Fundamental insights into gene regulation from genomic analyses: past successes and future challenges

Karen Adelman

NIH National Institute of Environmental Health Sciences

Gaps in our knowledge of genome regulation

- We need to better define what 'regulatory element' means both for genome annotation and functional studies
  - Not all enhancers or silencers are the same
  - Many intergenic regions are transcribed. Is this transcription meaningful? Is it regulated? Is it regulatory? How?
- Understanding how regulatory elements work will require the use of new techniques and read-outs
  - Move beyond descriptions of chromatin 'state' to more direct measurements of gene expression and non-coding RNA production
  - Not just RNA-seq. We can now study nascent RNA species. This gives direct insights into **transcription**- even of unstable RNAs
  - Not just to understand basic biology. Nascent RNA assays that can simultaneously, quantitatively, globally measure RNA production at promoters and enhancers could yield new insights into how sequence variants work

How to identify new regulatory elements? How to classify them and understand their function?

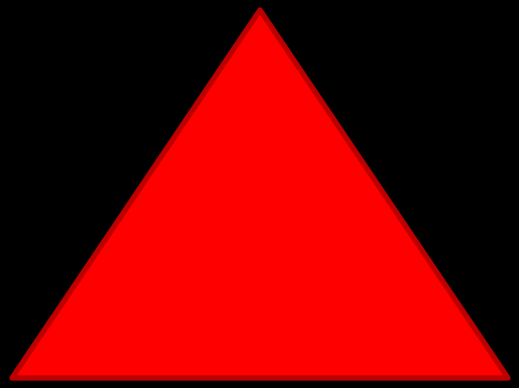
- Need to use assays that are sensitive measures of which gene are affected by regulatory elements – ideally not just the direct target in local region, but also the compendium of other genes that change along with target gene. This can begin to tease apart gene networks.
  - Should define both downstream targets and upstream regulators (like TFs, co-activators, etc.)
  - Nascent RNA methods can efficiently determine if effect occurs at level of initiation, elongation, RNA processing, stability effects
  - Can precisely define the transcribed region- hidden ORFs?
- Assays need to be high throughput, require a small number of cells, and ideally yield meaningful data with low sequencing depth to enable multiplex evaluation of multiple deletions/SNPs, cell types, conditions, time points

Let's look more closely at RNA!

Investigating nascent and newly produced RNA species is an efficient and direct way to define regulatory elements and elucidate function

Previous successes in this arena:
genomic analysis has already greatly improved our understanding of transcriptional mechanisms and
distal control of gene expression

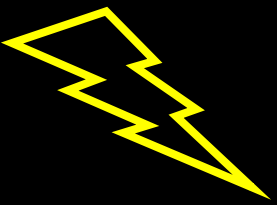
Fundamental biological questions

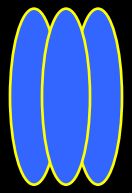


Technique development

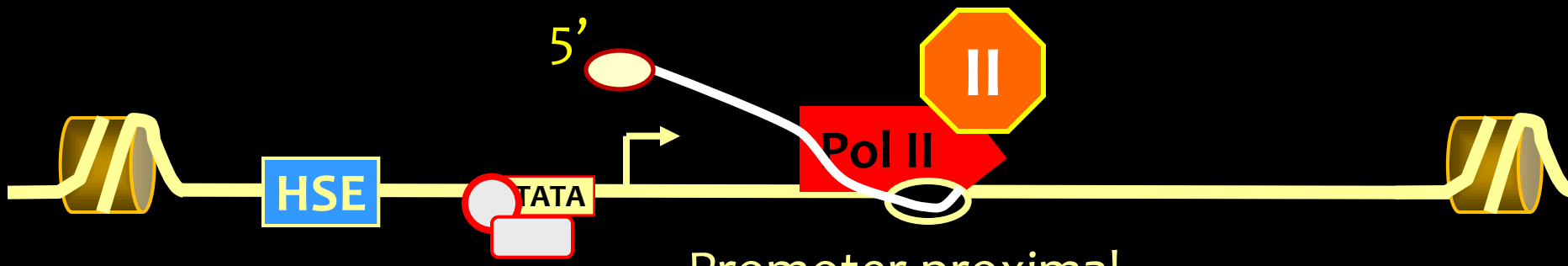
Genomic data collection

Promoter proximal pausing of RNA polymerase II at the heat shock genes in *Drosophila*





- Pol II pauses 20-60 nt into gene
- Pol II is released into productive elongation rapidly during heat shock

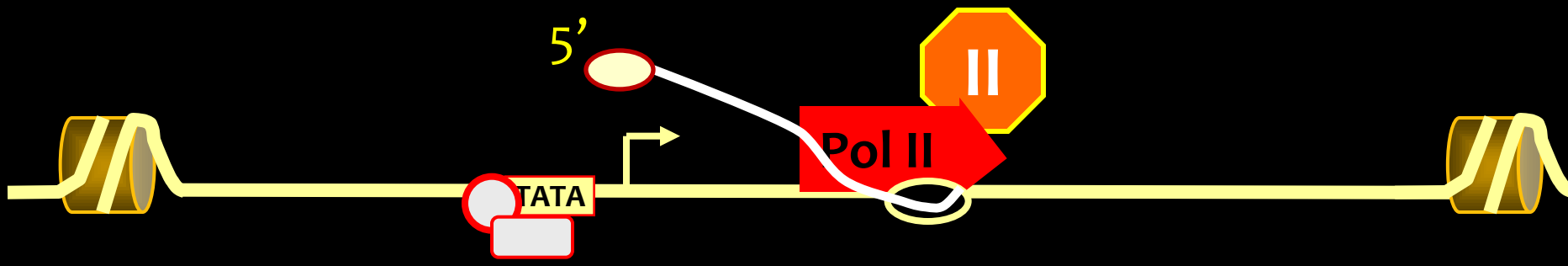


Promoter-proximal pausing

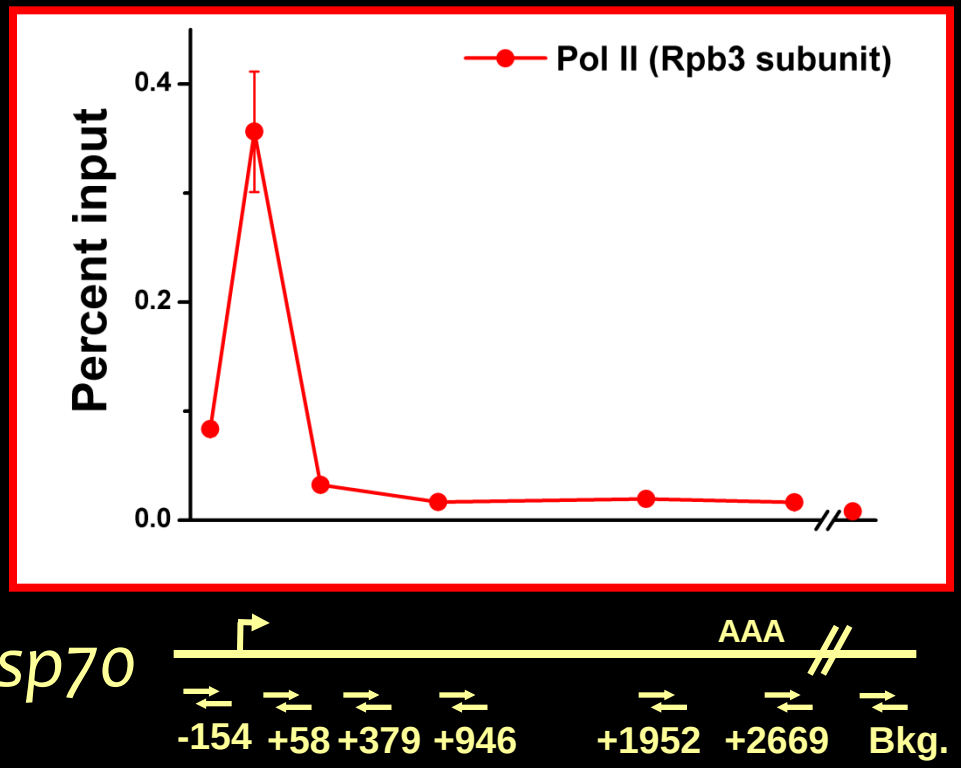
John Lis

A few other genes were shown to be regulated during early elongation

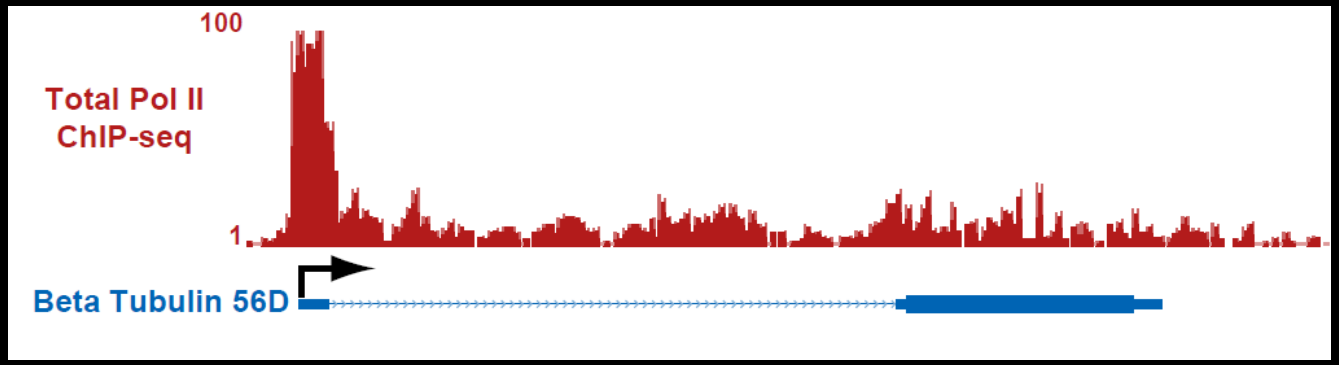
- Mammalian *c-myc*, *Fos*, *junB* (Groudine, Eick, Handa)
- HIV LTR (Peterlin)
- However, before the genomics era, pausing was considered to be a rare gene regulatory mechanism

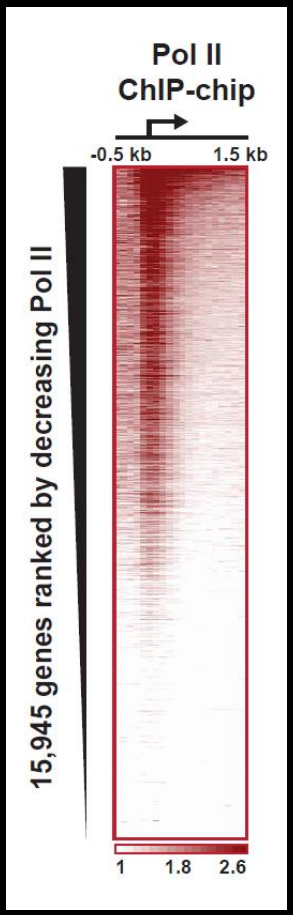


ChIP reveals Pol II in the promoter region of heat shock genes prior to induction



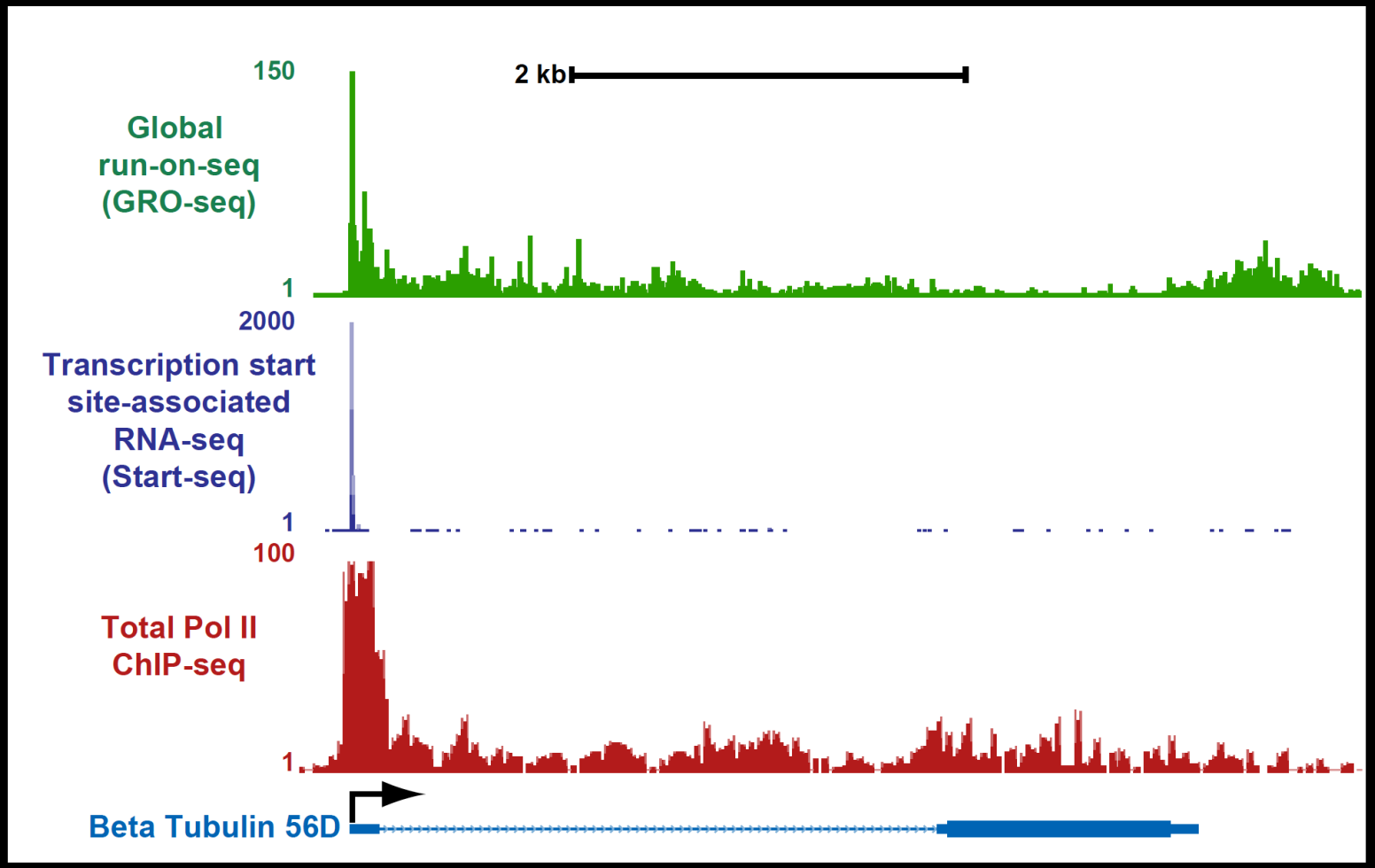
Pausing of RNA polymerase II is a widespread form of gene regulation in metazoa





Muse et al. (2007) *Nat. Genet.*; Zeitlinger et al. (2007) *Nat. Genet.*

These assays confirmed that the promoter-associated Pol II observed by ChIP is engaged in early elongation



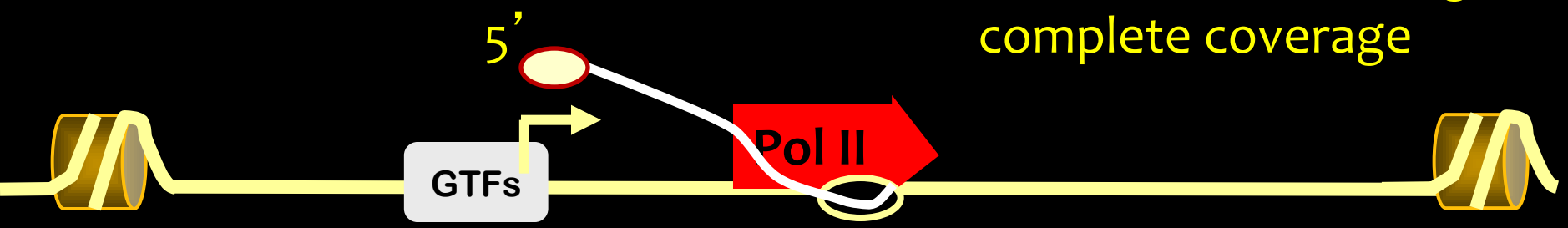
Core et al. (2008) *Science*; Nechaev et al. (2010) *Science*

Defining the 5'-end of nascent RNA with precision gives unique insights into coding and non-coding RNAs

Start-seq involves isolation and high-throughput sequencing of short transcription start site-associated RNAs

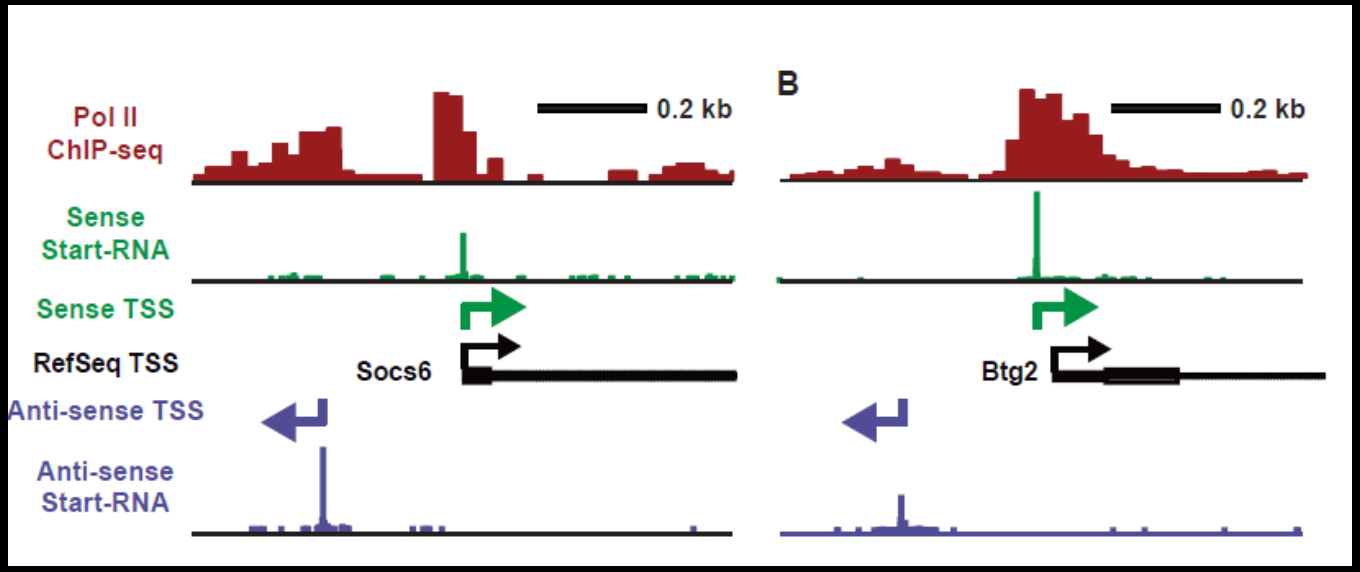
- Nuclear
- Short (<120 nt)
- Capped (when >25 nt)

Reads are all focused right at TSS (single nucleotide), so fewer reads are needed to get complete coverage

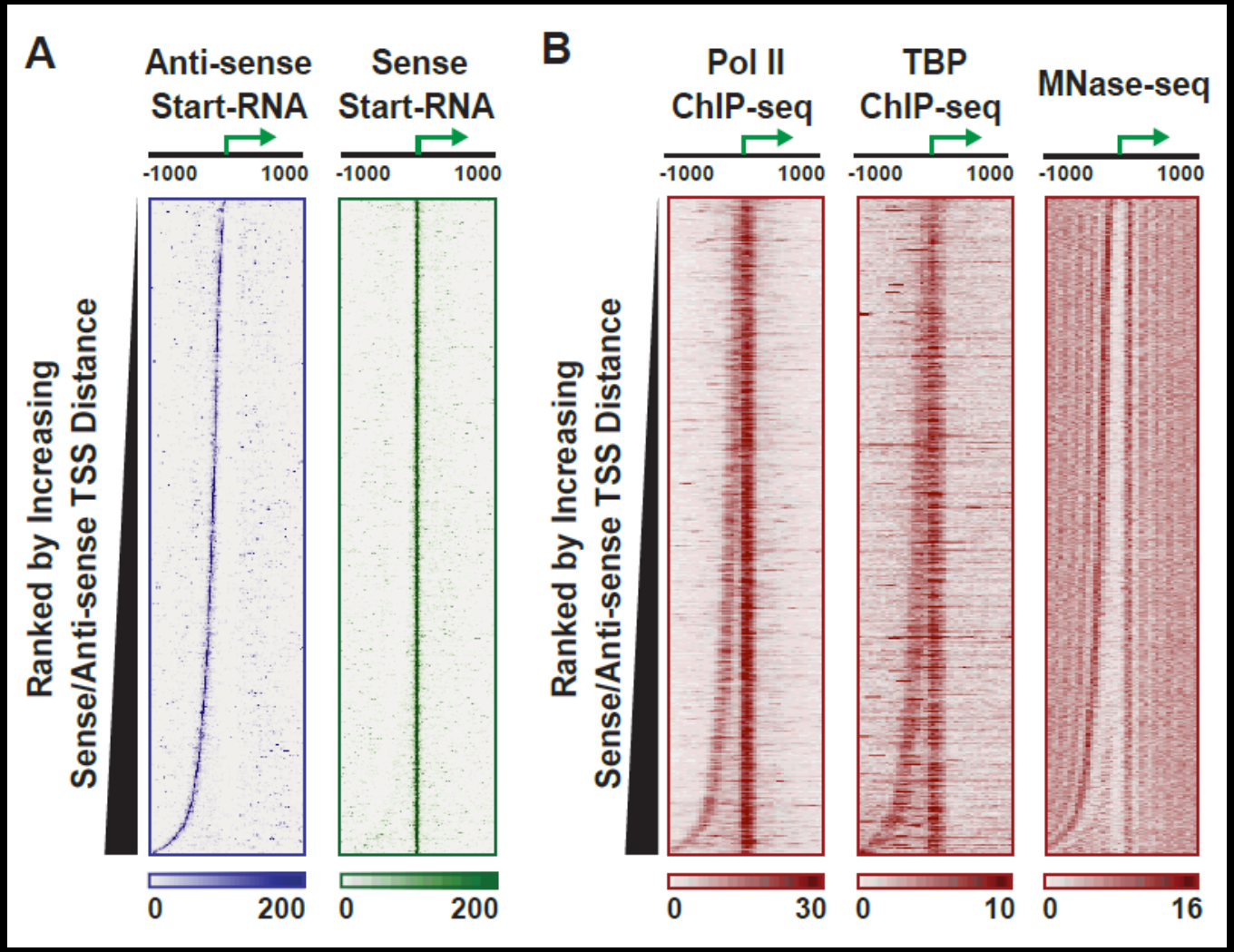


Will detect RNA if Pol II is paused, productively transcribing, arrested, terminating, etc.

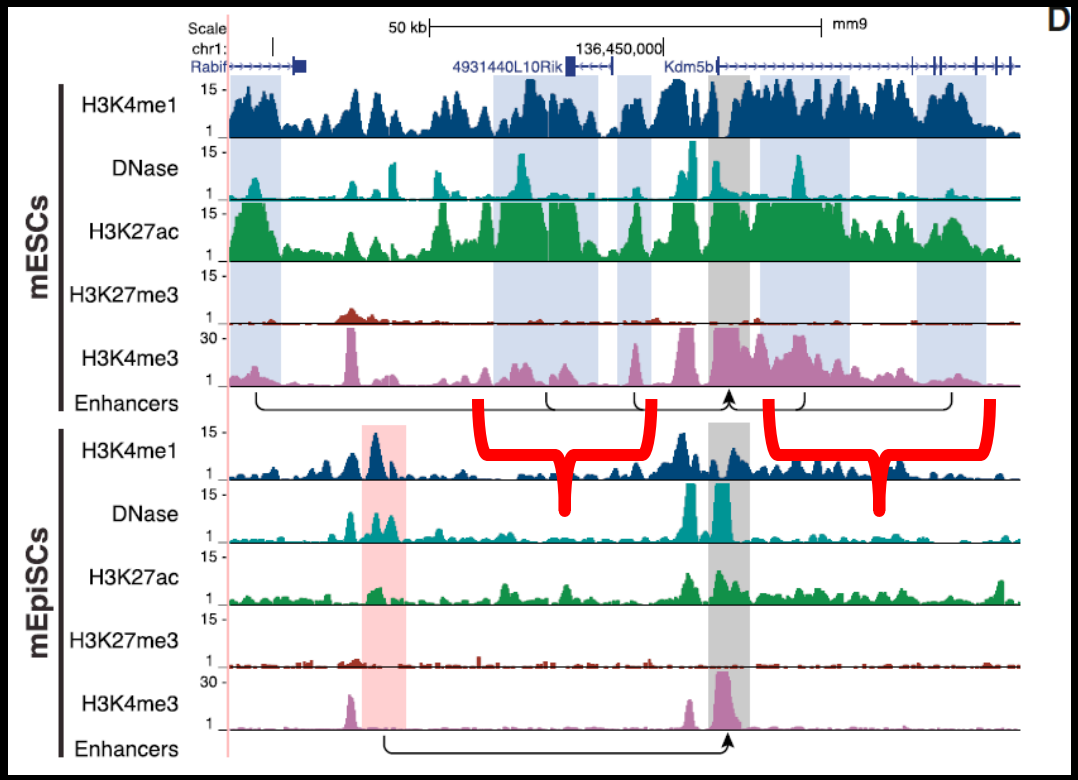
Start-RNAs and 5'-GRO-seq are sensitive measures of coding and non-coding transcription



Start-RNAs and 5'-GRO-seq are sensitive measures of coding and non-coding transcription



Enhancer regions are often described as broad regions…making it difficult to localize functional parts



Where is the action?

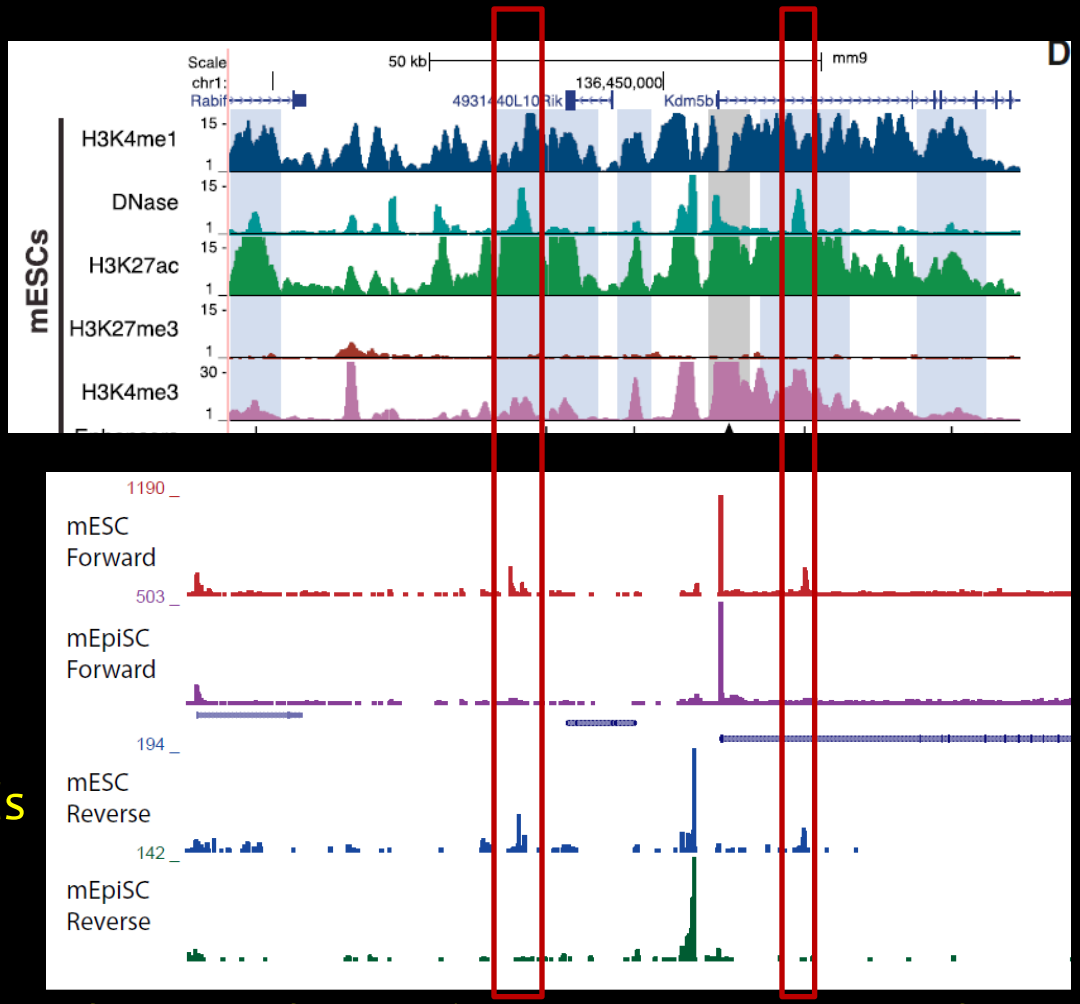
How do we define enhancers?

Are all DNase sensitive regions of H3K4me1 and K27ac the same?

What do these histone marks mean? What do they actually do?

Factor et al. (2014) Cell Stem Cell

Enhancer regions are often transcribed at a level that correlates with activity



one lane
on a miSeq
~15 million reads

Nascent RNA readouts readily reveal changes in enhancer activity in altered contexts

High-resolution definition of RNA 5'-ends localizes hubs of TF binding in region and includes (in same data set) level of transcription activity at mRNAs

Capitalizing on new technologies

- Technique development was driven by specific biological question and goal – e.g. to better measure early elongation complexes
- But the resulting techniques (GRO-seq, Start-seq, 4sU-seq) have broad appeal for agnostic data collection- they are sensitive, high-resolution measure of nascent RNA synthesis, ability to pinpoint enhancer RNA production, and give insights into RNA processing
- Efficiency allows for a number of contexts to be evaluated- dynamics!

This opens up a lot of new possibilities!

- NHGRI should both continue to support technology development and be quick to implement new techniques that change our views on genome structure and function
  - NHGRI can invest in newly emerging techniques, facilitating bringing them to fruition or making them generally available
  - Fund smaller projects to push techniques forward and then maintain flexibility in larger consortia to integrate new approaches

Summary

1) Why perform a comprehensive investigation of nascent and newly-transcribed coding and non-coding RNAs (start-RNA, GRO-seq, 4sU-seq) across multiple cell types and patient-derived cells?

- Efficient means to characterize cell-type specific regulatory regions
- Precisely identifies core of enhancer regions, extent of non-coding transcription
- Provides direct measures of gene activity, RNA processing and stability

2) Why NHGRI? Has experience and success with large scale initiatives, developing and enforcing quality standards, and fostering structured, collaborative projects

3) Transformative technological breakthroughs: improvements in single-cell sequencing (RNA recovery and sampling), development of robust, reproducible spikes for ChIP-seq to provide **quantitative** information about signals

4) Unbiased data generation that proves helpful: Higher resolution localization data, expansion into primary cells, perturbations to better define function

5) Optimal organizational structure: Some projects are best as looser, less top-down collaborations between wet labs and informatics groups. Needs structure through NHGRI to define protocols, standards, priorities (Cell types, assays, etc.) , but might benefit from flexibility over the course of the project rather than being locked into a specific protocol or regime.