Comparison of Histone Protein Locating Algorithms

Kate D. Fischl, Princeton University

Abstract— **Epigenetics encompasses the study of changes in gene expression caused by factors outside of the DNA sequence, and helps to strengthen understanding of disease and cell malfunctions. One such epigenetic endeavor involves using chromatin immunoprecipitation (ChIP) along with large-scale DNA sequencing (ChIP-Seq) to locate specific histone protein modifications whose existence is linked to gene expression. This project focuses on understanding the distinct ChIP-Seq software packages, all of which use a different form of peak detection to locate histone modifications. Because the algorithms vary, this project marks the first step towards developing an ideal algorithm for use in a collaborative cross-disciplinary effort here at the University of Maryland.**

Index Terms—**ChIP-Seq, Epigenetics, Histone Protein Modifications, MACS, SICER**

I. Introduction

Although knowing the gene sequence of a specific organism can be extremely helpful in understanding the gene expression within that organism's body, it is not always the key to complete understanding. Epigenetics studies changes in gene expression caused by factors outside of the DNA sequence and can help scientists better understand the reasons behind a specific cell malfunction or disease contraction [2, 12, 15]. Two widely studied epigenetic factors are DNA methylation and histone protein modifications as shown in fig. 1 [10]. This study focuses on locating specific histone protein modifications along a sequence of chicken DNA.

Chromatin immunoprecipitation (ChIP) along with large-scale DNA sequencing (ChIP-Seq) is a relatively new technology that locates histone protein modifications along a strand of DNA. Histone proteins have two main roles within the body. They help store DNA in a protected state so that when the DNA is not being utilized it is not damaged. Histone proteins act as the spool, if DNA is the thread wrapped around them for storage. Histone proteins also play a role in gene regulation. Many scientists believe that there is a link between modifications to histone proteins and whether or not certain genes are expressed [15]. This knowledge could potentially be used to identify predisposition to certain diseases or aid in the treatment of them [2, 6, 8, 9, 14, 15].

ChIP-Seq is a multi-step process that seeks to locate specific histone modifications. ChIP-Seq is composed of two main steps as depicted in Fig. 2. The first step, chromatin immunoprecipitation, involves mixing the DNA with an anti-body that is known to bind to the histone protein modification of interest. Once the antibody is bound, the DNA is fragmented into randomly sized pieces often using sonication. Next, DNA fragments bound to an antibody are separated out from those that have no antibody bound to them. The DNA is unwound from the histone proteins, and the antibody is unbound. At this point, only fragmented strands of DNA that were located near a bound antibody, and therefore also near the histone protein of interest, remain [6-10, 13].

The second aspect of ChIP-Seq involves high throughput sequencing of the identified fragments. Each fragmented strand of DNA has its ends repaired and ligated to a pair of adaptors. Afterwards, the fragments are replicated many times using the polymerase chain reaction (PCR) to increase the chance of correctly discovering the location of the histone protein modification of interest. Finally, the fragments are aligned to their locations along the DNA strand. A histogram is formed, containing the number of fragments, or read counts, that align to specific locations along the genome. High read count at a specific location indicates that an antibody was bound during ChIP-Seq and a likely location of the histone protein modification [6-10, 13].

Although the biological ChIP-Seq process produces a histogram from which likely locations for specific histone modifications can be deduced, this output can be quite noisy as well as contain artifacts and general bias. This is why the ChIP-Seq output is subsequently used as input to another program to further scrutinize the data for true peaks. This additional analysis tries to remove existing artifacts and locate true peaks in the data after forming a signal profile from the read counts. There are many open-source algorithms available for processing ChIP-Seq data. Even though these algorithms all operate differently, most use a form of peak detection to more accurately predict histone protein modification locations [1, 3-7, 9, 13].

This project is the first step in understanding currently available ChIP-Seq software for application here at the University of Maryland. It seeks to understand how different algorithms find peaks from the ChIP-Seq histogram output, how each algorithm determines the significance of the peaks it locates, and evaluate their accuracy. Two specific algorithms, SICER [4] and MACS [16] are examined in detail. This project is the beginning of an interdisciplinary effort to find an ideal algorithm for use here at the University of Maryland.

Manuscript received August 2, 2010. This material is based upon work supported by the National Science Foundation under Grant No. 0755224.

K. D. Fischl is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kfischl@princeton.edu).

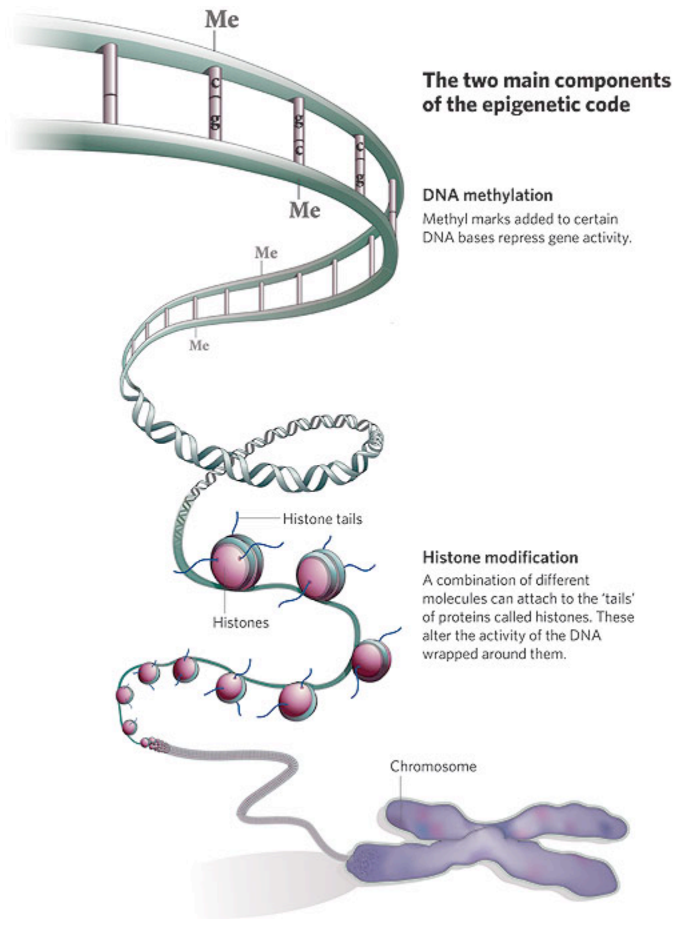

Me
The two main components of the epigenetic code
c
g
g
c
c
g
DNA methylation
Methyl marks added to certain DNA bases repress gene activity.
Me
Me
Me
Histone tails
Histone modification
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.
Histones
Chromosome

Fig. 1. Histone protein modifications are one of two main epigenetic components linked to DNA. Histone proteins act as the spool around which DNA winds as the thread, to be stored safely and securely in the body while it is not being used. Histone proteins also play a role in gene regulation, which is the main aspect of their epigenetic nature. (Figure from [10].)

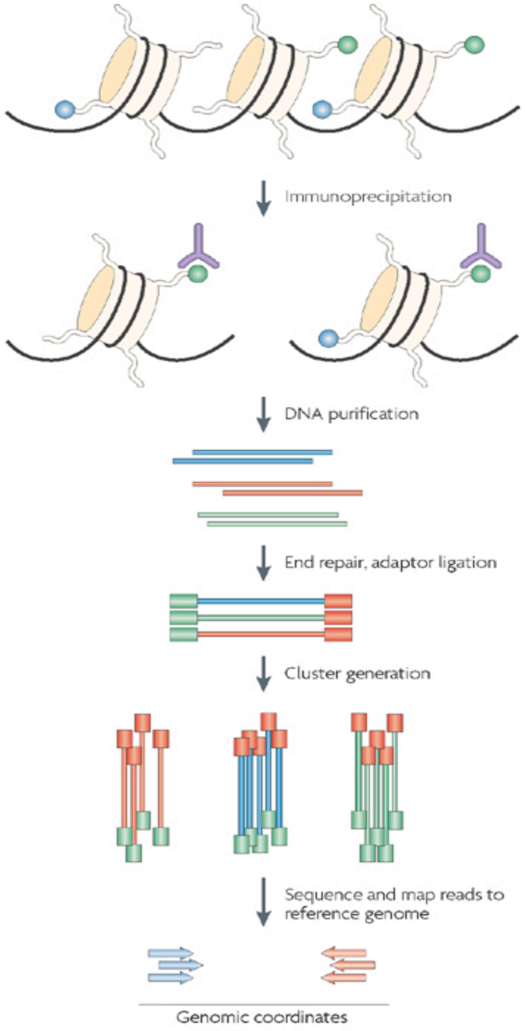


Fig. 2. The Chip-Seq process is composed of two steps and seeks to locate specific aspects along a genome. The first step includes chromatin immunoprecipitation (ChIP), which involves mixing the DNA with an antibody known to bind to the histone modification of interest and then fragmenting the DNA at random locations. The second step (seq) begins by separating out any DNA with an antibody bound to it, unbinding the antibody, and unwinding the DNA fragments. Then the ends of each fragment are repaired and ligated. Each fragment is replicated many times using PCR. Finally, the fragments are aligned against the genome to determine their original place in the DNA sequence and form a histogram of likely histone protein modification locations. (Figure from [6].)

II. Background

There are over thirty different open source ChIP-Seq algorithms available. Each algorithm has two main functions: 1) to locate statistically significant peaks in the read count data and 2) to determine the statistical significance of the peaks identified. Many algorithms report a false discovery rate (FDR), which evaluates the probability that peaks identified by the algorithms are the true peaks representing actual histone protein modification locations. Each algorithm accomplishes these two tasks in a unique way and with different statistical models. To locate statistically significant peaks, algorithms use a variety of statistical measures including sliding window algorithms, Markov models, Gaussian kernel density estimators, clustering approaches, etc [4]. To determine the statistical significance of the peaks that are found, the methods vary depending on if a control data set has been provided. If control data has been provided ChIP-Seq algorithms often use Poisson distributions, local Poisson distributions, t-distributions, conditional binomials, hidden Markov models, etc. to determine the significance of discovered peaks [4]. If no control data has been provided then the algorithms often generate a background through statistical means and for comparison use a null hypothesis, Poisson model, or negative binomial model to determine the significance of the peaks [4]. Control data is used to eliminate biases generated during the ChIP-Seq process. Some algorithms allow the user to input control data if desired, while others require control data as an input value. Algorithms are written in a variety of programming languages, with different input parameters, which consequently result in very distinctive levels of usability among the different ChIP-Seq programs [3, 5, 7].

The two algorithms that this study explores in depth are called SICER and MACS [4, 16]. SICER was chosen because it was previously used to analyze the given data in this study, and is one of the few ChIP-Seq processing algorithms developed specifically for locating histone protein modifications. MACS was chosen because previous studies have repeatedly shown it to be able to identify the most peaks at a high rate of accuracy [3, 7]. MACS also comes with explanatory documentation that is easy to download and understand. A third algorithm entitled F-Seq was also studied [1]. We were unable, however, to determine if F-Seq is compatible with our data sets and thus F-Seq's capabilities were not compared with SICER's or MACS's.

A. SICER

SICER stands for spatial clustering approach for the identification of ChIP-enriched regions. SICER was developed specifically to locate histone modifications, whereas other algorithms were initially developed to locate transcription factor binding sites but can also be used to locate histone protein modifications too.

The overall operation of SICER is shown in Fig. 3. SICER firsts partitions the genome into non-overlapping windows of equal size. SICER computes a score for each window using a Poisson distribution parameterized by the average number of reads in a window. The score represents the negative logarithm of the probability of finding the number of reads observed to be in the current window. This assumes that reads can align anywhere along the genome with equal probability. After all scores are computed, a window is deemed to be "eligible" if its actual read count is greater than a threshold read count, which is determined by a p-value requirement based on a Poisson distribution. The p-value denotes the probability of obtaining the observed read counts in that window by random chance. "Eligible" windows no more than the designed gap length away from other "eligible" windows are merged together to form islands. The user specifies this gap length at run-time. Islands' scores are the sum of the scores of the "eligible" windows that comprise them.

When control data is not available, a random E-value threshold specified by the user and a random background is then used to determine the statistical significance of each island. When control data is available, a lenient E-value is first used to determine each island's significance and then the islands are compared against the control data to further evaluate their significance. False discovery rate thresholds, fold changes, and p-values are lastly determined for all peaks. False discovery rate thresholds are the highest false discovery rate at which the island would still be deemed significant. Fold changes indicate a ratio between the true discovery rate and the false discovery rate, and p-values represent the statistical significance of this peak being a true peak. SICER takes into account enrichment information from neighboring nucleosomes, which gives it added sensitivity and specificity when compared to other algorithms. SICER has compared favorably against the ChIP-Seq algorithms entitled MACS, FindPeaks, F-Seq, and QuEST when evaluated by scaling the data repeatedly to test the significance of identified peaks after each scaling has been performed [4]. SICER is written in python and its source code and documentation can be downloaded from http://home.gwu.edu/~wpeng/Software.htm.

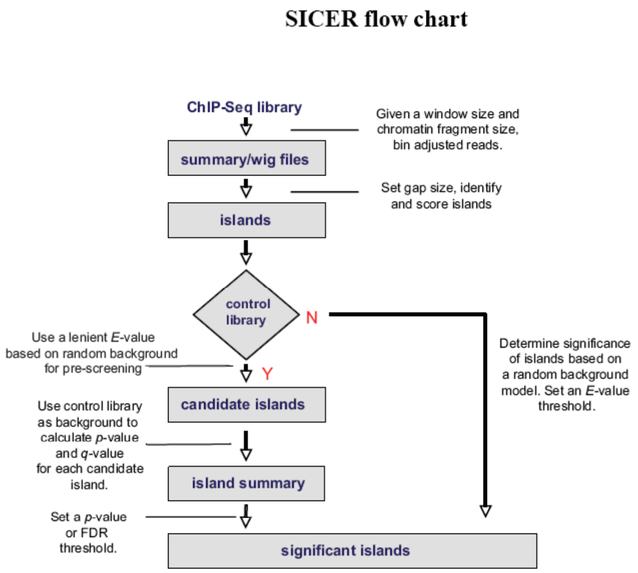
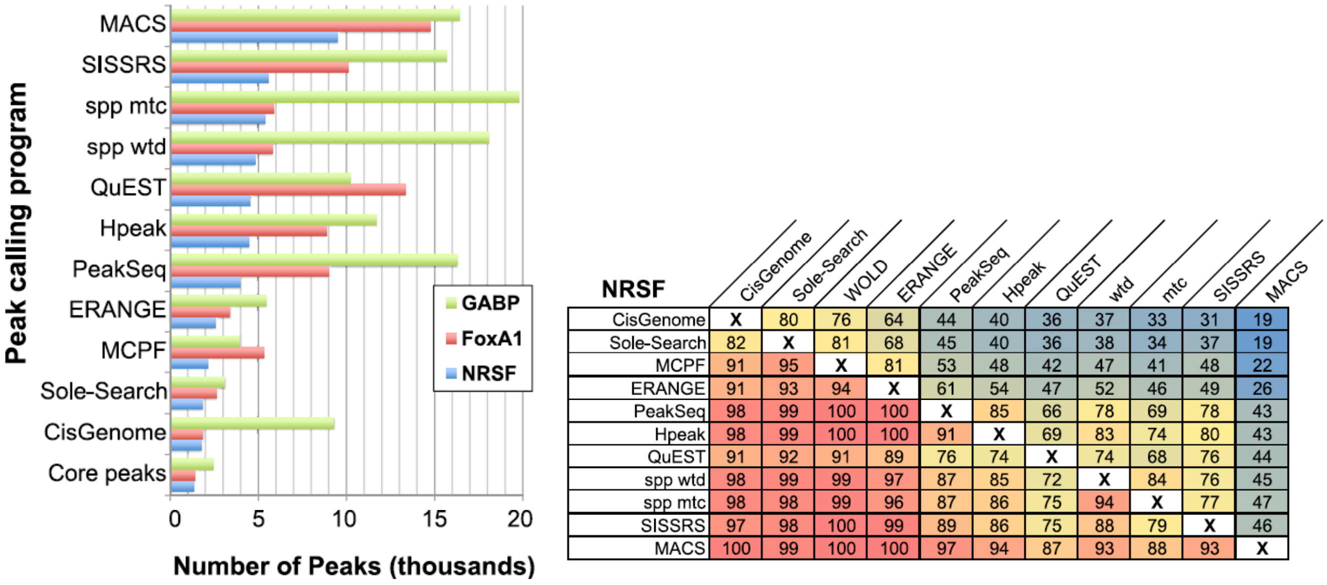

SICER flow chart
ChIP-Seq library
Given a window size and chromatin fragment size, bin adjusted reads.
summary/wig files
Set gap size, identify and score islands
islands
control library
N
Y
Use a lenient E-value based on random background for pre-screening
Determine significance of islands based on a random background model. Set an E-value threshold.
candidate islands
Use control library as background to calculate p-value and q-value for each candidate island.
island summary
Set a p-value or FDR threshold.
significant islands

Fig. 3. The SICER algorithm has two main methods of computation depending on if a control library is given. This diagram illustrates both types of processing that can occur with SICER. (Figure from [4].)

B. MACS

MACS operates very differently from SICER. MACS stands for Model-based Analysis of ChIP-Seq. MACS is a very widely cited algorithm that produces accurate and strong results whenever it is compared against other algorithms. It uses a sliding window approach to identify peaks. MACS is also written in python and provides great documentation and easily understood output files. MACS can be downloaded from http://liulab.dfci.harvard.edu/MACS/.

MACS finds peaks in the read count signal by first sliding a window across the data to find candidate peaks with significant enrichment determined by a local Poisson distribution. Using a local Poisson distribution allows MACS to overcome local biases cased by genome binding and determine locations with read counts significantly higher than that of random chance. Peaks that overlap are merged and the position within the peak with the highest amount of overlapping fragments is considered to be the summit. The summit is considered to be the specific location of the histone



NRSF	CisGenome	Sole-Search	WOLD	ERANGE	PeakSeq	Hpeak	QuEST	wtd	mtc	SISSRS	MACS
CisGenome	X	80	76	64	44	40	36	37	33	31	19
Sole-Search	82	X	81	68	45	40	36	38	34	37	19
MCPF	91	95	X	81	53	48	42	47	41	48	22
ERANGE	91	93	94	X	61	54	47	52	46	49	26
PeakSeq	98	99	100	100	X	85	66	78	69	78	43
Hpeak	98	99	100	100	91	X	69	83	74	80	43
QuEST	91	92	91	89	76	74	X	74	68	76	44
spp wtd	98	99	99	97	87	85	72	X	84	76	45
spp mtc	98	98	99	96	87	86	75	94	X	77	47
SISSRS	97	98	100	99	89	86	75	88	79	X	46
MACS	100	99	100	100	97	94	87	93	88	93	X

(a) (b)

Fig. 4. These two graphs are taken from [7]. The graph on the left (a) summarizes peaks located for each of the twelve algorithms compared. The chart on the right (b) contains the number of peaks found when locating the transcription factor NRSF. The chart shows the percentage of the peaks found by the algorithm on the top that are also found by the algorithm on the left hand side.

protein modification along the genome. If control data is available, MACS operates similarly to SICER and compares the two datasets to further eliminate peaks that are no longer significant. With or without the use of control data, MACS reports a significance rating for each peak found, which describes its likelihood of being an antibody-binding site.

C. Existing Comparative Studies

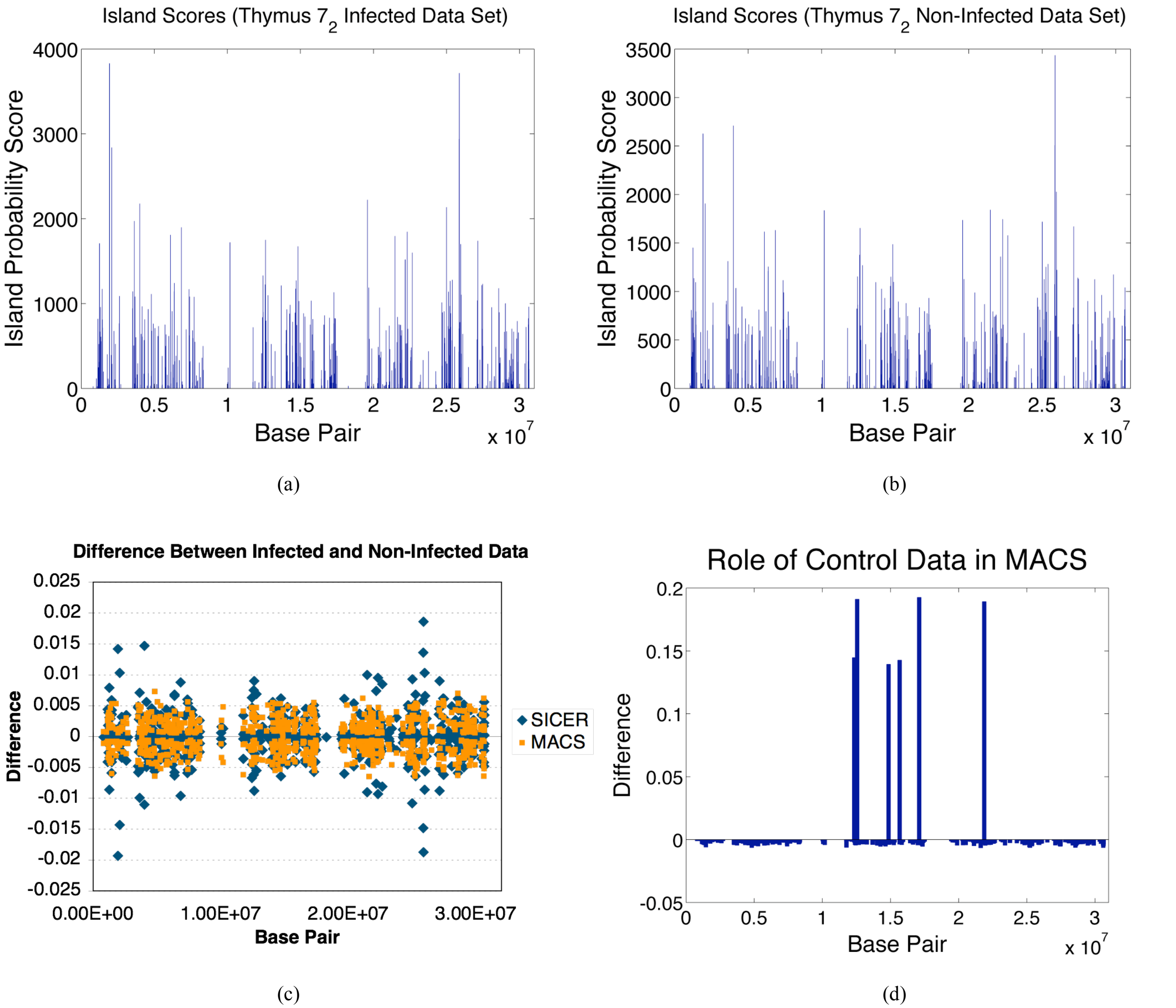
Although most ChIP-Seq algorithms are only a few years old, researchers have begun trying to find comparative ways to evaluate which of these methods are most accurate in locating histone protein modifications or transcription factor binding sites along a genome. These studies face many difficulties, however, because of the variation between the nature, input, and output for the ChIP-Seq algorithms. Studies often use existing or generated data with known peak locations to evaluate the accuracy of peak identification and significance prediction [3, 5, 7, 14].

One such study analyzed eleven open-source algorithms against three published transcription factor ChIP-Seq datasets with controls [7]. This study chose these eleven algorithms to represent the different types of algorithmic approaches available and ran all processing by using default controls that any average user would use. The study compared the peaks identified by each algorithm as well as which algorithms found the same peaks (Fig. 4). It also studied the accuracy with which each algorithm located peaks and the significance of those located peaks.

The study determined that Kharchenko et al.'s ChIP-Seq processing pipeline (spp), which uses directionality scoring, and MACS were the best at finding the location of the binding sites. It also found that most of the eleven algorithms chosen identified peaks at similar rates, which it concludes is not surprising since many of these algorithms are developed and tested with similar data. This study does not conclude that any one algorithm is superior, but instead indicates that better methods and testing data are needed to adequately determine an algorithm with the best performance.

The authors of a ChIP-Seq algorithm entitled, U-Seq, also performed a similar study in 2009 [3]. Authors and users of ChIP-Seq algorithms were asked to submit their algorithms as part of a "ChIP-Seq Community Challenge 1.0". Twelve different peak detection algorithms were submitted and run with data containing simulated ChIP-Seq reads added to real experimentally derived data. This study only tested for peak identification and significance prediction. The study concluded that MACS, USeq, Partek, SWEMBL, and ParkLab algorithms were the best at identifying peaks and that CisGenome, USeq, and MACS were the best at accurately predicting peak significance.

Neither of these studies included SICER in their list of participating algorithms but both rated MACS among the algorithms with the most accurate outputs. These studies begin to address a question that will become more important as a need for a reliable ChIP-Seq processing algorithm becomes more necessary. Both studies provide helpful comparisons but lack overarching and conclusive results.



(a)

(b)

(c)

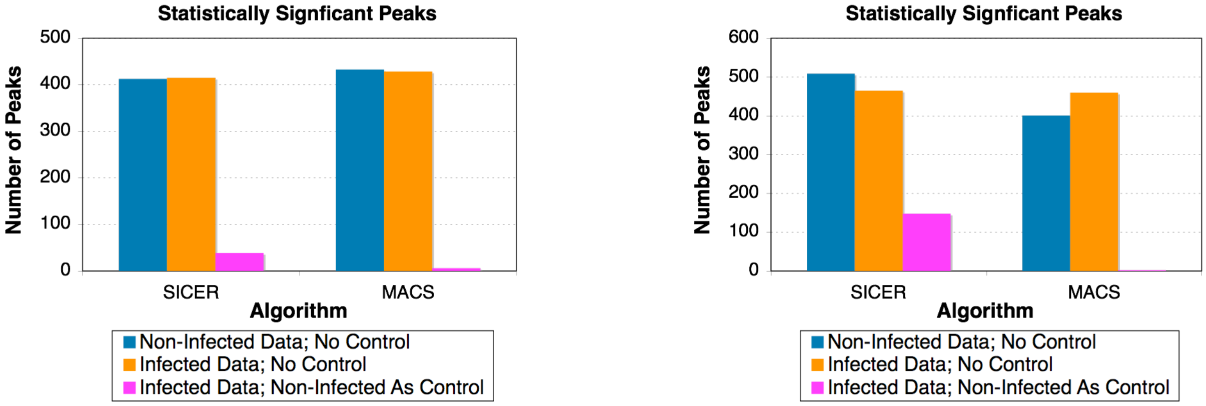
(d)

Fig. 5. (a) Output from running SICER with the data set from the thymus 7_2 infected line as input and with no control data. The graph displays all significant islands with their probability score. (b) Output from SICER for the data set from the thymus 7_2 non-infected line as input with no control data. (c) The difference between the output for both SICER and MACS, for the data sets from the non-infected and infected thymus 7_2 lines with no control data used. (d) The difference in MACS between the output for the data set for the infected thymus 7_2 line when the non-infected data set is used as a control and when no control is used.

III. Methods

The data used in this study comes from chickens that have been inbred so that they all contain the same DNA, or are effectively all identical twins. When exposed to disease some of these chickens become infected with disease, while some were more resistant to becoming infected with disease despite the fact that all chickens have the same DNA. This result is the essence of epigenetics, a gene expression caused by something besides the DNA sequence. ChIP-Seq is well suited for locating specific histone protein modifications responsible for these specific gene expressions. This study uses two separate sets of ChIP-Seq data, both from the eighth chromosome of the DNA found in the chicken thymus. One set is the 6_3 line and one is the 7_2 line. The read count profiles were compiled into BED file formats for analysis with the MACS and SICER algorithms. BED file formats are a type of file developed specifically for helping display aspects along a chromosome and are described in detail on the UCSC Genome Bioinformatics website.

Much of this study was spent on understanding how to process the data using the MACS, SICER, and F-Seq algorithms. All ChIP-Seq algorithms allow the user to specify



(a) (b)

Fig. 7. These graphs show the difference in the number of significant peaks identified in SICER and MACS for both (a) the data set from the thymus 7_2 chicken line and (b) the data set from the thymus 6_3 chicken line.

certain parameters, either as command line inputs or in the body of the program code. As previously stated, we were unable to discover how to alter F-Seq's parameters for use with the chicken genome. Most ChIP-Seq algorithms assume the user will be processing for the human genome as a default, but most offer options to change this setting for any organism of interest. For both SICER and MACS a window size of 200 base pairs was used for the initial scanning step at the beginning of processing. When running SICER, we used a gap size of 600 base pairs to form "eligible" islands, an E-value of 0.003 to determine island significance, and a fragment length of 190 base pairs as input parameters. The effective genome size used was 30671729 base pairs when running MACS. The effective genome size signifies the length of DNA that can be mapped and thus excludes any repetitive DNA, or DNA with no genetic consequence. When running SICER, the default value of 0.75 was used as the ratio between the effective genome size and actual genome size.

To analyze the data, both the 6_3 and 7_2 lines were processed in both SICER and MACS without control data as well as using the data from the non-infected chicken sets as control data from the infected chicken sets. The differences between the peaks identified in the data sets from the infected chickens compared to the data sets from the non-infected chickens were graphed, as well as differences in peaks identified in the data sets from the infected chickens when the data sets from the non-infected chickens was used as a control and when no control was used. The input parameters were also altered to test the effect, which in addition to considering suggested values described in each algorithm's "readme" file, helped us decide to use the parameters described above.

IV. Results

Because no confirmed data exists for the locations of the histone protein modifications of interest considered in this study, the results of this study mainly consist of comparisons between the two algorithms. To determine the accuracy of these algorithms in predicting the histone protein modification locations, further biological studies must be performed. Because the output data for SICER and MACS is displayed quite differently and contains different data, there are many aspects for comparisons between the two algorithms.

After graphing the output of SICER and MACS for both data sets with no control data, one can visually determine that differences exist between the data sets for the infected and non-infected chickens output by one algorithm (see Fig. 6(a) and 6(b)). One can also conclude by visual comparison, that the output for the same data by MACS and SICER differs. Graphing the differences between the data sets for the infected and non-infected chickens normalized against their total read counts, when neither is using control data, reveals areas of particular difference along the genome as seen in Fig. 6(c). Graphing these differences for SICER and MACS reveals that SICER outputs a greater difference between the data sets for the non-infected and infected chickens of the same line. By outputting larger differences between the data sets, SICER makes it easier to spot areas of difference in those datasets. Because the differences are less pronounced in the output from MACS, distinguishing between the two sets is much more difficult. SICER's output makes the areas of largest difference clear and easy to spot, despite the amount of data being processed.

As a measure of comparison, the difference of the outputs for the data sets for the infected chickens when the data sets for the non-infected chickens is used as a control and when no control is used can also be graphed. This comparison is only valid for MACS because SICER assigns a p-value to each island when control data is available and an island probability score when control data is not available. Graphing this comparison for MACS, results in illuminating just a few locations of significant difference between the data sets for the infected and non-infected chickens, as seen in Fig. 6(d). Although MACS may not illuminate areas of particular difference between the two data sets when the difference is graphed, as in Fig. 6(c), Fig. 6(d) provides an even easier method for spotting locations of distinct difference between the data sets for the non-infected and infected chickens. The graph shown in Fig. 6(d) locates six specific locations of statistically significance difference, in a way SICER cannot.

Lastly, Fig. 7 compares the number of significant peaks located by the two algorithms. In the 6_3 and the 7_2 chicken lines when no control data is used, MACS locates more significant peaks than SICER, but in the 6_3 chicken line without control data SICER locates more significant peaks than MACS. Because neither algorithm consistently locates more peaks than the other, more data is needed before further conclusions can be drawn. When control data is used, however, SICER locates many more peaks than MACS. Unfortunately, we cannot compare the accuracy at which the algorithms predict peaks to be true peaks, since we do not know the true locations of the histone protein modifications of interest. Additionally, other comparative studies have been unable to determine the importance of finding the most peaks in a data set [3, 7]. Therefore it is unclear if there is a value to locating more (or less) peaks.

Besides the differences in output data, SICER and MACS output their data in very different formats. Both provide detailed files generally explaining the output and what it means. MACS's output, however, is much more user-friendly and explanatory throughout all processing steps. MACS allows for many parameters to be altered on the command line, while SICER does not allow for much variation in terms of the inputs. Both algorithms, however, do have cited papers that explain in great detail how each algorithm operates. Nonetheless, after exploring different ChIP-Seq algorithms it is not surprising that MACS is widely used and praised for its capabilities.

V. Conclusions

Epigenetic studies are vital to the understanding of gene expression and malfunction. As technology improves and advances, scientists have discovered that simply knowing the DNA sequence of a specific organism is no longer enough to understanding its inner workings. Consequently, understanding epigenetics has become equally important in the world of fighting disease.

As the ChIP-Seq process becomes more streamlined and widely used, the corresponding algorithms also attempt to follow suit. Thus, it becomes paramount for ChIP-Seq process users to find algorithms that output reliable and true peaks. This study marks the first step in an interdisciplinary study at the University of Maryland towards an ideal algorithm for our specific project. Some conclusions can be drawn regarding the differences between the SICER and MACS algorithms but further study is required before an ideal algorithm can be defined.

VI. Acknowledgement

I would like to acknowledge the National Science Foundation CISE award #0755224 for making this study possible. I would like to thank my mentors: Professor Min Wu, Professor Jiuzhou Song, Avinash Varna, Wei-Hong Chuang, and Apratim Mitra, for all of their advice, guidance, and support. I would also like to acknowledge the MERIT-BIEN program for making this opportunity possible and providing so many valuable resources during this study.

References

[1] A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey, "F-Seq: a feature density estimator for high-throughput sequence tags," *Bioinformatics Advance Access*, Sept. 2008.

[2] C. D. Allis, T. Jenuwein, D. Reinberg, and M. L. Caparros, ed. *Epigenetics*. Cold Spring Harbor New York: Cold Spring Harbor Laboratory Press, 2007.

[3] ChIP-Seq Community Challenge. "Source forge: Find and Develop Open Source Software." Geeknet, Inc.

[4] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhau, and W. Peng, "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data," *Bioinformatics*, vol. 25, no. 15, pp. 1952-1958, Jun. 2009.

[5] D. A. Nix, S. J. Courdy, and K. M. Boucher, "Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks," *BMC Bioinformatics*, vol. 9, Dec. 2008.

[6] D. E. Schones and K. Zhao, "Genome-wide approaches to studying chromatin modifications," *Nature Reviews: Genetics*, vol. 9, pp. 179-191, Mar. 2008.

[7] E. G. Wilbanks and M. T. Facciotti, "Evaluation of algorithm performance in ChIP-Seq peak detection," *PLoS ONE*, vol. 5, no. 7, Jul. 2010.

[8] E. R. Mardis, "ChIP-seq: welcome to the new frontier," *Nature Methods*, vol. 4, no. 8, pp. 613-614, Aug. 2007.

[9] E. T. Liu, S. Pott, and M. Huss, "Q&A: ChIP-Seq technologies and the study of gene regulation," *BMC Biology*, vol. 8, no 56, 2010.

[10] J. Qiu, "Epigenetics: Unfinished Symphony," *Nature*, vol 441, pp. 143-145, May 2006.

[11] P. J. Park, "ChIP-Seq: advantages and challenges of a maturing technology," *Nature Reviews: Genetics*, vol. 10, pp. 669-680, Oct. 2009.

[12] R. Margueron and D. Reinberg, "Chromatin structure and the inheritance of epigenetic information," *Nature Reviews: Genetics*, vol. 11, pp. 285-296, Apr. 2010.

[13] S. Pepke, B. Wold, and A. Mortazavi, "Computation for ChIP-seq and RNA-seq studies," *Nature*, vol. 6, no. 11, pp. S22-S32, Nov. 2009.

[14] T. D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo, "A praticial comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments," *BMC Genomics*, vol. 10, no. 618, Dec. 2009.

[15] T. Kouzarides, "Chromatin Modifications and Their Functions," *Cell*, vol. 128, pp. 693-705, Feb. 2007.

[16] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, no. R137, Sept. 2008.