# DamID-Seq manual

Following manual contains brief description of DamID-seq data processing tools usage with examples of input and output data. These tools consist of three programs available as in Perl-scripts and as executable files (for windows-users). Algorithm of this pipeline is designed for DamID experiments with two biological replicates of both Dam-Protein (experiment) and Dam-only (control) samples.

## Input files

Input files containing reads alignment info should be in BED-format to be correctly processed by the first script. Such file should contain following columns: chromosome, start coordinate, end coordinate, read ID, score, strand and any other columns separated by space or tabular symbols. For example:

*chr2L 11920 11995 M02435:19:000000000-AB88P:1:1105:4378:12993 1 + 11920 11995 255,0,0*

*chr2L 12059 12134 M02435:19:000000000-AB88P:1:1105:4378:12993 1 - 12059 12134 0,0,255*

*chr2L 20185 20260 M02435:19:000000000-AB88P:1:1113:22228:4601 1 - 20185 20260 0,0,255*

*chr2L 22095 22170 M02435:19:000000000-AB88P:1:1103:6873:12106 1 + 22095 22170 255,0,0*

*chr2L 22279 22354 M02435:19:000000000-AB88P:1:1103:6873:12106 1 - 22279 22354 0,0,255*

*chr2L 22351 22426 M02435:19:000000000-AB88P:1:2116:26992:13487 1 + 22351 22426 255,0,0*

*chr2L 22433 22508 M02435:19:000000000-AB88P:1:2116:26992:13487 1 - 22433 22508 0,0,255*

*chr2L 23581 23640 M02435:19:000000000-AB88P:1:1102:14394:21491 1 + 23581 23640 255,0,0*

*chr2L 23681 23753 M02435:19:000000000-AB88P:1:1102:14394:21491 1 - 23681 23753 0,0,255*

*chr2L 24358 24432 M02435:19:000000000-AB88P:1:2116:13643:3145 1 + 24358 24432 255,0,0*

*chr2L 25026 25087 M02435:19:000000000-AB88P:1:2116:13643:3145 1 - 25026 25087 0,0,255*

Such file can be generated directly by some alignment programs or converted from BAM-format using any convenient tool. For example, **bedtools** package (<http://bedtools.readthedocs.io/en/latest/index.html>) contains **bamtobed** conversion utility which can be used for this purpose with default options:

*bedtools bamtobed -i input.bam >output.bed*

BED-files should be placed in the same folder with executable files (or Perl-scripts).

## Filtration of GATC-only reads

First script (**1\_GATC\_mapper**) should be launched four times (one for each BED-file). After launching it will ask to enter the input file name, for example:

*Dam-1.bed* - for first control sample

*Dam-2.bed* - for second control sample

*Protein-1.bed* - for first experiment sample

*Protein-2.bed* - for second experiment sample

After completion following files will be created in the same folder:

*“Dam-1\_reads\_per\_GATC\_filtered.txt”*

*“Dam-2\_reads\_per\_GATC\_filtered.txt”*

*“Protein-1\_reads\_per\_GATC\_filtered.txt”*

*“Protein-1\_reads\_per\_GATC\_filtered.txt”*

These files contain information about numbers of reads aligned on the edge of each GATC-fragment of the genome. For example:

chr2L 5302 6024 18

chr2L 6024 6880 6

chr2L 6880 6922 0

chr2L 6922 7693 0

chr2L 7693 7716 0

chr2L 7716 8552 13

chr2L 8552 8796 0

chr2L 8796 9694 1

chr2L 9694 9747 0

chr2L 9747 9802 0

chr2L 9802 12441 0

chr2L 12441 12693 54

chr2L 12693 12714 0

chr2L 12714 13066 39

Information about GATC-fragments coordinates is taken from file “All\_GATC\_list.txt” which was created for DM6 genome release. If DM3 release is preferable to DM6, file “All\_GATC\_list\_DM3.txt” should be renamed to “All\_GATC\_list.txt” with replacement of the previous one.

## Reproducibility filtering

Next script (**2\_dynSD**) filters out GATC fragments with too high difference in read-numbers between biological replicates. After launching it will ask to enter the first Dam-replicate filename (without “\_reads\_per\_GATC\_filtered.txt”):

*Dam-1*

Then the second Dam-replicate:

*Dam-2*

Then the first Dam-X replicate:

*Protein-1*

Then the second Dam-X replicate:

*Protein-2*

After completion following files will be created:

*“Protein-1+Protein-2\_vs\_Dam-1+Dam-2\_stat.txt”* - some statistics on Pearson correlation coefficients between biological replicates before and after filtering and numbers of passed filter and filtered out GATC-fragments.

*“Dam-1\_filtered.txt”*

*“Dam-2\_filtered.txt”*

*“Protein-1\_filtered.txt”*

*“Protein-2\_filtered.txt”* - four files similar to input files but with “NA” value of reads in filtered-out GATC-fragments. It should be mentioned that if one fragment is filtered out in one pair of biological replicates it would be also marked as “NA” in the second pair independently of its reproducibility there.

*“Dam-1+Dam-2\_filtered\_combined.txt”*

*“Protein-1+Protein-2\_filtered\_combined.txt”* - two files with combined over two replicates reads numbers for experiment and control samples. Filtered out GATC fragments marked as “NA”.

## Peak calling

The last script (**3\_Fisher\_auto-FDR**) uses one-sided Fisher's exact test to determine reliable GATC-fragments where Dam-X sample has more reads than Dam-only sample (“peaks”). After launching it will ask if you want to use your own P-value cutoff level or use automatically determined cutoff. This value (from 0 to 1, for example 0.0001) will determine the confidence level for Fisher's exact test which each GATC-fragment must reach to become a peak.

Automatic P-value is calculated as the biggest P-value at which the False Discovery Rate (FDR) reaches the desired cutoff value. This FDR-cutoff is entered in the next part of the script or the default value (0.05) is used.

Next the program will ask if you want to make minimal Dam-X over Dam excess level () as an additional requirement for GATC-fragment to become a peak. For example entering “2” will lead to peak calling in GATC-fragments reaching the minimal P-value and .

Next the name of overall experiment will be asked, for example:  
*MyProtein\_wild\_type*

Then the combined Dam-only data file name should be entered (without “*\_filtered\_combined.txt”)*:

*Dam-1+Dam-2*

Then the combined Dam-X data file:

*Protein-1+Protein-2*

Then the first Dam-X filtered repeat (without “*\_filtered.txt”)*:

*Protein-1*

Then the second Dam-X filtered repeat:

*Protein-2*

After completion following files will be created:

*“MyProtein\_wild\_type\_ROC.txt”* - this file would be generated if an automatic P-value option is used. It contains determined probability level and FDR-value at which this level was reached, for example:

target probability: 0.0116555663170057

or -log(10)= 1.93346662026649

with FDR: 0.0499766464269033

and full table with all examined probability levels and corresponding numbers of true- and false-positive peaks and FDR-values, for example:

probability True\_peaks False\_peaks FDR

0.0499792470100458 20357 2458 0.107736138505369

0.0464675847719192 20316 2384 0.105022026431718

0.0427161410544745 20252 2304 0.102145770526689

0.0368044047327221 20215 1915 0.0865341165838229

0.0292866963419953 19574 1670 0.0786104311805686

0.0162538699690401 19303 1151 0.0562726117140902

“MyProtein\_wild\_type\_at\_\*\*\*\_peaks.bed” - file with peaks coordinates (middle part of the name (\*\*\*) depends on input options). For example:

chr2L 1147020 1147278  
chr2L 2843534 2843825  
chr2L 3912276 3912457  
chr2L 3912572 3912996  
chr2L 3913342 3913556  
chr2L 3913556 3913851

“MyProtein\_wild\_type\_at\_FDR\_0.05\_probability\_full\_dataset.txt” - file with coordinates of all GATC-fragments of the genome followed by P-values (in  format) and Dam-X over Dam excess level (in  format), for example:

chrom start end Fisher\_probability log2(X/Dam)

chr3R 5702944 5704303 0.0791812460476249 0.446  
chr3R 5704303 5705701 0 MIN  
chr3R 5705701 5705861 -2.09848640124301 MIN  
chr3R 5705861 5705880 0 0  
chr3R 5705880 5706206 -7.11117736800484 MIN  
chr3R 5706206 5706524 NA NA  
chr3R 5706524 5706686 0.669006780958575 MAX  
chr3R 5706686 5706771 0 0  
chr3R 5706771 5707212 2.48919369894715 2.768  
chr3R 5707212 5707262 0 0  
chr3R 5707262 5707556 6.87297863222623 3.938  
chr3R 5707556 5708107 5.39080154682122 MAX  
chr3R 5708107 5708342 NA NA  
chr3R 5708342 5708516 2.19539641425107 4.147

Probability levels below 0 corresponds to cases when Dam-only sample has more reads in that GATC-fragment than Dam-X sample. In that case is printed.   
“MIN” and “MAX” values in the last columns correspond to cases when number of reads in experiment- or control-sample is equal to zero respectively.

“MyProtein\_wild\_type\_at\_FDR\_0.05\_probability\_track.wig” “MyProtein\_wild\_type\_at\_FDR\_0.05\_divided\_track.wig” - two track-files compatible with genome browsers like UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>) or Integrated Genome Browser (<http://bioviz.org/igb/index.html>).   
“Probability” track contains coordinates of all GATC-fragments having more than 0 reads in any of the samples and P-values for these fragments in the same format as in “full\_dataset” file.  
“Divided” track contains coordinates of all GATC-fragments having P-value above cutoff as for Dam-X>Dam and for Dam>Dam-X cases. The fourth column corresponds to value.