ChIPathlon: A competitive assessment for gene regulation tools.

Avi Knecht  
*University of Nebraska-Lincoln, avi@kurtknecht.com*

Adam Caprez  
*University of Nebraska-Lincoln, acaprez2@unl.edu*

Istvan Ladunga  
*University of Nebraska-Lincoln, sladunga@unl.edu*

Follow this and additional works at: [http://digitalcommons.unl.edu/ucareresearch](http://digitalcommons.unl.edu/ucareresearch)

Part of the [Bioinformatics Commons](http://digitalcommons.unl.edu/bioinformatics), [Cancer Biology Commons](http://digitalcommons.unl.edu/cancerbiology), [Other Computer Sciences Commons](http://digitalcommons.unl.edu/othercomputersciences), and the [Software Engineering Commons](http://digitalcommons.unl.edu/softwareengineering)

[http://digitalcommons.unl.edu/ucareresearch/76](http://digitalcommons.unl.edu/ucareresearch/76)

This Presentation is brought to you for free and open access by the UCARE: Undergraduate Creative Activities & Research Experiences at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in UCARE Research Products by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
ChIPathlon: a competitive assessment for gene regulation tools

Avi Knecht, Adam Caprez, Istvan Ladunga
Gene regulation: why do we care?

- When gene regulation of the cell cycle malfunctions, it frequently causes cancer.
- Adult, differentiated cells can be reprogrammed to induced pluripotent stem cells
  - Which can then be reprogrammed to heart muscle, skin, etc, to repair damaged tissue (to limited extent in clinical practice)
Mapping transcription factors & histone modifications to genome

- Genes, are regulated by transcription factors and proteins, that bind to specific sequences on the DNA.

- Transcription factors are mapped to the DNA by chromatin immunoprecipitation followed by next-generation sequencing.
Critical steps in data analysis

Peak calling

Using strand dependent bimodality in peak calling

Sharp binding

Broad binding

Challenges in mapping transcription factors to the genome

- Background correction is an open problem.
- DNA fragments can be much larger than the binding site.
- Sequencing read location does not follow any statistical distribution.
- Over 50 different methods are used for mapping, which produce different results.
Evaluate the performance of all transcription factor mapping (peak calling) methods.

To this end, we will develop a scalable and easy to use super computing pipeline to stage data, compare many different peak calling and differential binding site tools, and store all results into a single database.
MongoDB

- Works well with large data sets.
- Can handle incomplete data.
Pegasus

• Too much data for manual processing, need to create workflows.
• Built on condor, which is used by a variety of super computing centers.
Python

- Many bioinformatics packages already managed in python under bioconda.
- Has interfaces for both MongoDB & Pegasus.
YAML to Workflows I

Each individual job is defined in a plain text YAML file.

```yaml
zcat_awk_sort_peaks:
  inputs:
    - bed:
        type: file
    additional_inputs: null
  outputs:
    - bed:
        type: file
  command: zcat
  arguments:
    - "$inputs.0":
      changeable: false
      required: true
      has_value: false
    - "$outputs.0":
      changeable: false
      required: true
      has_value: false
  walltime: 2000
  memory: 2000
  cores: 1
```
Jobs are chained together by using module YAML files.
Users need to input a file selecting ENCODE experiment id’s for files to process, tools they want to use, and a path to a genome.

```yaml
runs:
- experiment: "ENCSR605MFS"
  align: bwa
  peak: spp
- experiment: "ENCSR605MFS"
  align: bowtie2
  peak: spp
- experiment: "ENCSR000ERE"
  align: bwa
  peak: spp
- experiment: "ENCSR000EGZ"
  align: bowtie2
  peak: macs2
genomes:
bwa:
  grch38p6: "/path/to/genome/base"
bowtie2:
  grch38p6: "/path/to/genome/base"
```
Conclusion

• The current pipeline handles all downloads, alignment of single or paired end reads, and peak calling.
• The modularity of the underlying architecture makes it very easy to add additional tools or processing steps without changing the workflow generation code.
• Workflows can be generated for any ENCODE experiment, making this a very versatile pipeline for comparing bioinformatics tools.
Workflow II

Remove Duplicates

samtools view -F 1804 -o 30 -b $input.bam > $filtered.bam

java -Xmx8G -jar $(PICARD) SortSam
INPUT=$filtered.bam OUTPUT=$sorted.bam SORT_ORDER=coordinate

java -Xmx8G -jar $(PICARD) MarkDuplicates
INPUT=$sorted.bam OUTPUT=$dupmark.bam METRICS_FILE=$dup.md
VALIDATION_STRINGENCY=LENIENT ASSUME_SORTED=true ...

bedtools bamtobed -i $input.bam $output.bed

Peak Calling

Rscript run_spp.nodups.R -c $experiment.nodup.bam
-l $control.nodup.bam -npeak=300000 -odir=$outputdir
-speak=110 -savm=$spp.narrowPeak -savp=$spp.pdf -out $spp.cscore -fdr=0.01

macs2 callpeak -t $experiment.nodup.bam -c $control.nodup.bam -f BED -n $macs2.prefix -g hs
-p 1e-2 --nomodel --shift 0 --extsize 110 -B --SPMR

zcat $spp.narrowPeak.gz | awk 'BEGIN{OFS="\t"}print $1,int($2),int($3),$4,$5,$6,$7,$8,$9,$10;' > $sorted.narrowPeaks

sort -k 8r,8r $macs2.prefix_peaks.narrowPeak | awk 'BEGIN{OFS="\t"}Sr="PEAK_NR" ; print $0}' > $sorted.narrowPeaks

Save To Database
Regulation of cRPGs (outer circle) and mRPGs (circle next from the outside).

Yellow circles: RPGs
Green diamonds: regulators
Icon size is proportional to the number of regulatory relationships.

Note the extreme density of regulatory relationships.