Program 'bedgraph2dmr.R'

Ryan Quan | rcq2102@columbia.edu

Created: August 14, 2014 | Modified: August 20, 2014

Version 0.1

Description Analyzes and plots bisulfite sequencing data using bsseq package.

Depends bsseq, stringr, magrittr

Author Ryan Quan

Maintainer Ryan Quan <rcq2102@columbia.edu

How To Use

1. In the same directory as `bedgraph2dmr_master.R`, create a folder named `data`.
2. (Optional) Place CSV of amplicon locations into current working directory. See below for how this CSV should be formatted.
3. Place all `.bedGraph` files in the `data` folder.
4. Open `bedgraph2dmr_master.R`.
5. Change parameters as needed.
6. Run `bedgraph2dmr_master.R`.

Inputs

- bedGraph files from Bismark methylation extractor
- (Optional) CSV of targeted amplicon locations

Note: You may name the file however you wish as long as it is the only CSV in the working directory. See figure below on how to format this file.

amplicon	batch	start	end	chr
APC_b_amp4	1	112043031	112043181	chr5
APC_c_amp4	1	112043203	112043359	chr5
APC_d_amp4	1	112043373	112043601	chr5
APC_e_amp4	1	112043736	112043968	chr5
BRCA1_a_an	1	41278260	41278381	chr17
BRCA1_b_an	1	41278355	41278501	chr17

Figure 1: Amplicon CSV Format

Outputs

Tables

- `dmr_all.csv` - all significant DMRs at specified FDR
- `dmr_subset.csv` - a subset of significant DMRs after applying cutoffs for mean methylation difference and number of sites
- `tstat_ttestValues.csv` - t-statistic for methylation loci

Plots

If amplicon file is supplied...

- Plots of each amplicon region with relative methylation values (points) and smoothed methylation values (lines).
- Tumor/cases are red while normals/controls are blue. The line at the bottom represents the t-statistic for the smoothed methylation values.
- Highlighted regions in red have been determined to be statistically significant at the pre-specified rate and meets all the assumptions provided by the user.

If amplicon file is not supplied...

- Plots each statistically significant DMR that meets all the assumptions provided by the user.

Parameters

`BSmooth`

- `ns` - minimum number of methylation loci in a smoothing window

- `h` - minimum smoothing window, in bases
- `maxGap` - maximum gap between two methylation loci, before the smoothing is broken across the gap.
- `mc.cores` - number of cores to use to apply smoothing algorithm

Defaults chosen using this publication.

`subset_by_type`

- `min_cov` - the minimum number of reads a methylation loci must have to be included in the analysis

`get_tstat`

- `est_var` - how the variance is estimated. T-statistics are formed as the difference in means between group 1 and group 2 divided by an estimate of the standard deviation, assuming that the variance in the two groups are the same (same), that we have paired samples (paired) or only estimate the variance based on group 2 (group2).

`export_dmr_data`

Settings for dmrFinder

- `FDR` - the false discovery rate
- `max_gap` - how dmrFinder determines CpG clusters. If set to 1, dmrFinder will output individual methylation loci that are significant.

Subsetting DMR results

- `cg-num` - minimum number of CpG sites the DMR should have
- `mean_diff` - minimum mean methylation difference between two DMRs

Settings for Plots

- `batch_num` - subsets the amplicon file by batch number to generate plots only for amplicon regions within that batch
- `min_cov` - minimum coverage of reads needed to be included on graph (applies to relative methylation values)

Session Info

```
R version 3.1.1 (2014-07-10)
Platform: x86_64-apple-darwin13.1.0 (64-bit)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] stringr_0.6.2        magrittr_1.0.1       bsseq_1.0.0
[4] matrixStats_0.10.0   GenomicRanges_1.16.4 GenomeInfoDb_1.0.2
[7] IRanges_1.22.10      BiocGenerics_0.10.0
```