Transcriptional regulation is one of the primary methods that organisms employ to control gene expression levels. In eukaryotes, a key contributor to transcriptional regulation is the chromatin architecture; that is, the arrangement of nucleosomes, transcription factors (TFs), and other proteins that bind along the genome at any point in time. One way to predict binding sites is by scanning the sequence for motifs commonly associated with TF binding; this method is fairly sensitive but also prone to a large number of false positives. Thus, to ascertain the *in vivo* chromatin architecture, scientists have developed experimental methods, such as ChIP-seq and DNase-seq, to pinpoint the genomic locations of protein-DNA interactions.

ChIP-seq is considered the “gold standard” for locating transcription factor binding sites (TFBSs). This assay identifies the locations at which a particular TF binds to DNA *in vivo*. However, ChIP-seq requires a separate experiment and antibody for every TF, making the procedure time-consuming and costly. An alternative method, DNase-seq, gauges the accessibility of DNA at specific regions in the genome. In this protocol, the nuclease DNase I makes cuts throughout the genome, and the locations of the cuts are then mapped to a reference genome for analysis. Since regions bound by proteins are less accessible to the DNase enzyme than unbound regions, composite plots of DNase digestion patterns at TFBSs reveal reduced levels of digestion at many TFBSs, creating “footprints” in the data (1). Thus, DNase-seq allows detection of binding sites for multiple TFs using just one experiment.

Recent studies showed that unique and dramatic fluctuations exist within the DNase footprints of specific TFs, often at the resolution of a single base pair, distinguishing any one TF’s footprints from those of other TFs (1). Some recent studies have suggested that these patterns mirror the small-scale biochemical interactions between the bound protein and the DNA molecule (1). If true, this might allow identification of a specific TF’s binding sites based on the DNase digestion pattern alone. However, follow-up studies challenged this claim, demonstrating that most high-resolution fluctuations in DNase digestion are an artifact of DNase’s inherent sequence binding preferences (2). Most TFs only bind certain DNA sequence motifs, and some specific positions within these binding site motifs are cut more often than others due to DNase’s own sequence preferences. In fact, some have suggested that most TF footprints are not even footprints at all, but rather false positives where DNase cut counts are depleted due to an inherently DNase-resistant sequence context (2).

I have analyzed DNase’s sequence biases using experimental data from DNase digestion of naked (protein-free) DNA. I determined the relative “cuttability” of every possible 6-mer nucleotide sequence and used this information to produce corrected TF footprint plots in the budding yeast *S. cerevisiae*. My results confirmed that some TFs had almost no visible footprint after bias correction, but almost half of the 102 TFs I tested still showed visible footprints. For those TFs that still had footprints, correcting the sequence bias produced smoother footprints with a lower variance in cuts at individual positions, but for some TFs, such as MCM1, a clear factor-specific footprint shape was still present even after bias correction. Furthermore, different transcription factors have different footprint widths. Thus, although the factor-specific footprint patterns are not as striking as was originally suggested, it is still quite likely that a statistical model could incorporate this information to infer which specific TF binds at a given location.

I plan to investigate how the chromatin landscape influences transcriptional regulation in eukaryotes, using yeast as a model organism in order to understand basic regulatory principles that might later be extended to human data. My project will explore how differences in this landscape across time and across cell types are correlated with differences in gene regulation. There are two stages to the project. I will first improve upon existing methods for predicting the locations of transcription factor binding sites and identifying the specific transcription factors that bind at each one. I will then apply my model to existing datasets to investigate to what extent changes in the chromatin landscape correlate with changes in gene expression levels.

I have spent over 2 years analyzing data from DNase-seq experiments, with the broad goal of determining transcription factor binding sites. To date, no model exists that explicitly models sequence bias in DNase experiments to fully harness the high resolution of the data. My goal is to develop a novel statistical method that uses DNase data to compute the likelihood that each position is bound by a particular protein. Specifically, I will use a hidden Markov model (HMM) in which the DNase cut counts at each position in the genome are the observed data, and each of the model's "hidden" states represents a specific bound protein, which may be a TF, a nucleosome, or no protein at all. The HMM will include separate parameters for the individual positions within each TF, in order to take full advantage of the high resolution of DNase data. To avoid confounding the "actual" footprints with the sequence biases of DNase, I will include the sequence at each position as an additional predictor of DNase cut counts, using expectation maximization (EM) to simultaneously infer the parameters that govern the bias and the parameters of the underlying footprint patterns. To validate my results, I will compare my predicted binding patterns with published results of ChIP-seq experiments and nucleosome positioning experiments previously performed on the yeast genome.

Once I have developed a reasonable model for inferring the chromatin landscape, I will extend my results to directly investigate how this configuration of proteins affects gene expression. Specifically, I will predict gene transcription rates using features derived from the adjacent protein landscape. Previous attempts to solve this problem have achieved moderate success, but only within a certain subset of test cases (3). Most of these models have used sequence data such as kmer counts as the sole predictors of gene expression (3). One downside of this approach is that these features lack clear biological significance: it is still difficult to understand the indirect process by which a given kmer's appearance in the promoter influences the level of gene expression. In my proposed machine learning model, the features, or predictive variables, will be easily interpretable features that represent attributes of the gene's surrounding protein landscape. These features will include the number and location of TFs and nucleosomes in a gene's promoter region, and I will obtain them using the model that I developed in the first phase of my project. Thus, the two phases form a pipeline: first I will infer the chromatin architecture, and then I will use the model output as predictors for modeling transcription rates.

Developing tools to elucidate chromatin architecture is crucial for understanding gene expression. Differential gene expression drives many critical processes within an organism, allowing cells to adapt to changes in the environment and even become specialized through the process of cell differentiation. Despite our awareness of gene expression's importance to an organism, the processes regulating gene expression are far from fully understood. My proposed research project will bring us closer to understanding these fundamental cellular dynamics.

The DNase-seq data for this project were generated by the Crawford Lab at Duke and the Stam Lab at the University of Washington, and are freely available online.

1) Hesselberth, J. R., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, *6*(4), 283-289.

2) Sung, M. H., et al. (2014). DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. *Molecular Cell*.

3) Meyer, P., et al. (2013). Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach. *Genome Research*, *23*(11), 1928-1937.