# How to do ChIP-seq analysis

## PREFACE

This tutorial aims to give you an introduction to ChIP-seq data and to ChIP-seq data analysis in Qlucore Omics Explorer. The tutorial uses an inbuilt dataset that comes with Qlucore Omics Explorer.

## Contents

## 1. REQUIREMENTS

This document is written for users that would like to start working with analysis of ChIP-seq data in Qlucore Omics Explorer and would like to get familiar with the functionality with an example that is easy to follow. It is required that you have a license of Qlucore Omics Explorer 3.7, base and NGS module, version 3.7 or later. A trial license is sufficient to do this tutorial.

## 2. CHIP-SEQ ANALYSIS TUTORIAL

This is a tutorial explaining what ChIP-seq is, what it can be used for and how Qlucore Omics Explorer can help with the analysis of ChIP-seq data.

### 2.1. CHIP & CHIP-SEQ

Chromatin immunoprecipitation (ChIP) is a laboratory protocol for purifying genomic regions associated or bound to specific proteins of interest, or even specific modifications of proteins as occur on histones. Combined with next generation sequencing (ChIP-seq), it is possible to map out genome-wide protein binding profiles to determine promoters or other binding sites, or to investigate epigenetic regulation.

### 2.2. CHIP-SEQ ANALYSES

ChIP-seq analysis typically involves the detection of 'peaks' in read coverage for regions where the protein binding sites are located. These peaks manifest from the targeted immunoprecipitation serving to enrich genomic content from bound regions. However, even in whole genome sequencing such peaks might be detected spuriously. It is therefore critical to capture this bias and to control for it during analysis. Therefore, modern analyses often include an 'input' sample fractioned off for whole genome sequencing just prior to ChIP enrichment. Some analytical methods require each ChIP-seq sample to have a corresponding input sample, while others share the input. The commonality, however, is the desire for robust detection of true binding sites. From this, it should be noted that many of the methods employed for ChIP-seq analyses closely mirror those of other protocols and it is only the initial enrichment protocol in the laboratory that changes significantly (e.g. ATAC-seq). With such a broad range of uses, one would think peak detection and analysis is further complicated by a plethora of software tools. While many tools certainly exist, one of the most adaptable and therefore popular toolkits is MACS2. As a general all-purpose toolkit, it ranks highly across publications and is only occasionally beat by more specialized tools. Its interface can be daunting to non-bioinformaticians and integrating ChIP-seq analyses with other omics data is especially challenging. The Qlucore Omics Explorer packages MACS2 into its NGS module to make peak analyses easily relatable and explorable.

### 2.3. CHIP-SEQ IN QLUCORE OMICS EXPLORER

#### Processing and data generation

Qlucore Omics Explorer support peak detection using MACS2. The resulting peak files can then be merged into a consensus bed file. A user supplied group of bed files can also be merged using our consensus bed algorithm.

The resulting consensus bed file can then be used as a feature file for generating a count matrix for the peaks. It is also possible to use a user generated bed file as the feature file. This enables the user to extend peak analyses beyond standard types and to evaluate custom regions of interest such as promoters, structural hotspots, pseudogenes, personalized/non-reference genomes, etc.

A count matrix can also be generated using a gtf-file as the feature file for a truly gene centered ChIP-seq analysis. In this scenario, a padding can be added before and/or after each gene to include proximal binding profiles that may be important for gene regulation.

If the experiment also includes RNA-seq data, it is possible to also generate a count matrix for that. It is then possible to swap between the RNA-seq and ChIP-seq count matrices during the analysis. Combined with the ability to annotate samples and variables, this functionality allows the user to investigate the behavior of both data types without running two separate instances of Qlucore Omics Explorer.

## Genome Browser

The genome browser in Qlucore Omics Explorer features powerful filtering mechanisms making it easy to find interesting features. The browser also allows the user to annotate peaks and export the result as a bed file. For better visualization, the genome browser now allows direct sample annotation as well as independent labeling, coloring, and sorting. Additional information about genes, variants, gene fusions and much more can be added for a more complete understanding of the data.

## Count analysis

For analysis of the peak count data, the full power of Qlucore Omics Explorer is available including many statistical tests, machine learning with clustering and classification and powerful visualization. The ChIP-seq workflow also include generation of powerful variable annotations to provide a connection between peaks and nearby genes.

## Combined analysis

In Qlucore Omics Explorer the powerful statistics and machine learning applied to the count matrix can be combined with the detailed view of the genome browser to provide the best of two worlds. Once the most interesting peaks have been found using statistics, machine learning or visual analysis of the count data they may be included in follow up analyses as ‘active variables’; effectively filtering out all other variables for a more targeted approach. These active variables can be used as a filter in the genome browser so that a detailed analysis of the peaks themselves can be performed, confirmed, and rendered into publication quality figures.

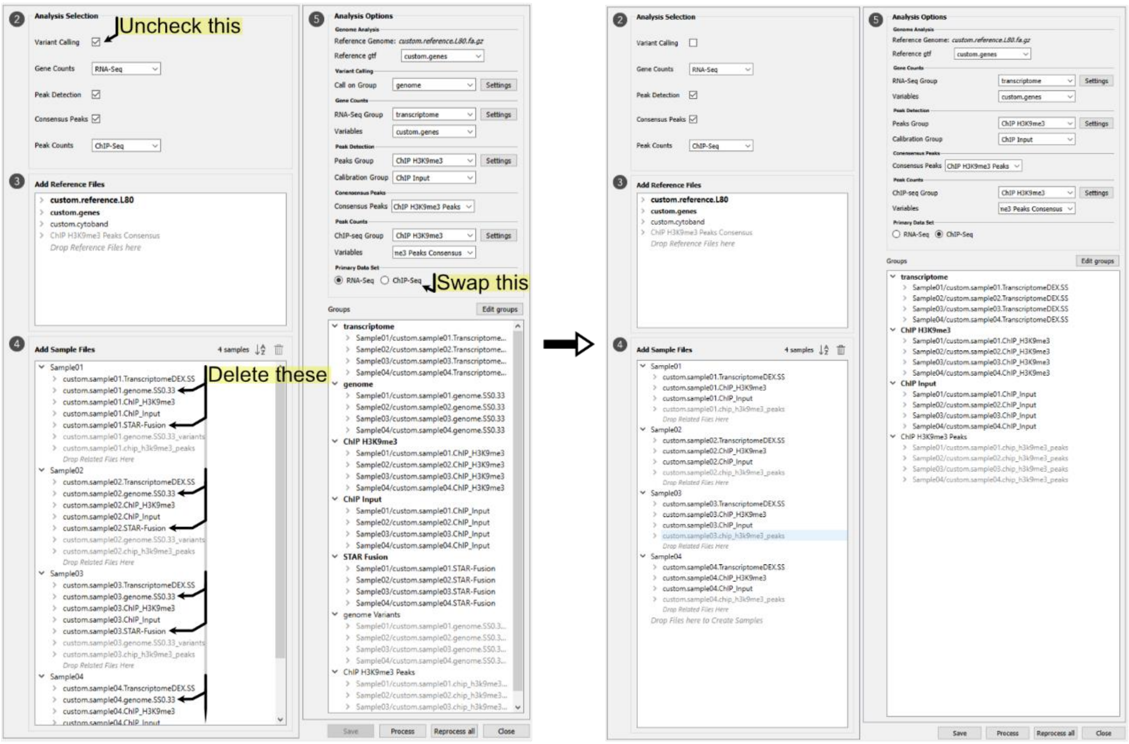
In the genome browser, on the other hand, interesting peaks can be annotated and exported as a variable list to further analyze and subset the count matrix. Peaks can also be added to a bed file for further use in the browser or downstream tools.

## Brief toy example

Qlucore Omics Explorer is distributed with a toy example showing, among others, ChIP-Seq H3K9me3 and corresponding transcriptomic data. We will run through a hypothetical exploratory analysis using this dataset. In the analysis, we will identify the ChIP-Seq peaks with the most variance, by read count, across all samples. We will find that these peaks cluster samples into two groups and examine the H3K9me3 pattern. We will then ask whether these peaks are significantly different between these sample groups and find other candidate peaks. We will note the genes proximal to these significant peaks before switching analysis to the transcriptomic data. We will find that the proximal genes are also differentially expressed between the same sample groups. Finally, we will note that one new gene is differentially expressed and examine its H3K9me3 profile to find that while read counts are similar, the binding profile is indeed different. This tutorial will show the ability of Qlucore Omics Explorer to adapt to your research questions and will highlight the power of combining different types of data. It is spread out over 6 parts and 38 steps. Let us begin:

## 3. TUTORIAL PART 1: GETTING SET UP

1. Open Qlucore Omics Explorer, version 3.7 or greater.
2. Open the example NGS dataset by clicking on *Help > Ngs Example Files > Example.qcproj*.
3. This will open the Project Manager as depicted on the left of screenshot 1, below. If you want your test project changes to be persisted, change the name of the project since the Example project is overwritten each time it is opened. As indicated in the screenshot, uncheck the “Variant Calling” box as we are not interested in the genetic differences between the four individuals of this example. Also, delete the genome and fusion files highlighted under “Add selected files” by clicking them and pressing the delete key on your keyboard. Finally, swap the radio button selection from RNA-Seq to ChIP-Seq. This will tell Qlucore Omics Explorer that you wish to primarily analyze the ChIP-Seq count matrix data first. The RNA-Seq data will still be available in the genome browser for visual comparison. If you switch to RNA-seq, that count matrix will be used for calculations instead, and the ChIP-Seq data would be available for visual comparison in the genome browser. You will do an RNA-Seq analysis later in this tutorial. For now, you should end with a project manager that looks the same as the right of screenshot 1.


Uncheck this

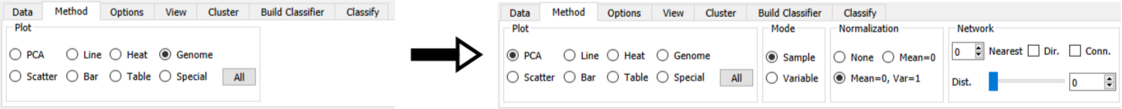
Swap this

Delete these

*Screenshot 1. Setting up the project manager (left: default, right: new)*

4. Take your time to familiarize yourself with the layout of the Project Manager, it is the central hub through which you will manage the lifecycle of your project. We will return to this step once more in the tutorial, but when analyzing a real-world dataset, it is likely you will return here fairly regularly to edit settings and rerun analyses.
    1. ***NOTE**: Feel free to check/uncheck analyses in step-2 (❷) and observe what options in step-5 (❺) become available to you. However, please make sure your Project Manager looks the same as shown in the right of screenshot 1 before continuing.*
    2. ***NOTE**: If you ever get lost, you can always close the Project Manager and reload the dataset from the Help menu.*

5. Click the “Process” or “Reprocess all” button to the lower right of the Project Manager. Since this dataset was already processed by us for you, then by clicking “Process” you will immediately load the results with the genetic data removed. The “Process” button is a great option if you have made a few changes within the project manager and simply want to update your dataset quickly. If you make large changes or are unsure of how they might affect your project then be safe and click “Reprocess all”. This will build all results from the beginning and will take some time.
6. After processing the data, you will be taken to the main viewport with a few small windows opened to show you diagnostics and statistics depending on what you chose in the previous step. Close these windows and maximize the main view if it is not already.
7. By default, the NGS browser view is showing because this is genomic data. We will return to this view in a few steps. Now we will analyze the variance and statistical significance of predicted consensus peaks using a PCA plot to help visualize the effect of filtering peaks on our ability to discriminate between samples. Change to the PCA view as shown in screenshot 2. *BONUS*: Open a second synchronized plot window and change it to a PCA keeping the original in Genome.



Plot: ○ PCA ○ Line ○ Heat ● Genome ○ Scatter ○ Bar ○ Table ○ Special

Plot: ● PCA ○ Line ○ Heat ○ Genome ○ Scatter ○ Bar ○ Table ○ Special

Mode: ● Sample ○ Variable

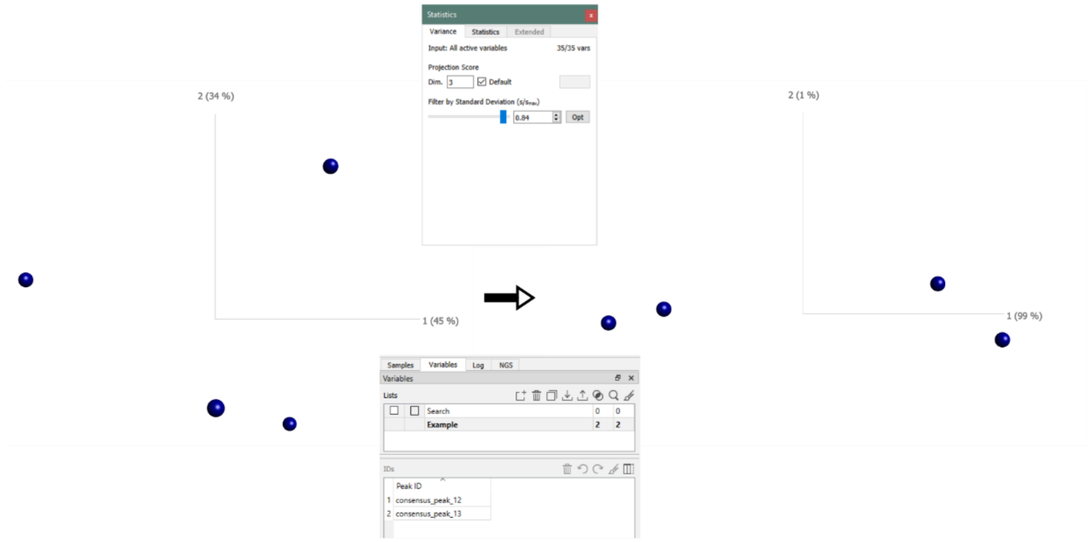
Normalization: ○ None ○ Mean=0 ● Mean=0, Var=1

Network: 0 Nearest □ Dir. □ Conn. Dist. 0

*Screenshot 2. Changing plot type, from Genome to PCA (left: default, right: new)*

## 4. TUTORIAL PART 2: STATISTICAL EXPLORATION AND DEFINING AN UNKNOWN PHENOTYPE

8. Once in PCA view you should observe a PCA plot like the one shown on the left of screenshot 3. You will see that there are a few aesthetic differences such as label font and line thickness. If you would like to have a similar or different look then you can play around with aesthetic mappings in the Plot Settings window accessible from the View tab.
9. Click the statistical analysis button () to open the Statistics window as shown in the top middle of screenshot 3. Change the tab from “Statistics” to “Variance” and begin sliding the Standard Deviation filter over to the right. Observe how the PCA begins to change and stop when you observe two tight clusters of two samples each, as shown on the right of screenshot 3. In this example, a threshold of 0.84 has been used resulting in the first component explaining 99% of the variance.
10. If you click the Variables tab on the main window and then click on the variable set labeled “Example”, you will find that you have removed all but two of the most variant consensus peaks explaining why a single component explains so much of the data (see bottom middle of screenshot 3).



Statistics

Variance | Statistics | Extended

Input: All active variables 35/35 vars

Projection Score

Dim. 3 ☑ Default

Filter by Standard Deviation ($s/s_{max}$) 0.84 Opt

Samples | Variables | Log | NGS

Variables

Lists

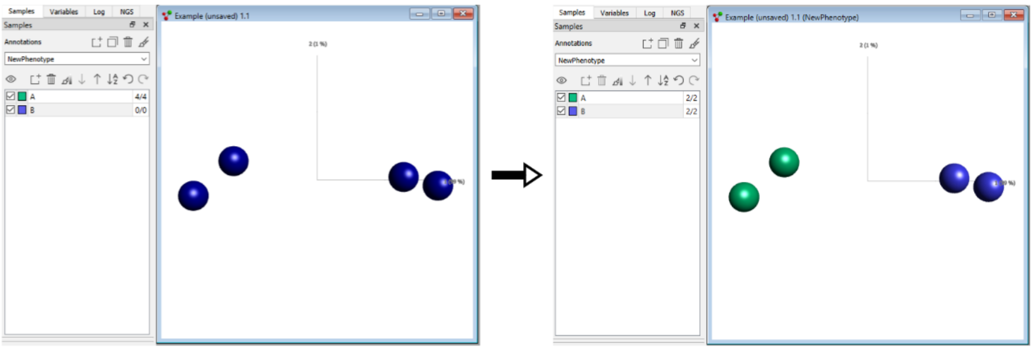
| | | | | |
|---|---|---|---|---|
| □ | □ | Search | 0 | 0 |
| | | **Example** | 2 | 2 |

IDs

| | Peak ID |
|---|---|
| 1 | consensus_peak_12 |
| 2 | consensus_peak_13 |

*Screenshot 3. Finding peaks with the greatest variance (left: unfiltered, right: filtered)*

11. You may also notice that the consensus peaks are labeled 12 & 13. Later, you will see that they are contiguous and belong to the same H3K9me3 signature.
12. It is quite likely that the methylation pattern identified here is significant between the two sample groups. Let us see if this is true, and at the same time, let's find out if there are any other significant peaks that may differentiate between these two groups.
13. You may lose or put aside the Statistics window for now. Click the Samples tab on the main window as shown in screenshot 4. Then, add a new sample annotation (□) to record our new potential phenotype. Label it, "NewPhenotype", and create two annotation values (□) within it. Label them, "A" and "B". You should have something like the left of screenshot 4 where sample A has 4 samples associated to it and sample B has none.
    3. ***NOTE**: The display settings have been altered a little to make the screenshot more readable. Specifically, the size of sample points have been increased and the aspect ratios of windows changed.*
14. Associate the two right clustered samples with annotation "B" by clicking the annotation in the list (☑ 🟦 B 0/0), then clicking the annotation tool (◉ Annot.) in the toolbox, and finally by clicking and dragging a circle around the two right samples. You may have to click on the paint icon (🖌) to recolor the PCA by the new annotation. You should end up with something like the right of screenshot 4. Well done, you have defined an annotation based on something interesting during your exploration of the data. It is time to turn this into a hypothesis for testing.



Samples

Annotations

NewPhenotype

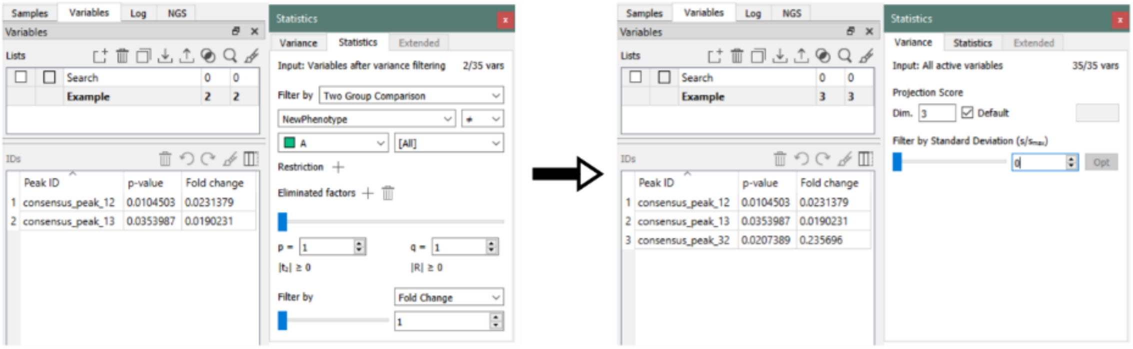
| | | | |
|---|---|---|---|
| ☑ | 🟩 | A | 4/4 |
| ☑ | 🟦 | B | 0/0 |

| | | | |
|---|---|---|---|
| ☑ | 🟩 | A | 2/2 |
| ☑ | 🟦 | B | 2/2 |

*Screenshot 4. Creating an artificial phenotype (left: original, right: new annotation)*

15. Focus on the Statistics window again or open it (⚠) if closed. Change the tab to "Statistics". Under filter by, select "Two group comparison" and choose the new sample annotation you made, "NewPhenotype". As a two group comparison defaults to a null

hypothesis of A=B. This test is done in the background. You can now view the p-values and fold-changes for the two peaks by clicking on the Variables tab in the main window and choosing which variable annotations you want to view using the column selector (▥). You should have the same results as shown on the left of screenshot 5 revealing that the peaks are indeed significant. Note that the q-values show significance is not maintained after correction for multiplicity.



| | | | |
|---|---|---|---|
| ☐ | ☐ | Search | 0 | 0 |
| | | **Example** | 2 | 2 |

| | Peak ID | p-value | Fold change |
|---|---|---|---|
| 1 | consensus_peak_12 | 0.0104503 | 0.0231379 |
| 2 | consensus_peak_13 | 0.0353987 | 0.0190231 |

Input: Variables after variance filtering 2/35 vars

Filter by: Two Group Comparison

NewPhenotype; A; [All]

p = 1 q = 1

$|t_2| \geq 0$ $|R| \geq 0$

Filter by: Fold Change 1

| | | | |
|---|---|---|---|
| ☐ | ☐ | Search | 0 | 0 |
| | | **Example** | 3 | 3 |

| | Peak ID | p-value | Fold change |
|---|---|---|---|
| 1 | consensus_peak_12 | 0.0104503 | 0.0231379 |
| 2 | consensus_peak_13 | 0.0353987 | 0.0190231 |
| 3 | consensus_peak_32 | 0.0207389 | 0.235696 |

Input: All active variables 35/35 vars

Projection Score

Dim. 3 ☑ Default

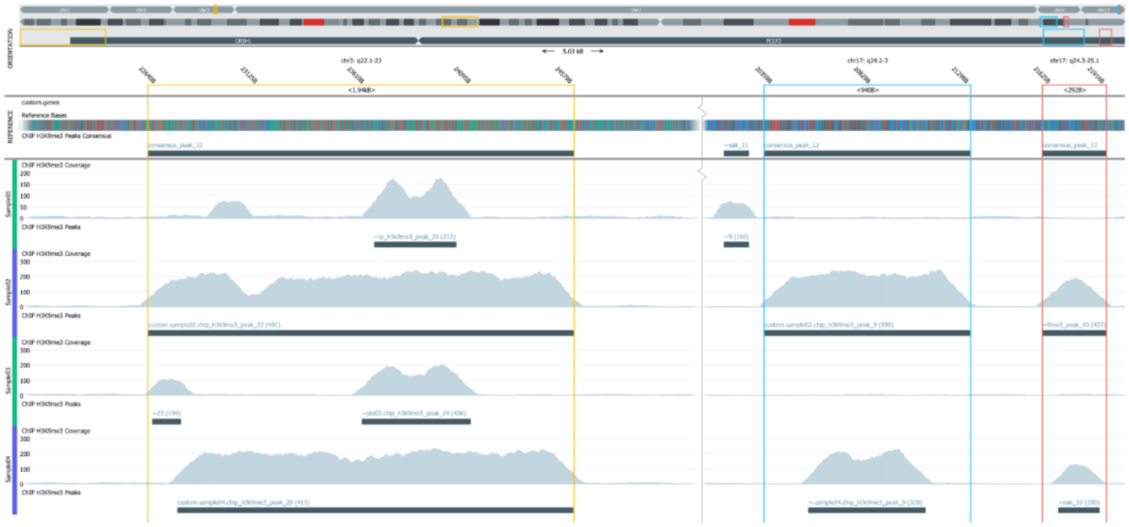
Filter by Standard Deviation ($s/s_{max}$) 0

*Screenshot 5. Evaluating significant peaks (left: high variance & significant, right: only significant)*

16. You can explore the statistical significance of all peaks, regardless of their variance (remember we set a very high variance threshold). To do so, set the p-value threshold to 0.05 in the Statistics tab and remove the variance threshold in the Variance tab by setting it to zero. You will observe the same results as shown on the right of screenshot 5. Notice that the PCA plot changed only slightly, and we discovered a new and potentially relevant peak. You may close the statistics window for now.

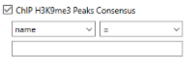
## 5. TUTORIAL PART 3: EXPLORING SIGNIFICANT PEAKS ON THE GENOME BROWSER

17. Now that we have identified significant peaks for our new phenotype, we can explore their profiles and positioning in the genome browser. As you will see, Qlucore Omics Explorer provides powerful built-in tools to rapidly narrow down on regions of interest and even display them side-by-side. To return to the genome browser view, click on the Method tab in the main window and select the Genome radio button (○ PCA ○ Line ○ Heat ● Genome ).
18. Click on the NGS tab ( Samples Variables Log NGS ) and then check Active Variables (☑ Active Variables). This will filter the genome view to only show the 3 significant peaks identified in step 16 side-by-side.
19. Clean up the data and make sure the zoom is sufficient to see all three peaks at once. Try to obtain a view similar to screenshot 6.
    a. You can clean up the data by not displaying the RNA sequencing data as tracks. We do not need this data yet. To hide these tracks look for the Sample tab under NGS ( Reference Sample ) and make sure that “transcriptome Coverage” is unchecked (☐ transcriptome Coverage ).
    b. You can adjust the viewport size (zoom) by clicking on the View tab ( Options View Cluster ) and clicking the zoom icons (🔍 🔍 🔍 🔍). In this example, you can also simply set the View Port Size to 5032.
    c. You can recolor the annotation with the NewPhenotype colors on the side by using the paint icon (🖌) as you did previously. *BONUS*: You might also use the “Color”, “Label”, and “Order” options in the main View tab further customize and annotate your figure.



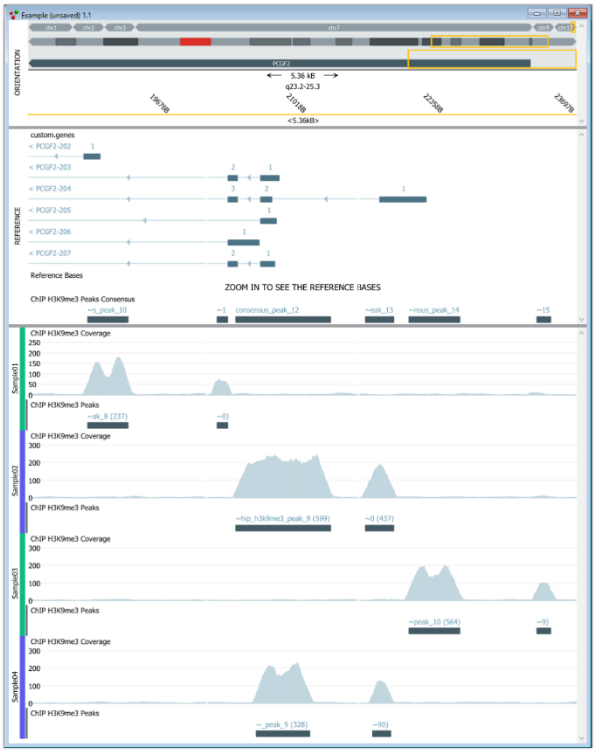
*Screenshot 6. Viewing peaks in the genome browser side-by-side*

20. Notice the three view port sections (yellow, peak_32; blue, peak_12; red, peak_13). Only peak_32 has prominent H3K9me3 profiles across all 4 samples. We will soon see, by focusing on each region, that peaks 12 and 13 belong to a gene with significant alternative transcription start sites.
21. To examine each region independently uncheck Active Variables. Then navigate to consensus_peak_32. You can do this by inputting the absolute coordinates of the peak in the Options tab. Or, as now demonstrated, by creating a filter on the variable name.
    a. Select the Reference tab under NGS ( Reference Sample ) and click-and-drag the "ChIP H3K9me3 Peaks Consensus" item up to where it says, "*Drag and drop annotation groups from panels below to filter.*". You should now see this →

    
    ☑ ChIP H3K9me3 Peaks Consensus
    name =

    b. Type into the empty field, "consensus_peak_32".
    c. The browser will move you to the genomic location and lock your viewport. To remain at the genomic location but remove the lock, simply uncheck the filter (☐ ChIP H3K9me3 Peaks Consensus ). You can create multiple filters like this to quickly navigate back and forth between several locations of interest.
22. Zoom out a little by increasing the size of the view port and you should notice a nearby gene, OR5H1, and no other H3K9me3 patterns nearby (disregarding low levels of background noise).
    *4.* ***Note****: Another observation to make is the difference between patterns for samples 1 & 3 and samples 2 & 4, both in terms of size and shape. In general, we know that H3K9me3 marks accumulate around active transcription start sites both slightly up and down stream. They also typically accumulate more upstream of the start site within the promoter region giving the characteristic pattern observed in samples 1 & 3.*
23. Now, navigate to consensus_peak_12 and consensus_peak_13 using your favorite method. They are very close to one another. Make sure your view port is not locked if using the filter method and zoom out until you observe other nearby histone marks.

24. You will now see that if you turn on “Display All Transcripts”, that these peaks do exist between all samples and that they perfectly align with the different transcription starts of PCGF2, see screenshot 7. To view all transcripts do:
    a. Under the main View tab (Options View Cluster) click the “show” button under Plot Settings.
    b. This will bring up a context sensitive window that will change its options depending on what you select on the screen.
    c. Click on the “custom.genes” track in the main window next to the vertically aligned “REFERENCE” section.
    d. You will now see a section in the middle of Plot Settings window labeled, “Display Mode”, select the radio button labeled, “All transcripts”, and close the window.
    5.
25. Now that we have some insight on the different histone marks, it is time to switch our focus to the transcriptomics side and of the equation and determine if there may be a link between the two. Indeed, H3K9me3 is a strong mark for active transcription. These PCGF2 transcripts all appear to have the same degree of tri-methylation, so without a transcript aware pipeline we are unlikely to detect differential regulation. However, consensus_peak_32, just upstream of OR5H1, certainly demonstrated significant differential tri-methylation aligned to the same transcript.



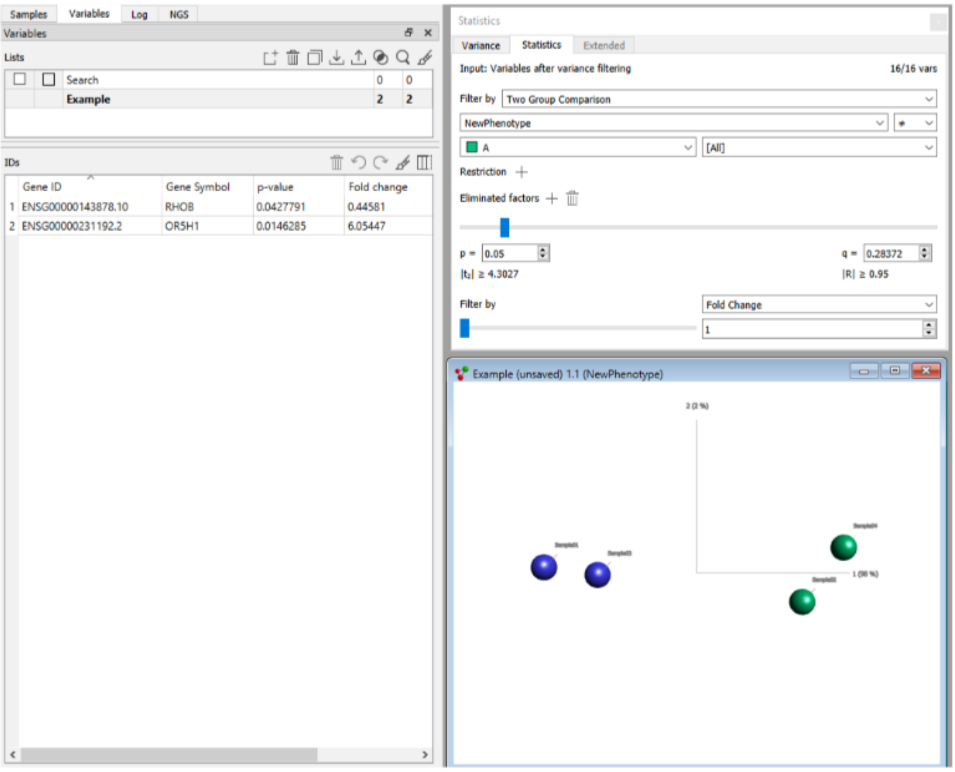
*Screenshot 7. Multiple transcriptional start sites*

## 6. TUTORIAL PART 4: SWITCHING ANALYSES TO EXAMINE THE TRANSCRIPTOME

26. To switch focus to transcriptomics we must first tell Qlucore Omics Explorer that the main variables of interest must be derived from our RNA-Seq experiments. To do this, open the Project Manager via the File menu. Once there, change the Primary Data Set from ChIP-Seq to RNA-Seq (● RNA-Seq ○ ChIP-Seq) and click the “Process” button on the bottom right.
27. The main view will reset, and you will notice the variables under Variables (Samples Variables Log NGS) now refer to gene IDs.
28. *28.* Once again, we will do some statistics before examining the genome browser. Change the plot type to PCA under the Method tab and open the Statistics window. *BONUS*: As before, try multiple synchronized windows showing both the PCA and Genome plots.
29. Make sure variance filtering is off by setting it to zero. Define your hypothesis to be consistent with the last, i.e. set *filter by* to “Two group comparison” and select the NewPhenotype filter.
30. Set the p-value threshold to 0.05 and then display it and the fold-change in the variables table. Note that OR5H1 has a six-fold upregulation between groups A and B. Also, there is no mention of PCGF2 as expected since we did not use a transcript-aware algorithm. However, an additional gene has revealed itself in RHOB with a significant ~2-fold downregulation (p-value 0.043; screenshot 8).
    a. There is certainly some evidence for H3K9me3 mediated effect on OR5H1 expression. However, statistical analysis of the correspondence between variance in methylation and variance in OR5H1 transcript abundance is beyond the scope of this tutorial (this sort of analysis is possible using the Extended tab in the Statistics window or by exporting data to another tool).

b. RHOB, on the other hand, is a new target that did not reveal itself in the ChIP-Seq analysis. In this analysis RHOB is borderline differentially expressed by a factor of 2. Time to look at the combined ChIP-Seq and RNA-Seq data to figure out why.

6.



|   |   | Lists | | |
|---|---|---|---|---|
| □ | □ | Search | 0 | 0 |
|   |   | **Example** | **2** | **2** |

|   | Gene ID | Gene Symbol | p-value | Fold change |
|---|---|---|---|---|
| 1 | ENSG00000143878.10 | RHOB | 0.0427791 | 0.44581 |
| 2 | ENSG00000231192.2 | OR5H1 | 0.0146285 | 6.05447 |

Input: Variables after variance filtering 16/16 vars

Filter by: Two Group Comparison

NewPhenotype

🟩 A [All]

Restriction

Eliminated factors

p = 0.05 q = 0.28372

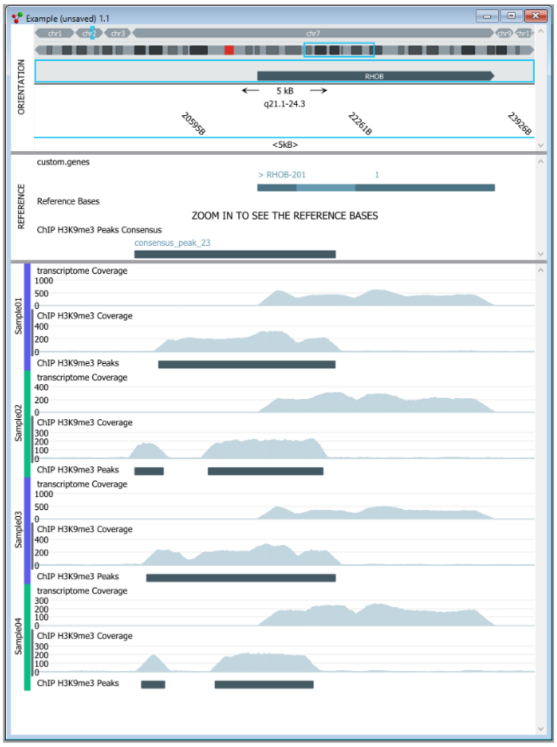
$|t_2| \geq 4.3027$ $|R| \geq 0.95$

Filter by: Fold Change 1

*Screenshot 8. Evaluating significant differentially expressed genes.*

## 7. TUTORIAL PART 5: ANALYZING RHOB ON THE GENOME BROWSER

31. Close the Statistics window and change the plot back to the genome browser.
32. Under the main NGS tab, create a filter for the “custom.genes” reference track and set “gene_name” equal to “RHOB”.
33. Make sure the view is not locked and zoom out until you can see all data for the gene and the H3K9me3 peaks.
34. On visual inspection we observe very little difference (screenshot 9) between the peaks in terms read count. The ChIP-Seq analysis confirms no significance in this regard. However, while samples 2 and 4 show typical tri-methylation patterns consistent with active gene expression, samples 1 and 3 show an atypical spread of methylation between upstream promoter and the transcriptional start site.



*Screenshot 9. Different H3K9me3 profiles, same depth*

## 8. TUTORIAL PART 6: SUMMARY

35. In this toy example we examined the ChIP-Seq and transcriptomic data from a simulation of H3K9me3. We did not have a predefined phenotype of interest and instead defined our own by exploring the data. We found a false positive result in consecutive peaks 12 and 13 in terms of read-count but showed that there is a difference in alternative transcription start sites. Statistical analyses revealed a new peak of interest near OR5H1. We switched to a transcriptomic analysis and showed that OR5H1 is differentially expressed suggesting that histone marks are a significant source of regulation for the gene (as indeed was originally simulated). Finally, we find a surprising but small effect between RHOB expression and the spread (not amount) of tri-methylation. This putative finding is tenuous and not statistically robust (but, for interest, the effect was indeed simulated).
36. This tutorial has walked you through a lot of functionality both new and old within Qlucore Omics Explorer for the analysis of ChIP-Seq and ChIP-Seq-like data. Each PART of this tutorial may be further expanded, and we invite you to read the documentation as each topic is thoroughly explained there.
37. Thank you and well done for completing this tutorial. If you have any questions, please send them through to **info@qlucore.com**.

## 9. USAGE, ACKNOWLEDGEMENTS ETC

Qlucore Omics Explorer makes use of the MACS2 library, copyright (c) 2019, Tao Liu lab at Roswell Park Comprehensive Cancer Center and Xiaole Shirley Liu lab at Dana-Farber Cancer Institute, All rights reserved.

## 10. DISCLAIMER

The contents of this document are subject to revision without notice due to continuous

progress in methodology, design, and manufacturing.
Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.
Qlucore Omics Explorer is only intended for research purposes.