RESEARCH

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

Pushpa Koirala, Carmen Calle Huerta, Christina Kirschbaum and Melika Moradi

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

1 Project

The goal for this week was to finish the Quality Control and start the Exploratory Data Analysis. Unfortunately, some problems arose and caused a delay. This delay results in us being behind our schedule.

More information about the progress and problems of this week can be found in the following two sections.

2 Quality Control ChIP-seq

2.1 Methods for ChIP-seq Quality Control

We actually wanted to conduct quality control in two ways. We simply wanted to see how the results varied when different tools were used. Using R's Bioconductor package and the Multiqc utility, we were able to learn more about it. Next, we'll have more outputs from the MultiQC tool and the R package, in addition to what we now have.

It is clear from the image ref:fig:compare1 that we can see a peak from every chromosome in our sample set, as is displayed in the histogram plot below.

First, we had prepared a short R code to know the chromosome's snipes numeric score.

```
Update all/some/none? [a/s/n]: GRanges object with 22733482 ranges and 1 metadata column:
             seqnames              ranges strand |     score
                <Rle>           <IRanges>  <Rle> | <numeric>
        [1]     chr1       10368-10468      * |   0.02429
        [2]     chr1       10469-10587      * |   0.13006
        [3]     chr1       10588-10688      * |   0.02429
        [4]     chr1       14864-15083      * |   0.02429
        [5]     chr1       15180-15399      * |   0.02429
        ...      ...               ...    ... .       ...
 [22733478]     chrY          59031995      * |   0.00992
 [22733479]     chrY 59032989-59032994      * |   0.01168
 [22733480]     chrY 59032995-59033033      * |   0.02429
 [22733481]     chrY 59033034-59033098      * |   0.13006
 [22733482]     chrY 59033099-59033104      * |   0.02429
 -------
 seqinfo: 74 sequences from an unspecified genome
```

Figure 1: Distribution of Peak denoting each chromosome snipes

Now we have assigned each peak to a chromosomal region by using the feature database. We have used the ChIPpeakAnno package. It has the function assignChromosomeRegion for calculating these overlaps by using an annotation database

From the plot below, we can see that the distribution of sizes using the hist() function. From here we can observe the frequency of gene peak size. According to the figure 2 we have, the peak do not seem to be normally distributed. We should now find a way to make them all normally distributed.

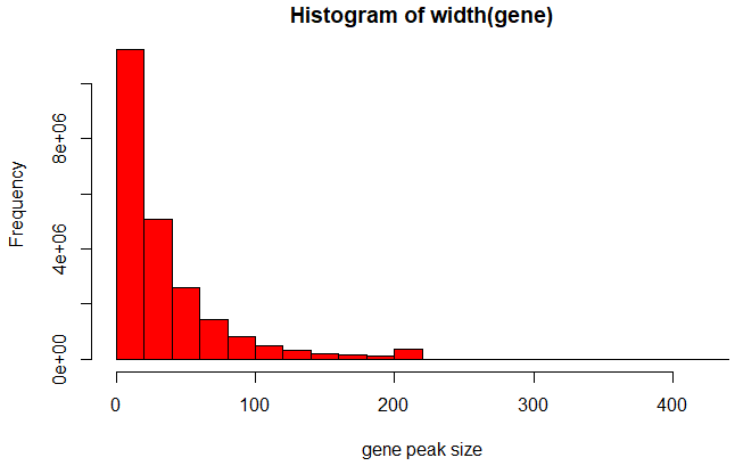

Figure 2: Distribution of Peak

2.2 Problems during ChIP-seq Quality Control

We tried this week to check the quality control of chip-seq. Because the chip-seq are in BED format, we have to convert them into for example bam file to read them in fastqc. The converting of the file was a little complex because we needed the reference genome to use the bedtools. We found The reference genome in NCBI (GRCh37_latest_genomic.fna.gz) but the problem was with missing Nucleotides. We removed them and made the bam file, but now they have bad quality in FastQC, and we do not know if they really have bad quality, or we did a something wrong by converting of files.

3 Quality Control RNA-seq

3.1 Methods for RNA-seq Quality Control

Firstly, all RNA sequences were ckecked with FastQC. The reports can be found in the repository. Afterwards, we wanted to process the sequences according to the main reference paper of Nativio et al. There, the reads were aligned to the human reference genome assembly GRCh37.75/hg19 using STAR with default parameters

and featureCounts was used to generate a matrix of mapped fragments per RefSeq annotated gene. Finally, all reports will be summarized with MultiQC.

Problems occurred with STAR. However, the code for the following steps was prepared while trying to fix it.

3.2 Problems during RNA-seq Quality Control

During the building of the Genome Index with STAR, many warnings arose. In a check with the most recent version of GRCh38, the warnings were absent. The problem was that the patches, which suggest how gaps in the genome sequence can be filled, are not integrated in the version GRCh37.75, but in the most recent version of GRCh38.

There were now two ways proposed to handle this problem:

1 Scientific approach: use GRCh38 and update the coordinates of the other omics data

2 Fast approach: remove the patches with the risk to distort the read counts because the reads that should align at the patches align somewhere else.

We decided on the fast approach, because of the timeframe of the project and because we are already delayed with the timeline.

4 Contributions

Member	Last week's work
Christina	FastQC for RNA-seq, STAR building Genome Index, prepare Code for STAR alignment and featureCounts
Pushpa	Quality Control Chip-seq
Melika	Quality Control ChIP-seq with FastQC
Carmen	FastQC for subset of ChIP-seq data (had difficulties opening the zip file and opening up FastQC), Research on pre-processing methods, Continued python tutorial

5 Plans for the upcoming week

- Finish STAR analysis and featureCounts for RNA-seq, summarize results with MultiQC
- Chip-seq needs to be explored
- Exploratory data analysis

Author details

References