Integrating DNA methylation and hydroxymethylation data with the mint pipeline

Raymond G. Cavalcante[1], Snehal Patil[1], Laura S. Rozek[2], and Maureen A. Sartor[1]

[1]Department of Computational Medicine and Bioinformatics, and [2]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI

Supplementary Information

Supplementary Table 1 – Example of sample metadata and covariate information to be used in setting up a mint project. The table should be tab-delimited and placed in the mint/projects folder with a filename of the form [projectID]_samples.txt. This ensures that the init.R initialization script looks at the proper metadata table for the project. The sampleID column will often not be human readable (i.e. automatically named .fastq.gz files provided by a sequencing core, GEO, or SRA). The humanID column is meant to connect human understandable names to the automatically generated IDs. The pulldown, bisulfite, mc, hmc, and input columns are binary where 0 means no and 1 means yes. In the example below, any bisulfite sample has a 1 in the mc and hmc columns to indicate that the platform (in this case ERRBS) cannot distinguish between them. The group column can contain multiple comma-separated numbers if a sample belongs to more than one group (e.g. "1,2"). Columns appearing after the group column are considered covariates to be used in the models used for differential methylation testing with csaw and DSS. Column headers for covariate columns *must match* the variables as they appear in the model and covariate columns of the comparisons table (Table S2).

projectID	sampleID	humanID	pulldown	bisulfite	mc	hmc	input	group	subject	age
test_hybrid	IDH2mut_1_hmeseal	IDH2mut_1	1	0	0	1	0	1	1	3
test_hybrid	IDH2mut_2_hmeseal	IDH2mut_2	1	0	0	1	0	1	2	3
test_hybrid	IDH2mut_1_hmeseal_input	IDH2mut_1	1	0	0	1	1	1	1	3
test_hybrid	IDH2mut_2_hmeseal_input	IDH2mut_2	1	0	0	1	1	1	2	3
test_hybrid	IDH2mut_1_errbs	IDH2mut_1	0	1	1	1	0	1	1	3
test_hybrid	IDH2mut_2_errbs	IDH2mut_2	0	1	1	1	0	1	2	3
test_hybrid	NBM_1_hmeseal	NBM_1	1	0	0	1	0	0	1	10
test_hybrid	NBM_2_hmeseal	NBM_2	1	0	0	1	0	0	2	10
test_hybrid	NBM_1_hmeseal_input	NBM_1	1	0	0	1	1	0	1	10
test_hybrid	NBM_2_hmeseal_input	NBM_2	1	0	0	1	1	0	2	10
test_hybrid	NBM_1_errbs	NBM_1	0	1	1	1	0	0	1	10
test_hybrid	NBM_2_errbs	NBM_2	0	1	1	1	0	0	2	10

Supplementary Table 2 – Example of comparison metadata and model information to be used in setting up a mint project. The purpose of this table is to encode information needed for testing differential methylation with csaw and/or DSS with a filename of the form [projectID]_comparisons.txt. The pulldown, bisulfite, mc, and hmc columns are as in Supplementary Table 1. Here, the input column takes values of TRUE or FALSE and indicates whether the input for a comparison of IP data should be used to filter out windows for analysis in csaw. The model column is used to build the design matrix. The contrast column should be a binary vector indicating which coefficient from the model to test in csaw and DSS. The covariates column lists the covariates used in the model formula (comma-delimited if more than one and NA if none). The entries of this columns should also match the column headings in the sample matrix (Table S1). The covIsNumeric column indicates whether the covariate is numerical (1) or categorical (0). The groups column indicates the group numbers from the sample matrix (Table S1) to use for the test. The interpretation is a comma-delimited list indicating what interpretation to give to regions with logFC (csaw) or methdiff (DSS) < 0 (first entry) or >= 0 (second entry).

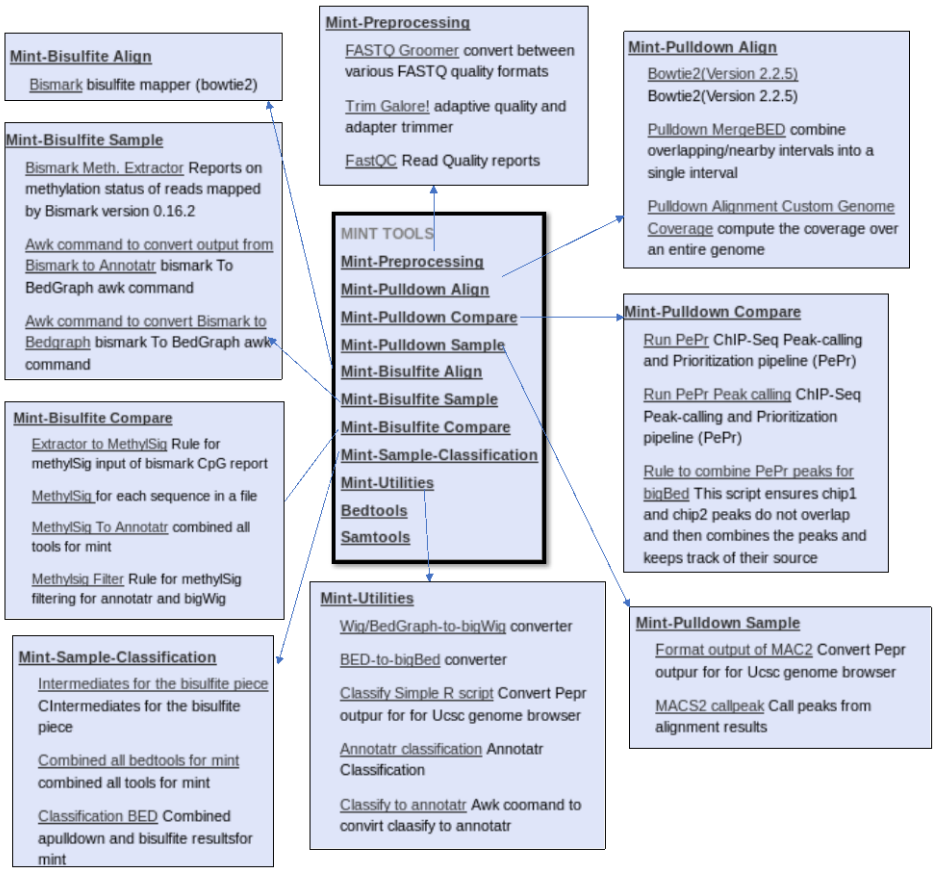
projectID	comparison	pulldown	bisulfite	mc	hmc	input	model	contrast	covariates	covIsNumeric	groups	interpretation
test_hybrid	IDH2mut_v_NBM	1	0	0	1	TRUE	~1+group	0,1	NA	0	0,1	NBM,IDH2mut
test_hybrid	IDH2mut_v_NBM	0	1	1	1	FALSE	~1+group	0,1	NA	0	0,1	NBM,IDH2mut
test_hybrid	IDH2mut_v_NBM_paired	1	0	0	1	TRUE	~1+group+subject	0,1,0	subject	0	0,1	NBM,IDH2mut
test_hybrid	IDH2mut_v_NBM_paired	0	1	1	1	FALSE	~1+group+subject	0,1,0	subject	0	0,1	NBM,IDH2mut
test_hybrid	IDH2mut_v_NBM_cont	1	0	0	1	TRUE	~1+group+age	0,1,0	age	1	0,1	NBM,IDH2mut
test_hybrid	IDH2mut_v_NBM_cont	0	1	1	1	FALSE	~1+group+age	0,1,0	age	1	0,1	NBM,IDH2mut

Supplementary Table 3 – Classification scheme for integrating methylation and hydroxymethylation data. **(A) The sample-wise classifier.** Rows are classifications given to 5mC + 5hmC signal from WGBS or RRBS and columns are 5hmC signal from hMeDIP-seq or hMe-Seal. The classifier operates on the intersection of the two signal tracks. Regions of no signal are determined either by the lack of coverage (5mC + 5hmC from WGBS or RRBS) or a lack of input coverage (5hmC from hMeDIP-seq or hMe-Seal). **(B) The comparison-wise classifier.** Rows are classifications given to 5mC + 5hmC differential methylation signal from DSS. Columns are 5hmC differential methylation signal from csaw. The classifier operates on the intersection of the two signal tracks. Hyper/hypo is written with respect to condition 1 of the comparison.

A.	**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

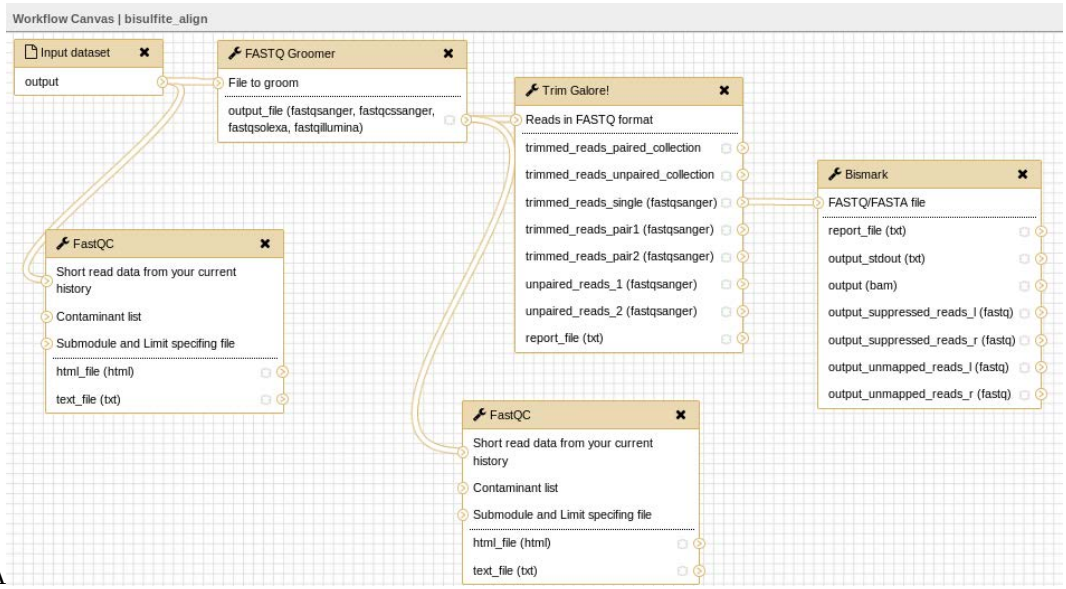
B.	Hyper hmc	Hypo hmc	No DM	No signal
Hyper hmc + mc	Hyper mc Hyper hmc	Hyper mc Hypo hmc	Hyper mc	Hyper mc
Hypo hmc + mc	Hypo mc Hyper hmc	Hypo mc Hypo hmc	Hypo mc	Hypo mc
No DM	Hyper hmc	Hypo hmc	No DM	No DM
No signal	Hyper hmc	Hypo hmc	No DM	unclassifiable

Supplementary Figure 1 – Overview of Galaxy tools used in the mint pipeline. The tools are organized in modules similar to the command line, with a flow based on Figure 1C. Users can create workflows based on the individual tools, or Galaxy workflows are provided (Figure S2).



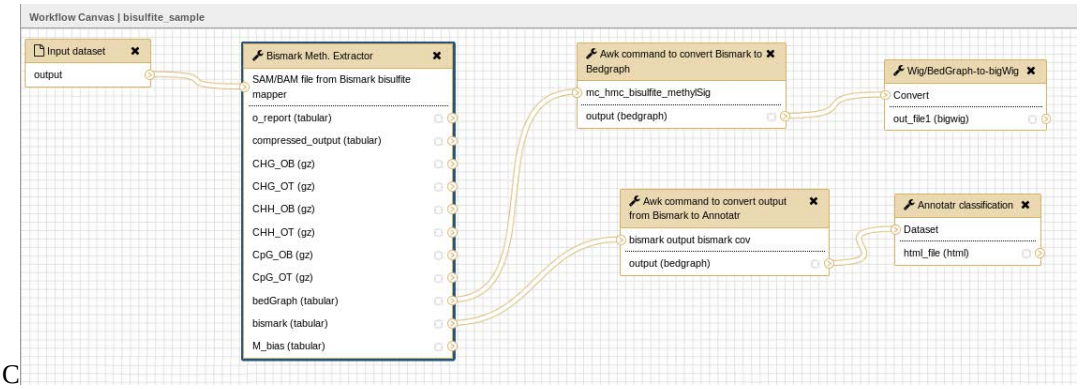
Supplementary Figure 2 – Galaxy workflow screenshots. Workflows in Galaxy function as pipelines, and the Galaxy implementation of mint includes workflows for each of the modules in the command line version. (A) Workflow for bisulfite_align, (B) pulldown_align, (C) bisulfite_sample, (D) pulldown_sample, (E) bisulfite_compare, (F) pulldown_compare, (G) sample_classification, and (H) compare_classification modules.

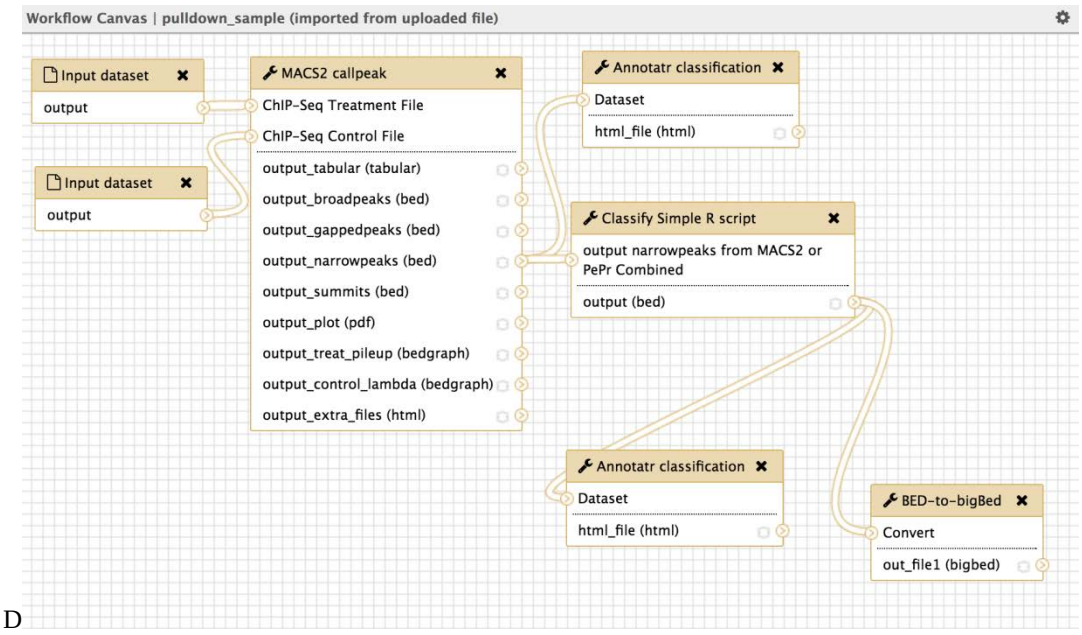
A



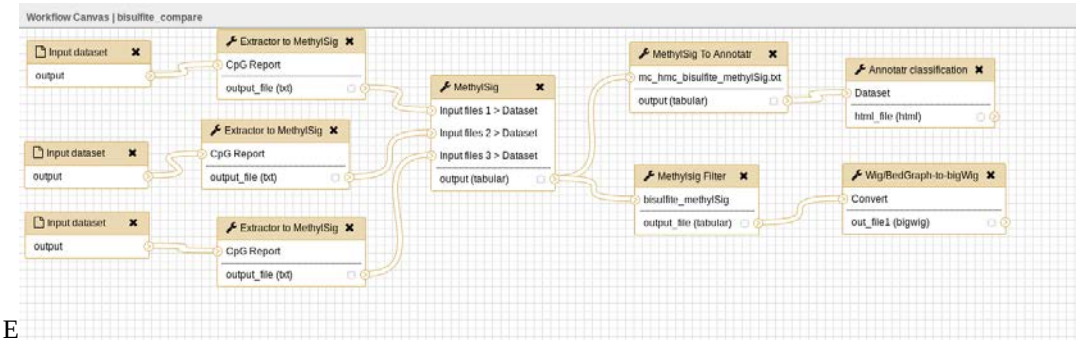
B

C

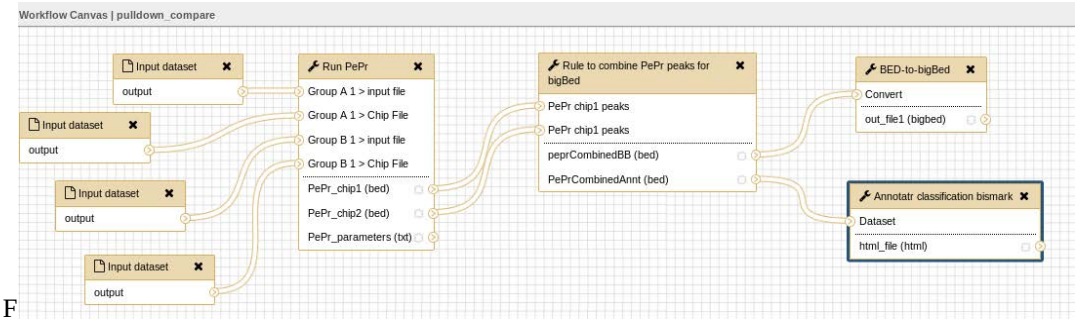



Workflow Canvas | pulldown_sample (imported from uploaded file)
Input dataset
output
Input dataset
output
MACS2 callpeak
ChIP-Seq Treatment File
ChIP-Seq Control File
output_tabular (tabular)
output_broadpeaks (bed)
output_gappedpeaks (bed)
output_narrowpeaks (bed)
output_summits (bed)
output_plot (pdf)
output_treat_pileup (bedgraph)
output_control_lambda (bedgraph)
output_extra_files (html)
Annotatr classification
Dataset
html_file (html)
Classify Simple R script
output narrowpeaks from MACS2 or PePr Combined
output (bed)
Annotatr classification
Dataset
html_file (html)
BED-to-bigBed
Convert
out_file1 (bigbed)

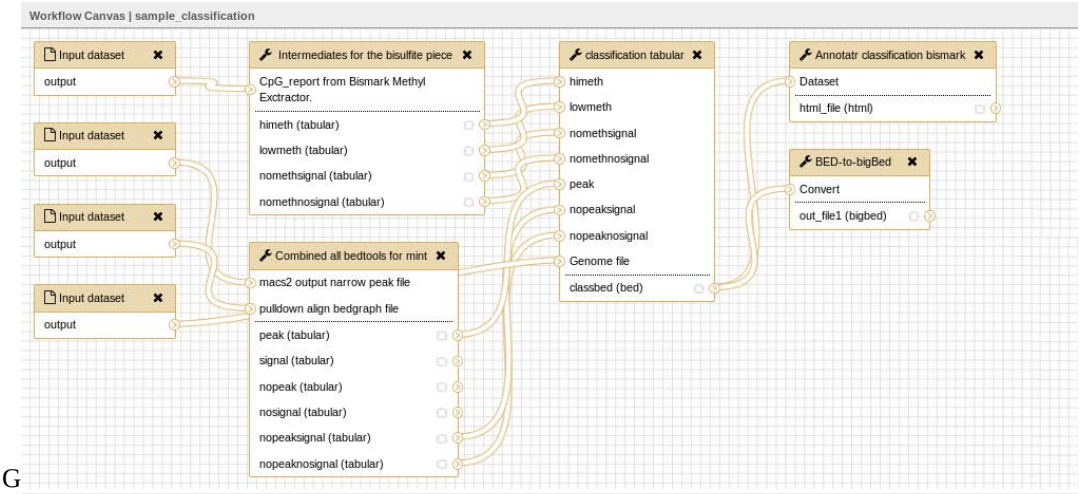
D


Workflow Canvas | bisulfite_compare
Input dataset
output
Input dataset
output
Input dataset
output
Extractor to MethylSig
CpG Report
output_file (txt)
Extractor to MethylSig
CpG Report
output_file (txt)
Extractor to MethylSig
CpG Report
output_file (txt)
MethylSig
Input files 1 > Dataset
Input files 2 > Dataset
Input files 3 > Dataset
output (tabular)
MethylSig To Annotatr
mc_hmc_bisulfite_methylSig.txt
output (tabular)
Methylsig Filter
bisulfite_methylSig
output_file (tabular)
Annotatr classification
Dataset
html_file (html)
Wig/BedGraph-to-bigWig
Convert
out_file1 (bigwig)

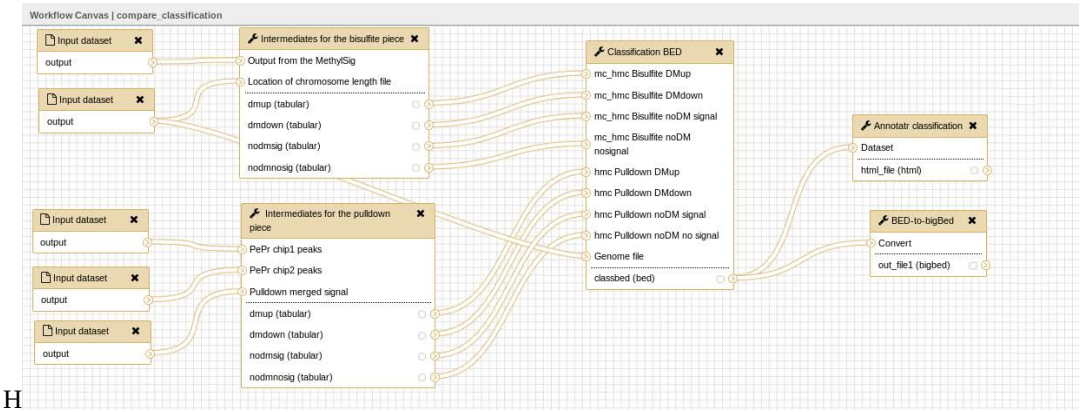
E


Workflow Canvas | pulldown_compare
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Run PePr
Group A 1 > input file
Group A 1 > Chip File
Group B 1 > input file
Group B 1 > Chip File
PePr_chip1 (bed)
PePr_chip2 (bed)
PePr_parameters (txt)
Rule to combine PePr peaks for bigBed
PePr chip1 peaks
PePr chip1 peaks
peprCombinedBB (bed)
PePrCombinedAnnt (bed)
BED-to-bigBed
Convert
out_file1 (bigbed)
Annotatr classification bismark
Dataset
html_file (html)

F

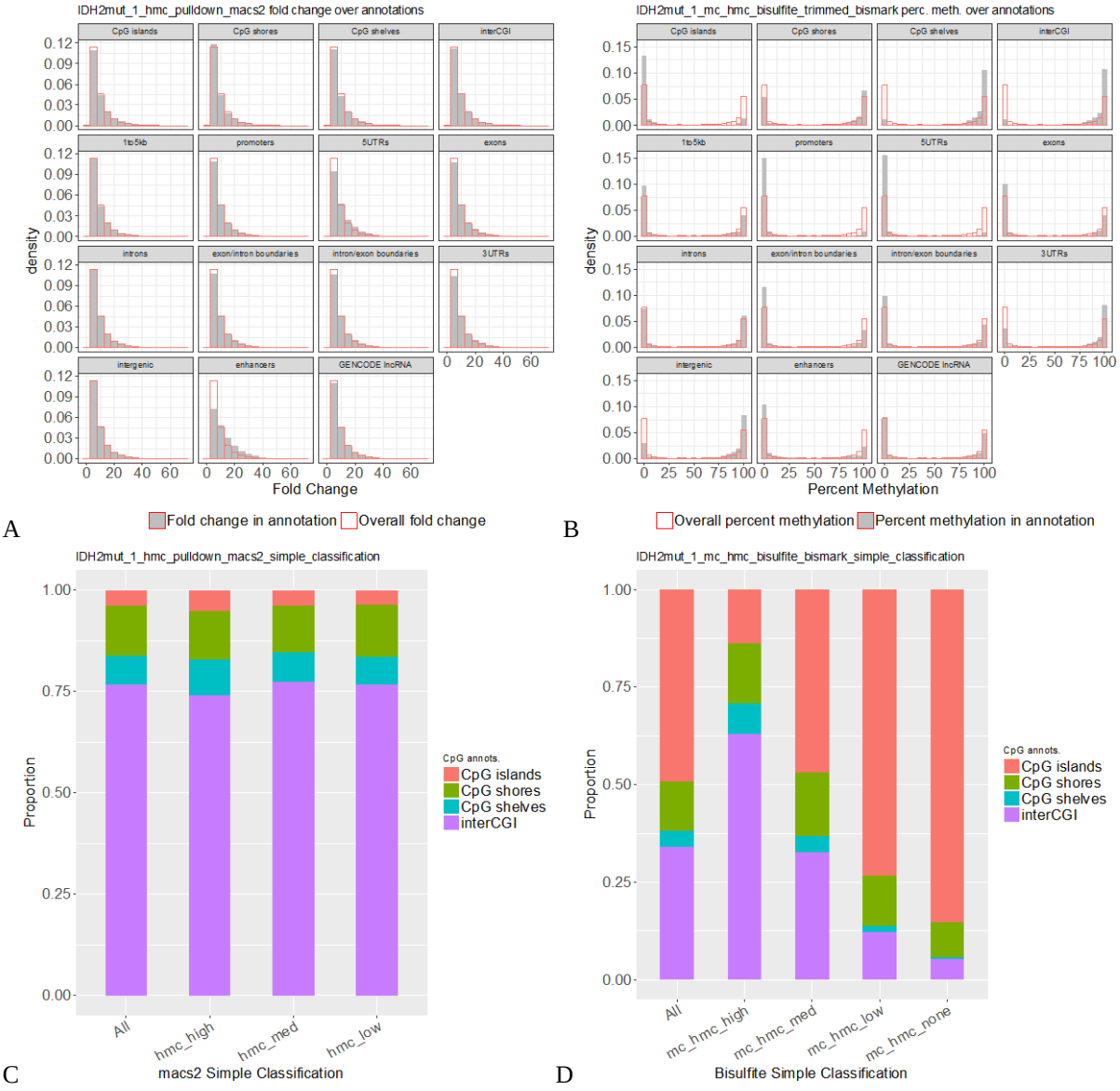

Workflow Canvas | sample_classification
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Intermediates for the bisulfite piece
CpG_report from Bismark Methyl Exctractor.
himeth (tabular)
lowmeth (tabular)
nomethsignal (tabular)
nomethnosignal (tabular)
Combined all bedtools for mint
macs2 output narrow peak file
pulldown align bedgraph file
peak (tabular)
signal (tabular)
nopeak (tabular)
nosignal (tabular)
nopeaksignal (tabular)
nopeaknosignal (tabular)
classification tabular
himeth
lowmeth
nomethsignal
nomethnosignal
peak
nopeaksignal
nopeaknosignal
Genome file
classbed (bed)
Annotatr classification bismark
Dataset
html_file (html)
BED-to-bigBed
Convert
out_file1 (bigbed)

G

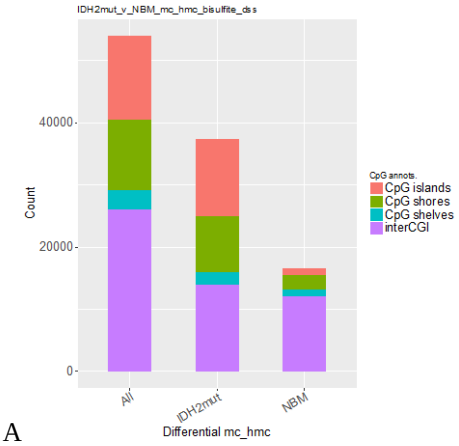

Workflow Canvas | compare_classification
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Intermediates for the bisulfite piece
Output from the MethylSig
Location of chromosome length file
dmup (tabular)
dmdown (tabular)
nodmsig (tabular)
nodmnosig (tabular)
Intermediates for the pulldown piece
PePr chip1 peaks
PePr chip2 peaks
Pulldown merged signal
dmup (tabular)
dmdown (tabular)
nodmsig (tabular)
nodmnosig (tabular)
Classification BED
mc_hmc Bisulfite DMup
mc_hmc Bisulfite DMdown
mc_hmc Bisulfite noDM signal
mc_hmc Bisulfite noDM nosignal
hmc Pulldown DMup
hmc Pulldown DMdown
hmc Pulldown noDM signal
hmc Pulldown noDM no signal
Genome file
classbed (bed)
Annotatr classification
Dataset
html_file (html)
BED-to-bigBed
Convert
out_file1 (bigbed)

H

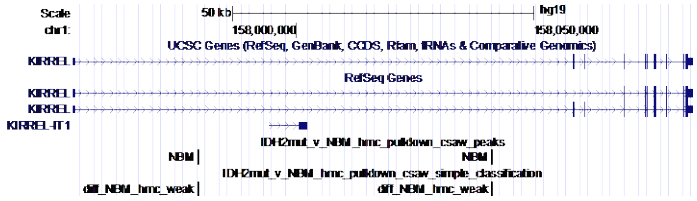
Supplementary Figure 3 – Selected outputs from the pulldown_sample and bisulfite_sample modules. Each graphic is part of the automatic output at the genome annotation step for both sample modules. Sample IDH2mut_1 is shown as the example. (A) Fold change of macs2 peaks measuring hydroxymethylation across genomic annotations. Gray bars denote the fold change distribution for peaks annotated to the feature labeling the facet, and red outlines are the overall distribution of peak fold changes. Of note is the hyper-hydroxymethylation present in peaks annotated to enhancers and 5'UTRs compared to background. (B) Percent methylation of CpGs across genomic annotations. Gray bars and red outlines are as in panel A. Of note is the hypo-methylation of CpGs annotated to enhancers and 5'UTRs relative to background, especially in light of corresponding hyper-hydroxymethylation. (C) Annotations of the 'simple classification' (low, medium, high) of hydroxymethylation peaks are similarly distributed across CpG features regardless of peak strength. (D) Annotations of the 'simple classification' (no, low, medium, high) of percent methylation of CpGs have different distributions across CpG features according to strength of methylation. In particular, as methylation weakens, it tends to be located more in CpG islands (orange), but less in CpG shelves (blue).



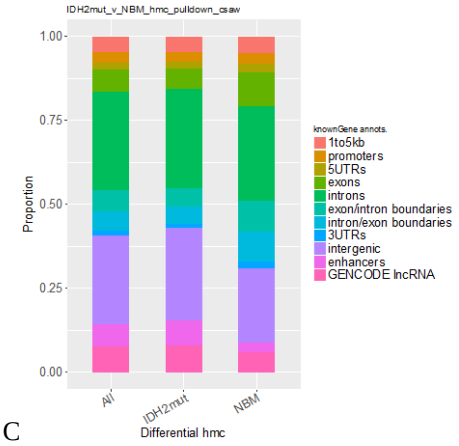
Supplementary Figure 4 – Selected outputs from the compare (A – E) and classification (F – H) modules comparing IDH2 mutants with NBM samples. (A) Number of DMRs found by DSS, and annotated to CpG island features. The number of hyper-methylated DMRs in the IDH2 samples is greater than those in NBM samples, in line with previous findings. **(B)** The Genome Browser with csaw track showing 'NBM' peaks (hypo-hydroxymethylated in IDH2 mutants) as was found in the paper originally describing the AML data. **(C)** Hypo-hydroxymethylated regions in IDH2 mutants occur more frequently at 5' ends of genes and exons than hyper-hydroxymethylated regions. **(D)** Conversely, hyper-methylated regions in IDH2 mutants occur more frequently at 5' ends of genes and exons than hypo-methylated regions. **(E)** Hyper-methylated regions occur more frequently at or near CpG islands than hypo-methylated regions. **(F)** Genomic annotation to CpG features for sample-wise classification of 5mC and 5hmC signals in the IDH2mut_2 sample. Combined 5mC (mc and mc_low) classifications occur more frequently in CpG islands (orange) than 5hmC. **(G)** Genomic annotation to CpG features of DhMR and DMR signal in the comparison of IDH2 mutant to NBM samples. **(H)** Genomic annotation to genic features, enhancers, and GENCODE lncRNA of the same integration in (G).



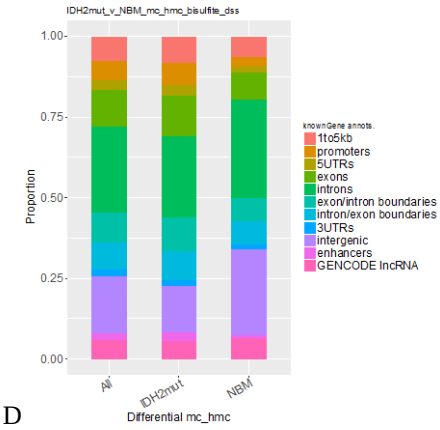
A



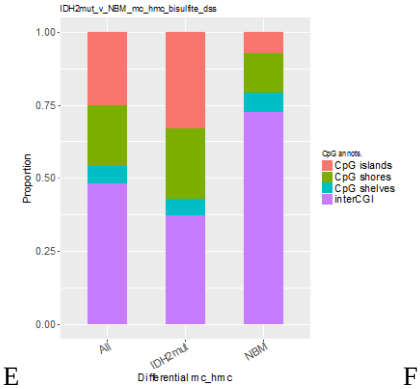
B



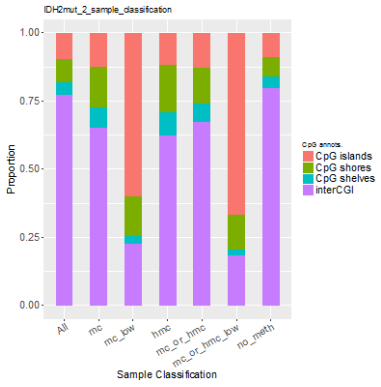
C



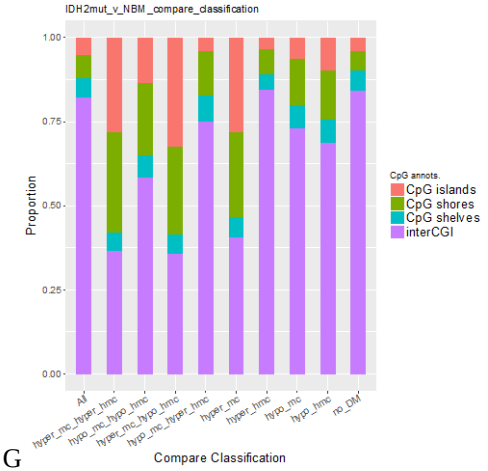
D


IDH2mut_v_NBM_mc_hmc_bisulfite_dss
Proportion
1.00
0.75
0.50
0.25
0.00
All
IDH2mut
NBM
Differential mc_hmc
CpG annots.
CpG islands
CpG shores
CpG shelves
interCGI

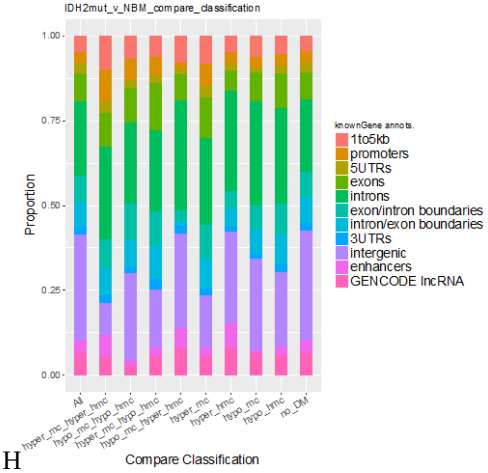
E


IDH2mut_2_sample_classification
Proportion
1.00
0.75
0.50
0.25
0.00
All
mc
mc_low
hmc
mc_or_hmc
mc_or_hmc_low
no_meth
Sample Classification
CpG annots.
CpG islands
CpG shores
CpG shelves
interCGI

F


IDH2mut_v_NBM_compare_classification
Proportion
1.00
0.75
0.50
0.25
0.00
All
hyper_mc_hyper_hmc
hypo_mc_hypo_hmc
hyper_mc_hypo_hmc
hypo_mc_hyper_hmc
hyper_mc
hyper_hmc
hypo_mc
hypo_hmc
no_DM
Compare Classification
CpG annots.
CpG islands
CpG shores
CpG shelves
interCGI

G


IDH2mut_v_NBM_compare_classification
Proportion
1.00
0.75
0.50
0.25
0.00
All
hyper_mc_hyper_hmc
hypo_mc_hypo_hmc
hyper_mc_hypo_hmc
hypo_mc_hyper_hmc
hyper_mc
hyper_hmc
hypo_mc
hypo_hmc
no_DM
Compare Classification
knownGene annots.
1to5kb
promoters
5UTRs
exons
introns
exon/intron boundaries
intron/exon boundaries
3UTRs
intergenic
enhancers
GENCODE lncRNA

H

Supplementary Figure 5 – A display of the entire UCSC Genome Browser track hub. All tracks from the track hub are displayed for the a genomic region containing MTA2 and EML3, which shows simultaneous hypo-hydroxymethylation and hyper-methylation in IDH2 mutants. Tracks have a default grouping based on the track type, but are easily rearranged by the user. For example, the track from the compare_classification module is grouped with the csaw and DSS tracks since the classification track is the intersection of the latter two.