Variability in Clusters

Jo Hardin

7/2/2018

Goal

The goal of the series of experiments is to assess the variability of the clusters. Both in terms of re-sampling from the data set as well as using a probability model to find an SE.

Up first is R-code (credit: Madison Hobbs) which imports the data, normalizes, connects genes, etc.

Dan's goal

is to show that the amount of RpoS affects the preferential binding. If we delete a different transcription factor, would that change the shape of the averages?

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: CentOS Linux 7 (Core)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/R/lib/libRblas.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] mclust_5.4                amap_0.8-16
##  [3] DESeq2_1.18.1             SummarizedExperiment_1.8.1
##  [5] DelayedArray_0.4.1        matrixStats_0.53.1
##  [7] Biobase_2.38.0            GenomicRanges_1.30.3
##  [9] GenomeInfoDb_1.14.0       IRanges_2.12.0
## [11] S4Vectors_0.16.0          BiocGenerics_0.24.0
## [13] cluster_2.0.6             stringr_1.2.0
## [15] purrr_0.2.5               readr_1.1.1
## [17] tidyr_0.8.1               tibble_1.4.2
## [19] ggplot2_2.2.1             tidyverse_1.1.1
## [21] corrr_0.2.1               dplyr_0.7.5
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-131           bitops_1.0-6           bit64_0.9-7
##  [4] lubridate_1.6.0        RColorBrewer_1.1-2     httr_1.2.1
##  [7] rprojroot_1.2          tools_3.4.1            backports_1.1.0
```

```
## [10] R6_2.2.2              rpart_4.1-11          DBI_0.7
## [13] Hmisc_4.1-1           lazyeval_0.2.0        colorspace_1.3-2
## [16] nnet_7.3-12           tidyselect_0.2.4      gridExtra_2.2.1
## [19] mnormt_1.5-5          bit_1.1-12            compiler_3.4.1
## [22] rvest_0.3.2           htmlTable_1.12        xml2_1.1.1
## [25] scales_0.4.1          checkmate_1.8.5       psych_1.8.4
## [28] genefilter_1.60.0     digest_0.6.12         foreign_0.8-69
## [31] rmarkdown_1.6         XVector_0.18.0        base64enc_0.1-3
## [34] pkgconfig_2.0.1       htmltools_0.3.6       htmlwidgets_1.2
## [37] rlang_0.2.1           readxl_1.0.0          RSQLite_2.0
## [40] rstudioapi_0.6        bindr_0.1.1           jsonlite_1.5
## [43] BiocParallel_1.12.0   acepack_1.4.1         RCurl_1.95-4.8
## [46] magrittr_1.5          GenomeInfoDbData_1.0.0 Formula_1.2-3
## [49] Matrix_1.2-10         Rcpp_0.12.17          munsell_0.4.3
## [52] stringi_1.1.5         yaml_2.1.14           zlibbioc_1.24.0
## [55] plyr_1.8.4            blob_1.1.0            grid_3.4.1
## [58] forcats_0.2.0         lattice_0.20-35       haven_1.1.0
## [61] splines_3.4.1         annotate_1.56.2       hms_0.3
## [64] locfit_1.5-9.1        knitr_1.16            pillar_1.2.3
## [67] geneplotter_1.56.0    reshape2_1.4.3        XML_3.98-1.9
## [70] glue_1.1.1            evaluate_0.10.1       latticeExtra_0.6-28
## [73] data.table_1.11.4     modelr_0.1.1          cellranger_1.1.0
## [76] gtable_0.2.0          assertthat_0.2.0      skimr_1.0.3
## [79] xtable_1.8-2          broom_0.4.2           survival_2.41-3
## [82] memoise_1.1.0         AnnotationDbi_1.40.0  bindrcpp_0.2.2
```

Variability

1. Sample variability

One of the things Madison did was to calculate ARI from the deep vs deep with 3 of the "bad" samples removed. As a first pass at understanding how things change will be to remove ranodm samples. If we remove 3 random samples, how will the ARI compare to the ARI that Madison found.

Working with the deep data only (at first?).

We remove the samples with 100% RpoS (100.00_A, 100.00_B, 100.00_C) because they have a different strain of E. coli than the rest of the samples. We also removed the samples with much lower median read count (0.00_B, 0.35_B, and 20.40_A).

Q: how different is it to remove the *low* count samples vs any random 3 samples.

Deep Data, all samples except 100% RpoS

Cluster genes that (1) have at least raw count of 50, and (2) are DE across conditions.

How Similar are the two Deep Clusterings?

ARI: Deep (0.00_B, 0.35_B, 20.40_A removed) vs. Deep (0.00_B, 0.35_B, 20.40_A included)

The clusterings for the deep data with and without 0.00_B, 0.35_B, and 20.40_A are more similar than the deep vs. shallow, but certainly don't agree wholeheartedly either.

Deep Data, 0.00_C, 0.35_C, 20.40_C and 100% RpoS removed

(chose 3 different samples, arbitrarily)

We remove the samples with 100% RpoS and 0.00_C, 0.35_C, 20.40_C.

How Similar are the two Deep Clusterings?

ARI: Deep (0.00_B, 0.35_B, 20.40_A removed) vs. Deep (0.00_B, 0.35_B, 20.40_A included)

The clusterings for the deep data with and without 0.00_B, 0.35_B, and 20.40_A are more similar than the deep vs. shallow, but certainly don't agree wholeheartedly either.

new 3 vs. original 3

new 3 vs. full data

2. Variability across deep & shallow.

a. correlate genes across deep and shallow. how does the sequencing depth affect those correlations? Are they within the SE that would be expected?

b. What if we did something like a t-test? would the same genes come up as significant in the deep and shallow if we were just doing t-test across no rPos vs. some. (although we don't have a control value, 100% doesn't give us a control)

my notes talk about CRBD... maybe there is something about dependence? like a paired t-test?

c. Somehow (ARI??) measure the variability of the clusters. DESeq and the Negative Binomial may give us a way to think about the variability of some of the measures.

3. Simulation

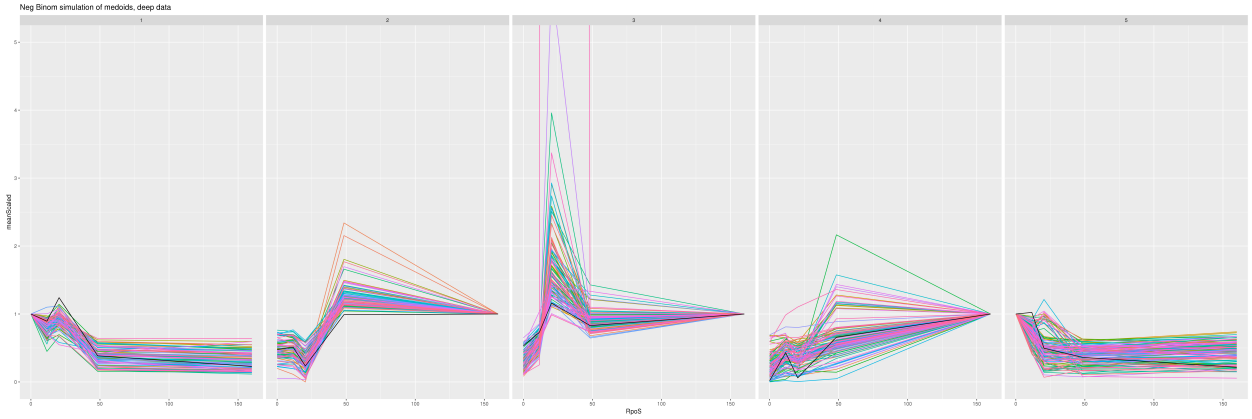
- simulate 1/10th of the deep data repeatedly
- does it "look" like the shallow data? what does it mean to look similar?
- use a parametric bootstrap w the multinomial.

```
## [1] "Mon Jul  2 18:10:03 2018"
```

```
## [1] "Mon Jul  2 19:40:19 2018"
```

There are 8550 genes above the 50 count threshold. The number of genes used in each replicate (above 50 and also DE) are summarized as:

```
## Skim summary statistics
##  n obs: 100
##  n variables: 1
##
## -- Variable type:integer -------------------------------------------------
##  variable   n    mean    sd   p0  p25  p50  p75 p100
##     numDE 100 2216.77 33.21 2132 2195 2219 2240 2295
```



The number of new medoids that correlate to each of the original medoid profiles is 83, 83, 101, 116, 117.

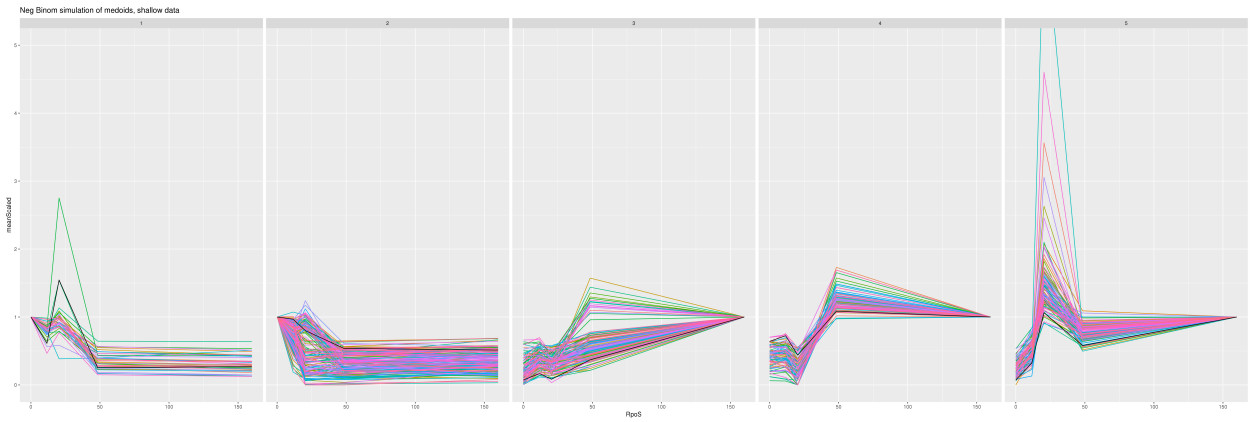
Repeating for shallow data...

```
## [1] "Mon Jul  2 19:42:29 2018"
```

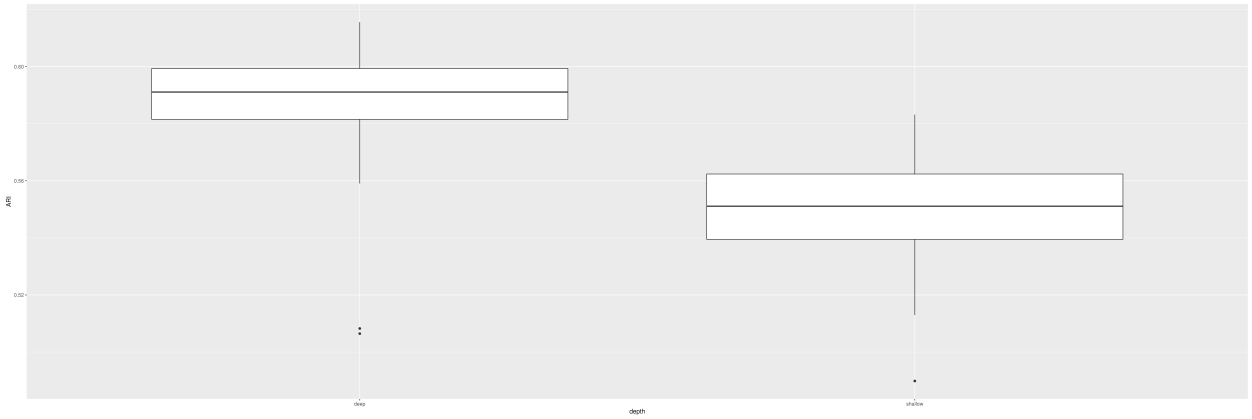
```
## [1] "Mon Jul  2 20:23:14 2018"
```

There are 10745 genes above the 5 count threshold. The number of genes used in each replicate (above 5 and also DE) are summarized as:

```
## Skim summary statistics
##  n obs: 100
##  n variables: 1
##
## -- Variable type:integer -------------------------------------------------
##  variable   n    mean    sd   p0     p25    p50    p75 p100
##     numDE 100 1570.96 29.67 1504 1548.75 1567.5 1590.5 1652
```



The number of new medoids that correlate to each of the original medoid profiles is 38, 164, 116, 82, 100.


0.60
0.56
0.52
ARI
deep
shallow
depth