Galaxy

Jean-Charles Cadoret oct 2012

Mapping reads and peak calling

One example, CTCF in GIR cells

from the web site : https://main.g2.bx.psu.edu/u/james/p/exercise-chip-seq

Step 1 : upload your datas

step 2 : your Fastq file are in the correct format ? To be sure, perform Fast Q groomer in **NGS: QC and manipulation**. Now you are sure that your file is in the correct format for the next steps

step 3 : What is the quality of your reads ? Do FastQC. Now You can select where you can trimm reads in according with the quality sequence (it's an option)

Step 4 : If it's a pair-end read, befor the mapping you need to split your file with FastQ spliter. If not, do anything and you can perform the mapping.

Step 5 : map with bowtie (different options. You can or you need to try different criteria or you can leave the default mapping options). Click the link to "Display at UCSC main" and you will be able to see the reads on the genome (chrom 19).

step 6 : the format of the file after the mapping is SAM, but to see something in a genome browser, it is better to transform in BAM format. Do that with SAM to BAM tool.

Step 7: Once are reads are mapped, you can call peaks with MACS. Use the "NGS: Peak Calling > MACS" tool. You should also change the tag size to the read length you observed in Step 2. Otherwise the default values should be reasonable.

Step 8: Once MACS completes it will produce two datasets. One is a report on the peak calling process. The other contains the positions of the peaks. How many peaks were found? Click the link to "Display at UCSC main" and you will be able to see the positions of the peaks on the genome.

Calling peaks with a control sample (create a workflow)

Next, we will incorporate an input DNA control, import the following dataset into your history and you can create a workflow between the step 2 to 6 to process with exactly the same way.

Step 1: At the top of the History panel, click "Options" and select "Extract Workflow". Here you have the chance to select which jobs will be included in the workflow.

Step 2 : Load the MACS tool again. Select your previous CTCF dataset for ChIP-seq tag file, but now select the mapped input DNA for "ChIP-seq control file". How many peaks are called this time? What is the effect of using the input control? What's the difference (use operate on genomic intervals tool)

Motif research

Now we would like to search a consensus sequence :

Step 1 : **Extract Genomic DNA** using coordinates from assembled/unassembled genomes

step 2 : use MEME or Nebula in Curie institute web site

Next ..... Friday 9th november