methdiffSatScan vignette

Laura Helmkamp

2015-09-18

Laura Helmkamp

2015-09-18

1: Introduction

DNA methylation plays critical roles in gene regulation and cellular specification without altering DNA sequences. It is one of the best understood and most intensively studied epigenetic marks in mammalian cells. Treatment of DNA with sodium bisulfite deaminates unmethylated cytosines to uracil while methylated cytosines are resistant to this conversion thus allowing for the discrimination between methylated and unmethylated CpG sites. Sodium bisulfite pre-treatment of DNA coupled with next-generation sequencing has allowed DNA methylation to be studied quantitatively and genome-wide at single cytosine site resolution. - change this; taken from MethylSig

The **methdiffSatScan** package allows users to find differentially methylated regions (DMRs) using the scan statistic. **methdiffSatScan** constitutes a convenient pipeline linking the sitewise likelihood-based differential methylation statistics calculated by the package **methylSig**, the free scan statistic software **SaTScan** (via **RSaTScan**), and the plotting capabilities of Gviz [1,2,3,4].

This document is a step-by-step user guide for the methdiffSatScan package.

2: Installation

2a: Installing methdiffSatScan

Install **methdiffSatScan** with the devtools R package:

```
library(devtools)
install_github('lhelmkamp/methdiffSatScan')
```

2b: Installing SaTScan

methdiffSatScan requires the stand-alone software **SaTScan**. This can be easily downloaded from http://www.satscan.org/download.html for Windows, Linux, or Mac OS X.

3: Sample data and data class

As sample data, we use a small portion of the sample data publicly available on the GEO database under accession number GSE61161. Data was obtained from 39 patients with chronic myelomonocytic leukemia (CMML), and we are interested in locating regions which could predict response to a commonly used treatment. Data for chromosome 18 is provided with the package.

```
library(methdiffSatScan)
data(CMML_chr18)
CMML_chr18
```

```
## methylSigData object with 65,503 rows
## --------------------------
##      chr start end strand coverage1 numCs1 numTs1 coverage2 numCs2 numTs2
## 1  chr18   796 796      +        70     70      0        67     63      4
## 2  chr18   816 816      +        72     46     26        67     59      8
## 3  chr18   817 817      +        66     46     20        55     53      2
## 4  chr18   824 824      +        70     64      6        66     63      3
## 5  chr18   825 825      +        72     70      2        55     55      0
## 6  chr18   830 830      +        70     67      3        66     56     10
## 7  chr18   831 831      +        70     68      2        57     55      2
## 8  chr18   832 832      +        69     66      3        65     56      9
## 9  chr18   833 833      +        71     68      3        57     56      1
## 10 chr18   840 840      +        58     58      0        51     51      0
##    coverage3 numCs3 numTs3 coverage4 numCs4 numTs4 coverage5 numCs5 numTs5
## 1         82     78      4        30     29      1       103     83     20
## 2         84     80      4        29     26      3       111    100     11
## 3         86     82      4        46     43      3       105     98      7
## 4         82     71     11        30     28      2       110    103      7
## 5         84     77      7        47     46      1       108    104      4
## 6         83     79      4        29     26      3       110    103      7
## 7         87     82      5        48     44      4       108    103      5
## 8         80     67     13        27     26      1       111    101     10
## 9         88     81      7        48     45      3       107     96     11
## 10       136    124     12        86     85      1       162    125     37
##    coverage6 numCs6 numTs6 coverage7 numCs7 numTs7 coverage8 numCs8 numTs8
## 1        128    124      4       177    171      6        96     88      8
```

```
## 2        128    118     10       178    171      7        98     94      4
## 3        171    161     10       111    107      4        82     73      9
## 4        127    122      5       177    163     14        99     90      9
## 5        176    174      2       116    115      1        80     76      4
## 6        128    109     19       176    130     46        96     89      7
## 7        175    158     17       115    112      3        82     79      3
## 8        128    103     25       175    127     48        98     95      3
## 9        177    159     18       116    112      4        83     70     13
## 10       144    142      2       138     77     61        62     62      0
##    coverage9 numCs9 numTs9 coverage10 numCs10 numTs10 coverage11 numCs11
## 1        107     98      9         81      74       7         NA      NA
## 2        108    101      7         82      75       7         NA      NA
## 3        123    114      9        108     102       6         NA      NA
## 4        106    104      2         82      76       6         NA      NA
## 5        122    118      4        111     106       5         NA      NA
## 6        107    102      5         81      74       7         NA      NA
## 7        124    114     10        110     107       3         NA      NA
## 8        107    102      5         81      73       8         NA      NA
## 9        124    118      6        111     102       9         NA      NA
## 10        88     86      2         77      74       3        203     203
##    numTs11 coverage12 numCs12 numTs12 coverage13 numCs13 numTs13
## 1       NA         NA      NA      NA         NA      NA      NA
## 2       NA         NA      NA      NA         NA      NA      NA
## 3       NA         NA      NA      NA         NA      NA      NA
## 4       NA         NA      NA      NA         NA      NA      NA
## 5       NA         NA      NA      NA         NA      NA      NA
## 6       NA         NA      NA      NA         NA      NA      NA
## 7       NA         NA      NA      NA         NA      NA      NA
## 8       NA         NA      NA      NA         NA      NA      NA
## 9       NA         NA      NA      NA         NA      NA      NA
## 10       0         94      93       1        176     173       3
##    coverage14 numCs14 numTs14 coverage15 numCs15 numTs15 coverage16
## 1          40      40       0         NA      NA      NA         NA
## 2          39      39       0         NA      NA      NA         NA
## 3          NA      NA      NA         NA      NA      NA         NA
## 4          39      38       1         NA      NA      NA         NA
## 5          NA      NA      NA         NA      NA      NA         NA
## 6          40      18      22         NA      NA      NA         NA
## 7          NA      NA      NA         NA      NA      NA         NA
## 8          40       7      33         NA      NA      NA         NA
## 9          NA      NA      NA         NA      NA      NA         NA
## 10        102      70      32         92      76      16         97
##    numCs16 numTs16 coverage17 numCs17 numTs17 coverage18 numCs18 numTs18
## 1       NA      NA         NA      NA      NA        114     108       6
## 2       NA      NA         NA      NA      NA        111      93      18
## 3       NA      NA         NA      NA      NA         79      71       8
```

```
## 4       NA      NA         NA      NA      NA        114      108        6
## 5       NA      NA         NA      NA      NA         81       81        0
## 6       NA      NA         NA      NA      NA        114      111        3
## 7       NA      NA         NA      NA      NA         81       79        2
## 8       NA      NA         NA      NA      NA        116      113        3
## 9       NA      NA         NA      NA      NA         81       80        1
## 10      97       0         49      49       0         70       70        0
##    coverage19 numCs19 numTs19 coverage20 numCs20 numTs20 coverage21
## 1          53      49       4         130     126        4         72
## 2          51      51       0         131      81       50         73
## 3          41      39       2          50      40       10         60
## 4          52      50       2         128     120        8         73
## 5          42      41       1          50      50        0         61
## 6          51      47       4         131      67       64         73
## 7          42      40       2          50      44        6         61
## 8          52      52       0         129      70       59         72
## 9          43      43       0          52      50        2         62
## 10        210     209       1         162     104       58        114
##    numCs21 numTs21 coverage22 numCs22 numTs22 coverage23 numCs23 numTs23
## 1       69       3         74      55      19         87       59       28
## 2       68       5         73      64       9         86       80        6
## 3       58       2        112      99      13         46       44        2
## 4       69       4         73      73       0         86       73       13
## 5       60       1        113     112       1         48       44        4
## 6       56      17         55      54       1         85       71       14
## 7       54       7        113     108       5         47       42        5
## 8       55      17         70      66       4         83       64       19
## 9       47      15        114     114       0         48       46        2
## 10     113       1        125     122       3        103       62       41
##    coverage24 numCs24 numTs24 coverage25 numCs25 numTs25 coverage26
## 1         172     123       49         80      76        4         62
## 2         172     163        9         80      73        7         60
## 3         129     123        6         84      76        8         15
## 4         170     162        8         79      77        2         62
## 5         132     132        0         83      81        2         15
## 6         169     163        6         78      68       10         62
## 7         133     133        0         84      82        2         15
## 8         169     148       21         78      54       24         62
## 9         133     122       11         84      60       24         15
## 10        233     202       31        118     111        7         56
##    numCs26 numTs26 coverage27 numCs27 numTs27 coverage28 numCs28 numTs28
## 1       62       0         46      41       5         18       18        0
## 2       60       0         45      36       9         18       18        0
## 3       14       1         44      40       4         12        8        4
## 4       62       0         45      43       2         17       17        0
## 5       15       0         44      44       0         12       12        0
```

```
## 6       58       4         46      43       3         18       18       0
## 7       15       0         45      42       3         12       12       0
## 8       58       4         44      35       9         18       15       3
## 9       15       0         45      35      10         12       11       1
## 10      32      24         44      44       0        221      211      10
##    coverage29 numCs29 numTs29 coverage30 numCs30 numTs30 coverage31
## 1          42      38       4         NA      NA      NA          40
## 2          42      42       0         NA      NA      NA          40
## 3          NA      NA      NA         NA      NA      NA          11
## 4          41      41       0         NA      NA      NA          40
## 5          NA      NA      NA         NA      NA      NA          12
## 6          40      39       1         NA      NA      NA          40
## 7          NA      NA      NA         NA      NA      NA          12
## 8          40      11      29         NA      NA      NA          40
## 9          NA      NA      NA         NA      NA      NA          12
## 10        216     171      45         84      82       2          81
##    numCs31 numTs31 coverage32 numCs32 numTs32 coverage33 numCs33 numTs33
## 1       40       0         96      92       4         43      42       1
## 2       39       1         94      92       2         43      39       4
## 3       11       0        101      96       5         58      54       4
## 4       39       1         92      90       2         43      42       1
## 5       11       1        101      97       4         60      59       1
## 6       37       3         94      91       3         42      34       8
## 7       10       2        101      97       4         60      58       2
## 8       34       6         95      89       6         41      26      15
## 9        7       5        101      90      11         60      46      14
## 10      57      24        114      98      16         36      36       0
##    coverage34 numCs34 numTs34 coverage35 numCs35 numTs35 coverage36
## 1         121      67      54         73      57      16          89
## 2         122     121       1         73      60      13          86
## 3         108     106       2         67      33      34          61
## 4         121     120       1         72      72       0          89
## 5         109     109       0         69      69       0          64
## 6         123     122       1         72      72       0          89
## 7         109     108       1         70      70       0          64
## 8         121      94      27         72      42      30          86
## 9         110      90      20         70      70       0          64
## 10         62      62       0        124     103      21          91
##    numCs36 numTs36 coverage37 numCs37 numTs37 coverage38 numCs38 numTs38
## 1       68      21        100      57      43         NA      NA      NA
## 2       84       2         99      96       3         NA      NA      NA
## 3       54       7         70      67       3         NA      NA      NA
## 4       82       7        100      93       7         NA      NA      NA
## 5       63       1         72      72       0         NA      NA      NA
## 6       84       5         99      94       5         NA      NA      NA
## 7       63       1         71      67       4         NA      NA      NA
```

```
## 8       85       1        100      94       6         NA      NA      NA
## 9       63       1         71      67       4         NA      NA      NA
## 10      88       3        180     176       4        149     143       6
##    coverage39 numCs39 numTs39
## 1          27      11      16
## 2          27       9      18
## 3          19      17       2
## 4          27      27       0
## 5          18      17       1
## 6          27      27       0
## 7          19      19       0
## 8          27      24       3
## 9          19      19       0
## 10         74      72       2
## --------------------------
## sample.ids: GSM1498786 GSM1498787 GSM1498788 GSM1498789 GSM1498790 GSM1498791 GSM1498792
## treatment: 0 0 0 0 0 1 1 1 1 1 1 0 0 1 0 0 1 0 0 0 0 0 0 1 1 1 1 0 0 0 1 0 1 0 1 1 1 1 1 1
## destranded: TRUE
## resolution: base
## options: maxCount=1000 & minCount=0 & filterSNPs=FALSE & assembly=hg18 & context=CpG
```

methdiffSatScan requires that input data be formatted as methylSigData object, as is output from the methylSigReadData() function in the methylSig package. If your data is not in this format, please see the Appendix.

4: Finding DMRs with methdiffSatScan

With data in a methylSigData object, the function **methdiffSatScan()** can be used to find DMRs. methdiffSatScan performs the following steps:

1. Obtain sitewise likelihood ratio statistics
2. Normalize the sitewise statistics using quantiles
3. Write the files needed by SatScan to a local directory
4. Run SatScan and view the results

We start with a simple call to **methdiffSatScan()** for the chromosome 18 data:

```
time0.a<-proc.time()
result.a<-methdiffSatScan(CMML_chr18, xvalues = "Index")
time1.a<-proc.time()
```

Here, `xvalues = "Index"` indicates that in the scan statistic, the data points will be treated as equally spaced rather than using the position of the site on the chromosome. We include timing code to allow the user to compare computational speed:

```
(time1.a-time0.a)/60
```

```
##      user    system   elapsed
##  2.648833  0.001500 13.404833
```

On our computer, where chromosome 18 takes about 13 minutes, analyzing the entire genome-wide dataset for the CMML data takes about 16 hours.

The output `result.a` is a list, the first element of which contains the significant DMRs:

```
result.a$DMRs
```

```
##       chr pos.start pos.stop ind.start ind.stop length nsites  pval
## 1   chr18  68685090 68686053     43903    44071    964    169 0.001
## 2   chr18  73090715 73092199     50540    50687   1485    148 0.018
## 3   chr18  65219178 65220130     42710    42859    953    150 0.018
## 4   chr18  43027822 43035090     29187    29442   7269    256 0.026
## 5   chr18   4443976  4445540      3358     3549   1565    192 0.035
## 6   chr18  30056605 30328285     23151    23463 271681    313 0.035
## 7   chr18  74839653 74841158     54786    54991   1506    206 0.037
## 8   chr18  18004331 18005631     16956    17075   1301    120 0.039
## 9   chr18  75487088 75487435     59937    59947    348     11 0.041
## 10  chr18  55038079 55091592     38084    38537  53514    454 0.043
```

Note that the result is also automatically output to `result`; we output the result as a precaution due to the long computational time of **methdiffSatScan** on large datasets. However, we recommend returning output to a more desciptive name as well, as `result` will be overwritten by subsequent calls to **methdiffSatScan**.

If multiple chromosomes are present in the input `meth` object, **methdiffSatScan** will automatically analyze each chhromsome separately. For human data, the user can also split each chromsome at the centromere with `splitcentromere = TRUE` and by specifying `build` from `"hg18"`, `"hg19"`, or `"hg38"`. This will speed up the analysis somewhat, and may be necesary for large datasets depending on system memory.

```
library(methdiffSatScan)
time0.b<-proc.time()
result.b<-methdiffSatScan(CMML_chr18, xvalues = "Index", splitchr = TRUE, splitcentromere =
time1.b<-proc.time()
```

```
(time1.b-time0.b)/60
```

```
##        user      system     elapsed
## 2.559166667 0.002833333 8.567500000
```

Comparing these new results to those obtained previously:

```
result.b$DMRs
```

```
##      chr pos.start pos.stop ind.start ind.stop length nsites  pval
## 1  chr18   4444894  4445540      3434     3549    647    116 0.002
## 2  chr18  68685090 68686053     28408    28576    964    169 0.001
## 3  chr18  73090715 73092199     35045    35192   1485    148 0.016
## 4  chr18  65219178 65220130     27215    27364    953    150 0.016
## 5  chr18  43027822 43035090     13692    13947   7269    256 0.021
## 6  chr18  30056605 30328285      7656     7968 271681    313 0.024
## 7  chr18  74839653 74841158     39291    39496   1506    206 0.026
## 8  chr18  18004331 18005631      1461     1580   1301    120 0.027
## 9  chr18  75487088 75487435     44442    44452    348     11 0.027
## 10 chr18  55038079 55091592     22589    23042  53514    454 0.036
```

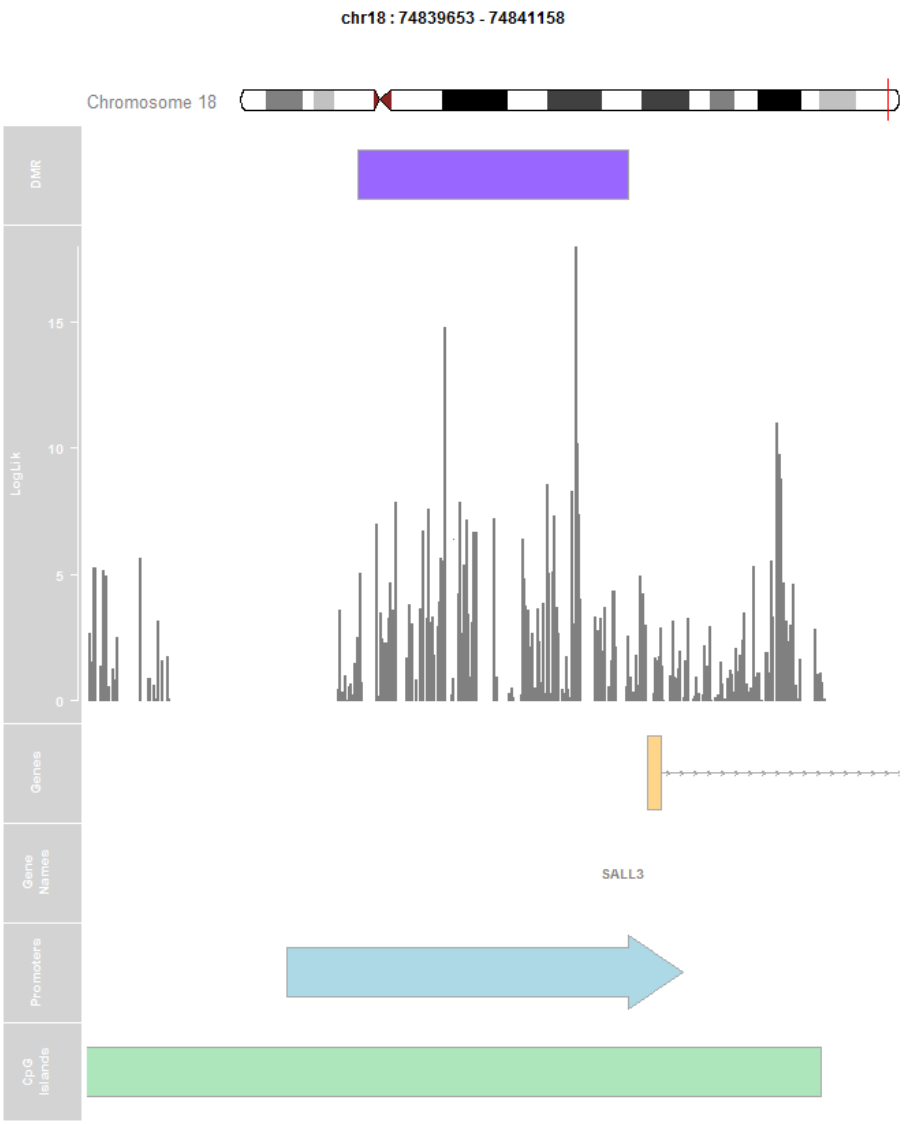
we note that the same number of regions was found, and the bounds of these regions are largely the same. However, the results have changed slightly. This is to expected, as SaTScan finds regions by minimizing the total variance of the data, and what we have specified as the "full" data has changed.

5: DMR visualization

We can now visualize the significant regions with the function **plotSatScanresult**. The results will be output to a pdf in `mydir`, if this was specified, or in the current working directory. The options `plottopn` and `plotpval` can be used to restrict plotting to the top few results by p-value or to results with a specified p-value more conservative than 0.05, respectively. The default is to plot all the significant results returned by **plotSatScanresult**. Alternatively, plots can also be output directly from the call to **plotSatScanresult** by specifying `plotresult=TRUE`.

```
plotSatScanresult(result.b)
```

For example the first region:



References

1. Park, Y., Figueroa, M. E., Rozek, L. S. & Sartor, M. A. MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics 30, 2414-22 (2014).

2. Kulldorff, M. SaTScan: Software for the spatial and space-time scan

statistics. http://www.satscan.org/

3. RSatScan
4. Gviz

Appendix: Creating a methylSigData object