A Modification Motif Analysis Demo

Pat Marks

April 27, 2012

1 Getting started

This is a demo for analyzing the modification detections from the Modification Detection module. Start by loading the raw modifications calls from `data/modifications.gff`

We install some packages from bioconductor:

We start by loading some utilities:

```
> library(ggplot2)
> library(cosmo)

Welcome to cosmo version 1.18.0

cosmo is free for research purposes only. For more details, type
license.cosmo(). Type citation('cosmo') for details on how to cite
cosmo in publications.

> source("scripts.R")
```

Notes about where to get the gff file from. Either from the SMRTportal job data location, or download the modifications.gff.gz from the SMRTportal job page.

2 Kinetic Data QC

The we load the gff file. The context field contains a 41 base context centered around the detected modification – pull out the center base. The context field contains a 41 base context centered around the detected modification. We pull out the center base. Look at the summary of the detected hits. For confident m6A detection you want coverages in the 15-20x (per strand) and above. Talk about the meanings of the columns from the GFF file.

```
> gffFile <- "/mnt/secondary/Smrtanalysis/userdata/jobs/040/040041/data/modifications.gff.
> hits <- readModificationsGff(gffFile)

Reading /mnt/secondary/Smrtanalysis/userdata/jobs/040/040041/data/modifications.gff.gz

> hits$CognateBase <- substr(hits$context, 21, 21)
> summary(hits)
```

```
  seqname             source            feature              start
Length:107758      Length:107758      Length:107758      Min.   :    132
Class :character   Class :character   Class :character   1st Qu.:1158852
Mode  :character   Mode  :character   Mode  :character   Median :2348764
                                                         Mean   :2332150
                                                         3rd Qu.:3497812
                                                         Max.   :4639410
     end              score           strand             frame
Min.   :    132   Min.   : 20.00   Length:107758      Length:107758
1st Qu.:1158852   1st Qu.: 23.00   Class :character   Class :character
Median :2348764   Median : 31.00   Mode  :character   Mode  :character
Mean   :2332150   Mean   : 50.89
3rd Qu.:3497812   3rd Qu.: 81.00
Max.   :4639410   Max.   :363.00
 attributes          coverage        context             IPDRatio
Length:107758      Min.   :  5.00   Length:107758      Min.   : 1.350
Class :character   1st Qu.: 41.00   Class :character   1st Qu.: 1.740
Mode  :character   Median : 47.00   Mode  :character   Median : 2.090
                   Mean   : 47.86                      Mean   : 2.848
                   3rd Qu.: 54.00                      3rd Qu.: 3.920
                   Max.   :223.00                      Max.   :10.910
 CognateBase
Length:107758
Class :character
Mode  :character
```

```
> hitsSubset <- hits[sample(nrow(hits), min(nrow(hits), 50000)),
+     ]
```

Plot the confidence scores and coverages to get a sense of what hits are significant. Figure 1

Make a histogram of the scores of the GFF entries, coloured by Cognate base. This will give you a sense of how strong your signals are and whether the strongest signals are enriched on any base.

You should see more high-scoring hits from the bases you expect to see methylated in your sample. This histogram can help you select a good filter for your hits. We'll pick a score threshold of 100. We sort the hits so we look at the most confident hits first.

```
> goodHits <- subset(hits, score > 70)
> goodHits <- goodHits[order(goodHits$score, decreasing = T), ]
> workHits <- goodHits
```

3 Motif Finding

Convert the context snippets into a cosmo compatible format. Pick a small selection of hits (cosmo will die if you provide too many at once). Feed the contexts to cosmo, look at the resulting logo plot. We found 'CTAT', starting at position 21-2 = 19

```
> p <- qplot(score, colour = CognateBase, geom = "freqpoly", data = hits)
> show(p)
```

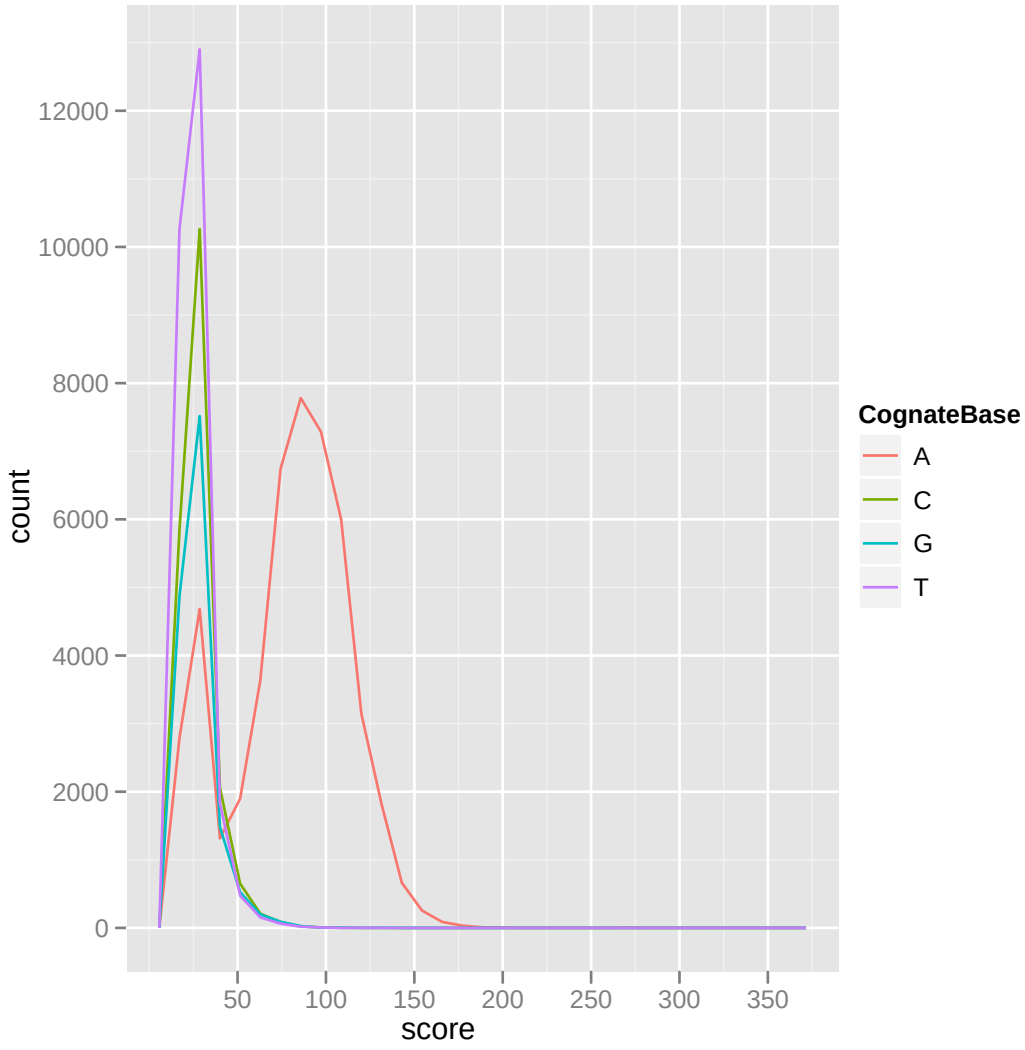

Figure 1: Modification Scores by cognate base

```
> p <- qplot(coverage, score, colour = CognateBase, alpha = I(0.3),
+     data = hitsSubset)
> show(p)
```

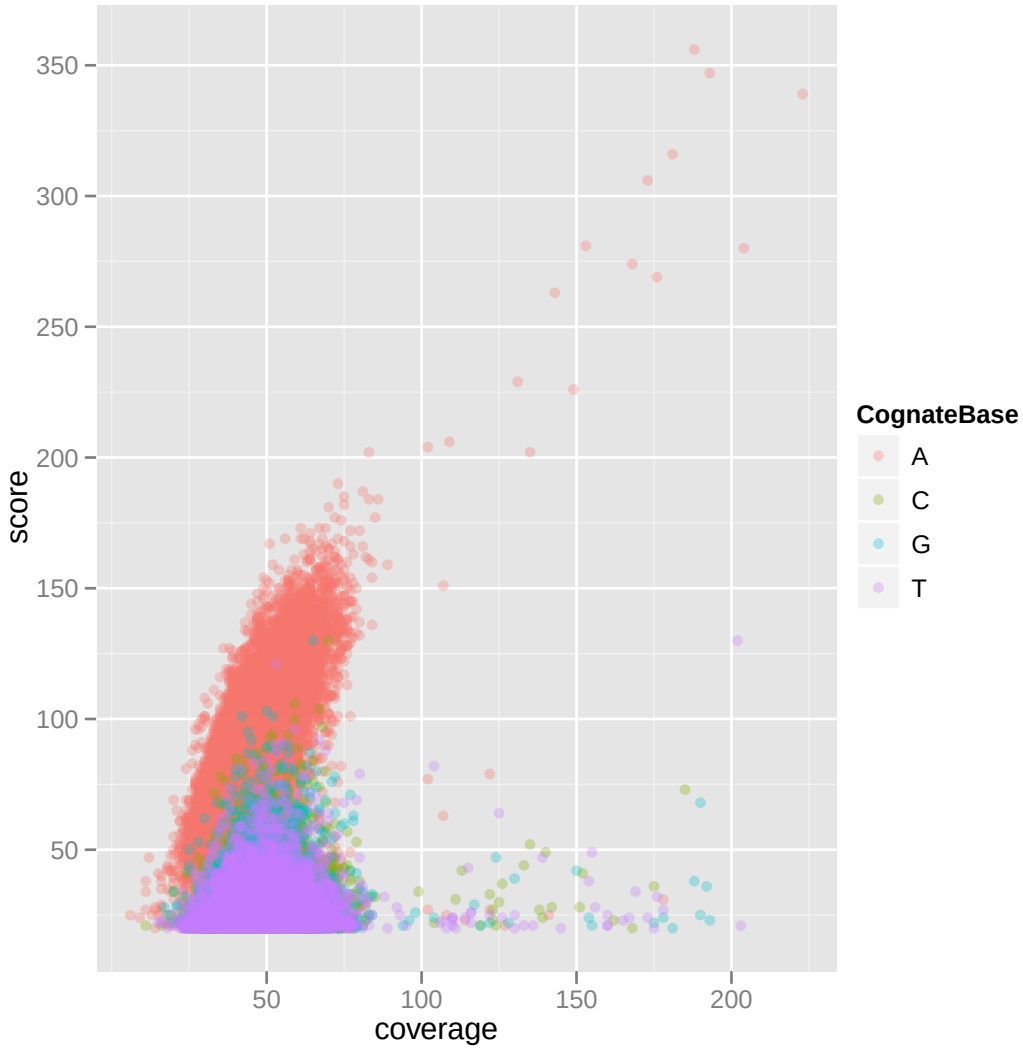

Figure 2: Score vs. Coverage

```
> sequenceList <- llply(workHits$context, function(x) list(seq = as.character(x),
+     desc = ""))
> smallHitList <- sequenceList[1:200]
> cc2 <- cosmo(seqs = smallHitList, minW = 4, maxW = 10, revComp = F,
+     silent = T)
> motif <- "CTATG"
> position <- 19
```

View the logo plot (Figure 3)

We found 'CTAT', starting at position 21-2 = 19. Pull out the -2 to +1 context window from the hits contexts. How many hits do we get with this motif. Roughly how many would we expect if everything was methylated in a random genome. We see the correct number of hits! We are now done with 'CTATG'. Remove the 'CTATG' hits from working list and try again. We still haev lots of unidentified motifs

```
> motif <- "CTATG"
> position <- 19
> snip <- substr(workHits$context, position, position + nchar(motif) -
+     1)
> observedHits <- sum(motif == snip)
> observedHits
```

```
[1] 0
```

```
> genomeSize <- max(hits$end)
> expectedHits <- genomeSize/(4^nchar(motif))
> expectedHits
```

```
[1] 4530.674
```

```
> workHits <- workHits[snip != motif, ]
> nrow(workHits)
```

```
[1] 33196
```

Not clear what's happening now – we probably have a mixture of motifs in similar proportion which will confuse cosmo. Continue on by hand, or wait for the fancy new motif finder!

4 Genome Annotation

We can load the genome sequence and annotate it instances of each motif, to determine the genome-wide methylation fraction of our motifs.

```
> seq_path <- "/mnt/secondary/Smrtanalysis/userdata/references/ecoli/sequence/ecoli.fasta"
> dna_seq <- read.DNAStringSet(seq_path)
> motifs = c("GATC", "GCACNNNNNNGTT", "AACNNNNNNGTGC")
> positions = c(2, 3, 2)
> motifAnnotations <- annotationDf(dna_seq, motifs, positions)
> head(motifAnnotations)
```

```
> plot(cc2)
```

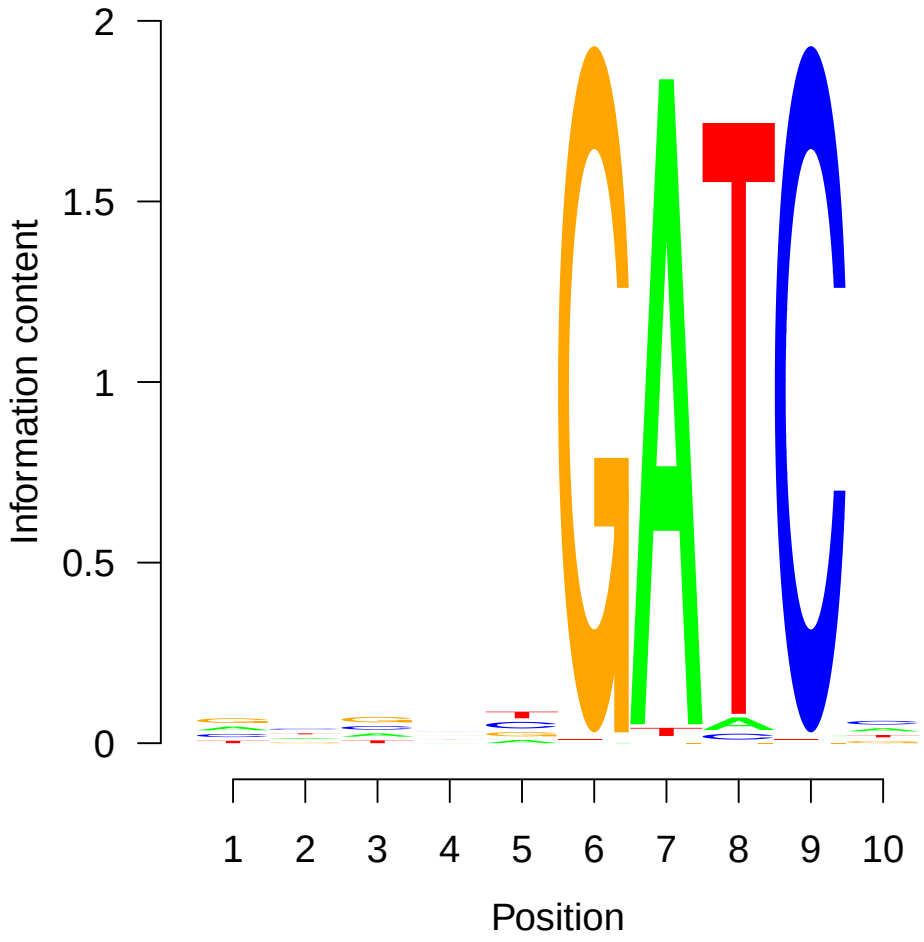

Figure 3: Logo plot

```
  strand  tpl  mod onTarget contigId
1      0  620 GATC       On        1
2      0  727 GATC       On        1
3      0  782 GATC       On        1
4      0  881 GATC       On        1
5      0 1168 GATC       On        1
6      0 1570 GATC       On        1
```

```
> table(motifAnnotations$mod, motifAnnotations$strand)
```

```
                    0     1
  AACNNNNNNGTGC   307   288
  GATC          19120 19120
  GCACNNNNNNGTT   288   307
```

5 Appendix

R can be downloaded for Linux, Mac and Windows from http://r-project.org We also recommend the graphical front-end RStudio, available from rstudio.org

The following R packages need to be installed to run through this demo. ggplot2 (available on CRAN), cosmo, Biostrings (available on Bioconductor)