RADMeth

Egor Dolzhenko & Andrew D. Smith, *University of Southern California*

RADMeth: Regression Analysis of Differential Methilation is a software for computing individual differentially methylated sites and genomic regions in data from whole genome bisulfite sequencing (WGBS) experiments.

Assumptions

For rapid differential methylation analysis, RADMeth should be run on a computing cluser with a few hundred available nodes, in which case it takes approxmately a few hours to process a dataset consisting of 30-50 WGBS samples. RADMeth can also be used on a personal workstation, in which case differential methylation analysis will take significantly longer. Note that the actual processing time depends on the coverage of each sample, the number of sites analyzed, and the number of samples in the dataset.

Input data

The input data consists of a proportion table and a design matrix.

- Consider the following **proportion table**:

	control_a	control_b	control_c	case_a	case_b	case_c
chr1:108:109	9 6	10 8	1 1	2 2	2 1	14 1
chr1:114:115	17 7	10 0	14 3	5 1	9 1	7 1
chr1:160:161	12 8	10 5	17 4	15 14	13 6	4 4
chr1:309:310	1 1	1 0	17 12	12 8	2 1	19 8
chr1:499:500	8 4	6 5	15 6	14 10	14 11	15 1
chr1:510:511	0 0	0 0	14 8	4 0	5 3	5 1

  As indicated in the header, this proportion table contains information about 6 WGBS samples: 3 controls and 3 cases. Each row of the table contains information about a CpG site and a proportion of reads mapping over this site in each sample. For example, the first row describes a cytosine within a CpG site located on chromosome 1 at position 108. This site is present in 9 reads in the sample `control_a` and is methylated in 6 of them.

- The design matrix for this dataset describes the structure of the experiment:

	base	case
control_a	1	0
control_b	1	0
control_c	1	0
case_a	1	1
case_b	1	1
case_c	1	1

The samples in this dataset are associated with two factors: `base` and `case`. The first column corresponds to the `base` factor and will always be present in the design matrix. Think of it as stating that all samples have the same baseline mean methylation level. To distingish cases from controls we add another factor `case` (second column). The 1's in this column correspond to the samples which belong to the `cases` group. You can use this design matrix as a template to create design matrices for two-group comparisons involving an arbitrary many samples.

Differential Methylation Analysis with RADMeth

Differential methylation analysis consists of (1) regression, (2) combining significance, and (3) multiple testing adjustment steps. In the following we assume that RADMeth resides in the root of the user's home directory (i.e. `~/radmeth`).

Regression Suppose that the proportion table and the design matrix described in the previous section are stored in files `proportion_table.txt` and `design_matrix.txt`. (These files are provided in the RADMeth's homepage `http://smithlabresearch.org/software/radmeth/`.) The regression step is run with the command:

```
~/radmeth/bin/wand -factor case design_matrix.txt proportion_table.txt > cpgs.bed
```

where `cpgs.bed` is the desired output filename. The `-factor` parameter specifies the factor with respect to which we want to test for the differential methylation. The test factor is `case`, meaning that we are testing for differential methylation between cases and controls. The output is a BED file with rows having the following format

```
chrom   start   end   c:log-odds-ratio:mean-meth-diff   pval
```

giving the methylation log-odds ratio, methylation difference between cases and controls, and the p-value from the (log likelihood ratio) test for differential methylation. Here are the first few lines of the output for our example:

```
chr1  108  109  c:-1.94116:0.449947     0.157971
chr1  114  115  c:-0.508053:0.0778541   0.559191
chr1  160  161  c:1.45929:0.325764      0.0951122
chr1  309  310  c:-0.712528:0.16905     0.239772
chr1  499  500  c:-0.252674:0.0629282   0.77014
chr1  510  511  c:-1.20397:0.285714     0.151844
                ...
```

We do not use these p-values directly, but instead we adjust the p-value of each CpG site based on the p-values of the neighboring CpGs.

Combining significance and adjusting for multiple testing Both of these steps are performed simultaneously. Given the `cpgs.bed` file from the previous step, run

```
~/radmeth/bin/adjust -bins 1:100:1 cpgs.bed > cpgs.adjusted.bed
```

Here, the only required parameter, besides the input file, is `-bins` whose value is set to `1:100:1`. This means that for each $n = 1, 2, \ldots 99$, RADMeth will compute correlation between p-values of cpgs located at distance n from each other. These correlations are used during significance combination step. In addition, bin sizes determine the window for combining significance. In contrast, if `-bins` is set to `1:15:5`, then the correlation is computed separately for p-values corresponding to CpGs at distances $[1, 5)$, $[5, 10)$, and $[10, 15)$ from one another.

The output of `adjust` has this format:

```
chrom   start   end   c:log-odds-ratio:mean-meth-diff:pval:combined-pval   fdr-pval
```

where all of the parameters, except `combined-pval` and `fdr-pval`, are as before; `combined-pval` is the p-value given by the Z test which combines `pvals` from proximal CpG sites and `fdr-pval` is the FDR corrected `combined-pval`.

Here is what the `cpgs.adjusted.bed` file looks like for our example dataset:

```
chr1  108  109  c:-1.94116:0.449947:0.157971:0.126815     0.405801
chr1  114  115  c:-0.508053:0.0778541:0.559191:0.126815   0.405801
chr1  160  161  c:1.45929:0.325764:0.0951122:0.126815     0.405801
chr1  309  310  c:-0.712528:0.16905:0.239772:0.239772     0.490947
chr1  499  500  c:-0.252674:0.0629282:0.77014:0.422756    0.608339
chr1  510  511  c:-1.20397:0.285714:0.151844:0.422756     0.608339
                ...
```

Invidual differentially methylated sites After completing the previous steps, individual differentially methylated sites can be obtained with the `awk` utility, present on virtually all Linux and Mac OS systems. To get all CpGs with FDR-corrected p-value below 0.01, run

```
awk  '$5 < 0.01 "{ print $0; $}"' cpgs.adjusted.bed > dm_cpgs.bed
```

Differentially methylated regions RADMeth can also join individually differentially methylated CpGs into differentially methylated regions. This can be achived with the command

```
~/radmeth/bin/dmrs -p 0.01 cpgs.adjusted.bed > dmrs.bed
```

The current DMR algorithm is concervative: it joins neighboring differentially methylated sites with p-value below 0.01 (set by the `-p` paramter).

The output format is

```
chrom   start   end   c:log-odds-ratio:mean-meth-diff   num-sites
```

where `log-odds-ratio` and `mean-meth-diff` are computed by averaging the corresponding parameters of individual differentially methylated sites in the region. The number of sites comprising the DMR is given by `num-sites`.

For our example, the output looks like this:

```
chr1  57409   57689   dmr:-2.79935:0.355322  18
chr1  58282   59009   dmr:-2.36703:0.309342  57
chr1  138548  139044  dmr:-1.95574:0.373333  22
                      ...
```

Obtaining the proportion table from the output MethPipe methylomes

RADMeth includes a program to combine the methylome samples generated by the MethPipe methylation analysis pipeline [1]. The MethPipe methylomes are specified like so

```
chr1  108  +  CpG  1.0    2
chr1  114  +  CpG  0.2    5
chr1  160  +  CpG  0.934  15
chr1  309  +  CpG  0.667  12
chr1  499  +  CpG  0.714  14
chr1  510  +  CpG  0.0    4
                ...
```

The information about each site includes chromosome, position, strand (always positive), type of the site, estimatated methylation level (obtained directly from the read proportion), and coverage.

The sample methylomes from our example can be combined into a proportion table with the command

```
~/radmeth/bin/make_table control_a control_b control_c case_a case_b case_c > proportion_table.txt
```

Note that the samples are included in the proportion table in order in which they are listed.

Splitting proportion tables for analysis on multicore systems

The regression step of the differential methylation analysis is by far the most time consuming. The analysis can be substantially sped up by splitting the proportion table into smaller tables, separately processing each individual table, and subsequently combining the results.

To split a proportion table stored in a file `proportion_table.txt` run

```
~/radmeth/bin/split_table -num_rows 1000000 proportion_table.txt
```

The proportion table is split into smaller tables `chunk_table_1.txt`, `chunk_table_2.txt`, ... each containing 1 million rows of the original table. The resulting tables can be processed in parallel as before

```
~/radmeth/bin/wand -factor case design_matrix.txt chunk_table_1.txt > cpgs_1.bed
```

Once each table is processed, the results must be combined together like so

```
cat cpgs_*.bed | sort -k 1,1 -k 2,2n -k 3,3n > cpgs.bed
```

The file `cpgs.bed` can now be processed as before.

References

[1] Song, Qiang, et al. "A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics." PloS one 8.12 (2013): e81148.