Week 13, Lecture 25

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center

Penn State

Final Project

- Will account for 50% of your the grade.
- It will consist of one or more datasets and a number of hypotheses that need to be tested
- Project given out Thursday, Dec 6th and is due Tuesday, Dec 11th

Data and scripts for this lecture

- Data, code and scripts are packaged in lec25.tar.gz
- have to master too many tools and it is easy to get stuck
- Use it as a guide - don’t just copy/paste
- Develop your own mini libraries of tools, shell scripts and methods

Chip-Seq peak-caller Recap

1. A peak caller transforms aligned read coverages to intervals of enrichment

![Graph showing peak-caller output]
The most common misconception

Caution: this a personal opinion that others disagree with

Most confusion in peak calling arises from combining the peak calling with statistical testing

• A peaks means a region that appears to have an enrichment.
• The opposite of peak is a region with no data
• Some peaks may be caused by random agglomeration of data – but that those are still peaks – only that they are peaks occurring by random chance
• Most published peak callers will often remove peaks based on often non-obvious reasons –

Crazy peaks by MACS

Understanding peak calling

• We don’t need to use the entire read! Only the 5’ end matters.
• Transform your data into reads that are 1 bp long around the start sites!
• See the README.txt for step by step instructions

README.txt in the data
Effects of increase of smoothing

Isolate peak calling by strands
• Tools that merge strands always need to make assumptions on the data – sometimes it is not obvious what these are
• The best approach is to operate on strand individually, then combine the results
• Caveat: low coverage, error prone data will be even more difficult to analyze once split

Investigating two error models

Peaks with no exclusion zone around them
How to get the average fragment size?

1. Need to find peak pairs
2. Compute distance between their limits

Step by step:
• Expand each fragment to the left by a limit
• Intersect fragment on the opposite strand
• Compute distances

Homework 25

• Run the pipeline described in README.txt
• Report the average fragment size that it generates.