# Establishing a quantitative framework for analyzing inducible gene expression in HeLa cells

Clayton W. Seitz

August 18, 2022

The three stages

- ▶ Select a set of genes that are differentially expressed after interferon treatment
- ▶ Mark genes which have special relevance in the literature e.g., PDL-1
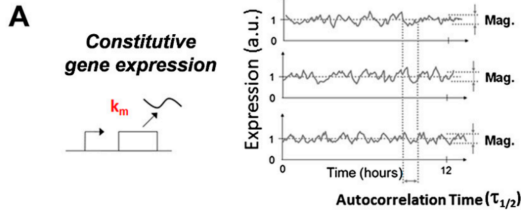- ▶ Iterative RNA FISH experiments, apply bursting models, spatial analysis
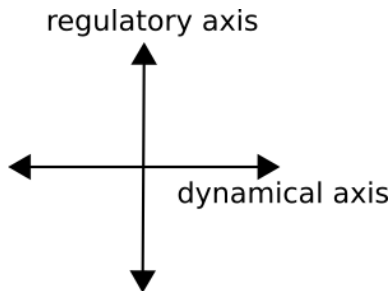
Interferon-γ induces differential expression of many genes

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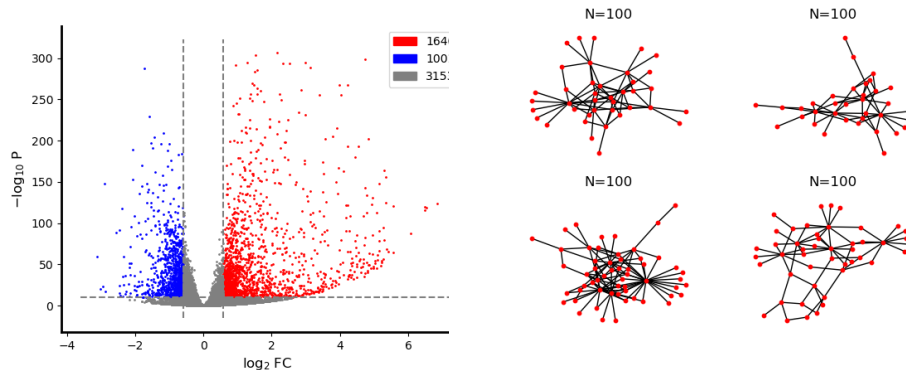

Randomly selected $N = 100$ genes with edges drawn for $I(X;Y) \geq 0.1$ using Kraskov's method: $I(X, Y) \approx \psi(k) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N)$

Promoter models are necessary for non-constitutive gene expression

A Constitutive gene expression, k_m

B Episodic 'bursty' gene expression, k_m, k_{on}, k_{off}

Expression (a.u.), Time (hours), Mag., Autocorrelation Time $(\tau_{1/2})$

Single-state models

▶ RNAs are 'born' at a fixed rate

▶ RNA counts are Poisson
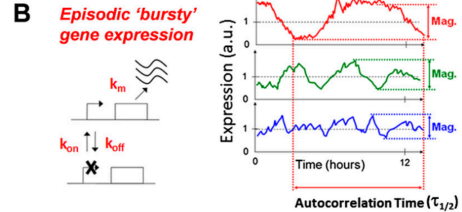
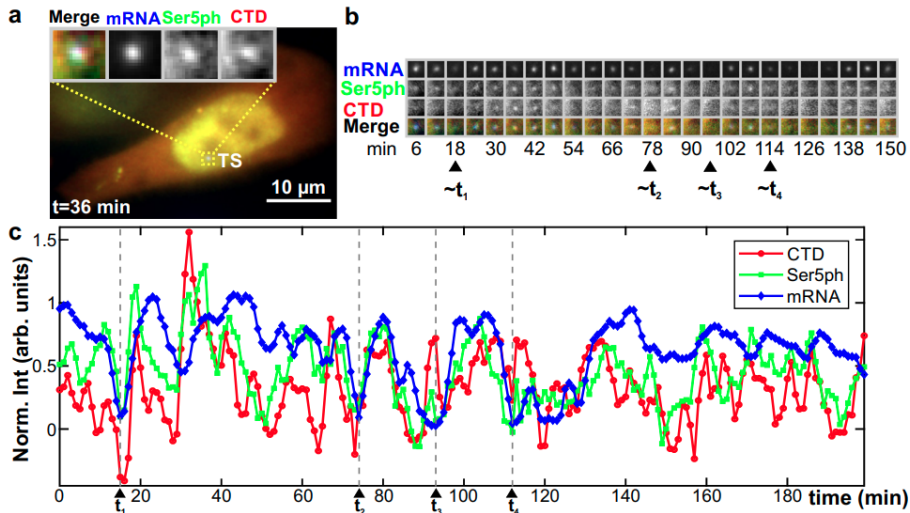## Multi-state models

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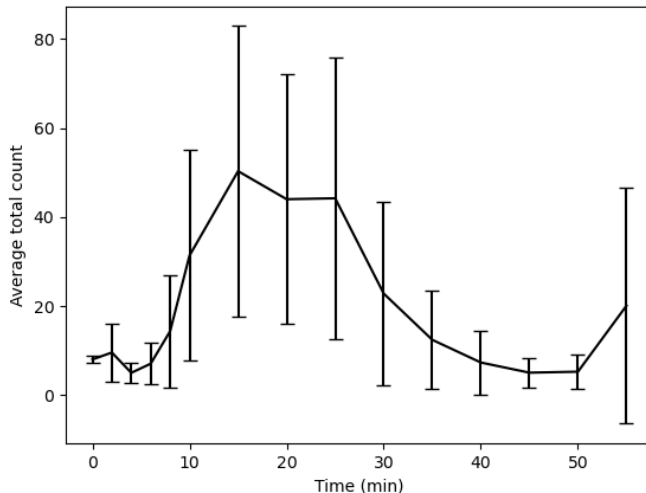
▶ RNA counts are not Poissonian

Single-state models tend to underestimate variance in RNA counts

Gene expression is stochastic (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

Example: variability in STL1 mRNA counts per cell at 0.4M NaCl

Error bars represent standard deviations from the mean
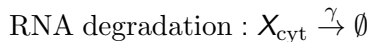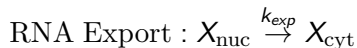Cells marked ON for > 3 STL1 mRNA in yeast

Assessing STL1 mRNA count variability at the transcription site

- ▶ Brightest spot in the nucleus defined as putative TS
- ▶ TS marked ACTIVE if $I > 2 * \text{med}$
- ▶ Nascent mRNA count is $\text{round}(I/\text{med})$
- ▶ Count variability suggests asynchrony
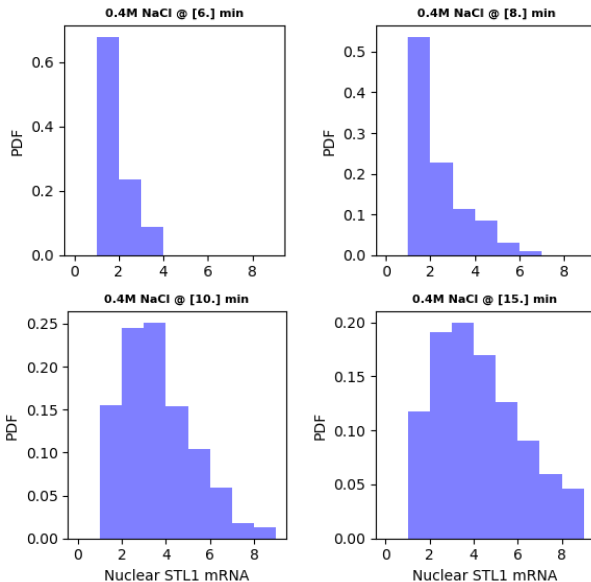
A spatial model for induced gene expression

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

$$\text{Gene activation}: G_{off} \overset{k_{on}}{\rightarrow} G_{on}$$

$$\text{Gene inactivation}: G_{on} \overset{k_{off}}{\rightarrow} G_{off}$$

$$\text{Transcription}: G_{on} \overset{k_t}{\rightarrow} G_{on} + X_{\text{nuc}}$$

$$\text{RNA Export}: X_{\text{nuc}} \overset{k_{exp}}{\rightarrow} X_{\text{cyt}}$$

$$\text{RNA degradation}: X_{\text{cyt}} \overset{\gamma}{\rightarrow} \emptyset$$

Raw data collected post induction can be used to infer parameters

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

Bayesian inference of model parameters

It is well-known that using just means and variances gives poor estimates of the model parameters (Munsky et al. PNAS 2018)

Let $\theta = (k_{on}, k_{off}, k_t, k_{exp}, \gamma)$. Using Bayes Rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)} \propto P(X|\theta)P(\theta)$$

Can infer θ if we know the likelihood $P(X|\theta)$ (the hard part) and specify a prior $P(\theta)$

Generally we have to resort to Monte Carlo methods to find $P(X|\theta)$

Kolmogorov's forward equation (chemical master equation)

To calculate the likelihood $P(X|\theta)$ one has to address the forward Kolmogorov equation

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

It is possible to find $P(\mathbf{x}, t|\mathbf{x}_0)$ in two main ways: (1) Finite state projection (2) Monte Carlo simulation

The former is exact, the latter is an approximation (see approximate Bayesian computation)