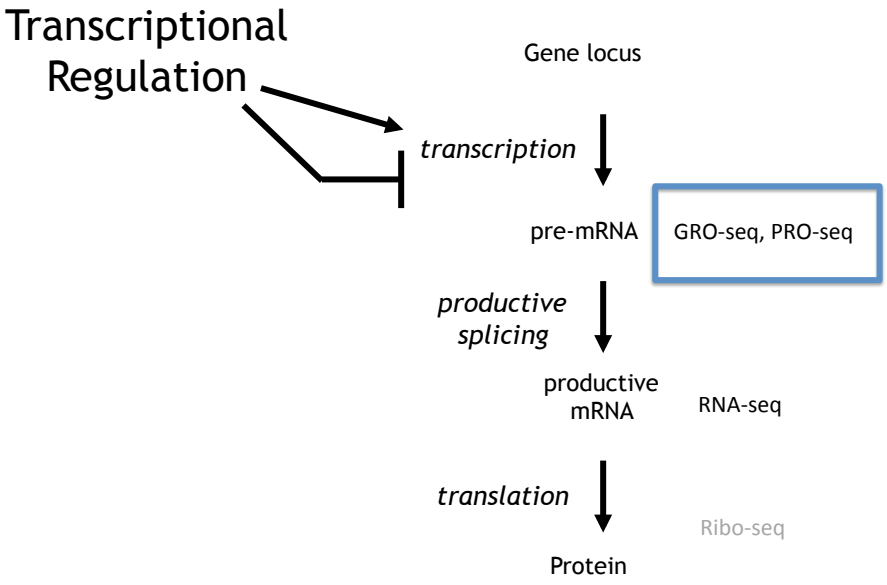
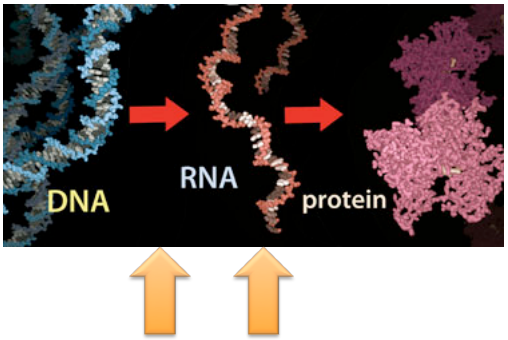
"Don't lower your expectations to meet your performance. Raise your level of performance to meet your expectations. Expect the best of yourself, and then do what is necessary to make it a reality."
--- Ralph Marston



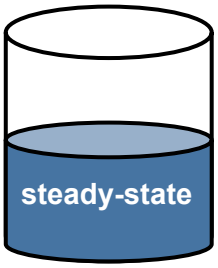
Despite years of expression studies, we have NOT been assaying transcription…



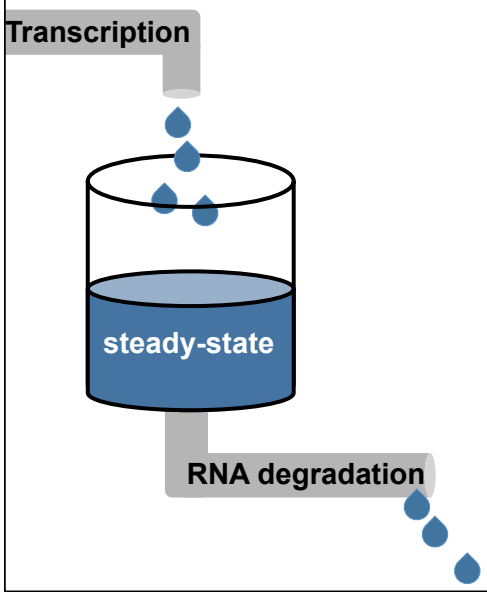
Nascent Transcription

Steady state RNA = Expression studies

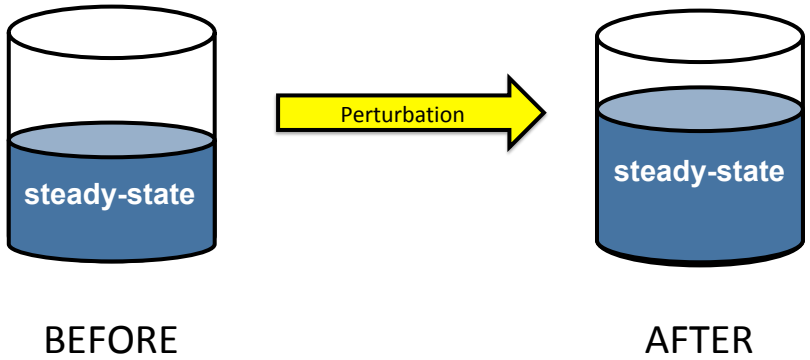
RNA-seq is a measure of steady state RNA



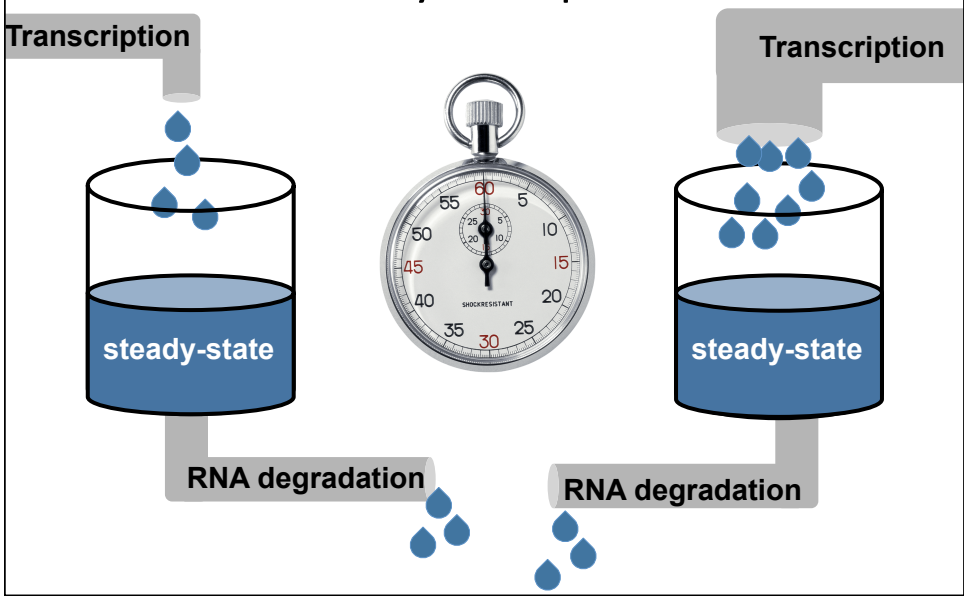
RNA-seq is a measure of steady state RNA



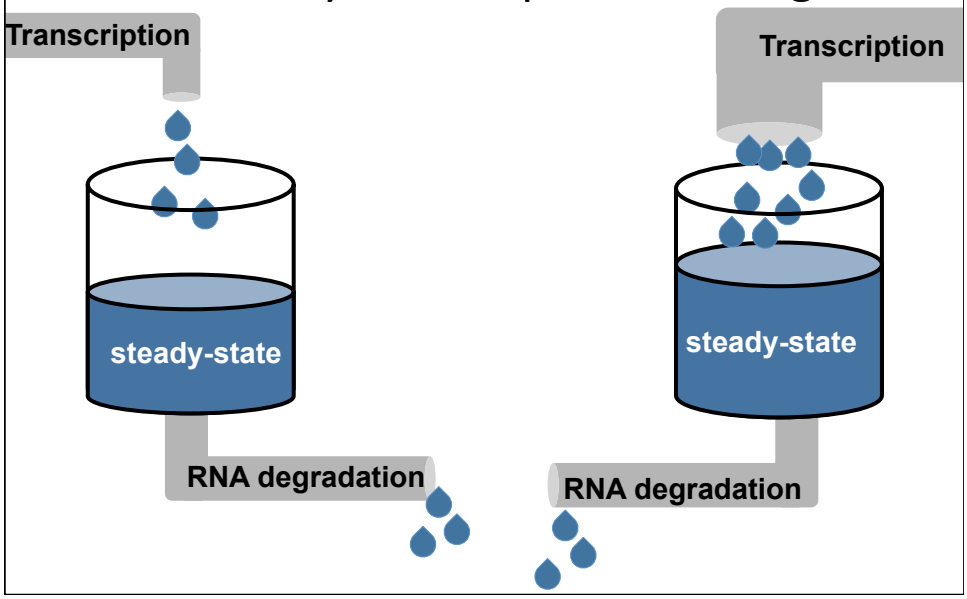
Typical RNA-seq experiment



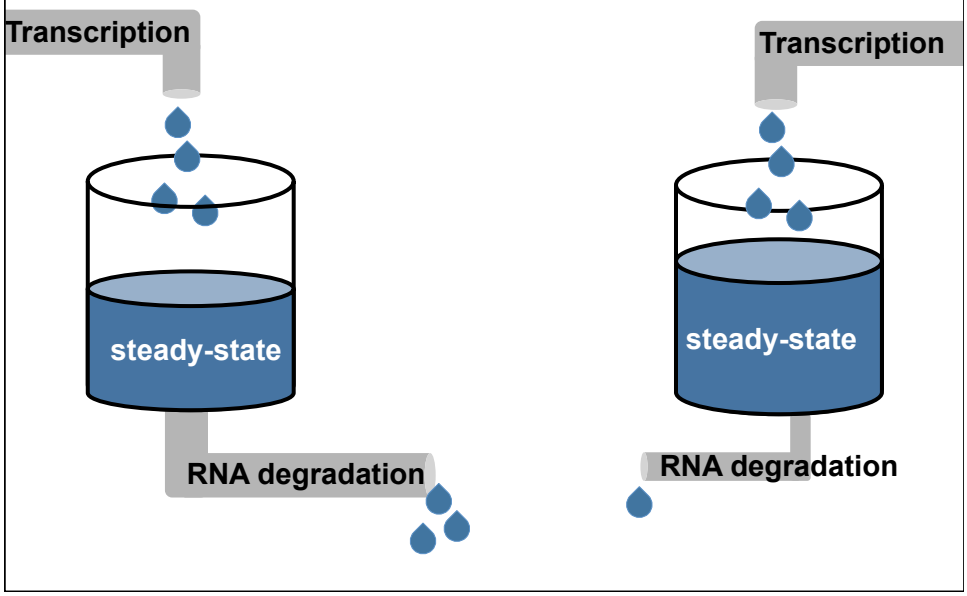
Alterations can not always be detected at early time points



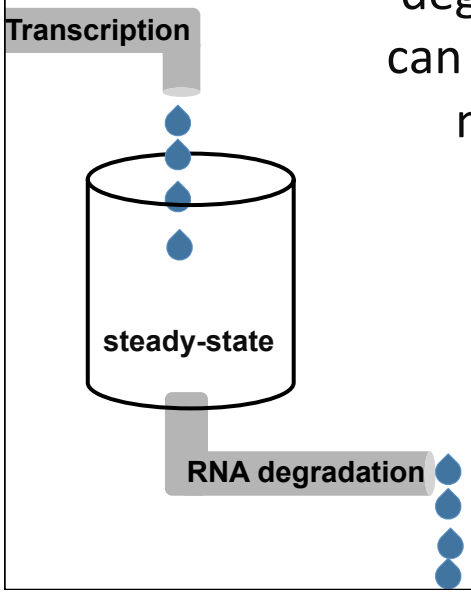
Changes in steady state are not necessarily transcriptional changes



Changes in steady state are not necessarily transcriptional changes

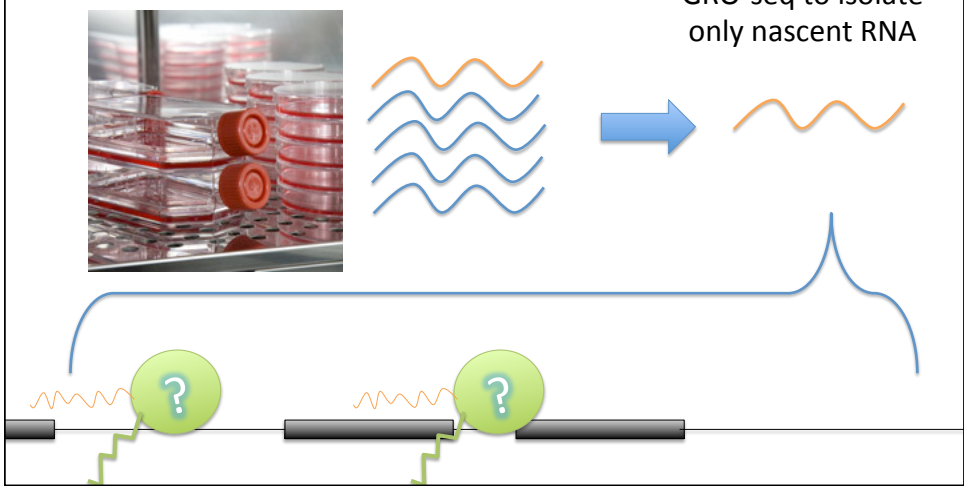


Furthermore, degradation of an RNA can occur so quickly it is not detectable by RNA-seq



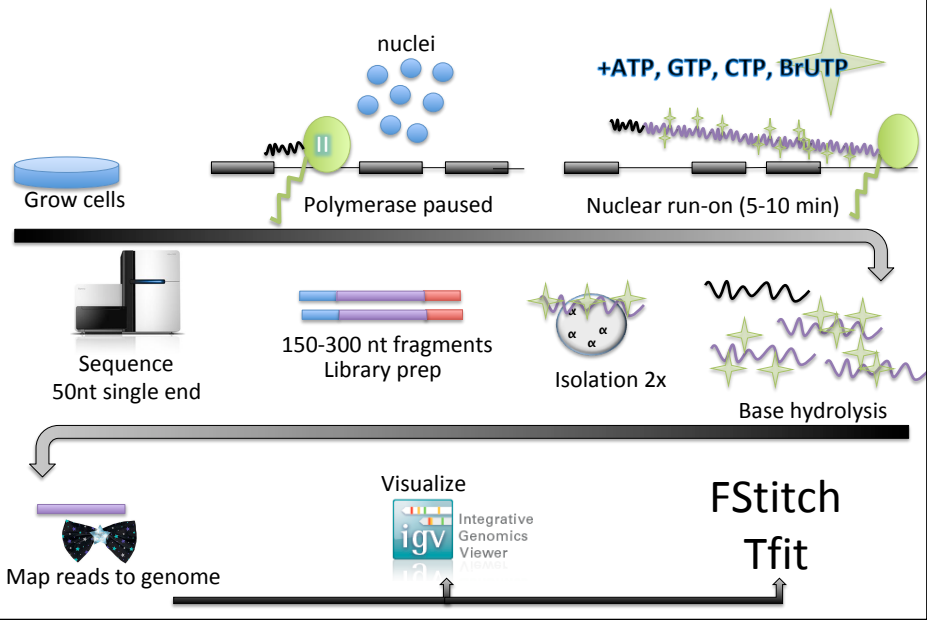
GRO-seq: A picture of the transcriptional landscape

Global Nuclear Run On Sequencing

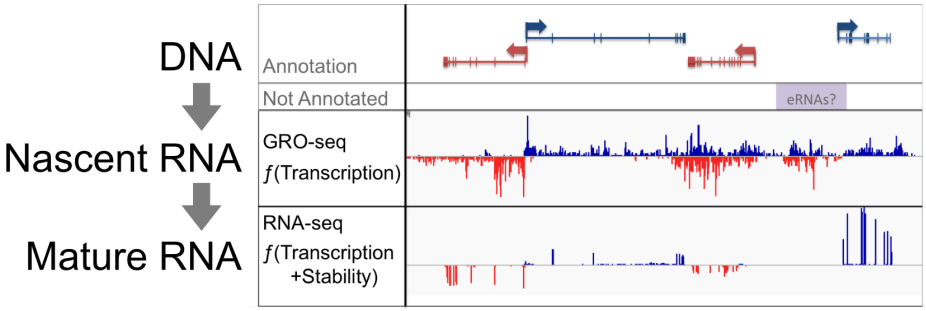


Global Nuclear Run On Sequencing

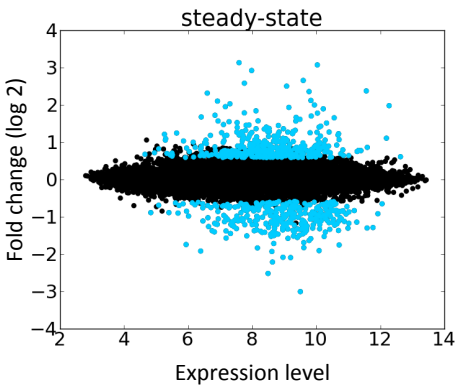
GRO-seq: isolating and sequencing nascent RNA

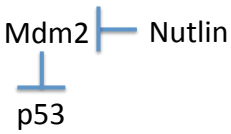


Nascent transcription paints a very different view of what's going on transcriptionally from RNA-seq



Unclear from expression studies what are direct versus indirect effects

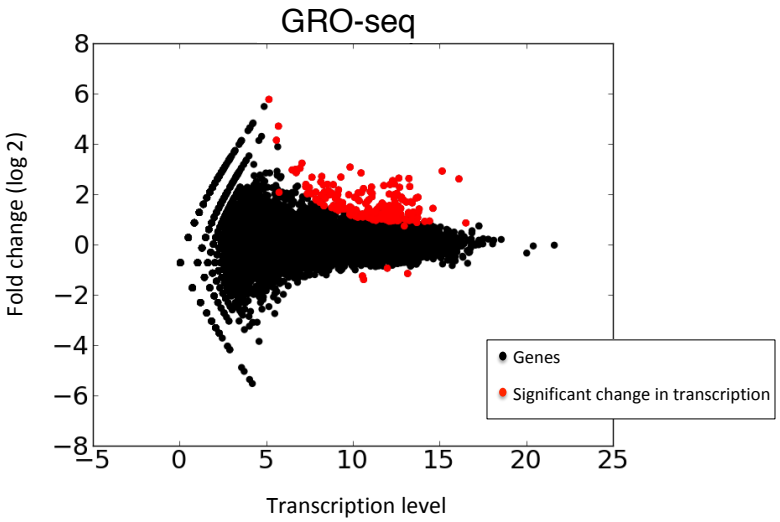




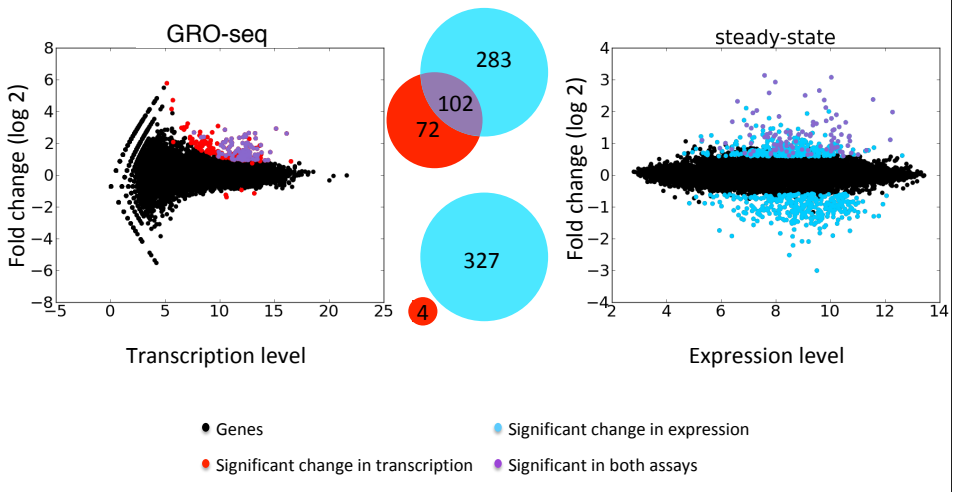
Genes

Significant change in expression

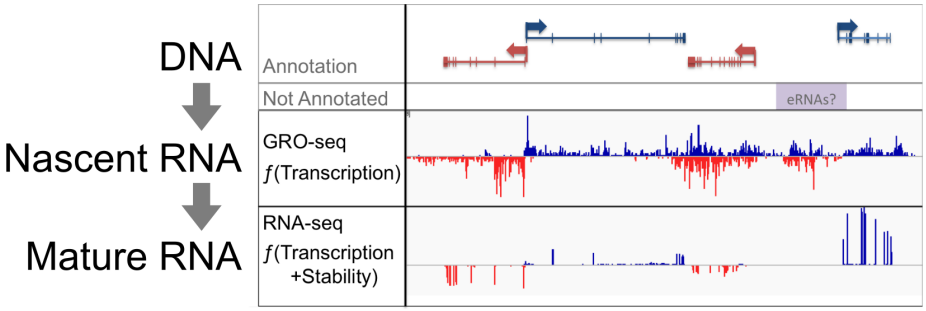
p53 appears to be solely an activator



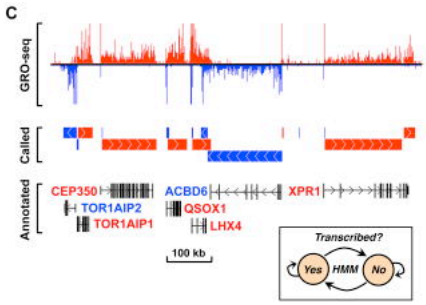
Different views of p53 function through different techniques



But looking only at annotated genes precludes one from “discovery”.

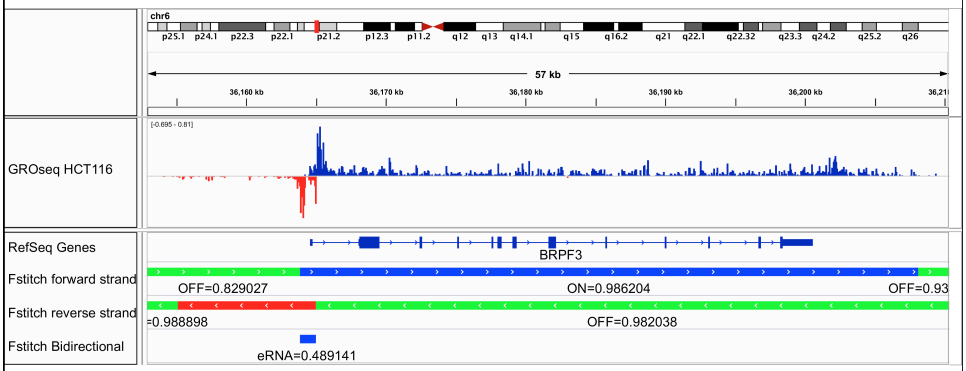


First approach to GRO-seq analysis was a simple two state HMM



- Ran independently on each strand.
- Trained using the known gene annotations.
- Emissions are yes/no of having read coverage.

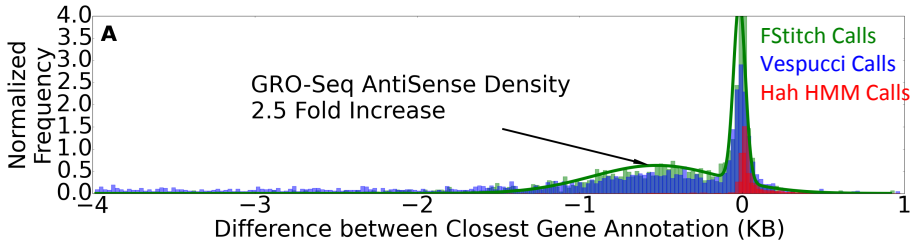
A slightly more sophisticated HMM (FStitch)



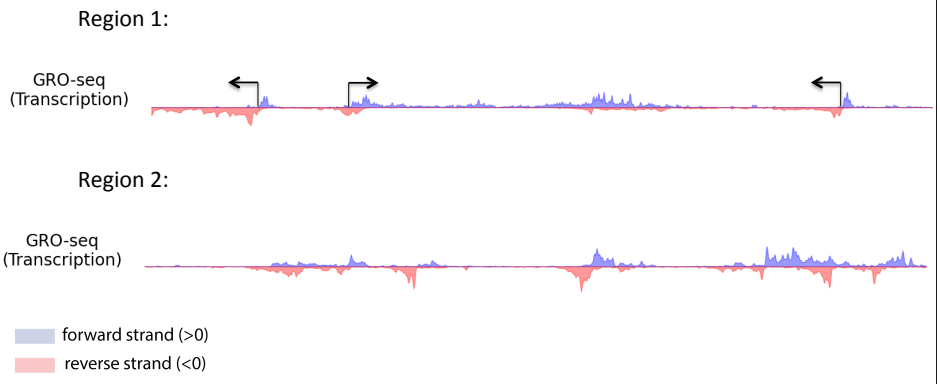
- Trained on manually developed training files.
- Emissions a "contig" representation of read data.

Azofeifa 2014; Azofeifa TCBB 2017

Alternative training methods drastically impact what you can learn from the data.

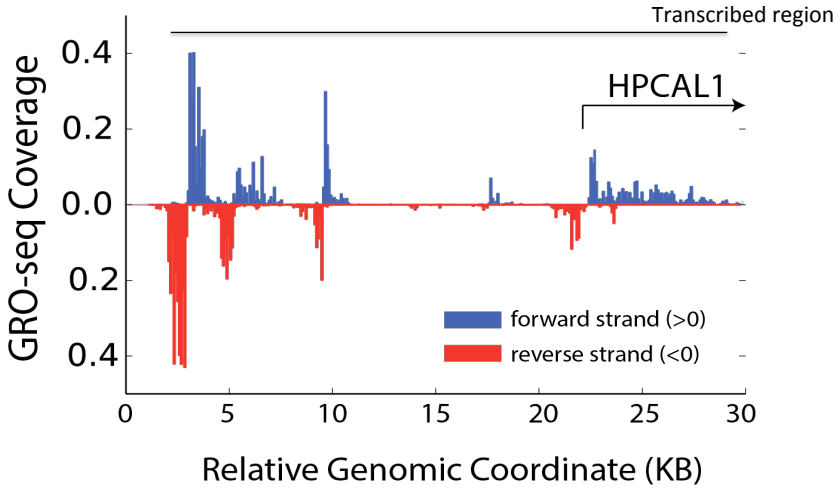


Relying on annotation even for training limits novel discovery



Therefore new algorithms are needed for the unique properties of nascent transcription

Transcribed regions appear to have substructure



Three Classes of Transcription in Eukaryotes

RNA polymerase I (pol I)

ribosomal RNAs (5.8S, 18S, 28S rRNA)

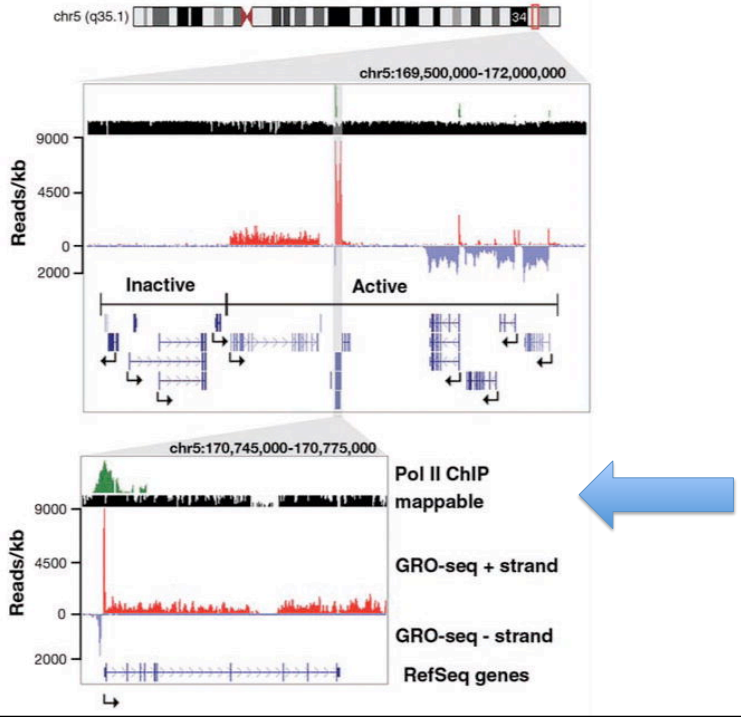
RNA polymerase II (pol II)

mRNAs
some small nuclear RNAs (snRNAs)
non-coding RNAs (mostly of unknown function)

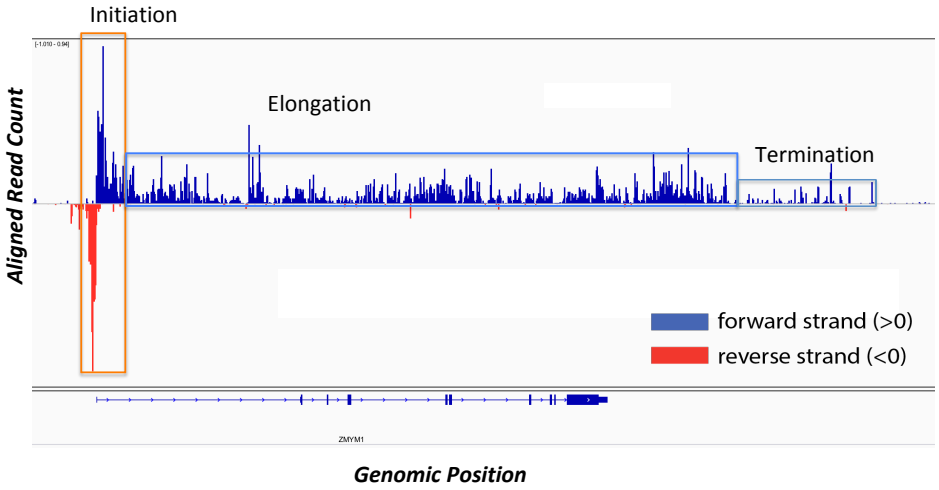
RNA polymerase III (pol III)

tRNAs
5S RNA
some snRNAs
small cytoplasmic RNAs (scRNAs)

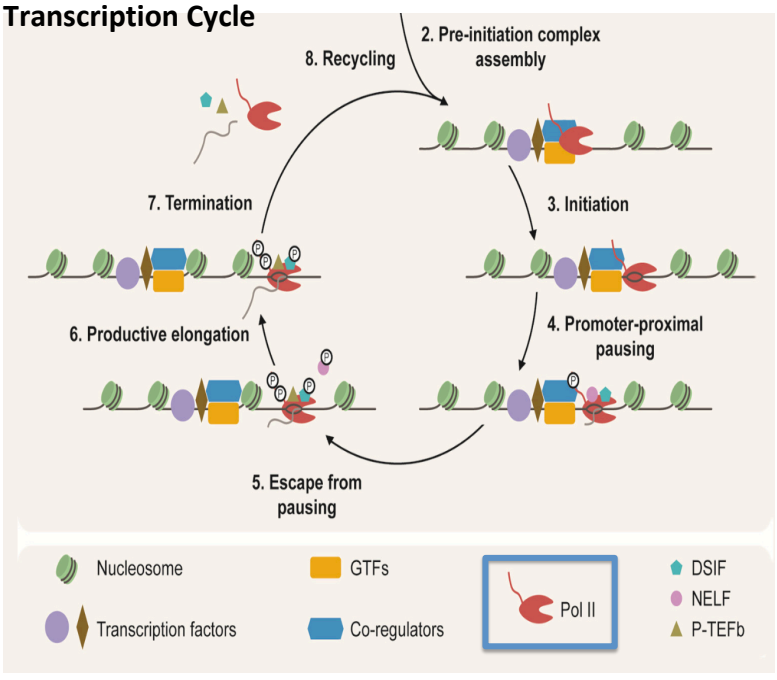
Nascent transcription tag the newly synthesized RNA – hence they obtain reads from ALL THREE classes of polymerase.



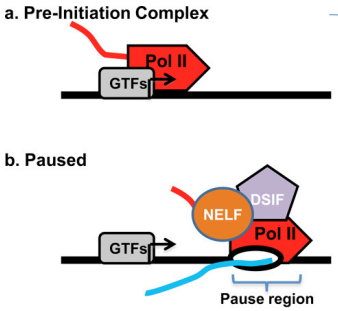
At genes, there are distinct patterns.



Pol II Transcription Cycle

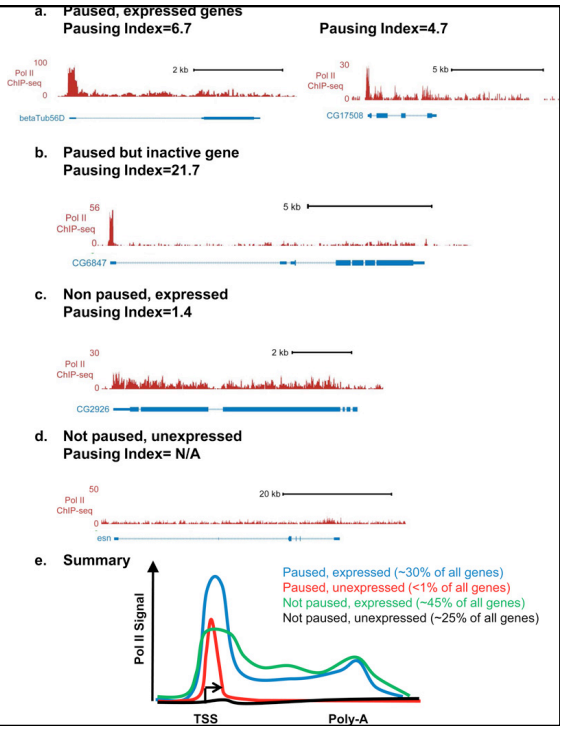


The shift from initiation to elongation can be a regulated event.

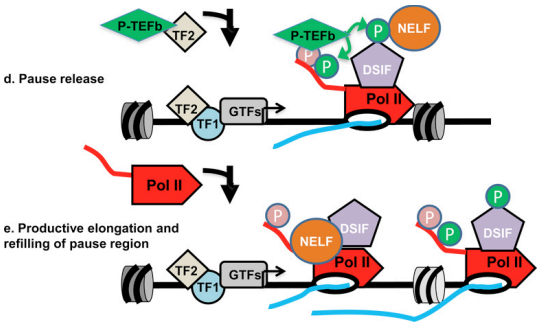


- **Release from pausing** can be the mechanism for induction of expression.
  - In *Drosophila*, the RNA polymerase can **pause** after synthesizing ~ 25 nucleotides of RNA in many genes.
  - under elevated temperature conditions, the **heat shock** factor stimulates **elongation** by release from pausing.
- This is in addition to regulation at initiation.

Predominant initial interest in GRO-seq: Pausing Ratios



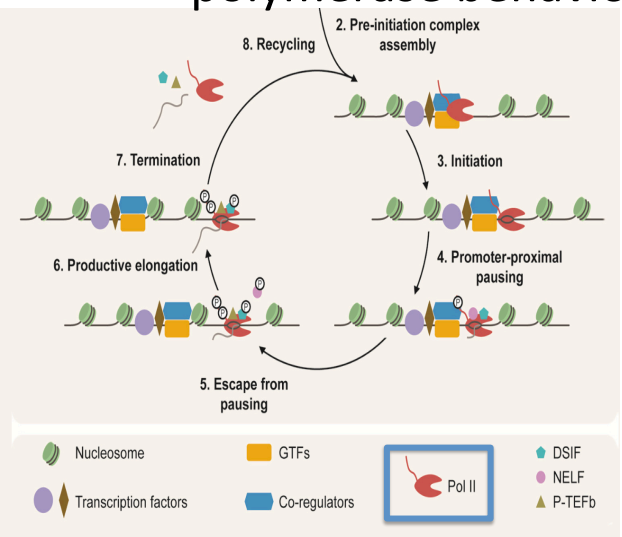
Elongation phase of transcription



- Requires the release of RNA polymerase from the initiation complex
- Highly processive

- Dissociation of factors needed specifically at initiation.
- Eukaryotic TFIID and TFIIA appear to stay behind at the promoter after polymerase and other factors leave the initiation complex

Developed a mathematical model of polymerase behavior



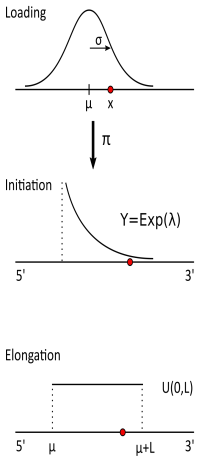
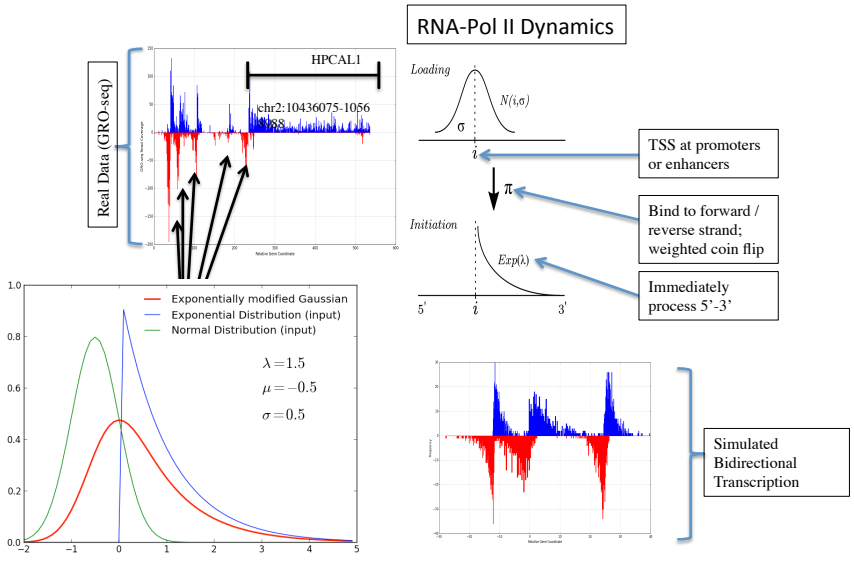
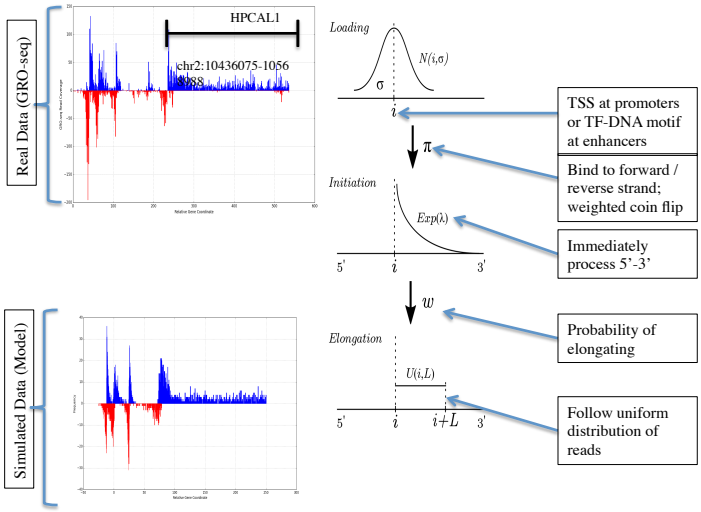


Image: Adapted from Fuda et. al (2009)

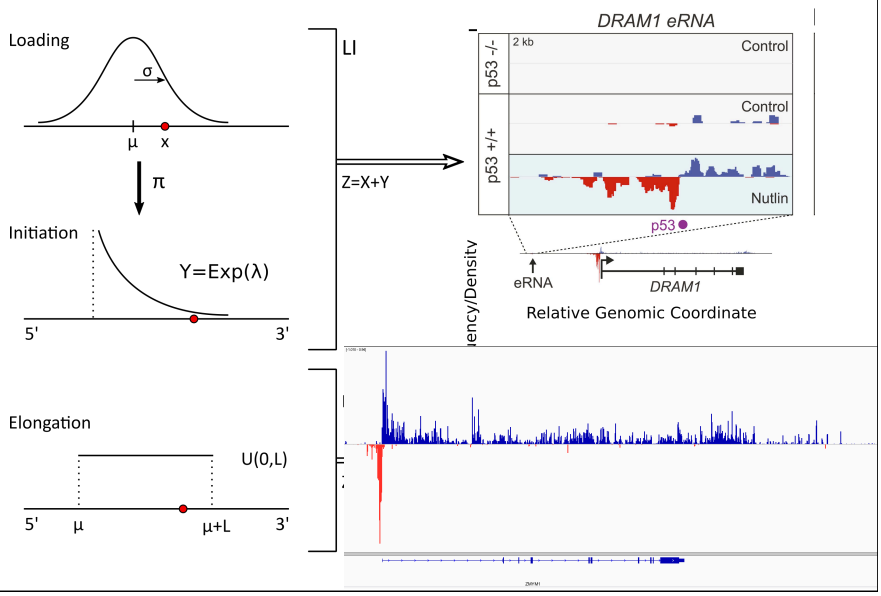
Our Generative Model approximates bidirectional transcription in GRO-seq



Our Generative Model approximates gene elongation in GRO-seq



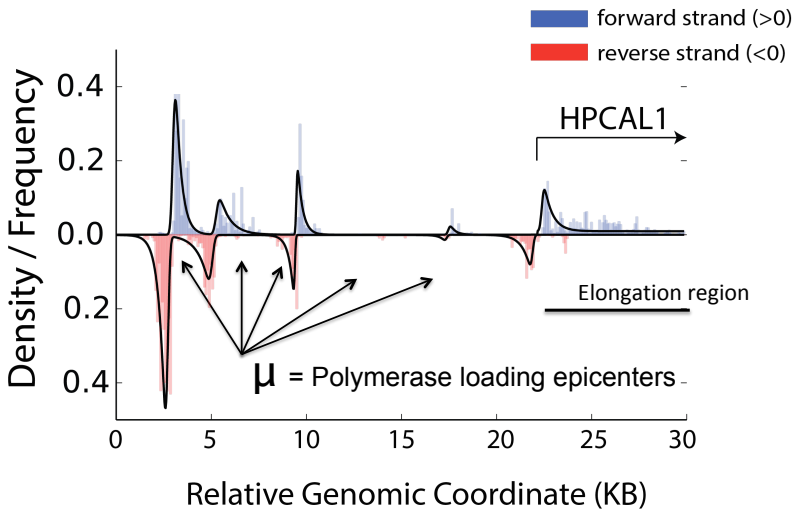
The model describes patterns observed within nascent transcription



Can we use our Generative model to infer the precise location of all enhancers and TSS given GRO-data?

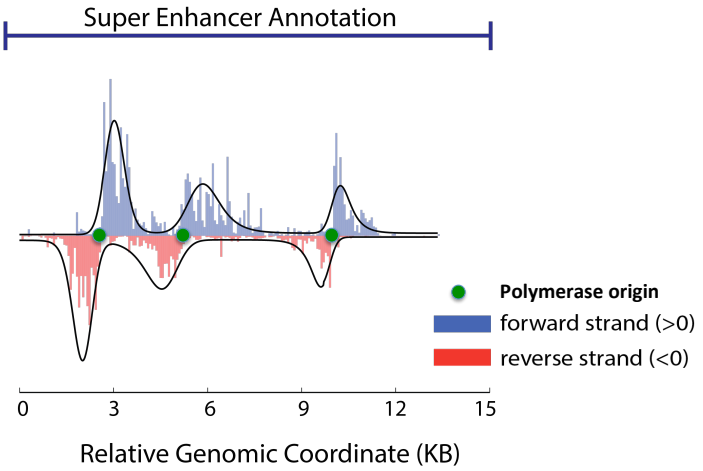
This is actually a fairly nasty Expectation Maximization algorithm.

We can fit the model to nascent transcription data



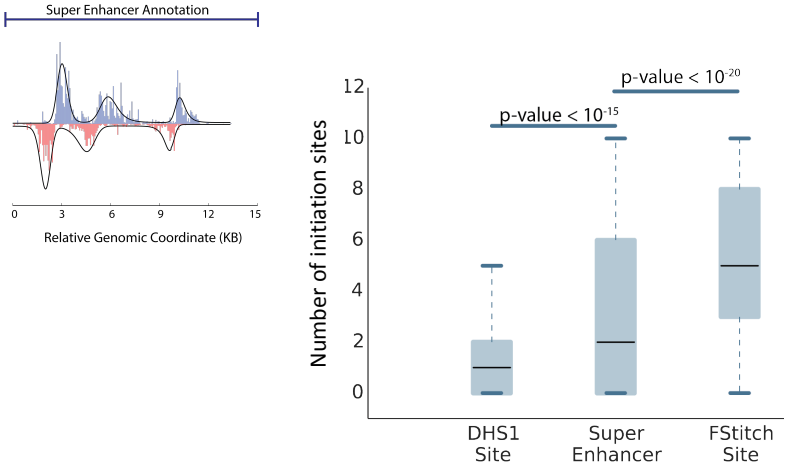
Azofeifa and Dowell. Bioinformatics. Jan 2017

We can fit the model to nascent transcription data



Azofeifa and Dowell. Bioinformatics. Accepted Aug 2016.

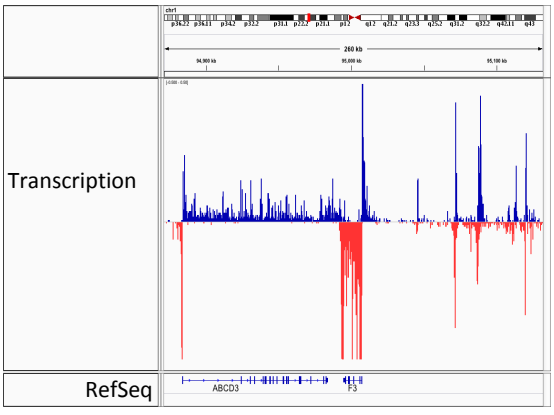
Transcribed regions (by HMM) contain several Pol II origins



ENCODE

~40-80% of the human genome is transcribed[1]

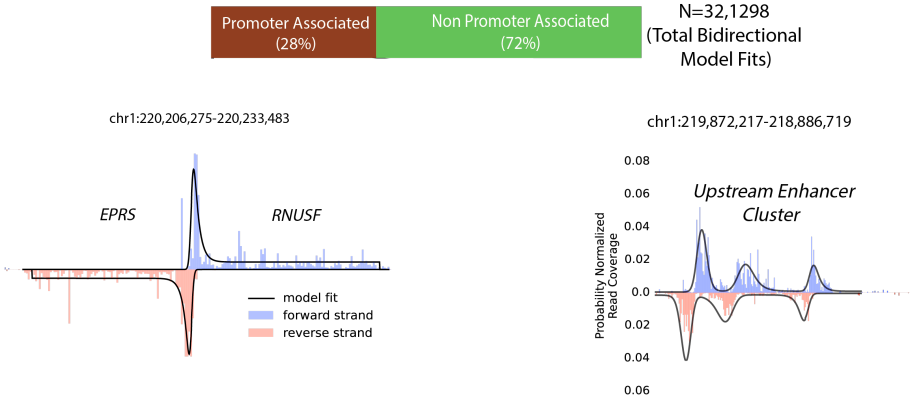
(…in some cell type)



Forward Strand
Reverse Strand

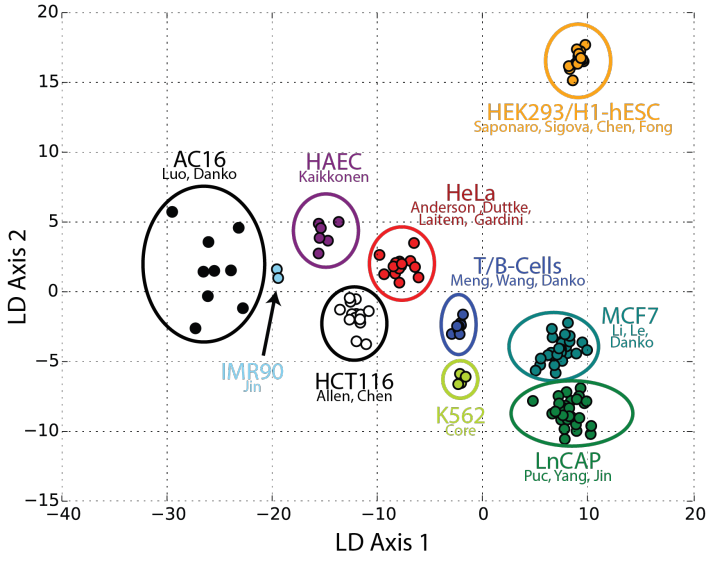
[1]**80% reported by ENCODE** (2012); we observe 40-60% across ~200 nascent transcription experiments

Most polymerase origins are not at protein coding genes

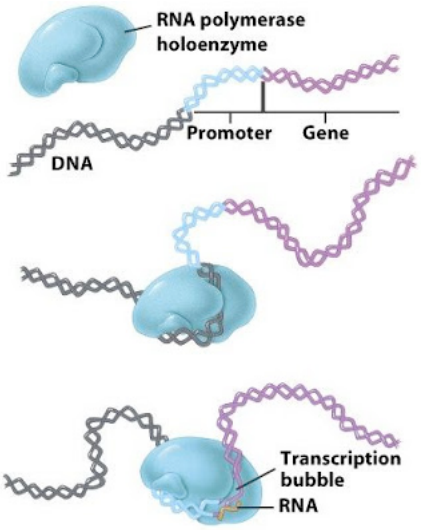


e.g. MOST transcription is non-coding and unstable

Patterns of bidirectional usage are a signature of cell type.



How does the cell regulate polymerase activity?



Images courtesy Sandwalk blog

The core polymerase binds DNA non-specifically as you might expect for a DNA binding protein that has to travel down a large number of different genes.

Then HOW does polymerase select "the right places" ???