Statistical analysis for ensemble snapshots of transcriptional bursting

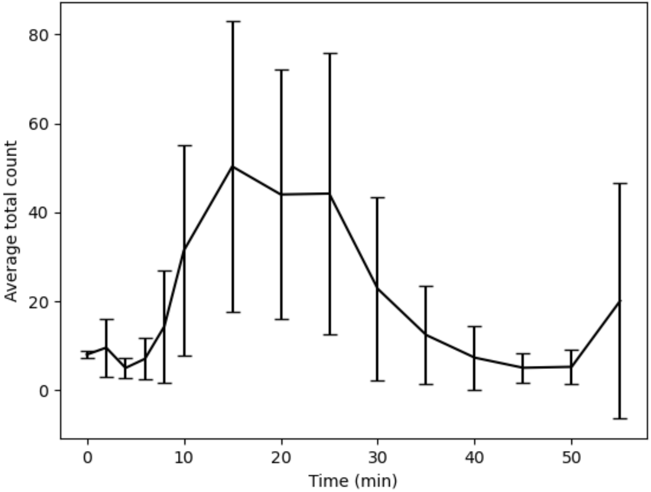
Clayton W. Seitz

June 28, 2022

Key questions

- Does IFNγ induce transcriptional bursts in HeLa cells?
- Which genes?
- What are the parameters of the burst (size, frequency, etc.)?
- In general, it possible to correlate spatial patterning with transcriptional bursting, using only ensemble snapshots (FISH)?

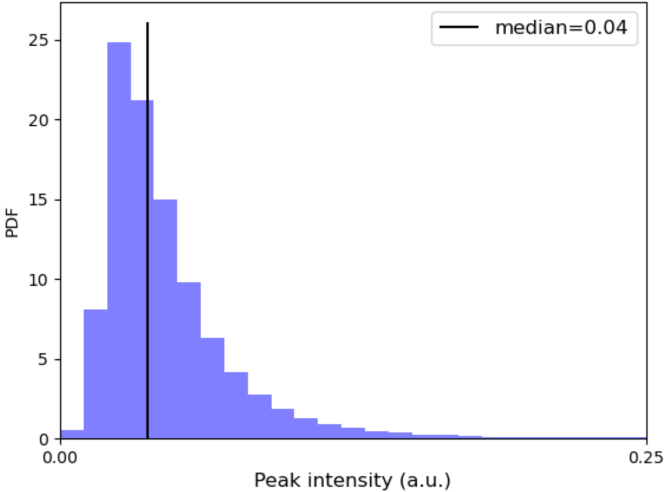
Significant variability in STL1 mRNA counts per cell at 0.4M NaCl



Error bars represent standard deviations from the mean

Cells marked ON for > 3 STL1 mRNA in yeast

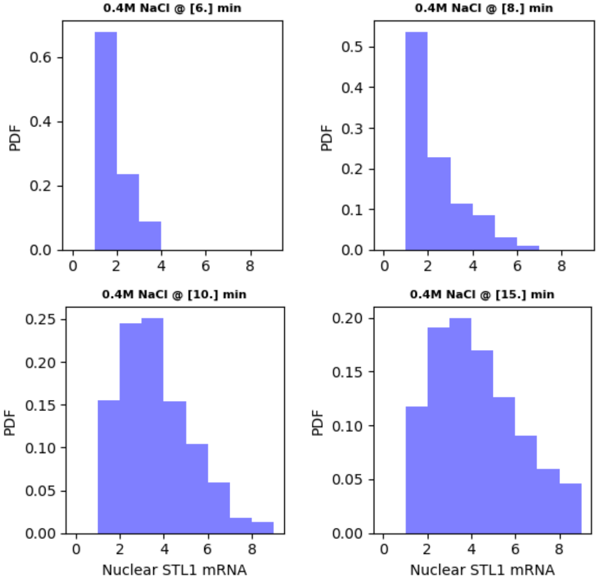
Assessing STL1 mRNA count variability at the transcription site



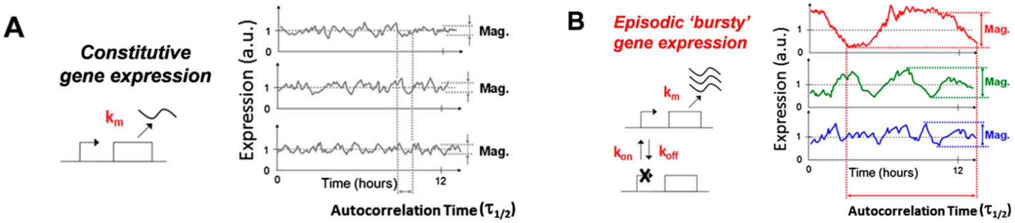
The median of the mRNA intensity distribution is used to determine the number of nascent RNA at the transcription site (TS)

Assessing STL1 mRNA count variability at the transcription site

- Brightest spot in the nucleus defined as putative TS
- TS marked ACTIVE if $I > 2 * \mathrm{med}$
- Nascent mRNA count is $\mathrm{round}(I/\mathrm{med})$
- Count variability suggests asynchrony



Gene expression is stochastic and non-constitutive


A
Constitutive gene expression
k_m
Expression (a.u.)
1
0
Mag.
Mag.
Mag.
0
Time (hours)
12
Autocorrelation Time ($\tau_{1/2}$)
B
Episodic 'bursty' gene expression
k_m
k_{on}
k_{off}
Expression (a.u.)
1
0
Mag.
Mag.
Mag.
0
Time (hours)
12
Autocorrelation Time ($\tau_{1/2}$)

Single-state models

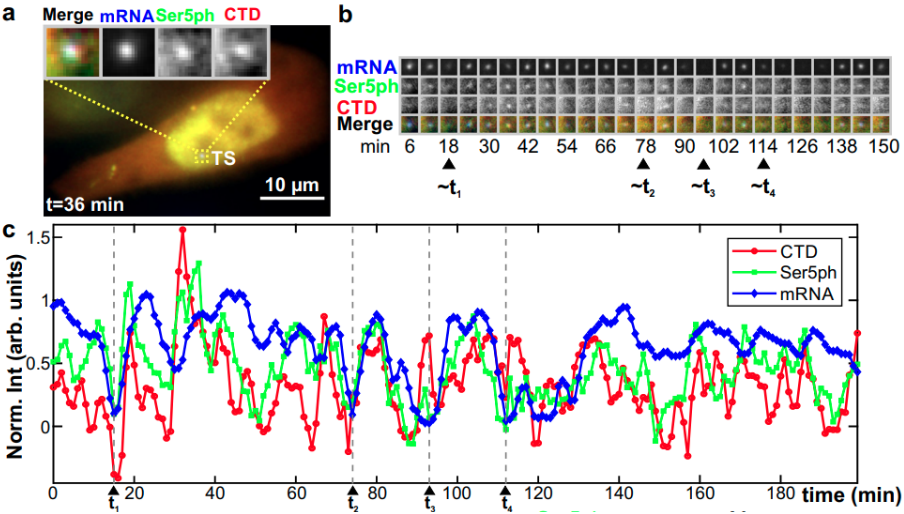
- RNAs are 'born' at a fixed rate
- RNA counts are Poisson

Multi-state models

- Promoter can be in multiple states (switching behavior)
- RNA counts are not Poissonian

Single-state models tend to underestimate variance in RNA counts

Gene expression is stochastic and non-constitutive (live-cell MS2-MCP)



Forero-Quintero, et al. *Live-cell imaging reveals the spatiotemporal organization of endogenous RNAPII phosphorylation at a single gene*. Nat Commun 2021

Ensemble averages and variances do not fully explain underlying transcription dynamics

- Transcription is stochastic, meaning that RNA counts can only be understood in terms of a probability distribution

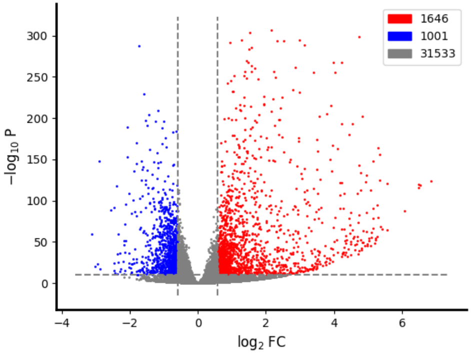
- High variance in mRNA counts suggests more complicated underlying dynamics which are not evident in ensemble averages

- We cannot assume that cells are bursting synchronously

- Bursting phase has implications for correlating bursting with spatial organization via ensemble data (FISH)
  - Classification of cells based on $P(X_{nuc}, X_{cyto}, X_{TS})$?
  - Can also look at the evolution of the spatial feature distributions

Interferon-γ induction in HeLa cells

Single cell transcriptome measurements of polyA mRNA for naïve HeLa cells (N=90), induced with interferon gamma (50ng/mL) for 24h



Siwek et al. *Activation of Clustered IFNγ Target Genes Drives Cohesin-Controlled Transcriptional Memory Cell 2020*

A compartment model for IFNγ induced gene expression

Let X represent an arbitrary RNA transcript of IFNγ induced gene G. Assume two chromatin states (on and off)

$$\begin{aligned}
\text{Gene activation} : G_{off} &\overset{k_{on}}{\rightarrow} G_{on} \\
\text{Gene inactivation} : G_{on} &\overset{k_{off}}{\rightarrow} G_{off} \\
\text{Transcription} : G_{on} &\overset{k_t}{\rightarrow} G_{on} + X_{\mathrm{nuc}} \\
\text{RNA Export} : X_{\mathrm{nuc}} &\overset{k_{exp}}{\rightarrow} X_{\mathrm{cyt}} \\
\text{RNA degradation} : X_{\mathrm{cyt}} &\overset{\gamma}{\rightarrow} \emptyset
\end{aligned}$$

Raw data collected post induction can be used to infer parameters

$$\theta = (k_{on}, k_{off}, k_t, k_{exp}, \gamma)$$

Bayesian parameter inference using ensemble snapshots

Likelihood-based methods can infer θ from ensemble snapshots (FISH data)

$$\theta = (k_{on}, k_{off}, k_t, k_{exp}, \gamma)$$

One way is through maximum a posteriori estimation (MAP):

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \; P(X|\theta)$$

A more robust (but harder) way is via Bayesian inference, which lets us infer θ from X while quantifying the uncertainty in our estimate:

$$P(\theta|\mathrm{x}) \propto P(\mathrm{x}|\theta)\pi(\theta) = \pi(\theta)\prod_t P(\mathrm{x}_t|\theta)$$

The likelihood $P(X, t)$ is the solution to the chemical master equation at time t

Kolmogorov's forward equation (chemical master equation)

Dynamics on biochemical reaction networks are inherently stochastic and the state space is discrete. We can only write probabilities over the state space

$$\begin{aligned} P(\mathrm{x}_i, t) &= \sum_j T_{ji}(\mathrm{x}_i, t|\mathrm{x}_j, t-\Delta t)P(\mathrm{x}_j, t-\Delta t) \\ &= \sum_k T_k(\mathrm{x}_i, t|\mathrm{x}_i - \nu_k, t-\Delta t)P(\mathrm{x}_i - \nu_k, t-\Delta t) \end{aligned}$$

where T_k is the probability of a reaction channel k firing in the interval $(t, t+\Delta t)$.

Taking the limit $\Delta t \to 0$ one can derive the forward Kolmogorov equation or chemical master equation (CME)

$$\frac{dP(\mathrm{x}, t|\mathrm{x}_0)}{dt} = \sum_k T_k(\mathrm{x}-\nu_k)P(\mathrm{x}-\nu_k, t) - T_k(\mathrm{x})P(\mathrm{x}, t)$$