A Comparison of Peak Callers Used for DNase-Seq Data

Hashem Koohy, Thomas Down, Mikhail Spivakov and Tim Hubbard
Spivakov's and Fraser's Lab

September 16, 2014

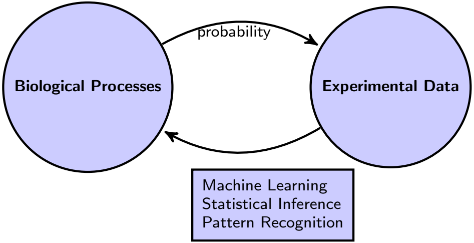
1. Introduction

2. DNase-Seq Data

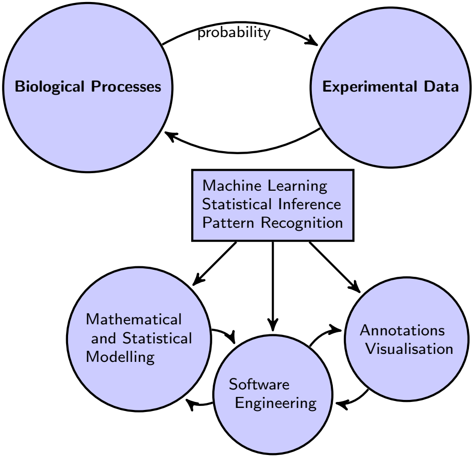
3. Results

4. Conclusion

Biology Becomes the most Data Intensive Science!



Biology Becomes the most Data Intensive Science!



ChIP-Seq Data Analysis

Sequencing, Mapping and Quality Controls

Sequencing is getting cheaper, providing us with more data! Mapping possibly is still the most computationally expensive part.

Peak Calling

Gauging the statistical significance of reads' enrichment which is generally known as "Peak Calling" is very central to ChIP-Seq data analysis.

Post Peak Calling Analysis

Different directions and purposes, including differential binding analysis, motif discovery, detection of regulatory regions, Genome segmentation and so on $\cdots$

Why Too Many Peak Callers?

Different protein classes have distinct mode of interactions:

Point-Source

These factors and chromatin marks are localised specifically and have high signal-to-noise ration

Broad-Source

These factors are associated with wide genomic domains, generating broad but more noisy signals; e.g. H3K9me3, H3K36me3

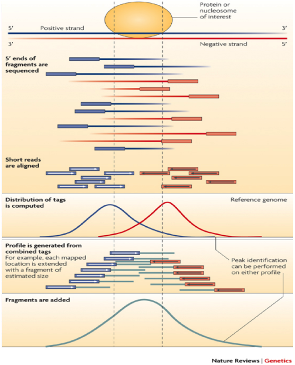
Mixed-Source

These factors show a point-source style signal at some regions whereas more broader in other regions e.g. RNA Pol II

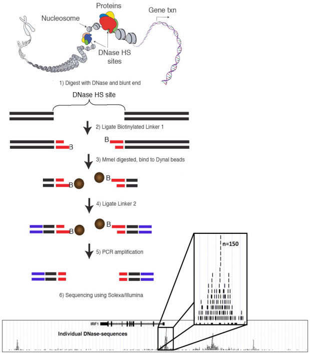
ChIP-Seq vs DNase-Seq

Note that DNase HS is different from its sister DNase Footprinting

ChIP-Seq: Nature Reviews, Peter J. Park, 2009


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added
Nature Reviews | Genetics

DNase HS: Duke Protocol


Proteins
Gene txn
Nucleosome
DNase HS sites
1) Digest with DNase and blunt end
DNase HS site
2) Ligate Biotinylated Linker 1
B
B
3) MmeI digested, bind to Dynal beads
B
B
4) Ligate Linker 2
B
B
5) PCR amplification
n=150
6) Sequencing using Solexa/Illumina
IRF1
Individual DNase-sequences

TF ChIP-Seq vs DNase-Seq

Some key differences between TF ChIP-Seq and DNase-Seq:

- In ChIP-Seq data, a protein is usually in "bound" or "unbound" position, whereas DNaseI shows a more generic behaviour, representing the openness of the chromatin to any regulatory feature;
- DNase HS are strand-independent and therefore no need to shift size or tag extension;
- DNase HS data sometimes shows less enrichment over wider regions (a kind of Mixed-Source).

Currently Existing DNase-Seq Protocols

Double Hit Protocol

Developed in John Stam Lab in University of Washington, and has been used greatly for detection of DHS in ENCODE project.

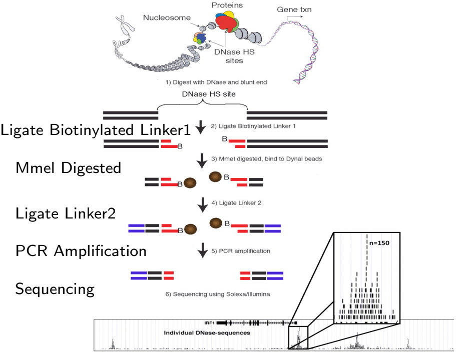
End Capture Protocol

Developed in Greg Crawford Lab in Duke University. It has been used for detection DHS in ENCODE. This protocol is also in great use by some other researchers world-wide.

ATAC-Seq

Developed in Greenland Lab in Stanford University. This is a very new protocol (published 2013) and has been reported to be very fast and very efficient.

“End Capture” (Duke) vs “Double Hit” (UW) Protocol



Ligate Biotinylated Linker1

MmeI Digested

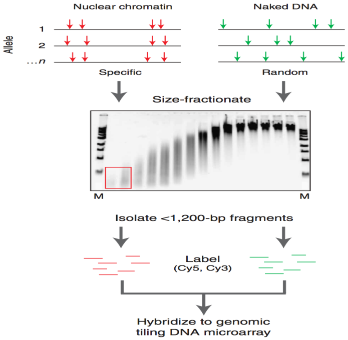
Ligate Linker2

PCR Amplification

Sequencing

Proteins
Gene txn
Nucleosome
DNase HS sites
1) Digest with DNase and blunt end
DNase HS site
2) Ligate Biotinylated Linker 1
B
3) MmeI digested, bind to Dynal beads
B
4) Ligate Linker 2
B
5) PCR amplification
6) Sequencing using Solexa/Illumina
n=150
IRF1
Individual DNase-sequences

End Capture Protocol: Greg Crawford Lab, Duke



Double Hit Protocol: John Stam Lab, UW

Study Design

- We Sought to assess four peak callers used for DNase-Seq data: Hotspot, F-Seq, MACS and ZINBA;
- The comparison was repeated on three human cell lines: GM12878, K562 and HelaS3, only on chr22;
- Raw data was obtained from ENCODE repository (from both Duke and UW protocols)
- Comparison was made in range of signal threshold (statistical significants of signals);
- All the remaining parameters kept as default (although we individually tried to assess them)
- The overlap level of detected peaks with TF binding sites was defined as the measure of comparison;
- The same process was repeated with Duke dat too.

DNase-Seq Peak Callers

Hotspot

The peak caller which is behind the ENCODE DHS.

F-Seq

F-Seq, Initially developed with DNaseI-Seq data in mind, but it has been used for TF ChIP-Seq data too.

MACS

Initially for TF ChIP-Seq, but has shown great performance for DNase-Seq data. This is the most used and cited peak caller.

ZINBA

Meant to be a generic peak caller for TF and Chromatin ChIP-Seq, DNase-Seq, RNA-Seq, FAIRE-Seq.

Hotsopt

- Tries to locally gauge the enrichment of tags by centring each tag in a small (250pb) and a large (50kb) window;
- The ratio of number of tags are assigned to each position;
- These scores are standardised (converted to $Z-$scores) by assuming a binomial distribution;
- Regions with $Z-$scores above the "threshold" is reported;
- This process is applied in two phases, the highly enriched regions are filtered and a second phase is applied to recover the regions which are overshadowed by monster peaks in phase one.
- FDR: some random tags are generated(uniformly distributed), then the ration of number of random tags to real tags for a specific $Z-$ score is reported as the FDR, for the given $Z-$score.

Hotspot Cont.

- The core of Hotspot has been implemented in C^{++} and its statistical analysis in R;
- It is wrapped up in python and bash script;
- It is relatively fast;
- I found it not well documented and not easy to work!

F-Seq

- An histogram-based (number of tags per bin) approach is, possibly, the most naivest for gauging the enrichment of short read tags;
- However, it suffers from some problems including boundary effects and selection of bin width;
- To overcome, F-Seq suggested in which a Kernel Density Estimator(with mean 0 and variance 1) is applied to obtain the distribution of reads:

$$p(x) = \frac{1}{nb} \sum_{i=1}^{i=n} K(\frac{x - x_i}{b})$$

- F-Seq has been implemented in Java, easy to use, though, doesn't support some commonly used file formats.

MACS

- The most used peak callers for ChIP-Seq data;
- It has been reviewed and benchmarked in different studies;
- At the time of development, the emphasis was on handling shift size and local biases from sequencability and mappability;
- A Poisson model is employed for identification of statistically signicant enriched regions;
- MACS has been implemented in python and is relatively fast. It is user friendly and fairly well-supported.

Zero Inflated Negative Binomial Approach: ZINBA

- ZINBA is a generic peak caller, and meant to be used for TF ChIP-Seq, histone ChIP-Seq, RNA-Seq and DNase-Seq(Both DNase and FAIRE);
- The short read tags are summarised into counts over non-overlapping windows (250pb) of the genome;
- Read counts per bin, G/C contents, mappablility scores and copy number variations are the parameters of its underlying mixture regression model;
- Based on this model, each region in the genome is assigned into one of the enriched, background and zero groups;
- ZINBA has been implemented in *R*

Two More Peak Callers

Two more peak callers for DNase-Seq are out now:

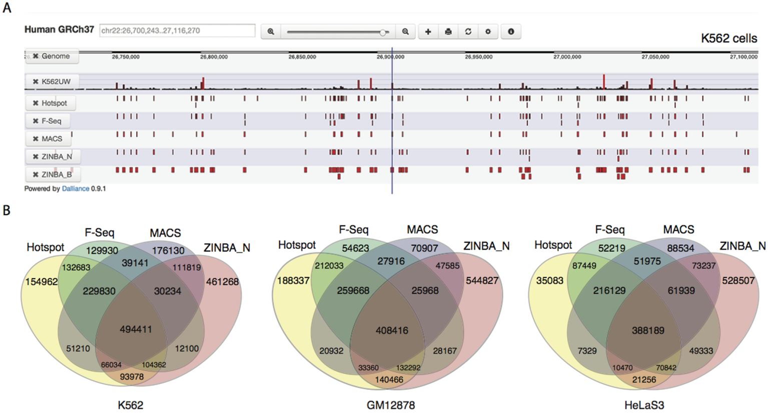
PeaKDEck

The idea behind PeakDEck is a kind of a combination of Hotspot (where they try to learn the local background) and F-Seq where they apply a Gaussian kernel to estimate the probability distribution! but surely has been more work!

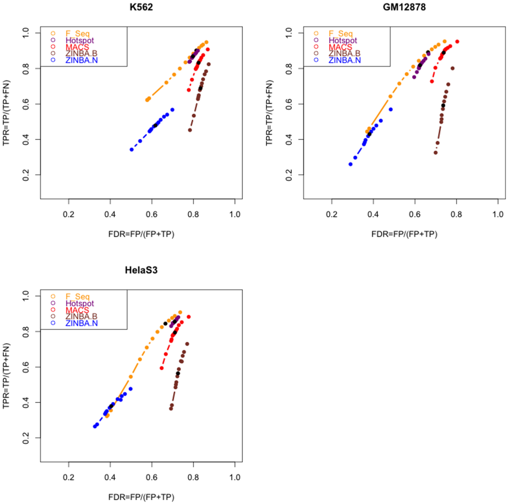
Dnase2hotspots

Dnase2hotspot is actually a modification of Hotpost; A key difference is that two phases of detecting hotspots in "Hotspot" is combined. It has also been claimed to be faster, more efficient!

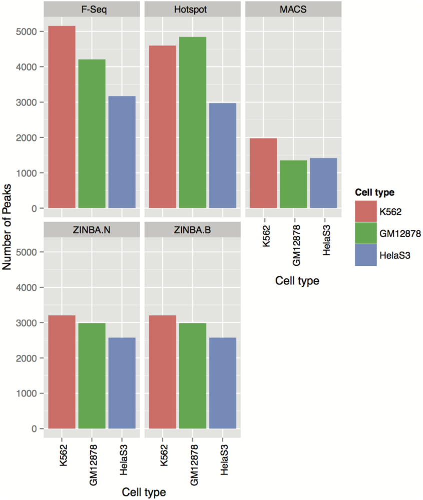
A Visual Inspection Shows Some Inconsistency



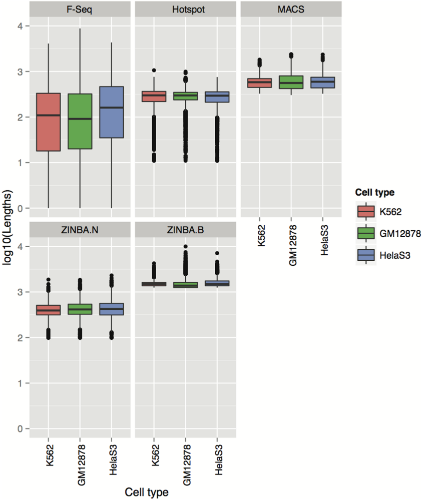
Sensitivity vs Specificity Shows up to 10% Difference



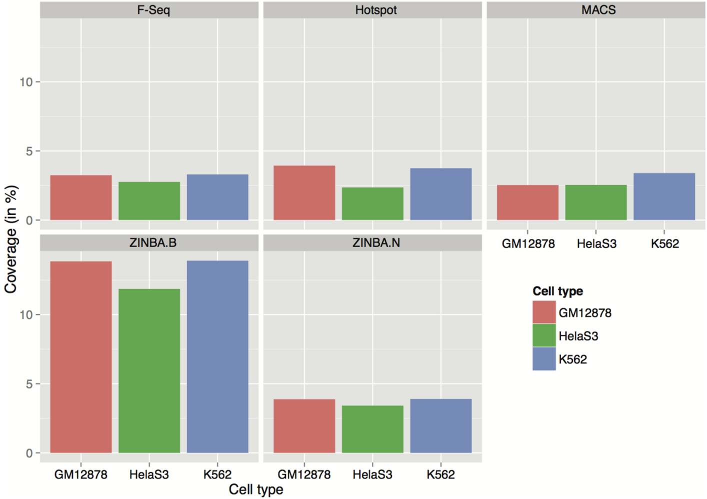
Number of Peaks Detected



Distribution of Peaks' Length



Chromosome-wide Coverage



F_β−Score: A Metric to measure the Performance of a Test

- F_β−Score is a commonly used measure for gauging the performance of a test;
- It is normally consistent with AUC;
- F_β−Score is defined as:

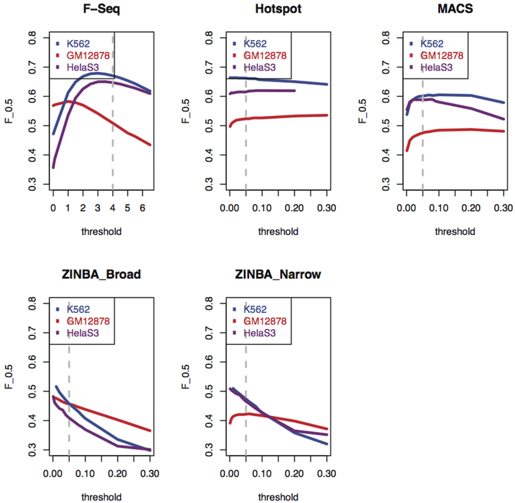
$$F_\beta = (1 + \beta^2).\frac{prec.recall}{(\beta^2.prec) + recall}$$

- Normally $\beta = 1$ but you can change it, depending on emphasising recall or precision(2 and 0.5) are very common.

Gold Standard Set

- It is generally accepted that open chromatin regions (DHS in ENCODE data) are accessible regions of the genome to TFs;
- Therefore it makes sense to compare the DNase peaks with TF Binding Sites;
- The problem is, though, set of TFBSs are incomplete;
- For each of the three cell lines in our study, there were more ChIP-Seq data of more 50 TFs;
- The union of the binding sites of these TFBSs were used as our "Reference Set";
- We set $\beta = 0.5$ to compensate for the incompleteness of our "Reference Set".

Improving the Performance by Adjusting the Parameters



Conclusion

- DNase-Seq is gaining popularity as a genome-wide chromatin accessibility analysis method;
- Its applications have led to new insights into genome function and variation;
- Robust peak detection on these data is therefore instrumental to the research community;
- They should be publicly available, well-documented and user-friendly softwares that can be easily used in any lab.

Acknowledgments

I am grateful to Spivakov's group members for their comments. This study was carried on during my transition from the Sanger Institute to the Babraham Institute. I therefore appreciate financial support from both institutes.