DATA SCIENCE IN LIFE SCIENCE

Identify epigenetic alterations associated with Alzheimer's disease and classification of gene expressions between healthy and sick patients Project Week 03 - Group 02

Pushpa Koirala, Christina Kirschbaum and Melika Moradi

Full list of author information is available at the end of the article

1 Project

This week, we wanted to finish the Quality Control. Besides the ChIP-seq dataset H3K9ac from Carmen, we achieved this goal. Our second goal for the week was the Data Analysis. We managed to start the Data Analysis with DESeq2 for the RNA-seq and looking at the peaks of the BigWig files of the ChIP-seq. However, a more in-depth Data Analysis for both types of omics data will be needed in the next week.

This means we are still delayed, but we started to catch up to our initial timeline in the last week.

2 Quality Control RNA-seq

After finishing the FastQC for the RNA-seq last week, the next step was STAR for alignment. As described last week, problems arose related to STAR, which could not solved after trying several different approaches. In the end, STAR was replaced by kallisto. featureCounts could be applied without further problems.

2.1 Problems during Quality Control

Because of computational restraints, it was not possible to build the Genome Index with STAR. Although using a computer with 16GB RAM and applying the recommended parameters, and even scaling them down once more to save memory, the process was still killed while sorting the suffix array. We decided to do a pseudoalignment with kallisto instead to create the bam files. This was possible without further problems.

2.2 Results of Quality Control

According to the FastQC results, the sequence quality histograms, which describe the mean quality value across each base position in the read, and the per sequence quality scores were good. The per base sequence content as well as for some sequences the per sequence GC content were bad.

The results of the pseudoalignment with kallisto were poor, only about 50% of the reads could be aligned for every sequence. The featureCounts assignments were good with consideration of the kallisto results.

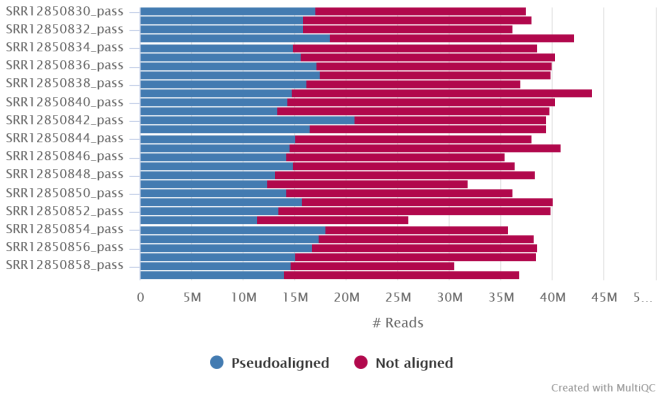


Figure 1: Results of pseudoalignment with kallisto

3 Quality Control ChIP-seq

The underlying fastq files could not be found anywhere, therefore we were all forced to convert them from bw files to fastq files, which was not an easy process. you can get access to the converted fastq files in our GitHub repository. There are a total of 30 ChIP-seq files in this group. We just cannot include all the pictures in this report. While checking ChIP-seq data, we uncovered a couple of unique things, therefore we've decided to explain some of them and summarize our findings.

Although it was challenging, we were able to convert them using the implemented pipeline shown below.

```
bigWigTowig GSM3752862_O-10A-H3K27ac.bw GSM3752862_O-10A-H3K27ac.wig
wig2bed -x  < GSM3752862_O-10A-H3K27ac.wig > GSM3752862_O-10A-H3K27ac.
    bed
bedtools getfasta -fi hg19.fa -bed GSM3752862_O-10A-H3K27ac.bed -fo
    GSM3752862_O-10A-H3K27ac.fa
perl fasta_to_fastq.pl GSM3752862_O-10A-H3K27ac.fa > GSM3752862_O-10A-
    H3K27ac.fq
```

3.1 Issues while checking the Quality

```
wig2bed < GSM3752845_Y-15T-H3K27ac.wig > GSM3752845_Y-15T-H3K27ac.bed
Error: WIG data contains 0-indexed element at line 8480708
       Consider adding --zero-indexed (-x) option to convert zero-
           indexed WIG data
```

Since wig2bed converts both variable - and fixed -step, 1-based, closed [start, end] UCSC Wiggle format (WIG) to sorted, 0-based, half-open [start-1, end) extended BED data. In the case where WIG data are sourced from bigWigToWig or other tools that generate 0-based, half-open [start-1, end) WIG, a –zero-indexed option is provided to generate coordinate output without any re-indexing. So we believe this might also be the reason of getting poor quality.

We added -x and it removes all the zero indexed files while converting them into wig2bed.

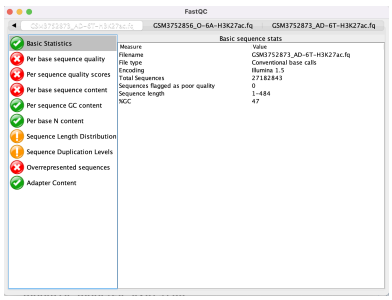
```
wig2bed -x < GSM3752845_Y-15T-H3K27ac.wig > GSM3752845_Y-15T-H3K27ac.
    bed
```

3.2 Results of the Quality control

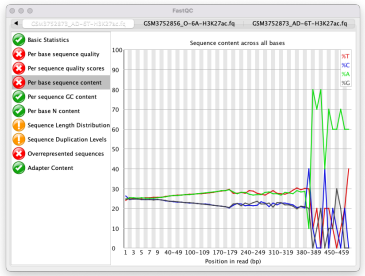
The quality of a bigwig file contains intervals and then indicates the coverage of that interval. That coverage again comes from the alignments covering that region. There is no direct connection between the interval and the fastq file, other than that the fastq file is the basis for the coverage calculation. Converting the interval to a fasta and then forcing it into a fastq file is technically possible, but has absolutely no practical meaning. We cannot recover sequencing reads from a bigwig, that information is lost once you convert your alignments (usually a BAM file) into a bigwig. Nonetheless, we have converted all those bw files to fastq files for the quality control check.

Despite the difficulties we got during conversion, we could control the quality of our fastq files with FastQC but as we already mentioned because of the conversion of the files from wig format to fastq, we lost some information. FastQC did not accept per base sequence quality, per sequence quality scores, Per base sequence content and overrepresented sequences but flagged no sequences as poor quality.

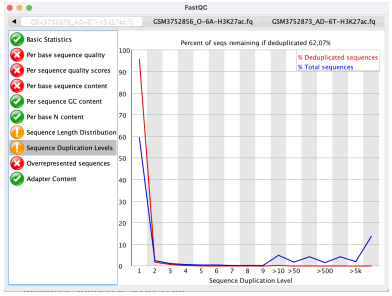
There are a few results shown in our report. As previously indicated, we assume that low quality results from the loss of data during conversion. We have not deleted any duplicates, because as long as it has a reasonable rate, removing duplicates is not a good idea. Since the goal of this component was to regulate its quality, we did not apply any function to delete their duplicates.



(a) Basic Statistics



(b) Per base sequence quality



(c) Duplication level

Figure 2: Insight into the results of Quality Control of the ChiIP-seq files

Finally, we run MultiQC on the report files that we get from FastQC (*_fastqc.zip) and MultiQC scanned all this files and make a report in regard to all the FastQC reports. The MultiQC report is availble in our Github repository.

4 Contributions

Member	Last week's work
Christina	Finished Quality Control for RNA, Start Data Analysis
Pushpa	Researched and implemented a pipeline to convert bw to fastq, Quality Control
Melika	Converted bw to fastq, ChIP-seq Quality Control in FastQC, Make final reports in MultiQC

5 Plans for the upcoming week

Carmen decided to quit the course this week, so we have to reorganize some things. She did not manage to do the steps for the ChIP-seq up to now, so Melika will make up for it, while Pushpa and Christina continue with the next steps.

- Process the ChIP-seq data H3K9ac
- More in-depth data analysis
- Machine Learning