A few offline examples and step-by-step instructions on how to use EaSeq

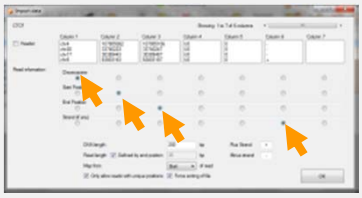
...More can be found in the integrated tutorials in EaSeq

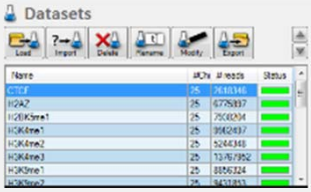
Table of contents

Guidelines on how to import data into EaSeq

1

In the ‘Dataset’ menu click the button ‘Import’ and select the files for import in the ‘file dialog’.

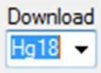




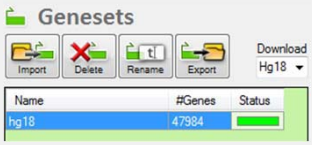
The ‘Import’ window will show the collumns that the dataset contains, and it will guess which of them that describe the chromosome, start, end, and strand. Various options such as the assumed size of the DNA fragments can also be set here.

The imported datasets will appear beneath the ‘Dataset’ menu. Green bars will show the progress of the file import. If the names are very long EaSeq will suggest to shorten them, since long names tend to result in messy figures.

2



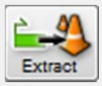
In the ‘Geneset’ menu select or type the reference genome to be downloaded from UCSC/Refseq.



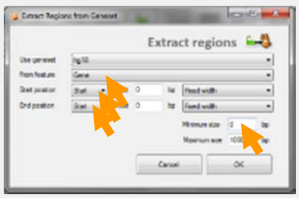
The ‘Import’ window will show the collumns that the geneset contains, and it will guess which of them that describe the chromosome, gene, CDS etc.

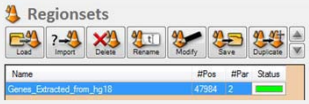
The imported genesets will appear beneath the ‘Geneset’ menu. Green bars will show the progress of the file import.

3



In the ‘Tools’ menu ´click the button ‘Extract’ to extract gene coordinates to a new set of regions.

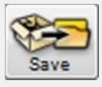




In the ‘Extract Regions from Geneset’ window select ‘Gene’ in the ‘From feature’ option as well as ‘Start’ and ‘Start’ for the start and end positions in order to extract coordinates corresponding to the Transcription start sites (TSS). Set minimum size to 0.

The imported regionsets will appear beneath the ‘Regionset’ menu. Green bars will show the progress of the file import. If the names are very long EaSeq will suggest to shorten them, since long names tend to result in messy figures.

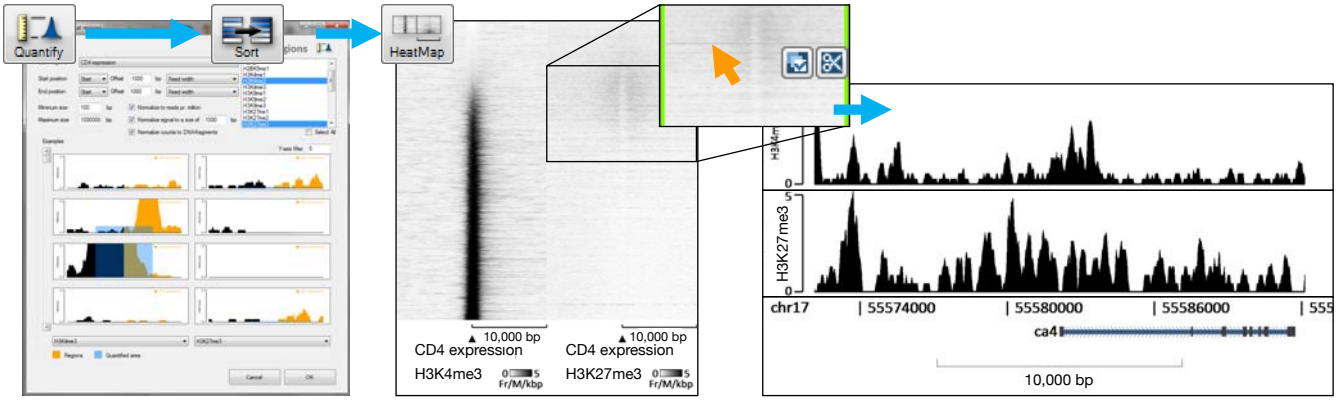
4



In the ‘Session’ Menu click the button ‘Save’ and specify location and name for the file in the file-dialog.

Workflow example 1:

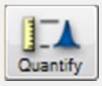
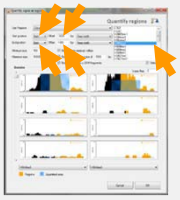
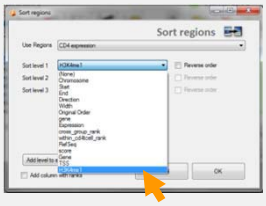
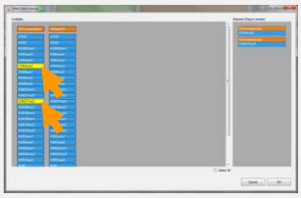
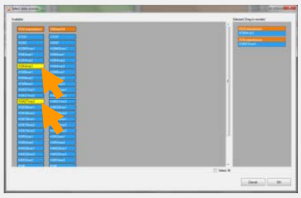
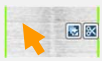
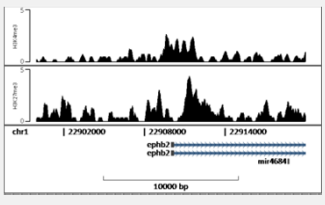
Quantitation, heatmaps, tracks, and interactive inspection



Making sorted heatmaps of genome-wide H3K4me3 and H3K27me3 profiles at TSS and use these to inspect tracks of genes individually.

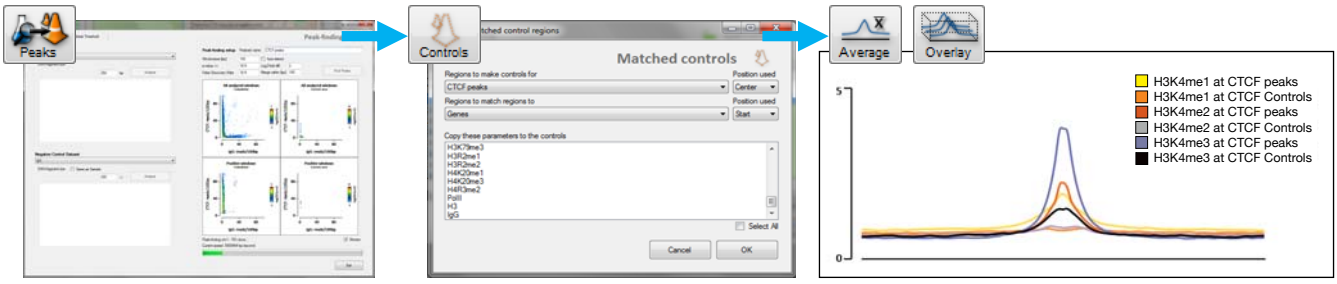
Here we use the ‘Quantify’ tool to control signal quantitation at a set of regions corresponding to 19,903 genes (left image), and we generate sorted heatmaps of H3K4me3 and H3K27me3 at the TSS of the 19,903 genes (middle image) and inspect individual loci by clicking on the heatmaps (right image).

Detailed instructions on how to carry out workflow example 1:

1		Select the regionset ‘CD4 expression’ and click the button ‘Quantify’.		In the ‘Quantify’ window: 1) Change start and end position to ‘Start’, 2) Set the Offset to -1000 and 1000 to quantify the signal from 1000 bp upstream to 1000 bp downstream of the start of the genes in ‘CD4 expression’. Select the ‘H3K4me3’ dataset for quantitation.
2		Select the regionset ‘CD4 expression’ and click the button ‘Sort’.		In the ‘Sort’ window select the ‘H3K4me3’ parameter for sorting.
3		Click the button ‘HeatMap’.		In the ‘Select data sources’ window select the ‘H3K4me3’ and ‘H3K27me3’ datasets in the column corresponding to the ‘CD4 Expression’ regionset.
4		Select the heat-maps and click the button ‘Settings’.		In the ‘Change settings’ window change the ‘Position’ setting to ‘Start’ for the heat-maps to depict the start of the regions in the regionset.
5		Select the heat-maps and click the button ‘Redraw’.		The heat-maps will now be updated showing the newly changed order.
6		Click the button ‘FillTrack’.		In the ‘Select data sources’ window select the ‘H3K4me3’ and ‘H3K27me3’ datasets in the column corresponding to the ‘CD4 Expression’ regionset.
7		Click a position in the heat-maps.		The newly generated tracks and gene annotation will autoupdate to show the selected position.

Workflow example 2:

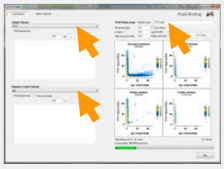
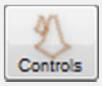
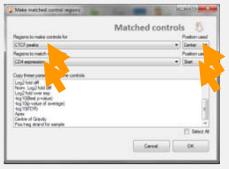
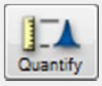
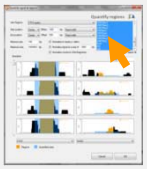
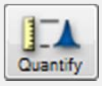
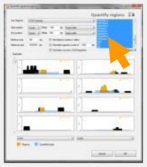
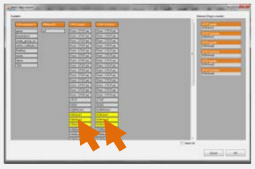
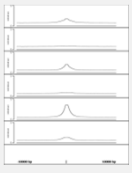
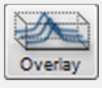
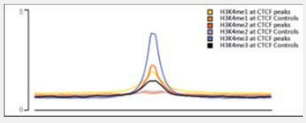
Peak-finding, controls, and average signal at peaks



Identifying CTCF peaks, generating a set of random control regions, and visualizing the average H3K4me1/me2/me3 levels at both sets of regions.

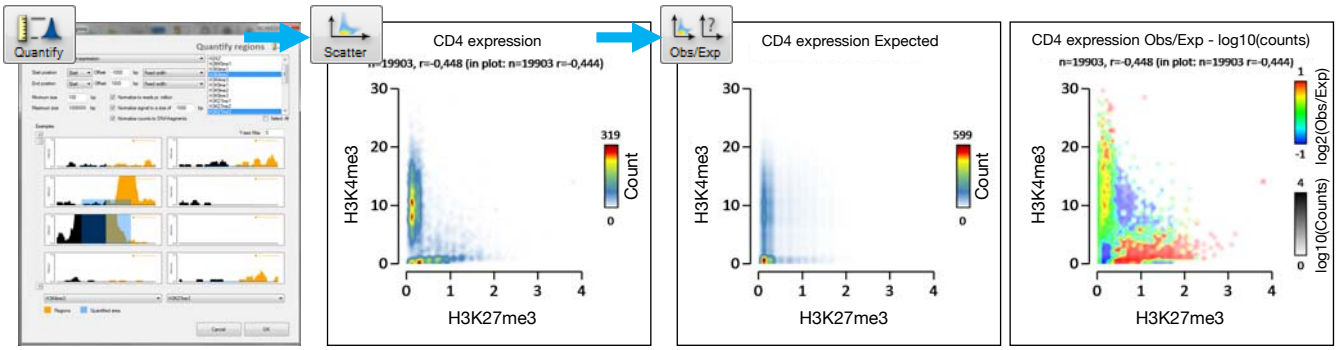
Here we find a set of CTCF peaks using the ‘Peaks’-tool (left image), generate a set of negative control regions that population-wise match the CTCF peaks on their relative orientation and distance to nearest TSS using the ‘Controls’ tool (middle image), and make six overlaid tracks showing average H3K4me1, me2, and me3 levels at both sets of regions using the ‘Average’ and ‘Overlay’ tools (right image).

Detailed instructions on how to carry out workflow example 2:

1		Click the button ‘Peaks’.		In the ‘Find peaks’ window: Select the ‘CTCF’ dataset as sample and ‘IgG’ as control. Name the peakset: ‘CTCF peaks’, and start the peak-finding. You can exit the window and follow the progress via the green bar of the new Regionset.
2		Once, the peak-finding has completed, click the button ‘Controls’.		In the ‘Make matched control regions’ window: Select the center of the ‘CTCF peaks’ regionset as the positions to make controls for and the start of ‘CD4 expression‘ (=TSS) for matching. Next, rename the regionset ”CTCF controls”.
3		Select the regionset ‘CTCF peaks’ and click the button ‘Quantify’.		In the ‘Quantify’ window: Select the ‘H3K4me1’, ‘H3K4me2’, and ‘H3K4me3’ datasets for quantitation.
4		Select the regionset ‘CTCF controls’ and click the button ‘Quantify’.		In the ‘Quantify’ window: Select the ‘H3K4me1’, ‘H3K4me2’, and ‘H3K4me3’ datasets for quantitation.
5		Click the button ‘Box plot’.		In the ‘Select data sources’ window select the ‘H3K4me1’, ‘H3K4me2’, and ‘H3K4me3’ parameters in the two columns corresponding to ‘CTCF peaks’ and ‘CTCF controls’ regionsets.
6		Select the Box plot, and either click the button ‘Settings‘ or click the ‘Settings’ icon (triangle) in the box plot.		In the ‘Change settings’ window change the ‘Maximum’ setting for the Y-Axis to 8 and deselect the ‘Auto’ and ‘Log scale’ options. Once the Settings window is closed, the Box plot will autoupdate to show the plot with the new settings.
7		Click the button ‘Average’ to get new average tracks.	 	In the ‘Select data sources’ window select the ‘H3K4me1’, ‘H3K4me2’, and ‘H3K4me3’ parameters in the two columns corresponding to ‘CTCF peaks’ and ‘CTCF controls’ regionsets. This will result in 7 plots – 3 showing the average normalized signal at the CTCF peaks, 3 at the CTCF controls, and 1 plot showing center and distances.
8		Select the 6 plots of the averages and click the button ‘Overlay’.		This will generate one new plot containing the values in all of the 6 plots overlayed in different colors.

Workflow example 3:

2D-histograms and plots of Observed / Expected values

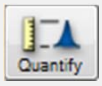


Making a scatter plot depicting the genome-wide levels of H3K4me3 and H3K27me3 at TSS and investigating if the levels of the- se two histone marks were correlated.

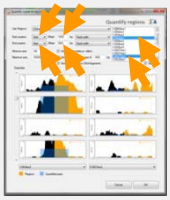
Here we quantify genome-wide H3K4me3 and H3K27me3 levels at 19,903 TSS positions using the ‘Quantify’ tool (leftmost image), visualize the relationship between the levels of the two marks in a 2D- histogram using the ‘Scatter’ tool (middle left image), and relate it to a 2D-histogram containing the densities that would be expected if there was no relationship between the levels of the two marks using the ‘Obs/Exp’ tool (middle right image). The result is shown in the rightmost image, which is pseudocolored 2D-histogram generatedby the ‘Obs/Exp’ tool and illustrate the relationship between the H3K4me3 and H3K27me3 signals relative to what would be expected by chance. 19,903 genes were depicted and positions on the X- and Y-axes were determined from the H3K27me3 and H3K4me3 levels segmented into 50 bins. Staining intensity was proportional to the density of regions in each bin, and pseudocoloring was determined from the log2 ratio of the number of observed (Obs) regions divided by the number of expected (Exp) regions in each bin. The plot suggests that regions with high levels of both marks occurred less frequently than expected by chance (blue coloring), and that regions that were positive for one mark, but for not the other mark, were more frequent than expected by chance (red/yellow coloring).

Detailed instructions on how to carry out workflow example 3:

1

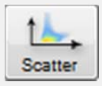


Select the regionset ‘CD4 expression’ and click the button ‘Quantify’

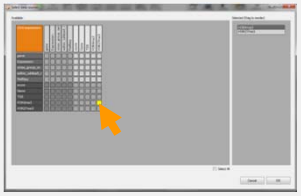


In the ‘Quantify’ window: 1) Change start and end position to ‘Start’, 2) Set ‘Offset’ to -1000 and 1000 to quantify the 2000 bp surrounding the start coordinates in ‘CD4 expression’. Select the ‘H3K4me3’ and ‘ H3K27me3’ datasets for quantitation.

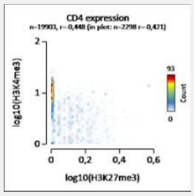
2



Click the button ‘Scatter’.



In the ‘Select data sources’ window select the combination of the ‘H3K4me3’ and ‘H3K27me3’ parameters.

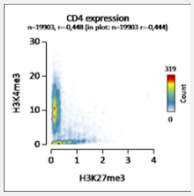


This will generate a new XY-scatter plot with each dot corresponding to a genomic region. The number of regions having each combination of H3K4me3 and H3K27me3 is color-coded, so that frequent combinations are yellow or red and infrequent are blue.

3

Select the scatter plot and click the button ‘Settings’ or click the ‘Settings’ icon (triangle) in the box plot.

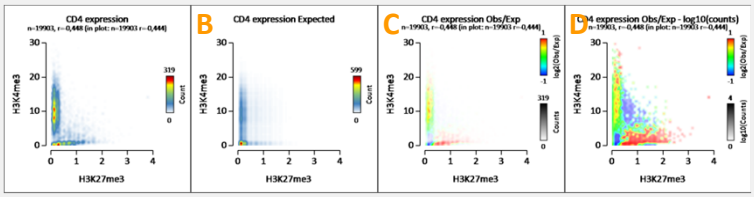
The plot show the log-values on the axes by default. To change to liniear scales, deselect the ‘Log scale’ options for the X- and Y-axes in the ‘Change settings’ window.



The XY-scatter plot will now show the linear values on the axes, and the two most prominent subpopulations (red areas) corresponds to K4me3-positive / K27me3-negative, as well as K4me3-negative / K27me3-negative transcription start sites.

4

Select the XY-scatter, and click the button ‘Obs/Exp’



This will generate 3 XY-scatter plots (marked by B, C, and D). B shows the expected number of regions having the various combinations between K4me3 and K27me3 if the two parameters were independent of each other. C and D both show the observed regions pseudocolored for the ratio between the observed and the expected regions with each combination (red: more, blue: less than expected). In C the density of the coloring is proportional to the observed number of regions with a particular combination, where the density in D is proportional to the log10 number of regions with each combination.