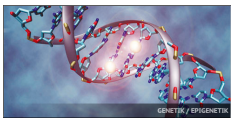
Epigenetic Data Processing Course, WS 2019/20
Prof. Dr. Jörn Walter
Abdulrahman Salhab, Michael Scherer, Kathrin Kattler
Department of Genetic/Epigenetics & Saarland University



UNIVERSITÄT DES SAARLANDES

Integrative analysis

1 Share assay-specific results with other groups

1. In principle you should be able to access all needed files from /vol/data/share/. However, feel free to ask the corresponding group for the data you need. Be careful in using the data, make sure that you don't change it or overwrite it.
2. You can explore the folder structure to have a general idea about the available data from the other groups.

2 IGV

If you created an IGV session yesterday, load it. Expand the current session to include the data listed here (one of the liver or kidney samples will be enough):

- ChIP-seq – signal tracks and segmentation
- WGBS – signal tracks and segmentation
- RNA-seq – signal track

1. Do you see a correlation between DNA methylation and histone marks at certain regions? what type of correlation (positive, negative)? save some snapshots.
2. Which states from DNA methylation segmentation (MethylSeekR) overlap with chromatin segmentation (ChromHMM). save some snapshots.
3. What are the methylation states of the promoters and the gene body in general (high, low)?
4. Can you find some examples for unexpressed genes. What are the methylation states of their promoters and bodies? Which histone marks are enriched at the promoters and bodies of these genes? save some snapshots.
5. Upload tracks of deferentially expressed genes (DEGs), differential methylated regions (DMRs) and the two chromatin segmentation tracks (16-state model). Can you find examples for genes that are regulated by DNA methylation and(or) histone marks. save snapshots.

3 Integrative analysis

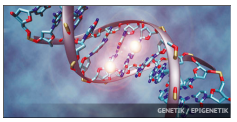
Below follows a list of various tasks. They do not have to be done in order. Feel free to prioritize what you find interesting for you.

1. **Summarize signals over regions of interest**
   a) Define regions of interest (ROI) – generate BED3 files[1] for one or more of the following regions:
      - Enhancers from the ChIP-seq segmentation (treat the different classes of enhancer similarly). Enhancer states have "Enh" in the label name (use the 16-state model).
      - Active promoters (AProm) from the ChIP-seq segmentation (use the 16-state model).
      - Promoters based on the gene annotation:

        `/vol/data/reference/mm10/rna-seq/gencode.vM2.annotation.gtf`

        Please, consider only "protein coding" genes, mind the gene direction and use 1500bp downstream and 500bp upstream of the TSS.

[1] https://genome.ucsc.edu/FAQ/FAQformat.html#format1



- Methylation segments generated through *RnBeads* (HMR, PMD, LMR and UMR)

b) Combine `computeMatrix` with `plotHeatmap`/`plotProfile` from deepTools2[2] to generate plots. e.g. histone mark/DNA methylation signals across all/some of the above mentioned ROI.

c) Based on you plots draw a biological conclusion about DNA methylation and histone marks/chromatin states/methylation segments.

2. **Chromatin states and DNA methylation**

a) Choose two samples (one liver and one kidney) and calculate the average DNA-methylation in the chromatin states (16-state model) of the corresponding samples.(hint: make use of bedtools intersect, groupBy to make a file with the following columns: chr, start, end, state, average methylation)

b) Read the resultant files into R and plot the average methylation per state per sample as boxplots (hint: ggplot(), geom_boxplot(), facet_wrap()).

c) Which states are the highest/lowest methylated and with which histone mark(s) is(are) associated with these two extreme methylation states (hist: use the emission heatmap from ChromHMM segmentation for the interpretation)?

d) Is there a difference in the average methylation between the two samples across the same chromatin states?

3. **How many DMRs are in heterochromatic regions (Het)?** (Hint: `bedtools intersect`)

4. **How many DMRs are in partially methylated domains (PMD)/Highly methylated Regions (HMRs)?** (Hint: `bedtools intersect`)

5. **Differential methylation and gene expression**

a) Annotate each DMR to the closest gene (hint: bedtools closest)

b) Do you see cases where multiple DMRs are assigned to the same gene? How often? In total how many unique genes are associated to DMRs? realize the number of DMRs and number of genes associated to. What do you conclude from these numbers?

c) plot two heatmaps to visualize mean methylation per group (from DMRs) and mean CPM values of the closest annotated genes (hint: heatmap or ComplexHeatmap in R).

d) What kind of correlation do you observe between DNA methylation and gene expression?

6. **Chromatin states and gene expression**

a) Create a BED6 file (see 1a) containing the following information for each gene:

```
chr    start    end    gene_ID    TPM    strand
```

You will need to combine the information from *genes.results files (from RNA-seq data) and the annotation file /vol/data/reference/mm10/rna-seq/gencode.vM2.annotation.gtf

b) Choose one sample (one replicate) and plot the coverage of all seven histone marks across the generated bed file ± 3kb (Hint: `plotHeatmap` from deepTools2).

c) Which marks are enriched at the TSS and which are enriched in the gene body?

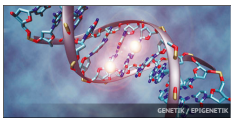
d) How are the genes ordered in the heatmap?

e) Try to cluster the genes into 6 classes with k-means. What do you observe?

f) Plot the gene expression values of the 6 clusters as boxplots. Which cluster contains the highest/lowest expressed genes and which are the most prominent histone marks associated with these two gene categories.

7. **Annotation and enrichment** – For the regions of interest from above (see 1a), and/or the list of differential genes try to utilize some of the following web services:

[2] https://deeptools.readthedocs.io/en/develop/index.html



- Gene IDs as input:
  - String – `https://string-db.org`
  - David – `https://david.ncifcrf.gov/`
  - Reactome – `https://reactome.org/`
- Regions as input:
  - LOLA – `http://lolaweb.databio.org/`
  - GREAT – `http://great.stanford.edu/public/html/`

8. **Enhancers and gene expression (Difficult!)**

   a) For each sample, extract a BED3 file (see 1a) with enhancers from the 14-state segmentation model (treat the different enhancer classes similarly)

   b) Create a union set of enhancers by merging all regions (Hint `bedtools merge`)

   c) Check length distribution and control extremely large regions in IGV

   d) Annotate each union region with absence/presence of the enhancer in each sample.

   - Suggestion: For each sample generate a tab separated file on this format (but without the header):

chromosome	start	stop	sample	measure	value
chr18	12334	12444	15_2	enhancer	1/0

     See the R packages `reshape2` and/or `data.table` and functions `dcast` and `melt` for inspiration. These files can then be concatenated and easily read into R.

   e) For each sample and histone mark, calculate the average signal for each histone mark. Either use the same format as above, or generate a big matrix, that later can be "melted".

   f) In the same manner, calculate the average methylation for each sample and enhancer

   g) For each enhancer, find the closest TSS and together with the distance.

   h) Together with the CPM value matrix from the RNA-seq experience, load all generated files into R and combine them into one large matrix.

   i) For both expression and histone marks, it might be relevant to apply log_2 scaling. Look to distributions to decide.

   j) As a first step set up a simple linear model with expression as dependent variable:

```
#Model:
expression ~ sample + group(early/late) + TSS_distance + enhancer \
    + methylation + H3K...
```

   k) Apply ANOVA to the linear model to see which variables are most explanatory

   l) Can we apply more complex machine learning approaches to this problem?

4 Preparation phase for presentation

As mentioned before, the examination of this workshop is a small presentation by each group. You will have 10 minutes to present your results followed by 5 minutes of discussion.

Please, take time to prepare slides during the workshop. When you present a figure, make sure you can explain the content if it isn't obvious.

As a guideline, have one slide/figure for each major step, narrow it down to 10 min and the most important results.