

Quelques exemples avec des données multi-omiques et multi-tissus en néonatalité porcine pour illustrer les différentes fonctionnalités d'Asterics.

Élise Maigné, Céline Noirot, Jérôme Mariette, Yaa Adu Kesewaah, Sébastien Déjean, Camille Guilmineau, Julien Henry, Arielle Krebs, Laurence Liaubet, Fanny Mathevet, Hyphen-Stat, Christine Gaspin, Nathalie Vialaneix









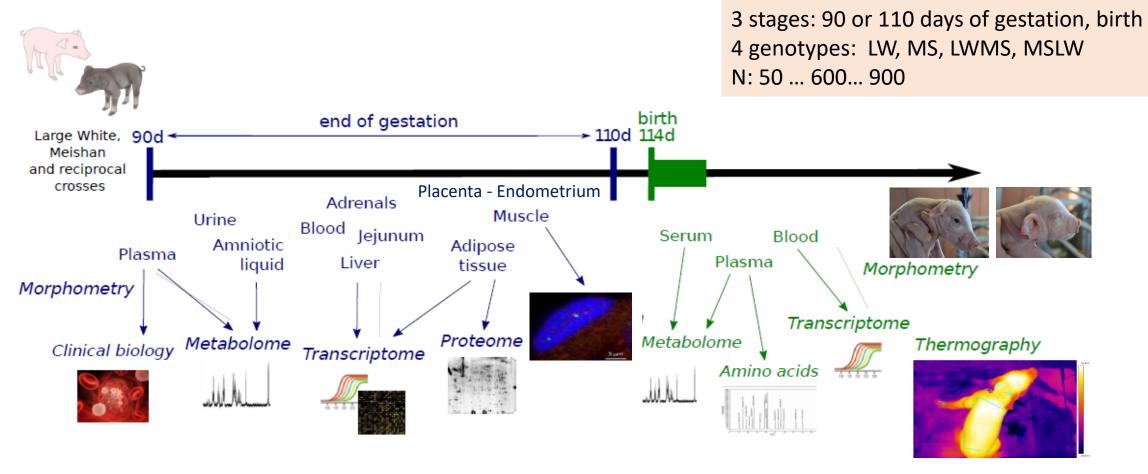








# Context: neonatal maturity and birth survival



Voillet et al, 2014, 2018; Yao et al.2017; Gondret et al, 2018; Marti-Marimon et al, 2018; Lefort et al, 2020,2021; Schmitt et al, 2021













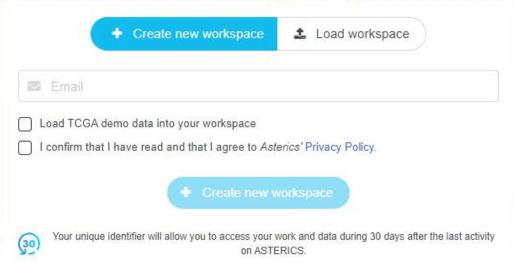




A tool for the exploration and integration of omics data

ASTERICS is an online tool designed to help you to perform your statistical and integrative analyses in an interactive and easy-to-use way.





https://asterics.miat.inrae.fr/

Project coordinators





### Funder



#### Partners







Dear user,

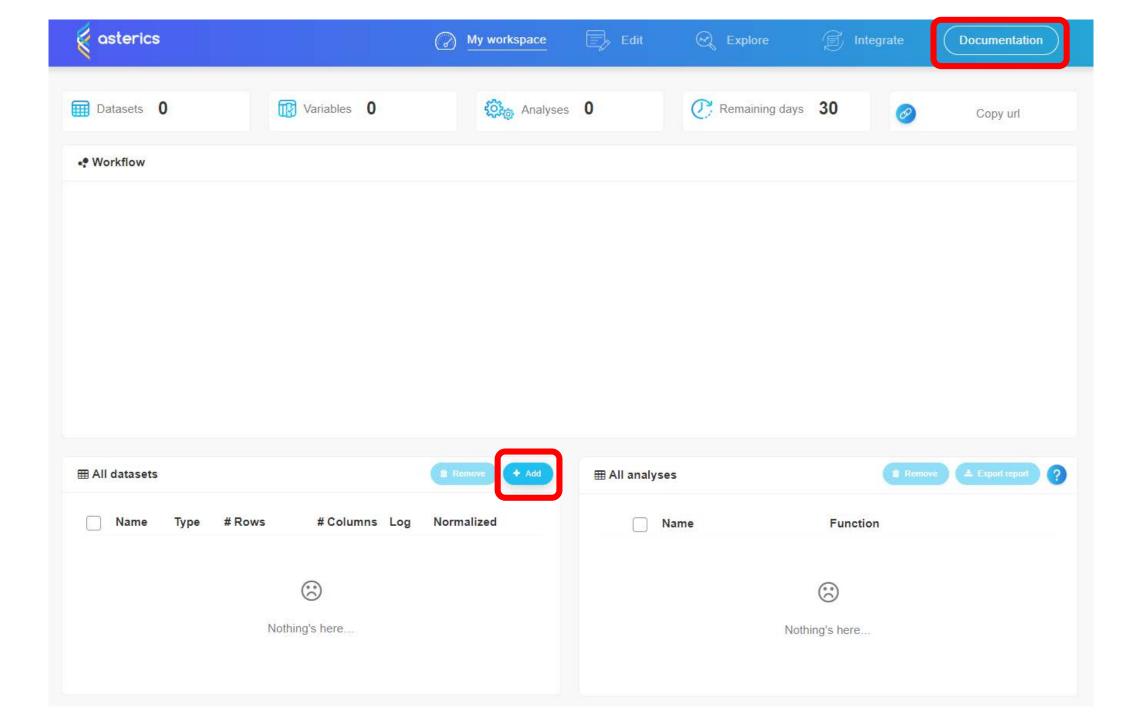
Your workspace has been successfully created at <a href="https://asterics.miat.inrae.fr/workspace/ba20fb6d-4196-45c0-999f-7ff6d522b109">https://asterics.miat.inrae.fr/workspace/ba20fb6d-4196-45c0-999f-7ff6d522b109</a>.

Thank you for using ASTERICS, The ASTERICS Team

Dear user,

your workspace <a href="https://asterics.miat.inrae.fr/78f3057f-ed70-42c8-ae5c-8d8b7be83484">https://asterics.miat.inrae.fr/78f3057f-ed70-42c8-ae5c-8d8b7be83484</a> will be deleted in 3 days.

thanks for using Asterics, The Asterics Team,



- 2.1 Create / load a workspace
- 2.2 Screen "My workspace"
- 2.3 Import your data
- 2.4 Retrieve an analysis
- 2.5 Export the data
- 2.6 Export report
- 3 What you can do with your data
- 3.1 Edit
- 3.2 Explore
- 3.3 Integrate

I Edit

- 4 Edit a dataset
- 5 Transformation
  - 5.1 Standard transformation
  - 5.2 Quantile normalization
  - 5.3 RNA-seq (and other count) d...
  - 5.4 Compositional data
- 5.5 Metagenomic data
- 5.6 Correcting batch effect with ...
- 6 Missing values
- 6.1 Explore missing values
- 6.2 Remove missing values
- 6.3 Impute missing values

II Explore

- 7 Explore variables
  - 7.1 Univariate
  - 7.2 Bivariate

#### **ASTERICS: User documentation**

Maigné Élise, Noirot Céline, Mariette Jérôme, Adu Kesewaah Yaa, Déjean Sébastien,

Guilmineau Camille, Henry Julien, Krebs Arielle, Liaubet Laurence, Mathevet Fanny,

Hyphen-Stat, Gaspin Christine, Vialaneix Nathalie

2022-08-17

#### Section 1 Introduction

This document is the user guide for the web tool ASTERICS. It is still a work-in-progress and mostly incomplete. Come back later for a better version. Thank you for your understanding.

Link to the application: http://asterics.miat.inrae.fr/

Support for ASTERICS can be obtained at asterics-tlse@inrae.fr.

Bugs can be reported here and suggestions can be made here.





# 13.2 Run PLS-DA

- 13.3 Explore individuals
- 13.4 Explore variables
- 13.5 Extract new data

#### 14 Multiple Factor Analysis (MFA)

- 14.1 Preprocessing
- 14.2 Run MFA
- 14.3 Explore individuals
- 14.4 Explore variables
- 14.5 Explore groups
- 14.6 Extract new data
- 15 Differential Analysis
- 15.1 Preprocessing
- 15.2 Multiple tests
- 15.3 Posthoc tests
- 15.4 Extract dataset

#### IV Case studies

#### 16 Case studies

- 16.1 Breast cancer
  - 16.1.1 Setup
  - 16.1.2 Edition
- 16.1.3 Principal Component A...
- 16.1.4 Hierarchical clustering
- 16.1.5 Data import
- 16.1.6 Normalization
- 16.1.7 Multiple Factor Analysi...
- 16.1.8 Partial Least Squares -...
- 16.1.9 Interpretation

#### References

### asterics

# Section 16 Case studies

#### 16.1 Breast cancer

This case study focuses on breast cancer, trying to link omics data to cancer subtypes. There are four breast cancer subtypes that somewhat represent sub-diseases, with different biological mechanisms. Their identification is particularly important when designing / selecting treatments. For this case study, we have mRNA and miRNA data on 988 individuals, as well as the associated subtypes. mRNA data is already normalized and is loaded in ASTERICS by default, whereas miRNA comes in the form of raw counts.

The data is available online (Vialaneix 2021).

#### 16.1.1 Setup

Let us open an empty session and load the TCGA demo data.

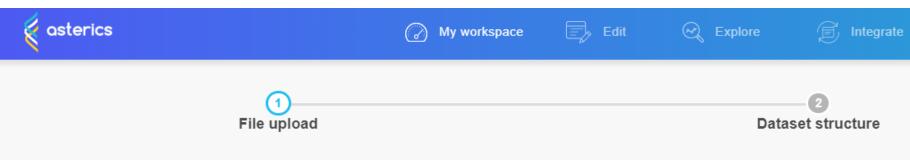


Since we will only use the mrna dataset, let us delete the other two.

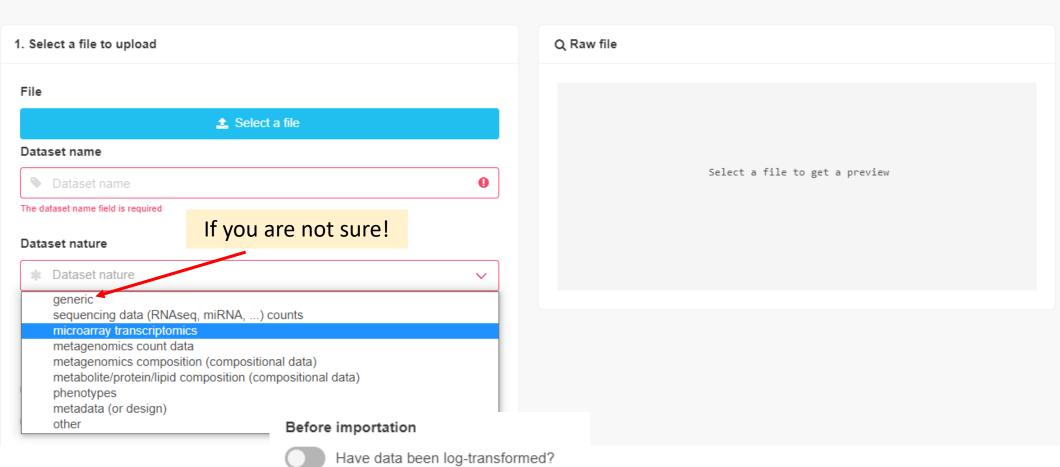


Here is mrna in the workspace:





Have data been normalized?



Documentation

#### 1. Select a file to upload

#### File

▲ TranscriptomeMuscle50.csv

#### Dataset name

TranscriptomeMuscle50

#### Dataset nature

microarray transcriptomics

If you don't know which nature to choose, use "generic".

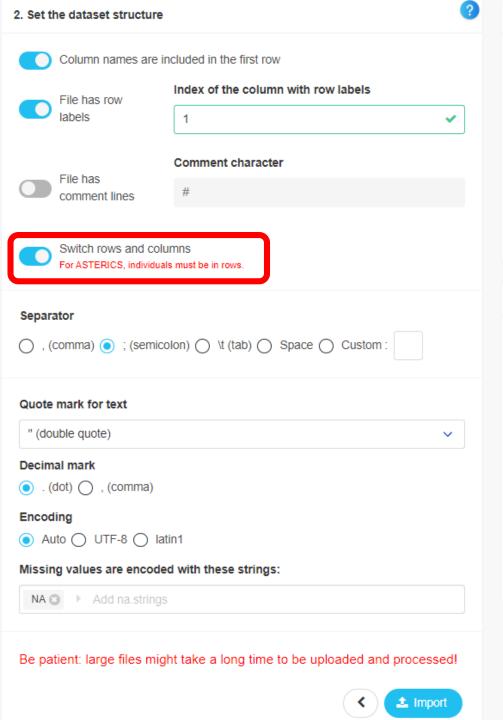
#### Before importation

- Have data been log-transformed?
- Have data been normalized?

 $\sim$ 

#### Q Raw file

;F405;F409;F411;F449;F455;F458;F469;F476;F481;F485;F498;F500;F524;F525;F53
A1CF\_52154;5.6803695954516;5.45610320624039;5.74484272573927;5.92019267767
A2M\_40313;5.50693239097091;5.76073759718855;5.5227941006065;5.663677602194
A2M\_10645;5.41017364325691;5.79477650004684;6.04237760287688;5.68692117450
A2ML1\_14639;5.4491082475774;5.36385961056552;5.17100665288747;5.6208513292
A4GALT\_23662;8.36976136324847;7.80328185134206;8.55421636844331;7.22995624
A4GALT\_9159;8.82297280814355;8.92285084244449;9.17229076620426;8.834796301
AAAS\_61230;5.66491543991508;5.70809268946484;5.85667371223367;6.9757878359
AAAS\_2894;5.89283524674721;5.6564212452558;5.66624190687793;6.090788421745
AAAS\_18735;7.74349641324072;7.70023645818195;7.43741001836278;8.7585108190
AAAS\_24237;6.41516988709249;6.40535808528583;6.82967565563798;7.0901431036



#### Q Raw file

;F476;F481;F560;F578;F688;F455;F485;F498;F738;F899;F900;F469;F557;F559;F67
A1CF\_52154;5.82599980707194;5.91662162096354;5.62770189995087;5.9748791179
A2M\_40313;5.20290966528126;5.71662080795284;5.73529430299502;6.27994060037
A2M\_10645;6.05043272137919;5.78608436449946;5.92397285829613;6.31840463378
A2ML1\_14639;5.09454305674891;4.89392730422894;5.06610396900029;5.109927232
A4GALT\_23662;7.44499238746051;7.88769803161222;7.75157714142366;7.85478800
A4GALT\_9159;8.80734586555401;8.74392533000205;8.55935004010872;8.291677940
AAAS\_61230;5.82599980707194;5.60366933006277;6.19088825000998;6.6908486193
AAAS\_2894;5.59207508123505;5.29881778254516;5.87420980899597;5.93550088274
AAAS\_18735;7.9656631095485;7.65139989082614;7.78415314320795;8.09455028881
AAAS\_24237:6.38767805826733:6.53943474153246:6.73107621612731:6.5142754375

Switch rows and columns!

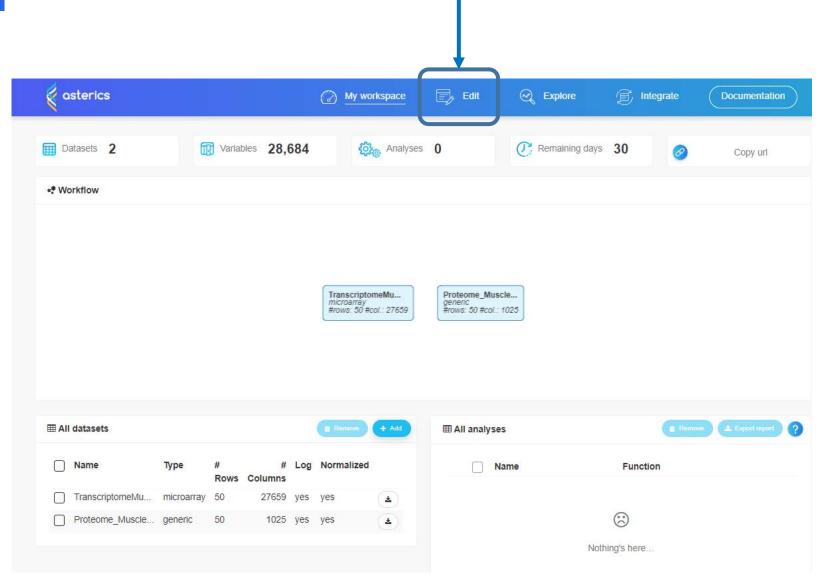
#### Q Dataset preview

	A1CF_52154	A2M_40313 numeric	A2M_10645	A2ML1_14639	A4GALT_23662	A4G
F476	5.826	5.2029	6.0504	5.0945	7.445	
F481	5.9166	5.7166	5.7861	4.8939	7.8877	
F560	5.6277	5.7353	5.924	5.0661	7.7516	
F578	5.9749	6.2799	6.3184	5.1099	7.8548	
F688	5.7889	6.434	6.5869	4.7516	8.0703	
F455	5.5397	5.5826	5.8344	5.648	7.4885	
F485	5.8868	6.1856	6.3453	5.661	7.7702	
F498	5.6354	5.81	5.7552	5.8236	7.6827	
F738	5.6221	5.1073	5.5711	5.6786	7.1388	
F899	5.5986	5.5457	5.8591	5.0093	7.3409	
F900	5.781	5.9127	6.1895	5.8663	7.7117	
F469	6.1845	5.7959	5.7232	5.5926	8.0633	
FFF7	E 0740	E 000E	C 0E 4	E EC20	7.0404	<b>→</b>

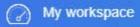
1 unique identifier per variable, so arrange the name to have the most relevant information!



Edit











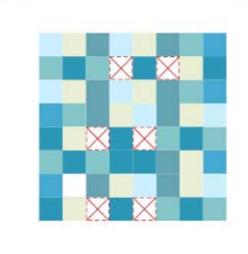


Documentation

# Let's edit data!

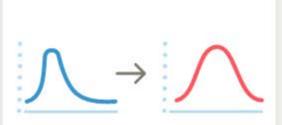






### Missing values

Explore, remove or impute missing values on a dataset.



#### Normalize dataset

Correct technical biases and prepare your data for further analyses.

#### **Dataset edition**

1.	Select	a	dataset	and	an	action	

Dataset

■ Infos\_50 ▼

#### Actions

- Transpose
- Change dataset nature
- Change variable (column) types
- Set individual (row) names
- Subset individuals (rows)
- Subset variables (columns)
- Rename categories
- Reorder categories

#### History

No history to display

Select an action.

#### ■ Dataset "Infos\_50"

# rows	# col.	# missing	% missing	# numeric	# cat.	# logic	# others
50	32	0	0	27	5	0	0

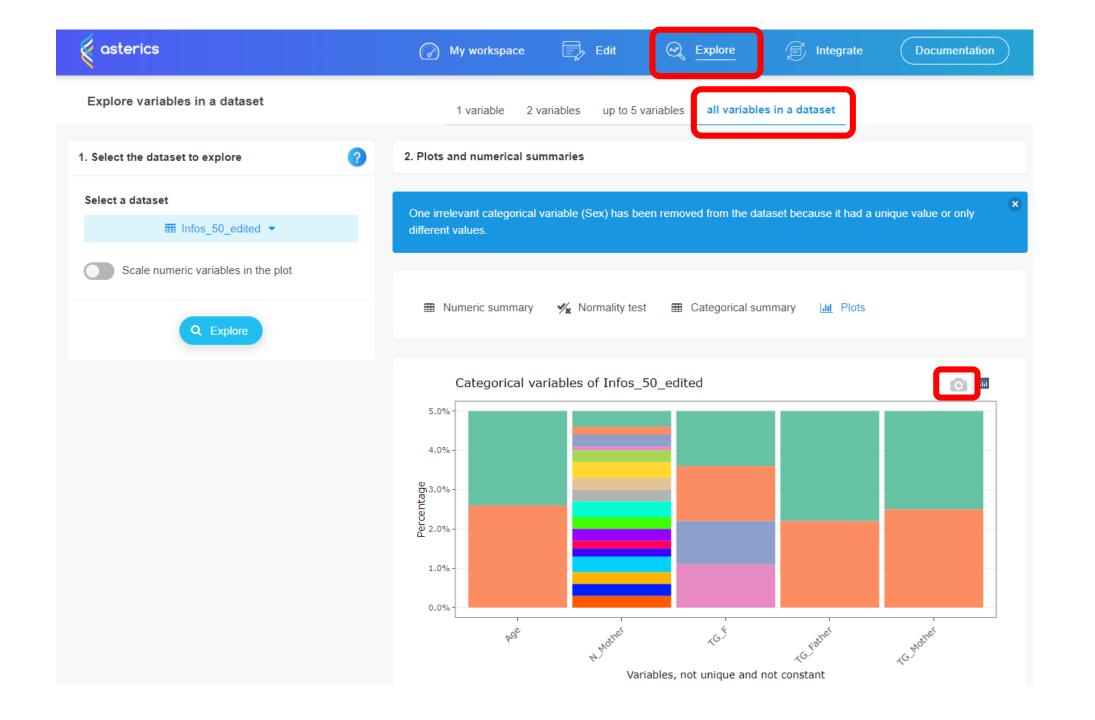
	Sex categorical 0.00%	N_Mother numeric 0.00%	Age categorical 0.00%	TG_Mother categorical 0.00%	TG_Father categorical 0.00%	TG_F categorical 0.00%	Weight numeric 0.00%
F449	M	6	90d	LW	LW	LWLW	733
F458	M	6	90d	LW	LW	LWLW	782
F500	M	8	90d	LW	LW	LWLW	597
F736	M	23	90d	LW	LW	LWLW	597
F744	M	23	90d	LW	LW	LWLW	784
F894	M	33	90d	LW	LW	LWLW	813
F895	M	33	90d	LW	LW	LWLW	389
F455	M	6	90d	LW	MS	MSLW	717
F485	M	8	90d	LW	MS	MSLW	754
F498	M	8	90d	LW	MS	MSLW	704
F738	M	23	90d	LW	MS	MSLW	522
F899	M	33	90d	LW	MS	MSLW	599



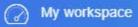


















Documentation

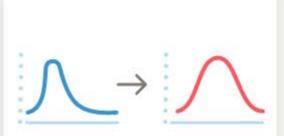
# Let's edit data!





# Missing values

Explore, remove or impute missing values on a dataset.



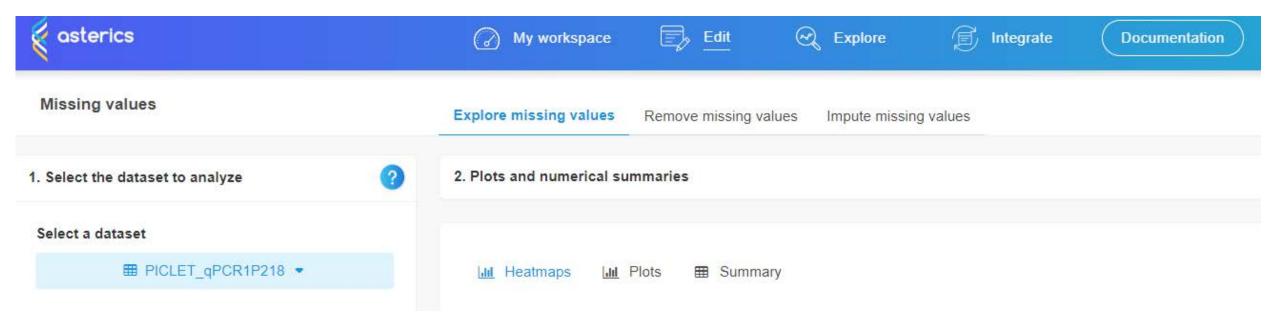
#### Normalize dataset

Correct technical biases and prepare your data for further analyses.

Which decisions should I take to handle missing values?

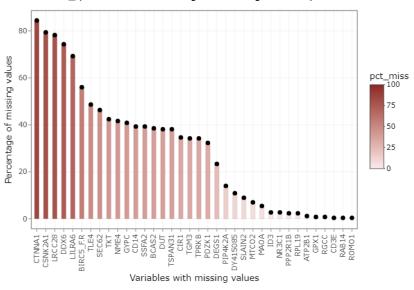


	ATP2B1 numeric 1.17%	BCAS2 numeric 38.52%	BIRC5_F.R numeric 56.03%	CD14 numeric \$5.30%	CD3E numeric 0.39%	CIR1 numeric 34.63%	CSNK2A1 numeric 79.38%	
P1	0.0559	0.0623	NA	0.1167	0.0191	0.0132	NA	N/
P2	0.0521	0.1207	NA	0.042	0.0828	0.0399	NA	N/
P3	0.153	0.0737	NA	0.0761	0.0773	NA	0.0826	N/
P4	0.0644	0.1384	NA	0.2347	0.0495	0.0555	NA	N/
P5	0.2821	0.246	NA	0.3472	0.2379	0.0441	NA	N/
P6	0.1404	0.1411	0.1522	0.0928	0.0943	NA	NA	N/
P7	0.5945	NA	0.1526	NA	0.2019	0.3797	NA	N/
P8	0.1211	0.1025	0.088	0.1648	0.141	NA	NA	NA

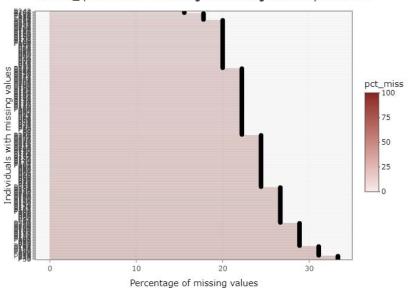


qPCR-1: Heatmap of missing values Missing values Pct missing 50

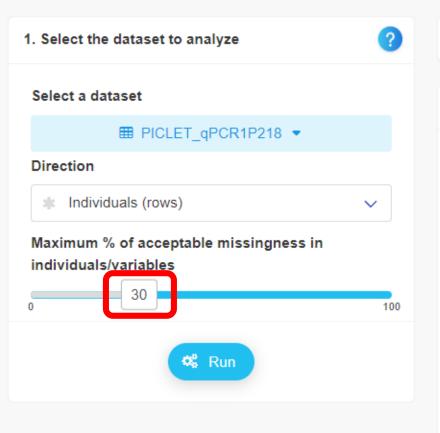
PICLET\_qPCR1P218: Percentage of missing values by variable



PICLET\_qPCR1P218: Percentage of missing values by individual



#### Missing values



Explore missing values

Remove missing values

Impute missing values

#### 2. Plots and numerical summaries

#### Purpose of Missing values:

#### Handle missing values in your datasets

- · Explore the distribution of missing values in your dataset
- · Impute missing values with PCA, k-means or by zero.
- OR remove individuals / variables with the largest proportions of missing values.

#### How to set options?

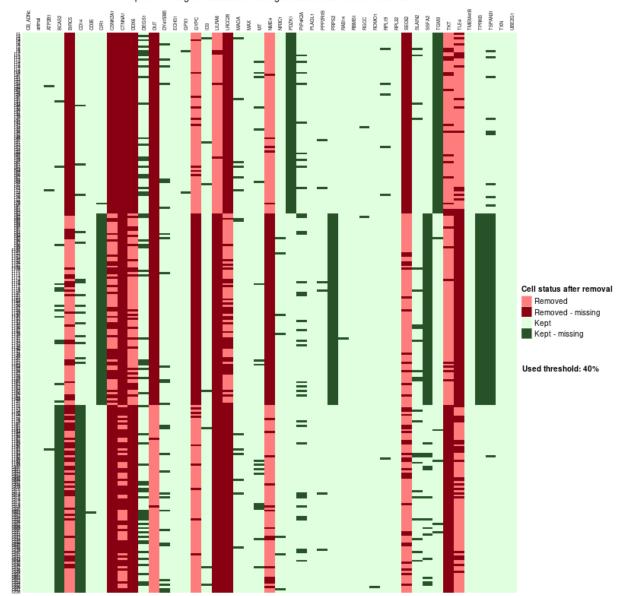
- . "Direction" is used to choose to remove either entire individuals (rows) or entire variables (columns).
- The chosen percentage corresponds to the maximum acceptable missingness. For instance, if 30% is chosen, it means that all individuals or variables with more than 30% of missing values will be removed.

Check help



for further advice

qPCR-1: original dataset showingvariablesthat will be removed

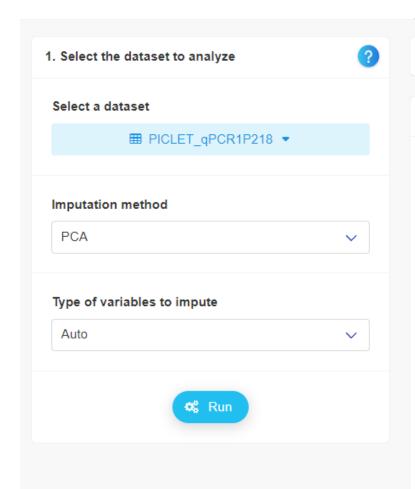








#### Missing values



Explore missing values Remove missing values

Impute missing values

#### 2. Plots and numerical summaries

#### Purpose of Missing values:

#### Handle missing values in your datasets

- . Explore the distribution of missing values in your dataset
- . Impute missing values with PCA, k-means or by zero.
- OR remove individuals / variables with the largest proportions of missing values.

#### How to set options?

Imputation method can be chosen according to what you know on the missing values and subsequent analyses:

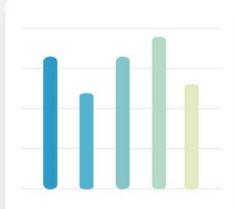
- **imputation by zeros** is dedicated to cases where missing values are due to a measurements below the detection threshold. It is a very basic approach to address this problem but certainly not the best;
- **imputation by PCA** is well designed when you want to use PCA, MFA, or PLS (for instance) afterwards because it best preserves the projection of your individuals on PC axes;
- **imputation by k-nearest neighbors** is based on the idea that two individuals that are similar on observed values also have similar values for unobserved variables. It best preserves the distances between individuals and is well adapted prior clustering.

The last two methods are only valid when data are missing at random.

In addition, for PCA and KNN you can choose to impute **only certain types of variables** (only numerical or only categorical variables). Setting this option to "Auto" imputes only variables of the most present type while "Mixed" imputes both numerical and categorical variables.

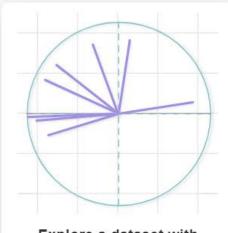
# Let's explore a dataset!





#### Explore variables in a dataset

Obtain numerical summaries and plots for a few variables.



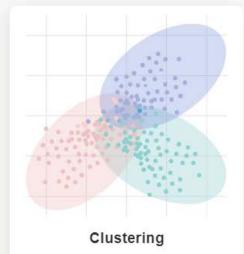
#### Explore a dataset with PCA

Perform Principal Component Analysis on a dataset.

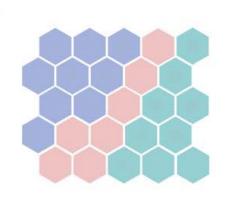


### Explore a dataset with a heatmap

Obtain the heatmap of a dataset.

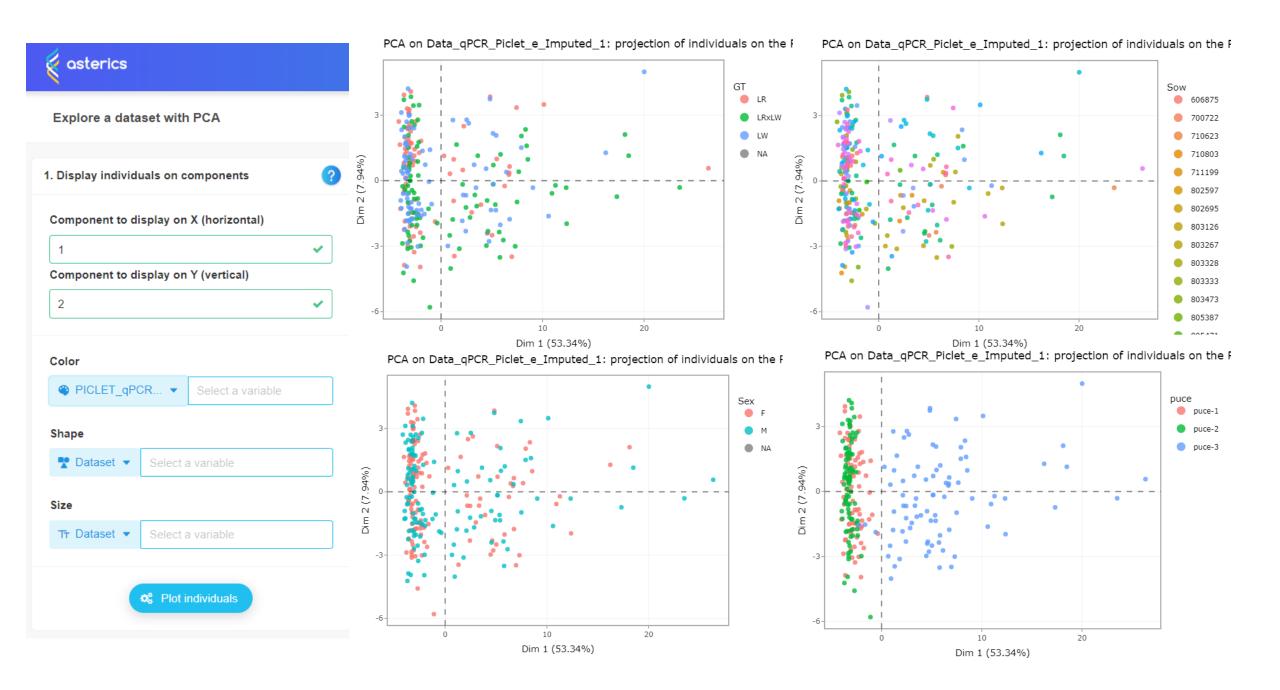


Cluster the individuals of a dataset.



#### Self-Organizing Map

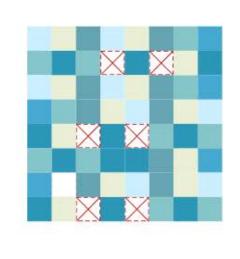
Use SOM as a clustering and visualization method.



# Let's edit data!

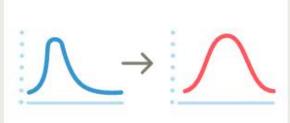






### Missing values

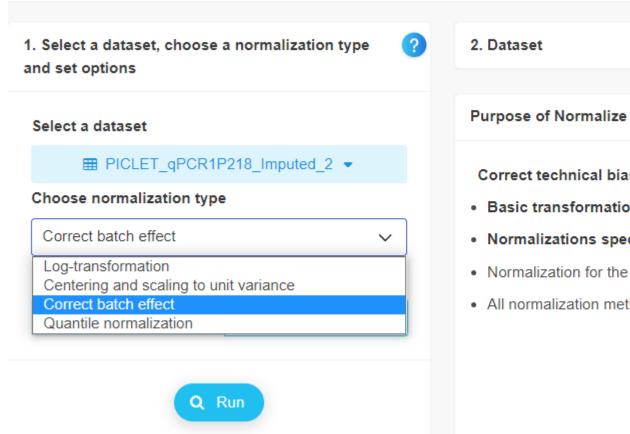
Explore, remove or impute missing values on a dataset.



### Normalize dataset

Correct technical biases and prepare your data for further analyses.

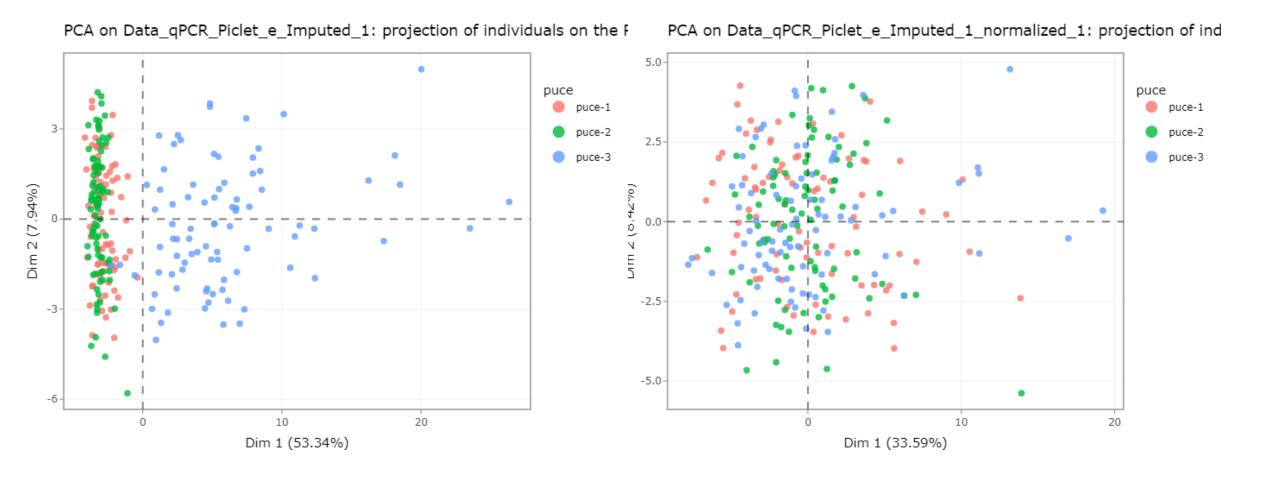
#### Normalize dataset

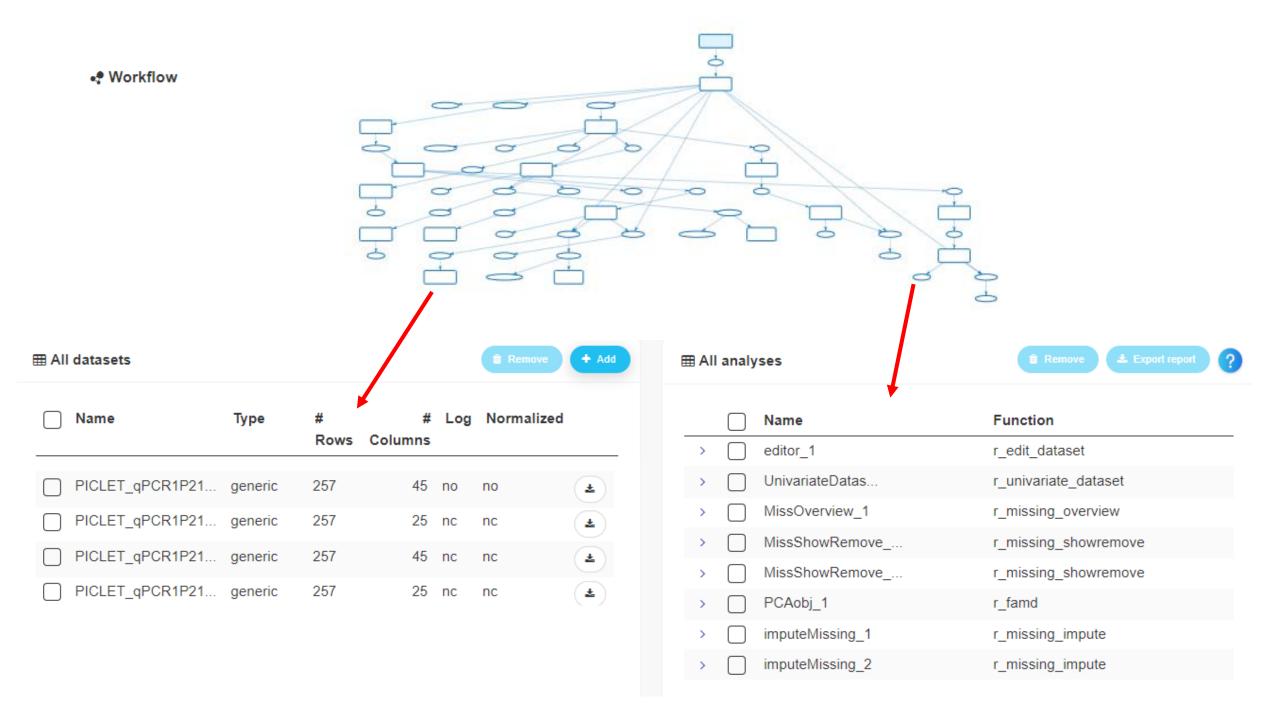


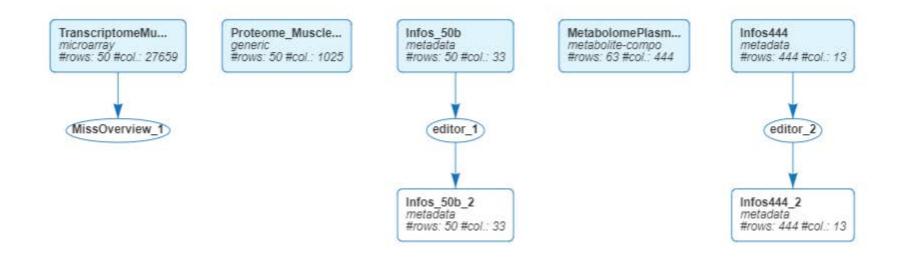
Purpose of Normalize dataset:

Correct technical biases and prepare your datasets for further analyses

- Basic transformations (log, ...).
- Normalizations specific of certain data types (RNA-seq, ...).
- · Normalization for the correction of an explicit batch effect.
- · All normalization methods come with diagnostic plots.

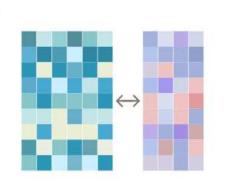






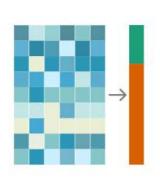
# Let's integrate data!





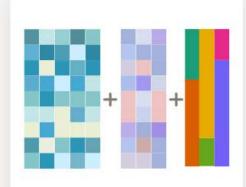
### Integrate two datasets with PLS

Performs Partial Least Squares analysis on two datasets.



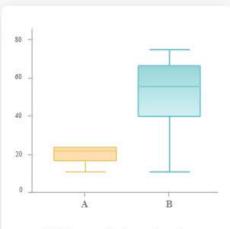
#### Integrate two datasets with PLS-DA

Performs Partial Least Squares analysis on two datasets.



#### Integrate datasets with MFA

Perform Multiple Factor Analysis on several datasets.



#### Differential analysis

Perform differential analysis for all numeric variables of a dataset.



**Datasets** 



My workspace





Integrate

**Documentation** 



Preprocessing

Run PLS

Explore individuals

Explore variables

Extract new data





2. Plots and summaries



■ General information



Venn diagram





Proteome\_Muscle\_50 
Infos\_50\_edited 
Infos\_50\_edited

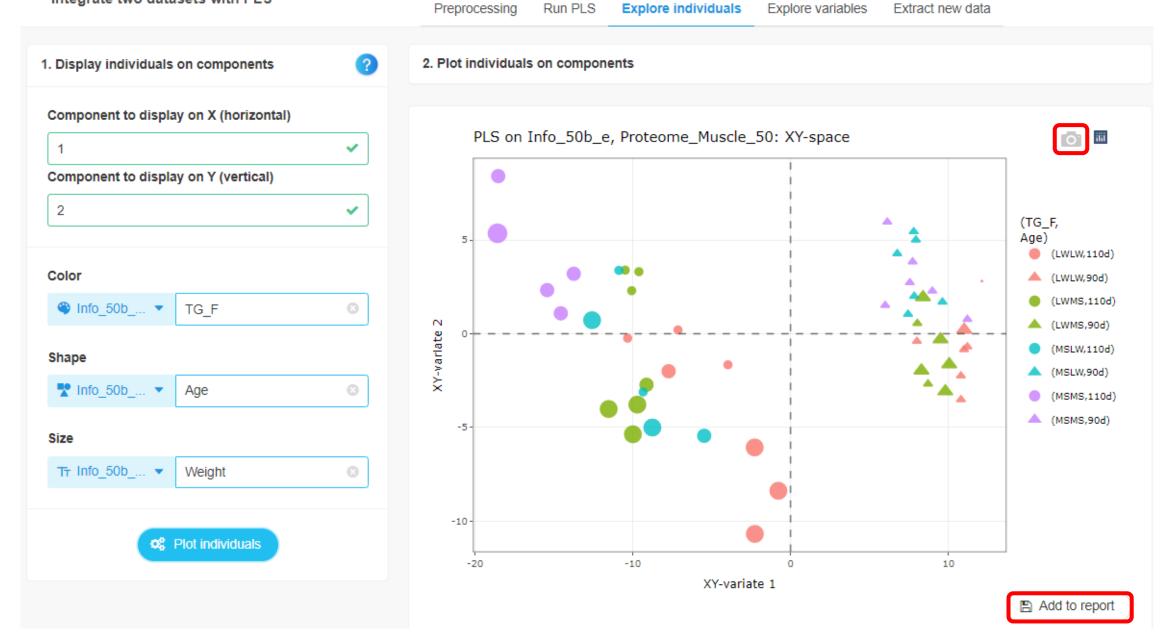
#### Description of datasets before filtering common individuals (rows)

Name	# rows	# col.	# missing	% missing	# numeric	# cat.	# logic	# others
Infos_50_edited	50	32	0	0	26	6	0	0
Proteome_Muscle	50	1025	0	0	1025	0	0	0

#### Description of datasets after filtering common individuals (rows)

Name	# rows	# col.	# missing	% missing	# numeric	# cat.	# logic	# others
Infos_50_edited	50	32	0	0	26	6	0	0
Proteome_Muscle	50	1025	0	0	1025	0	0	0





Documentation

Integrate two datasets with PLS

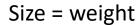
Preprocessing

Run PLS

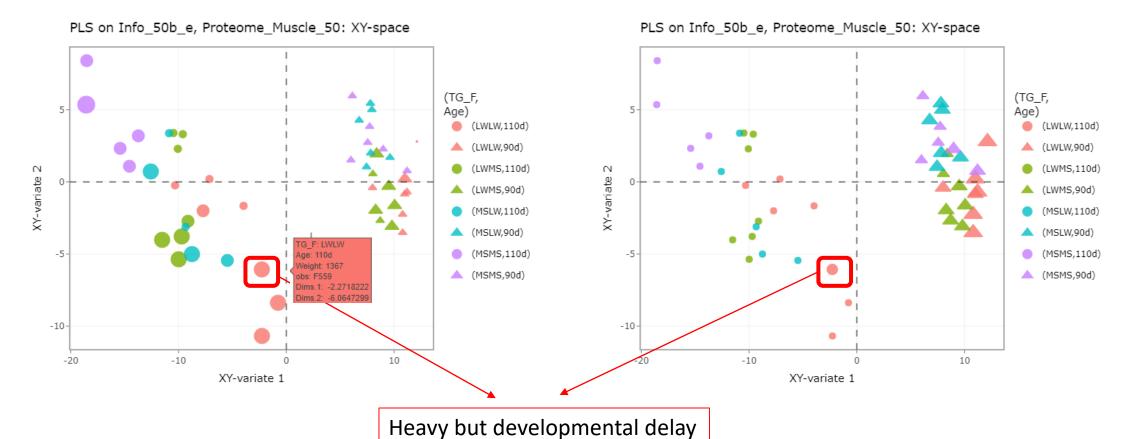
**Explore individuals** 

Edit

Explore variables



Size = MHC\_E





My workspace



**Explore** 



Documentation

(TG\_Father,

(LW,110d)

▲ (LW,90d)

(MS,90d)

(MS,110d)

Age)

10

Integrate two datasets with PLS

Preprocessing

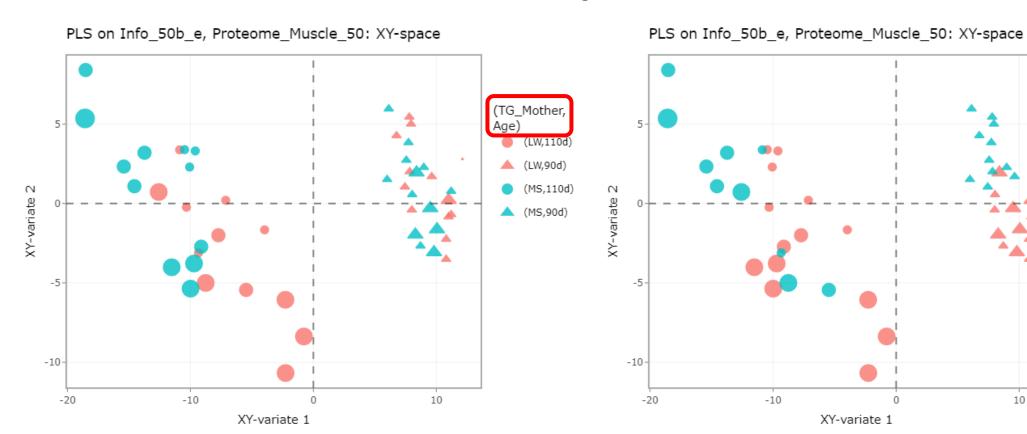
Run PLS

Explore individuals

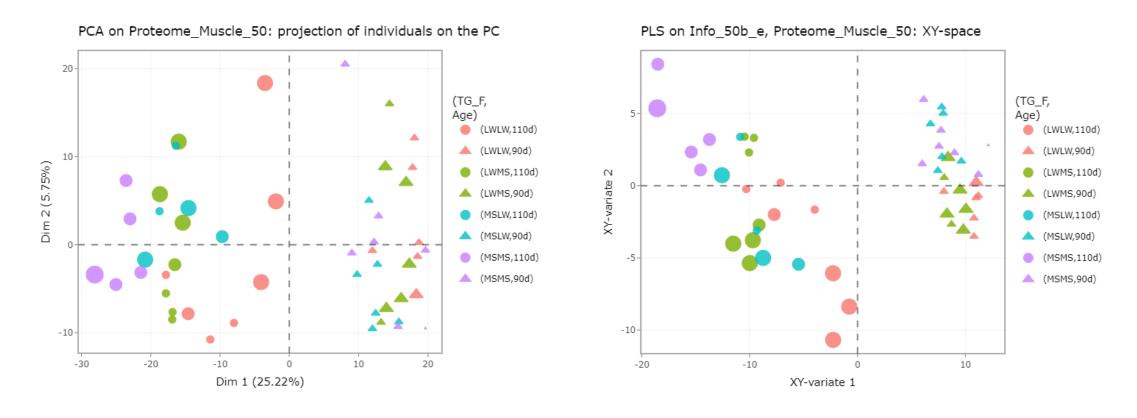
Explore variables

Extract new data

### Size = weight



### PCA vs PLS



Size = weight

### Integrate two datasets with PLS

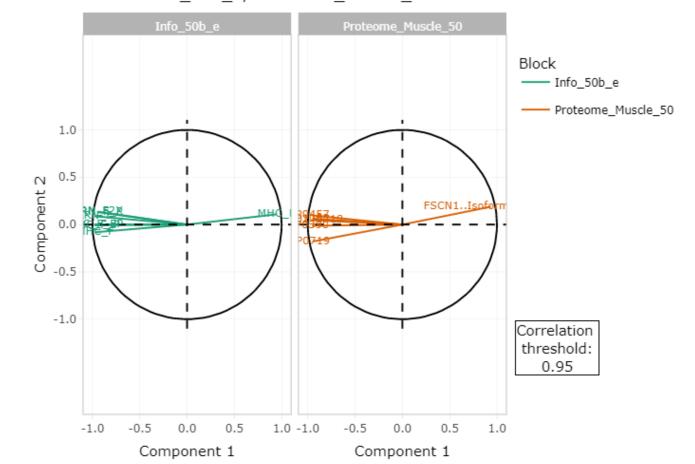
Preprocessing

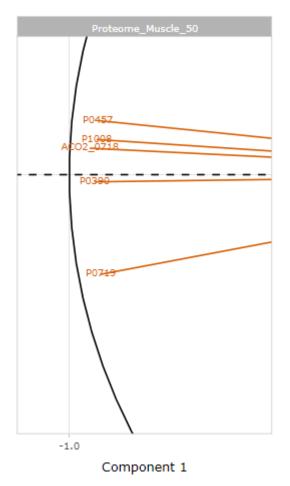
Run PLS

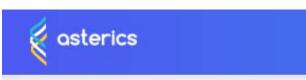
Explore individuals

Explore variables

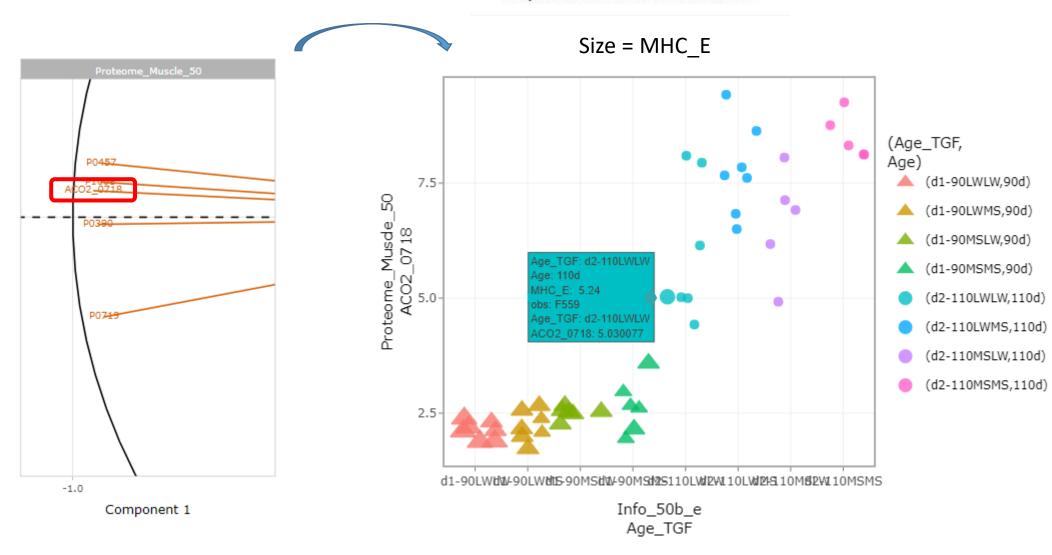
PLS on Info\_50b\_e, Proteome\_Muscle\_50: correlations of variables with







### Explore variables in a dataset





My workspace







Documentation

### Integrate two datasets with PLS

Component 1

Preprocessing

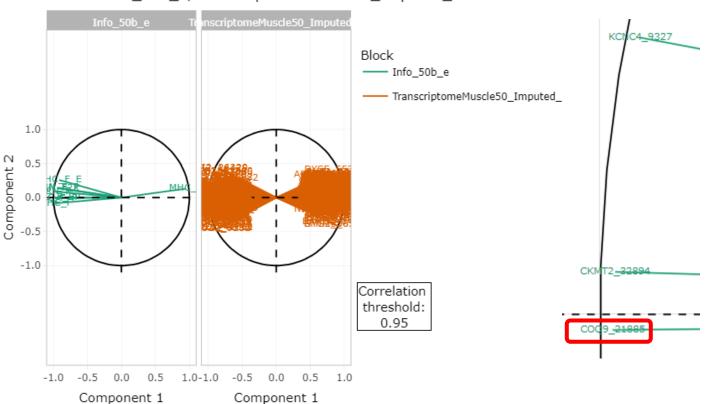
Run PLS

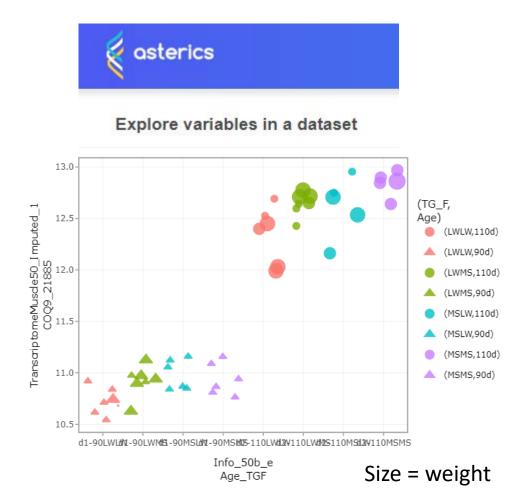
Explore individuals

Explore variables

Extract new data

#### PLS on Info\_50b\_e, TranscriptomeMuscle50\_Imputed\_1: correlations of

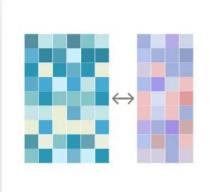






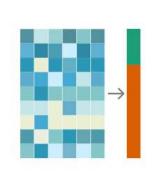
### Let's integrate data!





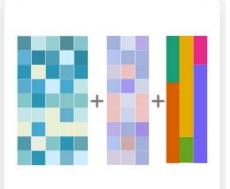
# Integrate two datasets with PLS

Performs Partial Least Squares analysis on two datasets.



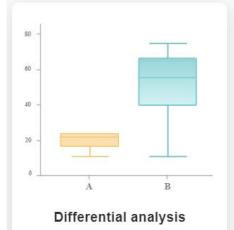
# Integrate two datasets with PLS-DA

Performs Partial Least Squares analysis on two datasets.



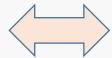
# Integrate datasets with MFA

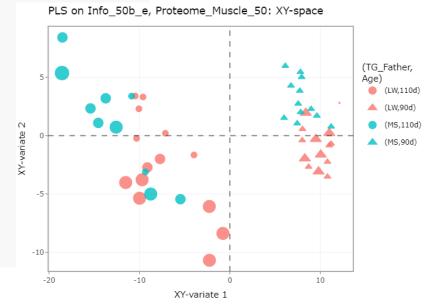
Perform Multiple Factor Analysis on several datasets.



Perform differential analysis for all numeric variables of a dataset.

It is easy to discriminate the age but what about the genotype?









My workspace







Documentation

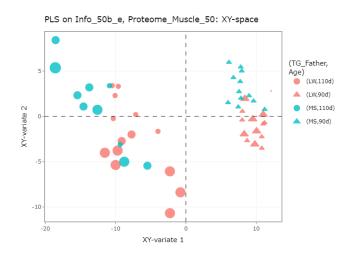
Integrate two datasets with PLS-DA

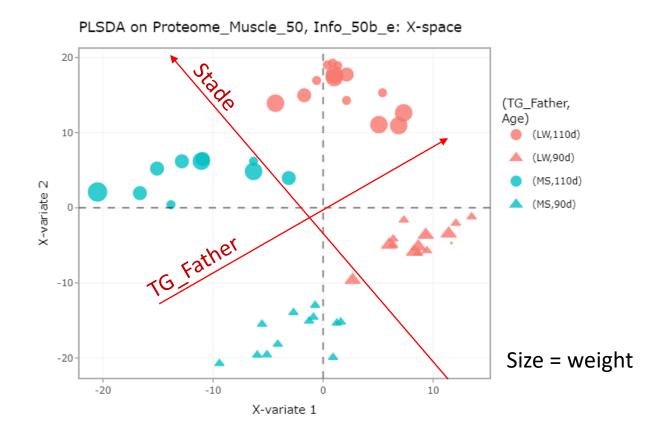
Preprocessing

Run PLS-DA

**Explore individuals** 

Explore variables







My workspace



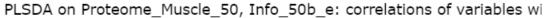
### Integrate two datasets with PLS-DA

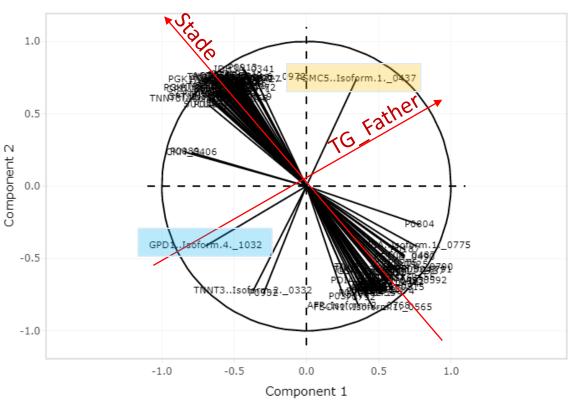
Preprocessing

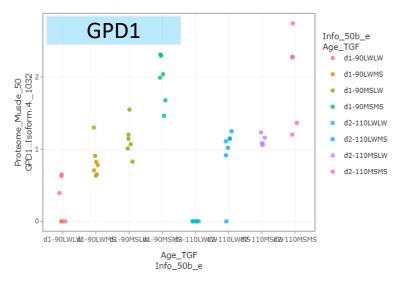
Run PLS-DA

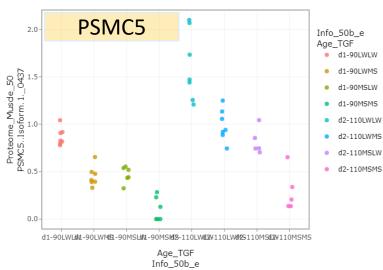
Explore individuals

Explore variables















Documentation

Integrate two datasets with PLS-DA

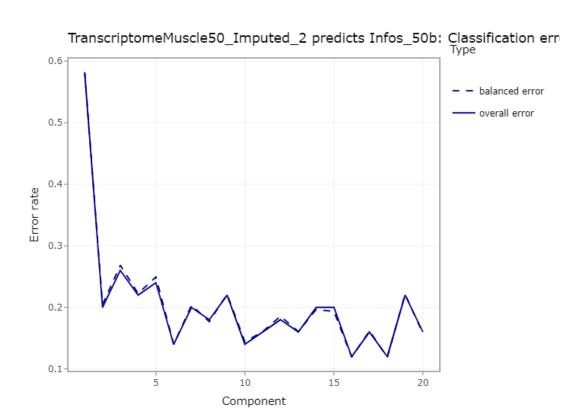
Preprocessing

Run PLS-DA

Explore individuals

Explore variables

Age\_TGF



PLSDA on TranscriptomeMuscle50\_Imputed\_2, Infos\_50b: X-space

