AMC

Manual EMERGE

B.A. de Boer

4/15/2015

Contents

Prerequisites:

Helper files

In the folder <EMERGE-DIR>/Genome program a file in which the sizes of the chromosomes and a file of all annotated genes for each genome is stored. These files will be loaded at startup of the application. When you want to use other genomes than those provided with the application you will have to create your own files in the Genome subdirectory of the program:

- A tab delimited <your genome>.sizes file with chromosome names with corresponding lengths
- Optional: A corresponding tab delimited <your genome>.genes file with the chromosome name, start position, end position, gene symbol and strand of each gene.

```
<your genome>.sizes

chr1    197195432
chr2    181748087
………
```

```
<your genome>.genes

chr1    3204562     3661579     Xkr4    -
chr1    4797973     4832908     Lypla1  +
…….
```

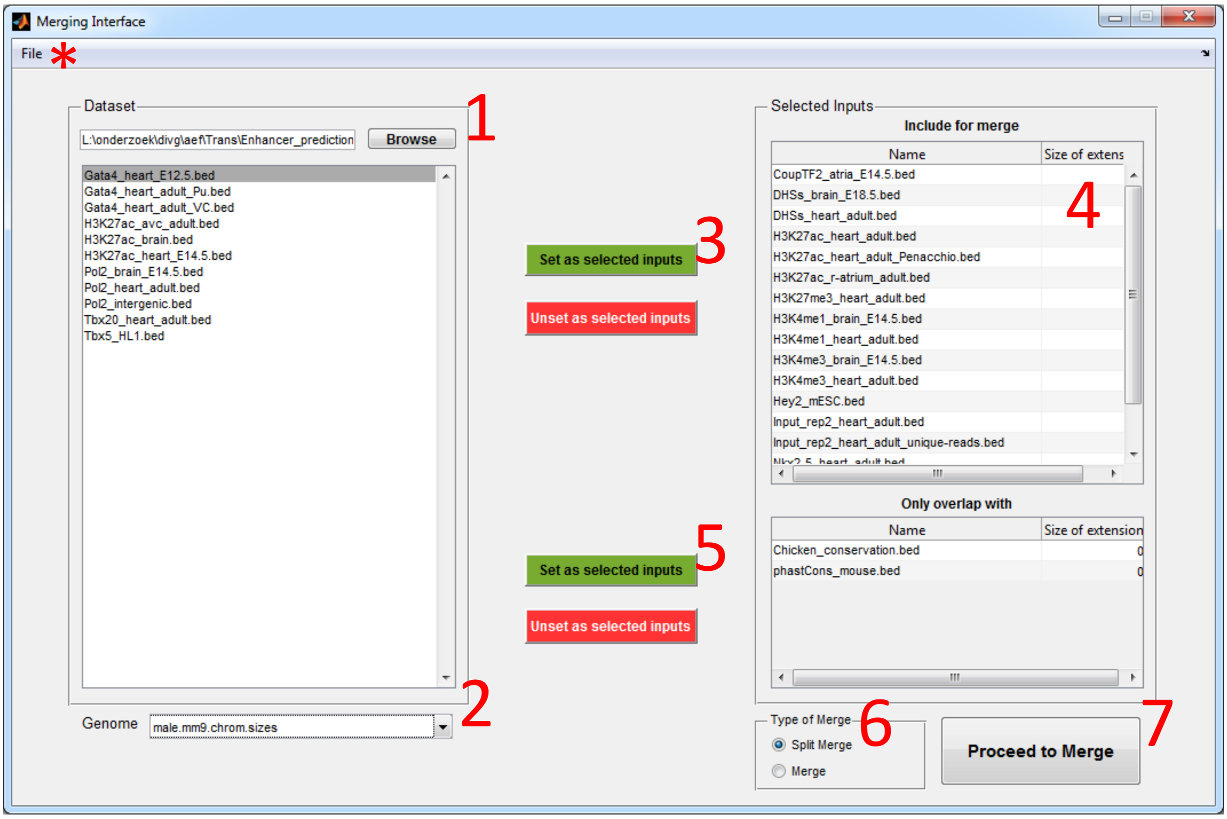
Input files

The program uses files in bed format. The first three columns of the bed file are obligatory whereas the fifth can be used to store a False Discovery Rate (FDR) or other significance value.

```
yourInput.bed

chr1    11972773    11973179    28    5.26    +    0    0    0,127,0
chr1    13364214    13364712    20    2.83    +    0    0    0,0,0
………
```

Merging files



Dataset

L:\onderzoek\divg\aef\Trans\Enhancer_prediction

Gata4_heart_E12.5.bed
Gata4_heart_adult_Pu.bed
Gata4_heart_adult_VC.bed
H3K27ac_avc_adult.bed
H3K27ac_brain.bed
H3K27ac_heart_E14.5.bed
Pol2_brain_E14.5.bed
Pol2_heart_adult.bed
Pol2_intergenic.bed
Tbx20_heart_adult.bed
Tbx5_HL1.bed

Genome: male.mm9.chrom.sizes

Selected Inputs

Include for merge

Name	Size of extens
CoupTF2_atria_E14.5.bed	
DHSs_brain_E18.5.bed	
DHSs_heart_adult.bed	
H3K27ac_heart_adult.bed	
H3K27ac_heart_adult_Penacchio.bed	
H3K27ac_r-atrium_adult.bed	
H3K27me3_heart_adult.bed	
H3K4me1_brain_E14.5.bed	
H3K4me1_heart_adult.bed	
H3K4me3_brain_E14.5.bed	
H3K4me3_heart_adult.bed	
Hey2_mESC.bed	
Input_rep2_heart_adult.bed	
Input_rep2_heart_adult_unique-reads.bed	
Nkx2.5_heart_adult.bed	

Only overlap with

Name	Size of extension
Chicken_conservation.bed	0
phastCons_mouse.bed	0

Type of Merge

● Split Merge

○ Merge

Figure 1 Merging interface

1. Select folder in which your bed files are stored.
2. Select the correct genome file (see Prerequisites how to add new genome files).
3. Select the files you want to merge.
4. Optionally, add an extension (will be added at 3' and 5' end) to the regions in your bed files. This extension is limited by the distance between the regions. To prevent regions to overlap the maximum applied extension is half the distance between two neighboring regions.
5. Optionally, add files you want to overlap with the merged datasets (Fig 2 panel c)
6. Select the type of merging (Fig 2 panel a and b)
7. Merge files

*. When you have already merged your files you can load them from disk

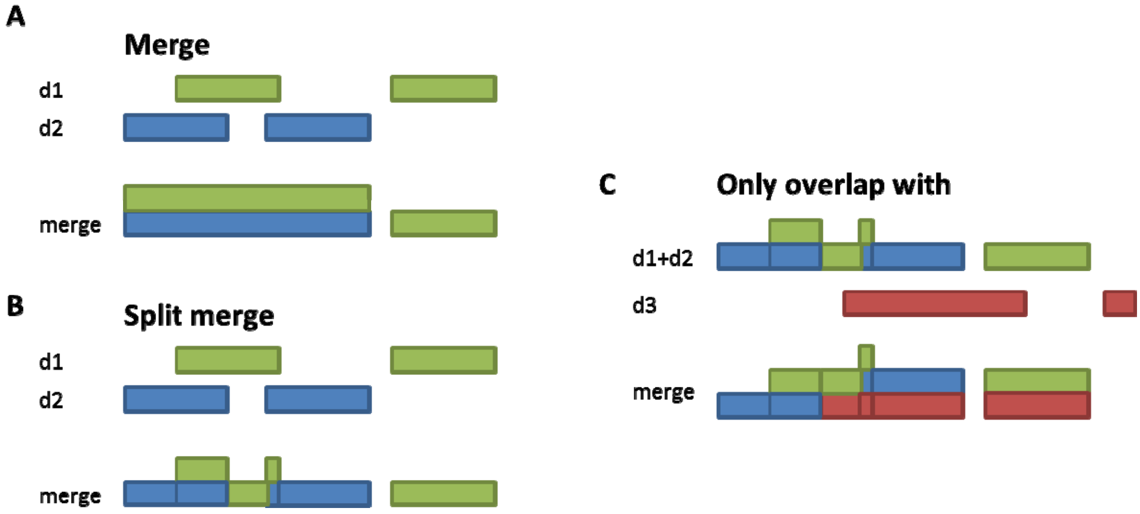
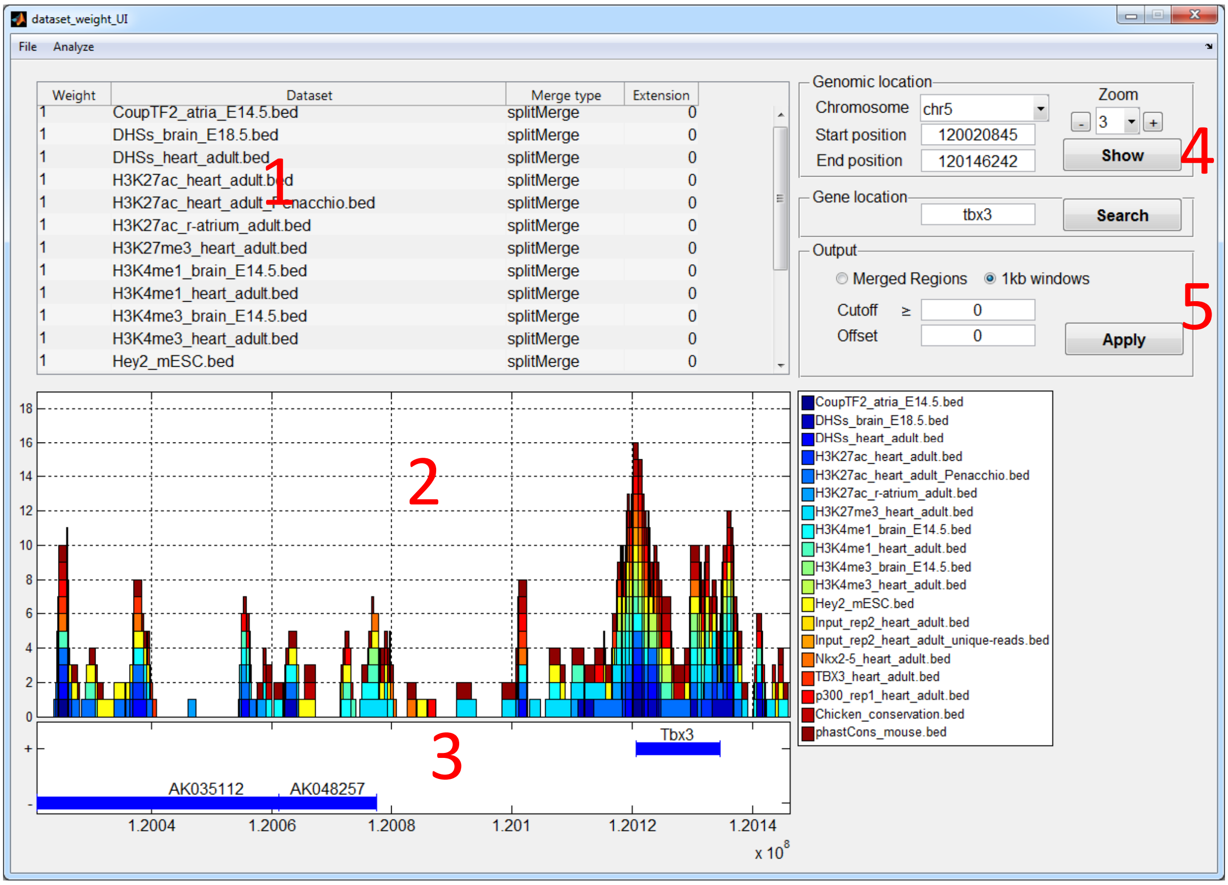


Figure 2 Types of merging

Dataset weight interface



Weight	Dataset	Merge type	Extension
1	CoupTF2_atria_E14.5.bed	splitMerge	0
1	DHSs_brain_E18.5.bed	splitMerge	0
1	DHSs_heart_adult.bed	splitMerge	0
1	H3K27ac_heart_adult.bed	splitMerge	0
1	H3K27ac_heart_adult_Penacchio.bed	splitMerge	0
1	H3K27ac_r-atrium_adult.bed	splitMerge	0
1	H3K27me3_heart_adult.bed	splitMerge	0
1	H3K4me1_brain_E14.5.bed	splitMerge	0
1	H3K4me1_heart_adult.bed	splitMerge	0
1	H3K4me3_brain_E14.5.bed	splitMerge	0
1	H3K4me3_heart_adult.bed	splitMerge	0
1	Hey2_mESC.bed	splitMerge	0

Genomic location
Chromosome: chr5
Start position: 120020845
End position: 120146242
Zoom: 3

Gene location: tbx3

Output
○ Merged Regions ● 1kb windows
Cutoff ≥ 0
Offset 0

Figure 3 Dataset Weight interface

1. Table in which the weight of each individual dataset can be set.

2. Overlap browser, when all weights are positive the output is shown as in Fig 3, when negative weights are included output is shown as in Fig 4. In the case that weights are automatically determined, the sum of weights can transformed to Odds ratios by taking the natural logarithm of the sum of weights (Fig 5).

3. Genome browser, shows all gene symbols. Coordinates are based on the maximum spanning region of all splice variants. Upper lane shows genes on the forward strand whereas the lower lane shows transcripts on the reverse strand.

4. Browse to exact position, zoom or search for a specific gene symbol.

5. A cutoff and an offset can be set. One can also determine the overlap of the different datasets in 1 kbp windows. This is done in overlapping windows in steps of 200 bp. The first time this option is selected, this has to be comptuted, which takes as while (in the order of 30 minutes). In the case that weights are automatically determined, the sum of weights can transformed to Odds ratios by taking the natural logarithm of the sum of weights (Fig 5). These settings are applied in the overlap browser as well as in the output in bedgraph format.

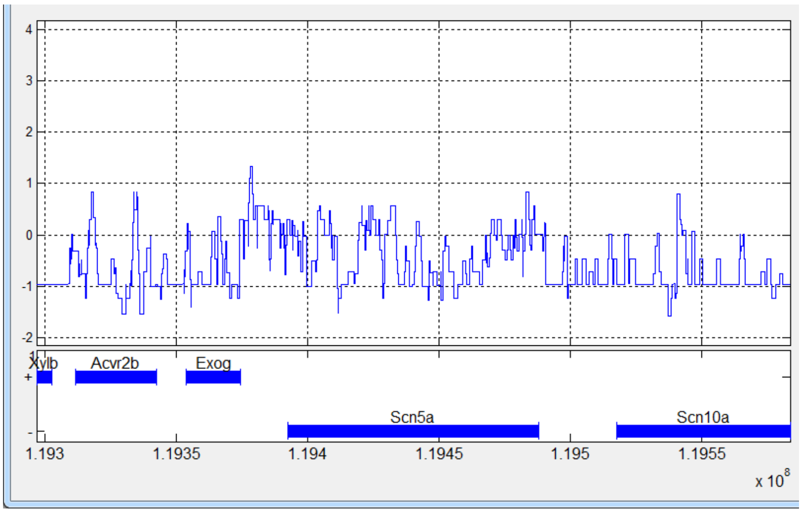
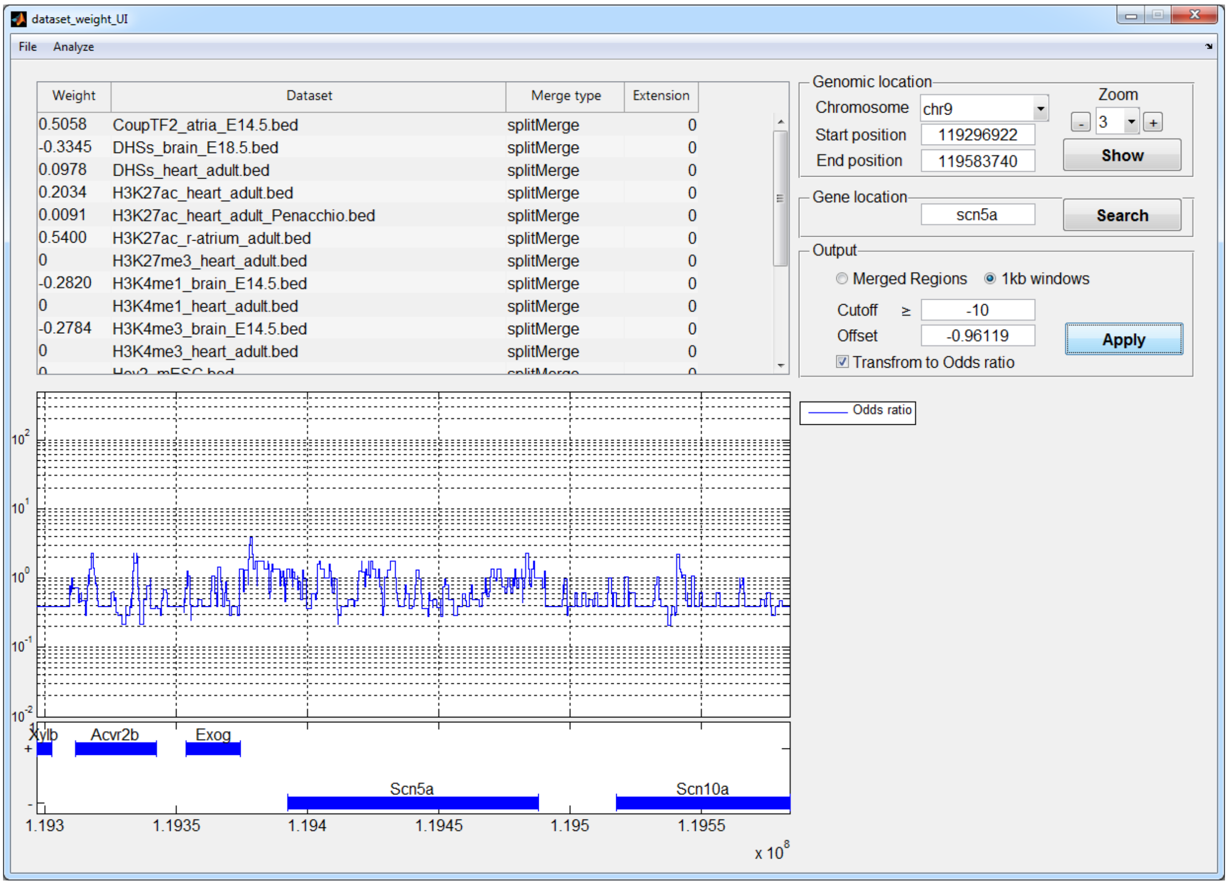


Figure 4 Overlap browser with negative weights



Weight	Dataset	Merge type	Extension
0.5058	CoupTF2_atria_E14.5.bed	splitMerge	0
-0.3345	DHSs_brain_E18.5.bed	splitMerge	0
0.0978	DHSs_heart_adult.bed	splitMerge	0
0.2034	H3K27ac_heart_adult.bed	splitMerge	0
0.0091	H3K27ac_heart_adult_Penacchio.bed	splitMerge	0
0.5400	H3K27ac_r-atrium_adult.bed	splitMerge	0
0	H3K27me3_heart_adult.bed	splitMerge	0
-0.2820	H3K4me1_brain_E14.5.bed	splitMerge	0
0	H3K4me1_heart_adult.bed	splitMerge	0
-0.2784	H3K4me3_brain_E14.5.bed	splitMerge	0
0	H3K4me3_heart_adult.bed	splitMerge	0
0	Hey2_mESC.bed	splitMerge	0

Genomic location

Chromosome chr9

Start position 119296922

End position 119583740

Zoom 3

Gene location scn5a

Output

○ Merged Regions ● 1kb windows

Cutoff ≥ -10

Offset -0.96119

☑ Transfrom to Odds ratio

Figure 5 Overlap browser with negative weights

Analyze menu

Determine optimal weights

Select the datasets you want to determine weights for and select bed files with True positive (TP) regions and True Negative (TN) regions. Weights will be determined based on using the complete TP and TP population. The ROC curve, which shows the predictive value of the model is based on 25 repetitions of splitting the TN and TP datasets into 75% for training and 25% for testing.

Determine ROC curve

Makes an ROC curve based on the weights of the selected datasets. One can make use of the an ordering based on significance levels (Excess ratio (ER), or FDR).

Save TN TP to Excel

Saves the occurrence of each merged dataset for all TN and TP regions to an excel sheet.

File menu

Load Mergefile

Loads a merged dataset, including the settings from the Dataset weight interface.

Export to bedgraph

Merge file is exported and can be used in external viewers such as the UCSC genome browser.

Save

Saves the current state of the program including the 1 kb bins and weights.