Rotation Report

Winter 2015

Kanishk Asthana kasthana@eng.ucsd.edu

Project Links: GitHub, Google Drive

Introduction:

HOMER is a motif discovery and Next-generation sequence analysis software package which was originally created in the Glass Lab. HOMER is used for multiple types of analysis in the lab including for calling peaks in Chip-Seq data. There are multiple programs currently available for calling Chip-Seq peaks, however it is not known how HOMER performs relative to these other programs.

The objective of this rotation was to compare HOMER to other popular Chip-Seq calling programs. Another objective was to assess whether the lab should continue using HOMER or switch to another program which performs better. The most comprehensive study done for comparing Chip-Seq programs was by Wilbanks et. al. and I have used many of the metrics introduced in that study to judge performance.

Chip-Seq Candidate Programs Chosen:

Given that there are so many Chip-seq calling programs available, the main limitation for choosing candidate programs became the availability (or lack of) of documentation for these program. Almost all other programs that I looked at lacked enough documentation to start using them. Given these limitations I chose the following two programs to compare with HOMER:

1. The most popular Chip-seq program so far is MACS. MACS was chosen for its popularity and wide adoption. I suspect it is this popular because it is very simple to use and requires only a single line command to generate an output. However, this comes at the cost of lesser flexibility and options as compared to HOMER. Moreover, MACS does not provide information about the strand the peak lies on +/-. This lack of information might hinder downstream analysis.
2. SISSRS is another candidate program that was chosen. I found it easy to install and use, however it does not provide information such as strand direction, a unique ID for each peak and a score for each peak found. This lack of information might hinder downstream analysis.

Pair wise comparison of shared peaks

For the pair wise comparison two datasets were initially chosen. The first was a PU.1 Chip-seq dataset generated in the Glass Lab (this dataset can be found at /data/home/kasthana/ mm10-C57BL7-ThioMac-PU1-notx.sam on the Glass Lab server). The second was an ENCODE data set (replicate no 2 was chosen). The R script used to make this comparison can be found here.

Table 1: Comparison of HOMER vs MACS.

Data Set	HOMER (% of Peaks shared with MACS)	MACS (% of Peaks shared with HOMER)
PU.1	93.24609	90.06348
ENCODE	74.55422	89.26311

Table 2: Comparison of HOMER vs SISSRS

Data Set	HOMER (% of Peaks shared with SISSRS)	SISSRS (% of Peaks shared with HOMER)
PU.1	82.0557	75.01613
ENCODE	80.47605	63.70693

Table 3: Comparison of SISSRS vs MACS

Data Set	SISSRS (% of Peaks shared with MACS)	MACS (% of Peaks shared with SISSRS)
PU.1	82.8864	87.56901
ENCODE	63.42505	95.92693

{Additional Comparisons made for HOMER vs MACS: for GAPB (Replicate2) and NRSF (Replicate 2)}

Looking at the shared peaks information it is clear that the number of shared peaks can be quite variable across programs and datasets. However, is this variability simply because of larger number of peaks called by one program with respect to the other? What about highly ranked peaks? If two programs are finding the same number or percentage of high confidence peaks then it might be that the larger number of peaks we see are simply because of more false positives.

So I decided to test the idea that the larger subset of peaks called by either program might be because of more garbage or false positive peaks found by one program with respect to another. If one program is detecting more garbage peaks then the percentage of shared peaks should increase for both programs for the top 25 percentile of peaks, and for top 5000 peaks with respect to the average. (Both top 25 percentile and top 5000 were arbitrary choices)

SISSRS does not provide a rank or score for its peaks so the analysis from this point onwards was done only for MACs and HOMER.

The following scripts were written to filter number of peaks for HOMER and MACS:

Returns peaks above x percentile: macsPeakFilter.java, homerPeakFilter.java

Return top x number of peaks: macsPeakFilterNumeric.java, homerPeakFilterNumeric.java

Returns peaks below x percentile: macsPeakFilterBelow.java, homerPeakFilterBelow.java

Example: ***java homerPeakFilter inputHomerPeakFile outPutFile 75***

Table 4 Comparison of HOMER vs MACS for top 25% and top 5000 peaks

Data Set	Type of Peaks	HOMER (% of Peaks shared with MACS)	MACS (% of Peaks shared with HOMER)
PU.1	**Average**	93.24609	90.06348
	Top 25 Percentile	90.48544	85.42430
	Top 5000 Peaks	82.57576	80.66000
	Bottom 25 Percentile	57.85524	51.35473
ENCODE			

NOTES:

Another, potential limitation of MACS is that it does not give information about which strand +/- the peaks lie on in its output. HOMER's motif analysis functionality requires strand direction index for detecting motifs. Moreover, to check whether the Chip-Seq worked correctly the lab often does a motif analysis on the detected peaks using HOMER. If the known motif for a transcription factor being ChiP-ed occurs in the region around the detected peaks then it is seen as a indicator that the ChiP worked as planned. One way to get around the problem that MACS does not provide strand direction information is to artificially introduce a column for the strand direction "+" in the output for MACS

Explain choice of chip-seq programs and what made you narrow down to the ones you did narrow down to. Introduce the paper in the references. Compare results with the paper. Introduce the different datasets you used dude. This is interesting.

Analysis methods: Check overlap: top 500, top percentile, overall. Metric, choose the one with better representation, not clear at all.

Next show the true positives stuff and explain the analysis and hg18 to hg19 conversion that you did. Next explain the true positives curve. And how they compare. Macs might just be better. Make table.

Next show motifs analysis results for the intersection of the two data sets. That is a lot of writing to do my friend lets begin.