Fitting the Parallel Mixed Model with the `PMM` R-package

Anna Drewek

October 30, 2018

1 Introduction

The Parallel Mixed Model (PMM) approach is suitable for hit selection and cross-comparison of RNAi screens generated in experiments that are performed in parallel under several conditions. For example, we could think of the measurements or readouts from cells under RNAi knock-down, which are infected with several pathogens or which are grown from different cell lines. PMM simultaneously takes into account all the knock-down effects in order to gain statistical power for the hit detection. As a special feature, PMM allows incorporating RNAi weights that can be assigned according to the additional information on the used RNAis or the screening quality. The theory behind PMM is shortly described in the second section (more details can be found in [1]). The third section shows the functionality of the `PMM` R-package by using an RNAi dataset as example.

2 Background

PMM is composed of a linear mixed model and an assessment of the local False Discovery Rate. The linear mixed model consists of a fixed effect for condition and of two random effects for gene g and for gene g within a condition c. We denote the readout or measurement result of the RNAi s silencing the gene g as y_{gcs} for the condition c. The linear mixed model of the PMM is defined as the following linear model

$$Y_{gcs} = \mu_c + a_g + b_{cg} + \beta X_{gcs} + \varepsilon_{gcs}$$

where μ_c is the fixed effect for condition c (typically close to 0 if the data is Z-Scored), a_g is the gene effect overall pathogens, b_{cg} is the gene effect within a pathogen and ε_{gcs} denotes the error term.
The effect of a certain gene g within a condition c is described by the sum of the two random effects:

$$c_{cg} = a_g + b_{cg}.$$

A positive estimated c_{cg} effect means that the RNAi readout for condition c is enhanced if the corresponding gene g is knocked down. A negative effect means that the RNAi readout is reduced. The linear mixed model is estimated by the `lmer` function from the `lme4` R-package.
To distinguish hit genes, PMM provides as second step an estimate of the local False Discovery Rate (FDR). We define the local false discovery rate as

$$\widehat{fdr}(c) = \frac{\widehat{\pi_0}\widehat{f_0}(c)}{\widehat{f(c)}}$$

where π_0 stands for the proportion of true hits, f_0 for the distribution of the readout for all genes that are hits, f_1 for the distribution of readout for all genes that are no hits and $f(c) =$

$\pi_0 f_0(c)+(1-\pi_0)f_1(c)$. The three quantities are separately estimated by using Maximum Likelihood, Poisson regression and moment estimation (for details see [2] and [3]).
Additionally, a sharedness score sh_g is offered for an easier cross-comparison of the results from PMM. The sharedness score indicates if a gene is a hit in only one condition or if the hit appears among all conditions. The sharedness score is a combination of two quantities:

$$sh_g = \frac{1}{2}\left((1 - \text{mean}(fdr_{cg})) + \sum_c (fdr_{cg} < 1)\right)$$

The first part defines the shift away from 1 and the second part describes how many pathogens support the shift.

3 Working Example

The `PMM` R-package contains the following functions:

Function	Description
`pmm`	fits the PMM
`hitheatmap`	visualizes the results of PMM
`sharedness`	computes the sharedness

Moreover, an RNAi dataset on infection with several pathogens is included in the R-package.

```
> library(pmm)
> data(kinome)
> head(kinome)
```

```
  GeneID GeneName condition siRNA   company CellCount InfectionIndex
1     25     ABL1     ADENO     1    Ambion  0.105927       1.127378
2     25     ABL1     ADENO     2    Ambion -0.319963      -1.517203
3     25     ABL1     ADENO     3    Ambion  1.830071       0.151446
4     25     ABL1     ADENO     1 Dharmacon  0.096761      -1.657932
5     25     ABL1     ADENO     2 Dharmacon  0.534097       0.440019
6     25     ABL1     ADENO     3 Dharmacon  0.123893      -0.230555
  weight_library
1              2
2              2
3              2
4              1
5              1
6              1
```

The dataset contains the readouts of 826 kinases knock-down experiments - each targeted by a total of 12 independent siRNAs coming from three manufactures: Ambion (3 siRNAs), Qiagen (4 siRNAs) and Dharmacon (4 siRNAs + 1 pool siRNA). After knock-down the cells were infected with a pathogen, imaged with a microscope and the infection rate, as well as the number of cells were extracted from the microscope images. All experiments were conducted for 8 different pathogens. For example, the normalized number of cells (0.105927) and the normalized infection score (1.127378) from the frist row of the data matrix are readouts of the microscope image from the siRNA experiment where gene *ABL1* was knocked down with siRNA 1 from the manufacturer Ambion (for details see [1]).

3.1 Input to PMM

In order to use `pmm` your data needs to be stored as `data.frame`. Each row should correspond to one independent RNAi experiment. The data frame should have at least the following three columns:

1. gene identifier
2. condition
3. RNAi readout

In our example, the column *GeneID* identifies the genes, the column *condition* corresponds to the pathogens which indicate the different conditions and as siRNA readout serve the columns *InfectionIndex* and *CellCount*.

Note:

1. The data should contain several independent siRNAs (different seeds) measurements per gene and condition.
2. The biological replicates (experimental results with identical siRNA (same seeds) and identical condition) should be averaged.

3.2 Fitting PMM

In order to fit the PMM, take the data frame with your measurements as input and specify the correct column names by using the arguments `gene.col` and `condition.col`. As example, we fit the PMM for the readout *InfectionIndex*.

```
> fit1 <- pmm(kinome, "InfectionIndex", gene.col = "GeneName",
+ condition.col = "condition")
> head(fit1)
```

```
  GeneID   ccg.ADENO fdr.ADENO ccg.BARTONELLA fdr.BARTONELLA ccg.BRUCELLA
1   AAK1  0.20302301         1     0.01690878              1   0.14275979
2   AATK  0.12641832         1    -0.16504408              1  -0.02594678
3   ABL1 -0.12880826         1     0.10988276              1  -0.05980182
4   ABL2 -0.17482037         1    -0.01058879              1  -0.12022731
5  ACVR1 -0.17032261         1    -0.13287511              1  -0.25434025
6 ACVR1B  0.03966076         1    -0.18706962              1  -0.24045426
  fdr.BRUCELLA ccg.LISTERIA fdr.LISTERIA    ccg.RHINO fdr.RHINO ccg.SALMONELLA
1            1   0.31750137    0.9190209 -0.008399173 1.0000000     0.10503464
2            1   0.03525291    1.0000000  0.419511429 0.9367450    -0.02181384
3            1  -0.11524458    1.0000000 -0.225982871 0.9968858    -0.15649114
4            1  -0.25343879    1.0000000  0.021600504 1.0000000     0.11198023
5            1  -0.35657568    0.9236887 -0.314185167 0.9269005    -0.17388909
6            1  -0.03998615    1.0000000 -0.091292436 1.0000000     0.07908277
  fdr.SALMONELLA ccg.SHIGELLA fdr.SHIGELLA ccg.VACCINIA fdr.VACCINIA
1              1   0.06197101            1   0.35902140     0.853158
2              1  -0.16183992            1  -0.08375321     1.000000
3              1  -0.33228704            1  -0.08383964     1.000000
4              1  -0.16282568            1  -0.07334072     1.000000
5              1  -0.41598223            1  -0.18011942     1.000000
6              1  -0.01247689            1  -0.09540113     1.000000
```

The default output of pmm is a matrix that contains the estimated c_{cg} effects for each condition c and gene g, as well as an estimate for the local false discovery rate. A positive estimated c_{cg} effect means that the response was enhanced when the corresponding gene was knocked down. A negative effect means that the response was reduced. Another version of the output giving some more information is also available by using the argument simplify:

```
> fit2 <- pmm(kinome, "InfectionIndex", gene.col = "GeneName",
+ condition.col = "condition", simplify = FALSE)
> class(fit2)

[1] "list"

> names(fit2)

[1] "ccg.matrix" "lmm"        "ccg"

> identical(fit1,fit2[[1]])

[1] TRUE
```

The non-simplified output of pmm is a list of three components. The first component contains the simpilified output, i.e the matrix with the estimated c_{cg} effects and the estimated local false discovery rate, the second component contains the fit of the linear mixed model and the third component contains the estimated a_g and the estimated b_{cg} values. Additional arguments of pmm can be used to add weights to the linear mixed model fit (weight) or change the number of minimal required siRNA replicates for each gene (ignore). Moreover pmm can deal with missing values. Missing values appear, for example, if your data doesn't contain a full set of combinations for conditions and genes, meaning that for each gene not every condition was performed. As an example, we set all measurements of the gene *AAK1* and the pathogen *Adenovirus* to NA. The result of the pmm fit looks as follows:

```
> kinome$InfectionIndex[kinome$GeneName == "AAK1" &
+     kinome$condition == "ADENO"] <- rep(NA,12)
> fit3 <- pmm(kinome ,"InfectionIndex", gene.col = "GeneName")
> head(fit3,3)

  GeneID  ccg.ADENO fdr.ADENO ccg.BARTONELLA fdr.BARTONELLA ccg.BRUCELLA
1   AAK1         NA        NA    0.005334682              1   0.13126340
2   AATK  0.1266126         1   -0.165151096              1  -0.02596045
3   ABL1 -0.1286538         1    0.110046576              1  -0.05973969
  fdr.BRUCELLA ccg.LISTERIA fdr.LISTERIA   ccg.RHINO fdr.RHINO ccg.SALMONELLA
1            1   0.30611328    0.9469859 -0.01998922 1.0000000     0.09394962
2            1   0.03528034    1.0000000  0.41979780 0.9348462    -0.02182476
3            1  -0.11521569    1.0000000 -0.22601982 0.9970148    -0.15648729
  fdr.SALMONELLA ccg.SHIGELLA fdr.SHIGELLA ccg.VACCINIA fdr.VACCINIA
1              1   0.05042401            1   0.34808772    0.8799419
2              1  -0.16194550            1  -0.08380437    1.0000000
3              1  -0.33238810            1  -0.08379270    1.0000000
```

The output shows now NA for the removed combination.

3.3 Visualization of Results

The results of PMM can be illustrated by a heat map using the function `hitheatmap`.

```
> hitheatmap(fit1, threshold = 0.2)
```

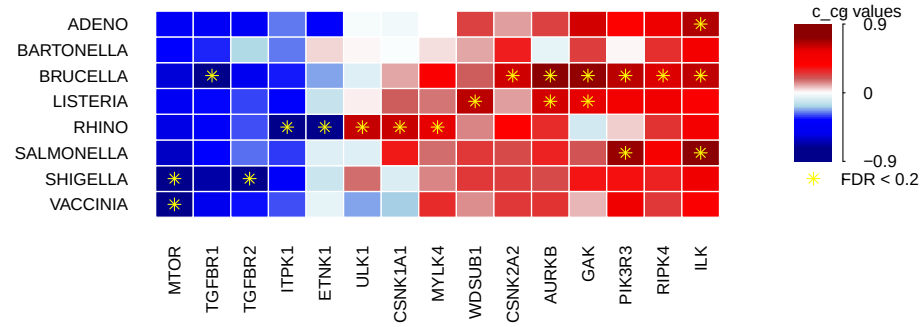

Figure 1: Visualization of PMM results for an easier cross-comparison between the different conditions.

Red color indicates a positive estimated c_{cg} effect, blue color a negative estimated c_{cg} effect. The darker the color, the stronger is the estimated c_{cg} effect. The heat map contains only the genes for which the local false discovery rate is below a given threshold for at least one condition. The yellow star indicates the significant genes. The plot can be modified by passing further arguments to the `plot` and the `par` function

```
> hitheatmap(fit1, threshold = 0.4, cex.main = 2,
+ main = "My modified plot", col.main = "white",
+ col.axis = "white", cex.axis = 0.8, bg = "black", mar = c(6,8,4,6))
```

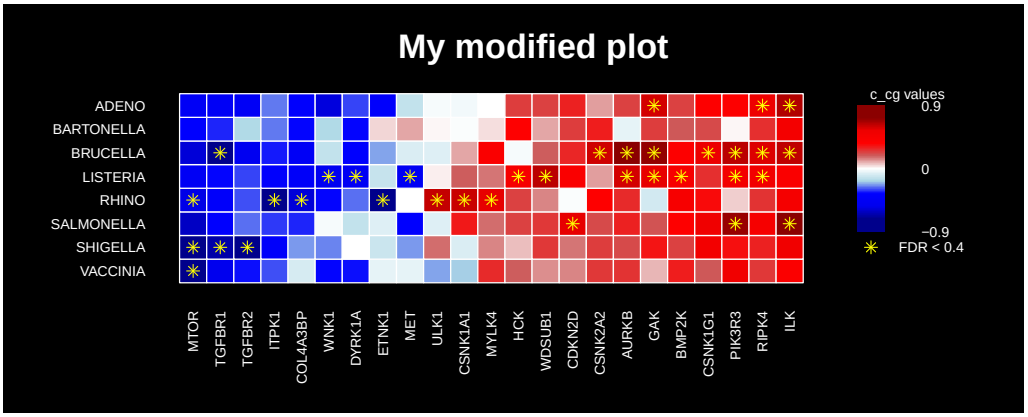

Figure 2: Modified Heat map.

Missing combinations are plotted in white color and marked by NA. Use the argument `na.omit = na.omit` to plot only complete combinations.

3.4 Adding Sharedness Score

The sharedness score returns a value between 0 and 1 for each gene. Score 0 indicates that a gene is not shared among the condition and score 1 that the gene is significant among all conditions. The sharedness score is only computed for genes that pass a given `threshold`.

```
> sh <- sharedness(fit1, threshold = 0.2)
> sh[order(sh$Sharedness),]
```

```
    GeneID Sharedness
14    ULK1  0.1214748
4    ETNK1  0.1249985
2  CSNK1A1  0.1866004
7    ITPK1  0.1998552
3  CSNK2A2  0.2517173
15  WDSUB1  0.2525156
13  TGFBR2  0.3190709
9    MYLK4  0.3270544
5      GAK  0.4207531
1    AURKB  0.4424351
11   RIPK4  0.4762744
10  PIK3R3  0.5137591
12  TGFBR1  0.5388681
8     MTOR  0.6198474
6      ILK  0.6240870
```

The sharedness score can also be visualized within the `hitheatmap`. Use the argument `sharedness.score = TRUE` to add a row for the sharedness score. The darker the green color, the stronger is the sharedness among the conditions.

```
> hitheatmap(fit1, sharedness.score = TRUE, main = "My hits found by PMM")
```

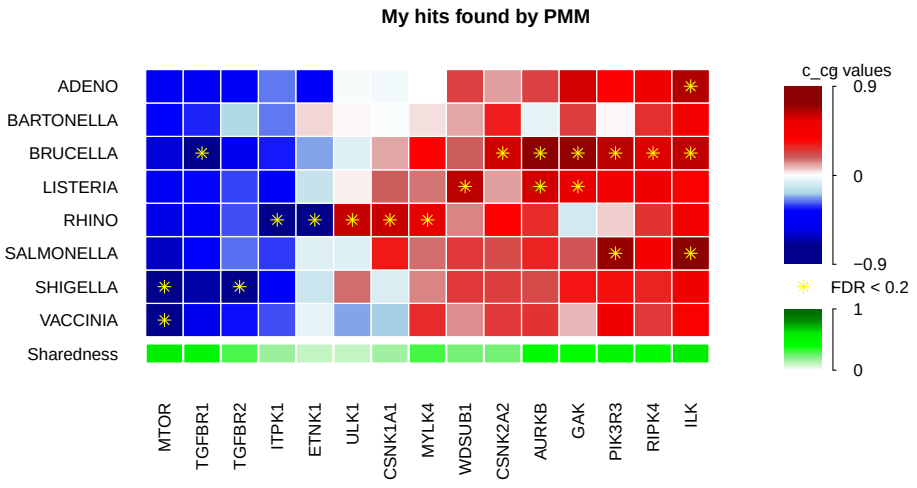

Figure 3: Visualization of PMM results with sharedness score.

References

[1] Rämö, P., Drewek, A., Arrieumerlou, C., Beerenwinkel, N., Ben-Tekaya H., Cardel, B., Casanova, A., Conde-Alvarez. R., Cossart, P., Csucs, G., Eicher, S., Emmenlauer, M. Greber, U., Hardt, W.-D., Helenius, A., Kasper, C., Kaufmann, A., Kreibich, S., Kübacher, A., Kunszt, P., Low, S.H., Mercer, J., Mudrak, D., Muntwiler, S., Pelkmans, L., Pizarro-Cerda, J., Podvinec, M., Pujadas, E., Rinn, B., Rouilly, V., Schmich F., Siebourg, J., Snijder, B., Stebler, M., Studer, G., Szczurek, E., Truttmann, M., von Mering, C., Vonderheit, A., Yakimovich, A., Bühlmann, P. and Dehio, C. *Simultaneous analysis of large-scale RNAi screens for pathogen entry*. BMC Genomics, 15(1162):1471-2164, 2014.

[2] Efron, B. *Size, Power and False Discovery Rates*. The Annals of Statistics, 35(4):1351-1377, 2007.

[3] Efron, B. *Empirical Bayes Methods for Estimation Testing and Prediction*. Cambridge University Press, 2014.