SCnorm: a quantile-regression based approach for robust normalization of single-cell RNA-seq data

Rhonda Bacher and Christina Kendziorski

November 28, 2016

Contents

1 Introduction

SCnorm (as detailed in Bacher* and Chu* *et al.*, *submitted*) is a quantile-regression based approach for robust normalization of single-cell RNA-seq data. SCnorm groups genes based on their count-depth relationship then applies a quantile regression to each group in order to estimate scaling factors which will remove the effect of sequencing depth from the counts.

2 Run SCnorm

Before analysis can proceed, the SCnorm package must be installed.

```
install.packages('SCnorm_x.x.x.tar.gz', repos=NULL, type="source")
```

After successful installation, the package must be loaded into the working space:

```
  library(SCnorm)
```

```
## Loading required package:  parallel
## Loading required package:  quantreg
## Loading required package:  SparseM
##
## Attaching package:  'SparseM'
```

```
## The following object is masked from 'package:base':
##
##    backsolve
## Loading required package:  cluster
## Loading required package:  moments
## Loading required package:  reshape2
## Loading required package:  ggplot2
```

2.1 Required inputs

Data: The matrix `Data` should be a $G - by - S$ matrix containing the expression values for each gene and each cell, where G is the number of genes and S is the number of cells/samples. The matrix should contain estimates of gene expression. Counts of this nature may be obtained from RSEM, HTSeq, Cufflinks, Salmon or a similar approach.

The object `ExampleData` is a simulated data matrix containing 5,000 rows of genes and 180 columns of cells.

```
data(ExampleData)
str(ExampleData)
```

```
##  num [1:5000, 1:180] 3.81 21.7 2.13 22.68 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:5000] "X_1" "X_2" "X_3" "X_4" ...
##   ..$ : chr [1:180] "C1_1" "C1_2" "C1_3" "C1_4" ...
```

Here we simulated data as in SIM 1 (as detailed in Bacher* and Chu* *et al.*, *submitted*) with K = 4 (four slope groups), each condition has 90 cells and condition 2 has been sequenced approximately 4 times as much as condition 1.

Conditions: The object `Conditions` should be a vector of length S indicating which condition each cell belongs to. The order of this vector should match the order of the columns in the `Data` matrix.

```
Conditions = rep(c(1,2), each= 90)
str(Conditions)
```

```
##  num [1:180] 1 1 1 1 1 1 1 1 1 1 ...
```

2.2 SCnorm: Check count-depth relationship

Before normalizing using SCnorm, it is advised to check the count-depth relationship in your data. If all genes have a similar relationship then a global strategy such as median-by-ratio in the DESeq package or TMM in edgeR will be adequate. However, in our paper we show that a count-depth relationship that varies among genes leads to poor normalization when using global scaling strategies, in which case we strongly recomend proceding with the normalization provided by SCnorm.

The function below will estimate the count-depth relationship for all genes, genes are first divided into groups based on their non-zero median expression, then the density of slopes for each group is plot. We recomend checking a variety of filter options, in case you find that only genes expressed in very few cells or very low expressors are the main concern.

```
checkCountDepth(Data = ExampleData, Conditions = Conditions, OutputName = "check_exampleData",
                PLOT=TRUE, FilterCellProportion = .1)
```

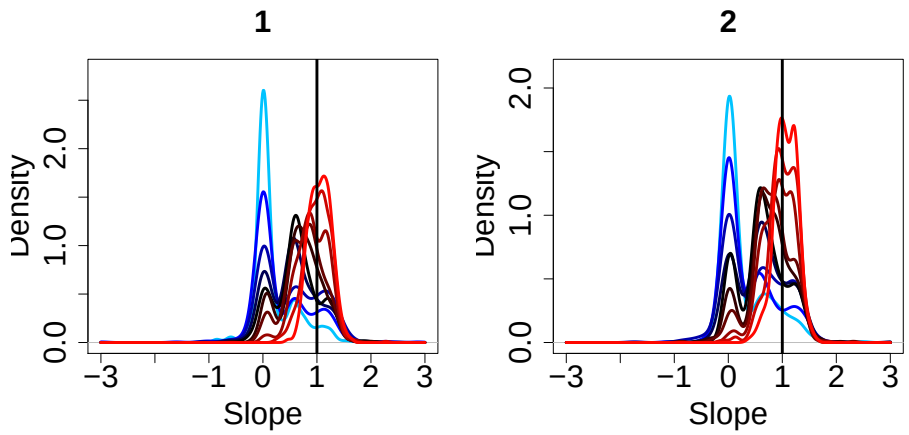

Figure 1: Evaluation of count-depth relationship in un-normalized data.

It can also be used to evaluate data normalized by other methods:

```
# Total Count normalization, Counts Per Million, CPM.
ExampleData.Norm <- t((t(ExampleData) / colSums(ExampleData)) * mean(colSums(ExampleData)))

checkCountDepth(Data = ExampleData, NormalizedData = ExampleData.Norm,
                Condition = Conditions, OutputName = "check_exampleDataNorm", PLOT=TRUE,
                FilterCellProportion = .1, FilterExpression = 2)
```

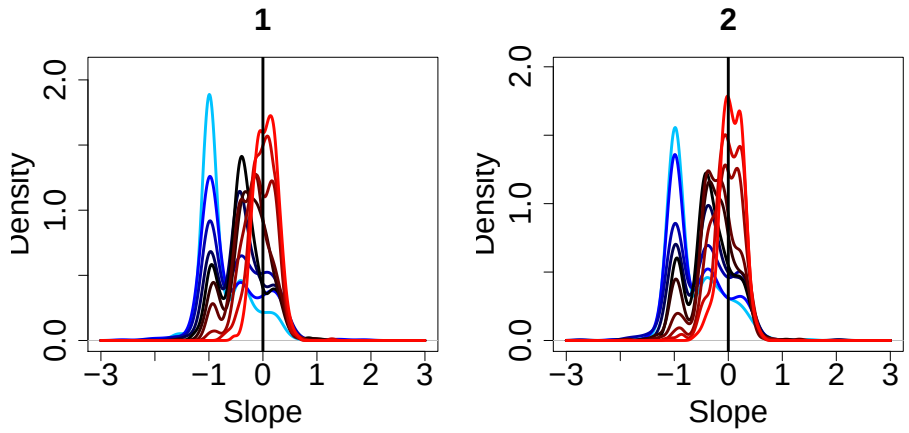

Figure 2: Evaluation of count-depth relationship in counts per million normalized example data.

Evaluating the bulk dataset included in the paper:

```
library(SCnorm)
data(bulkH1data)
Conditions <- rep(1, dim(bulkH1data)[2])
checkCountDepth(Data = bulkH1data, Condition = Conditions, OutputName = "check_bulkData",
                PLOT=TRUE, FilterCellProportion = .1, FilterExpression = 2)
```

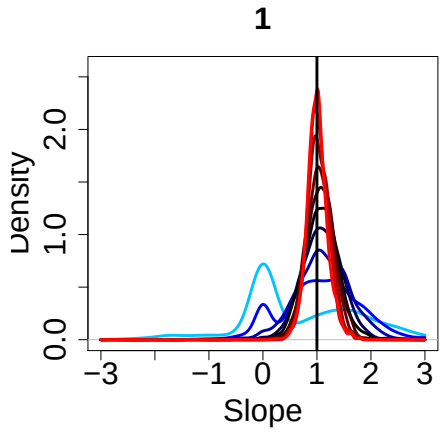

Figure 3: Evaluation of count-depth relationship in un-normalized bulk H1 data.

Evaluating the H1 single cell dataset included in the paper:

```
library(SCnorm)
data(scH1data)
Conditions <- rep(c("1M", "4M"), each=92)
checkCountDepth(Data = scH1data, Condition = Conditions, OutputName = "check_scData", PLOT=TRUE,
                FilterCellProportion = .1, FilterExpression = 2)
```

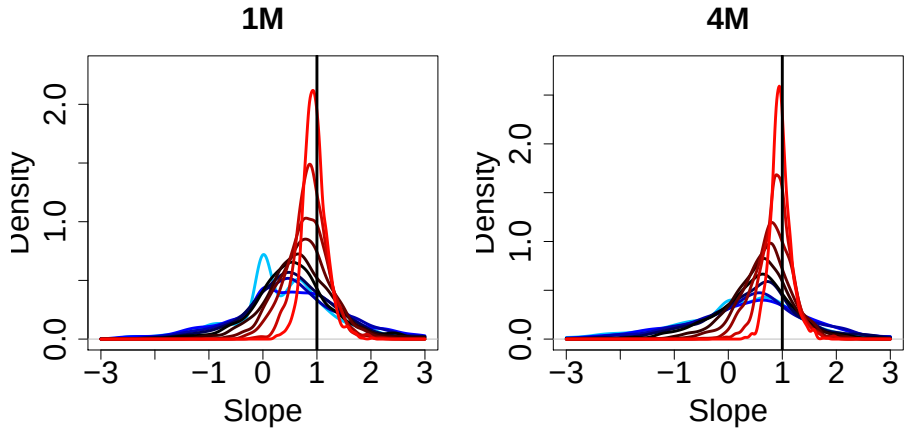

Figure 4: Evaluation of count-depth relationship in un-normalized H1 single cell data.

2.3 SCnorm: Normalization

SCnorm will normalize across cells to remove the effect of sequencing depth on the counts and return the normalized expression counts, a list of genes which were not considered in the normalization due to filter options, and optionally an additional matrix of scale factors (default = FALSE).The default filter for SCnorm only considers genes having at least 10 non-zero expression value. The user may wish to adjust the filter and may do so by changing the value of FilterCellNum.

```
Conditions = rep(c(1,2), each= 90)
DataNorm <- SCnorm(ExampleData, Conditions, OutputName = "MyNormalizedData",
                   PLOT=TRUE, FilterCellNum = 10)
str(DataNorm)
```

2.4 Evaluate choice of K

SCnorm first fits the model for $K = 1$, and sequentially increases K until a satisfactory stopping point is reached. For each value of K, SCnorm will estimate the count-depth relationship on the normalized counts. Gene evaluation groups are formed by splitting genes into 10 groups based on their non-zero median un-normalized expression and for each group the mode of the normalized count-depth relationship is estimated. If the absolute value of the maximum mode is ¡ .1, then K is selected, otherwise K is increase by one.

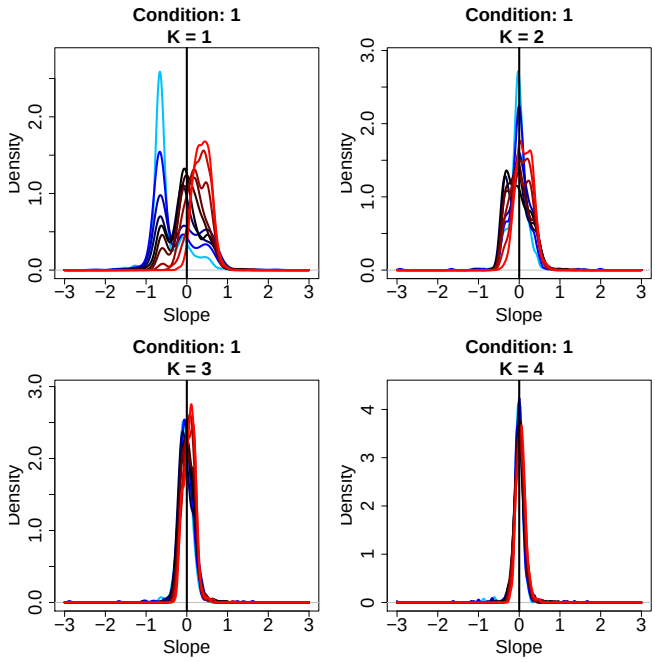

Figure 5: Evaluation of K

In Figure 5, $K = 4$ is chosen, once all 10 slope densities have absolute value of slope mode ¡ .1.

When more than one condition is present SCnorm will first normalize each condition independently then apply a scaling procedure between the conditions. In this step the assumption is that most genes are not differentially expressed (DE) between cells, that any systematic differences in expression across the majority of genes is due to technical bias and should be removed.

Generally the definition of condition will be obvious given the experimental setup. If the data are very heterogenous within an experimental setup it may be beneficial to first cluster more similar cells into groups and define these as conditions in SCnorm.

3 Session info

Here is the output of sessionInfo on the system on which this document was compiled:

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

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.11.6 (El Capitan)
##
## 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] parallel  stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] SCnorm_0.9     ggplot2_2.2.0  reshape2_1.4.2 moments_0.14   cluster_2.0.4
## [6] quantreg_5.29  SparseM_1.72   knitr_1.14
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.8        magrittr_1.5       munsell_0.4.3      colorspace_1.2-6
##  [5] lattice_0.20-33    highr_0.6          stringr_1.1.0      plyr_1.8.4
##  [9] tools_3.3.1        grid_3.3.1         gtable_0.2.0       MatrixModels_0.4-1
## [13] lazyeval_0.2.0     digest_0.6.10      assertthat_0.1     tibble_1.2
## [17] Matrix_1.2-6       formatR_1.4        evaluate_0.9       stringi_1.1.1
## [21] scales_0.4.1       BiocStyle_2.1.33
```