Kate Lachance
Biophysics 205: Final Proposal
April 24, 2017

Characterization of transcription factor-mediated pausing of RNA Polymerase II during transcriptional elongation

Introduction

Transcription by RNA Polymerase II (Pol II) is a tightly regulated and highly dynamic process. This complex control has only been recently appreciated with the advent of high throughput sequencing technologies capable of measuring RNA polymerase density genome-wide. Pol II ChIP-seq is can detect where polymerase is bound to DNA, but cannot distinguish actively transcribing Pol II from polymerase that is not engaged. With global run-on sequencing (GRO-seq) and precision run-on sequencing (PRO-seq), transcription proceeds in the presence of labeled nucleotides, so that the location and orientation of transcriptionally engaged polymerases can be specifically mapped across the genome (Core et al., 2008; Kwak et al., 2013). More recently, native elongating transcript sequencing (NET-seq) has been developed to leverage the stability of the RNA-DNA-RNA ternary complex to quantitatively purify all transcriptionally engaged Pol II and to sequence the 3' end of nascent RNA and map Pol II with strand specificity at nucleotide resolution (Churchman and Weissman, 2011; Mayer et al., 2015).

The nucleotide resolution provided by NET-seq is a vast improvement over the approximately 200 base pair (bp) resolution of ChIP-seq and the approximately 50 bp resolution afforded by GRO-seq. Run-on techniques are further limited by the necessity for stalled Pol II to resume transcription in the presence of labeled nucleotides, which varies by experimental condition and individual polymerase (Core et al., 2008; Weber et al., 2014). Pol II ChIP-seq and GRO-seq, despite their lower resolution, enabled the detection of strong transcriptional pauses approximately 50 bp downstream of many transcription start sites (TSSs), elucidating the prevalence of promoter-proximal pausing (Core et al., 2008; Kwak et al., 2013). Much focus has been devoted to elucidating the control of transcription initiation and the role of promoter proximal pausing. Comparatively little is known about how transcription elongation can be regulated, although NET-seq reveals that Pol II pauses throughout elongation as well; throughout well transcribed genes of *E. coli*, there is approximately one regulatory pause site per 100 bp (Larson et al., 2014). While these pauses may be caused by DNA sequence motifs, it has also been suggested that sequence-specific, DNA-bound transcription factors (TFs) could obstruct the elongation of Pol II in human cells (Larson et al., 2014; Mayer et al., 2015).

Genomic sequence and its encoded *cis*-regulatory function can be interpreted through the binding of TFs to their respective DNA motifs. These TF binding sites, and indeed many TFs, were first discovered upon the advent of DNase I footprinting (Galas and Schmitz, 1978; Dynan and Tjian, 1983). DNase I footprinting identifies regions of increased DNA accessibility and therefore increased fragmentation due to nuclease cleavage, termed DNase I hypersensitive sites (DHSs). DHSs often form around TFs bound to the regulatory DNA, as the TF displaces surrounding nucleosomes (Gross and Garrard, 1988); however, the bound TF occludes cleavage by DNase I, leaving nucleotide-resolution footprints indicating TF occupancy (Hesselberth et al., 2009). This technique can be extended genome-wide in human cells with DNase-seq and, in conjunction with previously annotated TF binding motifs, can be used to map TF occupancy for all characterized TFs at nucleotide-resolution across the genome (Neph et al., 2012).

This is a vast improvement over the thousands of ChIP-seq experiments, which each require a highly specific antibody for the target TF, that would be required to obtain a similar view of global TF occupancy in human cells. Using DNase-seq to map TF occupancy, however, does have limitations: DNase I as a nuclease does display sequence-specific bias that curtails the ability of DNase-seq data to discover *de novo* regulatory motifs as originally claimed (He et al., 2014; Neph et al., 2012). Furthermore, the identification of the TFs bound at DHSs are limited to previously annotated TF binding motifs. These motifs are discovered using a combination of universal protein binding microarrays

(PBMs), ChIP-seq, and high throughput SELEX sequencing (HT-SELEX) and are documented in databases such as UniPROBE and JASPAR, as well as in individual publications (Hume et al., 2014; Sandelin et al., 2004; Mathelier et al., 2015; Jolma et al., 2013). Even with these limitations, using DNase-seq to map TF occupancy allows for hundreds of different bound TFs to be accurately mapped across the human genome, representing a vast improvement over the entire contents of the ENCODE database of ChIP-seq data for TFs in the HeLa S3 cell line (The ENCODE Project Consortium, 2012; Sloan et al., 2016).

While DNase-seq can be used to map occupancy of many transcription factors, Mayer et al. (2015) focused their attention on two specific TFs, CTCF and YY1, to study their interaction with transcribing Pol II (**Figure 1**). Both CTCF and YY1 were hypothesized to obstruct elongating Pol II, as CTCF had been implicated in Pol II pausing and YY1 is thought to position +1 nucleosomes (Shukla et al., 2011; Vierstra et al., 2013). Indeed, when NET-seq and DNase-seq signal around CTCF and YY1 binding recognition sites were quantified, there was higher strand-specific Pol II density around the TF binding sites, suggesting that these TFs may induce Pol II pausing (Mayer et al., 2015).

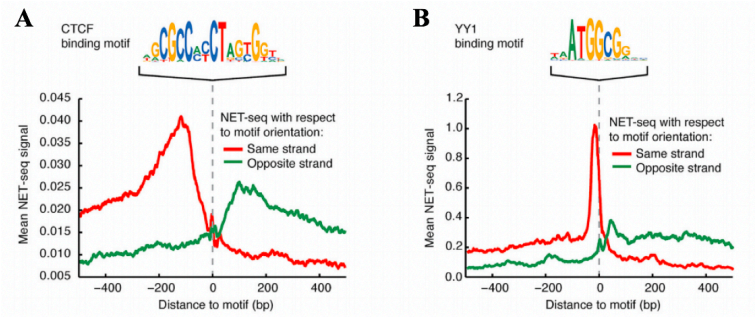

Figure 1. Pol II pausing associated with transcription factor occupancy from Mayer et al. (2015). Average Pol II occupancy around (A) 16,399 CTCF motifs and (B) 731 YY1 motifs located in DHSs.

This suggests an intriguing mechanism for transcriptional regulation through TF-mediated Pol II pausing during elongation. I propose to investigate this hypothesis by first recapitulating the Mayer et al. (2015) analysis before extending it to 500 TFs in the JASPAR database, validating a portion of these findings with TF occupancy data from ChIP-seq experiments deposited in the ENCODE database (Mathelier et al., 2015; Sloan et al., 2016). Next, I will experimentally validate these findings and suggest causation of pausing using an *in vivo* depletion in both *cis* and *trans* of specific TFs. Finally, I plan to categorize TFs based on their effect exerted on elongating Pol II and to investigate other positional causes affecting Pol II pausing around bound TFs, all with the goal of characterizing transcription factor-mediated pausing of Pol II during transcriptional elongation.

Aim 1: Computationally characterize Pol II behavior when transcribing through TF obstacles

The canonical model of sequence-specific TFs involves a TF binding to its DNA motif in the promoter region of a gene and subsequently regulating the expression of that gene. However, only 15.9% of TF binding sites, however, are in gene promoters and a substantial fraction of TF binding sites are found within introns (35.8%) and exons (2.9%) of genes (Stergachis et al., 2013). Elongating Pol II, therefore, often encounters DNA-bound TFs that it must circumvent while maintaining high fidelity throughout transcription.

I propose to investigate these obstacle navigation dynamics in HeLa S3 cells, as this ENCODE cell line has multiple publicly available datasets ripe for analysis and major features transcription regulation in humans remains poorly understood. I will first leverage DNase-seq data to identify DHSs where proteins are bound to DNA across the genome (Neph et al., 2012). As this data is unreliable for *de novo* motif discovery, I will identify TFs bound to DHSs by their known specific DNA binding motif, as catalogued in the JASPAR CORE database (He et al, 2013; Mathelier et al., 2015). In this way, I will annotate sites where there is a DHS identified overlapping a known TF DNA binding motif to create a genome-wide map of occupancy for 500 TFs found in the JASPAR CORE database (**Figure 2**). There are several limitations to this approach for mapping TF occupancy: 1) DNase I is subject to sequence-specific cutting bias, 2) the analysis is limited to TFs with well-characterized DNA binding motifs, and

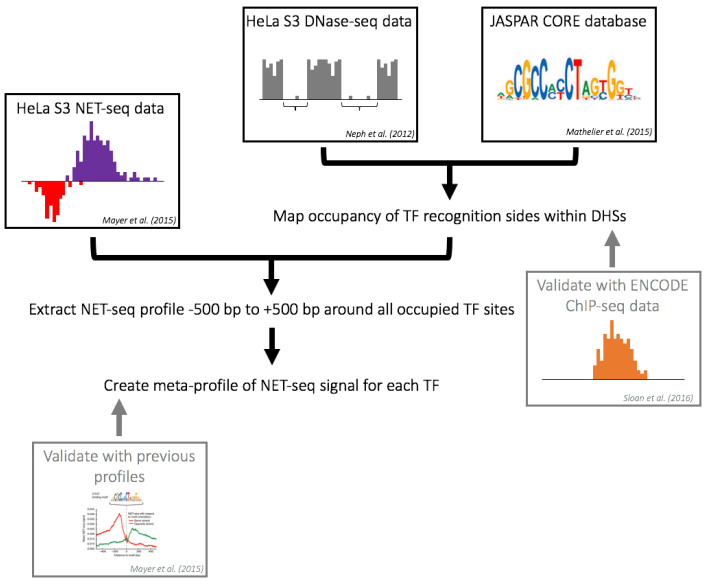

Figure 2. Overview of strategy to use NET-seq and DNase-seq data to create meta-profiles of NET-seq signal (Pol II occupancy) for each transcription factor.

3) there may be ambiguous results for TFs with very similar DNA binding motifs. These same limitations do not apply to ChIP-seq characterization of TF binding; to quantify the sensitivity and specificity of this approach, I will compare the occupancy maps for 20 TFs generated by this approach to the ChIP-seq profiles peaks for these specific TFs in HeLa S3 cells deposited into the ENCODE database (Sloan et al., 2016).

Once TF obstacles that Pol II encounters during transcription have been located, I propose to use publically available NET-seq data to create Pol II density profiles around bound TFs (**Figure 2**). This computational approach can be validated with the published Pol II density profiles around CTCF and YY1, but extends the analysis to characterize the transcriptional profiles around 500 TFs (Mayer et al., 2015). As two negative controls, where I would not expect to see high Pol II density, I would generate similar NET-seq meta-profiles around 1) unoccupied TF binding sites that do not overlap with a DHS and 2) the same number of sites of comparable size to occupied TF sites, but selected randomly from the genome. To better quantify any polymerase pausing observed for each TF, I would define a transcriptional pause as a three standard deviation enrichment in Pol II occupancy from the unoccupied control (Churchman and Weissman, 2011). This definition is standard, but will miss less extreme pausing events that may still have biological significance. I would also quantify the strand specificity of pausing events for each TF by normalizing each strand to a probability distribution and comparing with a Kolmogorov-Smirnov test for statistical difference. I would expect to observe strand specificity with respect to motif orientation for many TFs, as had been observed for both CTCF and YY1 (Mayer et al., 2015).

Aim 2: Experimentally measure Pol II pausing due to TF occupancy by abrogating specific TF binding

While computational analyses may indicate that TF occupancy correlates with Pol II pausing during transcription, correlation is not causation. To support the hypothesis that TFs are an obstacle to Pol II elongation and cause the polymerase to stall, further experimentation is required.

I propose to experimentally probe for Pol II pausing near conditionally occupied and unoccupied TF sites, using an RNAi knockdown system recently utilized by Duarte et al. (2016) to investigate whether GAGA-associated factor (GAF) causes Pol II to stall: I will knockdown specific 10 TFs of interest, chosen for their varied NET-seq meta-profile but all of which are correlated with Pol II pausing, using RNAi. As RNAi has varied efficiency, I would target TFs with known susceptibility to RNAi and confirm knockdown with a western blot in HeLa S3 cells (Cusanovich et al., 2014).

To assess the effect of knocking down a specific TF on Pol II pausing, I would conduct a series of ChIP-qPCR experiments. I would design primers for genes in which, based on the DNase-seq analysis, I expect the TF of interest to bind. Then, I would perform both Pol II ChIP-qPCR and ChIP-qPCR for the TF of interest in TF-RNAi-treated cells and, as a negative control, LacZ-RNAi-treated cells (**Figure 3**). While ChIP-qPCR provides neither the resolution nor whole-genome scale of ChIP-seq or NET-seq, these experiments serve as a cost- and time-effective mechanism by which to assess Pol II behavior in the presence or absence of multiple TFs. The ChIP-qPCR experiments for the TF of interest will confirm that the TF is bound to its expected motif in the control cells and absent from the site in the

experimental cells. As knockdown of these transcription factors is expected to have multiple downstream effects across the genome, I will quantify Pol II pausing near the TF binding site as a fraction of Pol II density in the promoter proximal region. By comparing the relative amount of Pol II pausing near the TF binding site in the experimental conditions, when the TF is bound to the DNA, relative to the control conditions, when the TF is absent, I would uncover evidence that there is an interaction between the TF of interest and Pol II that causes a pause in elongation. These experiments will validate the previous bioinformatic analyses that specific TF binding is correlated with increased proximal Pol II density; these experiments will not, however, definitely prove causation of polymerase pausing nor lend much insight into the mechanism of Pol II pausing. Further limitations of these experiments include: 1) can only evaluate the effect of TFs that are amenable to knockdown by RNAi and have specific antibodies commercially available for western blot and ChIP experiments and 2) the combinatorial effects of TFs that bind proximal to one another cannot be easily untangled. These experiments, however, provide a critical experimental foundation to further exploration of the interaction between TFs bound across the genome and elongating polymerase.

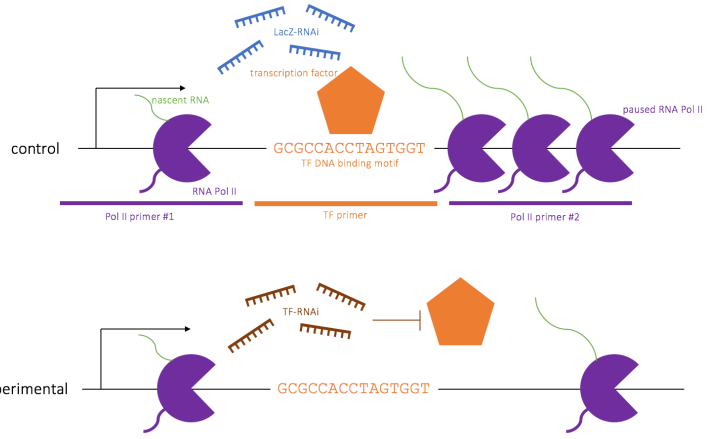

Figure 3. Experimental design for measuring Pol II density due to TF occupancy by abrogating specific TF binding through ChIP-qPCR.

Aim 3: Generalize Pol II behavior around DNA-bound TFs, categorizing by their effect exerted on Pol II and clustering sites based on their position with respect to genomic features

A large portion – some estimate nearly 10% – of human genes encode transcription factors (Babu et al., 2004). Most of these transcription factors, however, can be classified into families based upon common structural features (Mathelier et al., 2016). Because TFs share common structural features, they may also share common interactions with elongating polymerase. If this interaction with Pol II is the mechanism by which a TF exerts its regulatory function, then a classification of TFs based upon their effects on elongating Pol II may be a better representation and extend our understanding of TFs as a class of proteins.

To classify TFs based upon the observed polymerase behavior proximal to their binding sites, I propose to use k-means clustering on the meta-profile of NET-seq signal of each TF (**Figure 4 A & B**). This clustering will yield a classification of TFs based on their common correlation with Pol II pausing. Similarly, hierarchical clustering could be performed to build a phylogeny of TFs, illuminating the relationships between TF families. As a positive control, I would manually create a small set of TFs into a known number of clusters based on visually similar NET-seq meta-profiles before reproducing with computationally clustering. As a negative control, it is clear from previous work that CTCF and YY1 are correlated with very different Pol II density profiles, and should never be clustered together. This novel classification of TFs may be related to the previously annotated structural TF families; a comparison between novel clusters and structural families, if similar, could uncover the structural mechanism by which a group of TFs interact with Pol II.

Thus far, Pol II and TFs have only been considered in isolation; that is, the genomic context in which Pol II and TFs interact has not been taken into account. This context is, however, of vital importance in building a complete model of this interaction and its potential regulatory function. Much of the potential effects of genomic context may have been masked in the NET-seq meta-profiles by combining and averaging all instances of occupancy for a specific TF together. Considering each site

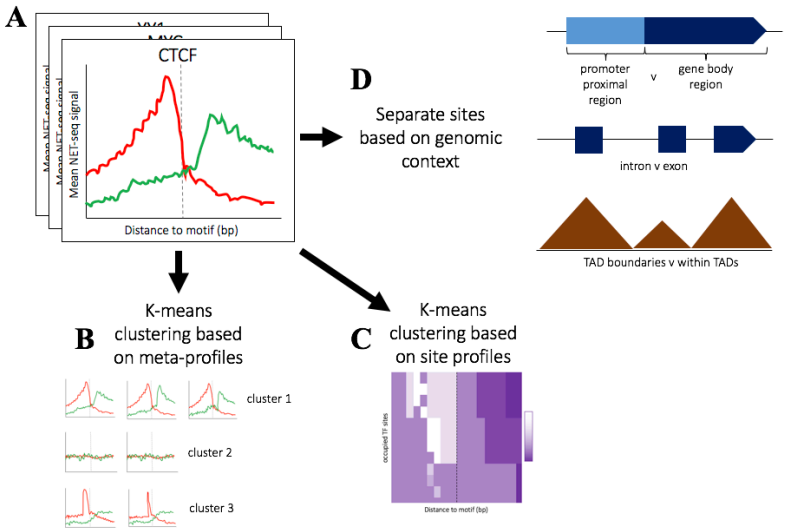

Figure 4. Analysis to generalize Pol II behavior around DNA-bound TFs while considering genomic context. NET-seq meta-profiles for each TF (A) can be clustered to generate a novel, potentially functional classification of TFs (B). These meta-profiles can also be separated into each individual instance of a specific TF bound to DNA, which can then be clustered to reveal the impact of nucleosome positioning (C) or separated based on genomic context (D).

separately could reveal important differences in Pol II behavior correlated to the genomic context of each individual site. Cellular environmental factors to consider include nucleosome position, promoter proximal regions of genes, intron-exon boundaries within genes, and topologically associating domain (TAD) boundaries, among many others.

To investigate the effects of nucleosome position, I propose to perform k-means clustering on the individual NET-seq profile for each occupied TF site of a specific TF (**Figure 4C**). Generating clustered heatmaps that deconstruct the meta-profiles could elucidate clusters of peak Pol II density within the broader pause observed in the meta-profile. One explanation of clusters of separate pauses is that each pause is separated by a single, well-positioned nucleosome; this is probable, as Pol II pausing has been observed upstream of nucleosomes in yeast and *Drosophila* (Churchman and Weissman, 2011; Mavrich et al., 2008; Weber et al., 2014). If nucleosomes are separating pauses, then the distance between pauses should be approximately the amount of DNA that wraps around a single nucleosome, 147 bp. The specific location of nucleosomes could then be confirmed with MNase-seq data, already generated in HeLa S3 cells (Auerbach et al., 2009).

To further characterize pausing effects due to genomic context, I propose to separate instances of TF occupancy specifically by their genomic contexts: whether the TF is bound 1) in the promoter proximal region (within 200 bp downstream of the TSS) or in the rest of the gene body, 2) in an intron or in an exon, and 3) near a TAD boundary or contained within a TAD (**Figure 4D**). After segregating TF sites based on their genomic context, I would regenerate NET-seq meta-profiles and compare between the two conditions using a Kolmogorov-Smirnov test, as before. Of particular interest for this analysis, YY1 is considered a promoter-centric TF, so its NET-seq profile is likely confounded by promoter proximal pausing; separating instances of YY1 near TSSs from those in gene bodies may illuminate separate pausing patterns (Xi et al., 2007). Additionally, recent studies have shown that splicing occurs co-transcriptionally and alternative splicing can be regulated by the kinetics of elongation (Dujardin et al., 2014; Fong et al., 2014); this suggests that there may be a regulatory role for differential pausing between introns and exons that could be uncovered by this analysis. Finally, CTCF is considered a key component of the proposed boundary complex separating and insulating TADs, as defined with Hi-C data (Rao et al., 2014); this analysis could decouple the pausing associated with bound CTCF at these boundaries compared with CTCF bound throughout the genome.

Together, these analyses could suggest a novel mechanism by which TFs interact with Pol II in a specific genomic context to potentially regulate gene expression.

References

Auerbach R.K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrançois, P., Struhl, K., Gerstein, M., and Snyder, M. (2009). Mapping accessible chromatin regions using Sono-seq. PNAS *106*, 14926-14931.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Op. in Struct. Biol. *14*, 283-291.

Churchman, L.S. and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y (2014). The functional consequences of variation in transcription factor binding. PLOS Genetics *10*, 1-13.

Duarte, F.M., Fuda, N.J., Mahat, D.B., Core, L.J., Guertin, M.J., and Lis, J.T. (2016). Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. Genes Dev. *30*, 1731-1746.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Munoz, M.J., Alló, M., et al. (2013). Transcriptional elongation and alternative splicing. Biochem. Biophys. Acta *1829*, 134–140.

Dynan, W.S. and Tjian, R. (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell *35*, 79-87.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.D., and Bentley, D.L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. *28*, 2663–2676.

Galas, D.J. and Schmitz, A. (1978). DNase footprinting: a simple method for the detection of protein-DNA binding specificity. Nuc. Acid Res. *5*, 3157-3170.

Gross, D.S. and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. *57*, 159-197.

* He, H.H., Meyer, C.A., Hu, S.S., Chen, M., Zang, C., Liu, Y., Rao, P.K., Fei, T., Xu, H., Long, H., Liu, X.S., and Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat. Methods *11*, 73-78.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., Fields, S., and Stamatoyannopoulos, J.A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods *6*, 283-289.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2014). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nuc. Acid Res. *43*, 117-122.

Jolma, A., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327-339.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.

* Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., and Weissman, J.S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science *344*, 1042-1047.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2015). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nuc. Acid Res. *44*, 110-115.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., et al. (2008). Nucleosome organization in the Drosophila genome. Nat. Genet. *39*, 1507-1511.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

* Neph, S., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Lieberman-Aiden, E. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1-16.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nuc. Acid Res. *32*, 91-94.

Shukla, S., Kavek, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Shandberg, R., and Oberdoerffer, S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74-79.

Sloan, C.A., et al. (2016). ENCODE data at the ENCODE portal. Nuc. Acid Res. *44*, 726-732.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M., and Stamatoyannopoulos, J.A. (2013). Exonic transcription factor binding directs codon choice and impacts protein evolution. Science *342*, 1367-1372.

Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J.A. (2014). Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat. Methods *11*, 66-72.

Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol. Cell *53*, 819-830.

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., and Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res. *17*, 798-806.

* paper discussed in class

Contributions

Heather Landry (PhD student, Churchman Lab) and Blake Tye (PhD student, Churchman Lab) both helped to refine the experimental validation protocol by discussing previous approaches as well as experimental feasibility.