Seminar

Reporter: Lichuang Huang

2022-07-07

Supervisor: Gang Cao

1. **RNA-seq analysis**

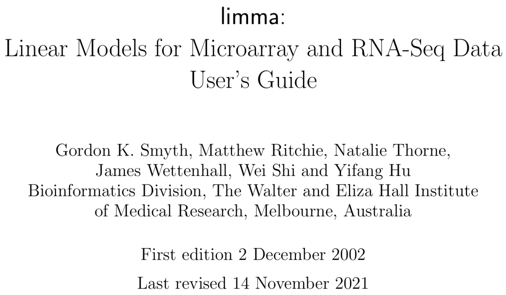
2. **AHR and Kidney**

3. **MCnebula in Website**

4. **Next Schedule**

RNA-seq analysis

Literature and Guidances



limma:

Linear Models for Microarray and RNA-Seq Data
User's Guide

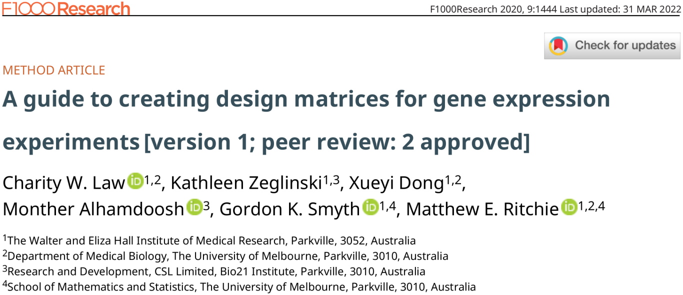
Gordon K. Smyth, Matthew Ritchie, Natalie Thorne,
James Wettenhall, Wei Shi and Yifang Hu
Bioinformatics Division, The Walter and Eliza Hall Institute
of Medical Research, Melbourne, Australia

First edition 2 December 2002

Last revised 14 November 2021

Figure 1: limma guidance

Literature and Guidances



F1000Research F1000Research 2020, 9:1444 Last updated: 31 MAR 2022

Check for updates

METHOD ARTICLE

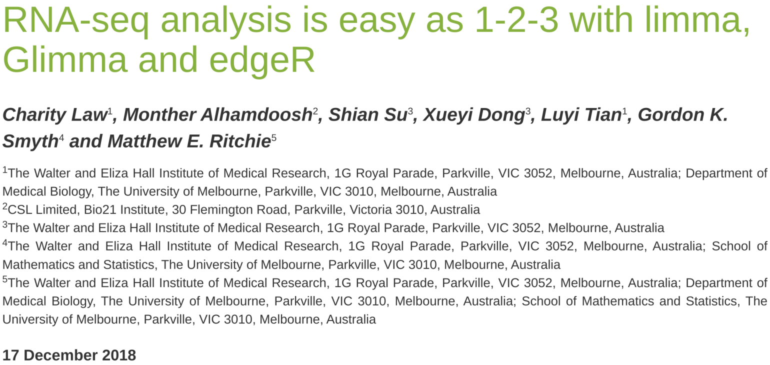
A guide to creating design matrices for gene expression experiments [version 1; peer review: 2 approved]

Charity W. Law[1,2], Kathleen Zeglinski[1,3], Xueyi Dong[1,2], Monther Alhamdoosh[3], Gordon K. Smyth[1,4], Matthew E. Ritchie[1,2,4]

[1] The Walter and Eliza Hall Institute of Medical Research, Parkville, 3052, Australia
[2] Department of Medical Biology, The University of Melbourne, Parkville, 3010, Australia
[3] Research and Development, CSL Limited, Bio21 Institute, Parkville, 3010, Australia
[4] School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010, Australia

Figure 2: design matrix

Literature and Guidances



RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR

Charity Law[1]***, Monther Alhamdoosh***[2]***, Shian Su***[3]***, Xueyi Dong***[3]***, Luyi Tian***[1]***, Gordon K. Smyth***[4] ***and Matthew E. Ritchie***[5]

[1]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Melbourne, Australia; Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Melbourne, Australia
[2]CSL Limited, Bio21 Institute, 30 Flemington Road, Parkville, Victoria 3010, Australia
[3]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Melbourne, Australia
[4]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Melbourne, Australia; School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Melbourne, Australia
[5]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Melbourne, Australia; Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Melbourne, Australia; School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Melbourne, Australia

17 December 2018

Figure 3: RNA-seq

Analysis route

1. Data packaging
   - Reading in count-data
   - Organising sample information
   - Organising gene annotations
2. Data pre-processing
   - Transformations from the raw-scale
   - Removing genes that are lowly expressed
   - Normalising gene expression distributions
   - Unsupervised clustering of samples
3. Differential expression analysis
   - Creating a design matrix and contrasts
   - Removing heteroscedascity from count data
   - Fitting linear models for comparisons of interest
   - . . .

Limma Workflow: read data

Raw counts data

Table 1: Raw counts

	10_6_5_11	9_6_5_11	purep53	JMS8-2	JMS8-3
497097	1	2	342	526	3
100503874	0	0	5	6	0
100038431	0	0	0	0	0
19888	0	1	0	0	17
20671	1	1	76	40	33
27395	431	771	1368	1268	1564

Limma Worklow: gene annotations

Table 2: Gene annotations

ENTREZID	SYMBOL	TXCHROM
497097	Xkr4	chr1
100503874	Gm19938	NA
100038431	Gm10568	NA
19888	Rp1	chr1
20671	Sox17	chr1
27395	Mrpl15	chr1

Limma Worklow: filter and normalization

Table 3: Normalization

	10_6_5_11	9_6_5_11	purep53	JMS8-2
497097	-4.309973	-3.851299	2.5254857	3.298898
100503874	-5.894935	-6.173227	-3.4350428	-3.040952
100038431	-5.894935	-6.173227	-6.8944744	-6.741392
19888	-5.894935	-4.588264	-6.8944744	-6.741392
20671	-4.309973	-4.588264	0.3629134	-0.401542
27395	3.858281	4.418296	4.5239053	4.567516

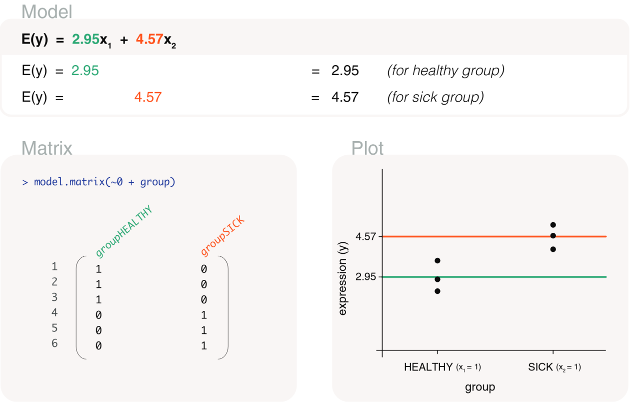
Limma Worklow: design matrix

Design matrix to create linear model

Table 4: Design matrix

Basal	LP	ML	laneL006	laneL008
0	1	0	0	0
0	0	1	0	0
1	0	0	0	0
1	0	0	1	0
0	0	1	1	0
0	1	0	1	0

Limma Worklow: design matrix



Model

$E(y) = 2.95x_1 + 4.57x_2$

$E(y) = 2.95 \quad = 2.95$ *(for healthy group)*

$E(y) = \quad 4.57 \quad = 4.57$ *(for sick group)*

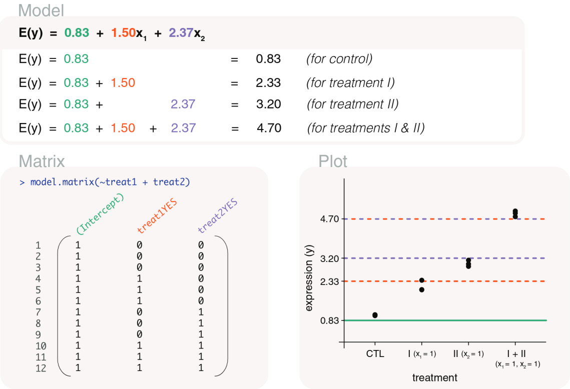
Matrix

```
> model.matrix(~0 + group)
     groupHEALTHY   groupSICK
1    1              0
2    1              0
3    1              0
4    0              1
5    0              1
6    0              1
```

Plot

Figure 4: Single factor

Limma Worklow: design matrix



Model

$E(y) = 0.83 + 1.50x_1 + 2.37x_2$

$E(y) = 0.83$	=	0.83	*(for control)*
$E(y) = 0.83 + 1.50$	=	2.33	*(for treatment I)*
$E(y) = 0.83 + \quad 2.37$	=	3.20	*(for treatment II)*
$E(y) = 0.83 + 1.50 + 2.37$	=	4.70	*(for treatments I & II)*

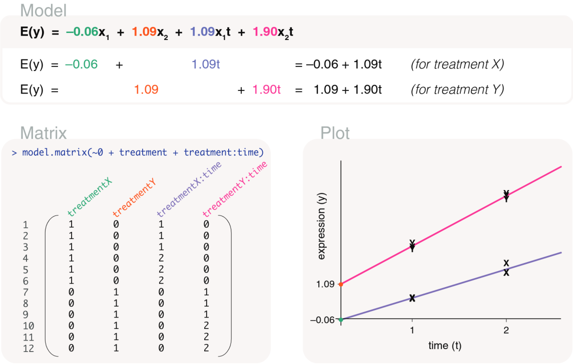
Matrix

```
> model.matrix(~treat1 + treat2)
   (Intercept) treat1YES treat2YES
1       1          0         0
2       1          0         0
3       1          0         0
4       1          1         0
5       1          1         0
6       1          1         0
7       1          0         1
8       1          0         1
9       1          0         1
10      1          1         1
11      1          1         1
12      1          1         1
```

Plot

Figure 5: multiple factor

Limma Worklow: design matrix



Model

$$E(y) = -0.06x_1 + 1.09x_2 + 1.09x_1t + 1.90x_2t$$

$E(y) = -0.06 + 1.09t = -0.06 + 1.09t$ *(for treatment X)*

$E(y) = 1.09 + 1.90t = 1.09 + 1.90t$ *(for treatment Y)*

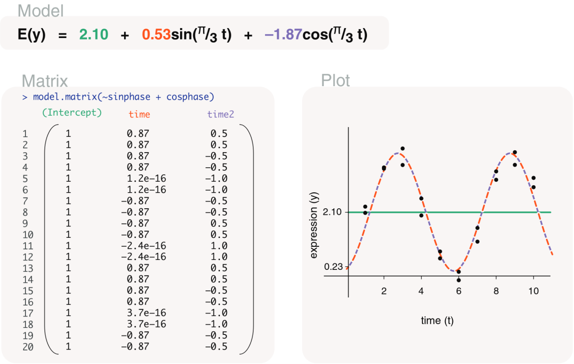
Matrix

```
> model.matrix(~0 + treatment + treatment:time)
   treatmentX treatmentY treatmentX:time treatmentY:time
1      1          0            1               0
2      1          0            1               0
3      1          0            1               0
4      1          0            2               0
5      1          0            2               0
6      1          0            2               0
7      0          1            0               1
8      0          1            0               1
9      0          1            0               1
10     0          1            0               2
11     0          1            0               2
12     0          1            0               2
```

Plot

Figure 6: covariate: time series

Limma Worklow: design matrix



Model

$$E(y) = 2.10 + 0.53\sin(^{\pi}/_{3}\,t) + -1.87\cos(^{\pi}/_{3}\,t)$$

Matrix

```
> model.matrix(~sinphase + cosphase)
     (Intercept)     time         time2

1         1          0.87          0.5
2         1          0.87          0.5
3         1          0.87         -0.5
4         1          0.87         -0.5
5         1          1.2e-16      -1.0
6         1          1.2e-16      -1.0
7         1         -0.87         -0.5
8         1         -0.87         -0.5
9         1         -0.87          0.5
10        1         -0.87          0.5
11        1         -2.4e-16       1.0
12        1         -2.4e-16       1.0
13        1          0.87          0.5
14        1          0.87          0.5
15        1          0.87         -0.5
16        1          0.87         -0.5
17        1          3.7e-16      -1.0
18        1          3.7e-16      -1.0
19        1         -0.87         -0.5
20        1         -0.87         -0.5
```

Plot

Figure 7: covariate: complex model

Limma Worklow: fit linear model

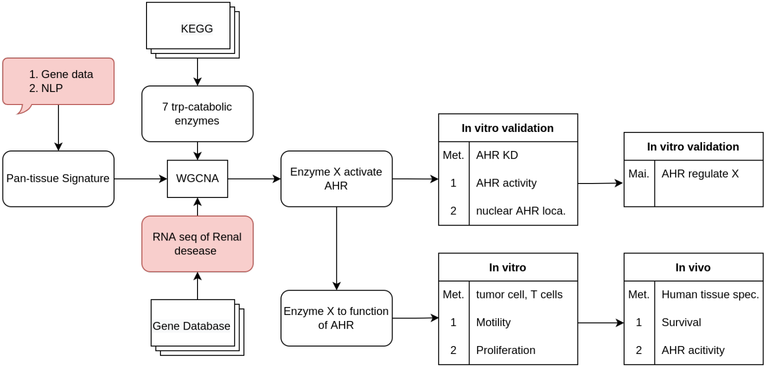
Binary comparison: basal.vs.ml

Table 5: Fitting

ENTREZID	SYMBOL	TXCHROM	logFC	AveExpr
242505	Rasef	chr4	-6.545602	5.117962
12521	Cd82	chr2	-4.699399	7.069340
20661	Sort1	chr3	-4.941593	6.704161
53624	Cldn7	chr11	-5.515495	6.295139
71740	Nectin4	chr1	-5.595622	5.164669
12759	Clu	chr14	-4.697829	8.856284

AHR and Kidney

Analysis route



In vitro validation	
Met.	AHR KD
1	AHR activity
2	nuclear AHR loca.

In vitro validation	
Mai.	AHR regulate X

In vitro	
Met.	tumor cell, T cells
1	Motility
2	Proliferation

In vivo	
Met.	Human tissue spec.
1	Survival
2	AHR acitivity

Figure 8: Route

AHR signature screening

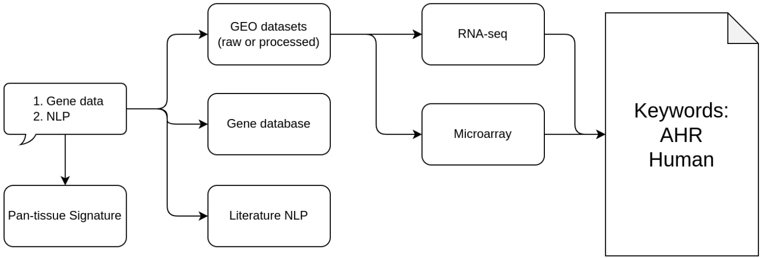


Figure 9: Route-ahr.sig

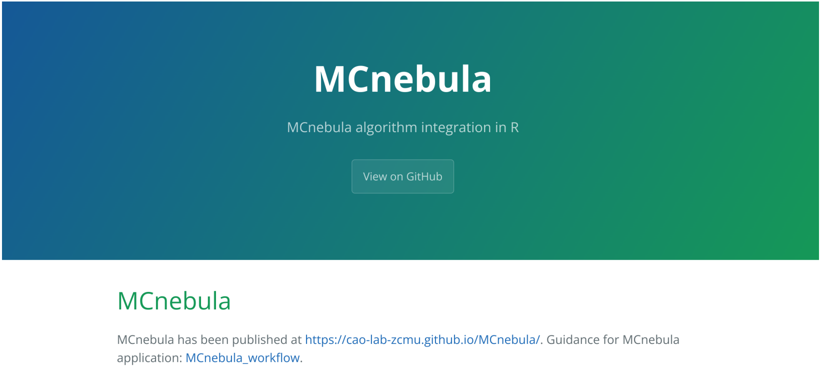
GEO data series for screening AHR signature

Table 6: GSE siries

Accession	Title	Series Type	Taxonomy	Sample Count
GSE18...	Analy...	Expre...	Homo ...	9
GSE18...	AhR a...	Expre...	Homo ...	15
GSE18...	Activ...	Expre...	Homo ...	16
GSE18...	Rutae...	Expre...	Homo ...	6
GSE18...	circR...	Expre...	Homo ...	6
GSE15...	Activ...	Expre...	Homo ...	36
GSE16...	Trans...	Expre...	Homo ...	33

MCnebula in Website

MCnebula mount at:

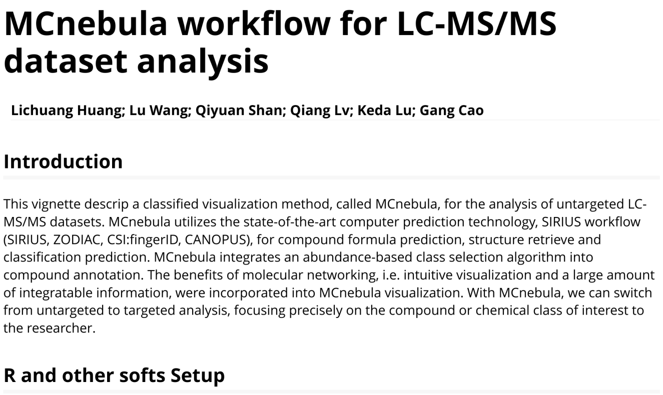


MCnebula

MCnebula has been published at https://cao-lab-zcmu.github.io/MCnebula/. Guidance for MCnebula application: MCnebula_workflow.

Figure 10: Website

Long documentation: vignette



MCnebula workflow for LC-MS/MS dataset analysis

Lichuang Huang; Lu Wang; Qiyuan Shan; Qiang Lv; Keda Lu; Gang Cao

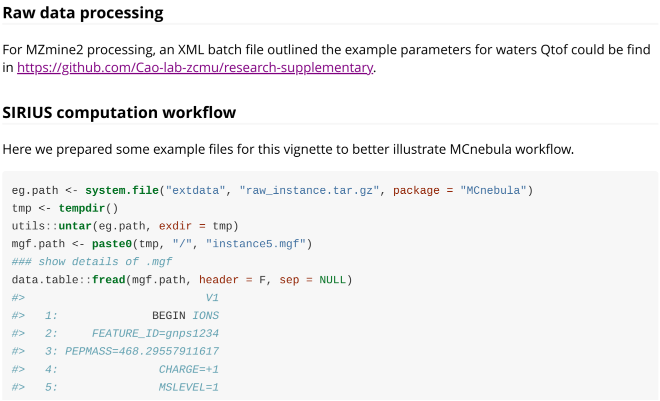
Introduction

This vignette descrip a classified visualization method, called MCnebula, for the analysis of untargeted LC-MS/MS datasets. MCnebula utilizes the state-of-the-art computer prediction technology, SIRIUS workflow (SIRIUS, ZODIAC, CSI:fingerID, CANOPUS), for compound formula prediction, structure retrieve and classification prediction. MCnebula integrates an abundance-based class selection algorithm into compound annotation. The benefits of molecular networking, i.e. intuitive visualization and a large amount of integratable information, were incorporated into MCnebula visualization. With MCnebula, we can switch from untargeted to targeted analysis, focusing precisely on the compound or chemical class of interest to the researcher.

R and other softs Setup

Figure 11: vignette

Long documentation: vignette



Raw data processing

For MZmine2 processing, an XML batch file outlined the example parameters for waters Qtof could be find in https://github.com/Cao-lab-zcmu/research-supplementary.

SIRIUS computation workflow

Here we prepared some example files for this vignette to better illustrate MCnebula workflow.

```
eg.path <- system.file("extdata", "raw_instance.tar.gz", package = "MCnebula")
tmp <- tempdir()
utils::untar(eg.path, exdir = tmp)
mgf.path <- paste0(tmp, "/", "instance5.mgf")
### show details of .mgf
data.table::fread(mgf.path, header = F, sep = NULL)
#>                            V1
#>   1:              BEGIN IONS
#>   2:     FEATURE_ID=gnps1234
#>   3: PEPMASS=468.29557911617
#>   4:               CHARGE=+1
#>   5:               MSLEVEL=1
```

Figure 12: vignette.2

Next Schedule

Gene informatics analysis

- GEO datasets analysis
- . . .