HST508 Final Project Report

Joshua Batson, Ina Chen, Scott Luro

December 2013

1 Introduction

We set out to find genomic features of TAD boundaries, and to determine if those signatures were strong enough to predict the locations of the boundaries.

2 TAD Boundaries

Rather than work with previously reported TAD boundaries, we elected to build our own as follows. We took 25-bin windows about the diagonal and calculated total hits to the left (L) and right (R) of the diagonal. We found that L/R yielded the sharpest profiles and defined the boundaries to be the area between a relative maximum (most contacts to the left) and relative minimum (most contacts to the right).

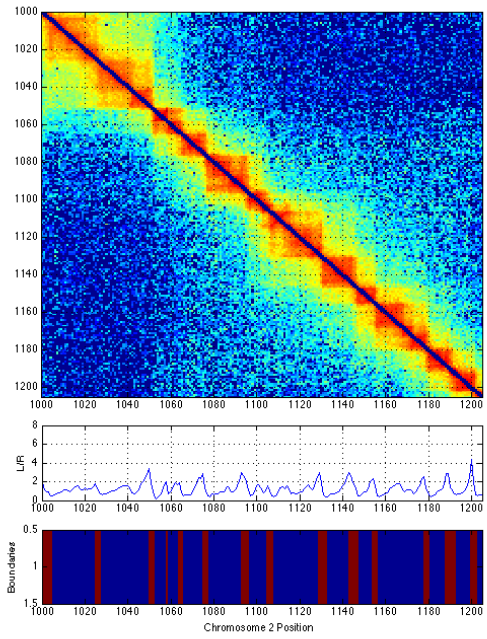

Figure 1: hES TAD boundaries used for analysis on an example stretch of Chromosome 2

We then attempted to classify TAD regions aside from boundaries and interiors by clustering the window profiles (left) but every metric pairs and triples tried produced amorphous clouds that did not have boundary points localized to a specific region (right).

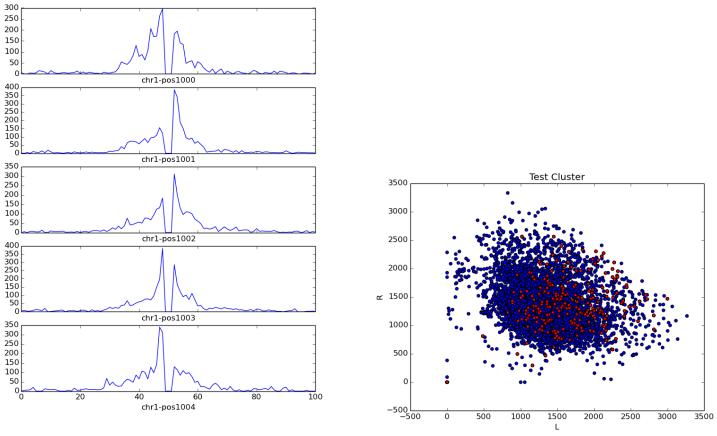

Figure 2: Left: Example profiles of the windows. Right: Scatter plot of points (blue) and boundary points (red) using two example metrics.

3 Sequence Search

TADs exist on the Mb scale, and their boundaries exist on the 40kb scale. We decided to begin our sequence search with 7-mers, since a given 7-mer will occur on average of 2.4 times in a random 40kb sequence of DNA. Long k-mers might occur too rarely to make out a pattern, and shorter kmers might carry too little information.

The most common sequences in TAD boundaries are not specific to TAD boundaries. For example, the most common 7-mers in TAD boundaries are also quite common in TAD interiors.

7-mer	frequency per 40kb (boundary)	frequency per 40kb (interior)
AAAAAAA	97	88
TTTTTTT	94	90
AAAAATA	24	24
TATTTTT	23	24
AAATAAA	23	23
random	2.4	2.4

These low complexity sequences are quite common throughout the genome, and therefore are not appropriate distinguishing marks between regions of the genome. To compensate for this, we defined a set of control regions by shifting the TAD boundaries right 1Mb. We then

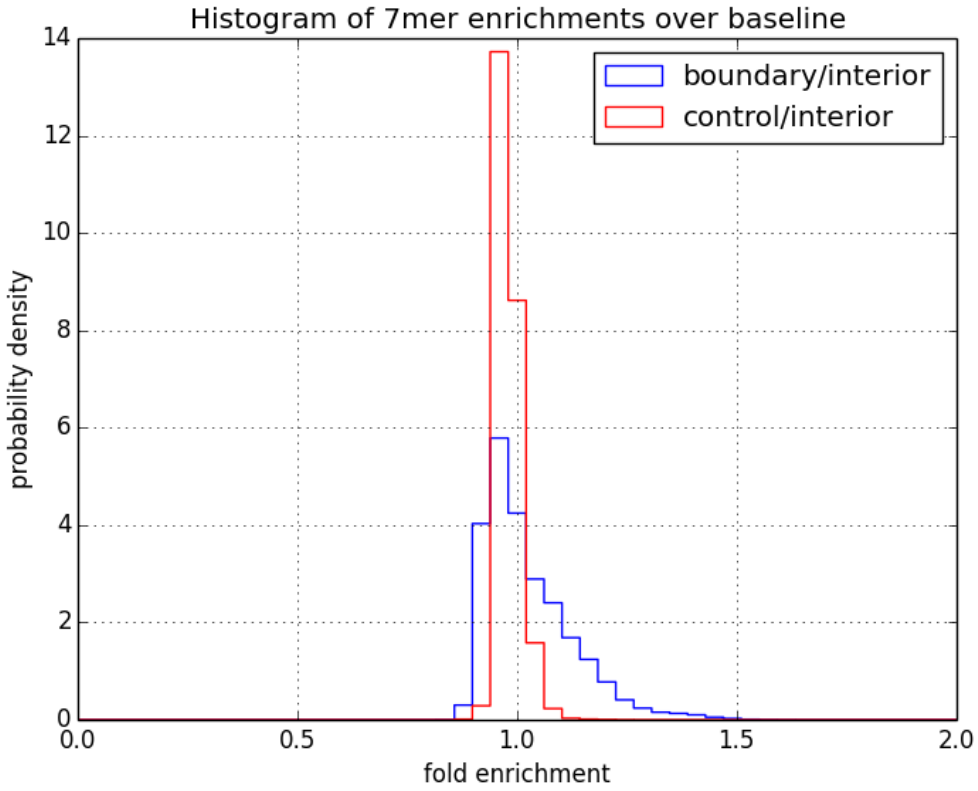

Figure 3: Histogram of 7-mer enrichments relative to TAD interiors, for TAD boundaries and controls.

computed the enrichment of each 7-mer in the TAD boundaries and in the control regions *relative to* the TAD interiors. Since TAD interiors are close to TAD boundaries, one might expect their 7-mer frequencies to be similar. In fact, the controls are much more similar to the TAD interiors than the boundaries are.

As you can see above in Figure 3, most 7-mers in control regions are only mildly enriched or depleted relative to TAD interiors. The 99^{th} percentile of enrichment for controls is just 1.08. TAD boundery enrichments extend much further to the right; the 99^{th} percentile of enrichment is 1.5. There are 500 7-mers with enrichment at least 1.4, which we will use below to try and construct a classifier. Not that you need convincing, but the p-value for a t-test to distinguish the two enrichment distributions is less than 10^{-300}; MATLAB reports it as 0.

All the area under the blue curve and above the red represents sequence information distinguishing TAD boundaries from interiors. Curiously, the most enriched 7-mers are quite rare; they are just even more rare in interiors than in boundaries.

7-mer	frequency per 40kb (boundary)	frequency per 40kb (interior)
ACCGACG	0.07	0.04
ACCGCGG	0.16	0.09
CGCAACG	0.05	0.03
CGCGCAA	0.05	0.03
CACCGCG	0.20	0.11
random	2.4	2.4

To visualize the distribution of enriched oligonucleotides, 160kb-regions centered about TAD boundaries were searched for the top five 7-mers. A histogram with bins of 2kb, smoothed over an average of 5 bins, displaying local boundary counts of hES cells is shown below in Figure 4. All non-discrete boundary calls were consolidated by defining the boundary position as the cluster midpoint. A control sequence with a lesser enrichment is shown for comparison (all enrichment values are parenthesized in the figure legend; N.B., top 7-mers differ from above table due to slight differences in boundary call thresholds). Enrichment is evident as mild peaks do appear near called boundaries, similar to the plots for ChIP-seq counts for CTCF and previously reported histone marks.

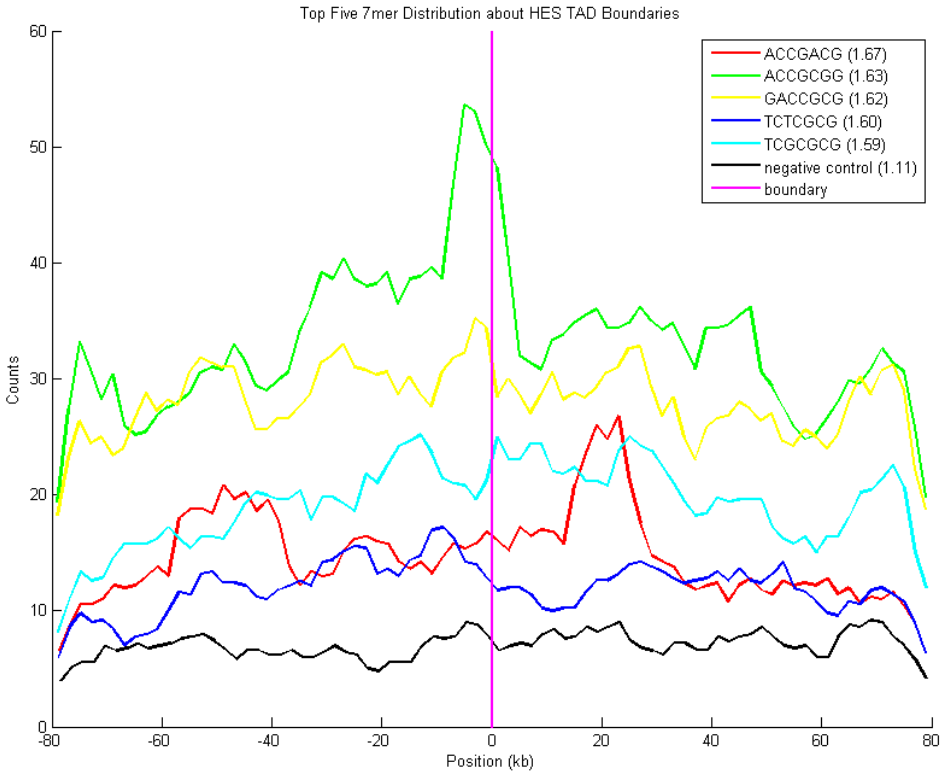

Figure 4: Positions of most enriched 7-mers relative to TAD boundaries

4 Boundary Prediction Using Sequence

Let us begin with an extremely simple model: a fixed 7-mer ω occurs at a generic location in a TAD boundary with probability p_1 and in a TAD interior with probability p_2. The count of ω in a 40kb bin then follows a Poisson distribution, with means λ_1 and λ_2, respectively. The likelihood of seeing a specific count m is

$$\mathbb{P}_\lambda(m) = e^{-\lambda}\frac{\lambda^m}{m!}.$$

The relevant log likelihood ratio is then

$$\log\frac{\mathbb{P}_{\lambda_1}(m)}{\mathbb{P}_{\lambda_2}(m)} = \log e^{-(\lambda_1-\lambda_2)}\frac{\lambda_1^m}{\lambda_2^m} = -(\lambda_1-\lambda_2) + m(\log\frac{\lambda_1}{\lambda_2}).$$

Note that the ratio $r = \lambda_1/\lambda_2$ is merely the enrichment of ω in TAD boundaries relative to interiors.

Given counts of a bunch of 7-mers $\omega_1, \ldots, \omega_N$ in a given 40kb bin, the log likelihood is

$$L = \Lambda + \sum_j m_j \log r_j,$$

where m_j is the count of ω_j, r_j is the enrichment of ω_j, and $\Lambda = -\sum_j \lambda_{1j} - \lambda_{2j}$ is a constant independent of the observations. Put another way, when building a naive classifier to distinguish between a bin belonging to a TAD boundary or a TAD interior, one should weight the count of each k-mer by the log of its enrichment.

We took the 500 7-mers most enriched in TAD boundaries (enrichment = 1.42 to 1.81), and computed the likelihood score for a collection of 10,670 bins from TAD boundaries and 9,838 bins from TAD interiors. The Receiver Operating Characteristic curve is shown at the end of the document in Figure 5.

At the optimal point on the ROC, the true positive rate is 53% and the false positive rate is 37%, for a difference of 16%. That lift is certainly nontrivial, but hardly a precise classifier.

Was our model too unrealistic? The Poisson assumption is quite good: for each of the top 500 k-mers, the ratio of variance of counts to mean of counts is in the range $(0.8 - 3)$, with most near 1. The sum of independent Poisson distributions is also Poisson, with mean the sum of the means. The distribution of total counts for the top 500 7-mers in TAD boundaries does look Poisson, and the hypothesis cannot be rejected using a chi-squared test.

One simple modification would be to stratify based on GC content, as nucleotide content should impact which sequences are common/important. However, re-running the above analysis on just the bins with similar GC content gives essentially the same picture.

5 Discussion

The DNA sequence is substantially different in TAD boundaries from other regions in the genome. Any p-value computed based on counts of 7-mers in any of the variations of the analysis above is extremely statistically significant. But there is no sequence (at least,

no short sequence) which is present almost exclusively in TAD boundaries. So any use of sequence to distinguish will rely on a combination of many sequence signals. Combining those in a linear fashion is doomed to fail, since any two Poisson distributions with similar means will have substantial overlap. We tried some fancier things (SVM), but those yielded similar results. There may be some higher-order combinatorial phenomenon (some of this sequence followed by some of that transcription factor followed some of this histone modification then another bit of sequence), but it will be quite difficult to detect from data on the 40kb scale.

So while TAD boundaries as an ensemble differ from other parts of the genome, each of the individual factors is noisy and scarce enough that it cannot be used to reliably distinguish TADs. The combination of such rare yet diffuse signals makes for a bad classifier.

In addition to trying to predict boundaries with sequences and epigenetic marks, we were also interested in the reverse: searching for novel DNA-associated proteins that may be related to establishing and/or maintaining TADs given called topological domain boundaries. Since growing k-mers greater than 10bp resulted in insufficient mapping to TAD boundaries, as aforementioned, we took the top 50 9-mers with the highest enrichment values and searched the JASPAR and UniPROBE databases. While our simple queries were not controlled, it is interesting that our best hit (database search E-value of 0.0048) was to RSC3, an active chromatin remodeling protein found in yeast. Polybromo 1 (PBRM1) is a human homolog of RSC3 and is the targeting subunit of the PBAF SWI-SNF chromatin remodeling complex (Reisman et al 2009 Oncogene). PBRM1 may assist chromatin organization relevant to TAD development and serves as a promising candidate for future study.

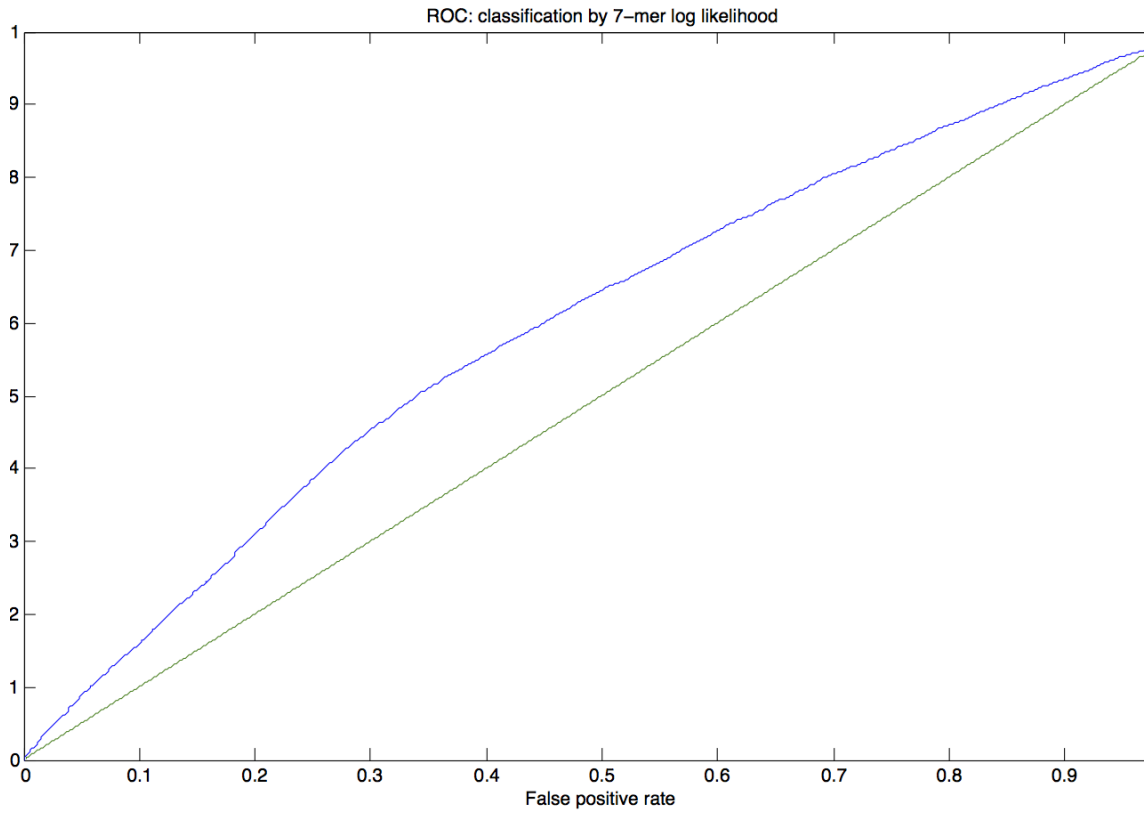

Figure 5: ROC curve of classification by 7-mer likelihood