

DNase Hypersensitivity Analysis for Identification and Localization of Non-Coding Functional elements

Submitted by:

Amandeep Sarao

Agenda

- **Terms**
- **Introduction**
 - KIBRA Gene
 - Goal of the Project
- **Overview of the Process**
- **Methodology**
 - Primary Data Analysis Methods
 - Statistical Data Analysis Methods
- **Results and Discussion**
- **Conclusion and Future Work**
- **References**

Terms

- **Quantitative polymerase chain reaction:** qPCR is a modification of the polymerase chain reaction used to rapidly measure the quantity of the target nucleic acid in a clinical specimen.
- **DNase-I:** DNase I is an enzyme, which at low levels preferentially digests nucleosome-depleted DNA, whereas tightly packaged chromatin is more resistant to cleavage.
- **Amplicon:** The amplified sequence of DNA in the PCR process.
- **Ct (threshold cycle):** It reflects the point during the PCR reaction at which a sufficient number of amplicons have accumulated. It is inversely correlated to the logarithm of the initial copy number.
- **Relative quantification:** A relative quantification is used to analyze changes in gene amplification in a given sample relative to another reference sample
- **Calibrator:** Reference sample used in each relative quantification studies.
- **Control:** Sample having known concentration of DNA.

Introduction

KIBRA Gene

On 19 Oct, 2006, the gene “ KIBRA” was discovered by researchers at Translational Genomic Research Institute to be associated with the Human memory performance.

This new understanding related to the memory performance will be helpful for Scientist to develop new drugs that will improve memory function by minimizing the effects of age-related memory loss associated with different diseases such as Alzheimer's.

Goal of the Project

To better understand the regulatory networks controlling the KIBRA gene by looking at non-coding functional elements also referred to as cis regulatory sequences .

To localize these non-coding functional elements using Quantitative PCR and DNase Hypersensitivity Analysis.

Overview of the Process

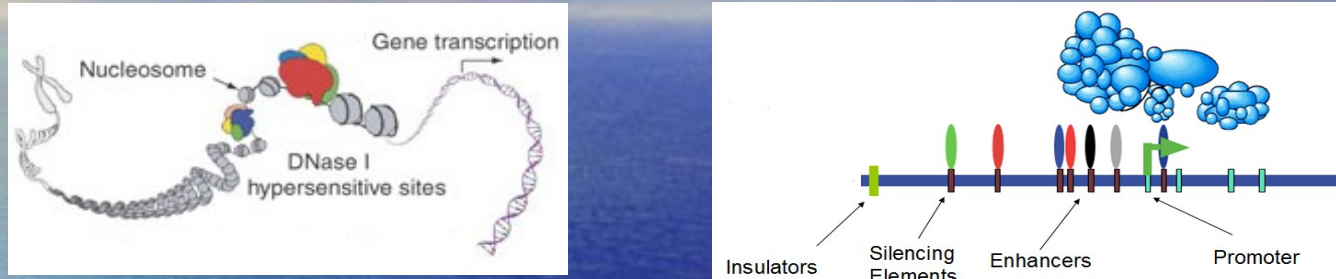


Figure: The graphs showing the Cis-regulatory Elements

The cis- regulatory sites i.e. Promoters, enhancers, suppressors, insulators, and locus control regions all have been shown to be associated with DNase-I Hypersensitive sites (HS sites). Therefore, we combined DNase Hypersensitivity detection assay and q-PCR technology to localize and quantify these sites in KIBRA gene. The idea is that these sequences can be detected by the presence of heightened sensitivity to DNase-I cleavage due to the absence of nucleosome structure.

Primary Data Analysis Methods

- Determination of Cycle Threshold values (Ct Values)
- Melting Curve Analysis
- Validation test for using Livak Relative Quantification Method
- Determination of DNase I Sensitivity Ratio using Livak Comparative C (t) method

Melting Curve Analysis

(Primary Data Analysis Cont'd...)

Basics:

- Following PCR amplification, the fluorescent signal of the reaction mixture is continuously monitored as it is heated to higher temperatures.
- Plotting this fluorescence as a function of temperature gives a “DNA melting curve”.
- Contaminating DNA or primer designs would show up as additional peak separate from desired amplicon peaks in these melting curve graphs.

Purpose:

It helps in distinguishing between specific versus non-specific PCR products.

Validation: Livak Relative Quantification Method

(Primary Data Analysis Cont'd...)

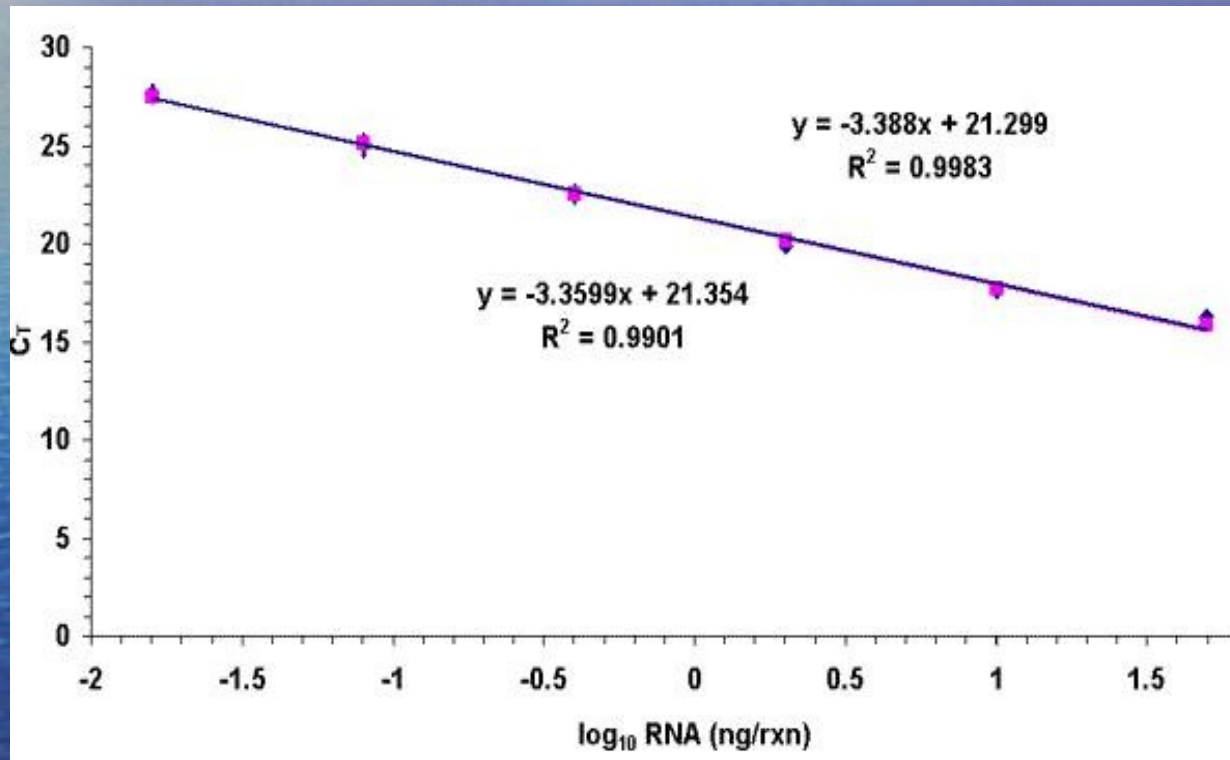


Figure: C (t) values of target and reference versus Log input amounts of DNA / RNA

Determination of Initial DNase I Sensitivity Ratio

(Primary Data Analysis Cont'd..)

Cycle threshold (Ct values) generated by Quantitative PCR is used to calculate DNA Sensitivity ratio using Livak relative quantification analysis steps :

1) Normalize the C (t) values of target using control and calibrator

$$\Delta C (t) = C (t) \text{ target} - C (t) \text{ control}$$

$$\Delta C (t) = C (t) \text{ target} - C (t) \text{ calibrator}$$

2) Calculate $\Delta C (t)$ Average

$$\Delta C (t) \text{ Ave} = \text{Average } \Delta C (t) \text{ of replicates}$$

3) Obtain the Fold difference

$$\Delta \Delta C (t) = \Delta C (t) \text{ Sample Ave} - \Delta C (t) \text{ Calibrator Ave}$$

4) Finally, we will compute the Initial DNase I Sensitivity Ratio

$$2^{-\Delta \Delta C (t)} = \text{Initial DNase I Sensitivity Ratio}$$

Statistical Data Analysis Methods

- 20 % Trimmed Mean
- LOWESS (Locally weighted least square smoother)
- MAD (Median Average Deviation Approach)
- Baseline Behavior Detection
- Computation of Signal to Noise Graphs

Trimmed Mean

(Statistical Data Analysis Methods Cont'd...)

- Performed 20% trimmed mean for Initial DNase Sensitive Ratio obtain using Livak method.
- It involves trimming of P (20%) percent observations from both sides of the dataset.

Purpose:

The purpose of taking trimmed mean is to detect and remove the glaring outliers from the dataset that can confound our final calculations.

Mean is too sensitive to extreme observation. The idea of taking trimmed mean is to solve that problem

LOWESS

Locally weighted Least Square Smoother (Lowess) is a non-parametric method based on linear and nonlinear least squares regression and was originally proposed by Cleveland in 1979. It is a procedure for drawing a smooth curve through a scatter diagram.

Purpose:

- This method helps in reducing systematic errors associated with the measured DNase sensitivity ratio.
- It helps smoothing the dataset by eliminating the lower and upper extreme values of the dataset.

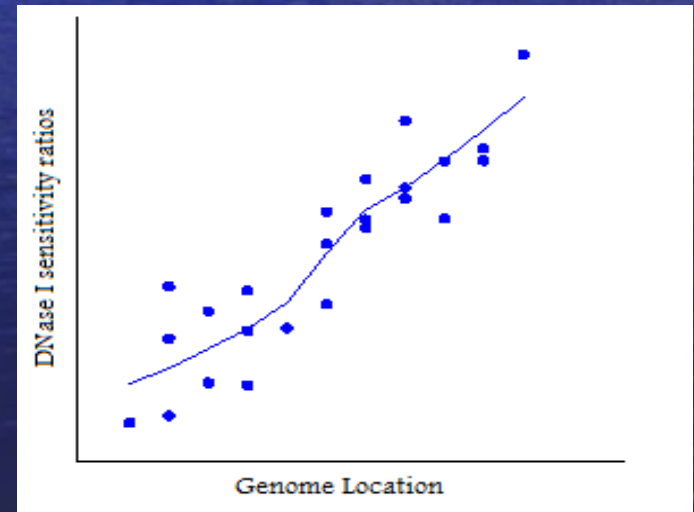


Figure: DNase Sensitive ratio versus Genomic locations showing smoothing curve

LOWESS

(LOWESS Cont'd...)

R-Code (LOWESS)

Lowess (x, f=0.2, iter=3, delta=.01*diff (range (x)))

f: the smoother span. This gives the proportion of points in the plot which influence the smooth at each value. Larger values give more smoothness.

iter: the number of robust iterations that should be performed. Smaller values of 'iter' will make 'lowess' run faster.

delta: values of 'x' which lie within 'delta' of each other replaced by a single value in the output from 'lowess'

The data point to be smoothed has the largest weight and the most influence on the fit and vice-versa. The weight function is given by:

$$w_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3$$

where x is the DNase Sensitivity ratio value

xi is the nearest neighbors for x

d (x) is the distance between the x and the most distant neighbor

MAD

Median absolute deviation (MAD) is a measure of estimated error. It is an estimate of how far off the actual values are from the average value, assuming that our measured data values are accurate. It is calculated by:

$$\text{MAD} = \text{median} (| Y_i - \tilde{Y} |)$$

where \tilde{Y} is the median of the data and $|Y_i|$ is the absolute value of Y

Purpose:

- X is declared as an outlier on the basis of the following criteria.
 $0.645 * |X-M| / \text{MAD} > 2.24$
where M is the median and MAD is the average median deviation
- It is a robust method in treating Non-Gaussian data values and eliminating extreme unwanted outliers.

Baseline Behavior Detection

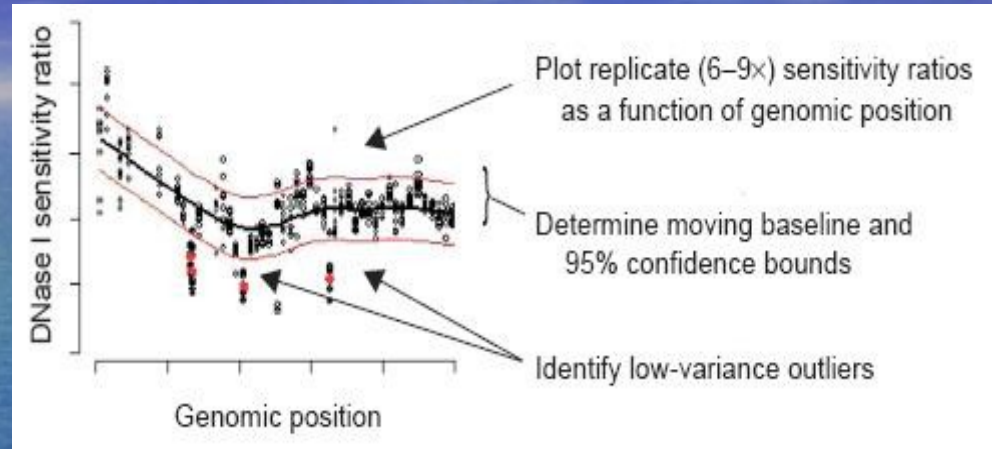


Figure: Plot of DNase-I sensitivity ratio versus Genomic positions

Purpose:

- To define Moving Baseline for the DNase Sensitivity Ratios
- To eliminate left out outliers (after applying trimmed mean, Lowess and MAD)
- To identify significant low variance outliers for which we will compute SNR to localize hypersensitive sites.

Computation of Signal to Noise Graphs

To compute SNR, we compute

$$S/N_i = |H S_i - B_i| / MAD_B * (1 + \sigma_c)$$

where SNR at site i is measured as the absolute deviation of the trimmed mean of the hyper sensitivity site ($H S_i$) from the interpolated baseline (B_i) divided by median average deviation of the centered baseline. The term in the denominator is the small correction factor that penalizes larger variances in hypersensitivity site clusters and rewards highly compact clusters.

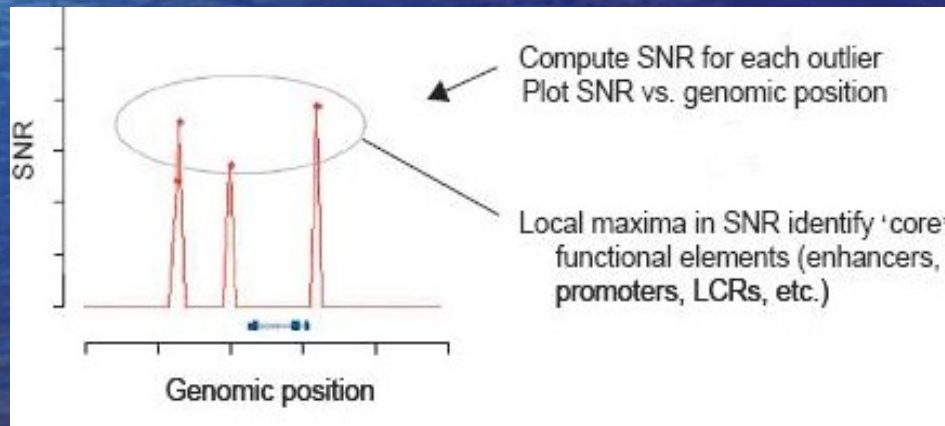


Figure: SNR versus Genomic LocationGraph for Significant outliers

Results and Discussion

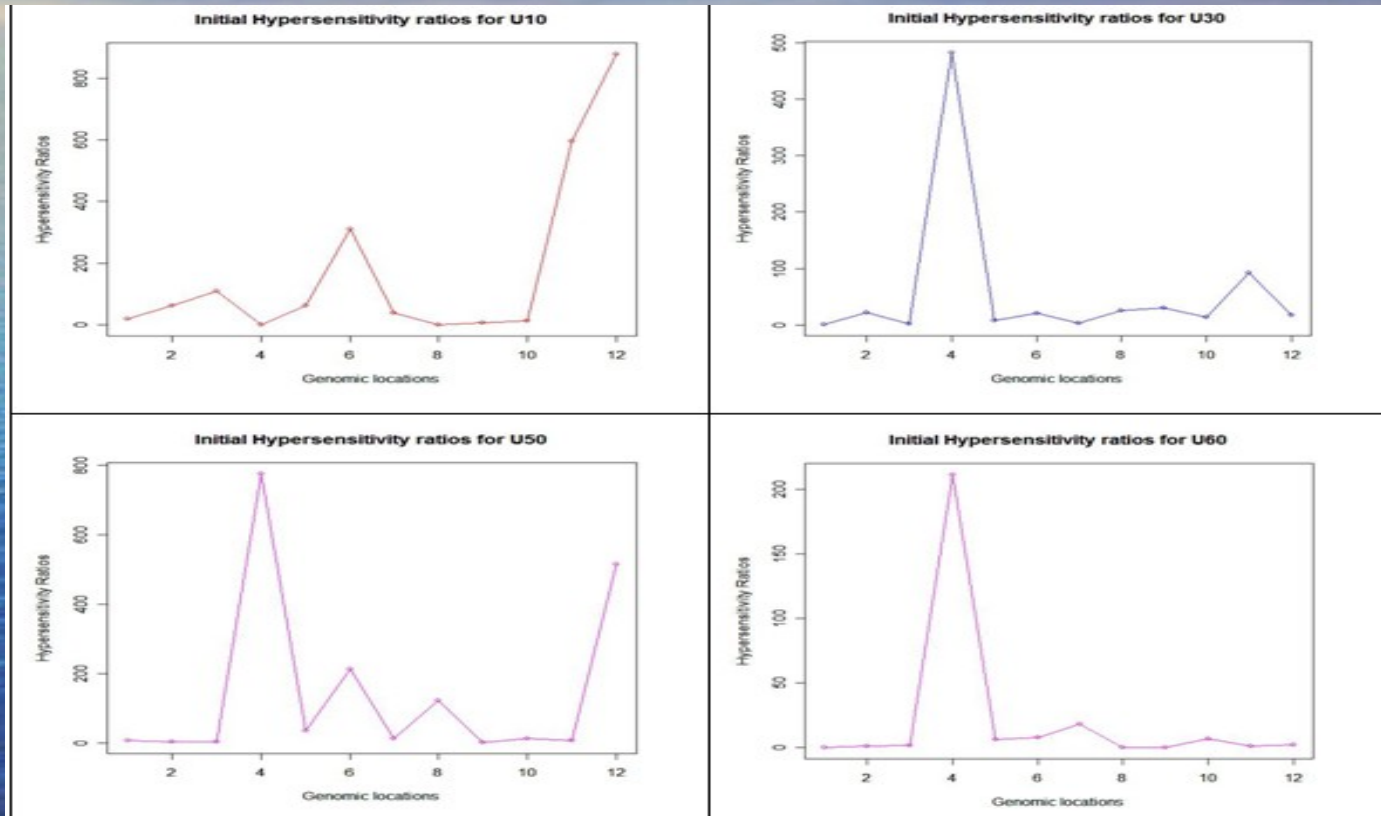


Figure: Graphs of DNase Sensitivity ratio versus Genomic Locations representing Initial Sensitivity Ratio

Results and Discussion

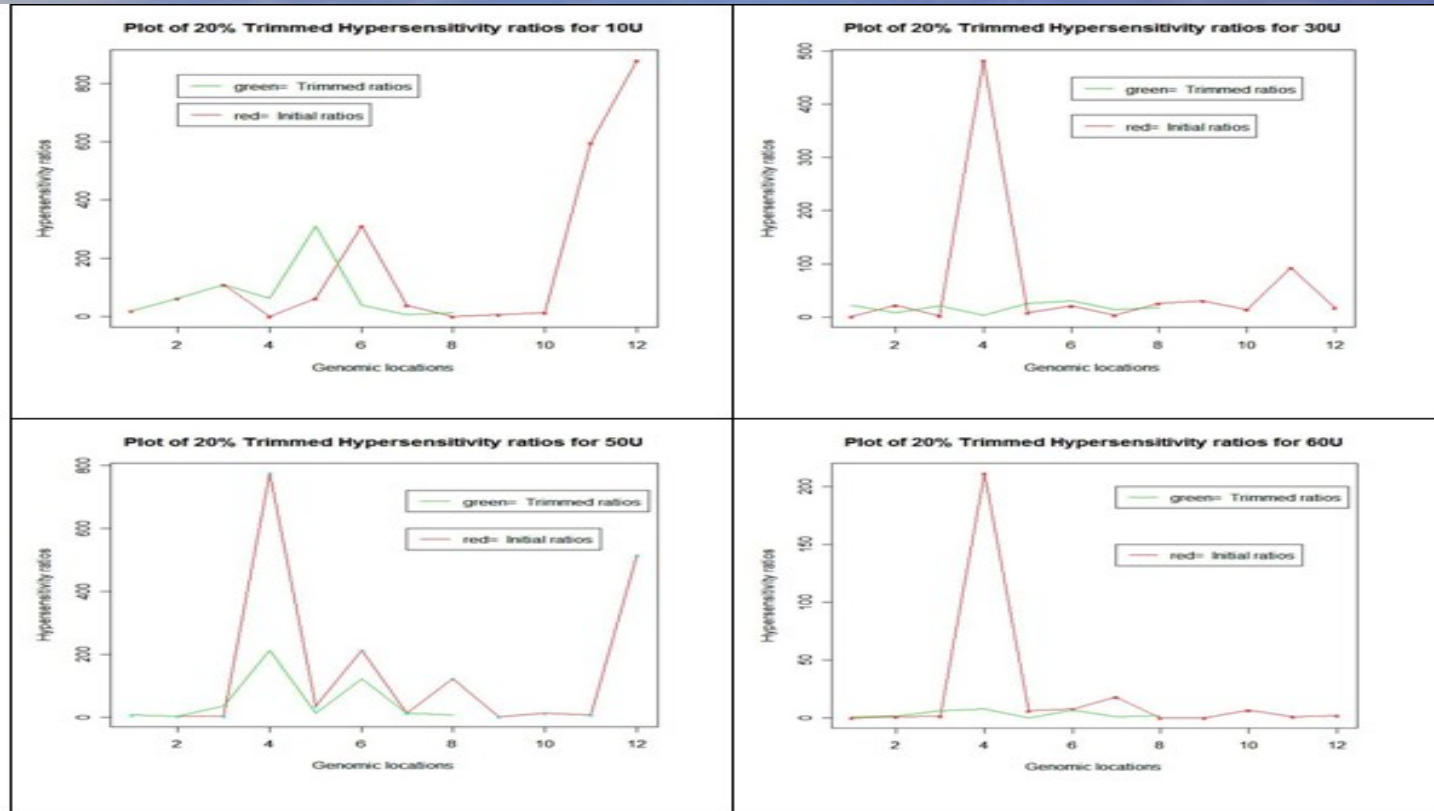


Figure: Graphs of DNase-I sensitivity versus Genomic locations showing Initial DNase Sensitive ratio and ratio obtained after taking 20% trimmed mean.

Results and Discussion

(Cont'd...)

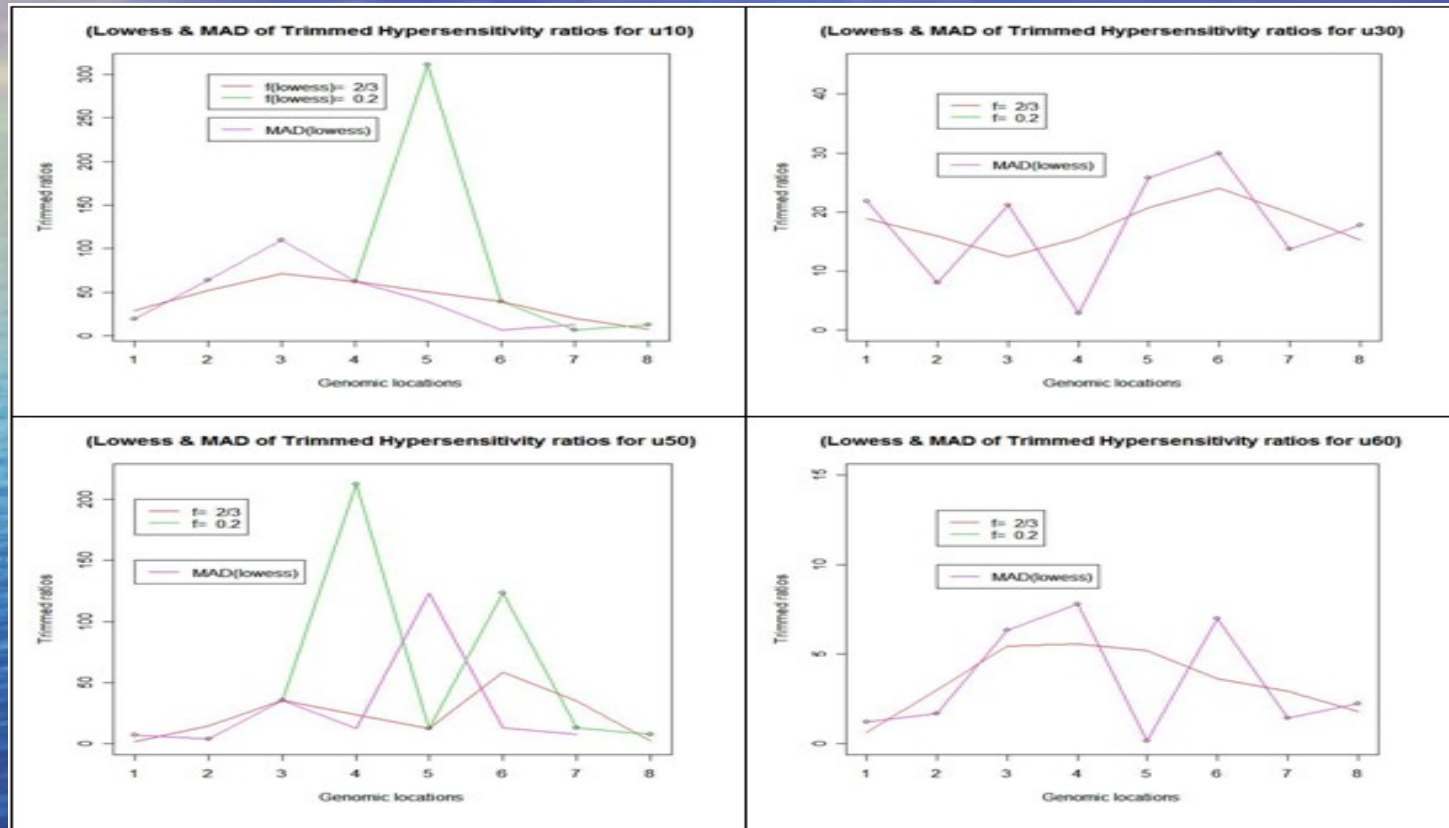


Figure: Graphs of DNase-I sensitivity versus Genomic locations showing Smoothed Curves obtained using Lowess and Median Average Deviation Approach (MAD).

Results and Discussion

(Cont'd...)

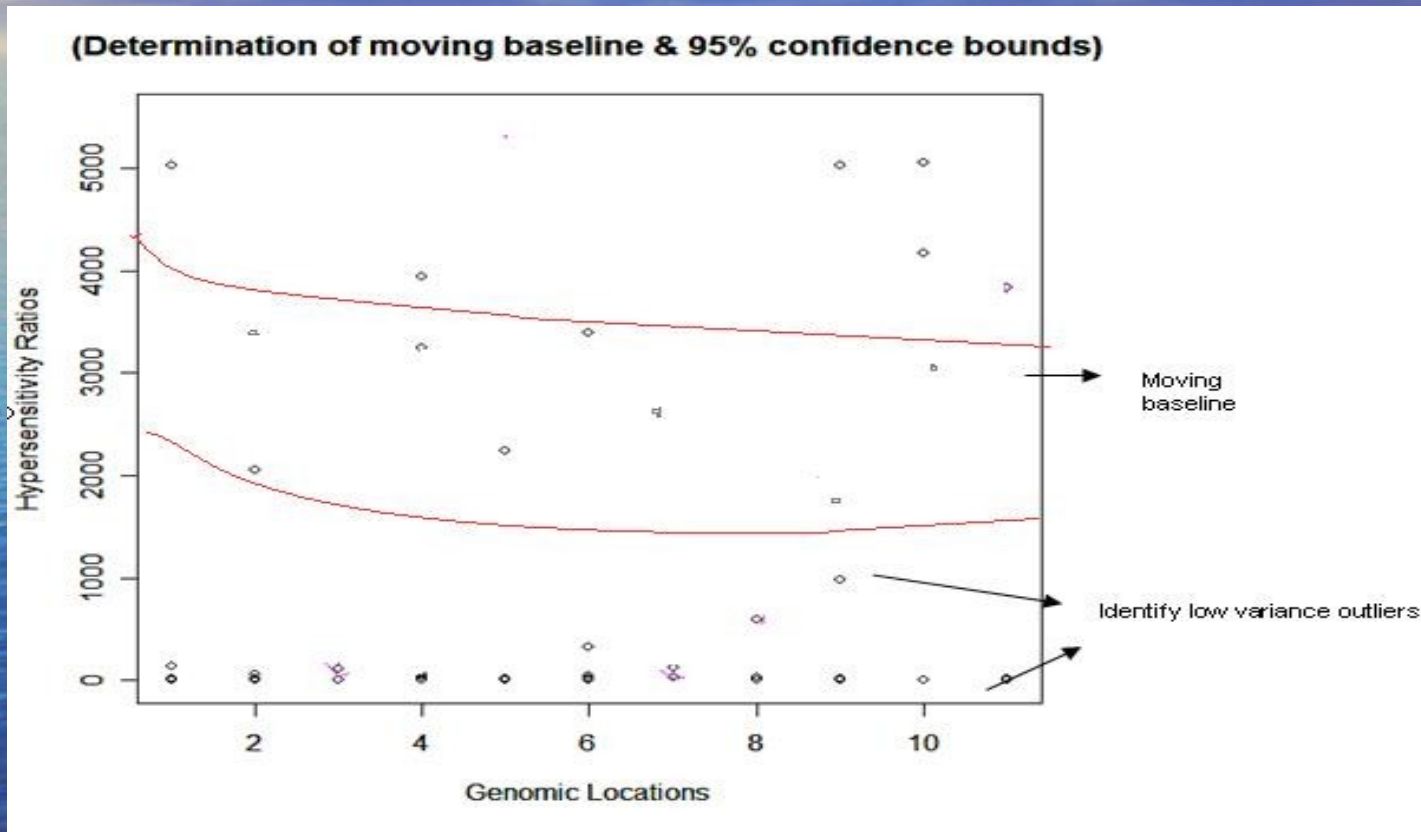


Figure: DNase-I Sensitivity versus Genomic locations Graph representing moving baseline and 95% Confidence bounds and the low variance outliers

Results and Discussion

(Cont'd...)

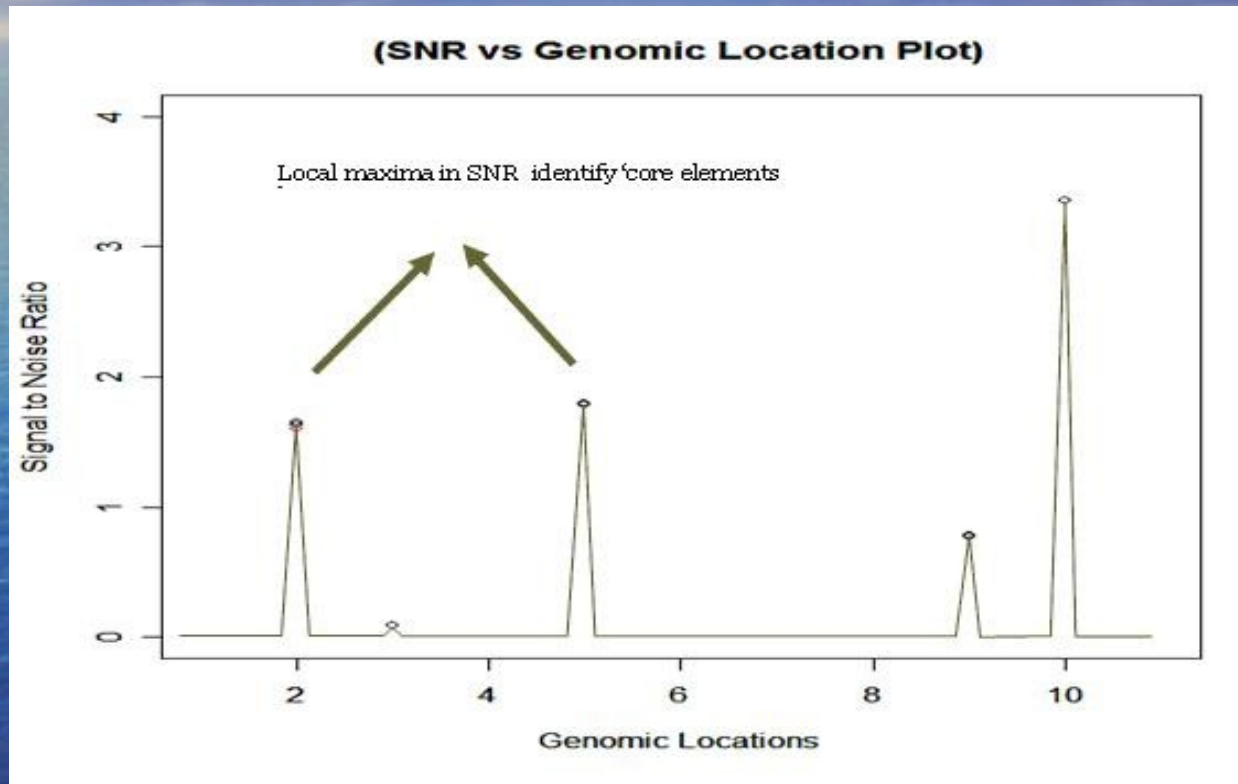


Figure: Graph of SNR (Signal to Noise ratio) versus Genomic Locations

Conclusion

The DNase-I Hypersensitivity analysis using quantitative PCR seems to be a high-throughput method for localization of intragenic and exogenic non-coding functional elements. But in our study, the exact location of these functional elements in the genomic sequences is still needed to be identified. Although we have obtained the SNR versus Genomic location graphs whose peaks reflects the presence of the core functional elements but still we have insufficient data to prove that we have actually identified these functional elements. We would need 6-8x replicates of each amplicon in order to obtain the SNR graphs and to correctly identify the presence of the core functional elements.

Future Work

Systematic, high-throughput identification of functional elements coinciding with DNase I hypersensitive sites will substantially improve our knowledge of transcriptional regulation and should simplify the search for non-coding genetic variation with phenotypic consequences. In other words, it will open the door for systematic exposition of the regulatory genome.

More importantly, these regions are known to be located in close proximity to genes for which loss of function or mutations result in severe abnormalities. But identification of these elements will ease the researchers and the scientist to come up with new therapies and drug to better target the diseases.

References

- Michael O Dorschner, Michael Hawrylycz, “High-throughput localization of functional elements by quantitative chromatin profiling”, *Nature Methods*-1, 219 - 225 (2004).
- C. A Heid, J Stevens, K J Livak, “Real time quantitative PCR”, *Genome Res.* 1996 6: 986-994
- Stefan and Augustinus Bader, “Quantitative Real-Time Polymerase Chain Reaction: Methodical Analysis and Mathematical Model”, *Journal of Biomolecular Techniques*, 15:107-111(2004)
- Kenneth J. Livak and Thomas D. Schmittgen, *METHODS*, “Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR”, 25, 402–408 (2001)
- Peter J Sabo, Michael S Kuehn, Robert Thurman, Brett E Johnson, “Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays”, *Nature Methods* 3,511-518(2006).
- Marisa L. Wong and Juan F. Medrano, “Real-time PCR for mRNA quantization”, *BioTechniques*® July 2005, Volume 39, Number 1: pp 75-85.
- Florian M. Pauler, Stefan H. Stricker, “Long-range DNase I hypersensitivity mapping reveals the imprinted Igf2r and Air promoters share cis-regulatory elements”, *Genome Res.* 15:1379-1387, 2005
- http://itl.nist.gov/div898/software/dataplot/refman1/ch3/lowess_s.pdf
- <http://hmg.oxfordjournals.org/cgi/content/full/15/13/2098>

Acknowledgements

I extremely appreciate my mentor and internship advisor, Dr. Jeffrey Touchman, for his valuable guidance and support throughout my project. I would also like to acknowledge Dr. Phillip Stafford and Dr. Chitta Baral for their consenting to be on my committee and providing me invaluable feedback. I am thankful to The Helios Foundation and TGen for funding my project.

Also, I would like to thanks Dr. Rosemary Renaut for providing me an opportunity to be a part of Computational Biosciences Program and her support throughout the program.