Bioinformatics
Date of completion: 13.05.2020
Software project report

Software project "Functional Genomics"

Integrative analysis of next generation sequencing data

Maximilian Otto, Lucas Rieckert and Kevin Zidane

Max Planck Institute of Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany

Associate Editor: Alena van Bömmel, Robert Schöpflin

Abstract

Motivation: With this project we want to examine the influence of histone modifications on the expression of developmental genes in the development of a mouse embryo and whether we can predict the gene expression based on the histone modification profile.
Results: The results of our study show that the histone modifications have a big influence on the expression of genes, since they change, for example, the spatial structure of the DNA. This is the way they change the accessibility of genes. Our study shows that the histone modification profiles are a good indicator for the expression profiles of genes.
Contact: max.otto@fu-berlin.de, acedon@zedat.fu-berlin.de, keviz94@zedat.fu-berlin.de

1 Introduction

It is generally known that every living being on our planet inherits his traits to his descendants through the information of the genetic code, stored in the DNA. But the sequence of the DNA-bases is not the only mechanism controlling the genetic information. The availability of the genes is one of the key elements in controlling genetic information. This availability is regulated by the modification of the spatial structure of the DNA-molecules. The probably most important kind of modifications is a modification of the histones. These are proteins with the DNA-helix winding around them to compress itself, consisting of 8 subunits. The histones can be modified by the binding of functional groups to one of its subunits. These bindings change the way the histones interacts with the DNA-molecule, for example an acetylation changes the electrostatic charge of the histone and by that how strongly the histone attract the DNA-molecule and therefore which genes are physical available to be transcribed. The different kinds of histone modifications and there influence on the expression of genes are the target of this report. We want to examine how 6 of these modifications influence the expression of genes responsible for the embryonic development of a mouse across different tissues and different stages of embryonic development. We orientated our work on the paper from David U. Gorkin *et al.*, 2017 to produce similar results.

2 Methods

2.1 Data

The data for our considered three tissues were retrieved and downloaded from Encode-portal (2020). For each tissue we picked data from ChIP-Seq-analyses, Control-ChIP-analyses, ATAC-Seq-analyses and RNA-Seq-analyses in ".fastq" format out from two researching universities (UCSC (Big Ren Lab) and Caltec (Babara Wold's Lab)) for the following development states of an embryonic Mouse:

- cells from embryos 14,5 days post conception (E14.5d)
- cells from completely developed mouses, postnatal (P0)

For the downloaded and uniformly renamed files we performed a quality control, to check, if the fastq-files retrieved from the various analyses reached acceptable quality scores. We used the FastQC and MultiQC software from Andrews *et al.* (2012) for this processing step.

2.2 Univariate analysis

2.2.1 Mapping

After validating the quality of our data, all the technical replicates belonging to one of the two biological replicates of each analysis data set were concatenated by a simple bash script [making usage of a for-loop calling a bash command], according to the ENCODE data processing flow charts corresponding to the overall accession numbers. The ChIP-Seq data were aligned to the mouse reference genome mm10 using Bowtie2 v2.2.1 (Langmead and Salzberg (2012)) according to the processing script

from the ENCODE 3 pipeline, producing BAM-files as output format. Afterwards, the BAM-files got sorted by SAMtools' sort (Li *et al.* (2019)), while removing unmapped reads and those under a certain the mapping quality score threshold (T = 30) (function call : "samtools -@ 50 view -F 1804 -q 30 <input file> -o <output file>") For calling the peaks in a further processing step, all duplicates of the previous created output files were marked by applying Picards' MarkDuplicates Jens Reeder (2018) and got removed through a SAMtools' "view" call.

We mapped the raw reads of the ATAC-Seq data on the mm10 reference genome using Bowtie2 v2.2.1 by Langmead and Salzberg (2012) following the instructions described in the ATAC-Seq-pipeline (2016) under the step "1a. Read alignment for paired-end ATAC-Seq Data" to produce the associated BAM-Files. After we used the view and sort function of SAMtools for these steps, we need to filter these mapped reads. First of all we removed the insufficient mapped reads; we used a threshold of 30 so every read with a score beneath this threshold got removed of our created BAM-Files. This was also performed by SAMtools view with the following parameters: "samtools view -F 1804 -q 30 -u <input file> | samtools sort /dev/stdin -o <out file>". After that, we removed all duplicates which are a result of the sequencing technique. First of all, we marked the duplicates using Picard from Jens Reeder (2018) and then removed these duplicates using SAMtools view for this removal. Since the downloaded ATAC-Seq data is paired-end sequencing data, we needed to process it once more, because we wanted to take a look at it as two single-end strands. To achieve that, we needed to edit the flags, more specifically: The flags which carry the information for a paired end. We did this using the Python3 package "Pysam" from Li *et al.* (2020) and edited the flags of every read. We saved that in a second BAM-File, so that we have one BAM-File with flags and one BAM-File without the mate flags for each originally downloaded file.

Mapping and sorting the RNA-Seq data sets was done with a different toolkit called STAR (Dobin (2013)), resulting in BAM files as output format. This tool mapped the reads to the mouse reference genome mm10 while using an annotation file in ".gtf" format. For further analyses was a file necessary, which contains the total read counts per gene, for being able to consider the expressed genes. A direct specification of the parameters of this tool create this file while mapping (parameter: "–quantMode GeneCounts"). Note: starting the program with the specified parameters to handle mapping, sorting and counting the reads per gene at once worked well for us with threads <= 16, otherwise it could result in a memory overflow or stop due to the system-preset limit of contemporaneous opened files.

2.2.2 Peak calling

After processing the data we used the "bamCoverage" function from Deeptools (Ramírez *et al.* (2016)) to create the corresponding bigWig-Files for every ATAC-Seq and ChIP-Seq BAM-File using the included parameter for the RPKM-normalisation. We used a bin-size of 25 base pairs (bp), ignored the gonosomes and extended the reads to 300 bp. For the ATAC-Seq data we additionally ignored the blacklisted regions. The bigWig-Files were used to perform a visualization of the coverage within IGV.

We applied the integrated function "callpeak" from MACS2 (Zhang *et al.* (2008)) on each BAM-file to call the peaks within a q-value of 0.005. We used the merged control-ChiP BAM-files of both biological replicates as control data in order to reduce the bias induced by noise occurring in sequencing procedures. MACS2 merges those replicates automatically, as far as they are given on the input, leading us to a rejection of the general suggested option to merge them with SAMtools in a previous step.

For the ATAC-Seq files we simply called MACS2 with the following parameters: "macs2 callpeak -t <input file> -n BASENAME". In this case, as there is no control data, MACS2 builds a model based on Poisson distribution to reduce the number of false positive peaks.

2.2.3 Correlation between histone marks

To get an overview of the relationships between our files, we computed correlation matrices. We made a matrix for each of the tissues containing the BAM-Files for each of the 6 histone modifications and the ATAC-Seq BAM-Files. For this task we used two functions from Deeptools. At first we calculated the correlation using the "multiBamSummary" function and then we plotted the matrices with the "plotCorrelation" function. We used them with the following parameters:
"multiBamSummary bins -b <input files> -o <out files.npz> -bl <blacklist file> plotCorrelation –corData <multiBamSummary matrix> -c pearson -p heatmap –skipZeros -T <matrix label> -o <out file> –colorMap summer"

2.2.4 Characterization of H3K27ac peaks

To survey common histone modification H3K27ac peaks between two different tissue types, the chosen form of visualization was a Venn-diagram. To create those diagrams, we needed the total number of peaks per tissue type and the amount of common peaks between two tissue types. To receive the total number of peaks for every tissue, we calculated the intersect of the corresponding peaks of both biological replicates of each tissue type in ".narrowPeak"-format with BEDTools' (AR and IM (2010)) "intersect" subroutine and stored the number of lines of the output file in a variable, hence every peak is represented in its own line in this file format. The pairwise common peaks between two different tissue types got extracted in a quite similar way, further, the intersect of both considered peak files of the different cell types, which were intersected in advance as described above, lead to the amount of common peaks. To finally create the Venn-diagrams, those numbers got passed to an R script for plotting. The same procedure was done for one tissue at two different development stages (E14,5d and P0) to conceive a rough temporal course.

To be able to compare the amount and coverage of H3K27ac peaks through different tissues we used deeptools to produce heatmaps. In a first step, we took the union of the H3K27ac peaks over all tissues in the following way: We treated the biological replicates of each tissue at both development states (E14,5d and P0) separately, by taking their intersection. We concatenated these intersections into one file, sorted it, and finally merged it making use of bedtools. Each tissue file was based on merged BigWig files of their biological replicates, to preserve the complete data covering one tissue at both times (E14,5d and P0) separately. The resulted respective BigWig-files are used as input-files for deeptools' "computeMatrix" subroutine, together with the union-file in ".bed"-format. We used the option "reference point" and the flag "referencePoint center" to stretch a window with the size of 2000 bp up- and downstream from the peak-center. Further, this matrix is used as input to build a clustered heatmap using k-means algorithm with 4 centers by deeptools' subroutine "plotHeatmap". In a second approach of visualizing the coverage of the H3K27ac histone modification peaks, we created a heatmap for one tissue of its H3K4me1, H3K4me3, ATAC-seq and H3K27me3 ChIP-Seq data, centered at H3K27ac peaks. This was done using the same routine from the previous heatmap, but with different BAM-files on the input and deactivated clustering. The required BED file remained the same.

2.3 Multivariate analyses with ChromHMM

The next step in our project was an multivariate analysis using a CHROMHMM (Ernst and Kellis (2012)). The ChromHMM is a Hidden Markov Model for the discovery and characterisation of chromatin states. We had to process our data once more before this model is usable. The ChromHMM needs a binarized version of our data on the input. The software provides a function for this task. So we binarized out data using

the "BinarizeBam" function from ChromHMM. We applied this function with the following parameters:
"java -mx4000M -jar ChromHMM.jar BinarizeBam -c <controlChip files> <CHROMSIZES-file> <histone ChiP-Seq files> cellmarkfile.txt <output directory>"
The mentioned "cellmarkfile.txt" is a file we had to create first. It is a simple table which lists, for every tissue, the to be investigated mark and the names of both input files (controlChiP-Seq and histone Chip-Seq respectively for every cell type).

After we successfully binarized our data, we were ready to train the Hidden Markov Model. The model got trained with different numbers of states (7, 10 and 15). We used the "LearnModel" function of ChromHMM with the following parameters for this:
"java -Xmx8G -jar ChromHMM.jar LearnModel -p 32 -noautoopen <binarized input files> <output directory> <number of states> mm10 logFile".
After the successful training of the model, we assigned each state number, shown in the emission matrix, to a functional name to transform it in a more incisive declaration, and proceeded with the segmentation tracks produced by ChromHMMs' "LearnModel" function and visualised them within IGV.

2.4 Differential gene expression analysis

Our next goal was to find differently expressed genes in pairwise comparisons between the three tissues we are working with. We performed three pairwise comparisons between two different tissue samples taken from an mouse embryo 14,5 days post conception, one comparison between the samples of one tissue taken from 14,5 days post conception and from a fully developed embryo. Since we want to look at the influence of histone modifications on the expression of development genes we needed to filter the list of expressed genes of the tissues. We only want to look at protein coding genes on the chromosomes 1-19. To perform these comparisons we first of all needed a list of all expressed protein coding genes in a mouse so that we can filter the list of expressed genes of each tissue, because we only wanted to look at the protein coding genes We used a R library called biomaRt (S *et al.* (2009)) for this task. We used the useMart() function to get a list of all expressed genes in a mouse. `useMart(biomart='ENSEMBL_MART_ENSEMBL', dataset='mmusculus_gene_ensembl')`

After we downloaded this list we had to filter for the protein coding genes. We used the getBM() function with the following parameters:

```
prot.gene <-
   getBM(attributes=c("ensembl_transcript_id",
   "ensembl_gene_id","gene_biotype","strand",
   "chromosome_name","start_position","end_position"),
   mart = mart object, filters ="biotype",
   values = "protein_coding")
```

With this list of all protein coding genes we were able to filter the ReadsPerGene file, created by STAR, in order to obtain the list of all expressed protein coding genes for each mouse tissue. We used a python script here, but the R command

```
prot.gene <-
prot.gene[prot.gene$chromosome_name %in% 1:19, ]
```

would have been suitable as well. Now, having a list of all expressed protein coding genes of a mouse, we were ready to perform the differential expression analysis of these genes, using the R library DESeq2 (MI *et al.* (2014)) from the Bioconductor package. After we loaded an annotation file corresponding to our data, we created a matrix for each pairwise comparison, consisting of the readsPerGene data for the two tissues. With this matrix and the corresponding annotation file we created a DESeq2 object with the following command:

```
DESeqDataSetFromMatrix(countData = dataMatrix,
 colData = annotation file, design = ~ condition)
```

Then we only kept those genes in the DESeq2 object that had at least 10 genes mapped to it. Subsequently we calculated the deferentially expressed genes with the "DESeq()"-function and extracted the log2FoldChange and the adjusted p-values, calculated through a Wald test using the "result()"-function from DESeq2. We performed the "lfcShrink()"-function on the DESeq2 object and plotted these shrinked results in a MA-Plot using the "plotMA()"-function for a better visualisation. We also created a vulcano plot showing the –log10 of the padj-value vs. log2FoldChange using the ggplot() function from the ggplot2 library. To finish the differential gen expression analysis we created heatmaps showing the normalized expression values of differentially expressed genes for the individual replicates. So we normalised the data using the "normTransform()"-function and the variance stabilizing transformation function "vst()". We picked the 100 genes with the largest log2FoldChange score and plotted these in heatmaps using the "pheatmap()" function from the pheatmap library for each pairwise comparison.

2.5 Multivariate statistical prediction model

As the last part of our research, we wanted to create a multivariate statistical model to predict the probability for an expression of a gene at a specified development state. This linear regression model (Y = X * beta) is based on all histone ChIP-Seq data per tissue type from E14,5d of the promoter regions of protein coding genes, written in an R script. First, the protein coding genes needed to be retrieved from the BioMart database using the "getBM()" function from the R-package "biomaRt" (S *et al.* (2009)). This data got filtered for retaining the information of chromosomes 1-19, all autosomal chromosomes, and handed them over to a GRanges object (data structured provided by "GenomicRanges" R-package (M *et al.* (2013))), which allowed us to store the promoter information of the chromosomes easily with an upstream of 2000bp and downstream of 500bp. For each tissue, the respective histone modification files in ".bam"-format got narrowed down to the raw read counts per gene, which fell into the examined promoter regions, through the application of a function called "bamCount()" (R-package: "bamsignals" (A and J (2009))). The same procedure was done for the control ChIP-Seq data. For further processing, the tissue wide histone modification counts had to be normalized with their belonging control ChIP-Seq data, scaled and centered at 1, with a pseudo count of 1. In explicit terms, we applied this formula

$$Snorm = log\left(\frac{S+1}{C+1} * \frac{1}{median(\frac{S+1}{C+1})}\right)$$

on every histone modification per tissue and scaled the counts with this call: "`Snorm_scaled <- scale(Snorm<tissue><state><modification>, center = 1, scale = TRUE)`".
Each tissue got a separate matrix assigned, containing the normalized data of every corresponding histone modifications. These are the feature matrices. For a more continuously prediction, the mapped RNA-Seq data, our dependant variable, needed a RPKM-normalization. This was done with the reads per gene count table loaded in R, where it got merged with the corresponding exon lengths. The exon lengths also were received from the BioMart database. On beforehand, some data preparation was necessary, to reach a state of being able to merge them with the ".tab"-file. This included an association of the gene IDs to the row names, deleting the not regarded columns with different counting values of the ".tab"-file and a transformation over an actual data frame for the exon length vectors. When all this was done, the RPKM values of the tissues

$$RPKM = \frac{countsPerGene}{\frac{exonLength}{1000} * \frac{\sum(countsPerGene)}{1000000}}$$

were calculated and stored as vector. To build, train and test a model, at least two different data sets were needed. We chose a proportion of 3:1 for creating training and test sets of each matrix and RPKM vector belonging to a tissue type. Through a random seed the data was chosen and assigned in a diverged way to the mentioned data sets. The linear prediction model was build and trained with this command, based on build- in R-functions:
"`fm <- lm( log(RPKM1) . ,data = trainSet)`"
To test the model, this function was applied:
"`testLM <- predict(fm, testSet)`"
In a further processing, the beta-coefficients were surveyed. For clarifying, to check if the models' predictions could be accurate enough, a R^2-Test and correlation check got implemented, to compare the estimated RPKM-vector with the true outcome values:

```
CorrelationTrain <- cor(log(trainSet[,"RPKM"]1), +
    fitted.values(fm))\newline
CorrelationTest <- cor(log(testSet[,"RPKM"]1), +
    predict(fm,testSet))\newline
RSquareTrain <- 1 - (sum((trainSet[,"RPKM"] - +
    fitted.values(fm))^2) / +
    sum((trainSet[,"RPKM"]) +
    - mean(trainSet[,"RPKM"])^2))
RSquareTest <- 1 - (sum((testSet[,"RPKM"] - +
    fitted.values(fm))^2) / +
    sum((testSet[,"RPKM"]) +
    - mean(testSet[,"RPKM"])^2))
```

Further, this was applied to compare the trained models with different Chip-Seq data, using other tissues' data sets.
This resulted in a visualization, respectively for the considered tissues, of the predicted gene expression values versus the gene expression of the real observed values using a smooth scatter plot, produced by the "smoothScatter()"-function with logarithmic RPKM-values+1.

3 Results

3.1 Data quality

The quality report summarized by MultiQC shows that 98 of the 99 samples have a sufficient mean quality score, as shown in figure 1. Only one ChIP-Seq experiment (H3K27ac in lung at P0, replicate 2) did not pass this specific criteria and "per base N content". The heatmap (figure 1) shows that, in average, most of the samples have passed the quality-checks with success. Only the RNA and ATAC-Seq experiments failed in section "per base sequence content".

Fig. 1: Heatmap showing results of different quality checks for each sample

3.2 Univariate analyses

3.2.1 Mapping

The mapping statistics shown in figure 2 point out that at least 80% of the reads of every sample were be mapped on the reference genome with success. We obtained the same mapping results for ChIP-SEQ data and control-ChIP data, as figured in appendix.

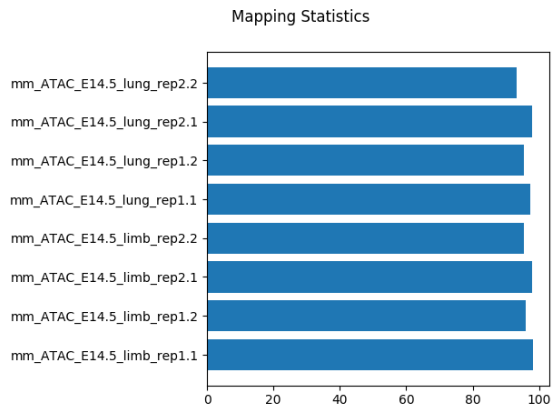

Fig. 2: Percentage of successfully mapped reads of ATAC-Seq data on reference genome with Bowtie2

3.2.2 Peak calling

The resulting files produced by MACS2 allowed us to count the number of significant peaks for the different histone modifications in each tissue-type at its development state. The different counts are shown in figure 32, attached to appendix. The amount of peaks is in general very similar between the different replicates. We found one exception concerning the number of significant peaks of H3K4me1 in replicate 2 of lung at P0 with a number of 129 against 82675 in replicate 1. This sample seemed to be corrupted by technical issues, caused by various sequencing procedures. This is also shown by a low signal value of this track shown in figure 3 but we could not find an origin of this in quality-check data.

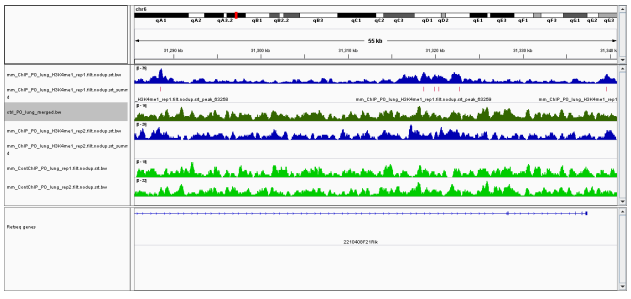

Fig. 3: IGV: Read coverage of lung in P0 for both replicates (blue tracks) with underlying peaks in red. The second replicate has a signal value of 26 against 36 for the first one. Resulting consequence of that is probably a loss of peaks. The dark green track in the middle is the merged control data. The bright green tracks are the control data for each replicate.

In addition we visualized through IGV the coverage of the reads by underlying the peaks and control data, as shown in figure 4, where we focus on a specifically in limb expressed gene called "Paqr3".

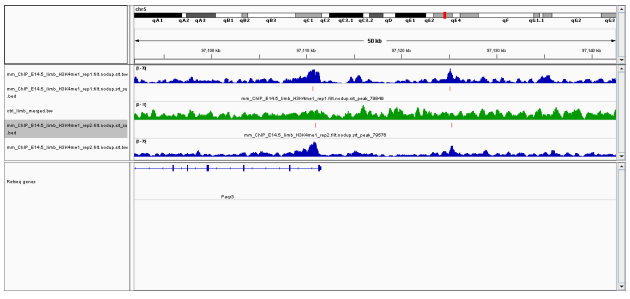

Fig. 4: Read coverage of both replicates from limb at E14.5 in blue on especially in limb expressed gene "Paqr3" . Merged control data in green.

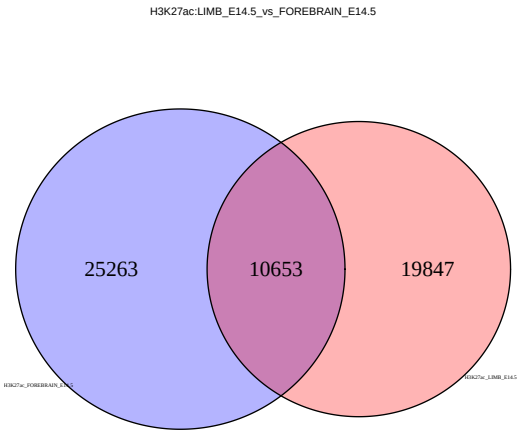

Fig. 6: ~~This figure shows the~~ total number of peaks for both tissues ~~distinctly~~ and the amount of common peaks in the middle.

3.2.3 Correlation between histone marks

As expected, the correlation matrices show a high correlation between the biological replicates of the same histone ChiP-Seq analysis. It also shows a relatively high correlation between the samples of the H3K27ac and H3K4me1 histone ChiP-Seq analysis, such as between H3K27ac and H3K4me3 samples. This isn't ~~really~~ surprising, because the functions of these histone modifications are familiar. H3K27ac modifications are associated with the higher activation of transcription, asociated with a mark for active enhancers. H3K4me1 modifications are enriched on active and on primed enhancers, so the correlation between these two enhancer associated modifications is obvious. H3K4me3 modifications are often involved in the modulation of genetic expression through chromatin remodelling and allows transcription factors to bind therefore it is an important factor in transcription initiation , which is the explanation for the high correlation between these two kinds of histone modifications.

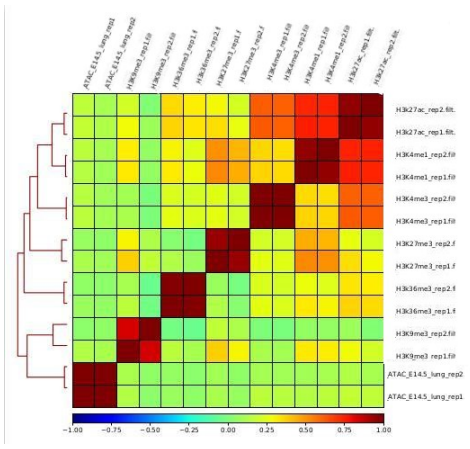

Fig. 5: Correlation Matrix of the Lung tissue ChiP-Seq data

3.2.4 Characterization of H3K27ac peaks

The heatmap shown in figure 7, produced by deeptools, points the coverage of reads in the region of H3K27ac-peaks in different tissue types at E14.5d and P0 ~~out~~. Due to the clustering with 4 centers, we are able to distinguish the profile of this marker from the different tissues. Indeed, there are much more similarities in coverage profiles of the same tissue types at different time stamps, than in different tissues at the same development state. So we could identify a cluster for lung and forebrain clearly, but less for limb. We observed the same trend by visualization through Venn-diagrams, which show the amount of peaks per tissue from ~~an~~ specific development state for two samples and the peaks they have in common as an overlapped area in the middle.

For the second analysis of the H3K27ac-peaks, pictured in a heatmap (figure 8), we unfortunately got black regions. There have been some difficulties in constructing an appropriate BED-File for this analysis. However, the biological meaning of this heatmap is not lost. It represents the coverage of different markers and ATAC-peaks in lung at regions belonging to H3K27ac-peaks. Regions of H3K27ac-peaks are associated with ATAC and H3K4me3-peaks. This finding accords to the logic that open chromatin, that is accessible for transcription, correlates with ATAC-peaks and with for enhancer and promoter specific markers, just like both above. The absence of H3K4me1 marker is in accord with this profile, because a signal in it would typically lead to suppression of transcription.

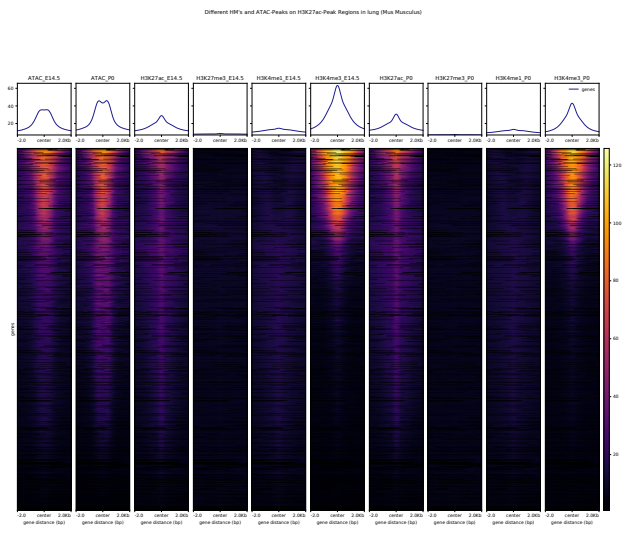

Fig. 8: Heatmap of read coverage of ATAC-SEQ, H3K27ac, H3K27me3, H3K4me1 and H3K4me3 in lung at regions of H3K27ac-peaks at E14.5 and P0.

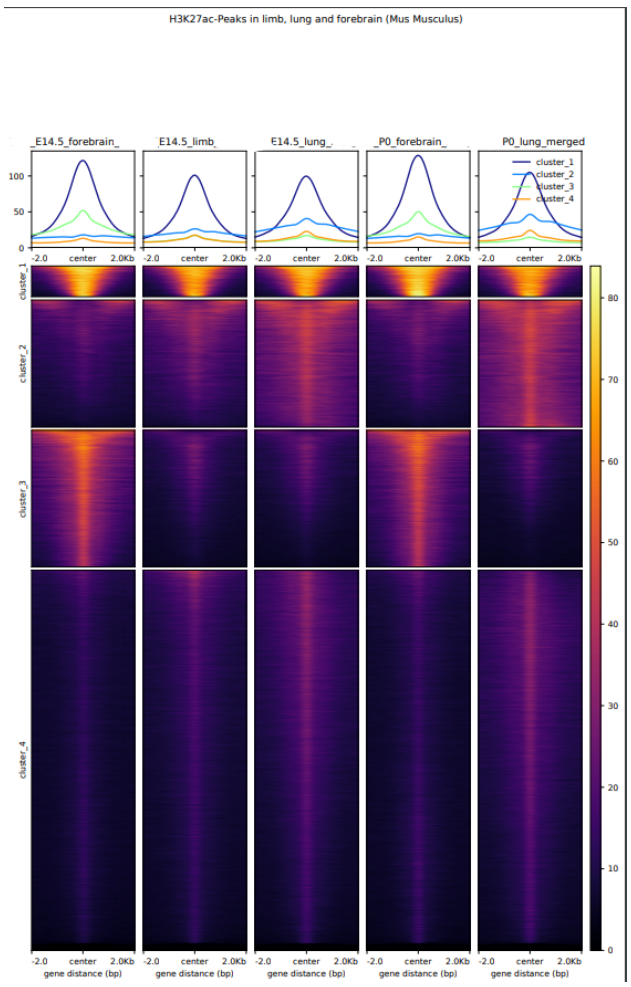

Fig. 7: Clustered heatmap of read coverage in common H3K27ac peaks of limb, lung and forebrain at E14.5 and P0 apart from limb in P0. We can observe a cluster for forebrain and lung, but less for limb.

3.3 Multivariate analyses with ChromHMM

The goal of of the multivariate analyses with ChromHMM was to predict current states of chromatin with the information of which histone modifications are present on the same gene. We learned the model with different number of states (7,10,15). The model consisting of 15 states was the most promising, since we got similar results like David U. Gorkin *et al.*, 2017 in their study on which our project is based. We were able to assign each state of the HMM to a functional chromatin state which is already known by the scientific society. The following figure 9 shows the emission matrix produced by the 15 state ChromHMM with the associated names of the chromatin states. We orientated our research on other chromatin state analysis to get this information, the one was the result of the study from David U. Gorkin *et al.*, 2017 and the other one was from Ernst and Kellis, 2018. We compared the patterns of occurring histone modifications per state with the patterns found in the mentioned studies. We checked if the potential states matched some specific peaks in our data to validate our chromatin states using IGV afterwards. For example: If a chromatin state signals the initiation of a transcription and a transcribed gene is located slightly downstream, then we took this annotation as validated.

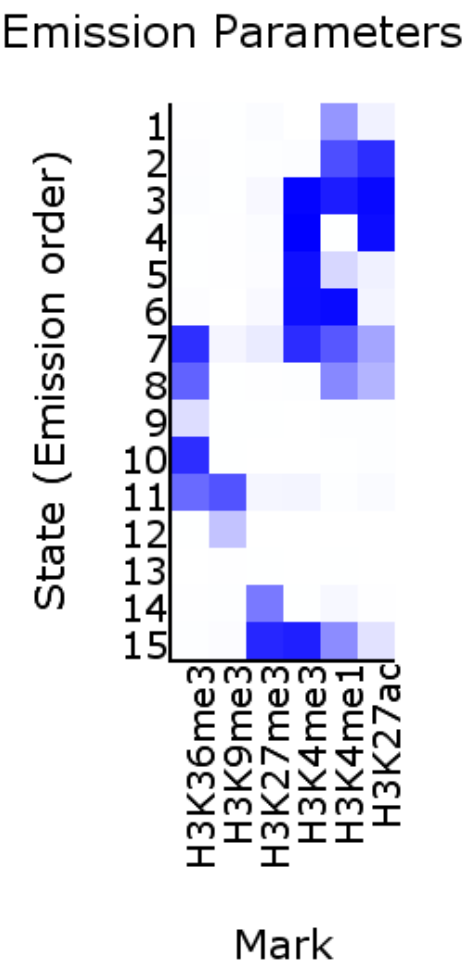

Fig. 9: This figure shows the emission matrix of the 15-state ChromHMM.states got following annotations:
1 = weak TSS dist. - Enhancer
2 = poised, TSS-dist. - Enhancer
3 = Strong, TSS-dist – Enhancer
4 = active TSS
5 = Poised, TSS-prox – Enhancer
6 = weak inactive – Promoter
7 = Strong, TSS-prox – Enhancer, Transk. for 5'-3'
8 = Initiation – Transkription
9 = Strong – Transkription
10 = H3K36me3 associated, but unknown functionality
11 = unknown functionality
12 = H3K36me3 associated - Heterochromatin
13 = no Signal
14 = Polycomb associated - Heterochromatin
15 = Bivalent - Promoter

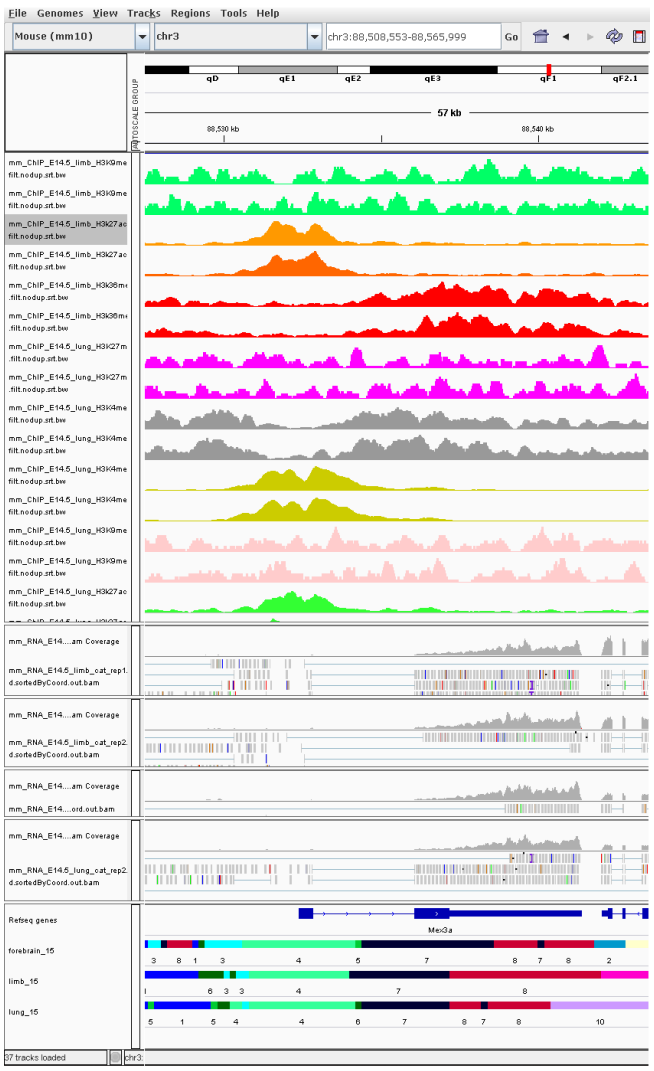

Fig. 10: IGV-visualization-~~snapshot~~ of histone modification peaks, transcribed RNA and segmentation tracks

This figure shows an excerpt of the IGV-viusalization of all ChIP-Seq peaks, respectively for the isogenic replicates, the RNA-coverage profiles and the segmentation tracks computed by ChromHMMs' 15-state model. Each histone modification of each tissue got a different colour, but remains the same for the according replicates. The numbering of the segmentation tracks is matching to Fig.9.

Differential expressed genes:

sample	overexpressed	underexpressed	total
forebrain vs limb	2383	2685	5068
forebrain vs lung	2302	3000	5302
limb vs lung	919	1427	2346
lungE14.5 vs lungP0	2809	2159	4968

Table 1. Table: Overview of differential expressed genes

The proportion of over and under expressed genes per comparison are in relative stable equilibrium. In total we have an amount of ca. 5000 differential expressed genes between the samples, but in limb vs lung only ca 2000.

So these results do not allow us to make any hypothesis about the relative expression between different tissues at the same development state compare to the relative expression between same tissue type at different time stamps. For this we would need visualization-techniques to show degree of differences in expression values.
Therefore, in a second phase we made different visualizations via MA-Plots, Volcano-Plots and heatmaps.

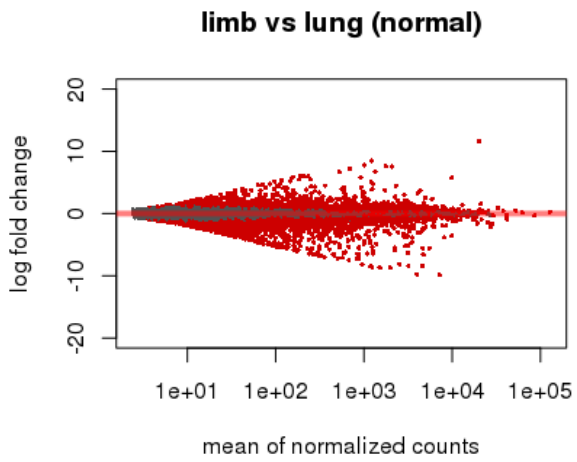

Fig. 11: Normalized MA-Plot comparing differential expressed genes in limb and lung at E14.5. MA-Plots compare two experiments by taking the average of normalized counts of their RNA-Seq expression values on the x-axis against their differences expressed in log2 fold-change on y-axis.

3.4 Differential gene analysis

As explained in the methods we used DESeq2 (MI *et al.* (2014)) to make an analysis of differential gene expression between protein coding genes of pairwise different tissues and also lung at E14.5 against P0. In a first step we filtered the differential expressed genes by a p-adjust value of 0.01 and an absolute value of log2-foldchange higher than 0.5, as we can see in table 1.

In figure 11 the density of the plot is lower than in the MA-Plots of the other experiments, shown in appendix. This observation is conform with what we could see in the expression table above. In addition we noted that the divergence of the plots is higher concerning experiments comparing pairwise different tissues than same tissues at different time states. For example the MA-Plot comparing lung in state E14.4 and P0 converge more on x-axis than the MA-Plot that compares forebrain and lung. So we find this trend not only in histone modification profiles (HMP), but also in RNA expression. This could be interpreted as an indication that HMP influences significantly RNA-expression, so transcription.

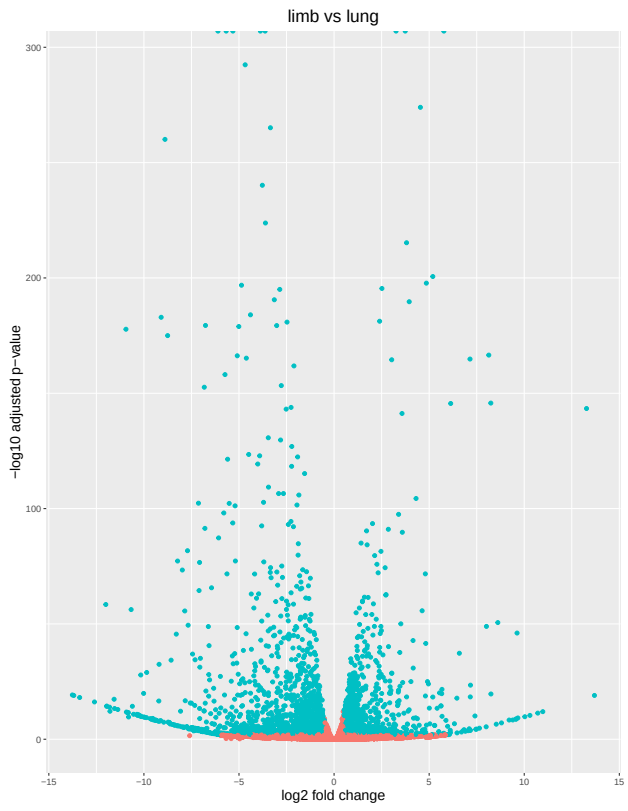

Fig. 12: Volcano plot of differential expressed genes in limb and lung at E14.5. This plot shows the logarithmic fold change of differential expressed genes on x-axis against their negative logarithmic adjusted p-value on y-axis. The blue dots are differential expressed genes, which passed the filter of an absolute logarithmic fold change value above 0.5 and an adjusted p-value of 0.01.

Volcano-Plots, as shown in figure 12, visualize the expressed genes filtered by log2-foldchange's absolute threshold value of 0.5 on x-axis, against an adjusted p-value of 0.01, transformed by a -log10 applied on the y-axis. So the blue dots shows the differential expressed genes. The red ones do not pass these filter parameters. The plots show through relative symmetric distributions the equilibrium of over- and under-expressed genes, conforming the above table. We could observe the same trend described in MA-plots, as the plot figured in 12 is not so dense on the extremities like in the plots comparing different tissue-types, which is validating our hypothesis.

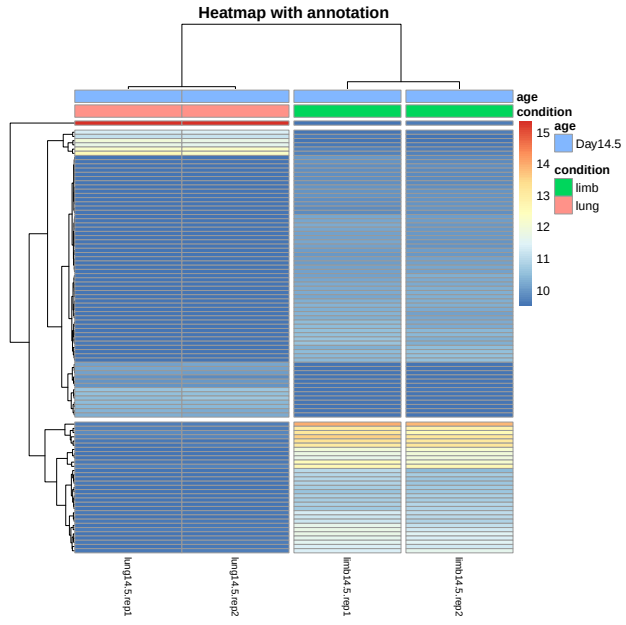

Fig. 13: This figure shows the heatmap containing the normalized expression values of differentially expressed genes for the 100 most expressed genes between the limb tissue and the lung tissue clustered in 4 clusters.

The heatmaps visualize differential expressed genes of two experiments by clustering them. So the differences of same degrees in gene expression are grouped together, to get an synoptic overview. As expected, the replicates are very similar, so they keep same colour on a same row. The heatmap of figure 13 shows less contrast than the heatmaps comparing different tissue types. So we determined this trend on all visualization-methods. The plots shown in appendix are comparing different tissue-types at same age. The most flagrant contrast appears more by comparing forebrain vs limb or forebrain vs lung than in lung vs limb. The differentiation of lung and limb in embryogenesis have probably a more common origin than in forebrain.

3.5 Linear prediction model

The last question to clarify is if we can predict the gene expression based on the epigenetic profile, thus the histone modification profile of a promoter. Therefore, we build a multivariate linear regression model to predict the expression of genes, just as described in Method. The examined results of the linear regression models of every tissue, were very similar to each other, so in the following we will set our focus on the results of the data obtained from limb tissues. The coefficients of the linear model have an intercept of 1.802. By observing the beta coefficients we can see that most of the examined histone modifications are indeed relevant for the gene expression. We got a the highest positive correlation between the H3K4me3 and the resulting gene products. H3K27ac has got an relatively high value too, which makes sense, because these both are very important in the activation of the transcription. Since H3K4me1 is an antagonist to H3K4me3 it makes sense that it got a low value when H3K4me3 has a high value. H3K27me3 and H3K9me3 are modification associated with the euchromatin, thus with the form of the chromatin which prohibits the expression of genes. We can see negative correlation between there occurrence and the amount of expressed genes, as expected. The H3K36me3 modifications got values below the significance level and seem to have no direct influence on the expression of developmental genes.

```
> coefficients(fmLimb)
(Intercept)     H3K27me3     H3K4me1     H3K4me3     H3K9me3     H3k27ac     H3k36me3
 1.80212084 -0.35844889  0.12644517  0.86499410 -0.34489949  0.41617806  0.01961764
> summary(fmLimb)

Call:
lm(formula = log(RPKM + 1) ~ ., data = trainLimb)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2782 -0.5521 -0.1922  0.4202  7.1680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.80212    0.03003  60.008  < 2e-16 ***
H3K27me3    -0.35845    0.01875 -19.112  < 2e-16 ***
H3K4me1      0.12645    0.02718   4.652 3.31e-06 ***
H3K4me3      0.86499    0.01863  46.442  < 2e-16 ***
H3K9me3     -0.34490    0.02518 -13.698  < 2e-16 ***
H3K27ac      0.41618    0.02000  20.811  < 2e-16 ***
H3K36me3     0.01962    0.01944   1.009    0.313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 15546 degrees of freedom
Multiple R-squared:  0.4549,    Adjusted R-squared:  0.4547
F-statistic:  2162 on 6 and 15546 DF,  p-value: < 2.2e-16

>
```

Fig. 14: R-Statistics of the Lin. model trained on Limb tissue data

The overall correlation is about 70%, which indicates the usefulness of the histone modification profiles for the prediction of gene expression. The selection of histone modifications seems to be extremely important to create a proper prediction model for the prediction of developmental gene expression profiles.

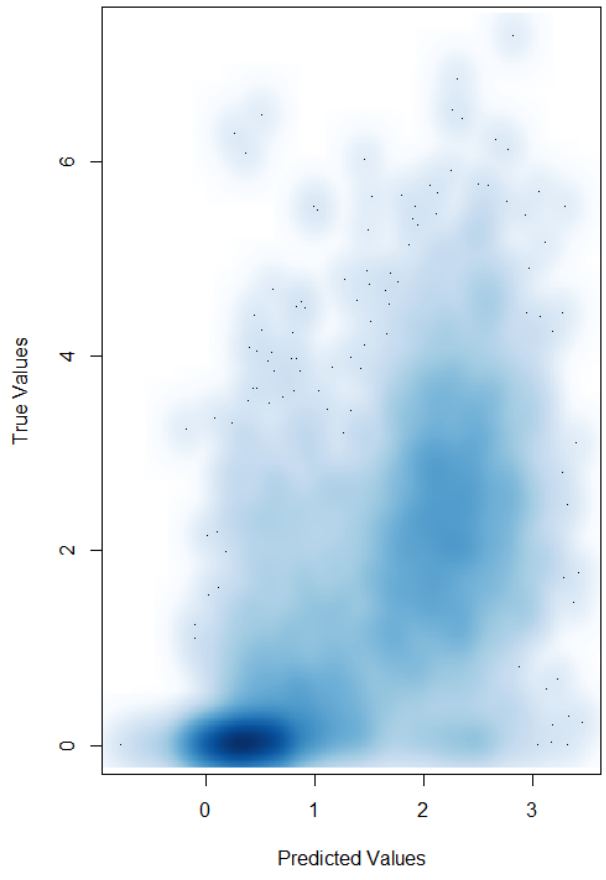

Fig. 15: Smooth scatter plot of the limb tissue data

Smooth scatter plot with the predicted values (the results of the linear model) versus the true RPKM-normalized values. These values are logarithmic for more clarity.

4 Discussion

First of all we can remark that the quality of the sequencing data was good enough to continue with downstream analysis. After mapping the reads to the mouse reference genome mm10, we determined an data set (tissue: lung, isogenic replicate 2) which had an overall bad mapping quality and the called peaks Visualizing the histone modification profiles (HMP) shows through clustering more similarity between same tissue type at different time stages, than between different tissue types at same age. As we observed this same trend in differential expressed genes, we can admit a causal relation between HMP and transcription. In addition our results of ChromHMM deliver similar chromatin states and emission probabilities as in passed studies from David U. Gorkin *et al.* (2017) and especially Ernst and Kellis (2018). This is a supplement indicator of the functional character of HMPs in transcription.

Which is also observable in the correlation matrix, where we have a high correlation between histone modifications with a similar biological function, just as the high correlation between H3K27ac and H3K4me1 which are both responsible for active enhancers.

The visualization of the histone modification signal tracks and the segmentation tracks provided enough information to verify the function of the different histone modification and the annotations of names to the chromatin state numbers.

As the regression model results in an overall correlation between the given and the dependant RPKM-values of around 70%, provided by the assumption that most of the 6 histone modifications on promoter regions of protein coding genes have an significant impact on gene expression. We reached similar high correlations by applying this prediction model on data of another tissue type, what suggests an tissue-type independent impact of HMP in promoter regions on gene expression. That is leading us to the conclusion, that our multivariate statistical regression model, based on the 6 histone modification data sets and actually examined corresponding transcribed RNA, produced a rough prediction for the expression of a genes. The case of an overfitted linear model and needs to be deliberated, due to our very high significance level of $p - value \leq 0,001$ for 4 of the 6 histone modifications and the aggregate R^2-value of 45,5%.To declare the accuracy and usability of this specific model some limitations should be resolved. This includes the relatively small spectrum of data obtained, the not significant histone modifications within this model. The major limitation is the missing comparison to different prediction model, because in this study, the data sets and linear models only got compared on different input files and we are missing a comparison of methodically different prediction models.

5 Conclusion

This study tries to evaluate if epigenetic factors as histone modifications impacts gene regulation in embryogenesis of mouse. The integrated analysis of ChIP-seq data, ATAC-Seq data and RNA-Seq data enabled us to clearly identify causal relation between them. Indeed, use of ChromHMM Software allowed us to build a functional segmentation model of chromatin states, that matched with transcripts-regions. In addition we build a prediction model that shows a considerable correlation of 70% on test-data. However, we probably need to consider other factors to improve the prediction model, cause in real cells there are a lot more known biochemical processes involved in gene regulation, as DNA-methylation, microRNA, random factors due to diffusion of transcription factors assumed by thermodynamic laws and so on.

Acknowledgements

We would like to thank Alena van Bömmel and Robert Schöpflin for supporting us through the whole project and providing further information, explanations, data and example R-scripts.

References

A, M. and J, H. (2009). Pysam: https://github.com/pysam-developers/pysam.

Andrews, S. *et al.* (2012). FastQC. Babraham Institute.

AR, Q. and IM, H. (2010). Bedtools: a flexible suite of utilities for comparing genomic features: https://github.com/arq5x/bedtools2/blob/master/README.md. Bioinformatics. 26, 6, pp. 841–842.

ATAC-Seq-pipeline (2016). ATAC-Seq pipeline v1 specifications:https://docs.google.com/document/d/1f0Cm4vRyDQDu0bMehHD7P7KOMxTOP-HiNoIvL1VcBt8/edit#heading=h.9ecc41kilcvq.

David U. Gorkin, Iros Barozzi, Y. Z. *et al.* (2017). Systematic mapping of chromatin state landscapes during mouse development. https://www.biorxiv.org/content/10.1101/166652v1.full.pdf+html.

Dobin, A., D. C. A. S. F. G. (2013). Spliced transcripts alignment to a reference. https://academic.oup.com/bioinformatics/article/29/1/15/272537.

Encode-portal (2020). Encyclopedia of DNA Elements – ENCODE: https://www.encodeproject.org/.

Ernst, J. and Kellis, M. (2012). ChromHMM: Chromatin state discovery and characterization: http://www.biolchem.ucla.edu/labs/ernst/ChromHMM/.

Ernst, J. and Kellis, M. (2018). Chromatin state discovery and genome annotation with ChromHMM. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945550/.

Jens Reeder, M. L. (2018). Picard toolkit. http://broadinstitute.github.io/picard/.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2.

Li, H. *et al.* (2020). bamsignals: Extract read count signals from bam files: https://bioconductor.org/packages/release/bioc/html/bamsignals.html.

Li, H. *et al.* (2019). Samtools: http://www.htslib.org/.

M, L. *et al.* (2013). Software for computing and annotating genomic ranges: https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html. PLoS Computational Biology, 9.

MI, L. *et al.* (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2: https://bioconductor.org/packages/3.11/bioc/html/DESeq2.html. Genome Biology, 15, 550.

Ramírez *et al.* (2016). deeptools2: a next generation web server for deep-sequencing data analysis: https://deeptools.readthedocs.io/en/develop/#. Nucleic Acids Research (2016): gkw257.

S, D. *et al.* (2009). "mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart: https://bioconductor.org/packages/release/bioc/html/biomaRt.html. Nature Protocols, 4, 1184–1191.

Zhang *et al.* (2008). Model-based analysis of chip-seq (macs). Genome Biol vol. 9.

Appendix

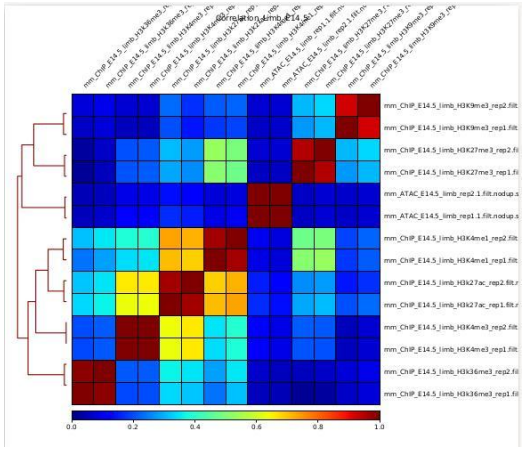

Fig. 16: Correlation Matrix of the Limb tissue ChiP-Seq data

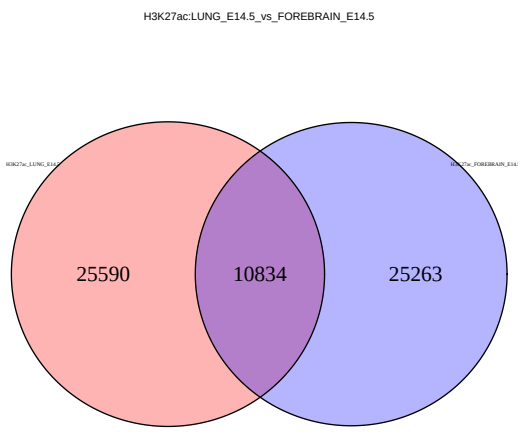

Fig. 17: This figure shows the total number of peaks for both tissues distinctly and the amount of common peaks in the middle.

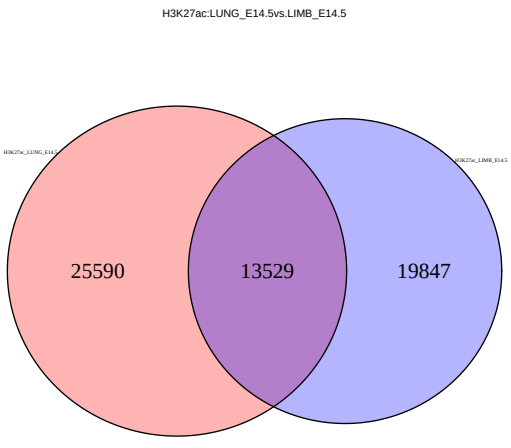

Fig. 18: ~~This figure shows the~~ total number of peaks for both tissues distinctly and the amount of common peaks in the middle.

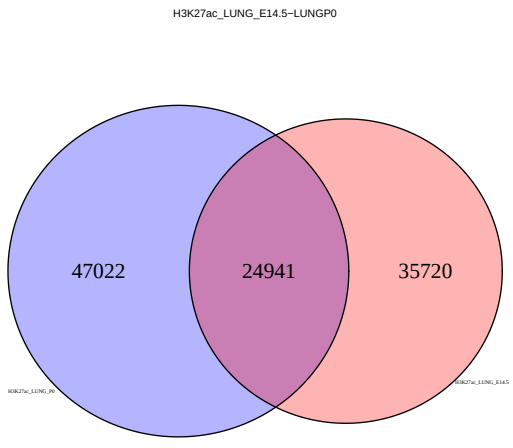

Fig. 20: This figure shows the total number of peaks for the lung tissue of development state E14.5d and P0 and the amount of common peaks in the middle.

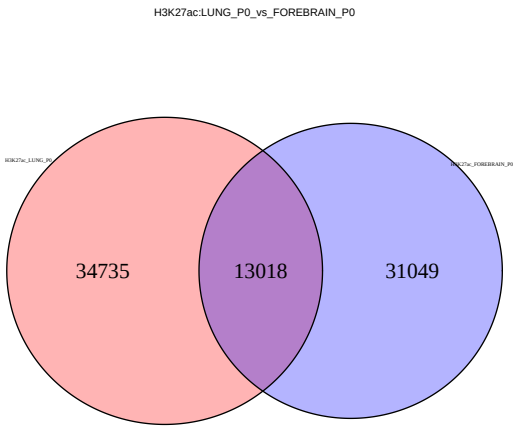

Fig. 19: This figure shows the total number of peaks for both tissues distinctly and the amount of common peaks in the middle.

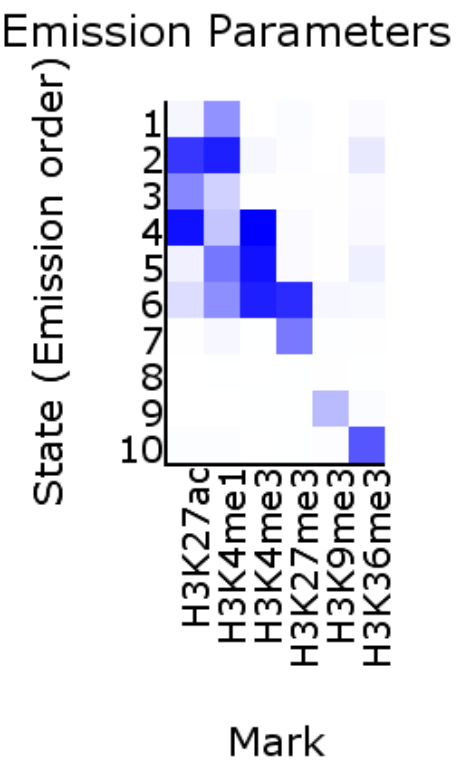

Fig. 21: This figure shows the emission matrix of the 10-state ChromHMM.

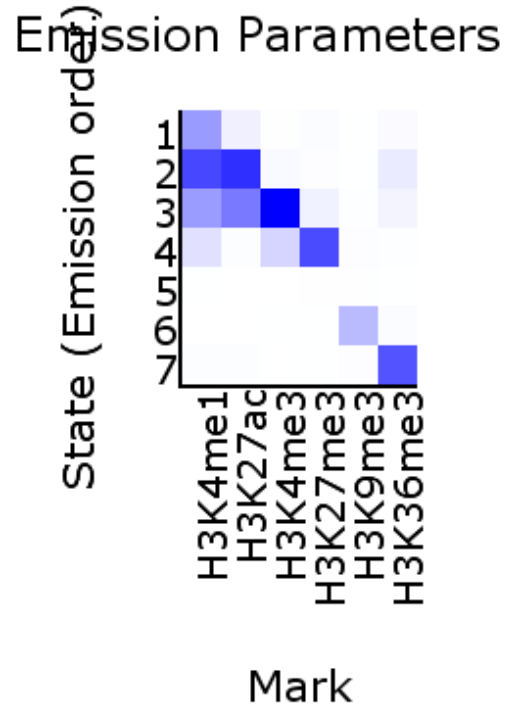

Fig. 22: This figure shows the emission matrix of the 7-state ChromHMM.

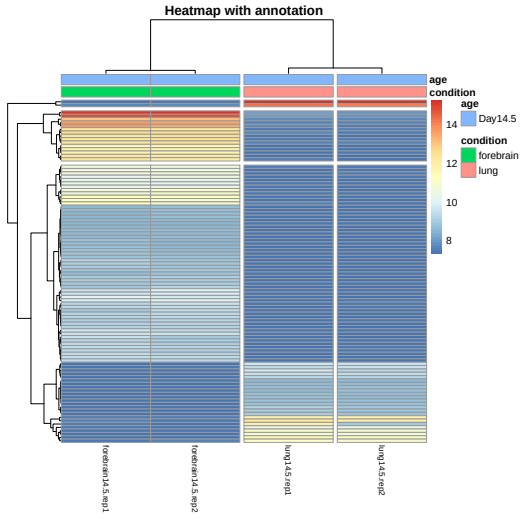

Fig. 24: This figure shows the heatmap containing the normalized expression values of differentially expressed genes for the 100 most expressed genes between the forebrain tissue and the lung tissue clustered in 4 clusters.

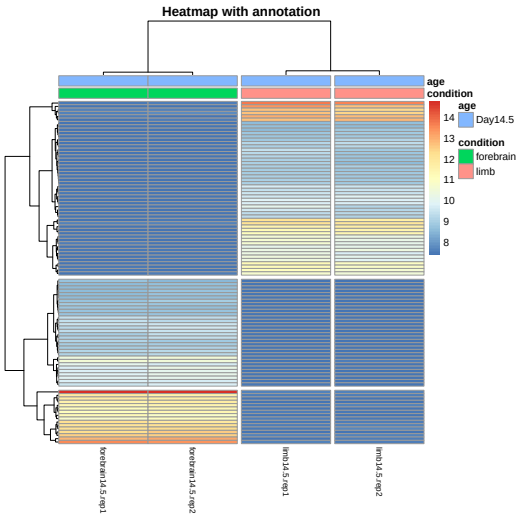

Fig. 23: This figure shows the heatmap containing the normalized expression values of differentially expressed genes for the 100 most expressed genes between the forebrain tissue and the limb tissue clustered in 4 clusters.

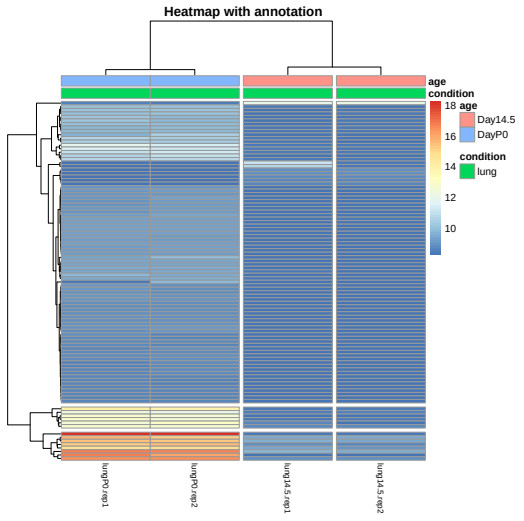

Fig. 25: This figure shows the heatmap containing the normalized expression values of deferentially expressed genes for the 100 most expressed genes between the lung tissue at time E14.5 and the lung at P0 tissue clustered in 4 clusters.

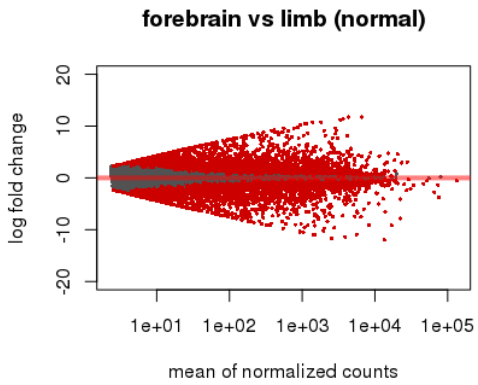

Fig. 26: Normalized MA-Plot comparing differential expressed genes in foerbrain and limb at E14.5

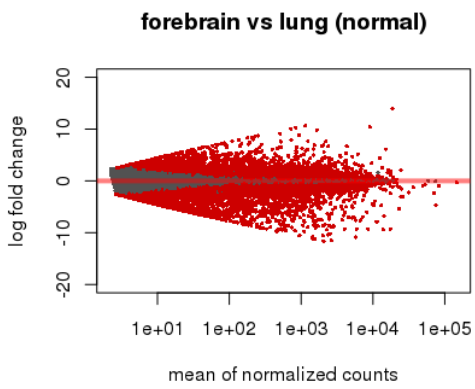

Fig. 27: Normalized MA-Plot comparing differential expressed genes in forebrain and lung at E14.5

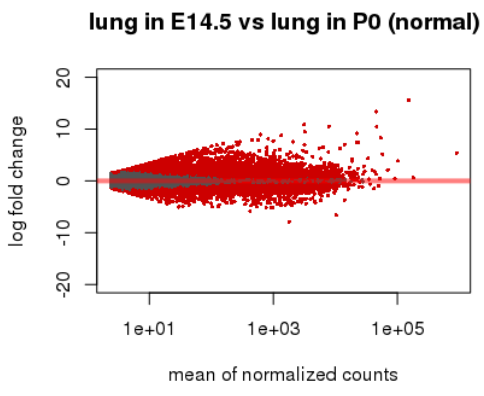

Fig. 28: Normalized MA-Plot comparing differential expressed genes in lung at E14.5 and P0

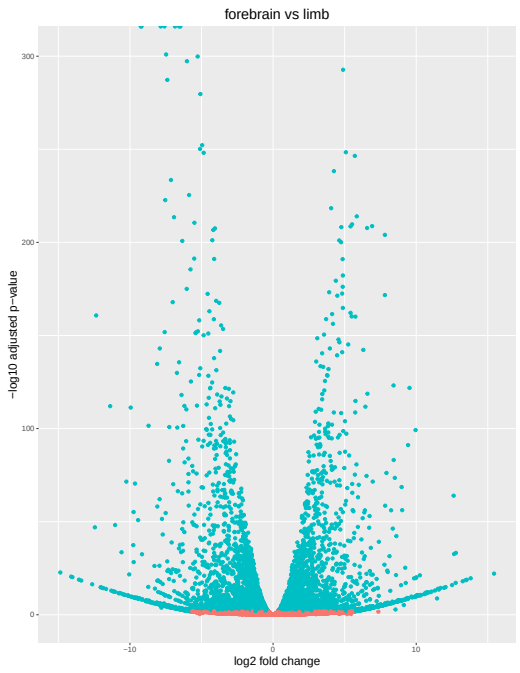

Fig. 29: Volcano plot of differential expressed genes in forebrain and limb at E14.5

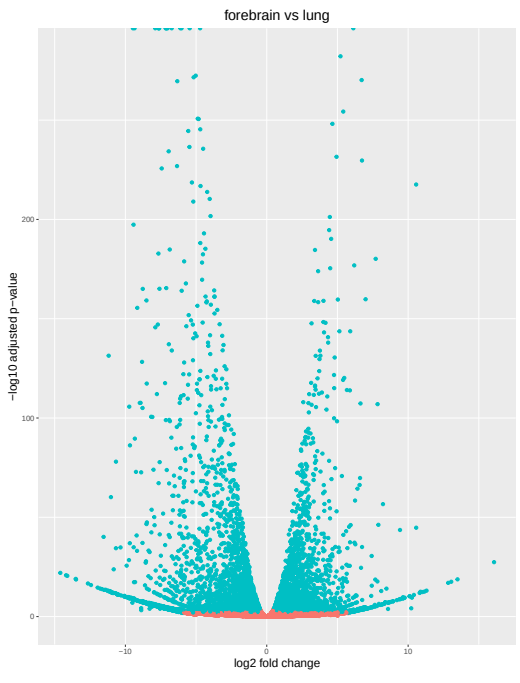

Fig. 30: Volcano plot of differential expressed genes in forebrain and lung at E14.5

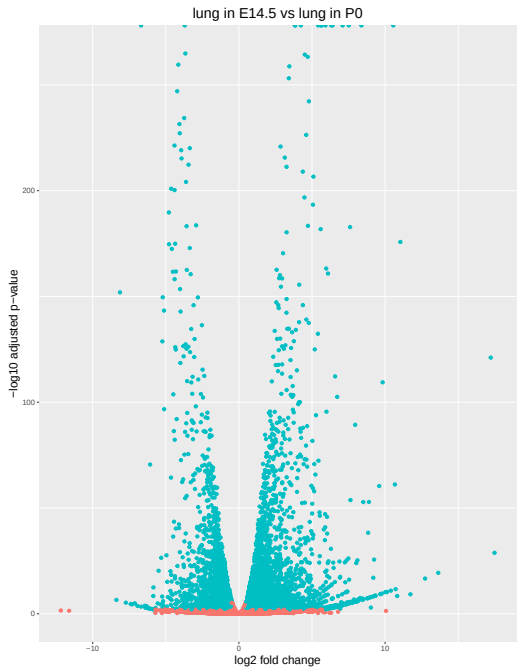

Fig. 31: Volcano plot of differential expressed genes in lung at E14.5 ans P0

```
> summary(fmLung)

Call:
lm(formula = log(RPKM + 1) ~ ., data = trainLung)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2402 -0.5410 -0.1616  0.4075  6.2340

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.61562    0.02856  56.566   <2e-16 ***
H3K27me3    -0.45060    0.01847 -24.401   <2e-16 ***
H3K4me1      0.02335    0.02823   0.827    0.408
H3K4me3      0.98223    0.01798  54.620   <2e-16 ***
H3K9me3     -0.41643    0.02537 -16.416   <2e-16 ***
H3k27ac      0.42748    0.02042  20.931   <2e-16 ***
H3k36me3     0.02990    0.02168   1.380    0.168
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9854 on 15548 degrees of freedom
Multiple R-squared:  0.4936,    Adjusted R-squared:  0.4934
F-statistic:  2526 on 6 and 15548 DF,  p-value: < 2.2e-16
```

Fig. 33: R^2 Statistik fpr the linear model trained on the lung tissue

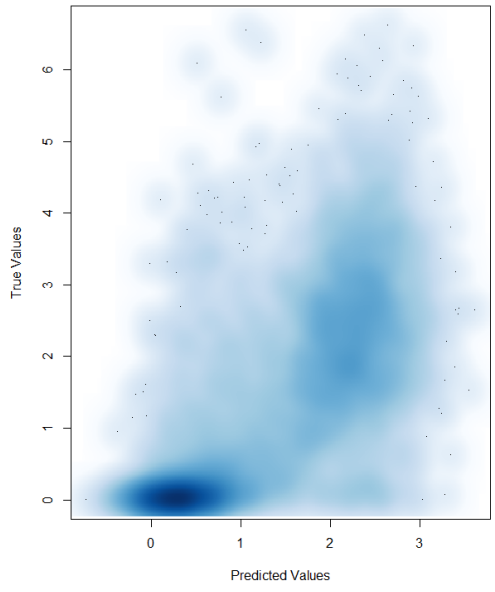

Fig. 34: Scatterplot for the lung tissue

```
Number of significiant peaks in mm_ChIP_E14.5_limb_H3K4me1_rep1:
109251
Number of significiant peaks in mm_ChIP_E14.5_limb_H3K4me1_rep2:
109383
Number of significiant peaks in mm_ChIP_E14.5_limb_H3K4me3_rep1:
23102
Number of significiant peaks in mm_ChIP_E14.5_limb_H3K4me3_rep2:
24119
Number of significiant peaks in mm_ChIP_E14.5_limb_H3k27ac_rep1:
31205
Number of significiant peaks in mm_ChIP_E14.5_limb_H3k27ac_rep2:
41014
Number of significiant peaks in mm_ChIP_E14.5_lung_H3K4me1_rep1:
36197
Number of significiant peaks in mm_ChIP_E14.5_lung_H3K4me1_rep2:
50472
Number of significiant peaks in mm_ChIP_E14.5_lung_H3K4me3_rep1:
22074
Number of significiant peaks in mm_ChIP_E14.5_lung_H3K4me3_rep2:
24260
Number of significiant peaks in mm_ChIP_E14.5_lung_H3k27ac_rep1:
37634
Number of significiant peaks in mm_ChIP_E14.5_lung_H3k27ac_rep2:
70856
Number of significiant peaks in mm_ChIP_P0_lung_H3K4me1_rep1:
82675
Number of significiant peaks in mm_ChIP_P0_lung_H3K4me1_rep2:
129
Number of significiant peaks in mm_ChIP_P0_lung_H3K4me3_rep1:
23143
Number of significiant peaks in mm_ChIP_P0_lung_H3K4me3_rep2:
21311
Number of significiant peaks in mm_ChIP_P0_lung_H3k27ac_rep1:
55216
Number of significiant peaks in mm_ChIP_P0_lung_H3k27ac_rep2:
58933
```

Fig. 32: Counts of significant peaks for each tissue

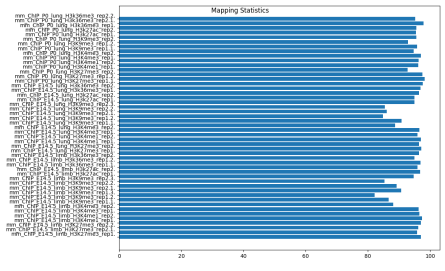

Fig. 35: Percentage of succesfully mapped reads on reference genome of mouse mm10 for each tissue type