User-friendly and interactive analysis of ChIP-seq data using EaSeq

Mads Lerdrup[1,2] and Klaus Hansen[2]

Affiliations:
1) Center for Chromosome Stability
Department of Cellular and Molecular Medicine
Faculty of Health and Medical Sciences
University of Copenhagen
Denmark

2) The Hansen Group
Biotech Research and Innovation Centre (BRIC)
Faculty of Health and Medical Sciences
University of Copenhagen
Denmark
e-mail: mads.lerdrup@sund.ku.dk

Abstract

ChIP-seq is a central method to gain understanding of the regulatory networks in the genome of stem cells and during differentiation. Exploration and analysis of such genome-wide data often leads to unexpected discoveries and new hypotheses. It therefore accelerates and improves the discovery phase, when scientists with biological understanding are enabled to analyse and visualize data. EaSeq (http://easeq.net) offers integrated exploration of genome-wide data in a visual, versatile, user-friendly, and interactive manner that connects abstract interpretations to the signal distribution at the underlying loci. Here we introduce the interface, data types and acquisition, and guide the reader through two example workflows. These workflows will enable the reader to perform genome-wide analysis and visualization of transcription factor binding sites and histone marks. This includes making basic plots; finding, annotating, sorting and filtering of peaks; using EaSeq as a genome browser; measuring ChIP-seq signal and calculating ratios; as well as data import and export.

Key words

ChIP-sequencing, Next generation sequencing, analysis, visualization, exploration, genomics, epigenetics.

1. Introduction

Chromatin Immunoprecipitation followed by DNA-sequencing (ChIP-seq) has been a workhorse for the identification of the transcriptional networks that drive cellular growth and differentiation and in gaining knowledge related to how histone marks are associated with the different chromatin states that maintain cellular identity[1-5]. Since the initial development of the method more than a decade ago[6-8], thousands of publications have used the method, and it is hard to imagine how the current detailed knowledge regarding transcription factor binding and histone modifications in the genome in different cellular stages could have been uncovered without the method.

ChIP-seq analysis is however often considered data-heavy and computationally intensive. The processing and analysis of data is therefore often done by computational experts, which are in high demand. Thus, this process tends to be time-consuming and test the patience of stakeholders. With genome-wide experiments such as those based on ChIP-seq, the exploration of the data also frequently leads to unexpected discoveries and the generation of new hypotheses. Data are therefore often reanalysed several times, while the biological understanding grows. In many cases, the disconnection between the biological understanding and the ability to analyse and visualize data impedes this process. Therefore, user-friendly software, which allow the biologists to perform quick and interactive genome-wide exploration and analysis, can accelerate and improve the discovery phase. We developed EaSeq[9] (http://easeq.net) to enable a more integrated exploration of genome-wide data in a visual, versatile, and user-friendly manner, which allows scientists with biological understanding to get more direct contact with the data and analyses, and to test hypotheses more rapidly. A secondary aim was to tightly connect abstract interpretations in e.g. clustered heatmaps or scatter plots to the signal distribution at the underlying loci, thus the interactivity in EaSeq enables users to check whether a point in a scatter plot actually represent the assumed state at the locus. EaSeq is aimed at facilitating and accelerating the explorative and often unpredictable part of the analysis work, whereas the preceding data processing can often be standardized and pipelined efficiently with existing tools such as Galaxy[10]. Given the low predictability and the frequent variation in how data are handled during this stage, EaSeq is designed to provide a “sandbox” with wide degrees of freedom and open workflows. Due to the diversity of data and the versatility of the uses, this chapter present advices on how to use the different functionalities on example data and there is no unified procedure on how to use EaSeq. As the program contains more than 50 plot and tool types and an exceeding number of possible actions, it is also limited what can be covered in this chapter beyond the most basic operations. Yet as demonstrated in 3.2.9 step “pp”, this is sufficient to make figures similar to those in high impact publications.

2. Materials

2.1 Installation and computer requirements

EaSeq was developed to be used on a windows PC. A zip-file containing the program can be downloaded free of charge from http://easeq.net. To use the program, simply unzip the files, and move the file easeq.exe into a suitable folder (avoid network drives as EaSeq will need to have write permissions to its home directory, and this might be disabled for network drives due to security reasons). EaSeq was designed to have no dependencies on any external code. The only requirement is Microsoft .NET, which is a standard component of Windows installations. Recent versions of EaSeq requires 64-bit windows, which are more or less standard on PCs today. It will run on most hardware, but efficient analysis of more than a handful of datasets would benefit from more than 4 GB RAM and a multicore CPU. EaSeq is designed for a monitor with a screen resolution of 1920x1080 or 1920x1200, and while all modern monitors will provide this resolution or better, some laptops might not fulfil this requirement and will benefit from an external monitor.

For PCs using MacOS or Linux, a "virtual machine" containing a Windows desktop will need to be installed first. Commercially available virtual machines for MacOS includes Parallels and VMWare Fusion. Once this has been installed together with Windows, the EaSeq installation is similar to that of a Windows PC. Another solution is to have a low-latency virtual Windows desktop running on a remove server. Company and university IT-departments might offer this option for its employees, and several commercial solutions exist for running such a solution on a reasonably priced pay-per-use basis.

2.2 Interface and basic concepts

Once EaSeq has been installed and started, the main window will be opened, and this is subdivided into a number of panels (Fig. 1). In the following, the different parts of the main user-interface will be referred to in [brackets], and thus [Tools/Quantify] implies that the button named "Quantify" in the "Tools" section within the main window should be clicked.

In the upper part of the interface, EaSeq offers tools to work on three types of imported data as well as so-called sessions, which are single-file assemblies of all the imported data, plots, and descriptions of the data handling that has been done. Below these sections, the interface also contains two toolbars for plots and analysis tools. The lower half of the main window contain a large grey area that can be filled with plots of all types with genome-wide relationships. Once the user has generated the intended plots to understand the data, they can all be exported as high quality images, which are ready for use in reports. In the right side, the interface there are buttons to open and overlay four different specialized panels. Finally, the procedures that involves a tool or a plot will often lead to the opening of a new window, which is usually closed again once the procedure is initiated.

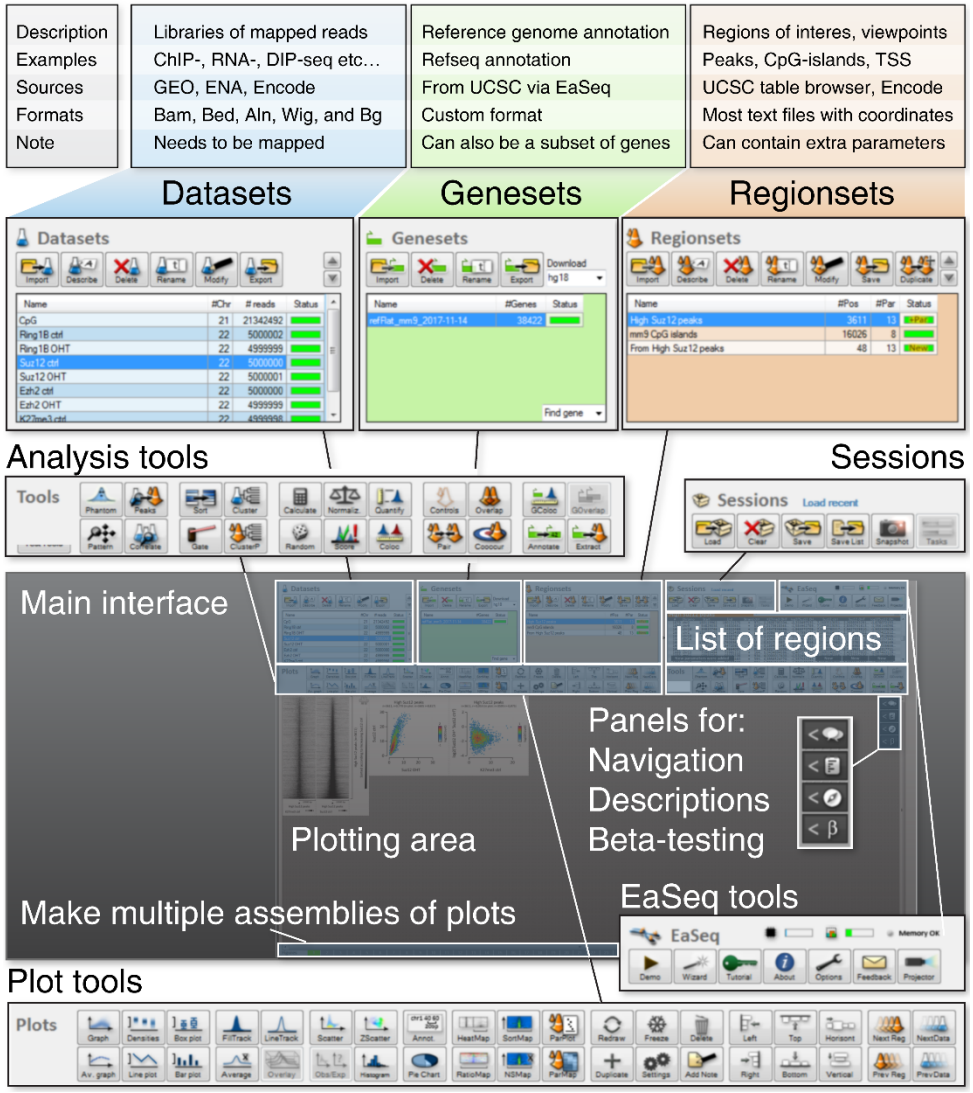

Figure 1: Overview of the main interface in EaSeq and enlargements of the central panels and areas.

2.2.1 Data types in EaSeq

The three types of data that EaSeq uses are (Fig. 1): 1) 'Datasets', which are the imported ChIP-seq or other types of next generation sequencing libraries, such as RNA-seq. Before import, the sequenced libraries need to be mapped (see 2.2.3. and 2.3.1). 2) 'Regionsets', which as the name implies are sets of regions in the genome that would be interesting to study (e.g. enhancers, CpG-islands, published peak-sets etc.). Regionsets could also originate from an imported analysis of RNA-seq or another type of genome-wide expression data (See Note 1). In many cases, Regionsets are used as the viewpoints for plots, but they can also be handled or analysed in relation to other Regionsets. 3) The final type of data, 'Genesets', contains the annotation of features for all genes in a reference genome (or in some cases subsets of such annotations). These data can be downloaded directly from UCSC[11,12] within EaSeq.

To illustrate this, imagine a user, who have generated ChIP-seq samples from mouse cells grown under different conditions and mapped these to the mm10 reference genome, and would like to study differences in these ChIP-seq signals at a published set of transcription factor binding sites and transcriptional start sites (TSSes). To do so, the user would need to import the libraries mapped to mm10 as datasets (Covered in 3.1.2) and the mm10 coordinates for the transcription factor binding sites as a Regionset (Covered in 3.2.1 step b). Either the mm10 coordinates for the TSSes can then be imported as a Regionset too, or the Refseq[13] annotation for mm10 genes can be downloaded from UCSC[12,11] and stored as a Geneset within EaSeq (Covered in 3.1.4 steps m-n). The coordinates for the TSSes within this Geneset then subsequently needs to be converted to a Regionset before further use (Covered in 3.1.4 step r). EaSeq also offers a range of other ways to convert one type of data to another (See Note 2).

2.2.2 The differences between Datasets and Regionsets

It is common for new users to be uncertain about the exact distinction between Datasets and Regionsets. As mentioned above, libraries of mapped reads should be imported as Datasets, whereas a previously generated set of peaks from the exact same data should be imported as a Regionset. In many plot types, Regionsets would often serve as the viewpoints for genome-wide presentations of e.g. a set of transcription factor binding sets, whereas as the Datasets would serve as the depicted signal (Fig. 2). Datasets are typically millions of lines long, whereas files containing such peaks would be a few thousand lines long, and EaSeq will try to warn the user, if it detects a potential mix of the data types based on file sizes. Regionsets could also be lists of established genomic features, and such data can vary a lot in their formatting. Therefore, EaSeq includes a wizard that helps to import atypical data. Besides genomic coordinates, Regionsets can also contain additional properties that can be visualized or e.g. used for subselection. In EaSeq these properties are called Regionset Parameters.

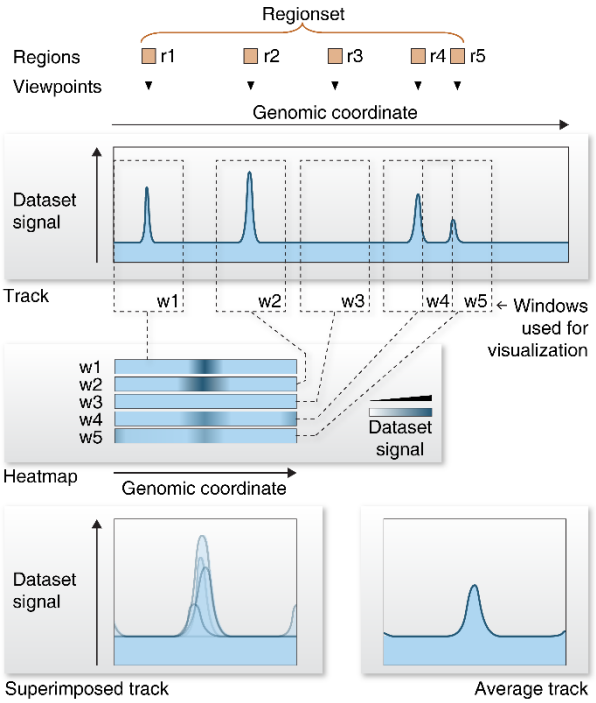

Figure 2: Upper part: Panels illustrating the relationship between Regionsets, viewpoints, visualized windows, and dataset signal. Lower part: examples of how the five Regions will be displayed by three different plot types.

2.2.3 Datasets: unmapped reads vs. mapped reads vs coverage vectors

Next generation sequencing of ChIP-seq samples (or other types of –seq e.g. DIP-seq, RIP-seq, RNA-seq, ATAC-seq, or CAGE-seq) leads to a high number of individual sequences, which range from 75 to 150 bp with the currently common sequencing technology. These sequences are commonly called reads, and before they can be used for most visualization and analyses, each read will need to be matched to a reference genome (currently named mm10 and hg38 for the mouse and human genome, respectively) in a process called aligning or mapping. Not all reads will be successfully matched, but those that are will contain a set of coordinates (chromosome, start, end, and strand) matching to the corresponding sequence in the reference genome, and these are often called mapped reads. Before data from a sequencing run can be imported into EaSeq, they will need to be mapped (See 2.3.1). During import, only the genomic coordinate is preserved, whereas other information e.g. sequence or quality metrics is discarded to save memory and storage.

Sometimes data have been processed further to give a coverage vector, which is a long list of values that reflect the aggregated number of reads along different positions within the chromosomes. Therefore, the information about each individual read is lost, once data from a library is converted to coverage vectors, and we recommend importing and using files containing individually mapped reads as they provide higher processing speed, better resolution and a more efficient normalization than coverage

vectors do. Files with coverage vectors are however needed e.g. when uploading custom tracks to UCSC (REF) or when depositing data on GEO (REF), and they can be generated in EaSeq as described in 3.1.6 step w.

2.2.4 Formats

Datasets containing individual reads can be imported from bed-, bam-, and aln-files, or most types of text files with columns containing the genomic coordinates (Fig. 1). Coverage-based Datasets can be imported from wig-files and bedgraph/bg-files. Often data that would be imported as a Regionset are also stored in the .bed format, and the use of the same format for mapped reads imported as Datasets and e.g. peaks imported as Regionsets, is of course an understandable source for confusion. Regionsets can however also be imported from a rich variety of file formats, and it is possible to import almost any type of text files where the information is separated into fields by tabs, spaces, semicolons etc. as long as the file contains information regarding the genomic coordinates (Covered in 3.2.1 step b). The format requirements for Genesets are a bit more stringent, as they will need to contain information for several gene features. If needed the best way to learn about the formatting requirements, is to download and export a Geneset within EaSeq and mimic the formatting in a custom-made file. It is important that all imported data adheres to the same conventions on chromosome names (See Note 3).

2.3 Getting data

2.3.1 Acquiring and processing libraries for import as Datasets

After sequencing, the typical processing procedure for the data includes the following steps (Fig. 3): 1) Basecalling, which is the process of obtaining a DNA sequence from the more raw data in the sequencer. This is often an integrated part of the processing done automatically by the sequencing machine itself or staff in the sequencing facility. This results in unmapped reads in the fastq format, which also contain information on the sequencing quality for each base. It is highly advisable to store fastq-files for submission into repositories such as GEO[14] during manuscript preparation. 2) Assessment of the overall sequencing quality using FastQC[15] and possible contaminants using FastQ Screen[16]. 3) Trimming of adaptor sequences or sequence cycles with low sequencing quality, e.g. using Trimmomatic, Trimgalore or cutadapt[17-19]. 4) Mapping (a.k.a alignment) to a reference genome. Typical input for this process is one or more fastq-files, which have been trimmed, and the output can be in many formats, most commonly bam- or sam-files. It is important to assure that the coordinates from all data analysed together are from the same reference genome (See Note 4). 5) For import into EaSeq, the files need to be converted to either bam-files or bed-files (the latter is often gzip compressed and can be stripped of information that is no longer needed and is therefore more compact). A useful tool for such conversions is samtools[20]. Some processing pipelines include one or more filtering steps to remove low quality reads and overly amplified PCR products, which are undesirable to have in the data. The latter part is also done per default when importing data into EaSeq, and this step is sometimes called deduplication. The procedures described above in steps 1 to 5 above often also needs to be applied to published data that have been deposited at repositories such as GEO[14] or ENA (https://www.ebi.ac.uk/) in an unmapped format such as fastq-files. These procedures typically require a bit of preceding experience with tools that are developed for Linux/UNIX, so it will often be helpful to get assistance from the sequencing or bioinformatics core facility on setting up the pipelines needed for this. Users, who are less comfortable with the command-line based UNIX environments, will often find that the popular and user-friendly GUI-based software Galaxy is an attractive alternative [10].

2.3.2 Where to get data?

As mentioned above, GEO[14] (https://www.ncbi.nlm.nih.gov/geo) and ENA

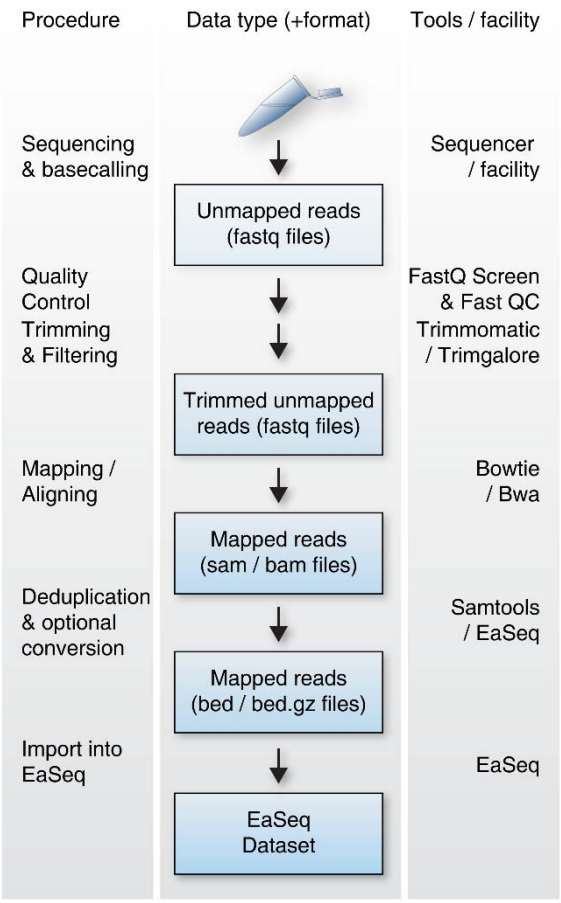

Figure 3: Overview of the acquisition and processing of libraries prior to import into EaSeq.

(https://www.ebi.ac.uk/) are excellent sources for libraries of unmapped reads, and most publications contain a GEO accession number that can guide readers to the location of the data that were used for the study. The GEO depositions also often contain coverage vectors mapped to a particular reference genome, and while wig and bedgraph/bg files can be imported directly in EaSeq, bigwig and bigbed files will first need to be converted by an external tool such as bigWigToWig (https://www.encodeproject.org/software/bigwigtowig/) before import. Consortia such as Roadmap Epigenetics[21] (http://www.roadmapepigenomics.org/) and Encode[22] (https://www.encodeproject.org/) offers resources with attractive collations of both mapped and unmapped libraries, as well as peaks, which have been identified from the data as well as other genomic regions that can be imported as Regionsets. A popular and comprehensive resource is also the Table Browser from UCSC[12] (https://genome.ucsc.edu/cgi-bin/hgTables), which can provide coordinates for a wide range of interesting genomic structures in a variety of reference genomes.

3. Methods

In the following sections, we will cover two example workflows, which are composed to illustrate the analyses and visualization of genome-wide transcription factors binding sites (3.1) and histone marks (3.2). The data used in the examples are parts of central and published stem cell studies[23,24], and we have generated zip-files containing mapped ChIP-seq libraries and other relevant data for these examples to help the reader getting started on the example workflows in a quick and easy manner. The first example (3.1) will cover peak-finding, annotation of peaks, saving data, descriptions, making a few plots, and using EaSeq as a genome browser. The second example (3.2) will cover data import, heatmaps, inspection of peaks, quantitation of ChIP-seq signal, ratio calculations, peak sorting, filtering, and a couple of additional plot types. The two example workflows can be handled independently of each other, whereas the individual steps within each workflow often requires actions that are done in the preceding steps. Therefore, it is recommendable to perform the steps in each workflow in a sequential manner.

3.1 Example 1: Peak-finding and -annotation from transcription factor ChIP-seq data

The aim of this example is to provide the reader with some basic experience on how to handle ChIP-seq data for transcription factors, including data import and export, peak-finding, annotation of peaks, and some basic visualization. A zip-file containing the data needed for this example workflow can be downloaded at http://easeq.net/wf1.zip. When the zip-file is decompressed, four .bed-files become available. These files contain mapped reads corresponding to three transcription factors (CTCF, Oct4, and Nanog) fused to GFP and enriched using an anti-GFP antibody as well as a control sample with GFP alone. These data were originally a part of the data presented in a pioneering ChIP-seq paper[24], and the sequencing reads were mapped to mm9 by us and down-sampled to reduce the file size and import time.

3.1.1 Optional disabling of hints

Whereas pop-up hints might be a useful way to learn the functionality of new software, they might be too much of a distraction during a guided procedure as this workflow. It is therefore advisable to disable the hints that shows up by default when running a new EaSeq installation.

a. This can be done by clicking on the white wrench icon in the popups and unchecking the “Show hints” box.
b. Hints can be enabled later by opening the “Options” panel and click on the tab called “Panels”.

3.1.2 Import of ChIP-seq data

c. To import the mapped data, click on [Datasets/Import] and select all four .bed-files: “CTCF.bed”, “Oct4.bed”, “Nanog.bed”, and “GFP.bed”.
d. When the import wizard shows up, it is able to recognize the files, so the default settings will suffice for these files. Click “OK” in the import form to start the process.
e. Keep an eye on the list in the “Datasets” panel as the imported datasets will appear here. Once the small progress indicator becomes green, then the Datasets are available for subsequent work.

3.1.3 Peak-finding

Peak-finding is the process of systematic identification of areas in the genome where the enrichment of a sample is strong, and this is almost always a requirement to e.g. identify transcription factor binding sites or areas with high levels of a histone modification. This procedure results in a set of “Peaks”, which essentially are a set of genomic regions that can be used as viewpoints in plots or for a diverse range of other tasks. In EaSeq, this is done by dividing the genome into “windows” of default 100 bp, and then identify those with significantly higher levels of signal. This process is repeated four times to avoid false negatives where peaks are divided between two windows. In EaSeq, this procedure requires two Datasets, a sample and a negative control, where the latter is used to model the unspecific background distribution in the library. Peak-finding typically takes 5-10 minutes and will result in a new Regionset, which can be used as viewpoints for subsequent genome-wide analysis.

f. To start the peak-finding, click on the [Tools/Peaks] icon, and in the window that appears.
g. Select “Oct4” as the “Sample Dataset” and “GFP” as the “Negative Control Dataset”.
h. *Optional*: Shorten the name of the resulting peakset, since long names tend to reduce the overview.

Three settings adjust the stringency of the peak finding, “p-value”, “False Discovery Rate”, and “Log2 fold diff.”. If one would like to detect areas enriched in broadly distributed ChIP-seq signals, such as the histone marks H3K27ac or H3K27me3, then it is recommendable to increase window sizes and merge distances from the default to larger sizes, such as 500 bp, 2000 bp, or more (Fig. 4).

i. For this example, use the default settings and start the peak-finding by clicking “Find Peaks”.

After a while, the plots in the window will start to monitor the currently analysed (right side) and cumulatively analysed (left side) parts of the genome. The plots show the level of positive signal and negative control signal in all windows (top) and the signals that were found to be enriched in the initial parts of the analysis (bottom) and considered positive. By studying the distribution of signal within these populations, it is possible to assess if the positive windows are rich in the signal from the negative control or not. If this is the case, it might be a reason to increase stringency of the peak-calling.

j. It is not required to keep the window open during the procedure, so you are welcome to close it by clicking “Exit”.

k. *Optional* tasks: Repeat the peak-finding for CTCF and Nanog as well.

l. *Optional* task: Monitor the progress of the peak-calling by clicking on “Tasks” in the main menu below “Sessions”.

Peak-finding will lead to the generation of a new Regionset, which contains all the coordinates and quality scores of the peaks. This can already be seen during the peak-calling, and it will appear in the list of Regionsets in the Regionset panel. Once the peak-finding completes, the Regionset will be marked as “New” in the Regionset list, and you can inspect the coordinates as well as additional parameters, such as false discovery rates (FDR) values and read counts of each peak by selecting the Regionset and observe and browse the list that appears in the panel to the right side of the Regionsets.

3.1.4 Annotation of peaks

Once a set of peaks is generated, an often-sought procedure is to learn the identity of neighbouring genes. To annotate the peaks (or any other Regionset) with gene names, accession numbers and distances, EaSeq will require a file with the coordinates of the genes in the used reference genome. EaSeq offers a quick and simple integration to acquire the Refseq[13] annotations directly from UCSC[11,12]. Please remember to cite these sources when using the data they provide.

m. To download gene annotations for mm9, select “mm9” in the dropdown box below the Genesets panel.

n. In the import wizard, click “ok” to the default settings, and EaSeq will import all the currently annotated mouse genes in the Refseq database and the mm9 coordinates of their features, and store this as a Geneset.

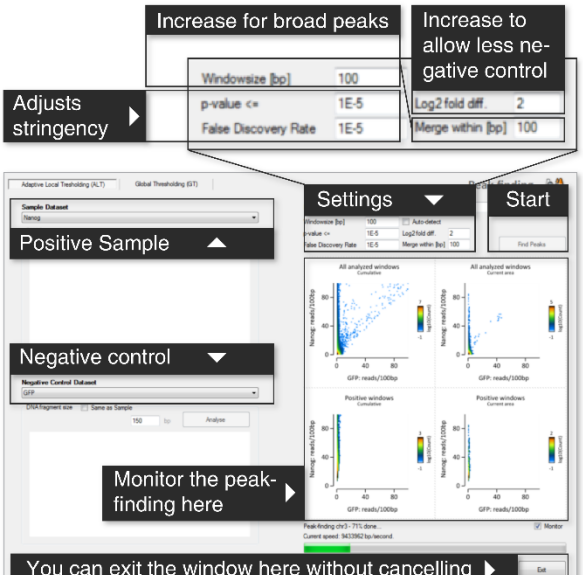

Figure 4: Example of the peak-finding window and annotation of its contents.

o. To run the actual annotation of the peaks, click [Tools/Annotate].

p. In the window that shows up, choose the “Start” of the genes as the position to use for the annotation. This can be altered, and a typical variation would be to select “Starts+Ends”, if one would like the nearest gene to be annotated to each peak regardless of which part of the gene that is closest.

q. The annotation will add a number of new parameters to the existing Regionset (in this case the Oct4 peaks), including gene names, accession numbers, and distances. Inspect these more closely in the viewer in the right side of the Regionset list.

r. *Optional*: A commonly used procedure is to visualize ChIP-seq signal at gene features such as TSS or the entire gene bodies. The methodology to do this will be covered in steps described in 3.2, but such a visualization will require that a Regionset corresponding to the gene feature of interest e.g. TSSes is generated. Now, with the mm9 annotation file downloaded and imported above in steps n-o above, it is simple to make such a Regionset from the mm9 Geneset. Just click on [Tools/Extract], change the end position to “start”, set the start offset to -500 and the end offset to 500, and click ok. This will result in a set of regions corresponding to the +/- 500 bp surrounding all the annotated TSSes. If kept unchanged, the default settings will result in a Regionset where each region correspond to the entire range from the transcription start (TSS) to the end sites of each annotated transcript.

3.1.5 Saving sessions

So far, the procedure has required a considerable amount of handling. To limit the risk of data loss and to accelerate later access to data and analyses, then save a session containing all the data and work. Such sessions are generally also more space efficient and loaded more quickly than the input data.

s. Save a session by clicking [Sessions/Save].
t. *Optional*: After the session is saved, try to close and reopen EaSeq, and then reimport the data in the session by clicking [Session/Load].

3.1.6. Exporting Regionsets and Datasets

Often it might be relevant to export a peak-set or another Regionset. In this case, do this for the newly generated and annotated Oct4 peaks as follows:

u. Select the Regionset containing the peaks and thereafter [Regionsets/Export]. This outputs a tab-separated text file containing several columns of peak coordinates, quality scores, and now also the identity of the gene with its TSS closest to each peak.
v. *Optional*: import and inspect these data in Excel or R[25].

An often-used procedure is to inspect and share data in the UCSC genome browser. For a better visualization, it is advisable to generate a coverage vector (See above) that can be uploaded to the UCSC genome browser[11].

w. In this case, export the Nanog ChIP-seq library (Dataset) as a "wig-file", by clicking [Datasets/Export] and use the default settings.
x. *Optional*: Upload the compressed Nanog.wig.gz file as a custom track in the UCSC genome browser.

3.1.7 Auto-generated descriptions

When exporting a coverage vector from the Nanog Dataset, another file named "Nanog.wig.description.txt" is stored in the same folder as the wig-file. This is an auto-generated description of the wig-file, including all handling that was done in EaSeq.

y. Try to open "Nanog.wig.description.txt" in the folder with the wig-file using a text editor.

These files can be exported along with all types of data exported from EaSeq, and they will be co-imported if the exported set of data is reimported later (provided that the description file and the data are kept in their original state). The auto-generated descriptions are saved as a part of session files and can also be accessed and studied from within EaSeq.

z. To inspect the auto-generated descriptions within EaSeq, open the "Description panel" by clicking on the notepad icon in the right side of the screen.

3.1.8 Making tracks

A key functionality in EaSeq is the abundance of visualization tools to study the imported data in an interactive manner. For a start, this example workflow will demonstrate how to make a set of tracks showing the one-dimensional distribution of ChIP-seq signal along the genome.

aa. To make one or more tracks, click [Plots/Filltrack].
bb. In the window that appears, select the desired combination(s) of Regionset(s) (horizontal) and Datasets(s) (vertical).

The default viewing style of the FillTracks is to show the cumulative selected ChIP-seq signal (Dataset) at the centre of all regions in the Regionset, but this can be changed to show single or subsets of loci (See later). If multiple combinations of Datasets and Regionset were selected in step bb, multiple plots will appear, and during this step their order can be changed upfront by dragging them in the panel in right side. All plots in EaSeq have a number of icons, which appears when the mouse in moved over the plot (Illustrated in figure 4c in [9]). The most important of these is the "Plot Settings", which gives the user many options to customize each plot (See 3.2.8 for instructions on how to do this).

cc. An important step is to export the plots made in EaSeq. A bitmap image (tif, jpeg etc.) can be exported by clicking on [Sessions/Snapshot]. To export a pdf-file, click on the icon in the right side of the user interface containing a beta sign to open the panel with recently added tools. In this panel click on "Export snapshot as pdf".

3.1.9 Using EaSeq to browse the genome

Once an EaSeq session contains Dataset(s), Regionset(s), and one or more track(s), EaSeq can be used as a genome browser similar to the UCSC Genome Browser[11], but often quicker as all computation is handled locally.

dd. To start navigating the genome in EaSeq, try to type the symbol of a gene in the dropbox below Genesets. This will cause the track(s) to change viewpoints to the selected gene.
ee. To alter the view, e.g. by zooming out, then open "Navigation panel" by clicking on the compass icon in the right side of the screen.
ff. Once the Navigation panel is open, several controls including a red knob appears. The red knob is a quick way to zoom in (drag up), zoom out (drag down), pan left (drag left), pan right (drag right), or any combination hereof.

A practical way to inspect the ChIP-seq signal at a set of peaks (see above), is to click on one of the peaks appearing in the list of regions on the right side of the Regionsets. This procedure optionally combined with the use of heatmaps (described in 3.2.2) can generally be applied to iteratively improving the quality of identified peak-sets (Fig. 5).

gg. Try to select the Regionset containing the Oct4 peaks.
hh. Click on one of the peaks
ii. Then zoom out by dragging the red knob down in the Navigation panel.

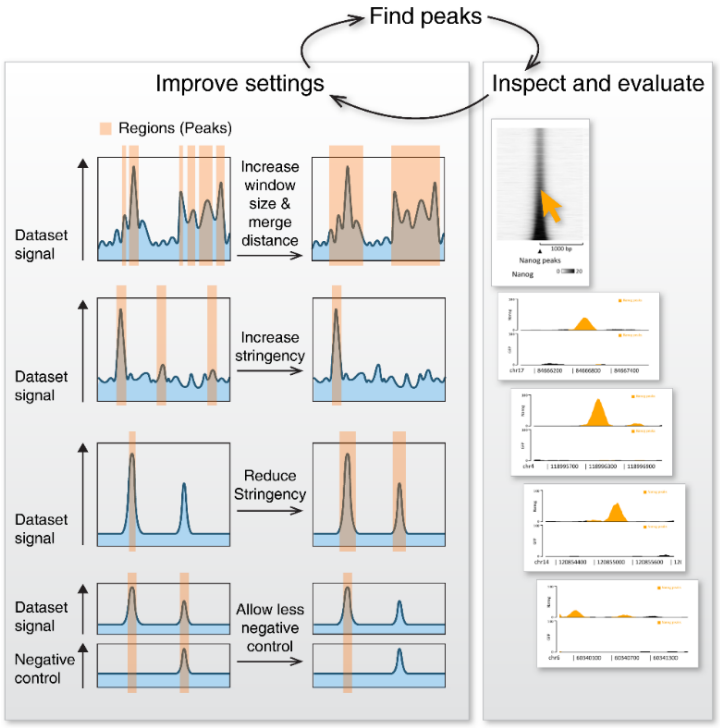

Figure 5: Upper part: Overview of the process of iteratively improving peak-sets through rapid inspection of the peaks.Lower part, some guidance on how to improve peak-calling in cases of suboptimal output peaks.

3.2 Example 2: Quantifying, calculating ratios, and visualizing histone mark ChIP-seq data

The aim of this workflow example, is to provide basic experience on how to handle ChIP-seq data for histone marks, including how to make plots such as heatmaps, how to inspect the signal at a predefined set of peaks (such as that generated in 3.1.3), how to quantify signal and calculate ratios, as well as how to sort and filter peaks based on this. A zip file containing the data needed for this example workflow can be downloaded at http://easeq.net/wf2.zip. When the zip-file is decompressed, a text-file with coordinates for CpG-islands and an EaSeq session file becomes available. The session file contains several already imported ChIP-seq datasets of a histone mark (H3K27me3) and chromatin modifiers from the group of Polycomb proteins (Ring1B, Suz12, and Ezh2). These are enriched under two different conditions, an inducible knock-out of the Ring1B polycomb protein (OHT) and control samples. These data were originally a part of central work in the polycomb field making use of ChIP-seq[23], and the sequencing reads were mapped to mm9 by us and down-sampled to reduce the file size and import time. To save time, the session file also contains a Regionset of Suz12 peaks that were identified previously (See figure 8 in [9]) based on an independent set of data in a procedure that was roughly as described in (3.1.3). These peaks will be used as viewpoints for the ChIP-seq signal and as our primary regions of interests for this example.

3.2.1 Import of ChIP-seq data and regions

a. Unpack the zip-file and import the session "Blackledge compact.eas" by clicking [Sessions/Load].
b. *Optional*: Data formats tend to vary a lot, but the import Wizard in EaSeq, offers a lot of flexibility that allows import of diverse formats. To get some experience with import of an interesting but more challenging set of viewpoints from an external source, then import "mm9 CpG islands.gz" from the zipped data as a Regionset by clicking [Regionset/Import]. The default view of the file browser is to look for text-files, so it will need to be instructed to look for all files or .gz-files. These CpG-islands were originally downloaded from the UCSC table browser[12] with contains a rich collection of relevant genomic loci for a large set of reference genomes. In this case, the import will lead EaSeq to ask whether it should skip the first line or not. This occurs because the first line of the file contains a hashtag (#), which often also is used for comments with content that should be ignored by software. In this case, it is instead used in the name of header, but EaSeq does not know this, and needs instructions on how to handle it. Therefore, instruct EaSeq that it should not skip the line. If the import goes well, then the wizard should identify that chromosome names, start coordinates, and end coordinates are found in columns 2, 3 and 4, respectively, and will suggest that the data in columns 2-4 should not be imported as parameters. In this case, just use the default suggestion and click OK. If EaSeq did not recognize the information structure properly, then it will need instructions from the user regarding in which columns the different coordinates can be found.

3.2.2 Make and use heatmaps for genome browsing and inspection of peaks

A quick way to get an overview of the signal in the imported datasets, relative to the Suz12 peaks, is to make heatmaps, which also can be used to explore the signal at the peaks in more depth.

c. To make one or more heatmaps, click on [Plots/Heatmaps].
d. This will open a window where it is possible to select the desired combination(s) of Regionset(s) (horizontal) and Datasets(s) (vertical).
e. To use the heatmaps as a starting point to explore ChIP-seq signal, then also make a set of tracks, by clicking [Plots/Filltrack].
f. Again, in the window that appears, select the desired combination(s) of Regionset(s) (horizontal) and Datasets(s) (vertical).
g. Click on a single point in the heatmap. The tracks will now update to show that particular locus. This is a quick and good strategy to assess if peak-

finding settings were appropriate or it should be repeated with more or less stringency (Fig. 5).

h. *Optional*: It is possible to combine this with manual navigation within the genome by opening the "Navigation panel" (click on the compass icon in the right side of the screen), and e.g. drag the red knob to zoom in (drag up), zoom out (drag down), pan left (drag left), pan right (drag right). In this way, the surroundings of many peaks can readily be inspected at any magnification to assess how strong the signal is at the peak relative to interesting genomic features or the general background.

i. To select multiple loci in the heatmap, and get the signal visualized in the tracks, press down the mouse button inside a heatmap, and keep it pressed down while dragging it over a subset of the peaks. Once it is released, the tracks will update to show the superimposed signal at all of the selected peaks. This superimposed track state is a useful way to get an overview of the signal distribution in the selected subpopulation of the peaks.

j. *Optional*: Select all peaks in the Regionset by clicking the small icon next to the selected area.

k. This is also a good occasion to consider the information provided in the heatmap. Heatmaps shows the signal from all loci within a Regionset, with the loci stacked vertically (Fig. 2). The horizontal center in the heatmap is by default also the center of each visualized locus, and the default view then shows the 10 kbp upstream loci to the left and the 10kbp downstream loci to the right. Both the visualized window as well as the vertical order of the heatmaps can be adjusted (see 3.2.5 and Note 5 for instructions on how to change this).

3.2.3 Quantify ChIP-seq at a set of peaks or other regions

In the following, we will quantify the signal of all imported ChIP-seq datasets at the midpoint and surrounding 1kbp of all Suz12 peaks.

l. To quantify ChIP-seq signal, click [Tools/Quantify].

This will open a new window (Fig. 6), which also offers an intuitive way to assess the signal distribution at peaks, and to evaluate if the applied settings will capture the signal appropriately. The eight tracks in the centre of the window (Fig. 6i) are visualizations of the ChIP-seq signal from the four first regions (vertical) in the Regionset that will be used for the quantification. The two columns of tracks shows two example Datasets, and changing those in the dropdown menu below the track examples (Fig. 6ii) has no effect on the final outcome. The datasets that are used for the actual quantitation can be selected in the list in the upper right corner (Fig. 6iii), and by default all datasets are selected. The Regionset to be used can be changed in the uppermost dropdown menu (Fig. 6iv), and the slider in the left side (Fig. 6v) will allow a quick inspection of a large number of regions from the selected Regionset. The window also contains several options to instruct specifically which area relative to each region in the Regionset that should be used as the basis for the quantitation (Fig. 6vi). Finally, the Quantify tool provides a range of options for how to normalize the quantified values (Fig. 6vii), and it will by default provide FPKM values (See Note 6). When most of these options are changed the visualization in the tracks (Fig. 6i) will change accordingly. The orange area (Fig. 6viii) show each region (and possibly neighbouring regions as well), whereas the blue transparent area (Fig. 6ix) gives an indication of the extent of each area that is used for the quantitation as well as how the normalization will affect the amplitude of the quantified signal.

m. *Optional*. Explore the functions and settings in the Quantify tool by changing the selection in the dropboxes or by changing the offset values – what do you see?

n. Once you have explored sufficiently, use the default settings for the quantitation, with start as start position, end as end position, and set both offsets to 0 (If many settings were changed in step m above, then it might be simpler to close and reopen the window). This will lead to the signal within the boundary of the peaks to be quantified, which might be an advantage if the peaks vary a lot in size. The default normalization will take the size of the peaks into account. If the start, end and offset settings are left as default, then the signal will be quantified within a window starting 500 bp upstream and ending 500 bp downstream of the peak centres, which in many

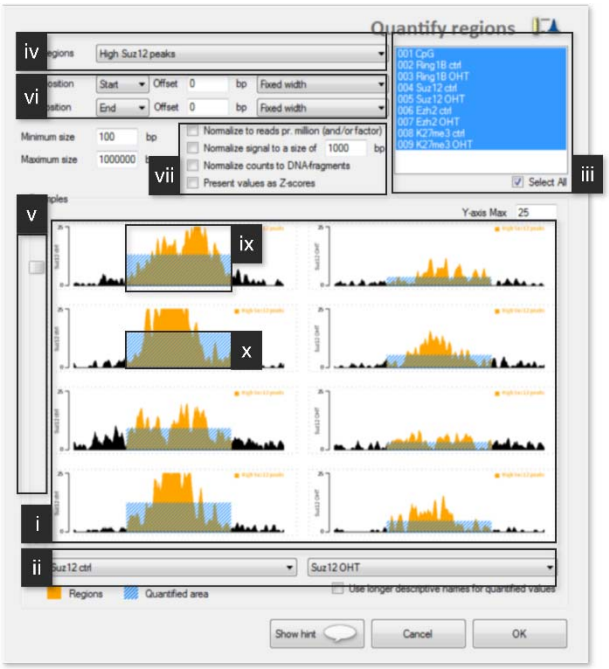

Figure 6: Example of the Quantify window and annotation of its contents.

cases also might be a useful option as the size of the quantified regions is kept constant.

The quantified values will be added to the Regionset as new "Parameters" (see 2.1.2), and the list of Regionsets will display that this Regionset has been updated.

o. *Optional*: Inspect these values by clicking on the updated Regionset in the table showing all the Regionsets, and then in the list of regions in the right side, new columns of data will appear (one for each Dataset that was included in the quantification).

3.2.4 Calculate ratios of quantified ChIP-seq signal at peaks

The next steps will demonstrate how to calculate the ratio between Suz12 in the conditional knock-out (OHT) and the control samples. Although this is a useful way to assess overall changes at a set of regions and to find those with the most pronounced changes, it is advisable do this with caution (See Note 7).

p. To calculate the log2 fold difference between the quantified Suz12 OTH and Suz12 control values, first open the Calculate tool by clicking [Tools/Calculate].

q. Then in the appearing window, select the "Division" radiobutton, check the "Logarithm" flag, from the dropdown boxes select the newly quantified values in the Suz12 OHT parameter as "x" and Suz12 ctrl parameter as "y", and click "OK".

As for the quantified values in 3.2.3, the ratios will now appear as a new "Parameter" (see 2.1.2), and the list of Regionsets will display that the Regionset has been updated.

3.2.5 Sort peaks according to quantified ChIP-seq ratios

To change the order of appearance of regions in certain plot types, such as heatmaps, they will need to be sorted first. To do this:

r. Open the Sort tool by clicking on [Tools/Sort].

s. In the appearing window, select the newly calculated ratios.

Any heatmap that shows the sorted Regionset will now automatically update to show the new and updated order, in this case the Suz12OTH/ctrl ratio.

3.2.6 Making 1D- and 2D-histograms

A useful way to get an overview of a population of quantified values is to make a histogram of these. If the relationship between two sets of quantified values or e.g. imported Parameters in a Regionset is in question, then a 2D-histogram is ideal. 2D-histograms are rather similar to scatter plots, but they use colour to show the density of regions with certain sets of values. They segment the plotted area into "bins" for this counting, similarly to what a 1D-histogram does, and the binning can be coarse or fine depending on data density, resulting in very different appearances.

t. To make a (1D) histogram of the quantified Suz12 OHT values, click [Plots/Histogram].

u. In the window that appears, select the desired combination(s) of Regionset(s) (horizontal) and Parameter(s) (vertical). If multiple combinations of Regionsets and Regionset Parameters are clicked, multiple plots will appear, and their order can be changed upfront by dragging them in the right side panel.

v. To make a 2D-histogram of the quantified Suz12 OHT and Suz12 ctrl values, select the Regionset containing the values that should be visualized and click [Plots/Scatter].

w. In the window that appears, select the desired combination(s) of Regionset Parameter(s) (both horizontal and vertical). If multiple combinations are clicked, multiple plots will appear, and their order can be changed upfront by dragging them in the right side panel.

3.2.7 Interactively exploring data in 2D-histograms

Once the new plots showing the Suz12 OHT values and the relationship between Suz12 OHT and Suz12 Ctrl levels have appeared. These plots are interactive and allow you to select one or more regions by clicking on the area in the plot. Once a region is selected it will be highlighted in other plots displaying the same Regionset, and tracks will automatically update to show the locus surrounding the selected regions(s).

x. To select a single region in the 2D-histogram, click on the dot in the plot.

y. To select a subset of multiple regions from the Regionset displayed by the 2D-histogram, press down the mouse button inside the 2D-histogram, and keep it pressed down while dragging it over the regions that you would like to select.

Plots can be moved, resized, and duplicated to build up an exportable (see 3.1.8 step cc) panel, which supports the scientific message.

z. To duplicate a plot, click on the small icon with a plus appearing inside the plot when the mouse pointer moves over it.

aa. To move a plot, press the appearing small icon with four arrows in the upper left corner of the plot and hold down the mouse button while moving the plot.

bb. To resize a plot, press the appearing small icon with the square and arrow in the lower right corner of the plot and hold down the mouse button while resizing the plot.

cc. *Optional*: If multiple plots are selected, then they can be moved and resized in the same way. To duplicate all plots, click on the [Plots/Duplicate] icon in the main menu. Duplicated plots will appear in a larger unoccupied area.

dd. *Optional*: Make some tracks by clicking on [Plots/Tracks], and select one or more regions in the 2D-histogram to show the signal at this/these regions.

3.2.8 Adjusting plot settings and a quick way to explore subsets of regions

Each plot type has multiple variations, and the possibilities to adjust settings and vary the plots are too comprehensive to cover in detail here. Inspiration and an overview of the possibilities can be found here: http://easeq.net/plots.pdf. In the steps below, we will only briefly cover how a few of these settings are accessed and changed.

ee. To enter the settings for the 2D-histogram generated in 3.2.6 steps v-w, move the mouse over the plot, and click on the small icon with the gears that appear in the upper right corner of the plot.

This will open a rather busy panel with many settings that can be changed, and when the cursor exits this panel, the changes will be applied and the plot will update automatically. Settings that are not relevant to the selected plot type are greyed out and inactive. The best way to learn about their functions is to duplicate a plot, change a setting, and study the outcome. Often, the effect can be deduced rather intuitively by comparing the plots.

ff. For the 2D-histogram, we will change the default log scale on the x- and y- axes from log (default) to linear. First duplicate the plot to compare the difference by clicking on the small icon with a plus appearing inside the plot when the mouse pointer moves over it. Then open the Plot Settings as described in step ee above, and uncheck the "Log scale" checkboxes beneath the X-axis and Y-axis headers.

gg. *Optional*: Click on some data within the 2D-histogram with linear scales on the axes, select an area, and observe the corresponding regions in the other plot with the log scales.

hh. *Optional*: Make another 2D-histogram of Ring1B OHT vs. Ring1B ctrl. This can be done either in 3.2.6 steps v-w or by duplicating the histogram as described in 3.2.7 step z open the Plot Settings as in step ee above and change the Parameters in the dropdown menus named "Parameter" beneath the X-axis and Y-axis headers to that of the Ring1B OHT and Ring1B ctrl values quantified in 3.2.3.

3.2.9 Making subpopulations of regions from plots and using the Gate tool

If you are interested in understanding or exploring a certain subset of regions further, it is a good idea to make a new Regionset consisting of this subset. This can be done either by:

ii. Select an area with regions of interest in the 2D-histogram made in 3.2.6 steps v-w showing the Suz12 OHT and Suz12 ctrl values.

jj. Click on the small icon showing a pair of scissor. A new Regionset should appear in the Regionset panel.

Alternatively, it can be done in this manner:

kk. Open the Gate tool by clicking [Tools/Gate]. This will open a new window, where the criteria that should be used can be defined.

ll. First, make sure that the selected Regionset in the uppermost dropdown menu is the Suz12 peaks, then to use the ratio calculated in 3.2.4 select the Parameter named "log2('Suz12 OHT' divided by 'Suz12 ctrl')", and to get the subset of regions with a ratio higher than 0 select the ">" operator in the dropdown menu next to the selected Parameter and type 0 in the textbox. This should result in a small Regionset consisting of a few regions (48 if done similar to our example).

3.2.10 Quick ways to make identical figures of different populations of regions

It is often useful to investigate ChIP-seq signal at several sets of Regions – either to compare signal e.g. at different populations of regions, e.g. enhancers, CpG-islands, and TSSes, or to explore a subset of regions as covered in 3.2.9.

mm. Start by selecting the heatmap(s) made in 3.2.2.

nn. A faster way to duplicate multiple plots, rather than clicking on the small duplicate icon in each of plots, is to click on [Plots/Duplicate].

oo. Then try to click on either [Plots/Next Reg.] or [Plots/Prev. Reg.] and observe what happens.

The selected heatmaps should show the next or previous Regionset on the lists. The same would happen to any other selected plot type, but for e.g. 2D-histograms the Regionset would need to have identical Parameters for the plots to show similar information.

pp. *Optional:* With the above steps done, you might be ready for a larger and more demanding exercise. Access the paper by Blackledge et al [23] and remake panels D, E, F, and H from Figure 4 of this paper using the data available in the session, the approaches described above, and some creativity.

4. Notes

1. An often-sought type of data integration, is to do combined studies of ChIP-seq and RNA-seq to reveal the relationship between histone modifications and transcriptional readout. RNA-seq data can be imported directly as datasets and visualized in e.g. heatmaps or tracks as for ChIP-seq data. However, to identify differentially expressed genes, we recommend using dedicated tools such as Deseq2[26] possibly within the Galaxy environment[10], and then import the coordinates for the genes together with the statistics as a Regionset into EaSeq for further analysis and visualization. Then this regionset can be used as a basis to explore the ChIP-seq data similar as how the Suz12 peaks were used in the above examples.
2. The different types of data in EaSeq; Datasets, Regionsets, and Genesets; can often be used to

provide another type of data. Conversion from one data type to another or making a derived set of the same type can be done by several tools, including (sorted according to usage frequency): [Tools/Extract] (Geneset -> Regionset), [Tools/Peaks] (Dataset -> Regionset), [Tools/Gate] (Regionset -> Regionset), [Tools/Controls] (Regionset -> Regionset), [Regionsets/Modify] (Regionset -> Regionset), [Datasets/Modify] (Dataset -> Dataset), and [Tools/Pattern] (Dataset -> Regionset).

3. To provide as much freedom as possible, EaSeq does not adhere to a specific chromosome nomenclature. The names within different types of data, which should be compared, do however need to be the identical. While it is most common to use chr3 when referring to chromosome 3, shorter names are sometimes used to save space, and data where chromosome 3 is referred to as "3" are not uncommon. If e.g. an imported Regionset uses a different naming than the already imported Datasets, then EaSeq opens a wizard that will allow you to check, if the chromosome names have sufficient correspondence and possibly allow the chromosomes to be renamed.
4. EaSeq does not keep track of which reference genome that has been used to derive the different Datasets and Regionsets. It will therefore not be able to point out if two incompatible reference genomes are used, and the user has to keep track of this.
5. When newly generated, heatmaps and tracks will show the centre of the visualized regions. For visualization of peaks and TSSes this is sensible, but for gene bodies this is not the case as the gene body sizes vary a lot, and the middle of the gene body is rarely an interesting viewpoint (see 3.1.4 step r to generate a Regionset of gene bodies and 3.2.2 to make an example heatmap). Therefore, for such plot types it makes sense to enter the plot settings (3.2.8 step ee), click on the checkbox named "Rel." to use relative window sizes, and change the "Window size" to e.g. 200 bp. With this setting the size of the visualized window will be adapted for each region so that it shows an area corresponding to 200% of the region size, so for a 30 kbp long gene this would be 60 kbp. In this manner, regions with highly dissimilar sizes can actually be fit into the same heatmap or track.
6. EaSeq contains several options to normalize data. Per default, tracks and heatmaps of datasets (which are not from coverage vectors, see 2.2.3) are denominated in FPKMs, which stands for Fragments per Kilobasepairs per Million Reads. This can be disabled in the plot settings if needed. The same applies for quantified values, but in the Quantify Tool (Covered in 3.2.3) the individual components of this normalization can be enabled or disabled on a custom basis. For quantitations that are intended for count-based statistical analysis such as in Deseq2[26], the normalization should be disabled to provide the raw read counts. A frequently used normalization type in ChIP-seq is the use of spike-in DNA as an internal control[27]. Constants from such spike-in normalization can be entered and used to adjust Dataset values in the menu controlled by [Datasets/Describe].
7. A few words of caution. a) ChIP-seq is inherently unsuited for detecting massive and global changes in the abundance of the studied target, although this can be alleviated using spike-in of e.g. drosophila DNA as an internal standard for global normalization[27]. b) ChIP-seq is subject to linear and non-linear biases, even within different replicates – we have an example in figure 4c here[9]. c) Stochastic variation likely contribute to quantified differences, so it is advisable to only use regions with a high density of reads to limit this variation. d) If one set of data is used for both peak-calling and to calculate the ratio of a signal, there is bound to be a bias as the stochastic variation will both contribute to the definition of the regions and the values used for calculating the ratios. This is the case even for two technical replicates, where the one used for peak-calling would have a higher amount of signal in the peaks than the other. Some good ways to limit these effects is to use independent replicates for the peak-finding and the quantitation (in this case in 3.2 the peaks were established earlier, see [9] figure 8), and to use ChIP-seq for studies of local changes, preferably by comparing the effects at one set of regions to those at another. Finally, it is always strengthening the scientific value of the data to have multiple biological replicates, potentially using one set of data for finding interesting relationships, and the rest of the replicates for validating these observations.

References

1. Atlasi Y, Stunnenberg HG (2017) The interplay of epigenetic marks during stem cell differentiation and development. Nature Reviews Genetics 18:643. doi:10.1038/nrg.2017.57

2. Barski A, Zhao K (2009) Genomic location analysis by ChIP-Seq. Journal of Cellular Biochemistry 107 (1):11-18. doi:10.1002/jcb.22077

3. Berdasco M, Esteller M (2010) Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry. Developmental Cell 19 (5):698-711. doi:https://doi.org/10.1016/j.devcel.2010.10.005

4. Holmberg J, Perlmann T (2012) Maintaining differentiated cellular identity. Nature Reviews Genetics 13:429. doi:10.1038/nrg3209

5. Ouyang Z, Zhou Q, Wong WH (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences 106 (51):21521-21526. doi:10.1073/pnas.0904863106

6. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129 (4):823-837. doi:10.1016/j.cell.2007.05.009

7. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316 (5830):1497-1502. doi:10.1126/science.1141319

8. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448 (7153):553-560. doi:10.1038/nature06008

9. Lerdrup M, Johansen JV, Agrawal-Singh S, Hansen K (2016) An interactive environment for agile analysis and visualization of ChIP-sequencing data. Nature structural & molecular biology 23 (4):349-357. doi:10.1038/nsmb.3180

10. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46 (W1):W537-W544. doi:10.1093/nar/gky379

11. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome research 12 (6):996-1006. doi:10.1101/gr.229102

12. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32 (Database issue):D493-496. doi:10.1093/nar/gkh103

13. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44 (D1):D733-745. doi:10.1093/nar/gkv1189

14. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30 (1):207-210. doi:10.1093/nar/30.1.207

15. S A (2010) FastQC: a quality control tool for high throughput sequence data. . Available online at: http://wwwbioinformaticsbabrahamacuk/projects/fastqc

16. Wingett SW, Andrews S (2018) FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res 7:1338-1338. doi:10.12688/f1000research.15931.2

17. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15):2114-2120. doi:10.1093/bioinformatics/btu170

18. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17 (1):3. doi:10.14806/ej.17.1.200

19. Krueger F (2015) Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16):2078-2079. doi:10.1093/bioinformatics/btp352

21. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28 (10):1045-1048. doi:10.1038/nbt1010-1045

22. The EPC, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S,

Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee B-K, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Ki Kim S, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge E, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry JS, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SCJ, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanari E, Tress ML, van Baren MJ, Walters N, Washietl S, Wilming L, Zadissa A, Zhang Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Reymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyengar S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Lamarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan K-K, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenenbaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, White KP, Auer T, Centanin L, Eichenlaub M, Gruhl F, Heermann S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin TV, Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JM, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AD, Kharchenko PV, Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan K-K, Yip KY, Birney E (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57. doi:10.1038/nature11247

https://www.nature.com/articles/nature11247#supplementary-information

23. Blackledge NP, Farcas AM, Kondo T, King HW, McGouran JF, Hanssen LL, Ito S, Cooper S, Kondo K, Koseki Y, Ishikura T, Long HK, Sheahan TW, Brockdorff N, Kessler BM, Koseki H, Klose RJ (2014) Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell 157 (6):1445-1459. doi:10.1016/j.cell.2014.05.004

24. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133 (6):1106-1117

25. R Development Core Team R (2011) R: A language and environment for statistical computing. R foundation for statistical computing Vienna, Austria,

26. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15 (12):550. doi:10.1186/s13059-014-0550-8

27. Bonhoure N, Bounova G, Bernasconi D, Praz V, Lammers F, Canella D, Willis IM, Herr W, Hernandez N, Delorenzi M, Cycli XC (2014) Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome research 24 (7):1157-1168. doi:10.1101/gr.168260.113