SegReg: Breakpoint analysis of time course expression data

Ning Leng, Ron Stewart

Contents

Overview

SegReg is a R package that can be used to perform breakpoint analysis on Microarrays or RNA-seq expression data with ordered conditions (e.g. time course, spatial course). For each gene or other features, SegReg estimates the optimal number of breakpoints as well as the breakpoints by fitting a set of segmented regression models. The top dynamic genes are then identified by taking genes that can be well profiled by its gene-specific segmented regression model. SegReg also implements functions to visualize the dynamic genes and their trends, to order dynamic genes by their trends, and to compute breakpoint distribution at different time points (e.g. detect time points with a large number of expression changes).

The model

To illustrate SegReg, here we use time course gene expression data as an example. Note SegReg may also be applied to other types of features (e.g. isoform or exon expression) and/or other types of experiments with ordered conditions (e.g. spatial course).

Denote the normalized gene expression of gene g and sample s is $X_{g,s}$. Denote the total number of genes as G and the total number of samples as S. For each gene, SegReg fits segmented regression models with varying numbers of breakpoints from 1 to n_k. In which n_k defaults to 3 but can also be specified by the user. The model with k breakpoints can then be written as:

$$M_g^k : X_g \sim \beta_0^k + \beta_1^k * I\{s : s \geq 1, s \leq b_{g,1}^k\} * s + \beta_2^k * I\{s : s \geq b_{g,1}^k + 1, s \leq b_{g,2}^k\} * (s - b_{g,1}^k) +, ...,$$

$$+\beta_{k+1}^k * I\{s : s \geq b_{g,k}^k + 1, s \leq S\} * (s - b_{g,k}^k)$$

For each k, the segemented regression estimates k breakpoints ($b_{g,1}^k$, $b_{g,2}^k$, ... $b_{g,k}^k$) between 1 and S. The segmented regression also estimate $k + 2$ βs. In which β_0^k indicates the intercept, and the other βs indicate slopes for the $k + 1$ segements separated by the k breakpoints. We denote the adjusted R^2 for this model as r_g^k.

For a given gene, among the models with varying k, SegeReg picks the optimal number of breakpoints for this gene by comparing the adjusted R^2s:

$$\tilde{k_g} = \text{argmax}_{k=1,...,n_k}(r_g^k)$$

To avoid overfitting, the optimal number of breakpoints will be set as $\tilde{k_g} = \tilde{k_g} - 1$ if any of the following happens: at least of one segements having less than c_{num} samples, or $r_g^{\tilde{k}} - r_g^{\tilde{k}-1} < c_{diff}$. The thresholds c_{num} and c_{diff} can be specified by the user; defaults are 5 and 0.1, respectively.

Then the gene specific adjusted R^2 and breakpoint estimates are then obtained from this optimal model: $r_g = r_g^{\tilde{k_g}}$; $(\beta_{g,0}, ..., \beta_{g,\tilde{k_g}+1}) = (\beta_{g,0}^{\tilde{k_g}}, ..., \beta_{g,\tilde{k_g}+1}^{\tilde{k_g}})$ and $(b_{g,1}, ..., b_{g,\tilde{k_g}}) = (b_{g,1}^{\tilde{k_g}}, ..., b_{g,\tilde{k_g}}^{\tilde{k_g}})$. Among all genes, the top dynamic genes are defined as those whose optimal model has high adjusted R^2s.

To compute the breakpoint distribution over the time course, SegReg calculates:

$$N_s = \sum_{g=1,...,G} \sum_{j=1,...,\tilde{k_g}} I\{b_{g,j} = s\}$$

The time points with hign N_s might be considered as time points with a large amount of expression changes.

SegReg also outputs fitted trend of each gene. For samples between the j^{th} and $j + 1^{th}$ breakpoint for a given gene, if the t statistic of $\beta_{g,j+1}$ has p value greater than c_{pval}, the trend of this segment will be defined as no change. Otherwise the trend of this segment will be defined as up/down based on the coefficient of $\beta_{g,j+1}$. The c_pval defaults to 0.1, but can also be specified by the user.

Installation

Install via GitHub

The SegReg package can be installed using functions in the devtools package.

To install, type the following codes in R:

```
install.packages("devtools")
```

```
library(devtools)
```

```
install_github("lengning/SegReg/package/SegReg")
```

Install locally

Install packages segmented and gplots:

```
install.packages(c("segmented","gplots"))
library("segmented")
library("gplots")
```

Download the SegReg package from:

https://github.com/lengning/SegReg/tree/master/package

And install the package locally.

Load the package

To load the SegReg package:

```
library(SegReg)
```

Analysis

Input

The input data should be a $G - by - S$ matrix containing the expression values for each gene and each sample, where G is the number of genes and S is the number of samples. The samples should be sorted following the time course order. These values should exhibit expression data after normalization across samples. For example, for RNA-seq data, the raw counts may be normalized using MedianNorm and GetNormalizedMat() function in EBSeq. More details can be found in the EBSeq vignette:

http://www.bioconductor.org/packages/devel/bioc/vignettes/EBSeq/inst/doc/EBSeq_Vignette.pdf

The object SegRegExData is a simulated data matrix containing 50 rows of genes and 40 columns of samples.

```
data(SegRegExData)
str(SegRegExData)
```

```
##  num [1:50, 1:40] 240 199 198 239 202 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:50] "g1" "g2" "g3" "g4" ...
##   ..$ : chr [1:40] "s1" "s2" "s3" "s4" ...
```

Run segmented regressions

The segreg() function can be used to run gene specific segmented regressions. Here we want to only consider up to 2 breakpoints for each gene. To do so we may specify maxk=2:

```
res <- segreg(SegRegExData, maxk=2)
res.top <- topsegreg(res)
# default adjusted R square cutoff is 0.5
res.top$radj
```

```
##        g3        g1       g28       g20       g15        g2       g10
## 0.9787382 0.9775005 0.9751380 0.9739715 0.9729747 0.9710139 0.9705118
##       g23       g14        g8        g5       g24       g17       g12
## 0.9701402 0.9694164 0.9691341 0.9689555 0.9656732 0.9652141 0.9644343
##       g29       g16       g22       g18       g25       g11       g30
## 0.9632348 0.9630272 0.9627092 0.9626837 0.9611528 0.9600736 0.9597989
##       g26        g7        g4        g9       g21        g6       g19
## 0.9572072 0.9529077 0.9420853 0.9377311 0.9304116 0.9291045 0.9259893
##       g27       g13
## 0.9183375 0.8576471
```

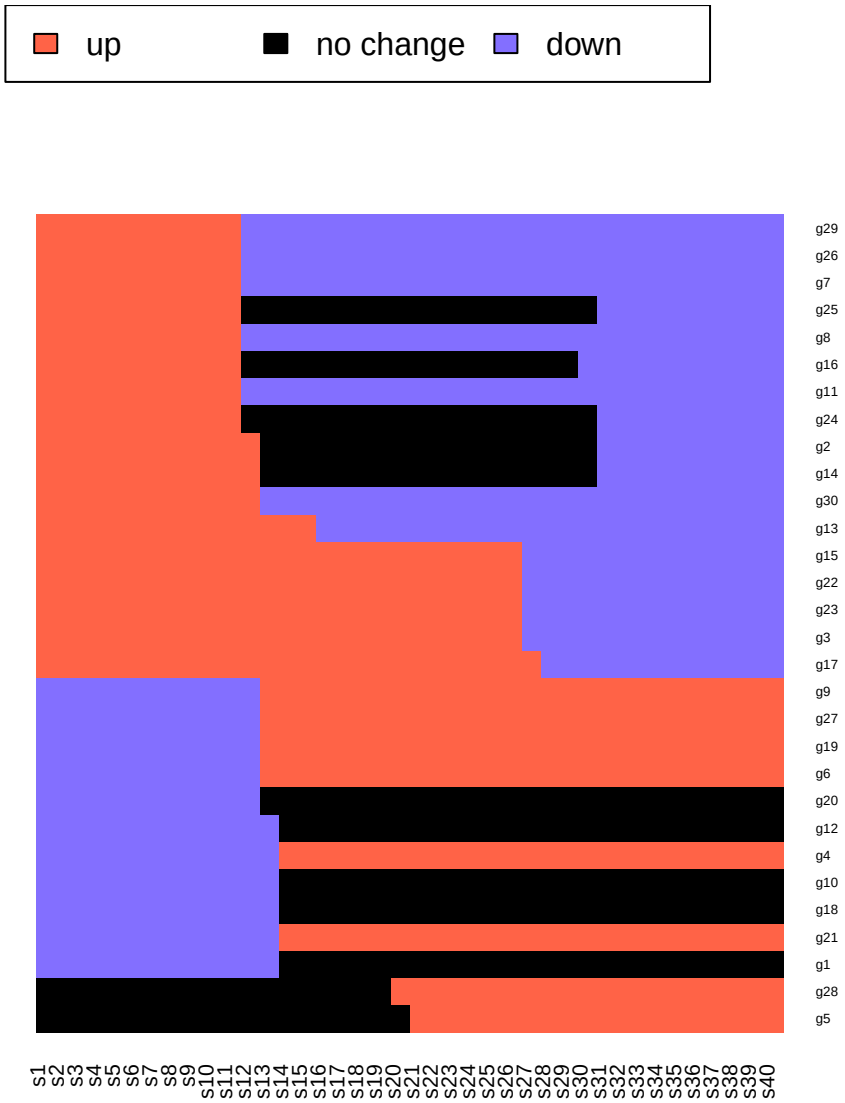
The topsegreg() function may be used to exract top dynamic genes. By default, topsegreg() will extract genes whose adjusted R^2 r_g is greater or equal to 0.5. To change this threshold, a user may specify the r.cut parameter in topsegreg() function. res.top$radj gives r_g of the top dynamic genes, sorted decreasingly by r_g.

By default the segreg() function only consider genes whose mean expression is greater than 10. To use another threshold, a user may specify the parameter meancut in segreg() function.

Visualize trends of the top dynamic genes

res.top$id.sign gives trend specification of the top genes. Function trendheatmap() can be used to display these trends:

```
res.trend <- trendheatmap(res.top)
```



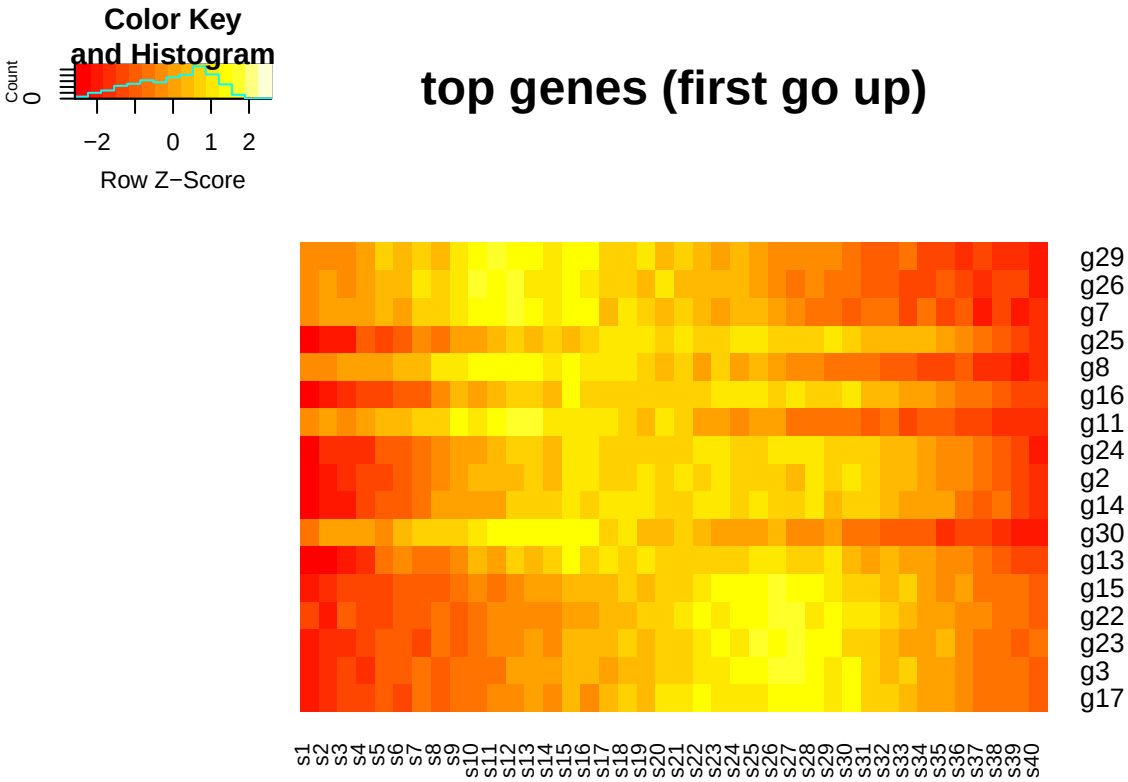

```
str(res.trend)
```

```
## List of 3
##  $ firstup      : Named num [1:17] 11.4 11.5 11.6 11.6 11.6 ...
##   ..- attr(*, "names")= chr [1:17] "g29" "g26" "g7" "g25" ...
##  $ firstdown    : Named num [1:11] 12.1 12.6 12.6 12.7 12.8 ...
##   ..- attr(*, "names")= chr [1:11] "g9" "g27" "g19" "g6" ...
##  $ firstnochange: Named num [1:2] 19 20.4
##   ..- attr(*, "names")= chr [1:2] "g28" "g5"
```

The trendheatmap() function classify the top dynamic genes into three groups: start with up, start with down and start with no change. Within each group, genes are sorted by the position of the first breakpoint.

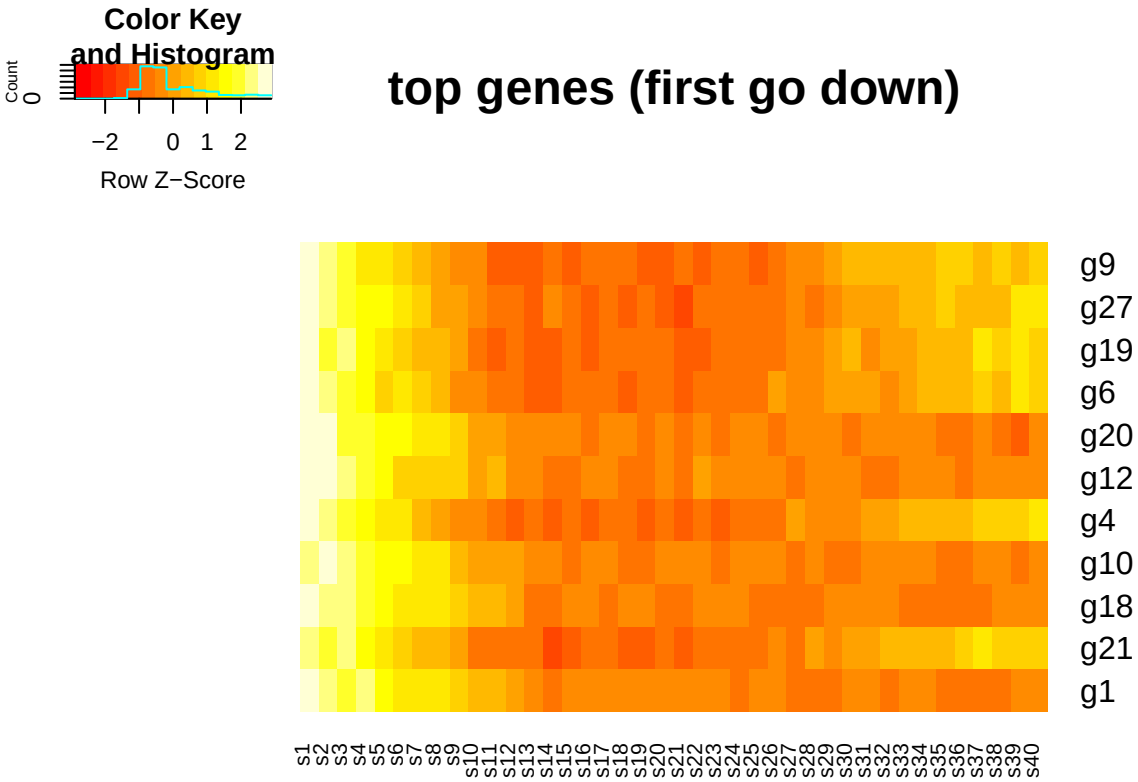
To generate expression heatmap of the first group of genes (first go up):

```
heatmap.2(SegRegExData[names(res.trend$firstup),],trace="none", Rowv=F,Colv=F,
                    scale="row", main="top genes (first go up)")
```



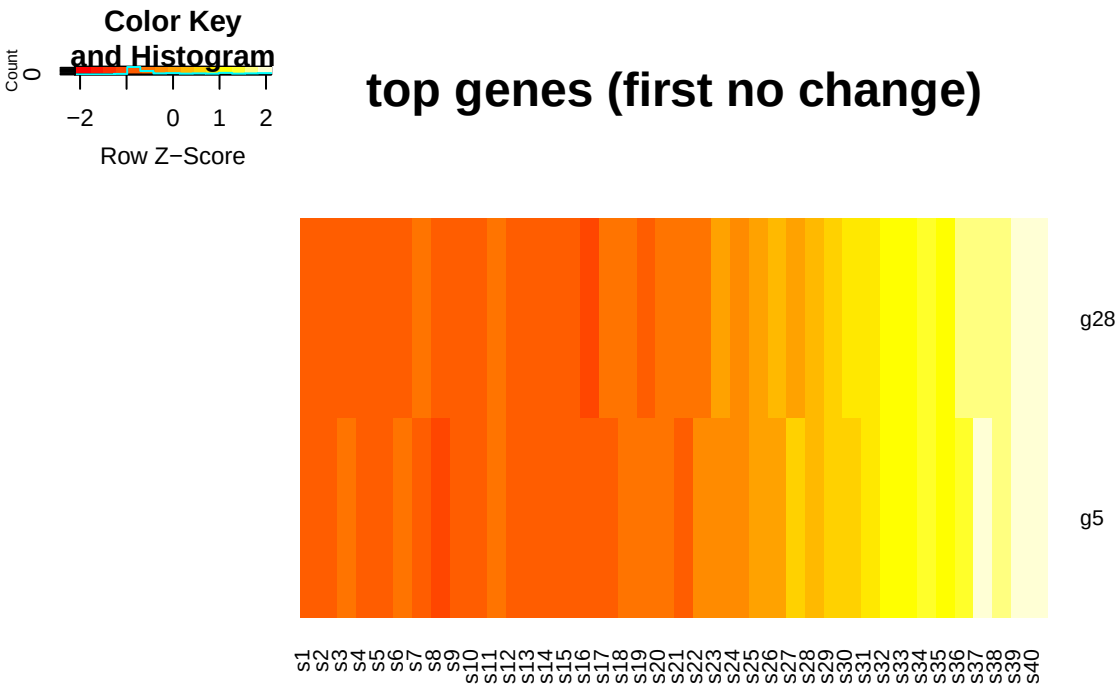
Similarly, to generate expression heatmap of the second group of genes (first go down):

```
heatmap.2(SegRegExData[names(res.trend$firstdown),],trace="none", Rowv=F,Colv=F,
                              scale="row", main="top genes (first go down)")
```



To generate expression heatmap of the second group of genes (first no change):

```
heatmap.2(SegRegExData[names(res.trend$firstnochange),],trace="none", Rowv=F,Colv=F,
                                        scale="row", main="top genes (first no change)",
                                                    cexRow=.8)
```

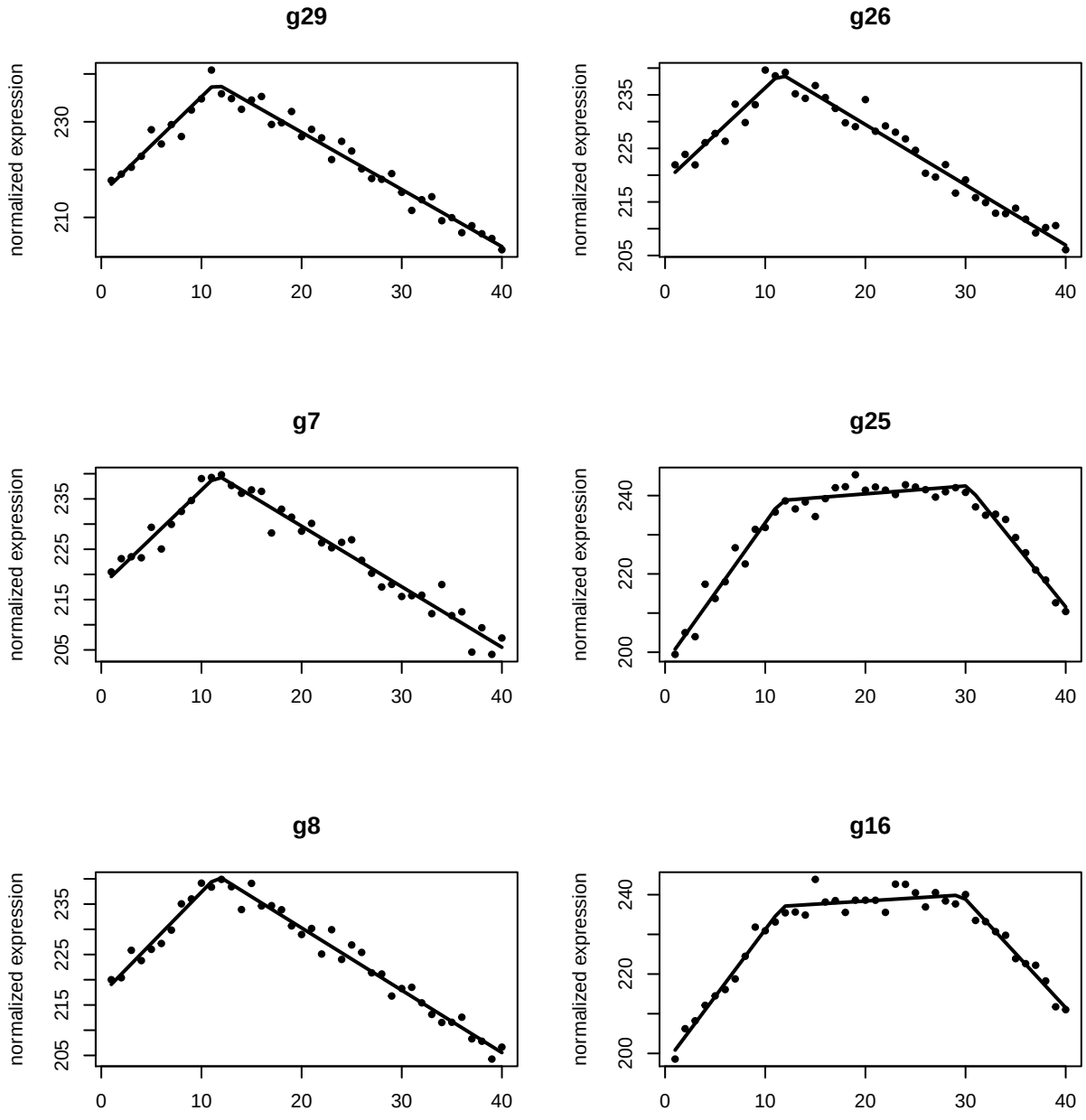


Visualize individual genes

The plotmarker() function may be used to plot expression of individual genes and the fitted lines.

For example, to plot the top 6 genes in the first group of genes (first go up):

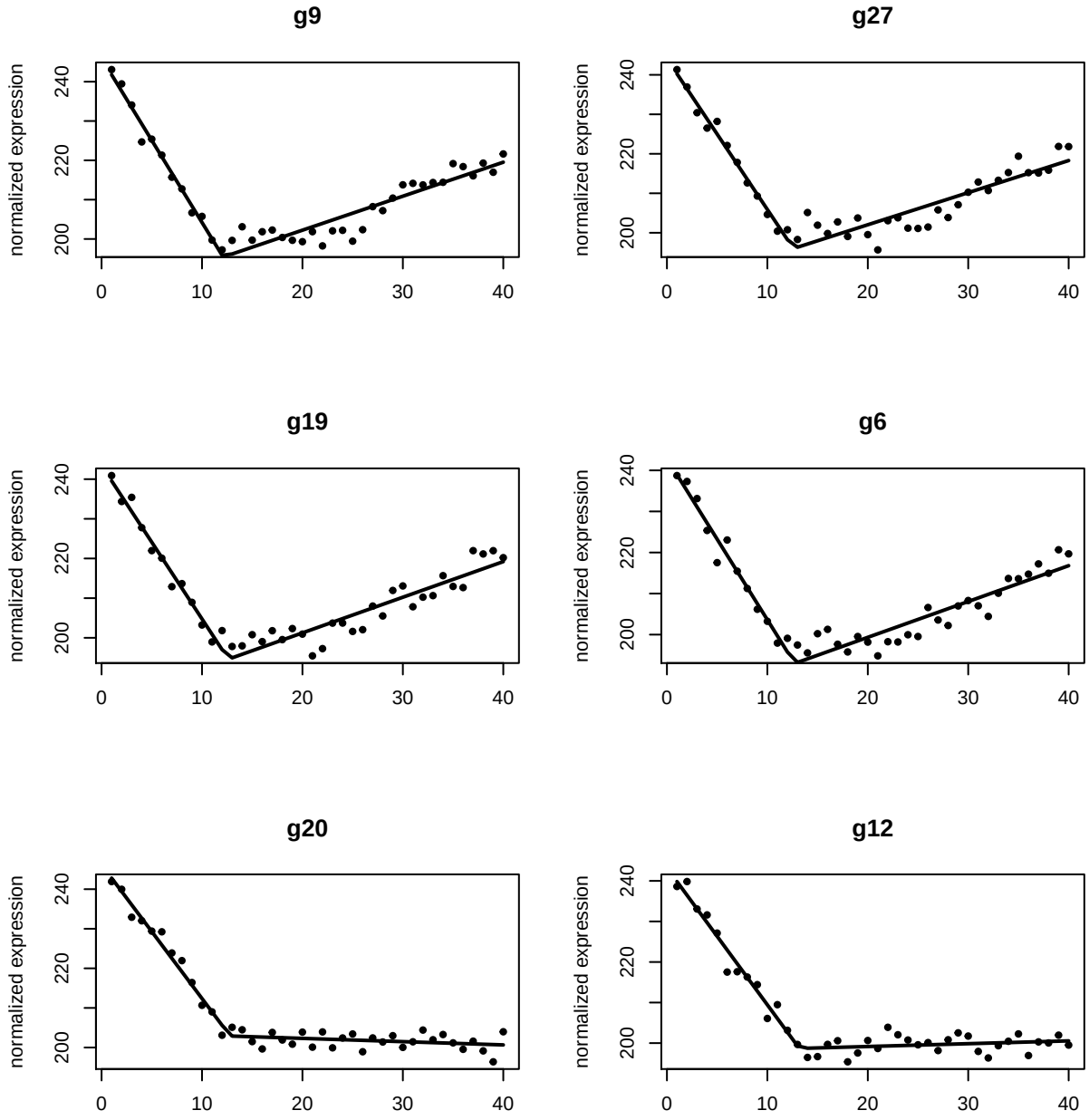
```
plot1 <- plotmarker(SegRegExData,listname=names(res.trend$firstup)[1:6],fittedres=res)
```



The input of function plotmarker() requires the expression data and a list of genes of interest. The parameter fittedres in function plotmarker() takes segreg() fitted results. If it is not specified, the function plotmarker() will run SegReg model on the genes of interest before plotting. Specifying fitted results obtained from previous steps will save time by avoiding fitting the models again.

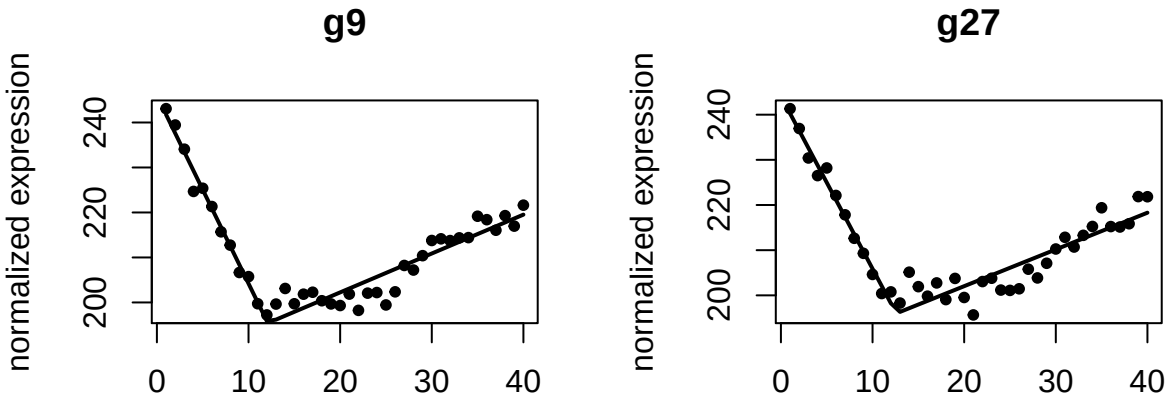
Similarly, to plot the top 6 genes in the second group of genes (first go down):

```
plot2 <- plotmarker(SegRegExData,listname=names(res.trend$firstdown)[1:6],
                                        fittedres=res)
```



To plot the 2 genes in the third group of genes (first no change):

```
plot2 <- plotmarker(SegRegExData,listname=names(res.trend$firstdown)[1:2],
                                        fittedres=res,par.param=c(1,2))
```



Gene specific estimates

For a given gene of interest, its estimated parameters can be obtained by (using g2 as an example):

```
print(res.top$bp["g2"]) # break points
```

```
## $g2
## psi1.t.use psi2.t.use
##   12.47356   30.14908
```

```
print(res.top$radj["g2"]) # adjusted r square
```

```
##        g2
## 0.9710139
```

```
print(res.top$slp["g2"]) # fitted slopes of the segments
```

```
## $g2
##  slope1  slope2  slope3
##  3.3110  0.0607 -2.9730
```

```
print(res.top$slp.pval["g2"]) # p value of each the segment
```

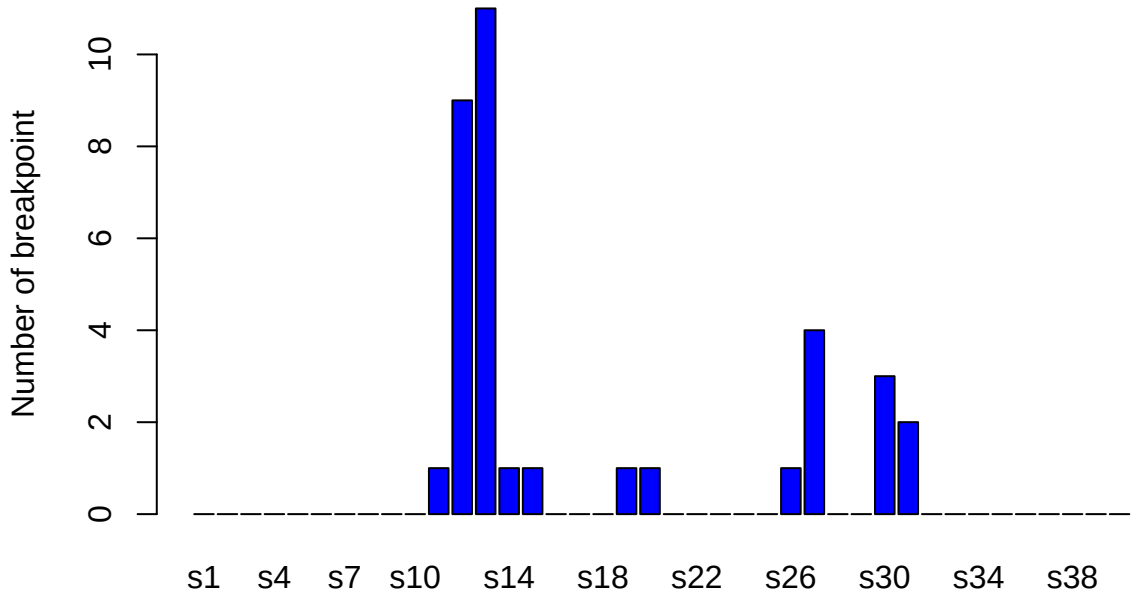
```
## $g2
##     slope1     slope2     slope3
## 0.01669386 0.31815050 0.02445599
```

The above printouts show that for gene g2, the optimal number of breakpoints is 2. Two estimated breakpoints are close to s12 and s30. The fitted slopes for the 3 segements are 3.31, 0.06 and -2.97.

Breakpoint distribution over the time course

To calculate number of breakpoints over the time course:

```
res.bp <- bpdist(res.top)
barplot(res.bp, ylab="Number of breakpoint", col="blue")
```



The bar plot indicates that many genes have breakpoint around s12 and s13.

More advanced options

In segreg() function, the thresholds c_{num}, c_{diff} and c_{pval} can be specified via parameters min.num.in.seg, cutdiff and pvalcut.

SessionInfo

```
print(sessionInfo())
```

```
## R version 3.2.1 (2015-06-18)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.5 (Yosemite)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] SegReg_0.0.1      gplots_2.17.0     segmented_0.5-1.4 devtools_1.11.0
## [5] rmarkdown_0.7
##
## loaded via a namespace (and not attached):
##  [1] codetools_0.2-11   gtools_3.5.0       digest_0.6.8
```

```
##  [4] withr_1.0.1        bitops_1.0-6       R6_2.1.0
##  [7] git2r_0.13.1       formatR_1.2        magrittr_1.5
## [10] evaluate_0.7       httr_1.0.0         KernSmooth_2.23-14
## [13] stringi_1.0-1      curl_0.9.1         gdata_2.17.0
## [16] tools_3.2.1        stringr_1.0.0      yaml_2.1.13
## [19] caTools_1.17.1     memoise_1.0.0      htmltools_0.2.6
## [22] knitr_1.10.5
```