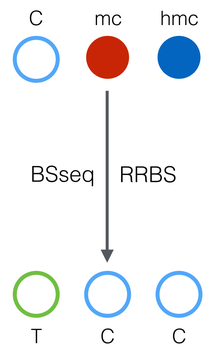
mint + annotatr: a pipeline to integrate and annotate DNA methylation and hydroxymethylation data

Raymond G. Cavalcante[1], Yanxiao Zhang[1], Yongseok Park[2], Snehal Patil[1], Maureen A. Sartor[1,3,4]

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, USA, [2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, USA [3] Department of Biostatistics, University of Michigan, Ann Arbor, USA, and [4] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, USA

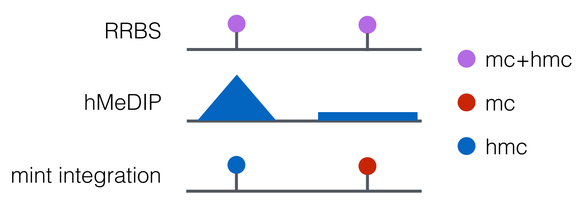
Motivation for integration

DNA methylation occurs in a variety of forms. 5-methylcytosine (mc) is the most widely studied, and there is ample evidence for its importance in development and gene regulation. 5-hydroxymethylcytosine (hmc) is a less abundant, but appears to be a good marker for epigenetic changes correlating with changes in gene expression.



The problem: Widely-used assays measuring methylation (e.g. BSseq & RRBS) do not distinguish mc and hmc. This may confuse biological interpretation because it is unclear which mark is playing a role relative to a phenotype.

Integrating assays measuring mc+hmc (e.g. BSseq & RRBS) with others measuring mc (e.g. oxBSseq & MeDIPseq) or hmc (e.g. TABseq & hMeDIPseq), helps differentiate the mc and hmc marks from one another.



The solution: We developed the mint pipeline to
- be a complete pipeline from raw reads to interpretation
- do sample methylation quantification
- do differential methylation analysis between groups
- differentiate between regions of mc vs hmc by integrating data from methylation assays using a simple classifier
- automatically generate UCSC genome browser tracks
- provide summaries of genomic annotations to facilitate biological interpretation.

Supported designs

Hybrid +

Bisulfite-based assay and pulldown assay.

or

Pulldown +

Pulldown assays for mc and hmc.

and

Sample-wise

Integration on a per-sample basis.

and or

Comparison-wise

Integration on a per-group basis.

Setting up and running mint

On a Mac or Linux system:

1. Get mint at github.com/sartorlab/mint
2. Setup dependencies and reference genomes.
3. Annotate the experimental design.
4. Initiate the project with:
   `Rscript init.R –project name –genome g –datapath path`
5. Customize the `config.mk` file with desired parameters.
6. Use `make` to run the analysis modules as per the design:

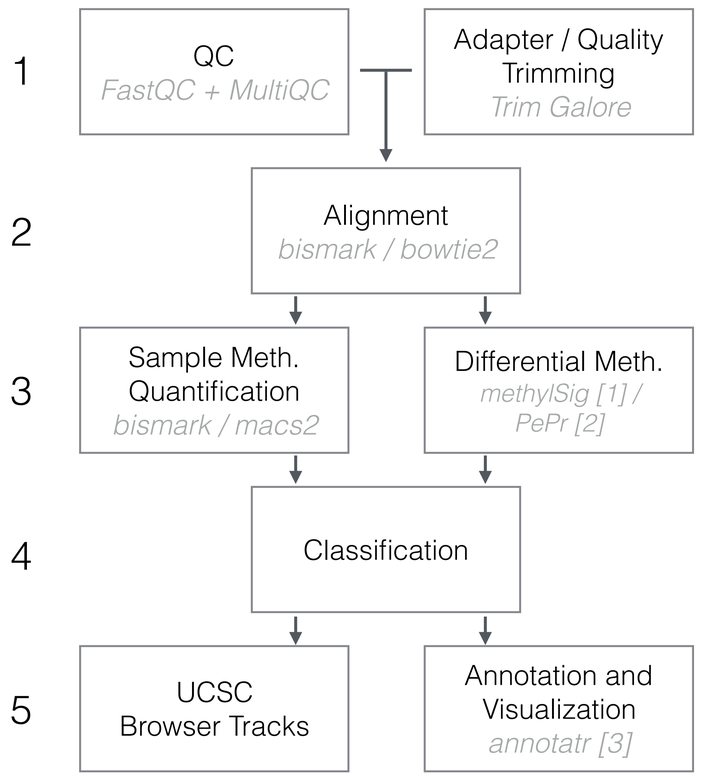
```
make bisulfite_align          make pulldown_align
make bisulfite_compare        make pulldown_sample
make compare_classification   make pulldown_compare
                              make sample_classification
```

References

1. Park Y, *et al.* (2014) “MethylSig: a whole genome DNA methylation analysis pipeline” *Bioinformatics.*
2. Zhang Y, *et al.* (2014) “PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data” *Bioinformatics.*
3. Cavalcante RG & Sartor MA (2016) “annotatr: Associating genomic regions with genomic annotations” *bioRxiv.*

This work was funded by NIH grants RO1CA158286S1 and T32GM070499.

Overall mint workflow



Classification schemas

Sample classification

	hmc peak	No hmc peak	No signal
High hmc + mc	hmc	mc	hmc or mc
Low hmc + mc	hmc	mc (low)	hmc or mc (low)
No hmc + mc	hmc	no methylation	no methylation
No signal	hmc	no methylation	unclassifiable

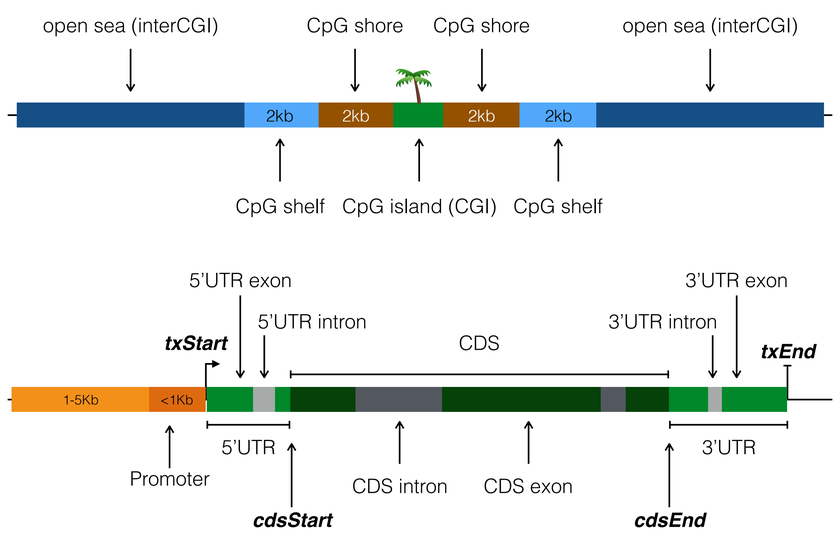
Comparison classification

wrt to condition 1	Hyper hmc	Hypo hmc	No DM	No signal
Hyper hmc + mc	Hyper mc Hyper hmc	Hyper mc Hypo hmc	Hyper mc	Hyper mc
Hypo hmc + mc	Hypo mc Hyper hmc	Hypo mc Hypo hmc	Hypo mc	Hypo mc
No DM	Hyper hmc	Hypo hmc	No DM	No DM
No signal	Hyper hmc	Hypo hmc	No DM	unclassifiable

annotatr: simple, fast & flexible annotation of genomic regions

We developed a general purpose R package, *annotatr* [3] that:
- annotates genomic regions in BED format to pre-built human and mouse CpG, genic (with Entrez Gene IDs and symbols), and enhancers (below), or custom genomic annotations
- reports <u>all</u> annotations overlapping input genomic regions to highlight multiple-annotations rather than using a prioritization
- summarizes and plots annotations with associated data (right)
- is faster than similar R packages, and on par with bedtools.

Get annotatr at github.com/rcavalcante/annotatr

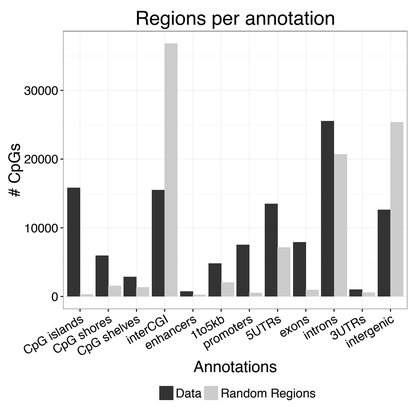


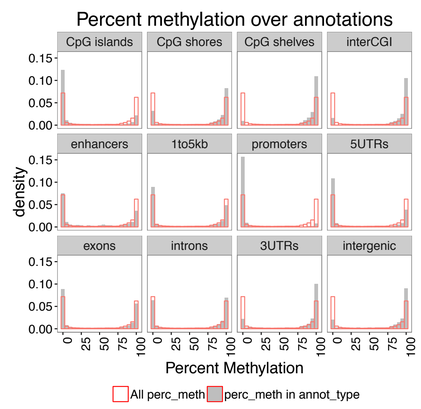
Annotation and Visualization

Quality control

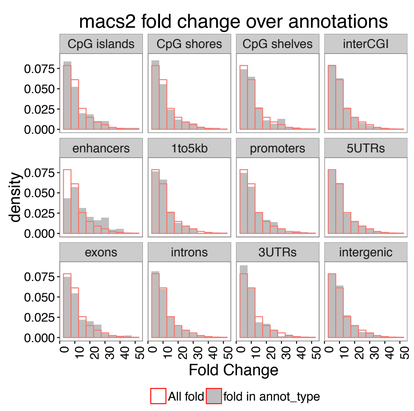
Sample Name	% mCpG	M C's	% Aligned	% Trimmed	% Dups	M Seqs
IDH2mut_1_mc_hmc_bisulfite				0.1%	85.6%	0.7
IDH2mut_1_mc_hmc_bisulfite_trimmed					84.9%	0.7
IDH2mut_1_mc_hmc_bisulfite_trimmed_bismark_bt2_SE_report			99.7%			
IDH2mut_1_mc_hmc_bisulfite_trimmed_bismark_bt2_splitting_report	37.5%	10.2				
IDH2mut_2_mc_hmc_bisulfite				0.0%	91.3%	1.9
IDH2mut_2_mc_hmc_bisulfite_trimmed					91.3%	1.9
IDH2mut_2_mc_hmc_bisulfite_trimmed_bismark_bt2_SE_report			99.9%			
IDH2mut_2_mc_hmc_bisulfite_trimmed_bismark_bt2_splitting_report	31.6%	29.0				

bismark methylation extractor

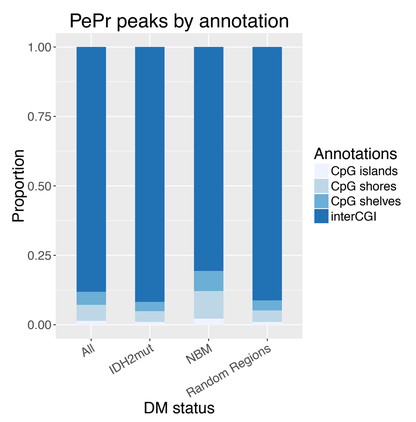




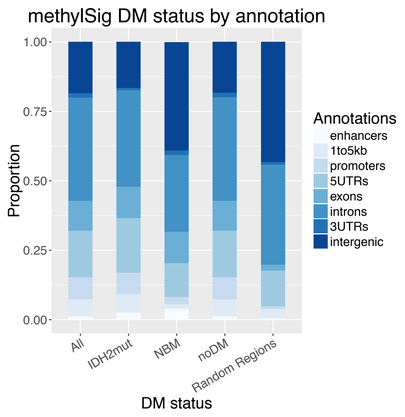
macs2

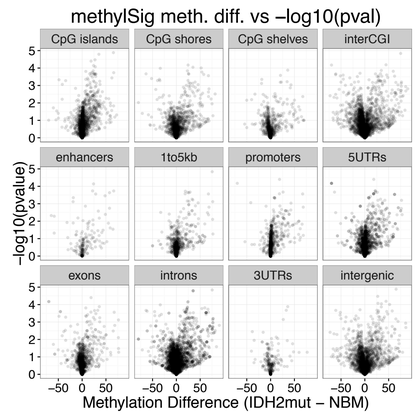


PePr

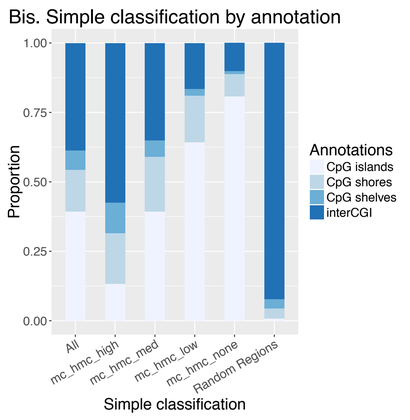


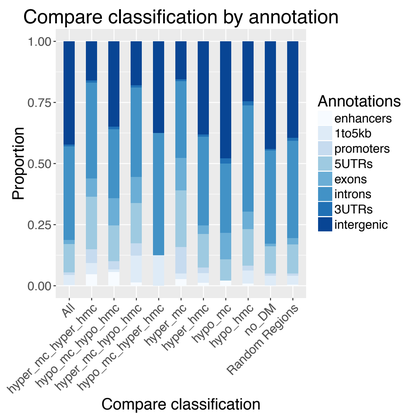
methylSig





Classifications





UCSC Genome Browser Track Hub