Wardrobe

Experiment Management System

Table of Contents

Introduction	2
Access	2
Contributions and Acknowledgements	2
Abbreviations	2
Preliminary Analysis	4
Adding Records	5
Quality Control Tab	8
RNA-Seq	8
DNA-Seq	g
Genome browser Tab	10
RPKM list Tab	10
Islands list Tab	11
Average Tag Density Tab	11
Islands Distribution Tab	11
Advanced Analysis	12
Differential Gene Expression and Gene Sets	13
Average Tag Density Profiles and Heatmaps	14
Differential ChIP-Seq Enrichment - MAnorm	17
Wardrobe Settings	18
Personal Settings	18
Folders	18
Catalogs	19
User Management	19

Introduction

Wardrobe is a software package for storing and analyzing ChIP-Seq, RNA-Seq and similar datasets. The Wardrobe workflow is separated into two parts, preliminary analysis and advanced analysis. During preliminary analysis, sequence data are downloaded, processed, evaluated for quality and mapped onto a local mirror of the genome browser. Advanced analysis involves comparing and integrating different datasets.

Access

Wardrobe source code and installation guide can be downloaded from https://code.google.com/p/genome-tools/

At CCHMC Wardrobe experiment management system (https://ems/) can only be accessed from within Cincinnati Children's secure network (i.e. on site or via VPN) by authorized users. A laboratory account and laboratory administrator rights can be set up with Artem Barski, PhD (artem.barski@cchmc.org). Laboratory administrators can add or remove users from their laboratory account. To encourage collaboration, data can be viewed on the browser, but not analyzed, by all accounts irrespective of whether that information has been placed in a shared folder.

Contributions and Acknowledgements

The Wardrobe software package and users' guide were generated collaboratively by Artem Barski, PhD and Andrey V. Kartashov, MS. Editorial assistance was provided by Shawna Hottinger, MS.

Abbreviations

ATDP: average tag density profile

CCHMC: Cincinnati Children's Hospital Medical Center

ChIP: chromatin immunoprecipitation

ChIP-Seq: chromatin immunoprecipitation sequencing

DESeq: differential gene expression analysis **DNA-Seq:** deoxyribonucleic acid sequencing

ERCC: External RNA Control Consortium (ERCC), used in the context of an established set of RNA

standards

FASTQ: a next-generation sequencing data format

MACS: model-based analysis of ChIP-Seq

MAnorm: a model for quantitative comparison of ChIP-Seq datasets

MWW: Mann-Whitney-Wilcoxon test

QC: quality control

RNA-Seq: ribonucleic acid sequencing **RPKM:** reads per kilobase per million

Seq: sequencing

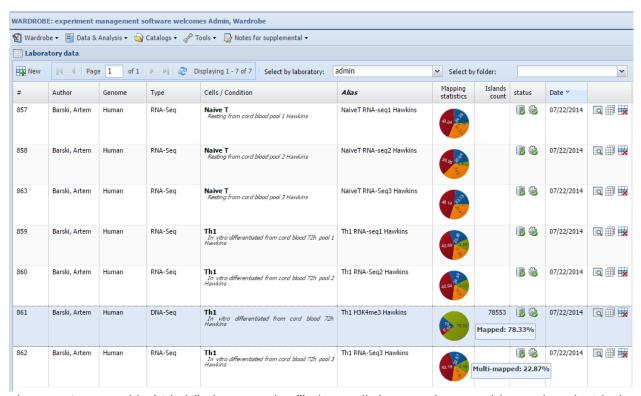
SRA: sequence read archive (previously known as short read archive), a type of data format

TSS: transcription start site(s) **VPN:** virtual private network

Preliminary Analysis

Preliminary analysis is performed, and results are shown in the Experiments window (Data & Analysis>Experiments).





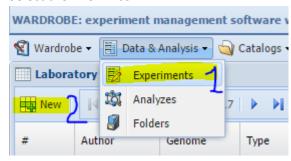
The Experiments table (titled "Laboratory data") shows all the records created by or shared with the user's laboratory. The datasets can be filtered by many variables, such as folder, owner and experiment type. The table shows the unique dataset identification number ("#") and owner of the record ("Author"), description parameters ("Genome", "Type", "Cells/Condition", "Alias") and some mapping statistics in the form of a pie chart ("Mapping statistics"). Pictograms show the status of the record download (HDD icon) and processing (gear icon). Please note that the status is updated every 10 minutes and that the system will not start downloading the record immediately. Depending on the number of new records entered at the same time, the records will have to wait in the queue for download or analysis.



Records can be viewed, duplicated or deleted using the three pictograms in the last column. Unneeded records can be deleted by clicking the delete icon (grid with red X icon). If the dataset has been downloaded from the CCHMC sequencing core, the record will be deleted, but the FASTQ file will be kept on the server. If the dataset was downloaded from the internet, it will be deleted.

Adding Records

- 1. Go to Data & Analysis>Experiments.
- 2. Select the "New" icon.



Tip: If the record to be entered is similar to one that already exists, it can be created by clicking the "Duplicate record" icon located in the last column of the table.

- 3. Fill in the Experiment form, which has four tabs: "General info", "Protocol", "Notes" and "Advanced".
 - a. Experiment Description (General info):
 - i. The following parameters in the General info tab are needed to describe your sample: Cells, Conditions, Fragmentation, Crosslink and Experiment Date. Donor field is not in use currently, but will be used for paired statistical analysis in the future.
 - ii. Genome Type: Select the genome to which the data will be mapped: Use the "+ spike" option only if you are using ERCC controls for RNA-Seq.



iii. Experiment Type:

- Select DNA-Seq for ChIP-Seq, DNase-Seq, MNase-Seq and all other experiments in which
 the data need to be mapped to the genome; select RNA-Seq for experiments in which
 mapping to a transcriptome is preferred.
- 2. Select paired-end ("pair") or single read. If paired-end read is selected, Wardrobe will expect two FASTQ/SRA files.
- 3. If the stranded RNA-Seq protocol was used, please select "dUTP", as this will affect RPKM calculations.



- iv. Antibody: Select your ChIP antibody or other protocol (e.g. DNasel). This selection will affect the pipeline parameters and island calling (e.g. narrow peaks vs. broad peaks in MACS). The pipelines and parameters for each antibody can be set up in the antibody catalog by laboratory-level administrators. You can also add antibody catalog # for your records.
- b. Experiment Arrangement (General info tab):



- i. Experiment's short name: Select a short name for the experiment. This name will display on the genome browser and will be used to designate the sample during analysis. Please select a name that is UNIQUE and SHORT but DESCRIPTIVE. DO NOT use "_" or "-" in lieu of spaces. The name can be edited after processing is complete.
- ii. Folders/Genome Browser folders: Folders are used to organize data on the browser and to share data between laboratories. Folders can be set up and shared by laboratory-level administrators (see Wardrobe Settings section). If the folder is not selected, the data will be downloaded but not analyzed. You can only select folders that belong to or have been shared with your laboratory. The data can be moved to another folder at any time, but relocating data will affect data visibility for other laboratories.
- iii. Checking the "Share data online?" box will deposit data on the external mirror of the genome browser if it has been set up. (Normally data are deposited on the browser available only on the internal network.)
- c. Data Source (General info tab):



Data can currently be obtained ("Download type") from a CCHMC sequencing core ("Core facility"), the internet (a direct URL to a FASTQ or SRA file) ("Direct link"), or a local file uploaded to the local wardrobe-upload shared drive ("local file").

- i. Core facility: Enter a filename or a code that will allow Wardrobe to find and download the file. Please note that the code needs to be unique (e.g. code such as AB1 will not work if there is also another sample named AB11)
- ii. Direct link: Please enter a URL for the file. The URL must end with the file name and extension. Wardrobe can process .fastq and .sra files that are uncompressed or compressed with .gz, .bz2, or .zip. Currently, .tgz files cannot be processed. If there are several files (e.g. sequencing data from several lanes), the URLs can be separated by semicolons (;). For paired-end data, the URLs for the first and second FASTQ should be separated by semicolons. *Currently, paired-end data coming from several lanes require manual processing.
- iii. Local file: The files should be copied to wardrobe upload share \EMS\WARDROBE-UPLOAD). To map the share go to Tools>Map network drive in Windows explorer. After copying the file, enter file name(s) in the URL field.
- d. Protocol and Notes tabs: You can enter protocol information and notes (including images and hyperlinks) in these tabs.



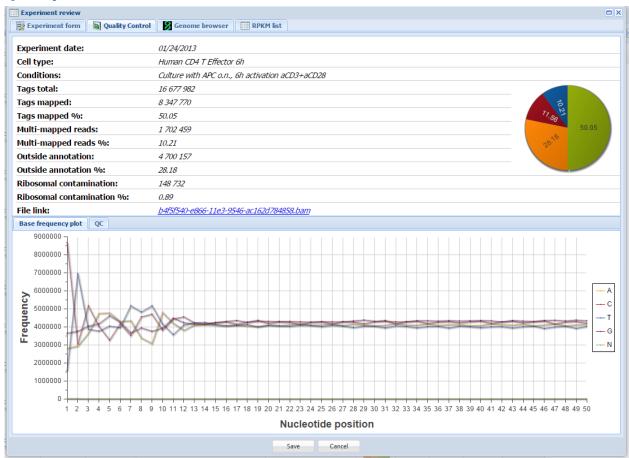
e. Advanced tab:

- i. ChIP-seq specific (Experiment additional info): Confirm or select the expected fragment size ("Expected Fragment Size"). In some cases, MACS does not determine the fragment size correctly. In these cases, you can force it to use the fragment size that you select by checking the "Force to use this fragment size?" box. Selecting the "Remove duplicates?" box will remove coincident reads (pairs of reads that start and end at the same position on the same strand) after alignment and before analysis.
- ii. Trim from the left/right (Experiment additional info): Selecting the "Trim from the left" or "Trim from the right" options allows getting rid of sequenced indices or poor-quality bases from either end of the read.

If you change any of the advanced options after finishing the initial analysis, select the "Force to repeat experiment analysis?" box (only visible after the initial analysis) and click the "Save" button to implement these changes in the analysis. Make and save the changes only ONCE and let the reanalysis finish before further modifying any of the options, otherwise the pipeline will break.

f. Click "Save".

Quality Control Tab



RNA-Seq

For RNA-Seq type data, the Quality Control tab shows the total number of reads ("Tags total") and the number of reads mapped in a unique fashion to transcriptome ("Tags mapped") or genome outside of the transcriptome ("Outside annotation"), multi-mapped ("Multi-mapped reads", which are discarded) and unmapped (percentage shown in red in the pie chart).

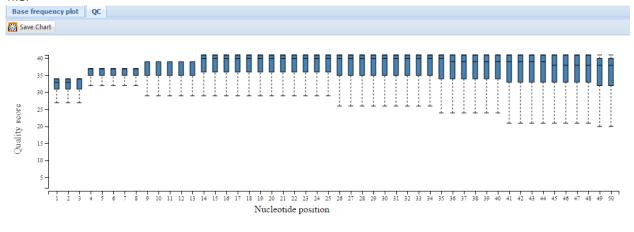
A high percentage of reads outside of the transcriptome (percentage shown in orange in the pie chart) may be due to DNA contamination and can inflate RPKM values by several points; this contamination is often a result of insufficient, on-column DNAse digestion.

Ribosomal contamination shows the number of reads that can be mapped to ribosomal DNA repeat. This reflects the completeness of ribosomal RNA removal during library preparation.

"File link" can be used to download the read mapping data (.bam) for use in other software. The same name can be used to find the sample's folder on the Wardrobe server.

"Base frequency plot" shows how often each base is present at each position of the read. Spikiness of the base frequency plot can help detect adapter contamination during library construction. The ratio of CG/AT can also indicate the presence of DNA contamination – in the human genome, the CG/AT is ~60/40, whereas in the human transcriptome, it is closer to ~50/50.

The "QC" box plot shows the distribution of Phred quality scores for each base as reported in the FASTQ file.



DNA-Seq

For DNA-Seq type data, the Quality Control tab shows the total number of reads ("Tags total") and the number of reads mapped in a unique fashion to transcriptome ("Tags mapped"), multi-mapped ("Multi-mapped reads", which are discarded), mapped duplicates ("Mapped duplicates", which may be discarded depending on the selected settings in the Advanced tab of the Experiment form tab) and unmapped (percentage shown in red in the pie chart).

In the case of single read sequencing **fragment size** is estimated by MACS ("Estimated fragment size") from the distribution of reads mapping to the top and bottom strands.

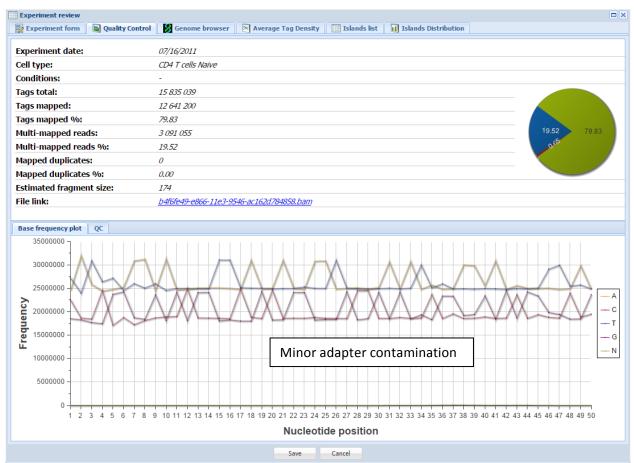
"File link" can be used to download the read mapping data (.bam) for use in other software. The same name can be used to find the sample's folder on the Wardrobe server.

"Base frequency plot" shows how often each base is present at each position of the read. Spikiness of the base frequency plot can help detect adapter contamination during library construction.

The "QC" box plot shows the distribution of Phred quality scores for each base as reported in the FASTQ file.

Genome browser Tab

The data are displayed on the genome browser as coverage per million reads. For single-read DNA-Seq, coverage is calculated after extending the reads to the estimated fragment size. For pair-end reads,



actual fragment sizes are used. Islands identified by MACS are shown under the signal track. For RNA-Seq, actual coverage by reads is used. For stranded RNA-Seq, reads mapping to the top and bottom strand are shown separately. Please note that for the dUTP method, the reads will map to the opposite strand.

RPKM list Tab

The RPKM tab displays for RNA-Seq type records and shows RPKM values for genes or isoforms. Total read number can also be shown (select the drop-down menu of one of the columns>Columns>TOT_R).

Values are initially determined by our algorithm for each RefSeq by isoform; however, they can also be summed up for common TSS (for promoter activity studies, e.g. average tag density profiles) and for genes (for functional analysis, e.g. GO) using the drop-down menu.

The table can be sorted, searched and filtered by clicking on the appropriate column headings. The "Jump" button opens the genome browser tab on the selected gene. The "Save" button allows the table to be downloaded in .csv format.



Islands list Tab

The Islands list tab displays for DNA-Seq type records and shows the list of islands as detected by MACS, statistics and the nearest genes. Islands are classified ("region" column) as intersecting with promoter, upstream (less than 20 kb), exon, intron or intergenic regions. The promoter radius can be adjusted; click the "apply" button after changes in the promoter size are made. If an island has multiple "summits" each summit coordinate ("start" and "end") is shown as a separate row together with the position of the whole island ("txStart" and "txEnd"). Checking the "Show uniq islands?" box will show each island (start and end position; "txStart" and "txEnd") only once and will not show summits. The "Jump" button opens the genome browser tab on the selected island. The "Save" button allows the table to be downloaded in .csv format.

Average Tag Density Tab

The Average Tag Density tab displays for DNA-Seq type records and shows the average tag density profile around all annotated TSS. Such graphs can be used to estimate the success of ChIP-Seq type experiments for some histone modifications (e.g. H3K4me).

Islands Distribution Tab

The Islands Distribution tab displays for DNA-Seq type records and shows distribution of islands between upstream, promoter, exon, intron and intergenic regions (in this order left to right). Area definitions in relation to the TSS are shown above the diagram for the upstream and promoter regions. The number and percentage of island counts per region are shown on each region in the diagram.

Advanced Analysis

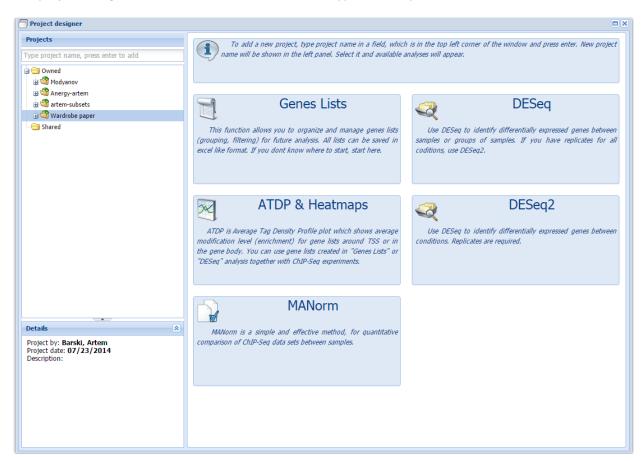
To open the Project designer window go to Data & Analysis>Analyze.



In the project designer, a new project can be created by typing its name in the top left field and then pressing enter on your keyboard.



The project designer window shows several available types of analysis:



Differential Gene Expression and Gene Sets

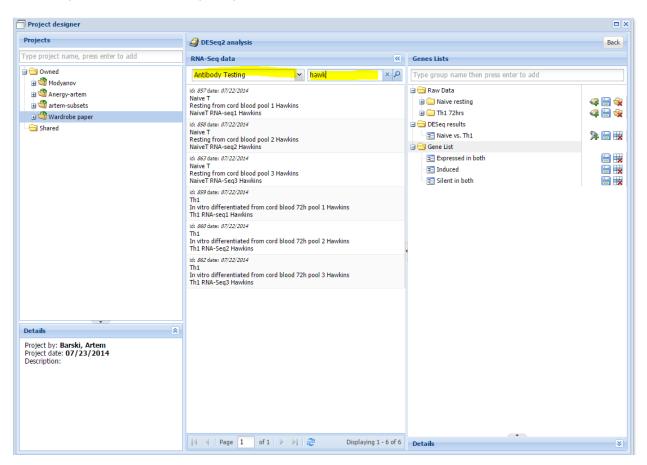
To perform the differential gene expression analysis, go to DESeq or DESeq2. DESeq allows the statistical analysis to be performed even if the data include monoplicates. If you have at least duplicates for all of the sample conditions, we recommend using DESeq2 because it utilizes more powerful statistical aproaches. As an example, we will perform the analysis described in our upcoming paper (Kartashov, Barski, in preparation).

To add raw data to the project, first create a condition by typing the name in the top right box (highlighted in yellow below).

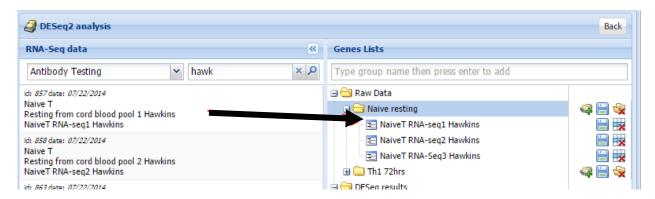
Next, identify datasets related to this condition. Open the RNA-Seq data tab by clicking the >> button.



In the "RNA-Seq data" tab, all of the RNA-seq experiments available to you are listed. These can be filtered by folder or searched by a keyword.



To add a dataset to the specified condition, drag it with your mouse from the "RNA-Seq data" tab to the appropriate subfolder of the Raw Data folder of the "Genes Lists" tab.



After defining the conditions and adding the desired raw datasets to these conditions, click the Fig. icon in the far right column to set up DESeq analysis. In the next window, list the conditions that you want to compare. To include additional conditions, use the "+". Provide a unique name for analysis and select whether you want to compare expression by isoform, TSS or gene. Series type is used when more than 2 conditions are compared. Assuming that we have 3 conditions (1, 2 and 3), selecting "Pairwise series" will perform all pairwise comparisons (1-2, 1-3, 2-3), whereas "Time series" will perform 1-2 and 2-3 and "Kinetics series" will perform 1-2 and 1-3. After setting up the conditions, click the run button.

DESeq analysis can take up to 10 minutes. After the analysis is complete, data can be saved in a .csv file by clicking the Save icon.

After the analysis is complete, we can filter genes and create gene sets by clicking the filter icon.

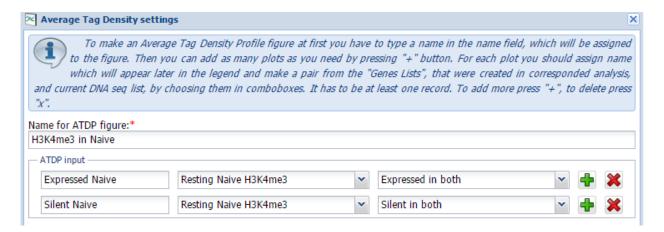


In order to create gene sets, DESeq results can be filtered using parameters such as p-value, p-adjusted, RPKMs, and chromosome, as well as logical operators (e.g. AND/OR).

Average Tag Density Profiles and Heatmaps

To compare the chromatin environment between gene sets using tag density profiles, select the "ATDP & Heatmaps" option in the Project designer window. Gene sets that were created as specified in the Differential Gene Expression and Gene Sets section of this guide are already present in this view. ChIP-Seq datasets can be added to the project in a manner similar to how RNA-Seq data are added (see Differential Gene Expression and Gene Sets section of this guide). To create average tag density profiles from an available gene set, select the graph icon of that gene set.

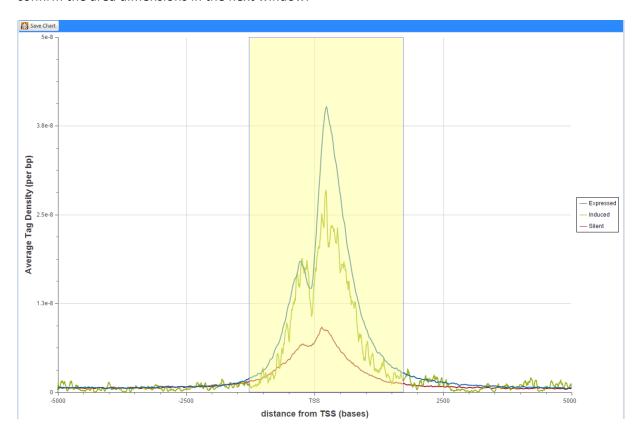
In the Average Tag Density settings window, the graphs can be set up in the "ATDP input" area by providing the name for the graph as a whole and a combination of gene set, ChIP-Seq dataset and line name for each plotted line.



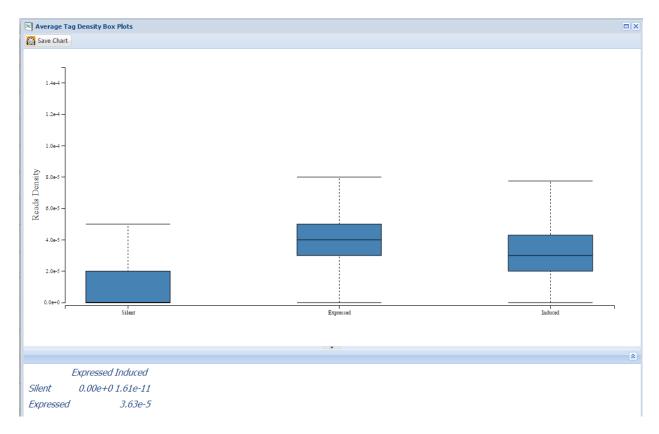
The graph will be calculated in a few minutes and can be viewed by pressing the magnifier icon.



Whether the level of modification (tag density) between gene sets is significantly different can be ascertained using the MWW test: highlight the area where you want to compare tag density and confirm the area dimensions in the next window.

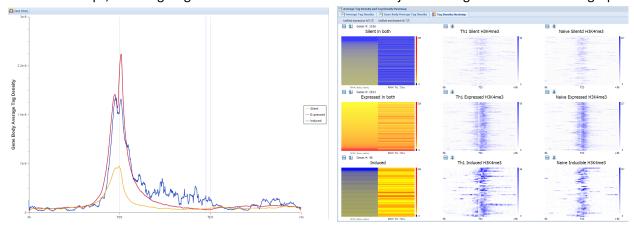


The resulting box-plot will show the tag density distribution between groups and MWW-based p-values.



Other tabs in the Average Tag Density Profile tab of the Project Designer window will show a similar graph for the gene body and tag density heatmaps.

In the heatmaps, ordering of genes and the color scale can be adjusted using buttons above the graphs.

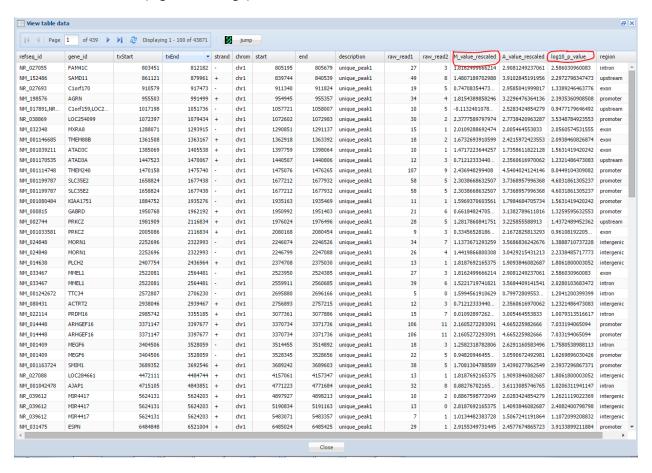


Differential ChIP-Seq Enrichment - MAnorm

Areas of the genome that are differentially modified can be identified using MAnorm. The set-up window can be opened by clicking the MAnorm icon in the Project designer window. Adding raw data, and identifying conditions are done similarly to the DESeq analysis (see Differential Gene Expression and Gene Sets section of this guide).



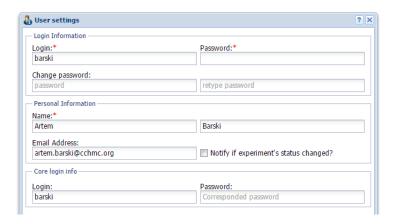
MAnorm analysis can take up to an hour. The results can be viewed or saved using the icons next to the analysis result. In addition to displaying the islands, the table shows the neighboring genes and where the island is located relative to these genes. We recommend filtering the list on the basis of both the p-value and rescaled M (log2 fold change).



Wardrobe Settings

Personal Settings

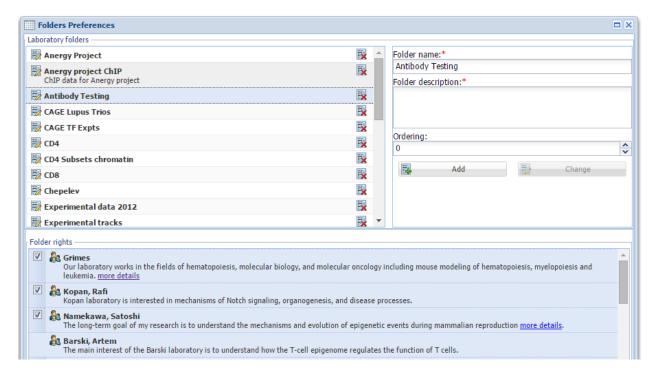
Go to Wardrobe>Personal Settings to add/change your personal information, including e-mail and password. Please make sure to enter your login name and password for the CCHMC sequencing core facility correctly.



Folders

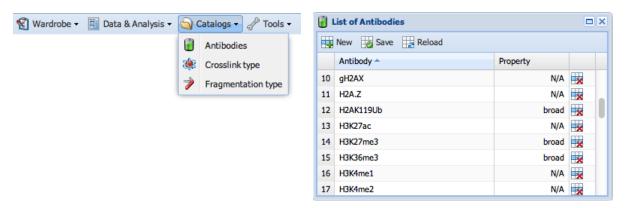
Folders can be added, shared and removed by laboratory-level administrators. To enter the folder management window, go to Data & Analysis>Folders.

Folders can be added by typing the new folder's name in field on the top right of the window. Folders can be shared by selecting the folder and adding a checkmark next to the laboratory name. When folders are deleted, the contents have to be copied into another folder (The copy dialog will appear after clicking the delete button).



Catalogs

Wardrobe uses catalogs to define several terms, such as conditions for fragmentation, crosslinking and antibodies. Only laboratory-level administrators can edit the catalogs. Currently, only the antibody catalog is used to define pipelines that will be utilized with a given antibody. The only selection is whether to use the narrow peaks or broad peaks method in MACS for a given antibody.



User Management

To add a user, go to Wardrobe>Users and groups. There will be the list of users from your laboratory. Here laboratory administrators can add users and reset their Wardrobe and CCHMC core facility passwords.

