

DMRforPairs vignette

Martin A. Rijlaarsdam

April 16, 2015

1 Overview

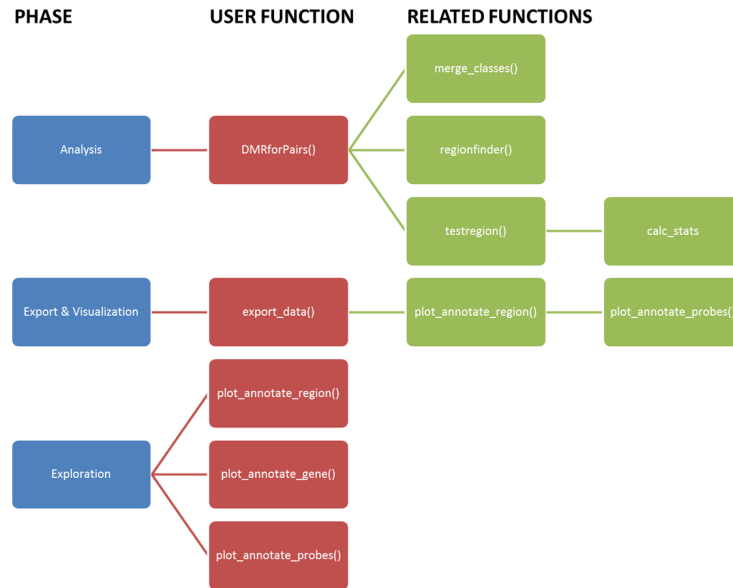


Figure 1: Flowchart of the DMRforPairs algorithm. The user progresses vertically through the pipeline while the algorithm uses the subsequent functions as indicated by horizontal connectors. The tuning function for the DMRforPairs parameters (`min_n` and `min_distance`) is not depicted, but can be used to explore the number of regions identified/probes included for various pairs of settings.

This is a demo illustrating the usage of DMRforPairs (Figure 1). DMRforPairs is designed to identify Differentially Methylated Regions between unique samples using array based methylation profiles. Regions are identified as genomic ranges with sufficient probes located in close proximity to each other and which are optionally annotated to the same functional class (see reference manual, `merge_classes()` function). Differential methylation is evaluated by comparing the methylation values within each region between individual samples and

(if the difference is sufficiently large), testing this difference formally for statistical significance.

The following remarks apply to the vignette:

1. the "annotate.significant" and "annotate" parameters have been set to FALSE to facilitate the speed of building the package/vignette and to allow running the vignette on computers that are not connected to the internet. The power of DMRforPairs is however greatly enhanced by the Gviz based visualizations that include annotation information from Ensembl. Therefore, it is advised to change these parameters to TRUE if internet is available. We recommend annotate.relevant to be set to FALSE at all times unless very small sections of the genome are analyzed (see documentation)
2. the vignette requires writing permissions in the working directory.
3. parallelization has been disabled in all examples as well as in this vignette. This is done to provide polite code for people sharing compute cycles.

2 Setting up the data and settings

2.1 The Data

Load the data. This dataset provides the average methylation values on chromosome 7 of two commercially available EBV transfected lymphoblastoid cell lines from healthy individuals (NA17105 (African American male) and NA17018 (Chinese female)). The dataset also contains this data for the breast cancer cell line MCF7 (Soule et al. 1973) and the HPV negative squamous-cell vulva carcinoma cell line A431 (Giard et al. 1973 and Hietanen et al. 1995). For a full description of the dataset (+references) and its format, please see the reference manual.

```
> library(DMRforPairs)
> data(DMRforPairs_data)
```

Columns 1 - 6 of the data indicate information about individual probes (n=29,974) and their annotation. Columns 7 - 10 indicate M-values for all samples and columns 11 - 14 represent the associated beta values.

```
> head(CL.methy,2)
```

```

      targetID chromosome position class.gene.related
cg00081087 cg00081087          7 34873912      Body;Body;Body
cg00087298 cg00087298          7 33149137          TSS200
      class.island.related      gene.symbol      A431.M      MCF7.M
cg00081087                      AAA1;NPSR1;NPSR1 -4.027028 -0.4932408
cg00087298      Island                      RP9 -13.872031 -12.2916392
      NA17105.M NA17018.M      A431.beta      MCF7.beta NA17105.beta
cg00081087 -6.300535 -2.475343 5.782836e-02 0.4153514568 0.012827340
cg00087298 -8.549864 -12.007505 6.669415e-05 0.0003185493 0.003003211
      NA17018.beta
cg00081087 0.1524629099
cg00087298 0.0004419503
```

2.2 The Settings

First, a number of possible settings for the `min_distance` (the maximal distance accepted between probes in a region) and `min_n` (the minimal number of probes in a region) are evaluated.

```

> parameters=expand.grid(min_distance = c(200,300), min_n = c(4,5))
> recode=1
> results.parameters = tune_parameters(parameters,
+   classes_gene=CL.methy$class.gene.related,
+   classes_island=CL.methy$class.island.related,
+   targetID=CL.methy$targetID,
+   chr=CL.methy$chromosome,
+   position=CL.methy$position,
+   m.v=CL.methy[,c(7:8)],
+   beta.v=CL.methy[,c(11:12)],
+   recode=recode,
+   gs=CL.methy$gene.symbol,
+   do.parallel=0)
> results.parameters
```

| | min_distance | min_n | n.regions | n.valid.probes | n.probes.included |
|------|--------------|-------|-----------|----------------|-------------------|
| [1,] | 200 | 4 | 40 | 589 | 178 |
| [2,] | 300 | 4 | 50 | 589 | 228 |
| [3,] | 200 | 5 | 32 | 589 | 154 |
| [4,] | 300 | 5 | 43 | 589 | 216 |

In the rest of this vignette, the default setting of minimally 4 probes per region is used. These have to be in < 200 bp distance of each other. The threshold for a relevant median difference in M value between the samples is set to 1.4. Benjamini Hochberg corrected p-values <0.05 are deemed significant. The parameter `experiment` sets name of the experiment which is reflected in the name

of the folder with results that will be created in the working directory.

```
> min_n=4
> d=200
> dM=1.4
> pval_th=0.05
> experiment="results_DMRforPairs_vignette"
> method="fdr"
> clr=c("red","blue","green")
```

3 Run DMRforPairs

The algorithm is most conveniently executed by calling the wrapper for the analysis part (DMRforPairs()) which returns the results of all the separate steps. DMRforPairs runs automatically, showing regular status updates. Analysis can take quite long, especially on a genome wide scale (several hours). The demo data should generally finish within a few minutes. The wrapper subsequently performs:

1. Recoding of the probe classes according to a custom or build in scheme.
2. Identification of regions with sufficient probe density (i.e. number of probes and proximity) over all genomic regions at which probes are annotated in the dataset .
3. Calculation of relevant statistics (e.g. median difference in M and beta values) and performing of formal tests to see if the difference is significant.

These steps are extensively described in the reference manual.

```
> output=DMRforPairs(
+ classes_gene=CL.methy$class.gene.related,
+ classes_island=CL.methy$class.island.related,
+ targetID=CL.methy$targetID,
+ chr=CL.methy$chromosome,
+ position=CL.methy$position,
+ m.v=CL.methy[,c(8:10)],
+ beta.v=CL.methy[,c(12:14)],
+ min_n=min_n,min_distance=d,min_dM=dM,
+ recode=recode,
+ sep=";",
+ method=method,
+ debug.v=FALSE,gs=CL.methy$gene.symbol,
+ do.parallel=0)
```

3.1 Examining the primary output of DMRforPairs

3.1.1 Recode probe classes (merge_classes())

Original probe classes...

```
> head(output$classes$pclass,3)
```

```
      [,1]  
cg00081087 "Body;Body;Body;"  
cg00087298 "TSS200;Island"  
cg00139681 ";"
```

Recoded probe classes...

```
> head(output$classes$pclass_recoded,3)
```

```
      [,1]  
cg00081087 "gene;NA;NA"  
cg00087298 "NA;tss;island"  
cg00139681 "NA;NA;NA"
```

Row numbers of probes without a recoded class...

```
> head(output$classes$no.pclass,10)
```

```
[1]  3  6 13 15 17 33 38 41 46 50
```

Classes used for recoding...

```
> output$classes$u_pclass
```

```
[1] gene  tss  island  
Levels: gene island tss
```

Merge classes returns a reduced set of probe data (annotation, M and beta values) including only probes associated with at least one recoded class. In case of recode=2 this implicates all probes in the dataset. This reduced set of probes is designated "valid" in the remainder of this vignette and in the reference manual.

3.1.2 Identify probe-dense regions (regionfinder())

Potential regions of interest...

```
> head(output$regions$boundaries,4)
```

| | chr | start_bp | end_bp | length_bp | n_probes | regionID | regionIDall | ClassAll |
|---|-----|-----------|-----------|-----------|----------|----------|-------------|-------------|
| 1 | 7 | 33080496 | 33080615 | 120 | 4 | 1 | 1 | gene |
| 2 | 7 | 34118464 | 34118935 | 472 | 5 | 2 | 2;38 | gene;island |
| 3 | 7 | 34873912 | 34874196 | 285 | 4 | 3 | 3 | gene |
| 4 | 7 | 105172664 | 105173132 | 469 | 5 | 4 | 4 | gene |

Probes with associated class after recoding (valid probes)...

```
> head(output$regions$valid.probes,2)
```

| | rowID | probeID | chr | position | pClass |
|------------|-------|------------|-----|----------|---------------|
| cg00081087 | 1 | cg00081087 | 7 | 34873912 | gene;NA;NA |
| cg00087298 | 2 | cg00087298 | 7 | 33149137 | NA;tss;island |

Associated m and beta values for all samples for each valid probe...

```
> head(output$regions$valid.m,2)
```

| | MCF7.M | NA17105.M | NA17018.M |
|------------|-------------|-----------|------------|
| cg00081087 | -0.4932408 | -6.300535 | -2.475343 |
| cg00087298 | -12.2916392 | -8.549864 | -12.007505 |

```
> head(output$regions$valid.beta,2)
```

| | MCF7.beta | NA17105.beta | NA17018.beta |
|------------|--------------|--------------|--------------|
| cg00081087 | 0.4153514568 | 0.012827340 | 0.1524629099 |
| cg00087298 | 0.0003185493 | 0.003003211 | 0.0004419503 |

Region to probe map: matrix of valid probes (rows) and recoded probe classes (columns) with either NA if not included in any potential region of interest or the ID of the region the probe is assigned to. By definition each probe can only be associated to one region per class. Region IDs are specific to a dataset and a set of DMRforPairs parameters. Region IDs are therefore not interchangeable between datasets/experiments and primarily serve as identifiers during exploration of the dataset.

```
> head(output$regions$perprobe,4)
```

| | gene | tss | island |
|------------|------|-----|--------|
| cg00081087 | 3 | NA | NA |
| cg00087298 | NA | 14 | 35 |
| cg00156506 | NA | NA | NA |
| cg00280235 | NA | NA | NA |

3.1.3 Calculation of relevant statistics and testing (testregion())

This output is structured like `output$regions$boundaries` but is supplemented with descriptive statistics and formal test results per region.

```
> head(output$tested,1)

chr start_bp end_bp length_bp n_probes regionID regionIDall ClassAll
1 7 33080496 33080615 120 4 1 1 gene
beta.median.MCF7.beta beta.median.NA17105.beta beta.median.NA17018.beta
1 0.9848456 0.003448186 0.003629177
m.median.MCF7.M m.median.NA17105.M m.median.NA17018.M
1 6.094547 -9.110979 -8.158476
median.delta.beta.MCF7.M.minus.NA17105.M
1 0.9693837
median.delta.beta.MCF7.M.minus.NA17018.M
1 0.9685009
median.delta.beta.NA17105.M.minus.NA17018.M
1 -0.0001267329
median.delta.m.MCF7.M.minus.NA17105.M median.delta.m.MCF7.M.minus.NA17018.M
1 14.6789 13.85861
median.delta.m.NA17105.M.minus.NA17018.M pairwise.p.MCF7.M.vs.NA17105.M
1 -0.5443935 0.02857143
pairwise.p.MCF7.M.vs.NA17018.M pairwise.p.NA17105.M.vs.NA17018.M
1 0.02857143 0.1142857
max.abs.median.delta p.value p.value.adjusted
1 14.6789 0.01246768 0.03562195
```

4 Export and visualization

The `export_data()` function performs a complete export of all results to TSV, pdf and png files for all (relevant) regions. These overviews are generated in increasing detail for:

1. all regions
2. regions with a relevant difference ($> dM$) and
3. regions with a significant difference.

HTML tables are used to access the results and describe the analysis (Figure 2). Thumbnails of the methylation pattern of a region are presented in the tables (2 and 3) as well as general statistics. Links to detailed statistics (tsv) and (pairwise) visualizations (pdf) are provided. Regions with a relevant difference can be looked up in the Ensembl database resulting in annotated figures of the methylation pattern. Also, direct links to the regions in the Ensembl and UCSC

DMRforPairs output generated on Mon Mar 31 09:31:40 2014. Identified regions were set to contain at least 4 probes with a maximum distance of 200 bp between individual probes (n=40). Regions in which median methylation levels (M-values) between the samples differed at least 1.41 (n=20) (=relevant) were tested for statistical significance (significant: p<0.05; multiple testing adjusted, n=7). The following samples were studied: MCF7.M,NA17105.M,NA17018.M. Chromosomal positions are indicated in bp. N indicates the number of probes in a region. ID indicates the region ID. Values in the columns per sample indicate median beta values. dM indicates the largest median difference (absolute) between any of the sample pairs. p indicates uncorrected p-value from Mann-Whitney U test (n=2) or Kruskal Wallis test (n>2). p.adj denotes the multiple testing corrected p-value (method fdr). Gene symbols of overlapping transcripts are listed for the exact region and within a margin of 10000 bp of the region.

| Thumbnail | Chr | Start | End | Links | Length | ID | Class | MCF7.beta | NA17105.beta | NA17018.beta | Gene.Symbol | dM | p | p.adj |
|-----------|-----|-----------|-----------|---|--------|-----|-------------|-----------|--------------|--------------|--|-------|---------|--------|
| | 7 | 106505688 | 106505961 | PDF STATS ENSEMBL UCSC | 274 | 346 | island | 0.966 | 0.01175 | 0.00440 | PIK3CG (margin: PIK3CG) | 15.99 | 0.00815 | 0.0272 |
| | 7 | 33080496 | 33080615 | PDF STATS ENSEMBL UCSC | 120 | 411 | gene | 0.985 | 0.00345 | 0.00363 | AVL9, NT5C3A (margin: AVL9, NT5C3A, AC074338.4, RNTSL505P) | 14.68 | 0.01247 | 0.0356 |
| | 7 | 107300939 | 107301048 | PDF STATS ENSEMBL UCSC | 110 | 529 | nas | 0.899 | 0.01680 | 0.02933 | SLC26A4, SLC26A4-AS1 (margin: SLC26A4, SLC26A4-AS1) | 8.96 | 0.00752 | 0.0272 |
| | 7 | 107300939 | 107301281 | PDF STATS ENSEMBL UCSC | 350 | 710 | gene:island | 0.899 | 0.01680 | 0.02933 | SLC26A4, SLC26A4-AS1 (margin: SLC26A4, SLC26A4-AS1) | 8.79 | 0.00120 | 0.0130 |

Figure 2: Example of the HTML output of DMRforPairs

genome browsers are presented. By default, DMRforPairs creates a folder (experiment.name) within the current working directory for the output (export_data() function). This is done because a complete export generates a large number of files. Visualization and export can take quite long depending on the status of biomaRt (Ensembl).

```
> tested_inclannot=export_data(
+   tested=output$tested,
+   regions=output$regions,
+   th=pval_th,min_n=min_n,min_dM=dM,min_distance=d,
+   margin=10000,clr=clr,method=method,experiment.name=experiment,
+   annotate.relevant=FALSE,annotate.significant=FALSE,
+   FigsNotRelevant=FALSE,debug=FALSE)
```

Please see the "results_DMRforPairs_vignette" folder in your working directory for the output of the vignette.

PIK3CG was one of the genes strongly differentially methylated in 1 of the samples relative to the other two. By clicking on the PDF link, the output region can be further studied (Figure 3). Additional statistics are accessible via the STATS link in the HTML table (significant.html).

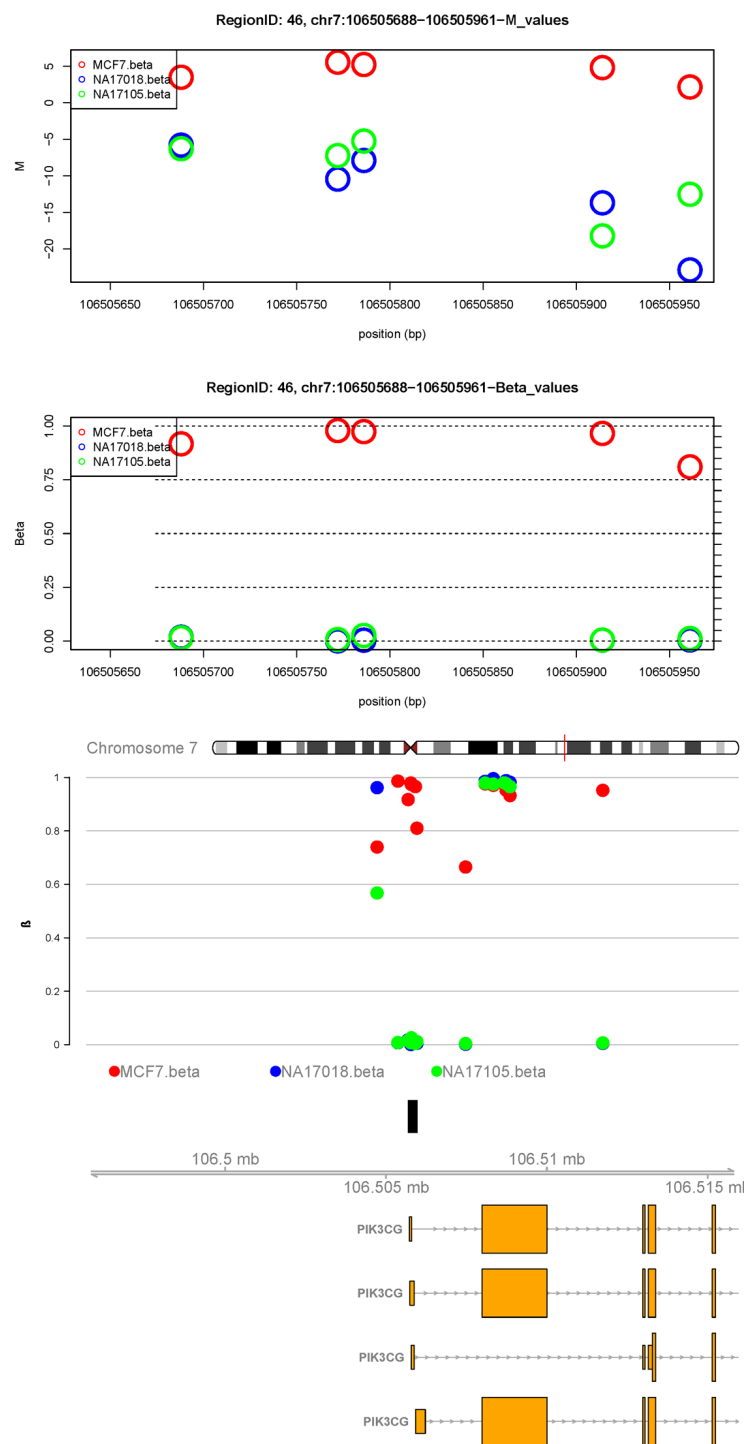


Figure 3: Differentially methylated region around the TSS of PIK3CG.

4.1 Examining the data further

There are also several functions to further explore the data based on the findings after export. These will be discussed in this section. For example, region 16 was highly relevant (median delta M=13.77). However, because of limited statistical power ($n=4$) the region did not survive correction for multiple testing. We might want to inquire this region further using the `plot_annotate_region()` function. By default relevant, but non-significant regions like 16 are not annotated. If we set `annotate` to `TRUE` in the example below we can appreciate that even though the number of probes is low (technical bias), the sudden consistent difference between MCF7 occurs right around the transcription start site of BMPER and the surrounding probes do not show this differential pattern (Figure 4). The `plot_annotate_region()` function also reports back the complete set of statistics and pairwise plots for the requested region.

```

> plot_annotate_region(output$tested,
+                       output$regions,
+                       margin=10000,
+                       regionID=16,
+                       clr=clr,
+                       annotate=FALSE,
+                       scores=TRUE,
+                       path=experiment)

$symbols_exact
[1] ""

$symbols_margin
[1] ""

$scores

```

| | [,1] |
|---|---------|
| beta.median.MCF7.beta | 0.9383 |
| beta.median.NA17018.beta | 0.0088 |
| beta.median.NA17105.beta | 0.0022 |
| m.median.MCF7.M | 4.0696 |
| m.median.NA17018.M | -7.5344 |
| m.median.NA17105.M | -9.4626 |
| median.delta.beta.MCF7.M.minus.NA17018.M | 0.9244 |
| median.delta.beta.MCF7.M.minus.NA17105.M | 0.9272 |
| median.delta.beta.NA17018.M.minus.NA17105.M | 0.0016 |
| median.delta.m.MCF7.M.minus.NA17018.M | 11.8639 |
| median.delta.m.MCF7.M.minus.NA17105.M | 13.7677 |
| median.delta.m.NA17018.M.minus.NA17105.M | 1.3083 |
| pairwise.p.MCF7.M.vs.NA17018.M | 0.0286 |
| pairwise.p.MCF7.M.vs.NA17105.M | 0.0286 |
| pairwise.p.NA17018.M.vs.NA17105.M | 0.6857 |
| max.abs.median.delta.m | 13.7677 |
| p.value | 0.0231 |
| n.probes | 4.0000 |

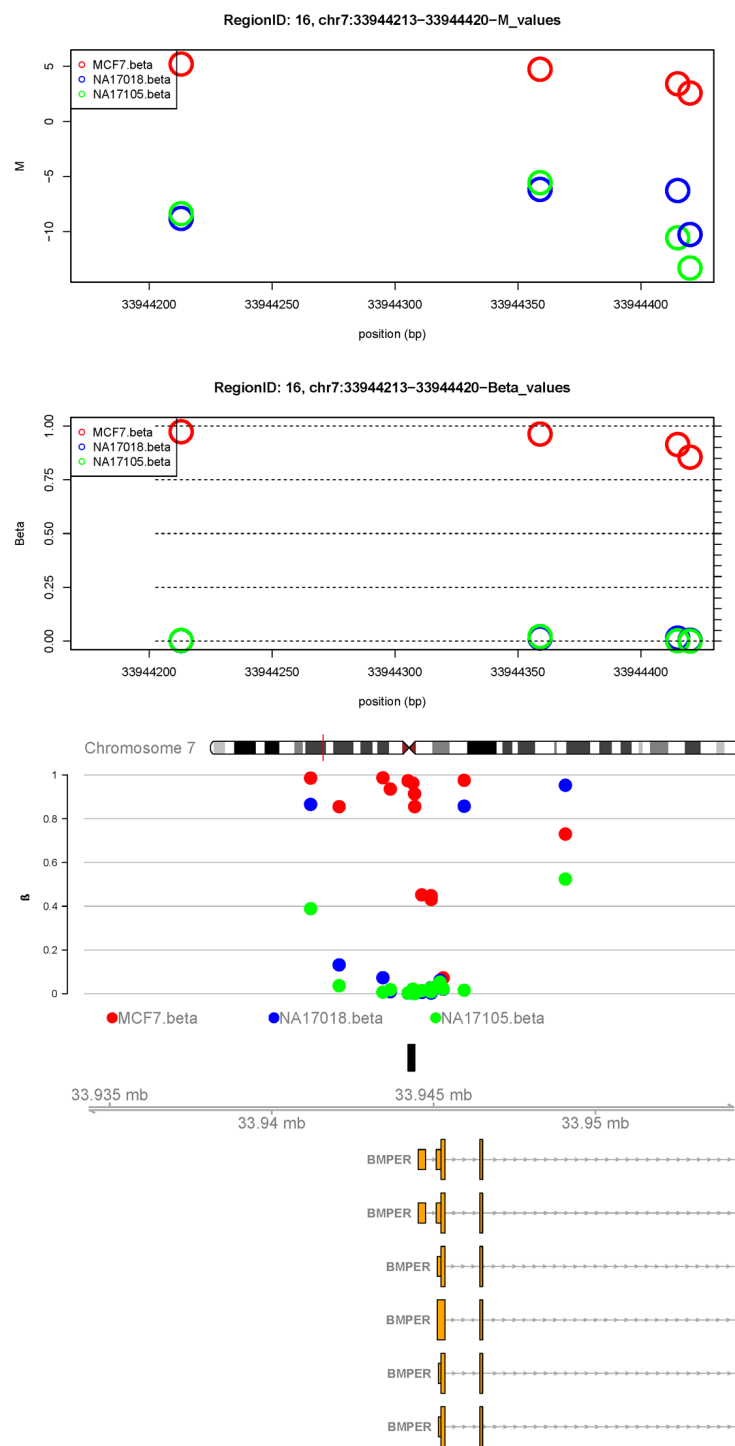


Figure 4: Differentially methylated region 16 - relevant, but not significant.

The DMRforPairs script also contains a wrapper to visualize the methylation pattern in and around a specific gene (gene symbol) as long as the gene symbol is annotated in the Illumina manifest. The `plot_annotate_gene()` function also reports back the complete set of statistics and pairwise plots for that gene. We will follow up on BMPER here (Figure 5).

```
> plot_annotate_gene(gs="BMPER",
+                     regions=output$regions,
+                     margin=10000,
+                     ID="BMPER",
+                     clr=clr,
+                     annotate=FALSE,
+                     path=experiment)

$symbols_exact
[1] ""

$symbols_margin
[1] ""

$scores
                                [,1]
beta.median.MCF7.beta          0.67506214
beta.median.NA17018.beta       0.27153857
beta.median.NA17105.beta       0.02429967
m.median.MCF7.M                1.05783234
m.median.NA17018.M             -1.44791040
m.median.NA17105.M             -5.35393171
median.delta.beta.MCF7.M.minus.NA17018.M  0.26227534
median.delta.beta.MCF7.M.minus.NA17105.M  0.36175373
median.delta.beta.NA17018.M.minus.NA17105.M 0.01144306
median.delta.m.MCF7.M.minus.NA17018.M     2.38364051
median.delta.m.MCF7.M.minus.NA17105.M     4.83587754
median.delta.m.NA17018.M.minus.NA17105.M  1.03137346
pairwise.p.MCF7.M.vs.NA17018.M            0.00006834
pairwise.p.MCF7.M.vs.NA17105.M            0.00000002
pairwise.p.NA17018.M.vs.NA17105.M         0.06660043
max.abs.median.delta.m                4.83587754
p.value                               0.00000017
n.probes                             40.00000000
```

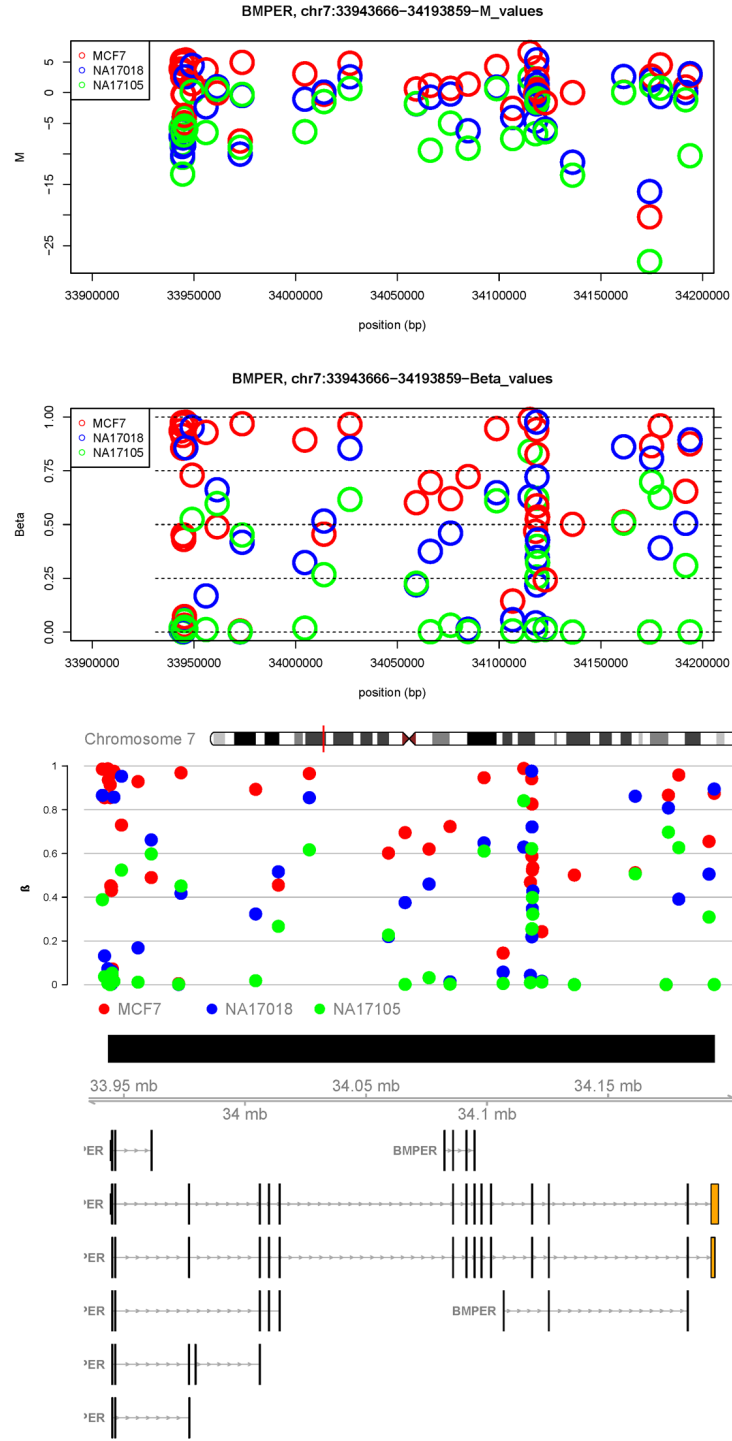


Figure 5: Methylation throughout the BMPER gene.

DMRforPairs can also visualize custom genomic regions. The example below basically generates a zoomed in version of the whole BMPER gene and shows that only the promoter region (before TSS) is differentially methylated (Figure 6). This overlaps with the region selected by DMRforPairs (region 16). The `plot_annotate_custom_region()` function also reports back the complete set of statistics for the requested custom genomic region.

```
> plot_annotate_custom_region(chr=7,
+                             st=33943000,
+                             ed=33945000,
+                             output$regions,
+                             margin=500,
+                             ID="BMPER_TSS",
+                             clr=clr, annotate=FALSE,
+                             path=experiment)

$symbols_exact
[1] ""

$symbols_margin
[1] ""

$scores
                                [,1]
beta.median.MCF7.beta           0.913892
beta.median.NA17018.beta        0.009059
beta.median.NA17105.beta        0.008536
m.median.MCF7.M                 3.409922
m.median.NA17018.M              -7.167840
m.median.NA17105.M              -6.908633
median.delta.beta.MCF7.M.minus.NA17018.M  0.900301
median.delta.beta.MCF7.M.minus.NA17105.M  0.912505
median.delta.beta.NA17018.M.minus.NA17105.M -0.000737
median.delta.m.MCF7.M.minus.NA17018.M    9.973318
median.delta.m.MCF7.M.minus.NA17105.M    10.288598
median.delta.m.NA17018.M.minus.NA17105.M -0.429142
pairwise.p.MCF7.M.vs.NA17018.M           0.000041
pairwise.p.MCF7.M.vs.NA17105.M           0.000041
pairwise.p.NA17018.M.vs.NA17105.M        0.931427
max.abs.median.delta.m              10.288598
p.value                             0.000170
n.probes                             9.000000
```

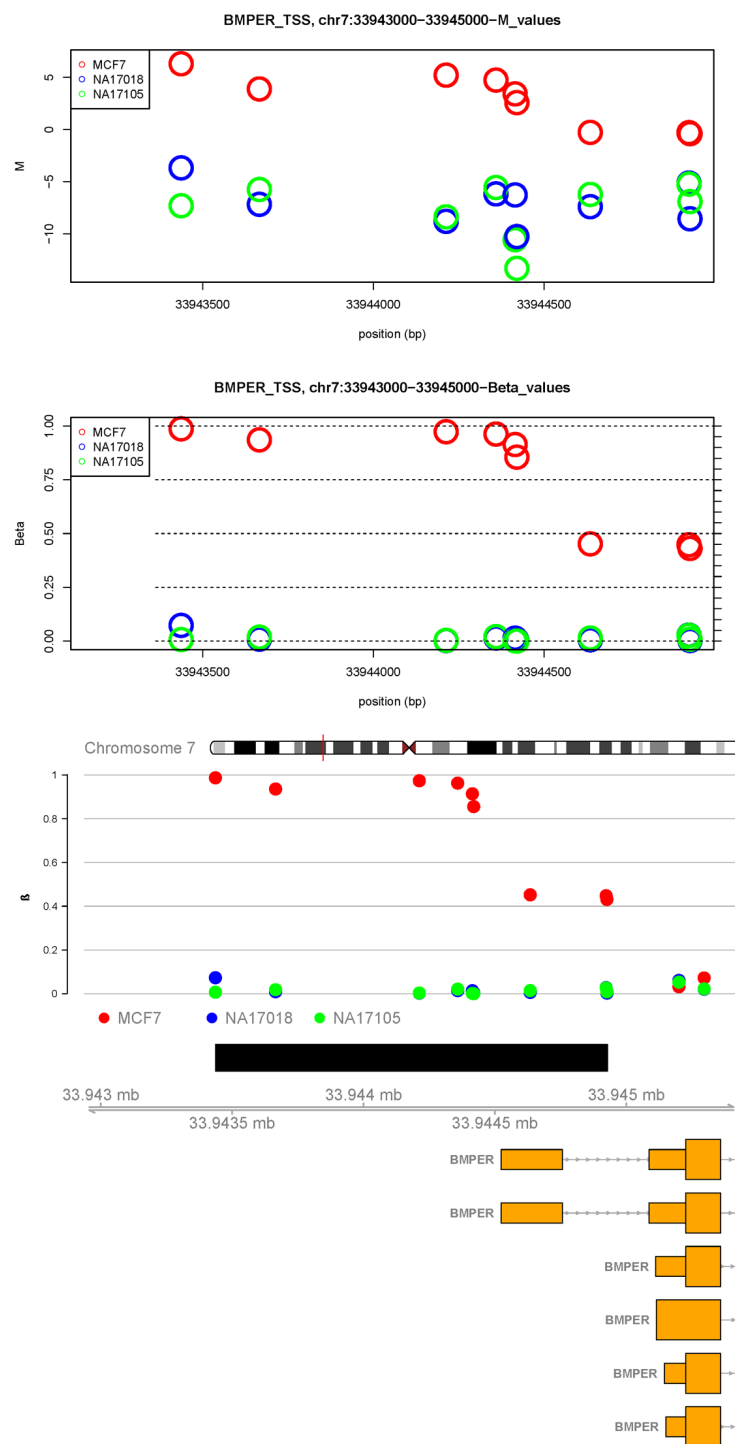


Figure 6: Methylation around the TSS of BMPER.