“Biosample-Agnostic” Prediction of the Impact of Chromatin Presence at CRISPR Binding Locations

# Introduction

I wanted to determine and predict the impact of chromatin presence on CRISPR/Cas9 performance. Problem is: I do not have a chromatin location dataset for any of the cell lines I have CRISPR knockdown data for.

So here I describe my effort to create a “biosample-agnostic” method by aggregating chromatin location data across many known biosamples. It works, but I’d like to revisit this effort using matched biosample CRISPR and chromatin data to really see how much the presence of chromatin matters to CRISPR performance. The work I present here is useful however—when combined with other machine learning methods—for prediction of CRISPR performance, but it is likely way too noisy for robust scientific inferences about chromatin’s impact.

# Method

I first searched ENCODE [1] for the keyword “chromatin” and then downloaded all the resulting “narrowPeak” BED files for hg19. Then combined the BED files by biosample type so that only one BED file is associated with each biosample type.

Next, I iterated through the combined BED files, counting the overlaps per base across the BED files. For example, suppose cell line A and cell line B contain evidence of chromatin presence at position 1000000 on chromosome one, and no other biosample types contained evidence of chromatin presence at that location. In that case I stored the count “2” for position 1000000 for chromosome one.

Once I calculated all the counts, I normalized them between 0 and 127 to enable ASCII encoding for the benefit of quick lookup.

When searching for a particular gRNA binding location’s counts, sometimes multiple hits in the genome occurred. In these cases I used the median count of all the hits since I do not know the intend targets of my CRISPR data set. (More accurately, I’m too lazy to figure it out. Besides, we likely won’t know users’ intended targets in these ambiguous cases).

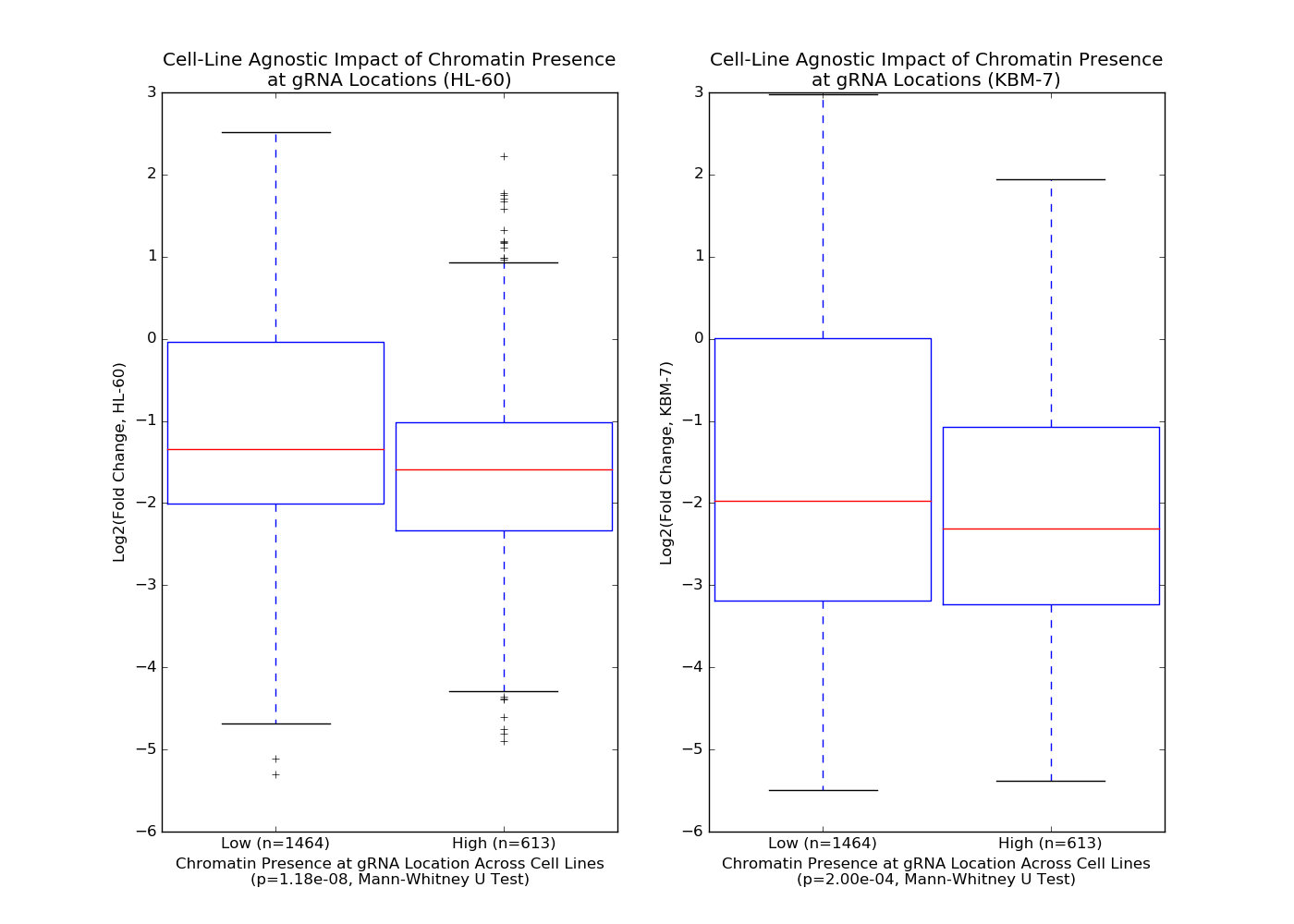
After some statistical juju, I decided that a normalized score cutoff of 40 differentiates the best from the worst cases most effectively.

I also developed a statistical model relating knockdown to counts, counts squared, and counts cubed. Cross-validated to control the risk of overtraining. This can be combined with an efficiency prediction model. Initially, since I am dealing with counts data, I chose negative binomial regression as the framework. Ultimately however, simple polynomial linear regression worked best.

# Results

I matched the human knockdown data for cell lines HL-60 and KBM-7 given by [2] with the chromatin location counts determined using the method described above. As stated earlier, in the case of multiple alignments of a given gRNA to the genome I used the median chromatin location count.

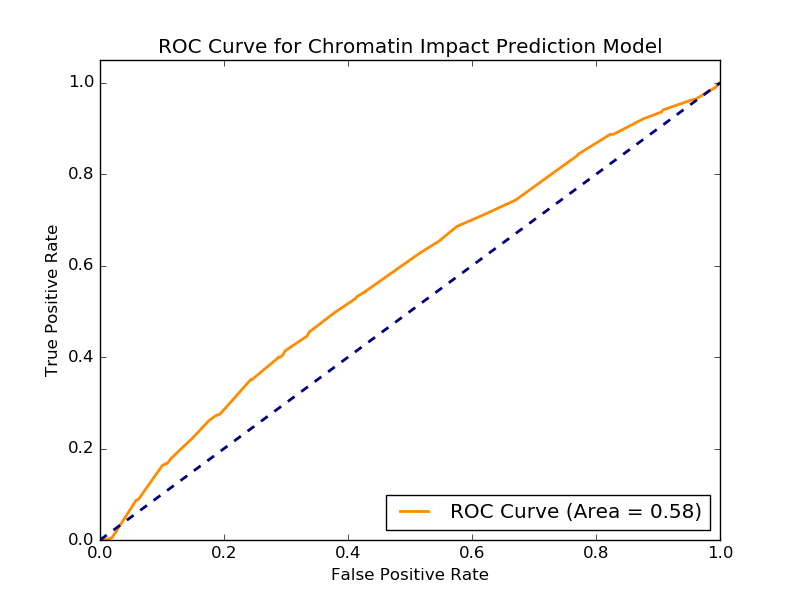
Splitting the counts at a threshold of 40 (“low” chromatin <= 40, “high” chromatin > 40) yields the following division:



There is definite separation in the knockdown results, and definite room for improvement. However, consider the following:

1. We are dealing with an aggregated view of chromatin location across many biosample types. Having chromatin location data for HL-60 and KBM-7 (which we don’t) would be more precise.
2. It may be that chromatin presence doesn’t matter that much—only diminishes CRISPR performance slightly. To find out for certain, I’d like to perform more robust inferences using biosample-matched data.

The predictive model described above proves a weak predictor (the authors of the CRISPR data provide a division between “good” and “poor” gRNAs):



Again, I’m not sure if this is due to chromatin presence not mattering much, or due to the use of aggregated chromatin location information across biosample types.

# Next Steps

First I’ll see if the mouse data contained in [2] has matching ENCODE data for its cell line.

If that doesn’t work, I’d like to conduct an experiment using a cell line with known chromatin locations as per the ENCODE data. Choose predicted high-efficiency CRISPRs that target an equal number of regions with and without measured chromatin presence. I’d have to work out a power calculation to determine appropriate sample size.

# References

1. <https://www.encodeproject.org>
2. <https://www.ncbi.nlm.nih.gov/pubmed/28587596>