Analysis of bisulfite amplicon MiSeq data using the aaRon package

Aaron Statham a.statham@garvan.org.au

April 27, 2015

Contents

1 Introduction

2 Alignment of raw data

3 Analysis of aligned data

The package *aaRon* may be downloaded directy from github using `devtools` if it needs to be installed or updated.

```
library(devtools)
install_github("astatham/aaRon")
```

Firstly we load the *R* package *aaRon*, and the *BSgenome* package of the organism our amplicon data was aligned to; in this case *BSgenome.Hsapiens.UCSC.hg19*.

```
library(aaRon)
library(BSgenome.Hsapiens.UCSC.hg19)
```

3.1 Loading the amplicon targets, and analysing the aligned reads against these targets

```
# load the amplicon info and display its format
amplicons <- read.csv("amplicons.csv")
str(amplicons)
```

```
## 'data.frame': 16 obs. of  5 variables:
##  $ Amplicon: Factor w/ 16 levels "APBB2","C9orf125",..: 13 16 11 9 2 10 14 7 5 3 ...
##  $ Target  : Factor w/ 16 levels "AGTAGAGGCTGGAGAAGGTGTTGTACTGGTCGATAGCCACGATGAGCTTGCACATGAGCTCCCCGAAGG
##  $ FW      : Factor w/ 16 levels "aggaagagagAGTAGAGGTTGGAGAAGGTGTTGTAT",..: 2 7 8 4 11 3 9 16 1 6 ...
##  $ RV      : Factor w/ 16 levels "cagtaatacgactcactatagggagaaggctAAAAAACTTAAATCACTAAACCCAAC",..: 12 9 2
##  $ Sequenom: logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```
# align the amplicons target sequence against hg19 and annotate it
amplicons <- ampliconGRanges(amplicons, Hsapiens, mc.cores=8)
amplicons
```

```
## GRanges object with 16 ranges and 11 metadata columns:
##        seqnames                 ranges strand   | Amplicon
##           <Rle>              <IRanges>  <Rle>   | <factor>
##    [1]     chr7 [ 24324841,  24325010]      +   |      NPY
##    [2]     chr2 [200335595, 200335716]      -   |    SATB2
##    [3]    chr10 [124902252, 124902386]      -   |     HMX2
##    [4]     chr7 [ 19184912,  19185078]      +   |   FERD3L
##    [5]     chr9 [104248506, 104248650]      +   | C9orf125
##    ...      ...                    ...    ... ...      ...
##   [12]     chr1 [200010070, 200010237]      +   |   CG242F
##   [13]     chr1 [211590146, 211590325]      +   |     LINC
##   [14]    chr12 [ 39299194,  39299348]      +   |    CPNE8
##   [15]     chr3 [125709660, 125709828]      +   |   ROPN1B
##   [16]     chr4 [ 40859230,  40859343]      +   |    APBB2
##
##
##    [1]           GAGCCTTCTGTGCCTGCAGATGCTAGGTAACAAGCGACTGGGGCTGTCCGGACTGACCCTCGCCCTGTCCCTGCTCGTGTGCCTGG
##    [2]                                                           GCCCCTGGCTGTAGCCTTTGGGACTTCTCTCCCGCGTC
##    [3]                                              GCCCCCTGGCTGGGTCACTGAGCCGCGGGAAGTGGACGACCCGAGTAAATG
##    [4]              GGGGAGGCCAGGGACAGGTCTGCGACGAAGTCCAGCACCGTAGTGTCCACGCAGCTCTCCGGATAGGCCGCCATCGCTTCGGC
##    [5]                                    CCAGGGCCCTGGATCTAGTGTCTGTCGCTGACCTTGGGCAGTCCCTGCCACGCTTGAGCCT
##    ...
##   [12]             GGCTGTGAGAGTCCCCTAGAGCTGAAGCCCCGGAGGCTGACCTGTGGGTCTGGCTGCTATGGGAACCCGGTTGGTCCAAAGAAG
##   [13] TGCTCAGGCTCAGCCTGTACCTGAGCTGCGCTGCGGCCACCTTCCTGCTGGGGACCCTGTTCGCCCTCGTCTGCCGGAGCCCGCGCGCCCCGCCGC
##   [14]                          CCACAGATGAGAGGGCAGGAGAGAGGGGAAGGGCTGCGTCCTCACCTGCAGGACACGGACACCTCCACCCG
##   [15]            CTGGGACAGAGGTGCAGGAAACACCTAATCTTACCTTCGTAACACAGCCGCTTGGTCTCTAGATGTGTTTCTGTTTAAATCGTCG
##   [16]                                                                   CCATTGGCTGCTGGGGAGGCGCGTGGCCCC
##                                             FW
##                                    <character>
##    [1]  aggaagagagGAGTTTTTTGTGTTTGTAGATGTTAGG
##    [2]    aggaagagagGTTTTTGGTTGTAGTTTTTGGGATT
##    [3]      aggaagagagGTTTTTTGGTTGGGTTATTGAGT
##    [4]         aggaagagagGGGGAGGTTAGGGATAGGTT
##    [5]    aggaagagagTTAGGGTTTTGGATTTAGTGTTTGT
##    ...                                    ...
##   [12]   aggaagagagGGTTGTGAGAGTTTTTTAGAGTTGAA
##   [13] aggaagagagTGTTTAGGTTTAGTTTGTATTTGAGTTG
##   [14]   aggaagagagTTATAGATGAGAGGGTAGGAGAGAGG
##   [15]      aggaagagagTTGGGATAGAGGTGTAGGAAATA
##   [16]          aggaagagagTTATTGGTTGTTGGGGAGG
##                                                                 RV  Sequenom    FW_len
##                                                        <character> <logical> <numeric>
##    [1]  cagtaatacgactcactatagggagaaggctCCGAATAATATCTAACCATATCCTCC      TRUE        27
##    [2] cagtaatacgactcactatagggagaaggctATAAACAACCTCCCACTTTAAAACTAA      TRUE        25
##    [3]   cagtaatacgactcactatagggagaaggctAACACCTAAACTCCCCTTAAAAATC      TRUE          23
##    [4]   cagtaatacgactcactatagggagaaggctTACCCCATCAAATTCAAAACTATTA      TRUE          20
##    [5]         cagtaatacgactcactatagggagaaggctTCCCTCCCAAAACCAAAAA      TRUE          25
##    ...                                                        ...       ...         ...
##   [12]       cagtaatacgactcactatagggagaaggctACTAAAATCCCCCAACCCAAC      TRUE          26
##   [13]   cagtaatacgactcactatagggagaaggctATAAATTTCCATACACAAAACTCCC      TRUE          28
##   [14]      cagtaatacgactcactatagggagaaggctCCTCCTCCCAATATAAACAACC      TRUE          26
##   [15]  cagtaatacgactcactatagggagaaggctAAAAAACTTAAATCACTAAACCCAAC      TRUE          23
##   [16]        cagtaatacgactcactatagggagaaggctACCCTAAACCAATCCCCTAA      TRUE          19
##           RV_len      size
##        <numeric> <numeric>
##    [1]        26       117
##    [2]        27        70
##    [3]        25        87
##    [4]        25       122
##    [5]        19       101
##    ...       ...       ...
##   [12]        21       121
##   [13]        25       127
##   [14]        22       107
##   [15]        26       120
##   [16]        20        75
##
##
##    [1]           TAACAAGCGACTGGGGCTGTCCGGACTGACCCTCGCCCTGTCCCTGCTCGTGTGCCTGGGTGCGCTGGCCGAGGCGTACCCCTCCA
##    [2]                                                          TCTCTCCCGCGTCTTGGGTCAGAGCAGCGTCCGCAGCAA
##    [3]                                         CGCGGGAAGTGGACGACCCGAGTAAATGGGGCAAAATGGAATCGTGGAGGGCAGCG
##    [4]      TGCGACGAAGTCCAGCACCGTAGTGTCCACGCAGCTCTCCGGATAGGCCGCCATCGCTTCGGCTTGGCCCTGCCTCTCATCGGTTTTCCGC
##    [5]                           CGCTGACCTTGGGCAGTCCCTGCCACGCTTGAGCCTCAGTTTCCCACCCGTAGGCCGAGCTGACCGTCTC
##    ...
##   [12]       GCCCCGGAGGCTGACCTGTGGGTCTGGCTGCTATGGGAACCCGGTTGGTCCAAAGAAGCCTTTCTTCCGGGCACCTGGAATTCCAGTTTA
##   [13] CGCTGCGGCCACCTTCCTGCTGGGGACCCTGTTCGCCCTCGTCTGCCGGAGCCCGCGCGCCCCGCCGCCCGACTTTGCCGCCGCCTGGAGCCGGCT
##   [14]                     GGAAGGGCTGCGTCCTCACCTGCAGGACACGGACACCTCCACCCGCGTGGCCGGGATGGCAGCGCTCAGCTGGTTC
##   [15]        CCTAATCTTACCTTCGTAACACAGCCGCTTGGTCTCTAGATGTGTTTCTGTTTAAATCGTCGTAGCCCTCTAGAGCGCCGGCTTTCGGG
##   [16]                                                     CGCGTGGCCCCGCCCCACCCGGCCGCACTCCCGTGAAGTGCTCC
##             CGs       primers
##          <list> <GRangesList>
##    [1] ########      ########
##    [2] ########      ########
##    [3] ########      ########
##    [4] ########      ########
##    [5] ########      ########
##    ...      ...           ...
##   [12] ########      ########
##   [13] ########      ########
##   [14] ########      ########
##   [15] ########      ########
##   [16] ########      ########
##   -------
##   seqinfo: 93 sequences from an unspecified genome
```

```
# Find BAM files which have been name sorted then overlapping read pairs clipped
bams <- dir("aligned_reads", pattern=".name.clip.bam", full.names=TRUE, recursive=TRUE)
names(bams) <- gsub(".*/", "", sub(".name.clip.bam", "", bams))
```

```
bams

##                                     1_S1                                     2_S2
## "aligned_reads/1_S1/1_S1.name.clip.bam" "aligned_reads/2_S2/2_S2.name.clip.bam"
##                                     3_S3                                     4_S4
## "aligned_reads/3_S3/3_S3.name.clip.bam" "aligned_reads/4_S4/4_S4.name.clip.bam"
```

The ampliconAnalysis is the real workhorse - it loads each supplied BAM file, filters reads on various flags and quality scores, creates a pileup of sequenced bases at each position of each amplicon and then creates methylation summaries.

```
results <- ampliconAnalysis(amplicons, bams, Hsapiens, paired=TRUE, mc.cores=8)

## Reading in aligned sequencing libraries
## Creating pileup at each base of each amplicon
## Calculating methylation and conversion ratios
```

That's it!

3.2 Exploring the analysed amplicon data

The results object returned by ampliconAnalysis is a *list* with the following elements:

- amplicons - a copy of the amplicons object passed to ampliconAnalysis.
- summary - Some summary statistics of the experiment, per amplicon.
- CpGs - Per CpG site methylation calls and counts of coverage.
- Cs - Per cytosine methylation calls and counts of coverage - useful for nonCpG methylation, or for estimating bisulfite non-conversion.
- all_bases - Counts and ratios of A/C/G/Ts seqenced per base of each amplicon.

Figure 1 shows some per-amplicon summaries of the experiment created from the results$summary object as follows.

```
library(reshape2)
library(ggplot2)
for (toplot in c("Reads", "Conversion", "Methylation")) {
        # Extract the metric of interest
        tmp <- results$summary[, paste0(amplicons$Amplicon, "_", toplot)]
        # Fix up the columns names
        names(tmp) <- sub(paste0("_", toplot), "", names(tmp))
        # melt() for ggplotting, again fix up the column names
        tmp <- melt(t(tmp))
        names(tmp) <- c("Amplicon", "Sample", toplot)
        # Plot
        p <- ggplot(tmp, aes_string(x="Amplicon", y=toplot, fill="Sample", group="Sample")) + geom_bar(sta
        p <- p + theme(axis.text.x=element_text(angle=90), axis.title.x = element_blank())
        print(p)
}
```

The methylation of the *FERD3L* amplicon looks variable across the four samples, so lets "zoom in" and plot the methylation of each individual CpG site using the results$CpGs object.

```
# Extract the FERD3L
tmp <- as.data.frame(results$CpGs[results$CpGs$amplicons=="FERD3L"])
tmp

##    seqnames    start      end width strand amplicons base    C.1_S1    C.2_S2    C.3_S3
## 1      chr7 19184934 19184934     1      +    FERD3L   CG 0.9301426 0.3332604 0.6425790
## 2      chr7 19184937 19184937     1      +    FERD3L   CG 0.8151505 0.2835984 0.5546423
```

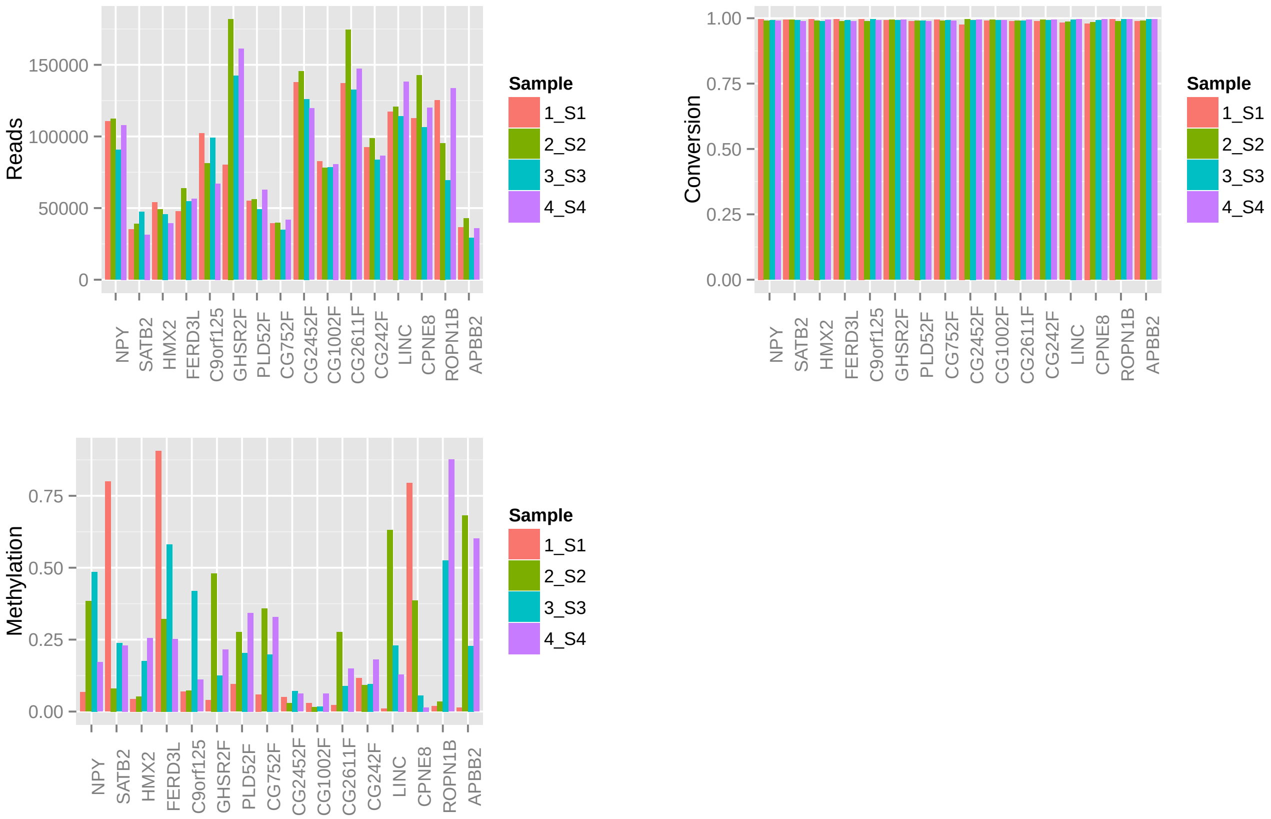

Figure 1: Per-amplicon summaries of sequencing coverage, bisulfite conversion and CpG methylation

```
## 3      chr7 19184950 19184950     1      +    FERD3L   CG 0.9458865 0.3568641 0.6798547
## 4      chr7 19184961 19184961     1      +    FERD3L   CG 0.9237819 0.2705802 0.5791630
## 5      chr7 19184971 19184971     1      +    FERD3L   CG 0.9643641 0.3997012 0.5973701
## 6      chr7 19184980 19184980     1      +    FERD3L   CG 0.9511014 0.3475279 0.5543482
## 7      chr7 19184986 19184986     1      +    FERD3L   CG 0.9583351 0.3198373 0.5311316
## 8      chr7 19184991 19184991     1      +    FERD3L   CG 0.9492011 0.3623085 0.6682803
## 9      chr7 19185012 19185012     1      +    FERD3L   CG 0.9611509 0.3372730 0.6236042
## 10     chr7 19185020 19185020     1      +    FERD3L   CG 0.9455229 0.3756777 0.6588640
## 11     chr7 19185032 19185032     1      +    FERD3L   CG 0.6251889 0.1549657 0.3004647
##       C.4_S4 cov.1_S1 cov.2_S2 cov.3_S3 cov.4_S4
## 1  0.2364581    24049    31984    27483    28301
## 2  0.1982975    24052    32003    27497    28311
## 3  0.2500442    24079    32029    27525    28311
## 4  0.2250397    24115    32094    27576    28355
## 5  0.2626163    24133    32129    27606    28376
## 6  0.2734501    24152    32219    27701    28422
## 7  0.2230201    24145    32207    27705    28410
## 8  0.2335013    24095    32177    27680    28381
## 9  0.2638512    24016    32152    27673    28391
## 10 0.4158056    23955    32094    27640    28357
## 11 0.2009840    23828    31949    27544    28251

# just want the "start" position and the "C" ratio for each sample
```

```
tmp <- tmp[,c(2, 8:11)]
names(tmp) <- c("Position", sub("C.", "", names(tmp)[-1]))
# Melt
tmp <- melt(tmp, id.vars="Position", value.name="Methylation")
# plot
p <- ggplot(tmp, aes(x=Position, y=Methylation, color=variable)) + geom_point() + geom_line()
p <- p + ylim(0, 1) + ggtitle("Methylation @ FERD3L") + xlab("CpG site position")
print(p)
```

3.3 Exporting an experiment ”bigTable”

```
# output summary
write.csv(results$summary, "summary.csv")
# output "bigTable" of all CpG sites
tmp <- as.data.frame(results$CpGs)[,-c(3:5)]
names(tmp)[1:4] <- c("chr", "position", "amplicon", "base")
write.table(tmp, "CpG_bigTable.csv", sep=",", row.names=FALSE)
```

4 Conclusions

5 Session info

```
sessionInfo()

## R version 3.1.1 (2014-07-10)
## Platform: x86_64-unknown-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=en_AU.UTF-8       LC_NUMERIC=C               LC_TIME=en_AU.UTF-8
##  [4] LC_COLLATE=en_AU.UTF-8     LC_MONETARY=en_AU.UTF-8    LC_MESSAGES=en_AU.UTF-8
##  [7] LC_PAPER=en_AU.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_AU.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
## [1] BiocStyle_1.4.1 digest_0.6.8    evaluate_0.5.5  formatR_1.0     highr_0.4
## [6] knitr_1.9       stringr_0.6.2   tools_3.1.1
```

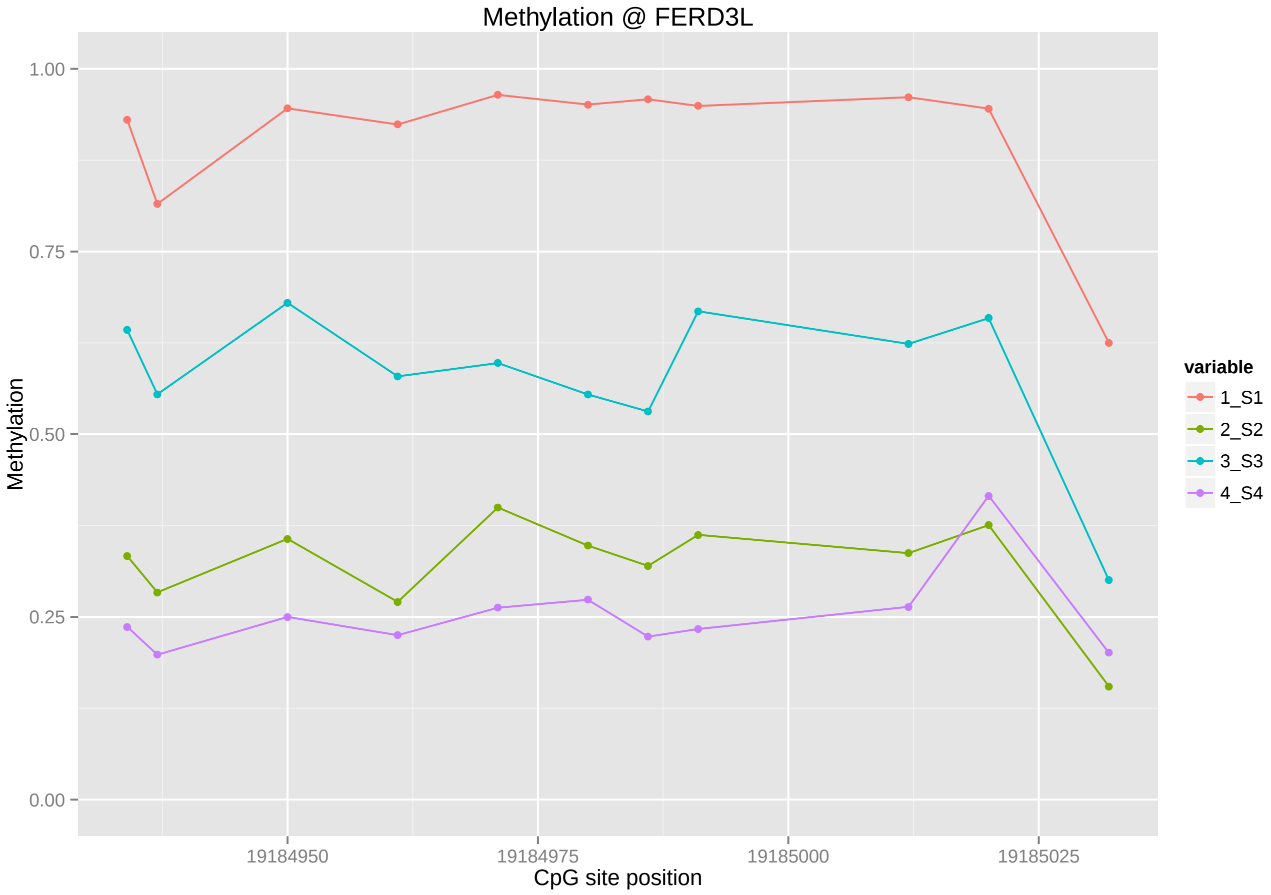

Figure 2: Methylation ratio for the four samples across the FERD3L locus