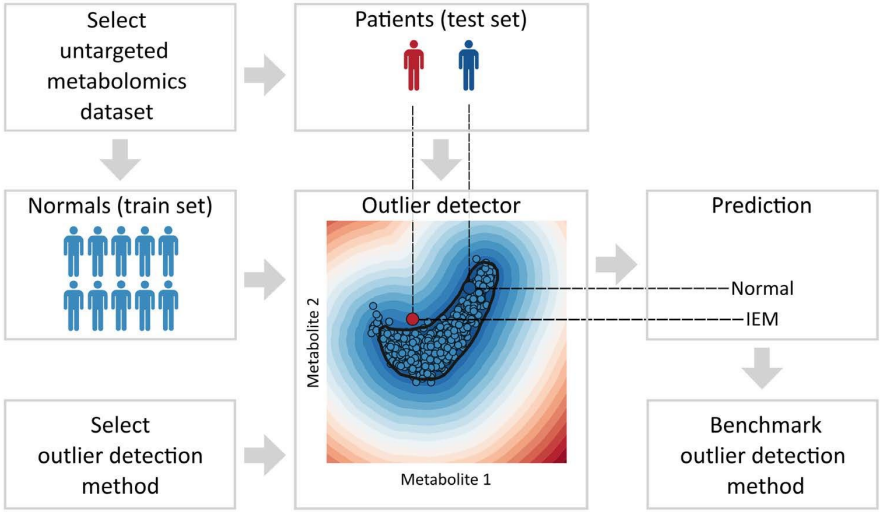


Figure obtained from:
Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." Metabolites 13.1 (2023): 97.
https://doi.org/10.3390/metabo13010097

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?

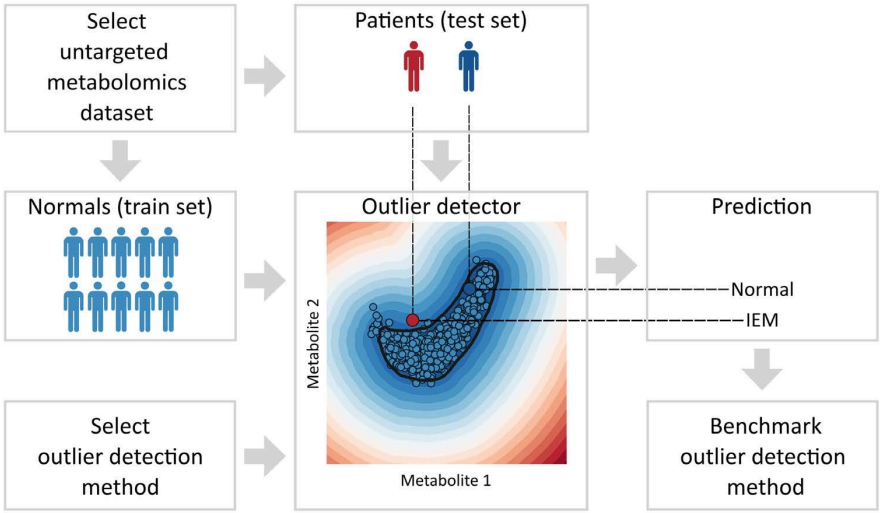
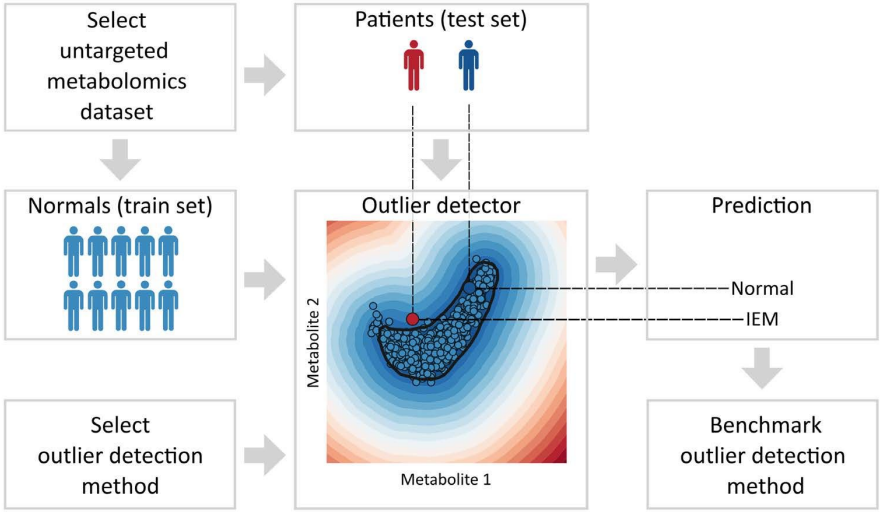


Figure obtained from:
Bongaerts, Michiel, et al. "Benchmarking Outlier Detection Methods for Detecting IEM Patients in Untargeted Metabolomics Data." Metabolites 13.1 (2023): 97.
https://doi.org/10.3390/metabo13010097

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.



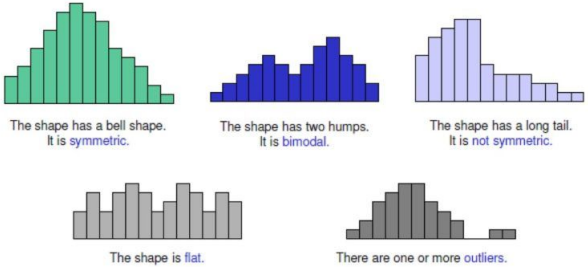
Which one to pick?

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

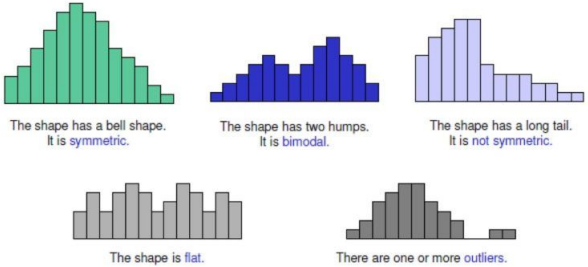
Data processing steps: distribution



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:

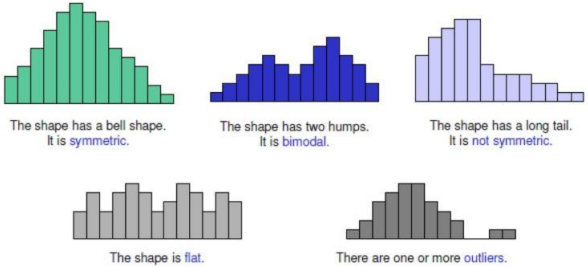
- Normal Distribution (Gaussian Distribution)
- Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Log-Normal Distribution
- …

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
- …

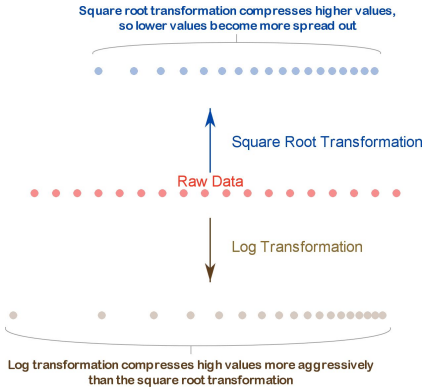
But, which one is it?

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?

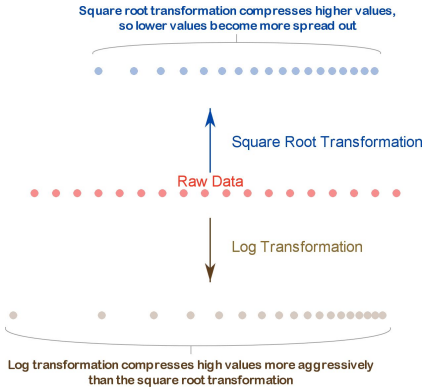
Data processing steps: transformation



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
-

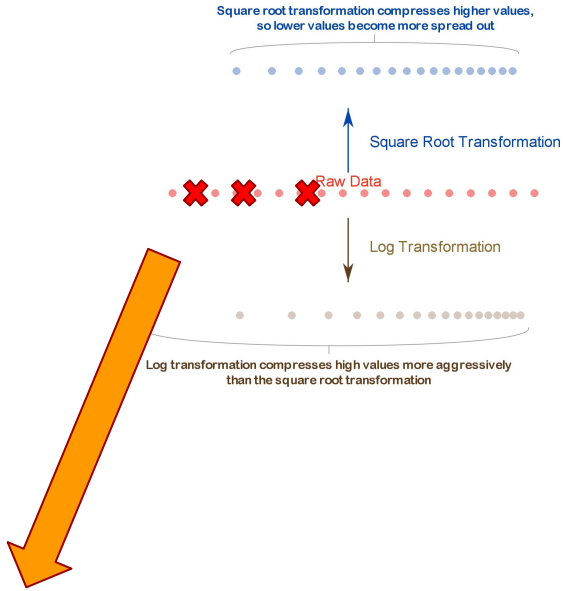
General rule of thumb:

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

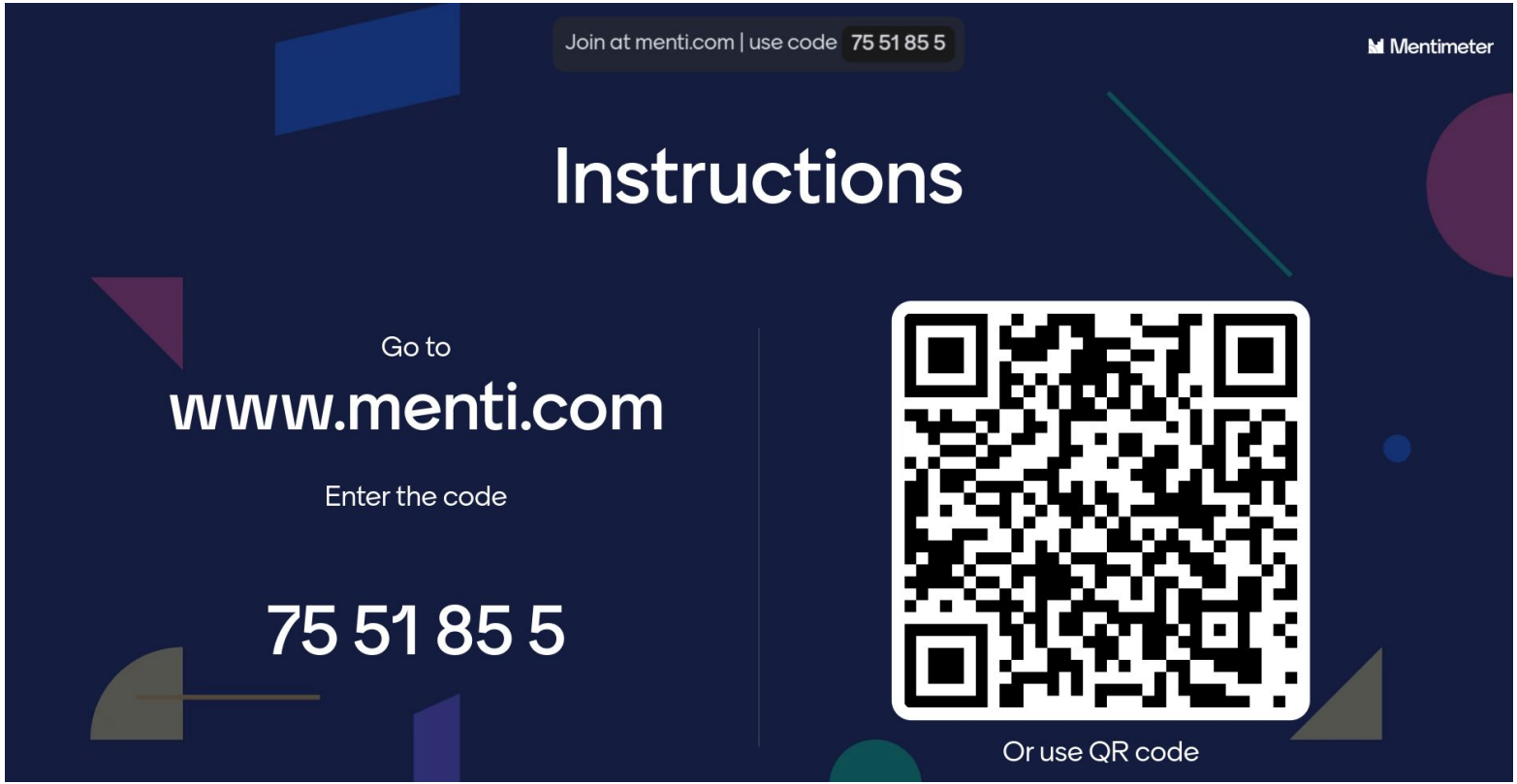
Highlighted all the missing values in the dataset

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

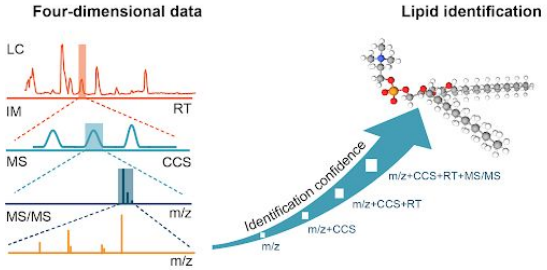
Highlighted only the standard missing values in the dataset

- 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:

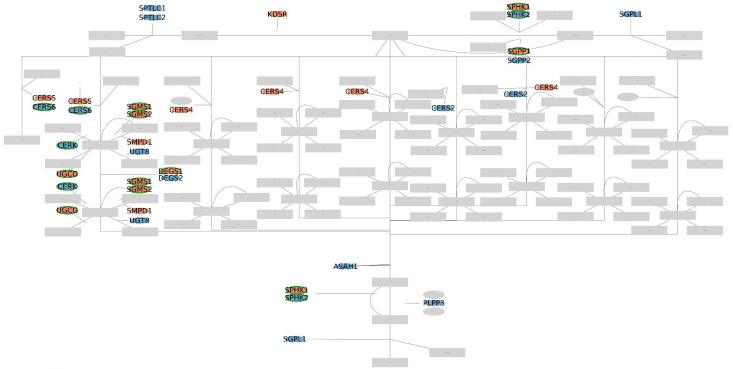
- Identify if there are duplicates Names/ IDs
- Review how these duplicates came part of the dataset
- Remove redundant rows
- Merge duplicates for multiple instances of same entity (average of values)
- Retain unique duplicates from entry errors or inconsistencies (label them)

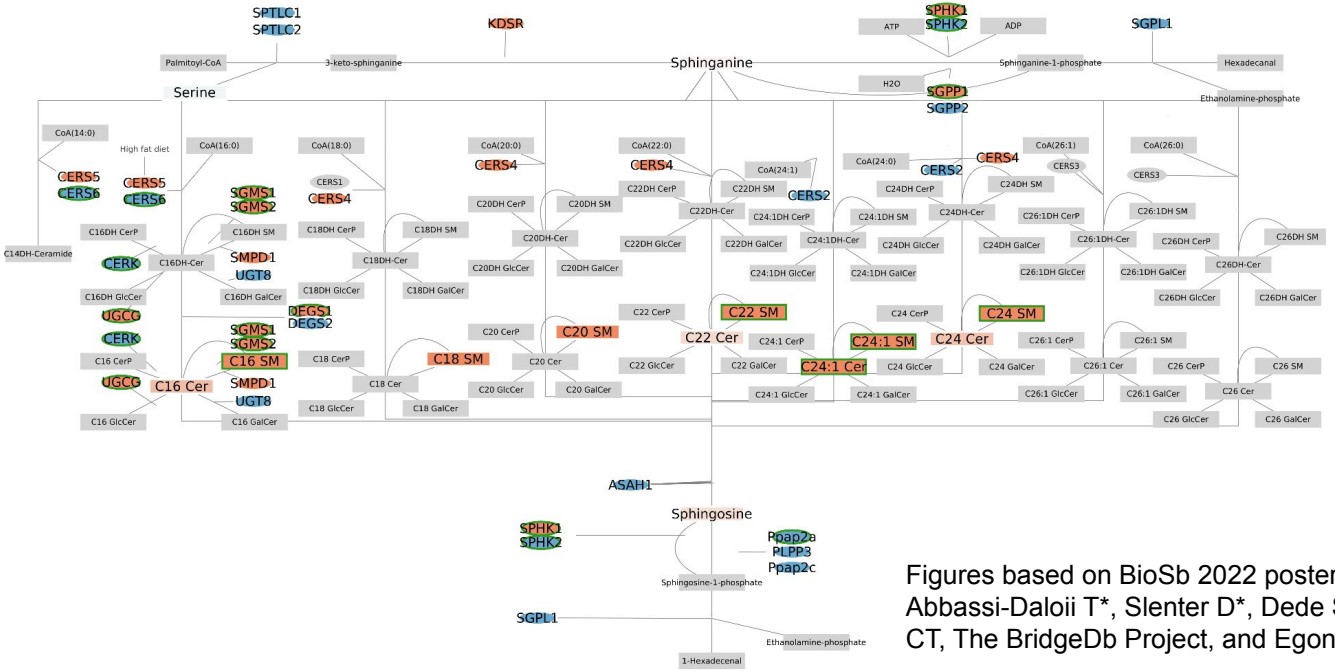
- Don’t forget about secondary IDs…

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

14:00-16:30

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides

Hands-on session - setting up laptop

14:00-16:30

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides

First; does everyone have a GitHub Account?

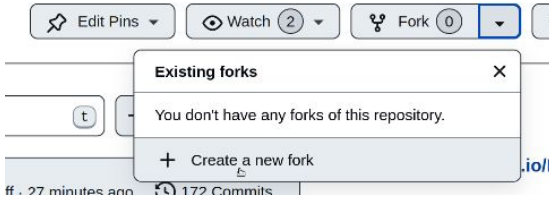
https://docs.github.com/en/desktop/installing-and-authenticating-to-github-desktop/setting-up-github-desktop

Hands-on session - setting up laptop

14:00-16:30

Requirements: R, Rstudio, GitHub Desktop

Live Demo/Slides



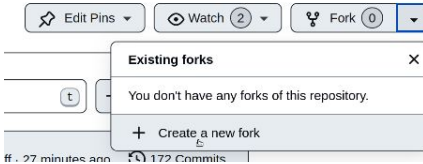
Second, find the repository and make a FORK

https://github.com/DeniseSI22/PETcourseMetabolomics

Hands-on session - setting up laptop

Second, find the repository and make a FORK

https://github.com/DeniseSI22/PETcourseMetabolomics



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 (*).

Your Github user name

Owner * Choose an owner / **Repository name** * PETcourseMetabolomics

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

Description (optional)

GitHub repository for the analysis of metabolomics data

☑ Copy the `main` branch only

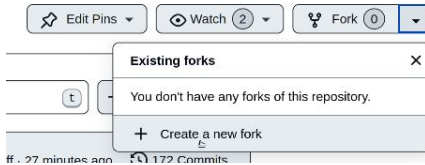
Contribute back to DeniseSI22/PETcourseMetabolomics by adding your own branch. Learn more.

Create fork

Hands-on session - setting up laptop

Second, find the repository and make a FORK

https://github.com/DeniseSI22/PETcourseMetabolomics



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 (*).

Your Github user name

Owner * Choose an owner / **Repository name** * PETcourseMetabolomics

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

Description (optional)

GitHub repository for the analysis of metabolomics data

Copy the `main` branch only

Contribute back to DeniseSI22/PETcourseMetabolomics by adding your own branch. Learn more.

Create fork

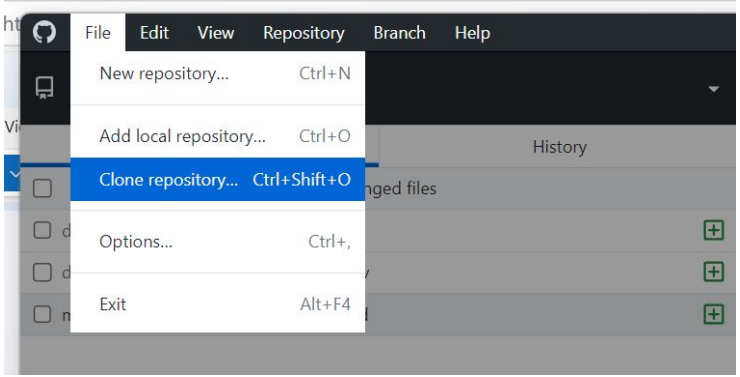
Now you are allowed to make changes to the available content!

Hands-on session - GitHub Desktop

- Select File/Clone repository

- In the pop-up menu, select GitHub.com (first option from the top), and add:

  username/PETcourseMetabolomics



Hands-on session - GitHub Desktop

- Select the folder where you want to store the repository under:

Local path

- And click Clone