# **User guide for ExNormalizeMets:**

For this guide, a windows machine is being used. No prior software other than Microsoft Excel and a web browser are required for installation. The guide will explain how to install R to be able to use the NormalizeMets R package through its ExNormalizeMets Excel interface (no direct interaction with R or command line functions are needed).

A detailed guide with illustrations for using the software is also provided, showing its main functionalities in a 'walk through' guide of an example project.

For support contact: olshansky.g@unimelb.edu.au

## **Getting the files ready:**

To get the required files, go to <u>https://metabolomicstats.github.io/ExNormalizeMets/</u> and download the latest version of the *ExNormalizeMet*:

ExNormalizeMets
Software for Normalising and Analysing Metabolomics data View on GitHub Download .zip Download .tar.gz

ExNormalizeMets

They are all the files needed for installing R and loading NormalizeMets through excel.

Once downloaded, open the folder:



Click on ExNormalizeMets\_0.24 to open the excel worksheet but otherwise don't modify the files in this folder (the downloaded folder can be copied to a different location on your computer for convenient access although it is only needed for first use).

## General Workflow:



## **Installing R:**

1. Install R from CRAN (The Comprehensive R Archive Network) by going to the following page: <u>https://cran.r-project.org</u>

	The Comprehensive R Archive Network
	Download and Install R
	Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
CRAN	Download R for Linux
Mirrors What's new?	Download R for (Mac) OS X     Download R for Windows
Task Views	D is not of many Linux distributions, you should check with your Linux nockase mensagement system in addition to
Search	the link above.
About R R Homepage	Source Code for all Platforms
The R Journal	Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source
Software	code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
R Sources R Binaries	• The latest release (2017-11-30, Kite-Fating Tree) R-3.4.3 tor oz, read what's new in the latest version
Packages Other	
Oller	• Sources of <u>R appna and beta releases</u> (daily snapsnots, created only in time periods before a planned release).
Documentation Manuals FAOs	<ul> <li>Daily snapshots of current patched and development versions are <u>available here</u>. Please read about <u>new features</u> and <u>bug fixes</u> before filing corresponding feature requests or bug reports.</li> </ul>
Contributed	• Source code of older versions of R is available here.
	Contributed extension packages
	Questions About R
	<ul> <li>If you have questions about R like how to download and install the software, or what the license terms are, please read our <u>answers to frequently asked questions</u> before you send an email.</li> </ul>

What are R and CRAN?

R is 'GNUS' a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and

	R for Windows
Subdirectories:	
base	Binaries for base distribution. This is what you want to install R for the first time.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for $R < 2.13.x$ ; managed by Uwe Ligges)
<b>T.</b> 1	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build
Download R 3.4.3 f	R-3.4.3 for Windows (32/64 bit)
Installation and other in New features in this ve	nstructions rsion
If you want to double-che .exe to the <u>fingerprint</u> on t	ck that the package you have downloaded matches the package distributed by CRAN, you can compare the <u>md5sum</u> of the he master server. You will need a version of md5sum for windows: both <u>graphical</u> and <u>command line versions</u> are available.
	Frequently asked questions
<ul> <li><u>Does R run under n</u></li> <li><u>How do I update pa</u></li> <li><u>Should I run 32-bit</u></li> </ul>	<u>uy version of Windows?</u> ckages in my previous version of R? or 64-bit R?
Please see the R FAQ for	general information about R and the <u>R Windows FAQ</u> for Windows-specific information.
	Other builds
<ul> <li>Patches to this relea</li> <li>A build of the devel</li> <li><u>Previous releases</u></li> </ul>	se are incorporated in the <u>r-patched snapshot build</u> . opment version (which will eventually become the next major release of R) is available in the <u>r-devel snapshot build</u> .
Note to webmasters: A sta < <u>CRAN MIRROR</u> >/bin/y	ble link which will redirect to the current Windows binary release is <u>vindows/base/release.htm</u> .

-

2. Once the installation is complete, run the downloaded file R-3.4.3-win.exe, follow the installation instructions and choose the location where R should be installed.

🔂 Setup - R for Windows 3.4.3	
Select Destination Location Where should R for Windows 3.4.3 be installed?	R
Setup will install R for Windows 3.4.3 into the following folder.	
To continue, dick Next. If you would like to select a different folder, dick	Browse.
C:\Users\user\Documents\R\R-3.4.3	Browse
At least 1.2 MB of free disk space is required.	
< Back Next >	Cancel

Click *Next* for default location (Recommended) or choose location manually. Make sure you **remember** where **R** is installed as you will need to locate this folder later to link R to Excel.

3. Click *Next* to install with the different required settings:

🕞 Setup - R for Windows 3.4.3	- • -
Select Components Which components should be installed?	R
Select the components you want to install; clear the components yo install. Click Next when you are ready to continue.	ou do not want to
User installation	
Core Files	83.7 MB
32-bit Files	48.7 MB
64-bit Files	50.2 MB
Message translations	7.3 MB
Current selection requires at least 190.6 MB of disk space.	
< Back Nex	kt > Cancel

4. Click *Next* to install with default settings:

🕞 Setup - R for Windows 3.4.3	
Startup options Do you want to customize the startup options?	R
Please specify yes or no, then dick Next.	
Yes (customized startup)	
<ul> <li>No (accept defaults)</li> </ul>	
< Back Nex	ct > Cancel

5. In the next screen check the *Don't create a Start Menu folder* icon if you don't intend to use R by itself and click next. Also uncheck the *Create a desktop shortcut* in the next screen:

🔂 Setup - R for Windows 3.4.3	
Select Start Menu Folder Where should Setup place the program's shortcuts?	R
Setup will create the program's shortcuts in the following Start N To continue, click Next. If you would like to select a different folder, click	Menu folder. Browse.
R	Browse
Don't create a Start Menu folder	
< <u>B</u> ack <u>N</u> ext >	Cancel



6. Wait until installation is done, the following screen should appear



## Linking R to excel and installing the required packages (first use only):

After installing R open the excel file ExNormalizeMets.xlsm, this will open the Excel interface onto the settings sheet:

	File	н	ome Insert Page Layout	Formulas Data Review	View Developer	♀ Tell me what you wa	ant to do				Sign in	₽, Sha	ire
	D34			* : × √ fr									*
	1	A	В	D		E	F G	н	I	J K	L	м	
1	3 4 5		Install R - Import Data	Visualize the Data	Nor	malize	ldent Biomar	ify rkers	Cluster Analysis	Generate Report			-
2	6 7 8 9 10 11	* *	Locate Rscript.exe to Link R Inastall required R packages	Log Transform Data & Remove Missing Values	- Norr choose fro normalizati	nalize - om several on methods	Fit Linear (Use dircetl metho	Model ly for ruv ods)	UV RLA Plots	Classification			
	12 13 14 15		Set directory for the project	Generate RLA Plots		Current	Compare V Plot	Volcano ts	Compare PCA Plots	SVM Fit Classifier			-
	16 17 18 19		Import featuredata	Generate PCA Plots	Generate RLA Plot	s O norm1	Compare R	RLA Plots	Generate	Correlation			-
	20 21 22		Import sampledata	Generate Dendrogram		⊂ norm3	Compare I Histogr	P-Value rams	Dendrogram	Compare			
	23 24 25 26		Import metabolitedata (optional)	Generate Interactive HeatMap	Generate PCA Plot	s o norm4	Generate V	/enn Plot	Interactive HeatMap	Coefficients			-
3	27 28 29 30	+			Manual - For instruct	ions and detailed exp	planation of methods an	d functions pre	ess here				
	31 32		sattings (A)							: 51			
	Ready	0									1	+ 110	66

The *settings* sheet is your **'Control sheet'**, any function you want to run, from importing data, Normalizing, viewing results and opening the manual can be done from here.

This manual can be accessed at any stage by clicking on (3) but more on this late. For now, first locate the Rscripts.exe file so that excel will know how to run R commands it generates.

#### **Locating Rscript:**

Press (1. 'Locate Rscript.exe to Link R') to locate the Rscript file in the window that opens up. Make sure to select the Rscript file in the bin folder of your R installation:



After selecting the file, to install all the required R packages and set up the needed dependencies, press (2. '*Install required R packages*').

et Up R		
All required R package your internet connecti	s will be installed. This might tak on and machine speed. For instal	e some time depending on lation, the Excel file you
Press OK to start instal	ation	

Pressing ok will start the installation, this might take a few minutes if you are using R for the first time as many of the base packages will need to be installed.

The installation window looks like this:

. C: Joseis Joisnansky, g Juocuments /n /n-5.4.5 / Dim /nscript.exe
downloaded 45 KB
trying URL 'http://cran.us.r-project.org/bin/windows/contrib/3.4/crmn_0.0.20.zip
Content type 'application/zip' length 246551 bytes (240 KB)
trying URL 'http://cran.us.r-project.org/bin/windows/contrib/3.4/knitr_1.18.zip'
Content type 'application/zip' length 913875 bytes (892 KB)
trying UKL 'http://cran.us.r-project.org/bin/windows/contrib/3.4/rmarkdown_1.8.z ip'
Content type 'application/zip' length 2291306 bytes (2.2 MB)
downloaded 2.2 MB
package 'GGally' successfully unpacked and MD5 sums checked package 'plotly' successfully unpacked and MD5 sums checked package 'ggplot2' successfully unpacked and MD5 sums checked package 'htmlvidgets' successfully unpacked and MD5 sums checked
ownloaded 872 KB rying URL 'http://cran.us.r-project.org/bin/windows/contrib/3.4/rmarkdown_1.8.2 ontent type 'application/zip' length 2291306 bytes (2.2 MB) controloaded 2.2 MB wackage 'GGally' successfully unpacked and MDS sums checked ackage 'ggploi2' successfully unpacked and MDS sums checked

When the installation is done, a window with the message Done! will appear.

Now that the installation is complete, NormalizeMets is ready for use!

## **Using NormalizeMets**

Data used in the following examples is provided with NormalizeMets (alldata\_eg in R), it is located in the *ExampleData* folder in your downloaded *ExNormalizeMetsSetup* file. Future references in this guide refer to this data by default.

An example excel document containing the data used for the tutorial with all settings identical to those in the tutorial is provided (MyFirstNormalizeMetsProject\_example.xlsm), if using this document, make sure to set the Rscript location and working directory for your machine.

#### Starting a new project:

To start a new project, you will need to set up or use an existing directory where the project will 'live', data and plots generated by excel will all be saved to that folder.



Before selecting the folder, you will be prompted to save a copy of the current version of the excel file, it is recommended to save it with a new name to make sure a 'clean' version always stays in your ExNormalizeMetsSetup folder.

ĺ	Microsoft Excel	×
	You are encouraged to save a copy of this excel file to be used for the current project!	
	ОК	

After saving the workbook under the name of your choice:

🔀 Save As	
🚱 🗢 📲 « Documents > Metab	olomics - 4, Search Metabolomics P
Organize 🔻 New folder	8H <b>•</b> 🔞
Documents library	Arrange by: Folder 🔻
Name ^	Date modified Type Size
File name: MyFirstNormalize	AetsProject 👻
Save as type: All Files	•
Authors: Add an author	Tags: Add a tag
<ul> <li>Hide Folders</li> </ul>	Tools   Save Cancel

Select the working directory where all files should be generated. We recommend making a new folder for each project.

X Please select a	Working Directory f	or the pro	ject						×
<b>GO-</b>	Libraries + Docu	ments 🕨	Metabolomics	•	-	<b>4</b> 7	Search Metabolo	mics	٩
Organize 🔻	New folder							888 🕶	0
XII Micro	Name	î.		Dat	te modifi	ed	Туре	Size	
	🎍 FirstProject			15/	01/2018	5:00 PI	M File folder		
Elibra									
M Vid *									
	Folder name:	FirstProj	ect						
					Tools	- (	OK	Cancel	

#### Loading data:

To load data for the project, in turn click on the following to load the relevant data:

	A A	B	( D	E		F G H	I	J K
Import	2 3 4 5	Install R - Import Data	Visualize the Data	Norma	lize	Identify Biomarkers	Cluster Analysis	Generate Report
featuredata	6 7 8 9	Locate Rscript.exe to Link R	Log Transform Data &	- Normal	ize -	Fit Linear Model (Use dircetly for ruv	UV RLA Plots	
Import	10 11 12 13 14	Inastall required R packages Set directory for	Remove Missing Values Generate RLA Plots	normalization	methods	Compare Volcano Plots	Compare PCA Plots	SVM Fit Classifier
sampiedata	15 16 17 18 19	Import featuredata	Generate PCA Plots	Generate RLA Plots	• norm1	Compare RLA Plots	Generate	Correlation
Import metabolitedata	20 21 22 23	Import sampledata Import	Generate Dendrogram Generate	Generate PCA Plots	C norm8	Compare P-Value Histograms	Generate	Compare Correlation
(optional)	24 25 26 27 28	metabolitedata (optional)	Interactive HeatMap	Manual	0 norm#	Generate Venn Plot	Interactive HeatMap	Coefficients
	29 30 31			Manual - For Instructions	and detailed explan	ation of methods and functions pre	as nere	

Loaded data needs to be in .csv format. After clicking on the required file, it will open a new sheet, showing the imported data. Select the setting sheet import more data and get back to the options.

For *featuredata*, set metabolites in columns and samples in rows. Unique sample names should be provided as row names.

						MyFict	VormalizeMetsProjer						
													n A Share
A1				1 × ~	f.e								÷
	A	8	с	D	E	F	G	н	1	J	к	L	M
1		m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_10	m_11	m_12
2	5_1	10485.86719	33220.5625	1112.979492	1408.639648	455.7529297	1100.402344	122.3631592	3855.804688	1637.482422	6194.292969	28793.65625	5200.75
3	s_2	8960.46875	29995.51563	926.2890625	529.0449219	873.609375	2201.246094	173.0802002	5090.636719	2011.500977	8868.859375	31683.09375	6183.91
4	s_3	10160.44531	28559.64063	1230.333008	1306.320313	1027.507813	2066.591797	269.7998047	4483.90625	1644.607422	7776.335938	44494.59375	10505.5
5	s_4	8794.476563	27593.75	901.762207	1800.083008	675.0395508	1675.849609	287.9992676	4949.261719	1508.802734	8405.171875	37030.40625	10726.6
6	s_5	8956.921875	28161.76563	979.1723633	818.0366211	904.0253906	1245.991211	167.4133301	4302.933594	1460.012695	6976.429688	29579.53125	5421.98
7	5_6	9092.257813	31685.3125	634.6899414	478.3811035	980.4233398	1259.563477	97.5625	4406.34375	2349.587891	7090.136719	39835.71875	5657.80
8	s_7	2271.044922	7692.175781	143.1688232	109.8399048	730.6171875	1387.791992	149.0966797	1543.40625	961.7314453	2356.685547	20212.70313	4483.5
9	5_8	7850.402344	26462.79688	822.5991211	1341.21582	583.5830078	2164.126953	227.7696533	5030.6875	1604.274414	6241.457031	30873.14063	8130.19
10	s_9	10969.0625	21605.14063	408.0004883	1499.479492	452.2375488	916.5283203	234.4975586	4657.253906	1847.6875	6606.15625	34395.96875	7142.32
11	5_10	1743.932617	7647.644531	218.4199219	127.4294434	408.2912598	999.5717773	74.82696533	1510.011719	895.9853516	2407.103516	18256.14063	2358.10
12	s_11	1284.412109	4819.996094	61.62762451	15.51902008	394.8364258	1112.143555	28.35443115	1163.742188	628.7016602	2646.070313	10544.94531	1728.:
13	5_12	2272.923828	6813.679688	243.9420166	280.7983398	434.3505859	1167.960938	86.25360107	1665.297852	1009.748047	3927.849609	21808.45313	6546.99
14	s_13	9938.0625	28498.21875	577.8486328	1302.723633	350.9599609	1200.636719	134.2883301	4406.34375	1777.506836	8467.085938	30871.17188	6432.08
15	5 14	6795.753906	24264.64063	745.7470703	465.0424805	990.1313477	1794.898438	119.0358276	4313.539063	1480.686523	6845.367188	32930.875	8029.3
16	s_15	1836.59375	6009.238281	142.5959473	37.65142822	491.6550293	1753.573242	70.06390381	1526.198242	941.4536133	3346.365234	20822.03125	2188.4
17	s_16	1699.289063	6458.503906	96.57415771	200.644165	310.729248	774.0219727	183.9802246	1690.515625	900.3242188	3143.683594	14266.88281	1877.80
18	5_17	1497.520508	5609.605469	77.26855469	169.138916	226.6845703	567.8823242	79.01464844	1362.202148	830.9599609	2024.548828	9131.3125	1730.23
19	s_18	9033.570313	21372.0625	1250.067383	1564.307617	892.3813477	1589.588867	252.1553955	4251.226563	1510.931641	6424.839844	38623.5625	7712.2
4		featuredata	settings	۲									E.
Reath	10							Average: 11023.3937	1 Count: 19478	5um 211649159.3			+ 127%

sampeldata should have sample information matching featuredata (samples in rows).

بە 🗄	ð- <del>:</del>						MyFirstNormal	izeMetsProject	- Excel						ॼ _ □	×
File							Developer								gn in 🧏 Sh	are
A1			¥ 1	×	fer .											~
A	ВС	D	E	F	G	н	I	J	к	L	м	N	0	Р	Q	
1	batch gende	r Age	bmi													
2 s_1	Batch 2 code	1 58.7	22.2													
3 s_2	Batch 2 code	0 76.7	23.7													
4 s_3	Batch 2 code	0 56.2	28.2													
5 s_4	Batch 2 code	0 77.2	26.2													
6 s_5	Batch 2 code	0 74.3	26.4													
7 s_6	Batch 2 code	1 66.8	26.4													
8 s_7	Batch 1 code_	0 65	29													
9 s_8	Batch 2 code_	1 66.5	26.1													
10 s_9	Batch 2 code_	0 70.2	27.3													
11 s_1	0 Batch 1 code_	1 55	25													
12 s_1	1 Batch 1 code_	0 51	28													
13 s_1	2 Batch 1 code_	1 55	29													
14 s_1	3 Batch 2 code_	0 79.2	30.7													
15 s_1	4 Batch 2 code_	1 52.7	25.4													
16 s 1	5 Batch 1 code	0 42	32	_												
>	featuredata	sampled	ata settin	P	÷											)
Ready	ái (							Aver	age: 44.0198333	3 Count 754	Sum: 13205	.95 🔠	•	+	+ 12	0%

Optional *metabolitedata* should have metabolite information matching featuredata with metabolite names in rows.

⊟	<b>5</b> • ∂					1	MyFirstNormali	eMetsProject -	Excel					æ	- 🗆 ×
File					ata Revi		Developer								n 🤉 Share
A1				* : ×	√ fe	names									~
	A	B C		D	E	F	G	н	I	J	к	L	м	N	0 •
1	names	IS neg_cor	trols pos_co	ntrols_gender											
2	m_1	0	1	(	D										
3	m_2	0	1	(	D										
4	m_3	0	1		D										
5	m_4	0	0	(	D										
6	m_5	0	0		D										
7	m_6	0	0		D										
8	m_7	0	0		D										
9	m_8	0	1		D										
10	m_9	0	1		D										
11	m_10	0	1		D										
12	m_11	0	0	(	D										
13	m_12	0	0		D										
14	m_13	1	1		D										
15	m_14	0	0		D										
		featuredat	a sampledata	metabolited	lata set	ings (+	)							4	Þ

*metabolitedata* can include any metabolite information such as grouping structures, internal standard metabolites, negative control and positive control metabolites.

After the data is loaded, you are ready to proceed to analyse the data!

# **NormalizeMets Workflow:**

## Visualize the Data

The following section refers to the visualize part:

nstall R - Import Data	Visualize the Data	Norma	lize	ldentify Biomarkers	Cluster Analysis	Generate Repo	
Locate Rscript.exe to Link R Inastall required R packages	Log Transform Data & Remove Missing Values	- Normal choose from normalization	lize - 1 several 1 methods	Fit Linear Model (Use dircetly for ruv methods)	UV RLA Plots	Classification	
Set directory for the project	Generate RLA Plots		Current	Compare Volcano Plots	Compare PCA Plots	SVM Fit Classifier	
Import featuredata	Generate PCA Plots	Generate RLA Plots	( nam1	Compare RLA Plots	Generate	Correlation	
Import sampledata	Generate Dendrogram		O norm2	Compare P-Value Histograms	Dendrogram	Analysis	
Import metabolitedata (optional)	Generate Interactive HeatMap	Generate PCA Plots O norma		Generate Venn Plot	Generate Interactive HeatMap	Compare Correlation Coefficients	

#### Log Transforming the data and removing missing value (mandatory):

A frequent issue in metabolomics data sets is the occurrence of missing values. It is important to reduce the number of missing values as much as possible by using an effective pre-processing procedure. For example, a secondary peak picking method can be used for LC-MS data to fill in missing peaks which are not detected and aligned.



"knn" - use k nearest neighbours method to replace missing values.

"replace" - replaces missing values by half the minimum value in featuredata.

#### Clicking 'Remove Missing Values & Log Transforming' the following appears:

	featuredata	Generate	K - Number of 10 teaturemax 0.8 prcelation ysis	
	sampledata	Dendrogram	Seed For R random 100 Missing Values Removed! View results in featuredata.nomissing, lation	
	Import metabolitedata (optional)	Generate Interactive HeatMap	number generator sampledata.nomissing and metabolitedata.nomissing	
			Remove Missin Trans	
)	featuredata.lo	og featuredata.nom	issingnolog featuredata.nomissing sampledata.nomissing metabolitedata.nomissing settings 🕘	:
1				

Note the new sheets that appeared, they have respectively the regular log transformed featuredata, the log transformed featuredata with missing values removed, featuredata with missing values removed and without log transformation, sampledata with rows removed corresponding featuredata.nomissing, metabolitedata with rows removed corresponding to featuredata.nomissing. Unless you are interested to view or copy any of this data, those sheets are only going to be used for further internal functions.

Now the plots in *Visualize the Data* can be generated! The data generated is also going to be used for the Normalization section.

#### **RLA plots**

One way of visualising the log transformed metabolomics data is the use of *across group* or *within group* relative log abundance (RLA) plots (De Livera et al. 2012 De Livera et al. (2015)).



Setting groupdata to *batch* and selecting *Generate plot* opens the interactive plot in the default web browser and saves a copy in the working folder.



Interactive plot opens in the default browser, to zoom in, simple select the required part

			1.4		1 ET1	0
	Organize 👻 👩 Open	<ul> <li>Share with</li></ul>	er		8== •	
Copy of all plots together with all other	🚖 Favorites	Documents library FirstProject		Arran	ge by: Folder	•
generated files are stored in the working	Downloads	Name	Date modified	Туре	Size	
directory	Recent Places	InteractiveRIaPlot_files	16/01/2018 2:25 AM	File folder		
	Libraries	Featuredata_c	16/01/2018 2:25 AM	Microsoft Excel C	225 KB	
	Documents	featuredata_log	16/01/2018 12:46	Microsoft Excel C	319 KB	:
	A Music	D InteractiveRlaPlot	16/01/2018 2:25 AM	Firefox HTML Doc	740 KB	
	Pictures	metabolitedata_c	16/01/2018 12:46	Microsoft Excel C	2 KB	
	Videos	nomissing_featuredata	16/01/2018 12:46	Microsoft Excel C	222 KB	
		nomissing_metabolitedata	16/01/2018 12:46	Microsoft Excel C	2 KB	
	Computer *	nomissing_sampledata	16/01/2018 12:46	Microsoft Excel C	6 KB	

#### **PCA plots**

The following function can be used to obtain multiple plots for exploration of the principal components of the *featuredata* matrix: a bar plot indicating the variance explained by each principal component, scores and loading plots with specified axes (interactive and non-interactive), and a pairs plot of the first *n* principal components. These plots are useful in identifying any outlying samples and getting a preliminary understanding of the structure of the data.



All produced plots are stored in the working directory, with interactive plots opened in the browser and static plots located in the new *plots* sheet.

#### Dendrogram

Generates a dendrogram to visualise clustering structures in the data, many different methods ae available.



#### **HeatMap**



The HeatMap produced can reveal interesting structures in the data.

## **Normalization**

Normalization methods presented in this package are divided into four categories, as those which use (i) internal, external standards and other quality control metabolites (*NormQcmets*) (Sysi-Aho et al. 2007, Redestig et al. (2009), De Livera et al. (2012), De Livera et al. (2015), Gullberg et al. (2004)) (ii) quality control samples (*NormQcsamples*) (Dunn et al. 2011), (iii) scaling methods (*NormScaling*) (Scholz et al. 2004, Wang et al. (2003)), and (iv) combined methods (*NormCombined*) (Kirwan and Broadhurst (2013)).

The normalization methods are accessible in the following section:



Clicking on the Normalize button opens the following menu enabling the choice of different normalization methods.



#### NormQcmets

The approaches in *NormQcmets* use internal, external standards and other quality control metabolites. These include the *is* method which uses a single standard (Gullberg et al. 2004), the *ccmn* (cross contribution compensating multiple internal standard) method (Redestig et al. 2009), the *nomis* (normalization using optimal selection of multiple internal standards) method (Sysi-Aho et al. 2007), and the remove unwanted variation methods (J. A. Gagnon-Bartsch, Jacob, and Speed 2014) as applied to metabolomics using "ruv2" (De Livera et al. 2012), "ruvrand" and "ruvrandclust" (De Livera et al. 2015). Note that *ruv2* is an application specific method designed for identifying biomarkers using a linear model that adjusts for the unwanted variation component.

To Normalize:

	Normalization Methods based on Quality Control metabolites	control samples
- NormQcmets - Normalization using quality	Choose Normalization Method: Choose column with metabolite QC indicators (qcmets): For 'CCMN' method elect Factors for 'ccmn' method elect Factors for 'ccmn' method winder + batch + gender without for Clear Tactors NULL	Choose from columns in metabolitedata. Should have for QC metabolite and 0 for regular metabolites
See details about the <i>ruv</i> method for an explanation of those parameters	For All ruv methods:     Number of unwanted K:     NULL       For 'ruvrandclust' method     Number of unwanted k:     NULL       Maximum Iterations     200     IambdaUpdate     F       nUpdate     300     p - number of dusters     2       For 'ruvrand' method:     Normalisation parameter lambda     0.01	

After Clicking the Normalise button the screen asking you where the normalized data is to be saved appears.

	Choose what to do with the normalised data Open Normalised data into the Excel Workspace?	There are 4 'slots' for saving normalised data inside the excel sheet. Data saved in these sheets, 'norm1' 'norm4' can later be accessed by excel.
	Open in norm1     Open in norm2	
Normalise	C Open in norm3 C Open in norm4	
	C Don't Show data Matrix, a local copy will be save	However as it is also possible to save a copy of the normalised data in the worksheet without
	Continue	importing it back to excel.

Upon clicking continue, you will return back to the settings sheet but you can notice some changes:



#### NormQcsamples

This function is based on the quality control sample based robust LOESS (locally estimated scatterplot smoothing) signal correction (QC-RLSC) method as described by Dunn et al. (2011) and impletemented statTarget (Luan 2017)

To Normalize:		
	Settings For Normalization using quality control samples Normalisation Method based on Quality Control samples 'rics'	View the references above for detailed explanation of the method
- NormQcsamples -	Set smoothing d Set degree of 2 -	
Normalization using quality control samples	Choose column corresponding to batch number (Batch)	
	Choose column corresponding to run order	*
	Normaliae Cancal	This method uses a unique form of data.

#### NormScaling

The scaling normalization methods (Scholz et al. 2004, Wang et al. (2003)) included in the package are normalization to a total sum, normalisation by the median or mean of each sample, and are denoted by *sum, median*, and *mean* respectively. The method *ref* normalises the metabolite abundances to a specific reference vector such as the sample weight or volume.

To Normalize:

	Normalization Settings	Choose from available scaling methods
- NormScaling - Normalization using scaling methods	Choose Normalisation Method:  Ter 'ref' method Column indicating reference vector Select number of reference vector Normalise Cancel	Can manually set reference vector, shouldn't be change for most purposes

#### NormCombined

In some circumstances, researchers use a combination of the above normalizations (i.e., one method followed by another). This can be achieved using the *NormCombined* function. The function defaults to employing 'rlsc' approach followed by the `median'.

To Normalize:

		Normalization Settings	initian - chaosa 2 Mathada	
•	- NormCombined - Normalization using combination of two methods	Method 1 Yisc1	Method 2 "medan"  Cancel	Choose methods and press Normalize to customize settings for selected Method 1 and Method 2

Note that normalizing the data is not necessary to proceed to fitting a linear model although it is highly recommended to try a few normalization methods when analysis data.



# Assessing and choosing normalization methods

The criteria for assessing and choosing a normalization method implemented NormlizeMets have been described in detail by De Livera et al. (2012), De Livera et al. (2015) and J. A. Gagnon-Bartsch, Jacob, and Speed (2014).

## **Identifying Biomarkers**

To view and compare the biomarkers identified, first a linear model has to be fitted to the data.



#### **Fit Linear Model**

A linear model has to be fitted for every Normalization method that is to be used down the line for Biomarker identification. Setting from one 'run' of the Fit Linear model will be saved for the next.

To Fit Linear Model:



	ignin A Share												
									V &	* 1 X			1
1	M	£	ĸ	)	1.1	. H	6	T I	1	0	c	1.000	A 4
	t stat Age t	t stat gendercode 1.1	t stat batchBatch 2 1	t stat (intercept) 1	coeff bmi	coeff Age	coeff gendercode 1	coeff batchBatch 2	coeff (Intercept) (	Adjusted Fip value ic	p value A	Fstat F	
	0.825784504	1.764943076	0.960005727	42.80583065	0.005229044	0.001454973	0.055865111	0.041846357	8.347035337	9.905-80	1.855-80	367.6640084	m 1
	0.106085024	1.431743816	0.250133436	54.69338644	-0.005601774	0.000173457	0.04205567	0.010118266	9.897252322	6.228-94	6.586-95	599.0273381	m 2
	-0.672814342	3.705851365	0.183890643	13.55311266	-0.016077832	-0.003033747	0.300187581	0.020513485	6.753388886	1.67E-25	1.296-25	39.83815528	ma
	1.776037426	2.712439324	0.390438834	7.735175731	0.027199876	0.010342764	0.283769155	0.05625139	4.985353126	1.76E-11	1.536-11	14.54006231	m_4
	-3.33111604	-1.14780064	1.581976002	18.91529068	0.011823617	-0.010603696	-0.065637851	0.124584175	6.66379303	3.238-38	2.05E-38	74.81961081	ms
	-2.15682434	-3.398042368	0.532692825	25.62430539	-0.00401972	-0.005932454	-0.167907096	0.036348681	7.800329522	2.516-52	1.16E-52	134.6606152	m_5
	0.454696412	-0.780593194	0.426049287	8.650757264	0.018530789	0.002152121	-0.066372874	0.049888541	4.531485299	4.04E-12	3.488-12	15.47626238	m_7
	0.18047732	-0.867447684	0.69416094	85.50233518	-0.000384273	0.000154624	-0.013351148	0.014713317	8.107255999	2.47E-120	4.02E-122	1462.387307	m_8
	-0.107033545	-0.291196344	0.450638161	78.30533966	0.002335586	-8.91E-05	-0.004356857	0.009903318	7.217708722	3.23E-115	1.05E-116	1226.56955	m.s
	1.50409845	1.295362839	-0.121375792	51.50038812	-0.000353099	0.00222685	0.034452986	-0.004445725	8.43852463	1.988-90	2.57E-91	531.2501461	m 10
	-1.467517268	1.824191114	1.515629134	39.15272664	0.016136544	-0.003386947	0.07563383	0.086539384	10.00065299	8.79E-75	1.935-75	309.130316	m_11
	1.714875421	1.35177297	0.496640156	20.89799348	0.02113052	0.005668586	0.080272464	0.039796591	7.645185192	3.23E-42	1.84E-42	89.17162335	m_12
	-1.192604835	2.001015074	0.407048264	20.9522018	-0.002977858	-0.004184627	0.126133714	0.035334734	8.13637728	3.68E-42	2.13E-42	88.93191195	m_14
	-0.807383006	3.297606306	-0.624667412	18.21328348	-0.0193363	-0.002562198	0.187997712	-0.049043076	6.396797354	1.37E-36	9.158-37	69.47791984	m_15
	1.004869076	0.791485203	1.170553694	14.65952391	0.031933538	0.003252398	0.046021169	0.093730574	5.251160527	5.96E-28	4.31E-28	45.54787626	m_16
-	-0.047653038	3.817067487	0.472646389	27.01482123	0.007441495	-0.000119121	0.171414195	0.029229971	7.473768302	5.81E-55	2.60E-55	149.098252	m_17
	-0.800552027	0.693838481	0.93163199	12.64444705	0.019401339	-0.002834625	0.044135203	0.081610536	4.95504007	2.858-22	2.296-22	32.98089506	m_18
	3.307272003	0.221861431	1.790631028	18.72237101	0.009311617	-0.009050349	0.010906812	0.121226408	5.670188116	1.00E-37	6.458-38	73.17804288	m_19
	1.690139137	2.385685586	-0.869503121	20.28596453	-0.002883421	0.004511866	0.134409506	-0.057424947	5.99336342	6.796-41	4.09E-41	84.18842903	m_20
	0.591887814	2.099125195	1.711110084	6.131752041	0.058851358	0.002453416	-0.156311434	0.175471041	2.81292127	1.71E-10	1.515-10	13.1283853	m_21
	1.792375048	1.319611638	0.06450134	36.06223625	0.003493443	0.003571628	0.047239371	0.003179813	7.953004173	4.36E-70	1.24E-70	261.150688	m_22
	-0.466092325	-3.043820023	1.879217797	47.47869662	-0.014272242	-0.000611448	-0.071734387	-0.060990281	6.89330977	6.18E-86	9.058-87	455.838848	m_23
	2.41043648	7.445599934	1.054710241	29.70463669	0.019395348	0.004704707	0.26107016	0.053411949	6.416561638	1.596-61	6.22E-62	190.9438527	m_24
	-0.776292248	2.830692881	0.195014779	42.74344536	-0.004939575	-0.001594997	0.104490507	-0.009912837	9.71953441	1.02E-79	2.00E-80	367.2450831	m 25
	-0.242043202	4.089108082	-0.555124672	30.70246805	-0.007441411	-0.000525045	0.190003558	-0.035522111	8.788748138	1.118-61	4.245-62	192.1130141	m_26
	-0.477658863	-1.828512586	1.874542384	10.38363815	0,085224219	-0.003191538	-0.219482973	0.309863512	7.678440315	9.63E-19	7.915-19	26.26732607	m_27
21	-0.135511198	-2.64711454	2.269684461	16.18465913	0.07814742	-0.000650212	-0.228184862	0.269435414	8.594848786	1.668-33	1.135-33	60.18175638	m_28

A new sheet is created, it stores the linear model output. This data should not be altered manually as other functions will need to access



#### **Volcano Plots**

overwritten

Volcano plots are useful in identifying biomarkers and generally assessing the normalization.

#### Compare Volcano Plots:



The plot is saved in the working directory and opens in the default browser.



#### **Compare RLA plots**

Used to assess normalization by comparing relative log abundance plots, similar input to the *Generate RLA* plots function

Compare Rla Plots:



#### **Compare P-Value Histograms**

Compare histograms of the coefficient's p-values. The distribution of the p-values should be used to assess the success of the normalization.

Compare P-Value histograms:



#### **Generate Venn Plot**

Generates a Venn plot that compares the biomarkers identified by the different normalization methods.

Generate Venn Plot:

i	Idantifiz	Set Venn Plot Settings	Venn Diagram
	Biomarkers	Generate Venn Plot	Sun, Amus (10)
	Fit Linear Model (Use dircetly for ruv methods)	Choose up to 3 methods for the Venn Plot Only available data that has a linear model fitted to it (using the Linear Model Fit function) will apear in the options	
	Compare Volcano Plots	Plot 1         Plot 2         Plot 3           norm1 (nomis)         •         norm3 (median)         •	
	Compare RLA Plots	P - Imit 0.05 Plot Title "Venn Dieoram"	
_	Compare P-Value Histograms	Text size 1 Aspect 1	
	Generate Venn Plot		
		Generate Plots Cancel	aansi aasaa (0) Aasaa (1)

# **Cluster Analysis**



### UV RLA plots

Unwanted Variation relative log abundance plots enable visualisation of the unwanted variation removed by each normalization method.



#### **Compare PCA Plots**

Compare principal component multi-plots for differed normalization methods.



<u>Generate Dendrogram and Generate Interactive HeatMap</u> are identical to those discussed in the *Visualize the Data* section. The user has to choose the normalized data to be used.

### **Classification**

Classification accuracy is a good way to assess the success of a given normalization method

#### SVM-Fit

The Support Vector Machines method is used to classify the data, the classification accuracy is then assessed based on the factor specified.

	SVM Fit Classification Accuracy settings	
Classification	SVM Fit Classification Accuracy	
SVM Fit Classifier	Normalization Source: Factor to use for classification:           norm2(runnad)         •	▶
	Tune Hyperparameters	a socia
	Crosswalidete Classification True 💌 K (fold cross validation) 5	3 -
	Generate Flots Cancel	2 - V

## **Correlation Analysis**

It is important to look at correlation coefficients when normalizing data.

### **<u>Compare Correlation coefficients</u>**



# **Generate Report**

The *Generate Report* function generates an interactive report based on basic user input. There is a choice of up to 3 normalization methods to be included together with the unadjusted data. The report includes various plots and diagnostic to assess the normalization. Guidance on interpretation of the various plots, together with notes of what the user should look for when assessing the results is provided in the generated document.

Generate Report:



An example report is available in your downloaded ExNormalizeMets docs folder.

## **Full Package Vignette:**

For package vignette with detailed explanations of methods and workflow, follow the link to: <u>https://cran.r-</u>project.org/web/packages/NormalizeMets/vignettes/NormalizeMets\_vignette.html