tigaR: Integrative significance analysis of temporal differential gene expression induced by genomic abnormalities

V Miok, **SM Wilting**, **MA van de Wiel**, **A Jaspers**, **P Noort**, **R Brakenhoff**, **P Snijders**, **R Steenbergen** and **WN van Wieringen**

January 14, 2016

1 Introduction

Package tigaR is developed to perform integrative differential gene expression analysis of the time-course omics data. Package allow to study differential gene expression in microarray and RNA-seq data. For the microarray(continious) data, package may be used to two types of integrative analysis:

- Identification of the features with temporal differential gene expression
- Identification of the features with temporal differential gene expression driven by DNA copy number

On the other hand, for RNA-seq(count) data, package is not designed for integrative analysis, thus only features with temporal differential gene expression can be identified. In this vignette, on a microarray and sequencing data sets, we intend to illustrate how the analysis can be performed.

2 Microarray data

The tigaR package is demonstrated on data from HPV-induced transformation experiment(Miok et al., 2014). HPV-immortalized in vitro cell line model system comprise four cell lines, where two cell lines are affected with HPV16 and two with HPV18. Employing microarray technique cell lines are measured for DNA copy number and mRNA gene expression at 8 different time points (Miok et al., 2014). Data comprise 2202 features, corresponding to third chromosome. The pre-processing of the data is described in Miok el al. 2014. After preprocessing matching procedure Wieringen et al. 2012 is applied to matches probes from these molecular levels on the basis of their genomic location.

Data are loaded in the following way:

```
> library(tigaR)
> data(cervixGE)
> data(cervixCN)
```

2.1 Design matrices

Before we start with integrative differential expression analysis we need to create design matrices. First we specify the number of the time points, groups (cell lines), knots and degree of the splines, as well as whether same or different spline per cell to be used (Miok et al., 2014).

- `DiffSpl` = `TRUE` or `FALSE`(default)

If `DiffSpl` = `TRUE` function create design matrix with different spline per cell line, otherwise is the same spline par cell line.

```
> numKnot = 3
> deg = 3
> numtp = 8
> numgroup = 4
```

Function `GetDesign` using this parameters create design matrices for the covariates, which are later required in the functions for fitting and testing.

```
> desMat = GetDesign(numKnot, deg, numtp, numgroup, DiffSpl=FALSE)
> groupfac <- desMat$groupfac
> timefac <- desMat$timefac
> DsgGroup <- desMat$DsgGroup
> ZSpline <- desMat$ZSpline
```

2.2 Model parameters estimation

In order to identify temporal differential gene expression induced by abnormalities in the DNA copy number it require fit three models. For each of this model shrunken dispersion related parameters need to be estimated. First, is the full model, which comprise group(cell line), copy number and time effect. Second is full model without copy number effect. Third is full models without time effect. Employing empirical Bayes shrinkage we estimated dispersion related parameters of the group, copy number and time effect. In addition, the Gaussian error standard deviation is shrunken by default.

First we fit the full model. Formula used by INLA specified the fixed and random effect of the model.

```
> form <- y ~ 1 + x + groupfac + f(timefac, model="z", Z=ZSpline,
                        prior="loggamma", param=c(1,0.00001))
```

Function `ShrinkPar` shrink the dispersion related parameters.

```
> shrinksimul <- ShrinkPar(form=form, dat=cervixGE, dat1=cervixCN,
                  shrinkfixed="x", shrinkrandom="timefac", ncpus=2,
                  orthogonal=FALSE, shrink=TRUE)
```

Fit the model with default settings:

```
> full <- FitIntAllShrink(form, dat=GEdata, dat1=CNdata, fams="gaussian",
  shrinksimul=shrinksimul, ncpus=2, orthogonal=FALSE, shrink=FALSE)
```

The followings arguments of the functions for shrinkage and fitting of the model, offer flexibility for modeling:

- fams = "gaussian"(default), "poisson", "zip"(zero-inflated Poisson), "nb"(negative binomial) or "zinb"(zero-inflated negative binomial).
  Specify which likelihood to be used.
- ncpus = 2(default).
  Number of cpus to use for parallel computations.
- orthogonal = TRUE or FALSE(default)
  Specify whether the design matrix of the random effect is orthogonalized onto copy number effect.
- multivar = TRUE or FALSE(default)
  Specify whether the multivariate spatial priors are imposed onto copy number effects.

Estimation of the shrunken dispersion related parameters and fitting of the model without copy number effect:

```
> form <- y ~ 1 + groupfac + f(timefac, model="z", Z=ZSpline,
          prior="loggamma", param=c(1,0.00001))
> shrinksimul <- ShrinkPar(form=form, dat=cervixGE, dat1=NULL,
          shrinkrandom="timefac", ncpus=2, orthogonal=FALSE, shrink=TRUE)
> noCN <- FitIntAllShrink(form, dat=GEdata, dat1=NULL, fams="gaussian",
        shrinksimul=shrinksimul, ncpus=2, orthogonal=TRUE, shrink=FALSE)
```

Estimation of the shrunken dispersion related parameters and fitting of the model without time(spline) effect:

```
> form <- y ~ 1 + x + groupfac
> shrinksimul <- ShrinkPar(form=form, dat=cervixGE, dat1=cervixCN,
                ncpus=2, orthogonal=FALSE, shrink=TRUE)
> noTime <- FitIntAllShrink(form, dat=GEdata, dat1=CNdata, fams="gaussian",
 shrinksimul=shrinksimul, ncpus=2, orthogonal=FALSE, shrink=FALSE)
```

2.3 Hypothesis testing

With parameter estimates at hand, we describe how relevant hypothesis may be tested. We are interested to answer two questions: i) is there features with temporal differential gene expression, and ii) is there features affected with abnormalities in DNA copy number. In order to answer this questions likelihood ratio test is employed. The former question can be answered comparing the full model and model without time effect using function TDGE. On the other hand, identification of the genes affected with DNA copy number is obtained comparing full and model without copy number effect using function TDGEindCN. The following argument of the function for testing offer flexibility for testing:

- annotation = matrix of two columns which represent gene and probe name.
- multivar = TRUE or FALSE(default)
  Specify whether the multivariate spatial priors are imposed onto copy number effects.
- sortby = 5 Specify the column which is used to sort the data, in this case is adjusted p-value.

Test for temporal differential expression:

```
> Result <- TDGE(dat=cervixGE, dat1=NULL, FitAlt=full[[1]],
  FitNull=noTime[[1]], DsgGroup=DsgGroup, DsgSpl=ZSpline,
  annotation=annotation, multivar=FALSE, sortby=5)
```

Test for temporal differential expression affected with abnormalities in DNA copy number:

```
> Result <- TDGEindCN(FitAlt=full[[1]], FitNull=noCN[[1]], dat=cervixGE,
   annotation=annotation, multivar=FALSE)
```

3 Sequencing data

We use the head-and-neck cancer data to illustrate applicability of our tigaR package on time-course sequencing data (Miok et al. 2014). Data set comprise transcript levels of the 2×6 samples. In Miok et al., 2014 experiment and data pre-processing are described.

Data are loaded in the following way:

```
> library(tigaR)
> data(seqRNA)
```

3.1 Design matrices

Before we start with integrative differential expression analysis we need to create design matrices. First we specify the number of the time points, groups (cell lines), knots and degree of the splines, as well as whether same or different spline per cell to be used (Miok et al., 2014).

Initial specification of the number of the time points, groups, knots and degree of splines, as well as type of spline is required.

- `DiffSpl` = `TRUE` or `FALSE`(default)

If `DiffSpl` = `TRUE` function create design matrix with different spline per cell line, otherwise is the same spline par cell line.

```
> numKnot = 3
> deg = 3
> numtp = 6
> numgroup = 2
```

Function `GetDesign` using this parameters create design matrices for the covariates, which are later required in the functions for fitting and testing.

```
> desMat = GetDesign(numKnot, deg, numtp, numgroup, DiffSpl=FALSE)
> groupfac <- desMat$groupfac
> timefac <- desMat$timefac
> DsgGroup <- desMat$DsgGroup
> ZSpline <- desMat$ZSpline
```

3.2 Model parameters estimation

Identification of the temporal differential gene expression require to fit two models. First, is the full model which comprise group(cell line) and time effect. Second is full models without time effect. In order to get more stable estimates empirical Bayes shrinkage of group, time and Gaussian error is performed.

First we fit the full model. Formula used by INLA specified the fixed and random effect of the model.

```
> form <- y ~ 1 + groupfac + f(timefac, model="z", Z=ZSpline,
                        prior="loggamma", param=c(1,0.00001))
```

Function `ShrinkSeq` shrink the dispersion related parameters.

```
> shrinksimulA <- ShrinkSeq(form=form, dat=Datax, fams="zinb",
                    shrinkrandom="timefac",  ncpus=2)
```

Fit the model with default settings:

```
> full <- FitIntAllShrink(form, dat=Datax, dat1=NULL, fams="zinb",
          timefac=timefac, groupfac=groupfac, ZSpline=ZSpline,
          shrinksimul=shrinksimulA, ncpus=2)
```

The followings arguments of the functions for shrinkage and fitting of the model, offer flexibility for modeling:

- `fams` = `"gaussian"`(default), `"poisson"`, `"zip"`(zero-inflated Poisson), `"nb"`(negative binomial) or `"zinb"`(zero-inflated negative binomial).
  Specify which likelihood to be used.
- `ncpus` = 2(default).
  Number of cpus to use for parallel computations.

Estimation of the shrunken dispersion related parameters and fitting of the model without time(spline) effect:

```
> form <- y ~ 1 + groupfac
> shrinksimulN <- ShrinkSeq(form=form, dat=Datax, fams="zinb", ncpus=2)
> noTime <- FitIntAllShrink(forms=form, dat=Datax, dat1=NULL, fams="zinb",
            timefac=timefac, groupfac=groupfac, ZSpline=ZSpline,
            shrinksimul=shrinksimulN, ncpus=2)
```

3.3 Hypothesis testing

Employing likelihood ratio test full model is compared against model without time effect in order to determine features with temporal differential gene expression. See the Miok et al., 2014 for more details on the test.

The following argument of the function for testing offer flexibility for testing:

- `annotation` = matrix of two columns which represent gene and probe name.
- `sortby` = 5 Sepcify the column which is used to sort the data, in this case is adjusted p-value.

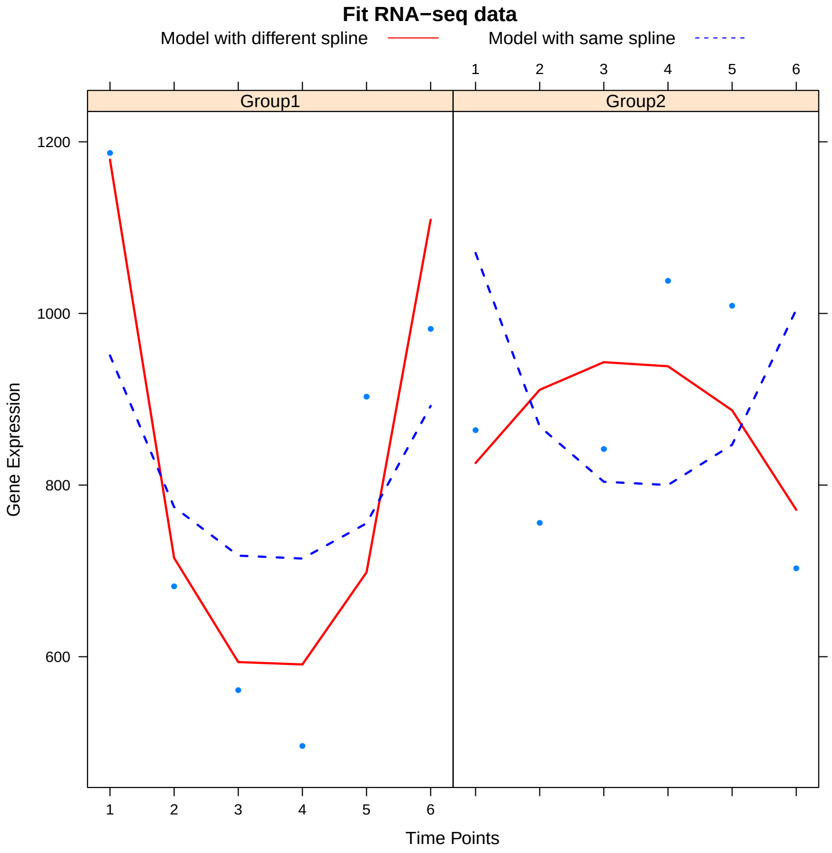
Test for temporal differential expression:

```
> Result <- TDGE(dat=Datax, dat1=NULL, FitAlt=full[[1]], FitNull=noTime[[1]],
      DsgGroup=DsgGroup, DsgSpl=ZSpline, annotation=rownames(Datax),
      sortby=5, numtp=numtp)
```

4 Plotting

Plot of the RNA-seq counts for the model with same and different splines:

```
> label <- c(rep("group1",8), rep("group2",8))
> k=123  # select the feature
> Plot_inlafit(fit=full[[1]][[k]], fit1=noTime[[1]][[k]], numtp,
        numgroup, label, Around0=FALSE, lattice=TRUE)
```


Fit RNA−seq data
Model with different spline
Model with same spline
Group1
Group2
Gene Expression
Time Points