FG1 - Learning Objectives

Written by: Klaas, Simon, Siebren, Christa . . .

Contents

Some general notes

- Question-/Problem-driven learning –> formulate candidate questions per week.
- Example dataset: monocytes, CD4 and CD8 T cells, B cells and NK cells).
- Start in UCSC genome browser with one cell type (eg monocytes). Introduce same data of other cell types in week 2.

Overall

Focus/paradigms/hypotheses from lectures

- Specific functional genomic regions (eg promoter, enhancers etc) are associated with distinct combinations of histone modifications. Also, chromatin and heterochromatin are also characterized by distinct histone modifications.

- Active and repressed genes are associated with distinct sets of histone modifications. Recognition of a histone modification may lead to regulation of gene transcription. Eg The basal transcription factor TFIID interacts H3K4me3 via its PHD finger domain.

- Vice versa, using epigenetic data (eg presence/absence of histone marks and chromatin accessibility) we can predict the transcriptional state of a gene or functionality of a region.

- Heatmaps based on genomic coordinates - in which the intensity of the color indicates the strength of the signal - allow for a very compact and powerful visualisation and therefore interpretation of epigenomics data.

- Presence of bivalent chromatin domains marks develomentally regulated genes, thought to mark genes taht will need to be switche don in the near future and are 'poised' for transcription.

'Research goals' of this practical

- Identify and visualize which epigenetic marks or combinations thereof are found at distinct genomic elements.

- Vice versa, use the combination of histone PTM data and chromatin accessibility to identify functional genomic elements (eg promoter, enhancers etc) and predict gene activity (active or repressed).

- Check predictions on gene activity in single-cell RNA-seq data (or with bulk RNA-seq?).

- Use the epigenetic data to explain gene expression differences detected in a single-cell RNA-seq experiment.

Global learning objectives

1. Understand how data are generated from a genomics/ATAC-/RNA-ChIP-sequencing experiment.

2. Know how to display and browse through genomics data in the UCSC Genome Browser and formulate some basic hypotheses using this tool.

3. Tell apart different genomics datasets based on their pattern int he UCSC Genome Browser.

4. Quantify and visually summarize the genomic distribution of histone ChIP-seq peaks and signal strength in r.

5. Statistically test for enrichment of histone marks in functional genomic elements.

6. Summarize genome-wide ChIP-seq signal in a heatmap and use this heatmap to identify identify functional genomic elements and predict gene activity.

7. Check hypotheses about the epigenetic landscape based on single-cell RNA-seq data with actual epigenomics data, and vice versa (check hypotheses on gene activity in single-cel RNA-seq data).

8. Formulate hypotheses about (cell-specific) TF function using histone ChIP-seq and ATAC-seq data.

Additional goals may inclule

- Use combinations of histone PTMs to identify heterochromatin and euchromatin regions in these cells.

- Find/Hypothesize/suggest TFs that could drive the expression of important marker genes of a given blood cell? (by looking at motif enriched in promoters and enhancers and TF gene expression)

- Presence of bivalent chromatin domains marks at genes that are suggested to be 'poised' for transcription.

Week 1

- Focus: how we make data out of a ChIP-seq experiment, how to use the UCSC Genome Browser, and familiarize ourselves with the UCSC Genome Browser visualization.

- *seq2science, UCSC Genome browser*

- Global objectives #1-3.

- Perform in H5P?

- Discuss ChIP-seq in presentation

- Follow seq2science workflow

- Peakcalling: note that it is not black/white, not absolute, thresholds are set. NOT into technicalities and mathematics of read piling/peak models, narrow-wide peaks

Learning Objectives (*after this tutorial the student should understand/be able to...*)

1. Know how raw read output (from eg ChIP-esq) is processed to obtain trustworthy data.

2. Understand how to mov ("*browsing*") around in the UCSC genome browser and know how to go to regions and genes of interest.

3. Recognize the following elements/regions in the UCSC genome browser: *chromosome, centromere, telomere, gene, exon, intron, utr, promoter, enhancer/repressor.*

4. Tell apart UCSC genome browser tracks of H3K27ac, H3K27me3, H3K36me3, H3K4me1 and H3K4me3 (and H3K9me3) histone ChIP-seq and ATAC-seq signal (and ChIP-seq input) based on their coverage profile and their location with respect to genes and other annotated genomic/chromosomal elements.
5. Can interpret - and thus understand - ChIP- and ATAC-seq signal in the UCSC genome browser.

Working questions

1. What histone mark marks the promoter of TNF gene? Let's go and look for it ;).

2. Now your turn, what about CD3 is this gene active in monocytes? (no it is not, also logical, it is part of the T cel receptor present on the cellsurface of T cells)

3. Look at the enhancer located 20kb upstream of .... [gene] .... Based on the histone and chromatin accessibility data would you say this enhancer is active or inactive? What histone marks do you find around it?

4. *Summarizing* What marks are generally found at promoters, at gene bodies, active enhancers and at repressed domains?

Week 2

- Focus: *Specific functional genomic regions (eg promoter, enhancers etc, (and their activity ?)) are associated with distinct combinations of histone PTMs.*

- *Bioconductor, count peak distribution and compare marks among regions*

- Global objectives #4-5.

In week 1 we have estimated which mark is enriched in promoters, gene bodies or other intergenic regions. This week we will put a number to these estimates. More specifically,..

Question from my side, what would be a good first step to look at for students?:
- Count the occurrences of a histone PTM per genomic element and then state something about the enrichment of that mark in a particular genomic element?
- Or, count the overlap of histone PTMs with each other and with ATAC-seq signals and assign a function

to these combinations (eg active enhancer, repressed enhancer, active promoter, gene, bivalent promoter)
- introduce heatmap in this week (question 8) or do it next week?

Learning Objectives

1. Move around in GRanges object which holds and can handle tables with genomic coordinates.

2. Obtain annotation of your reference genome of interest (in TxDb object) and know how to extract the features important to you.

3. Perform intersections with the r objects from (1) and (2).

4. Find and count the overlap (using intersection) among histone ChIP-seq peaks and genomic elements (`countOverlaps(query, subject)`, `findOverlaps(query, subject)`).

5. Plot this quantification (ggplot + geom_bar).
6. Statistically test for enrichment of histone marks in (a) particular genomic region(s).

7. Optionally, quantify and plot the overlap among histone marks (a tiny bit like chromHMM but much less fancy, ggplot + geom_bar with index at the bottom for the different combinations or plot the tabular output as heatmap with nubmer display).

Working questions

1. Look in the benome browser for the H3K4me3 in monocytes or look back at your answer of question 5 in week 1. This time we will count all the overlap between H3K4me3 peaks and different genomic elements to find out in which region most of the H3K4me3 peaks occur?
2. Is this enrichment in promoters more than you would expect by chance?

3. And what if we look at H3K27ac or H3K36me3?

4. Do these numbers differ (significantly?) when we look at T cells instead of monocytes?

5. Now we will count the overlap among histone marks in monocytes. What genomic elements are associated with each combination? (one question but will take some time for students)

The bar graph or heatmap we made for question 4 does not take along peak height (signal strength) which could hold additional information does not allow a locus-specific comparison among different cell types. For that we would need to, somehow, include the genomic location int he plot.

Luckily, we have another option to plot the occurrences and co-occurrences of histone marks around a chosen genomic landmark (eg TSS) that *does* take along the genomic locus and signal strength: a heatmap. We will do so next week.

Week 3

- Focus: *Histone PTMs positioning is cell-specific* or *Different cell types show distinct, cell-specific histone enrichment patterns (associated with cell-specific gene expression and the cell-specific activity of TFs (if we want to go into motifs))*

- Global objectives #4-6
- Heatmap using package ngs.plot, EnrichedHeatmap, deepTools, ???

- Use heatmap to compare marks or cells? Or both? Do first all marks in one cell

In week 1 we looked at the localization of histone marks (peaks) and we quantified this in week 2 of peaks. You might have seen that some histone marks show partial overlap with other marks. This week we will summarize and, importantly, compare histone markings genome-wide *among different cell types*

Learning Objectives

1. Be able to visually summarize read coverage (*signal*) of different histone marks around genomic landmarks (coverage plot around TSS or over gene body).

2. Summarize genome-wide ChIP-seq signal *of different histone marks* in a heatmap and cluster this heatmap (for 1 or 2 cell types only, not more. Centered on identified ATAC-summits or any peak midpoint?) and use this heatmap to identify/annotate functional genomic elements.

3. In a related way, use a heatmap to identify differentially marked promoters among different cell types (centered at TSSs, plot H3K4me3 signal).

4. And identify differently marked enhancers among different cell types (centered on ATAC peaks not in promoters(?), using H3K4me1 and/or H3K27ac).

5. Assign enhancers to genes (and understand the assumptions/decisions you make when you do this).

Working questions

1. Does H3K36me3 signal start at the TSS or at the ATG/start codon and does its signal increase or decrease downstream from ATG/start (vs upstream)? First look in the GB to formualte a hypothesis. Then plot.

2. Are H3K4me3 and H3K27ac signal higher or weaker directly upstream of the TSS compared to directly downstream of the TSS? Do these marks behave the same?

3. What histone marks are found around ATAC peaks and do these reflect different functional genomic elements?

4. Do NK cells look more like monocytes or like T cells? Based on H3K4me3 signal in promoters.

5. Find the top 3/5/10 differently marked promoters in each cell type?

6. The [TF] is essential for T cell phenotype. Based on histone markings at the promoter, would you argue that it is exclusively expressed in T cells or not (only compare with the cells we have data for here).

Week 4

- Focus: *Interpret single-cell RNA-seq data, formulate hypotheses about epigenome and test hypothesis by plotting the data for these genes*

- or *find or count motif occurrences using pattern matching from GRanges*

- Use PBMC3k dataset, get UMAP and top marker genes per cluster/cell type from Satija lab.

- We continue with clusters that differentially mark T cells from monoyctes [..] and assign them to closest gene.

- Open/active regions only?

- Compare these to + and - marker genes from scRNA-seq

nb:
- Platelets are anuclear but rich in RNAs including mRNAs (Schubert et al. 2014 Blood)

Learning Objectives

1. Can interpret UMAP and gene epression plots (FeaturePlots) from scRNA-seq.

2. Understand the added value of scRNA-seq over bulk rna-seq (lecture not this workshop).

3. Use the methods and plotting strategies from the first three weeks to test

Working Questions

1. As we discussed before, CD3 is part of the T cell receptor and should be expressed by all T cells. Can you confirm this with the scRNA-seq data?

2. What is the most likely explanation for the lack of CD3 epxression in some dots (a dot is a cell) that are clustered as T cell?

3. From week 3 we had a list of top differently marked promoters per cell type. For which are you able to confirm cell specific gene expression?

4. These [file] are the top marker genes from scRNA-seq experiment. Which of the following hypotheses about the epigenetic markings of these genes, do you agree with?

5. Let's look at ChIP-seq signals in around their TSSs using a heatmap (hopefully there are some bivalent promoters). What is the most likely explanation for detecting repressive marks?