Babraham Bioinformatics

Exercises: Analysing High Throughput Sequencing Data with SeqMonk

Version 0.2

Licence

Exercise 1: Creating a Project and Importing Data

- Create a new project based on the *Mus musculus* GRCm38 genome.
- Use the BAM/SAM import to read in the GSM307618.bam and GSM307619.bam files - extending each read by 250bp.
- Rename the GSM307618 DataSet to ES.H3K4me3 and GSM307619 to ES.H3K27me3
- Change the annotation to just show gene features and CpG islands.
- Try out the movement and visualisation controls in the chromosome view
    - Try zooming in and out of some sequence clusters to see how they look
    - Try using the different packing levels and strand splitting options
    - Go to Chr17 35437220 - 35577497
    - Find all homeobox (hox) genes and make these into an annotation track

Exercise 2: Quantitation

- Generate promoter probes (upstream of mRNA) from -1000bp - +1000bp excluding exact duplicates.
- Do a simple read count quantitation (without log transforming).
- View a histogram of the distribution of values in the K4 data.
- Requantitate the data using a log transformed read count.
- View the modified histogram to see if it's clearer.
- Draw a scatterplot of the two data sets to see how they compare.
- Draw a probe trend plot to see what pattern you observe relative to the transcription start site.

Exercise 3: Filtering

- Use the probe value histogram to pick a cut-off value which will filter out only promoters showing enrichment in H3K4me3 (those in the right half of the bimodal distribution).
- Apply a simple values filter to select just these enriched probes.
- Make this enriched list into a new annotation track and add it to the display.
- Filter from this enriched list promoters which do and don't overlap with CpG islands. Draw trend plots for both of these and see if there's a difference.

Exercise 4: Unbiased Quantitation

- Save your project.
- Use the MACS peak caller to find a set of enriched peaks using the default parameters, with no input store selected.
- Quantitate the probes using a log transformed read count.
- Look at some of the peaks which were found. Do you agree with all of the detection?
- Identify probes which are not within 2kb of one of your original enriched promoters. What proportion of all clusters do these represent? How many of them are situated within a known gene?

Exercise 5: Generating Reports

- Generate an annotated probe report linking your candidate peaks to their nearest gene.
- Order the report according to the number of reads and view the most promising candidates.
- Generate a probe group report which groups together probes within 500bp of each other.
- Order the report by the number of included probes and look at the most probe dense regions.
- Pick one of your top hits and export a view of the chromosome around the hit.