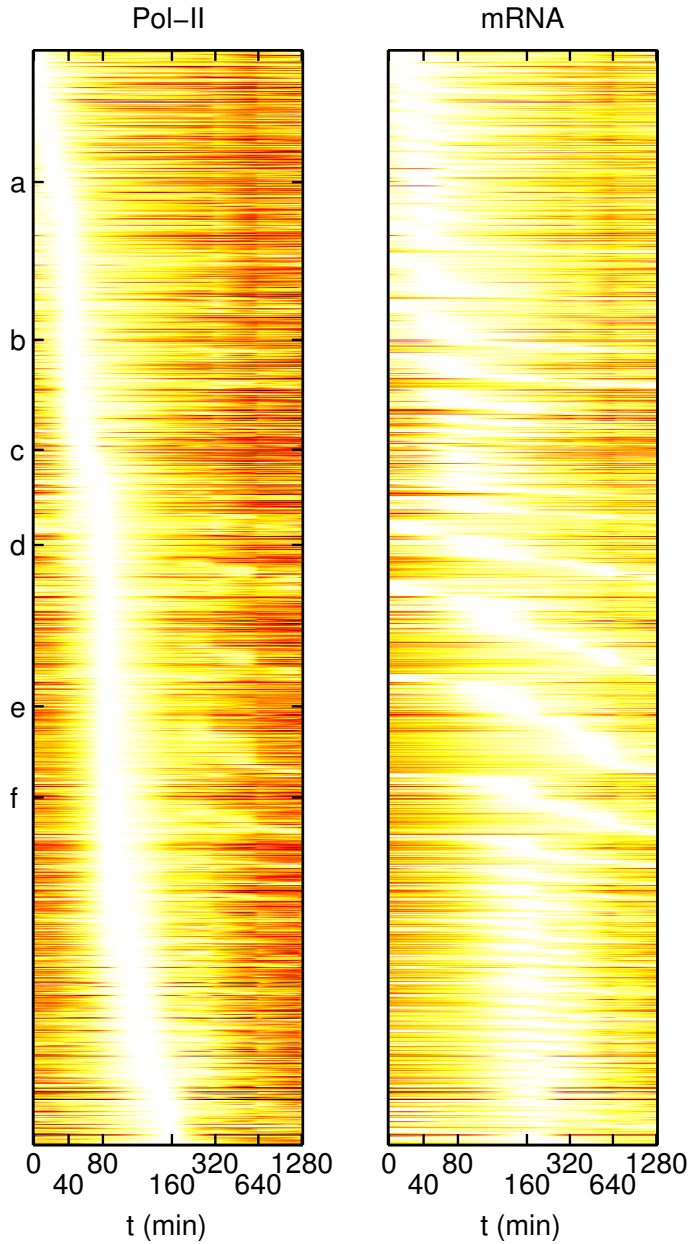
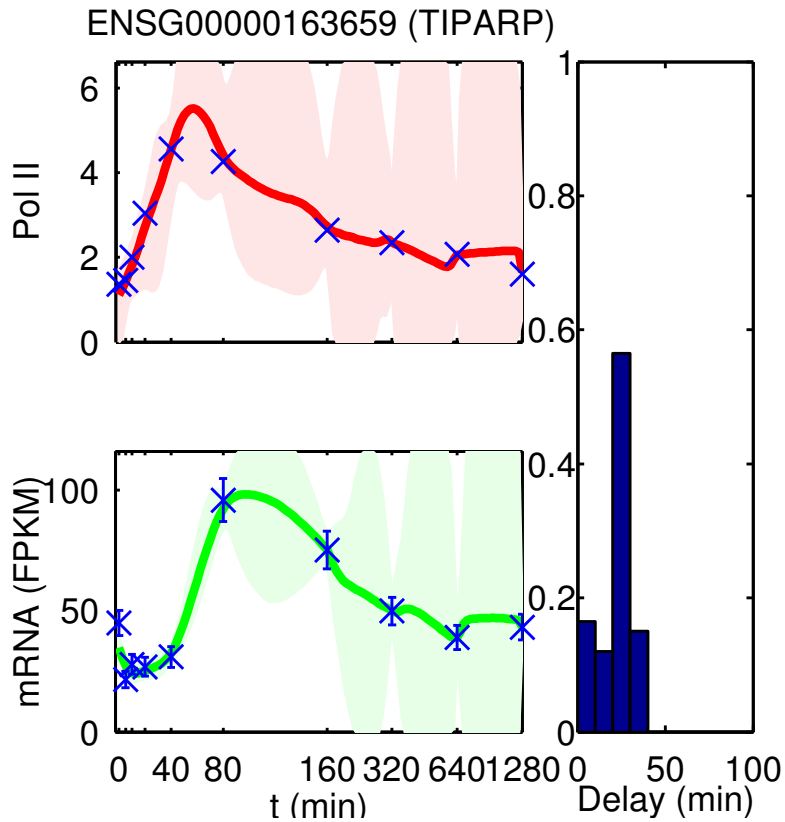
Probabilistic modelling of omic time course data

3rd Machine Learning for Personalised Medicine Summer School
Museum of Science and Industry
Manchester Sep 21st, 2015

Magnus Rattray
Professor of Computational & Systems Biology
Faculty of Life Sciences, University of Manchester

Transcription is a highly regulated dynamic process





Response to estrogen receptor stimulation in MCF7 cells

Talk Outline

Background: Introduction to Gaussian Processes

- From Gaussian distributions to Gaussian processes
- Gaussian processes for regression

Part 1. Modelling Pol-II elongation dynamics

- Representing promoter activity as a Gaussian process
- Inferring the time required for elongation
- Inferring and clustering promoter activity profiles

Part 2. Linking Pol-II activity to mRNA profiles

- Representing mRNA production rate as a Gaussian process
- Inferring RNA processing delays
- Delay link with splicing: evidence from intronic reads

Part 3. Inferring transcription factor targets

- Modelling TF activity as a Gaussian process
- Inferring targets by fitting regulation models

Gaussian processes: flexible non-parametric models

Probability distributions over functions

$$f(t) \sim \mathcal{GP}\left(\text{mean}(t), \text{cov}(t, t')\right)$$

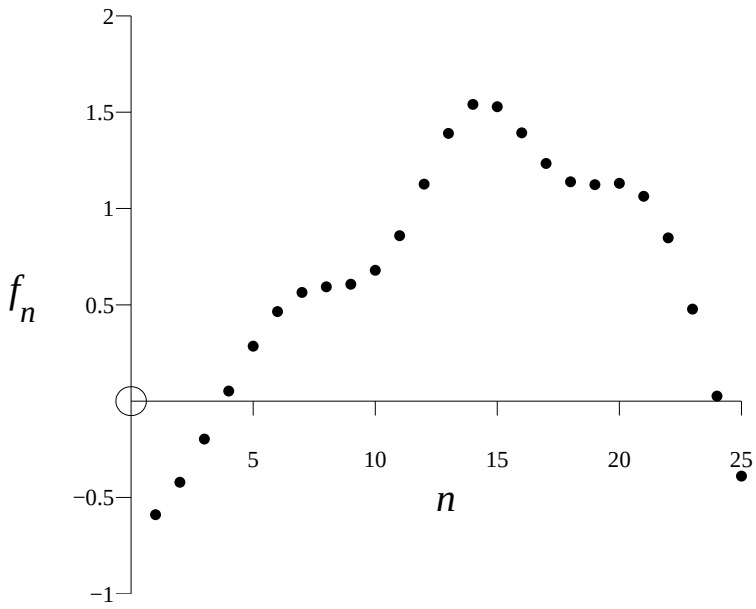
Covariance function $\text{cov}(t, t')$ defines typical properties,

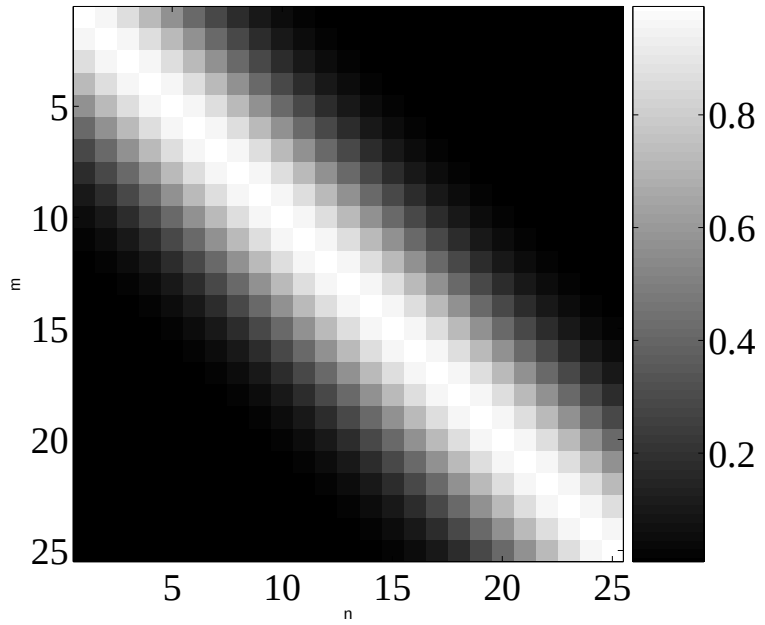
- Static . . . Dynamic
- Smooth . . . Rough
- Stationary. . . non-Stationary
- Periodic. . . Chaotic

The covariance function has parameters (called "hyper-parameters") tuning these properties

Gaussian processes

Samples from a 25-dimensional multivariate Gaussian distribution:



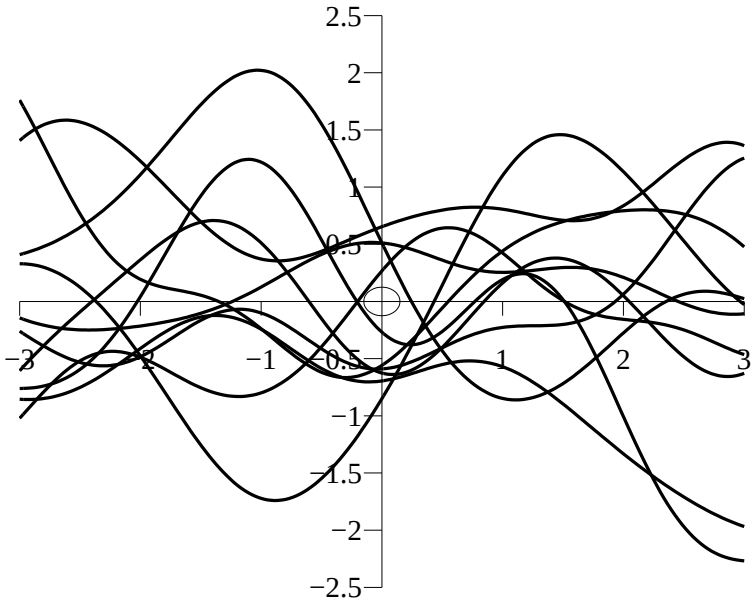


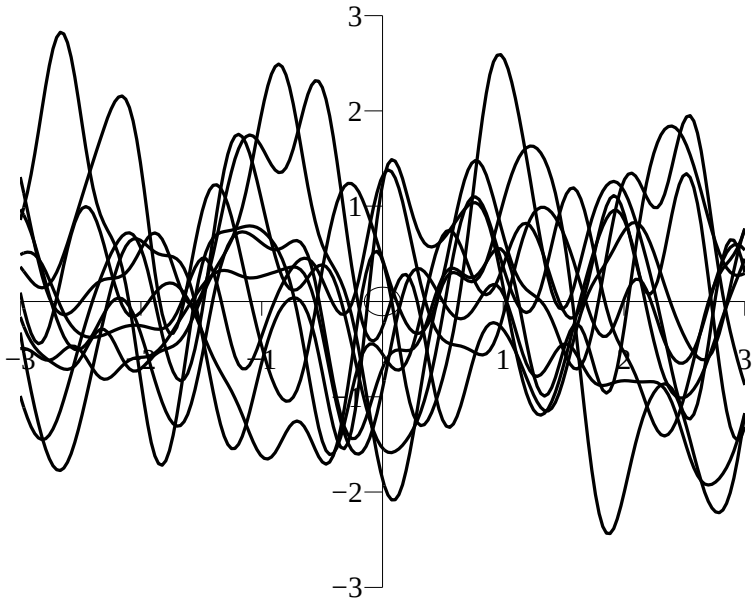
$$[f_1, f_2, \ldots, f_{25}] \sim \mathcal{N}(0, C)$$

Learning and Inference in Computational Systems Biology, MIT Press

Gaussian processes

Take dimension $\to \infty$

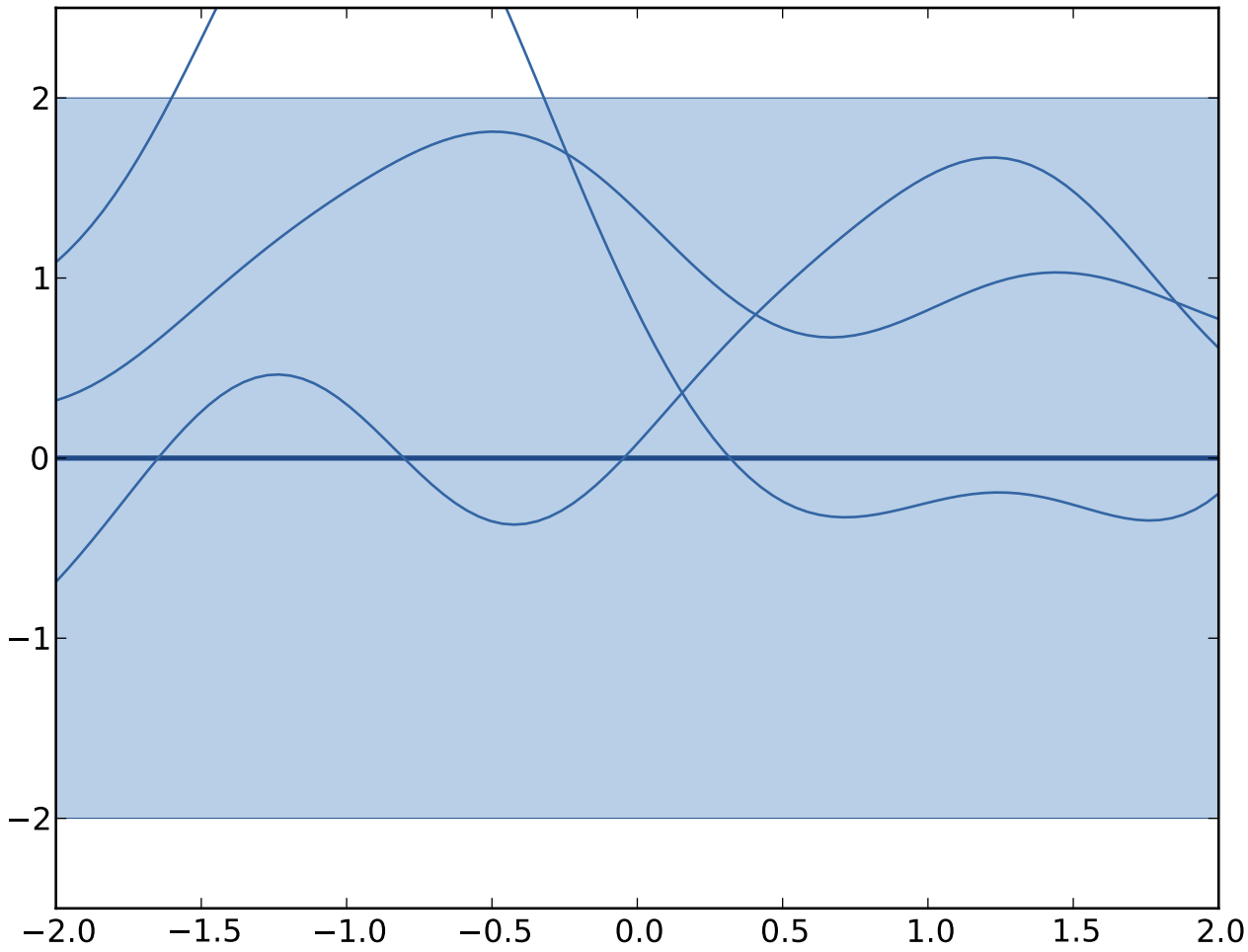




$$\mathrm{cov}\left(t, t'\right) = \exp\left(-\frac{(t-t')^2}{l^2}\right)$$

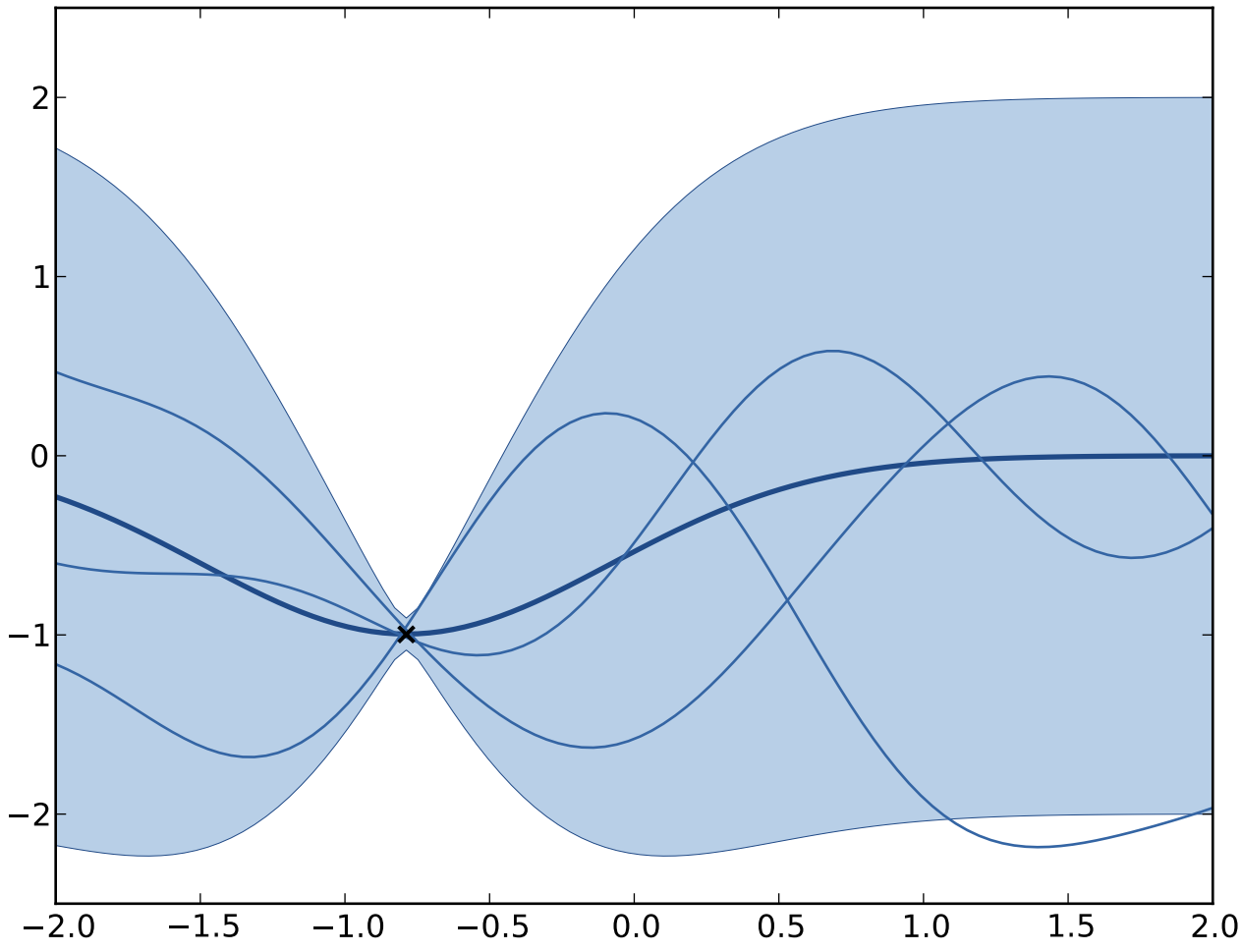
Learning and Inference in Computational Systems Biology, MIT Press

Gaussian processes for inference: Bayesian Regression



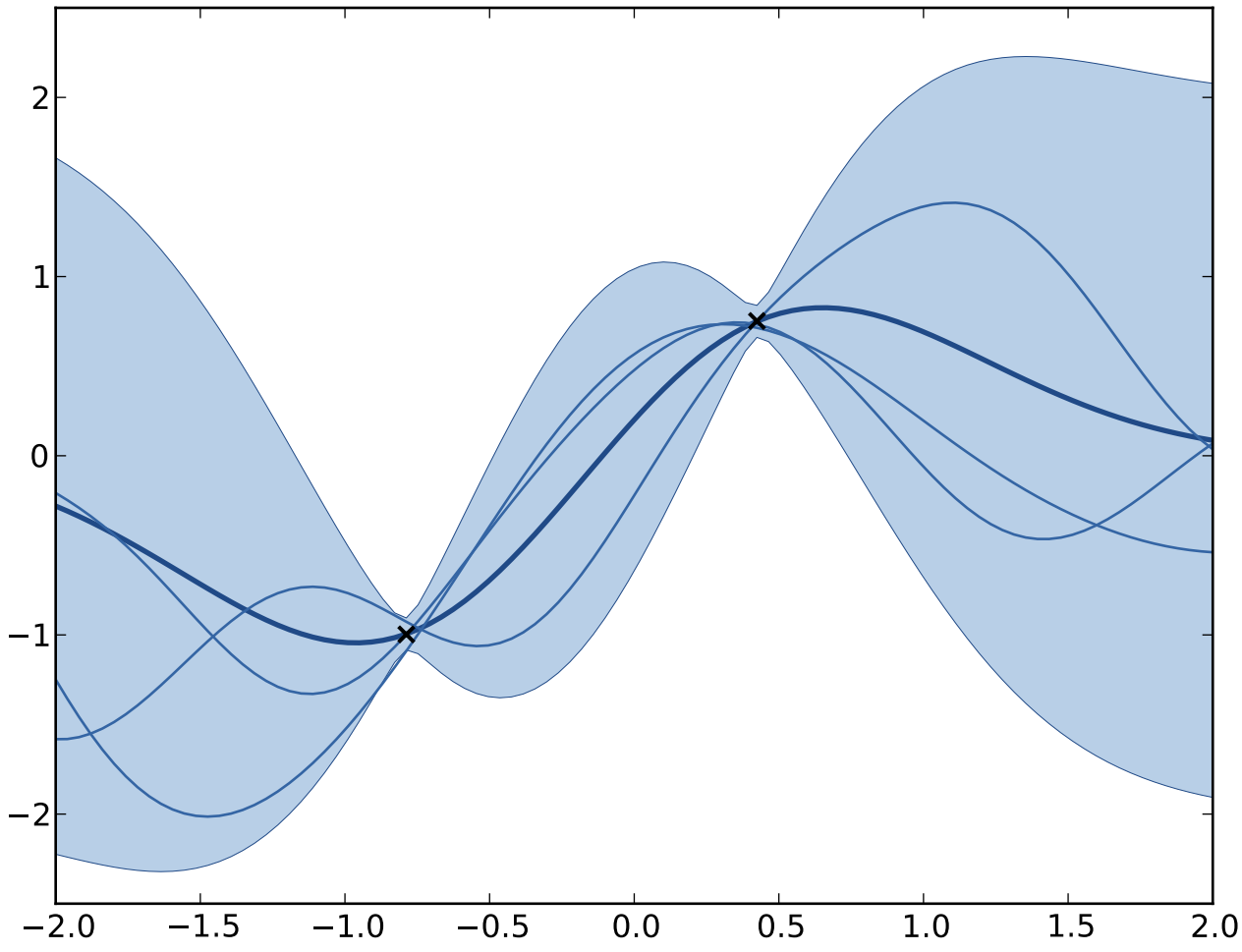
Figures from James Hensman

Regression example



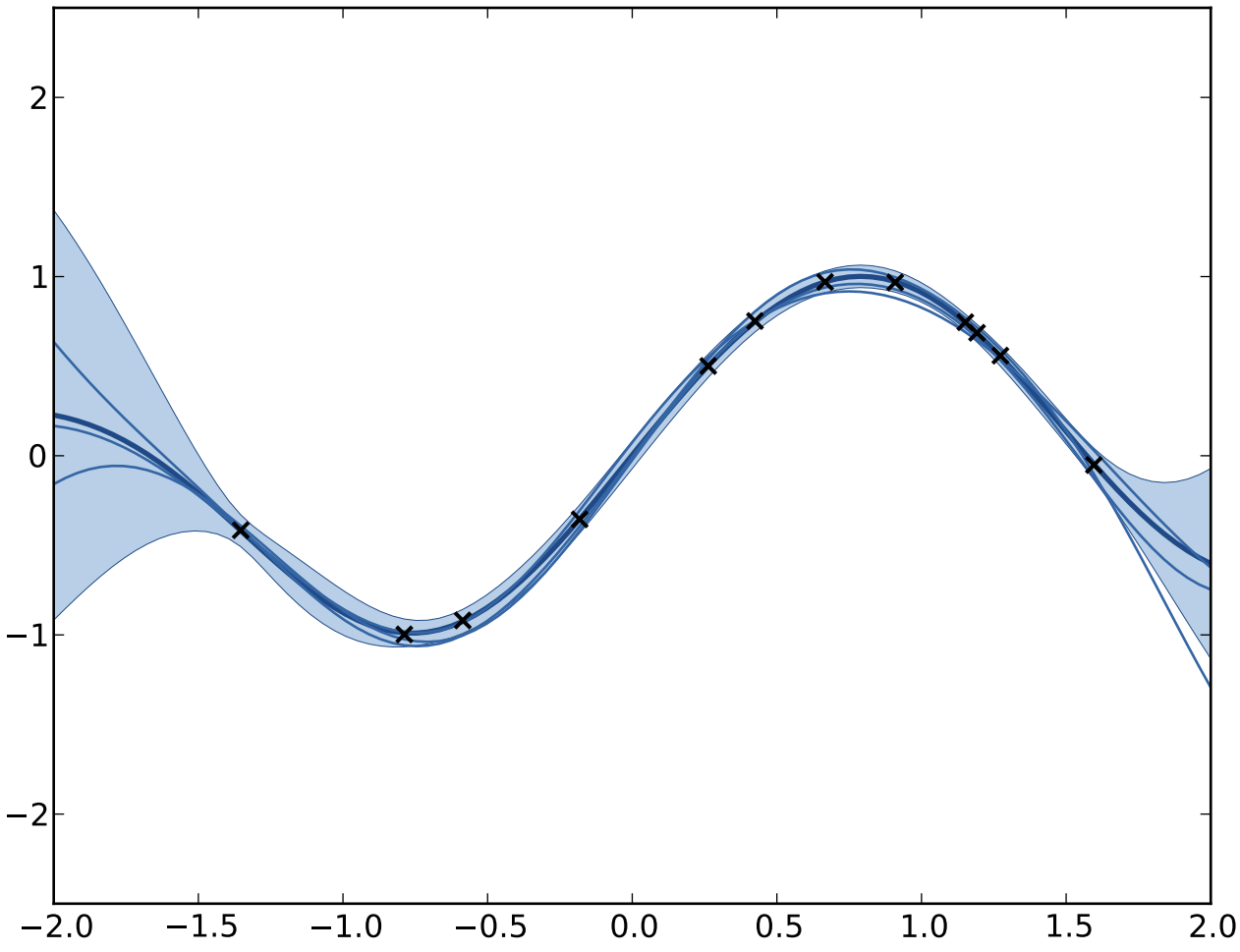
Figures from James Hensman

Regression example



Figures from James Hensman

Regression example



Figures from James Hensman

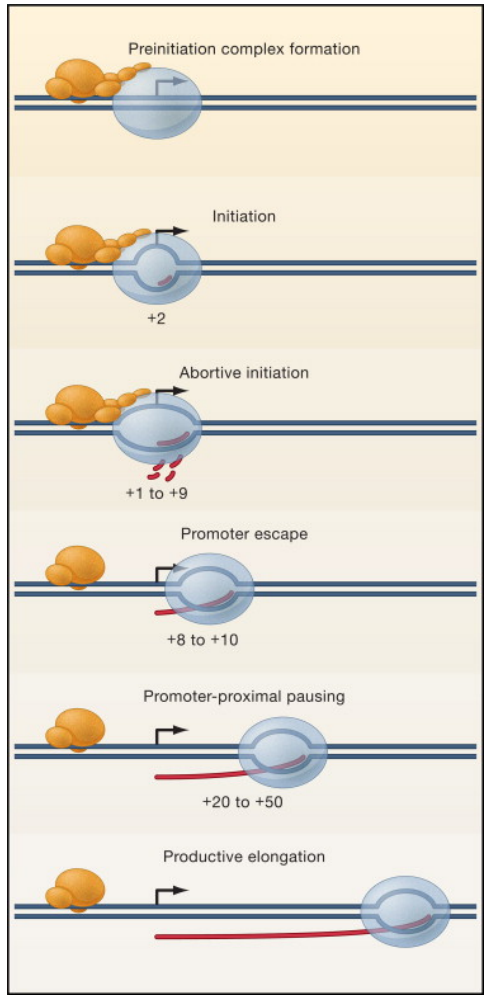
Part 1. Modelling RNA polymerase dynamics

Eukaryotic genes are transcribed by RNA polymerase II (RNA pol-II)

We model the dynamics using convolved Gaussian Processes (C. wa Maina et al. *PLoS Computational Biology*, 2014)

Joint work with Ciira Maina, Antti Honkela and Neil Lawrence

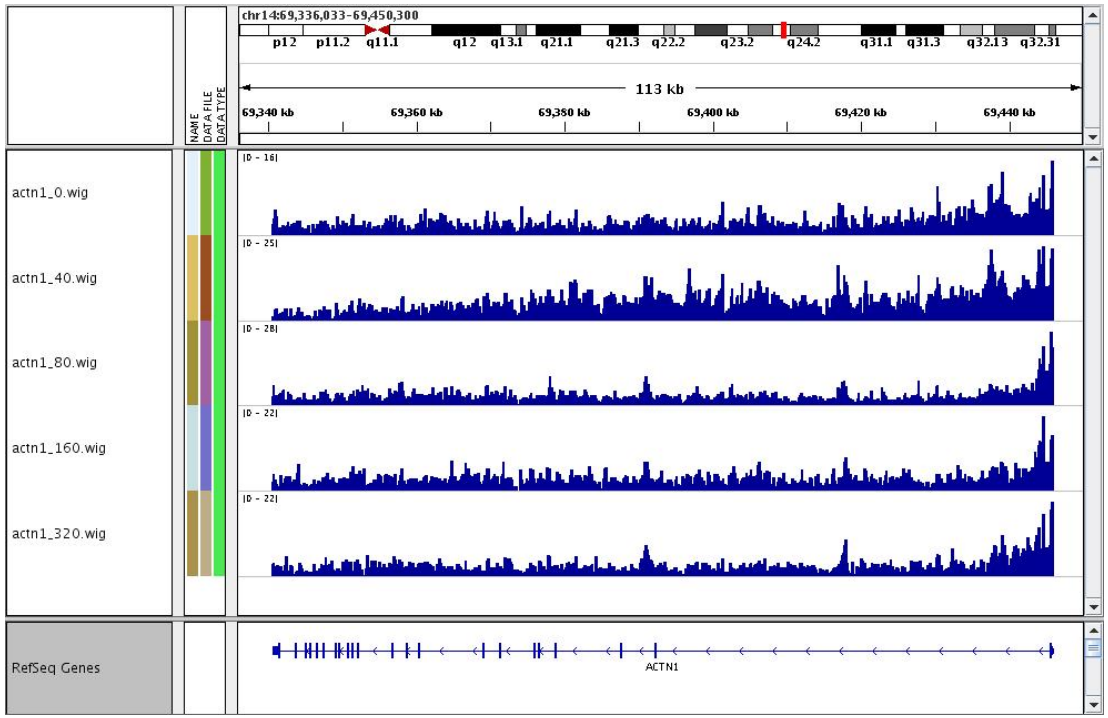
Data from Henk Stunnenberg and George Reid



[Margaritis and Holstege, Cell 133]

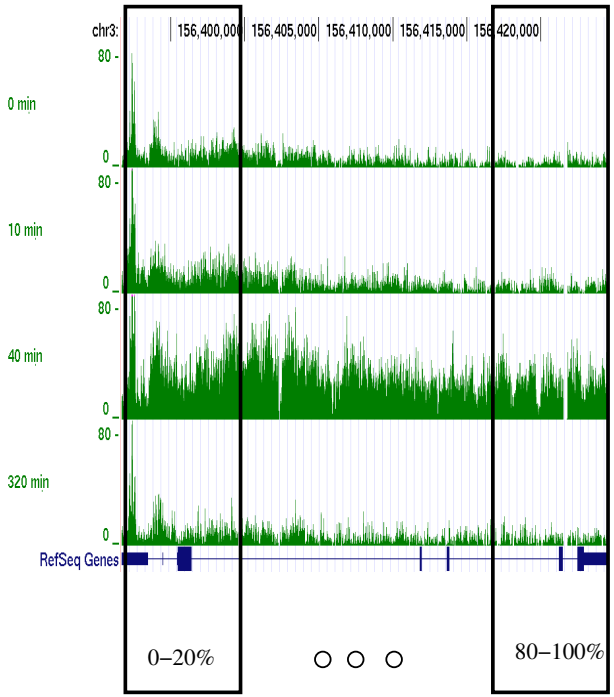
Modelling RNA polymerase dynamics

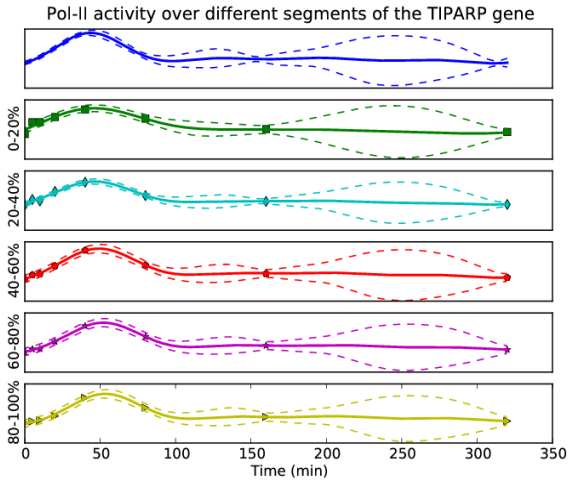
- MCF7 cells stimulated by estradiol (E2)
- Pol-II occupancy measured using ChIP-Seq
- 8 time points between 0 and 320 min (log scale: 0,5,10,20,40 etc)



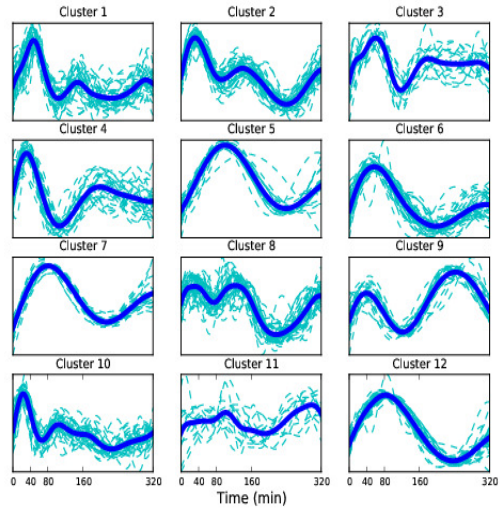
Pol-II occupancy of ACTN1 from 0 to 320 min showing a ‘transcription wave’

Modelling RNA polymerase dynamics





Promoter Activity Clusters



1) E2 responsive genes
2) Transcription speeds
3) Promoter activity profiles

Modelling RNA polymerase dynamics

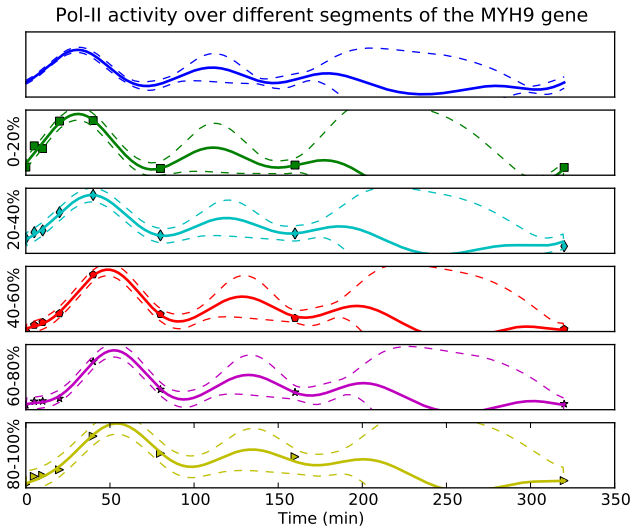
- Divide gene into 5 segments and consider Pol-II occupancy as a function of time for each region $i \in \{1, \ldots, 5\}$
- Occupancy for the ith segment is modelled as

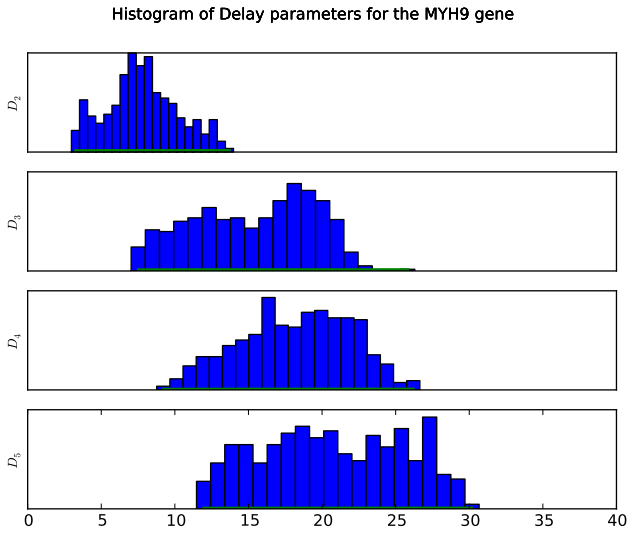
$$y_i(t) = \alpha_i \int f(t - \tau) k_i(\tau - D_i) d\tau + \epsilon_i(t)$$

- $f(t) \sim \mathcal{GP}(0, k_f)$ is the transcriptional activity at promoter
- $f(t - \tau)$ is time-lagged version to model transcriptional delay
- Smoothing kernel $k_i(\tau - D_i)$ models "spreading out" over time
- D_i determines the transcription speed
- Bayesian parameter estimation provides uncertainty ("posterior probability") of parameter estimates

Modelling RNA polymerase dynamics

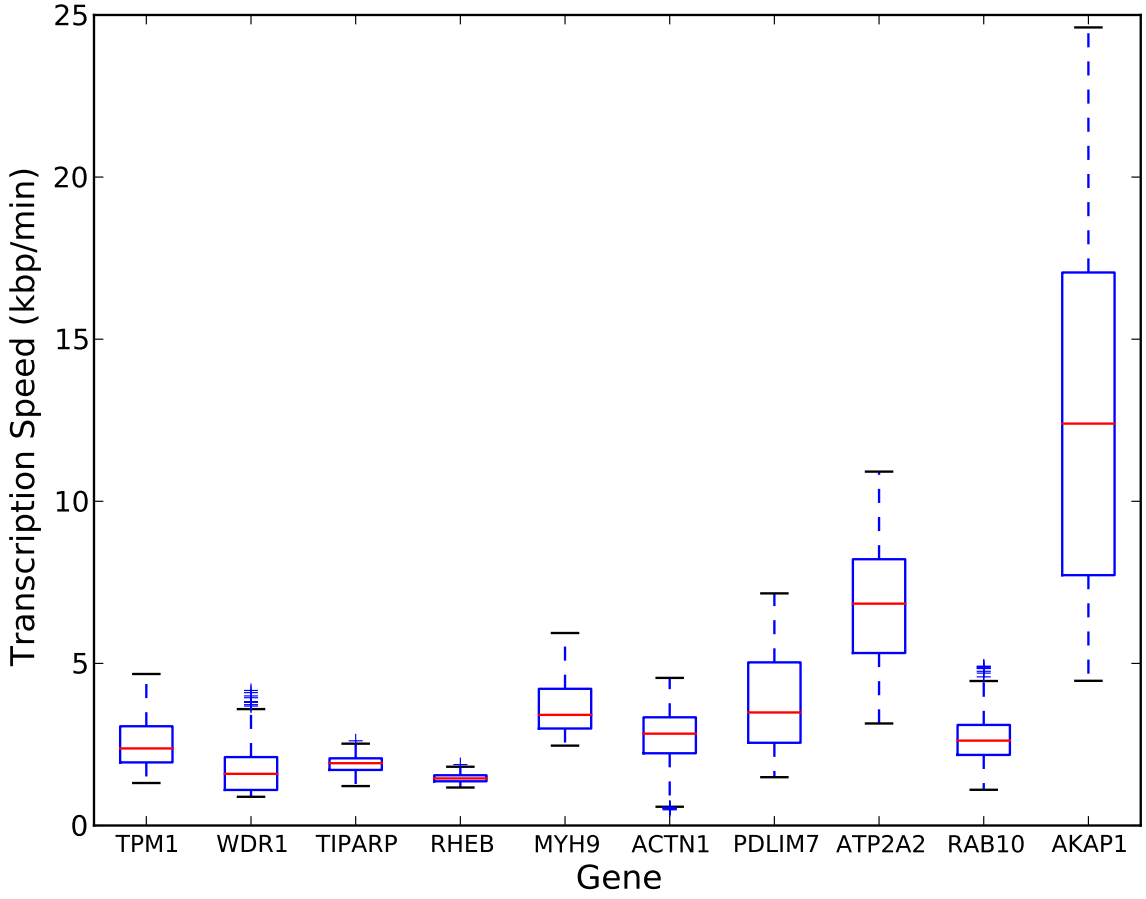
- Below we show the model fit for MYH9 (length 106741bp)





- Left: Transcriptional activity profile for each segment modelled as convolved Gaussian processes
- Right: Posterior probability for the delay parameters

Inferred transcription speeds



C. wa Maina et al. *PLoS Computational Biology 2014*

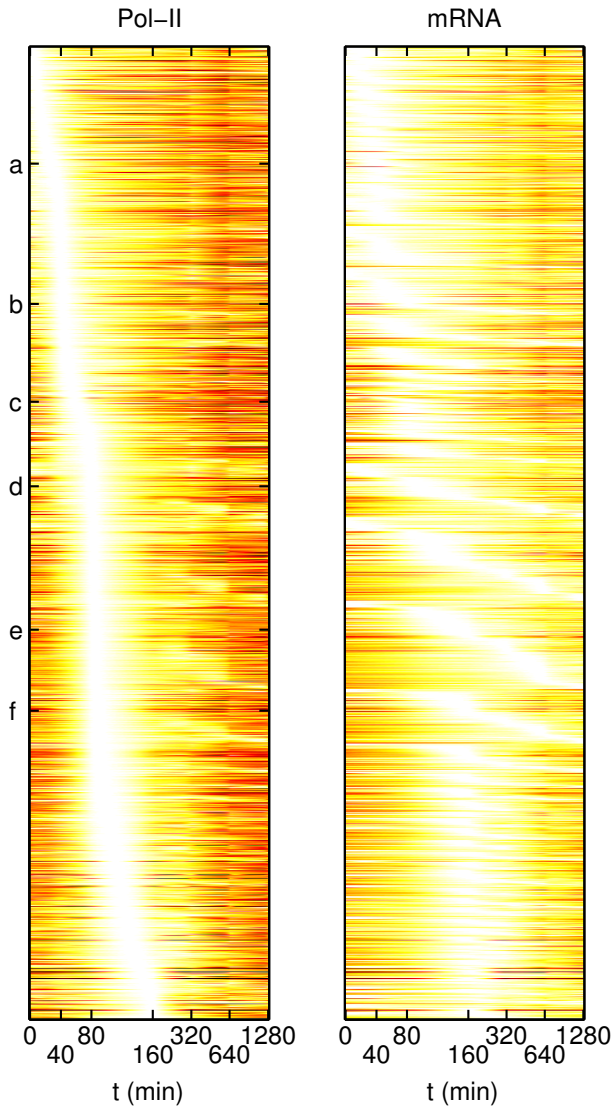
Inferred promoter activity clusters

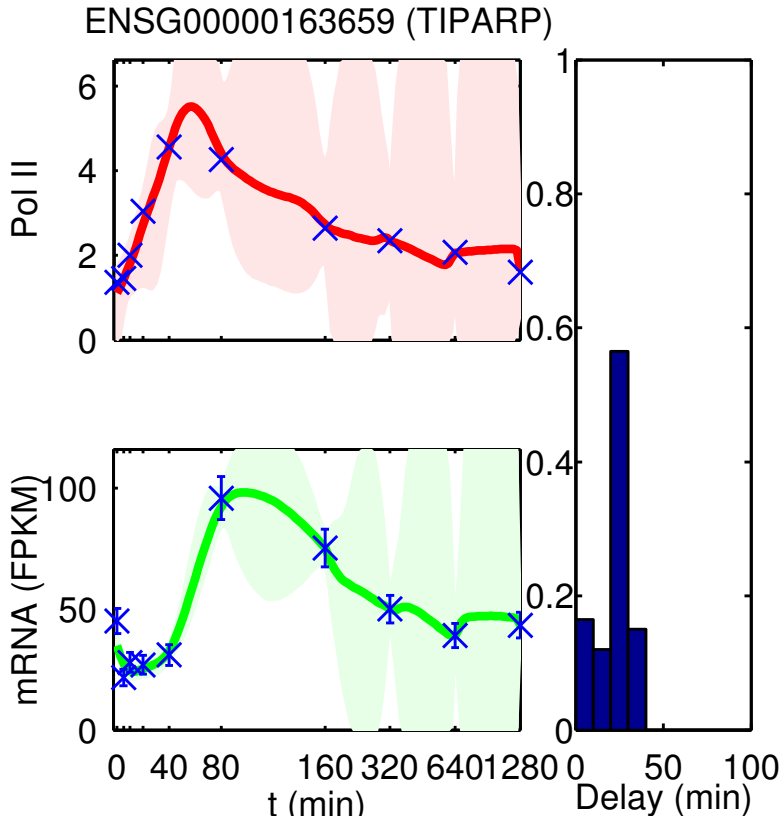
- Nearby binding in public ChIP-Seq data in the same system
- Significant enrichment shown in red
- Fast clusters (1,2,4,10) most enriched for ESR1 and FOXA1

Cluster	TFs						
	ESR1	FOXA1	c-FOS	c-JUN	MYC	SRC-3	TRIM24
1 (37)	27 (0.0)	14 (0.028)	16 (0.001)	6	4	25 (0.007)	27
2 (47)	31 (0.003)	19 (0.005)	16 (0.022)	7	7 (0.034)	36 (0.0)	38 (0.015)
3 (18)	11	5	7	5 (0.029)	6 (0.001)	11	12
4 (29)	20 (0.007)	11 (0.048)	9	7 (0.023)	2	18	23
5 (27)	15	4	6	8 (0.004)	9 (0.0)	16	19
6 (40)	27 (0.003)	8	12	7	4	25 (0.027)	31
7 (24)	10	6	5	6 (0.029)	3	13	19
8 (47)	32 (0.001)	10	14	14 (0.0)	8 (0.011)	31 (0.005)	40 (0.002)
9 (26)	18 (0.01)	7	11 (0.01)	11 (0.0)	3	12	22 (0.025)
10 (38)	30 (0.0)	14 (0.036)	15 (0.006)	2	1	29 (0.0)	32 (0.008)
11 (13)	5	2	7 (0.008)	4 (0.036)	2	7	13 (0.004)
12 (37)	19	8	12	11 (0.001)	4	23 (0.037)	29

C. wa Maina et al. *PLoS Computational Biology 2014*

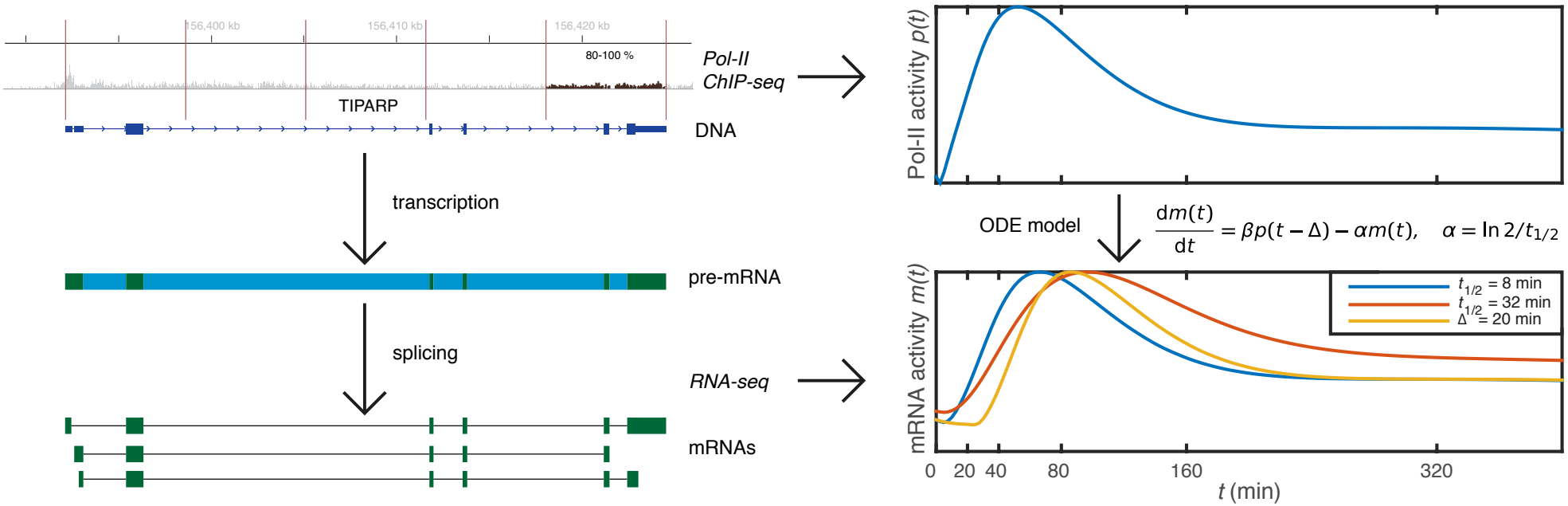
Part 2. Linking Pol-II activity to mRNA profiles





Joint work with Antti Honkela, Jaakko Peltonen, Neil Lawrence

Linking Pol-II activity to mRNA profiles



Honkela et al. "Genome-wide modelling of transcription kinetics reveals patterns of RNA processing delays" *PNAS* 2015 (in press)

Linking Pol-II activity to mRNA profiles

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \beta p(t-\Delta) - \alpha m(t)$$

- $m(t)$ is mRNA concentration (RNA-Seq data)
- $p(t)$ is mRNA production rate (3' pol-II ChIP-Seq data)
- α is degradation rate (mRNA half-life $t_{1/2} = 2/\alpha$)
- Δ is processing delay

Linking Pol-II activity to mRNA profiles

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \beta p(t-\Delta) - \alpha m(t)$$

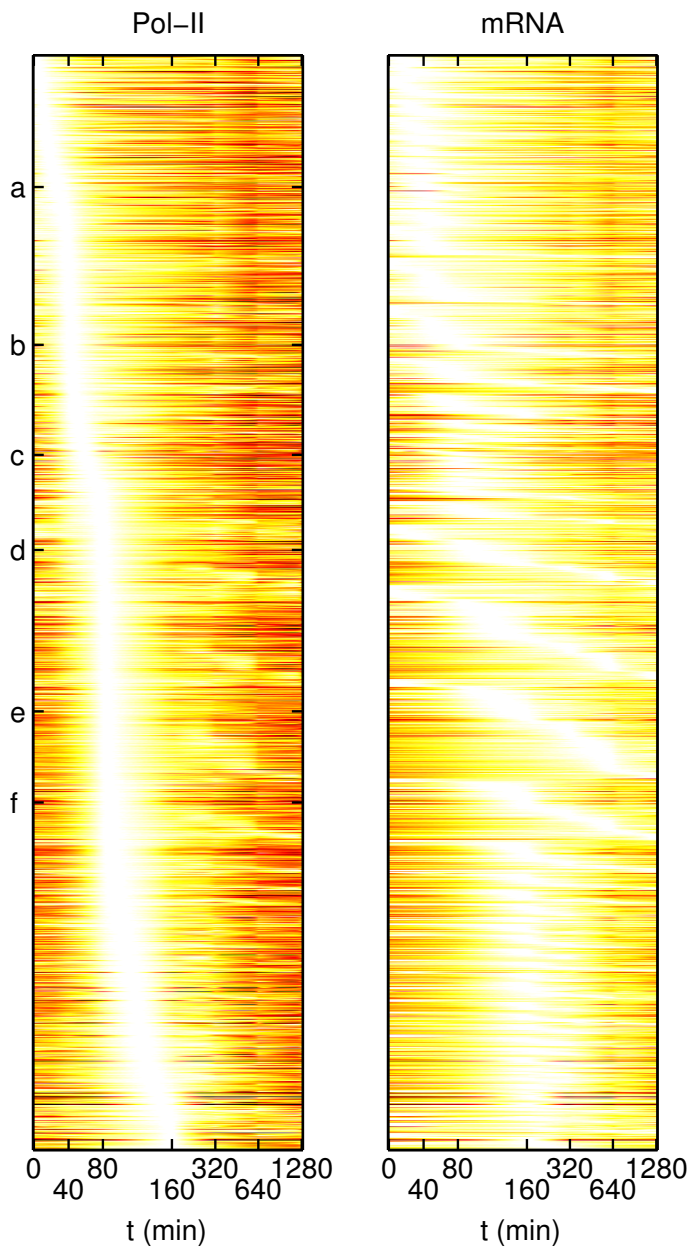
- $m(t)$ is mRNA concentration (RNA-Seq data)
- $p(t)$ is mRNA production rate (3' pol-II ChIP-Seq data)
- α is degradation rate (mRNA half-life $t_{1/2} = 2/\alpha$)
- Δ is processing delay

We model $p(t) \sim \mathcal{GP}(0, k_p)$ as a Gaussian process (GP)

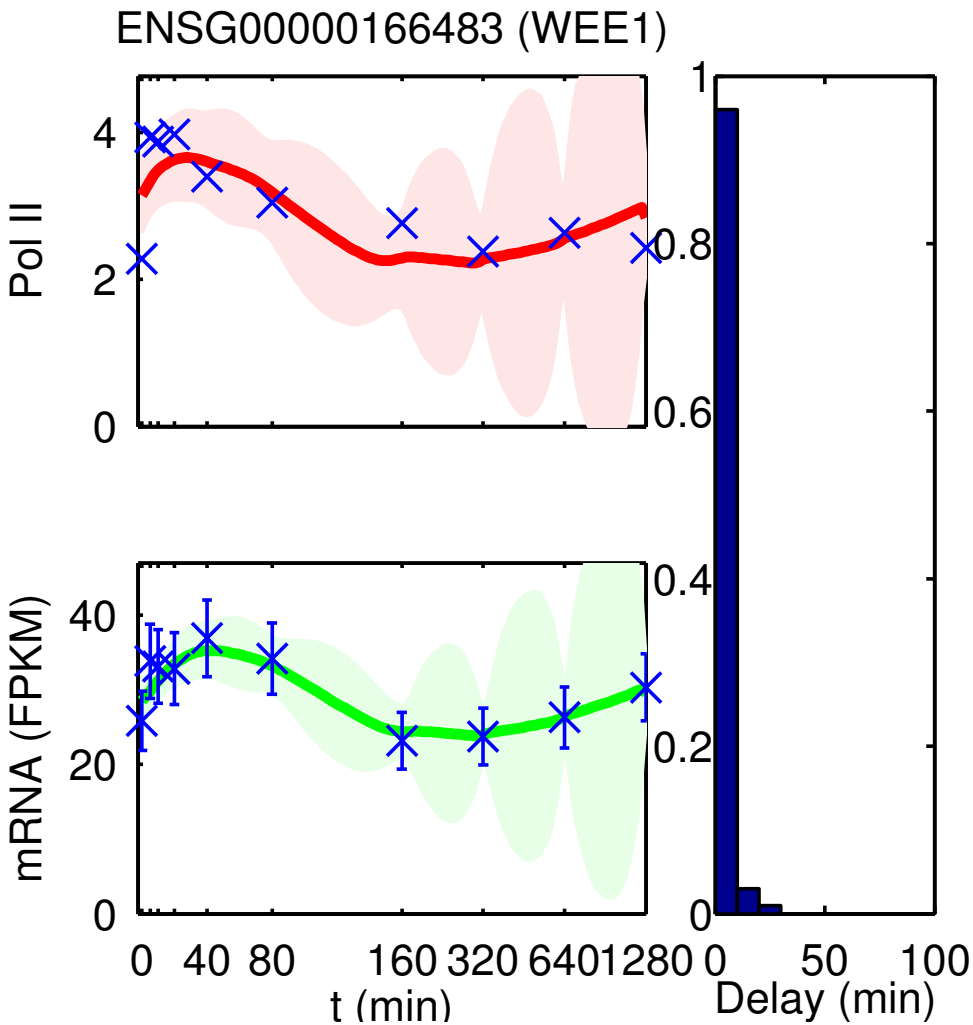
A linear operation on a GP is another GP, so that the likelihood $P(m, p|\alpha, \Delta)$ is tractable (similar to Honkela et al. *PNAS* 2010)

Bayesian MCMC used to estimate parameters α, β, Δ and GP covariance (2) and noise variance (1) parameters

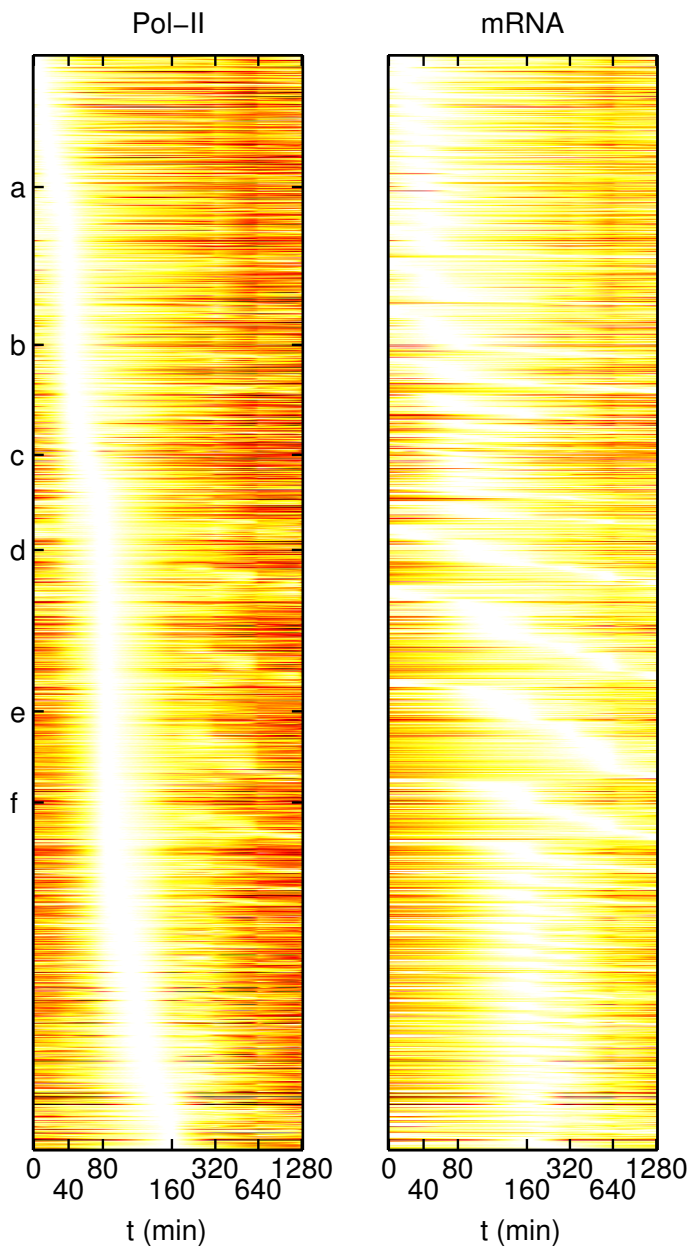
Example fits



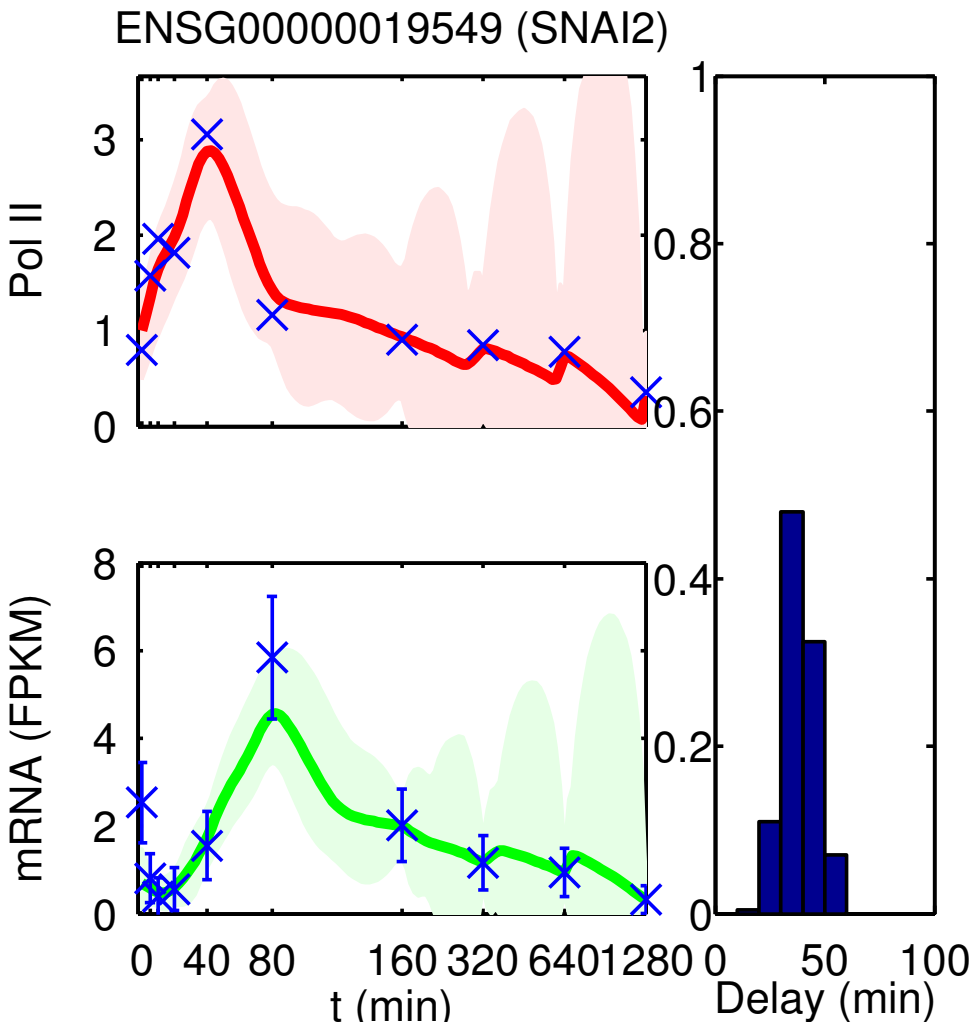
a: Early pol-II, no production delay



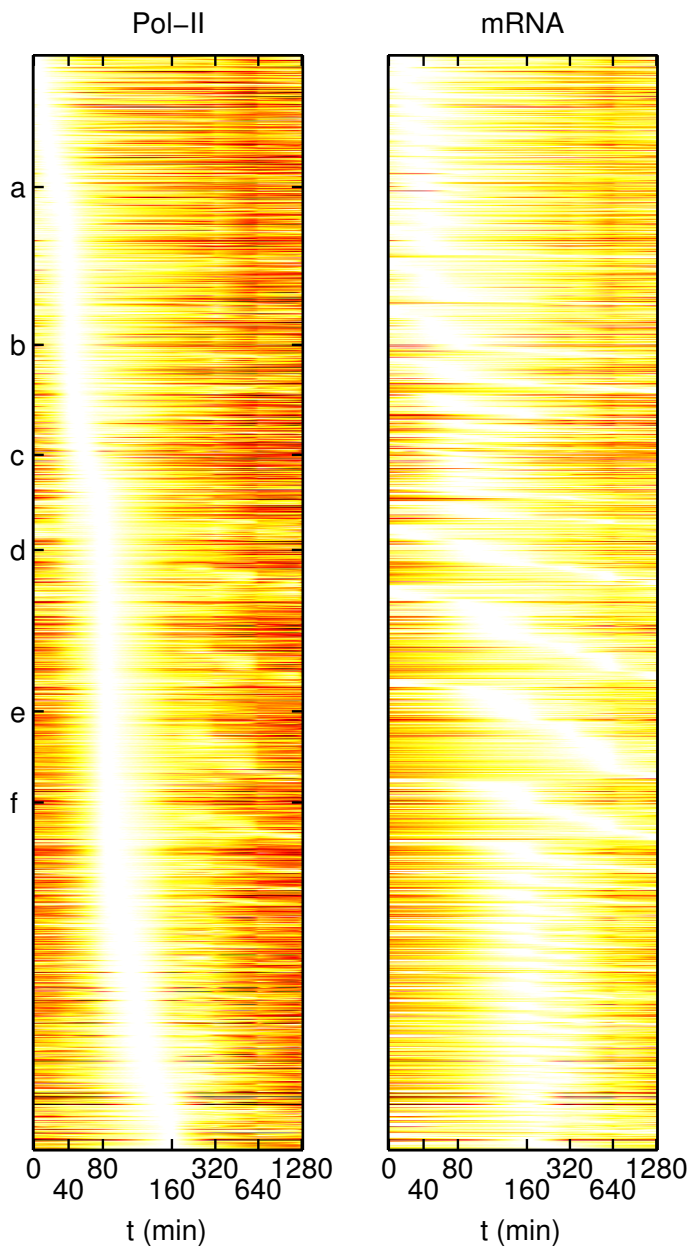
Example fits



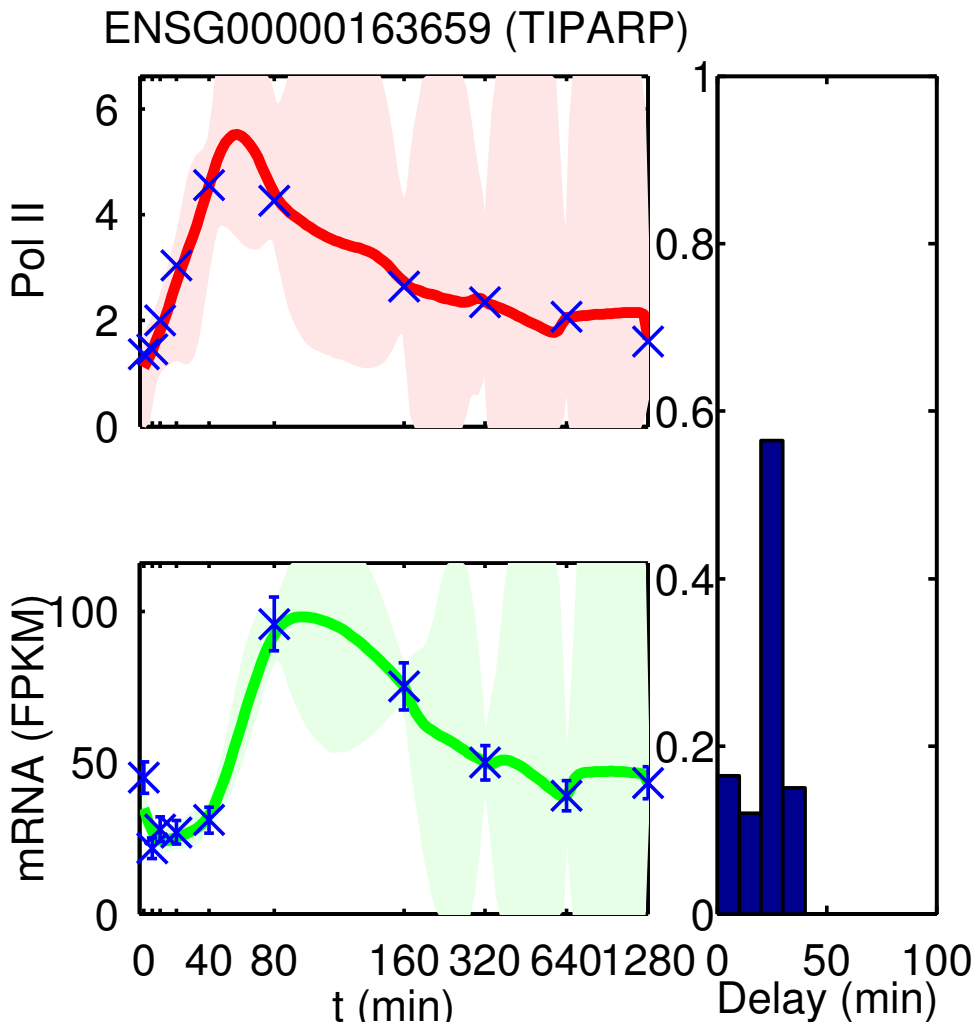
b: Early pol-II, delayed production



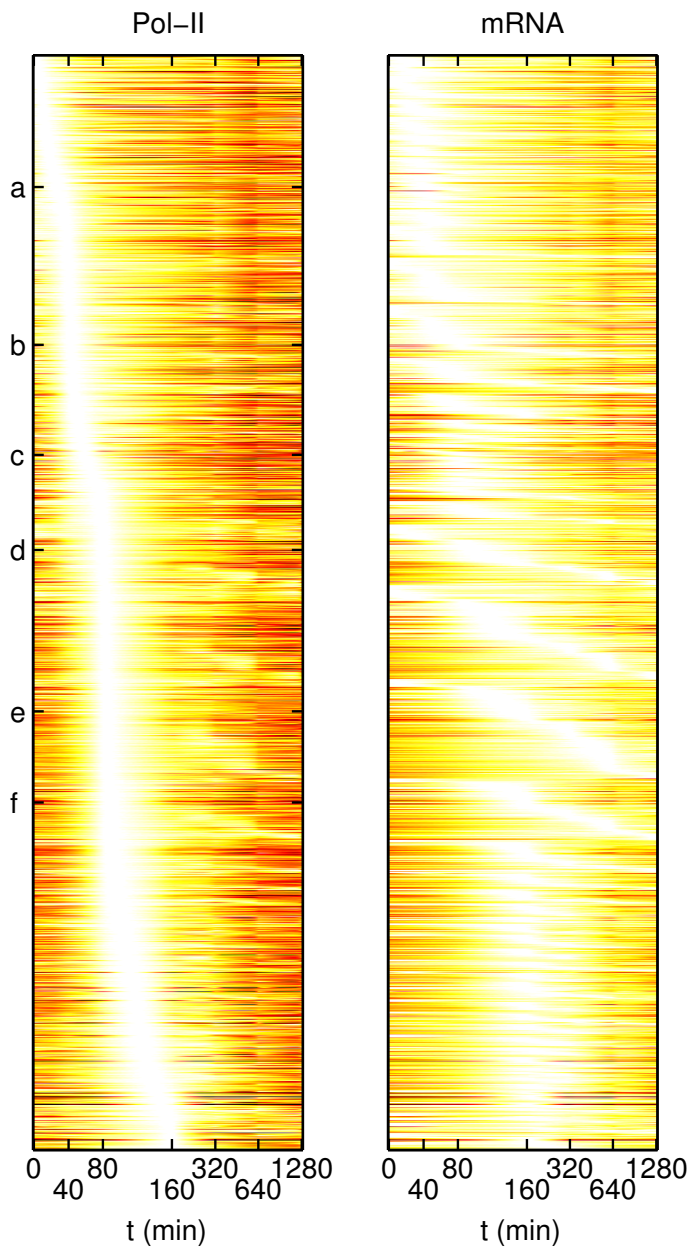
Example fits



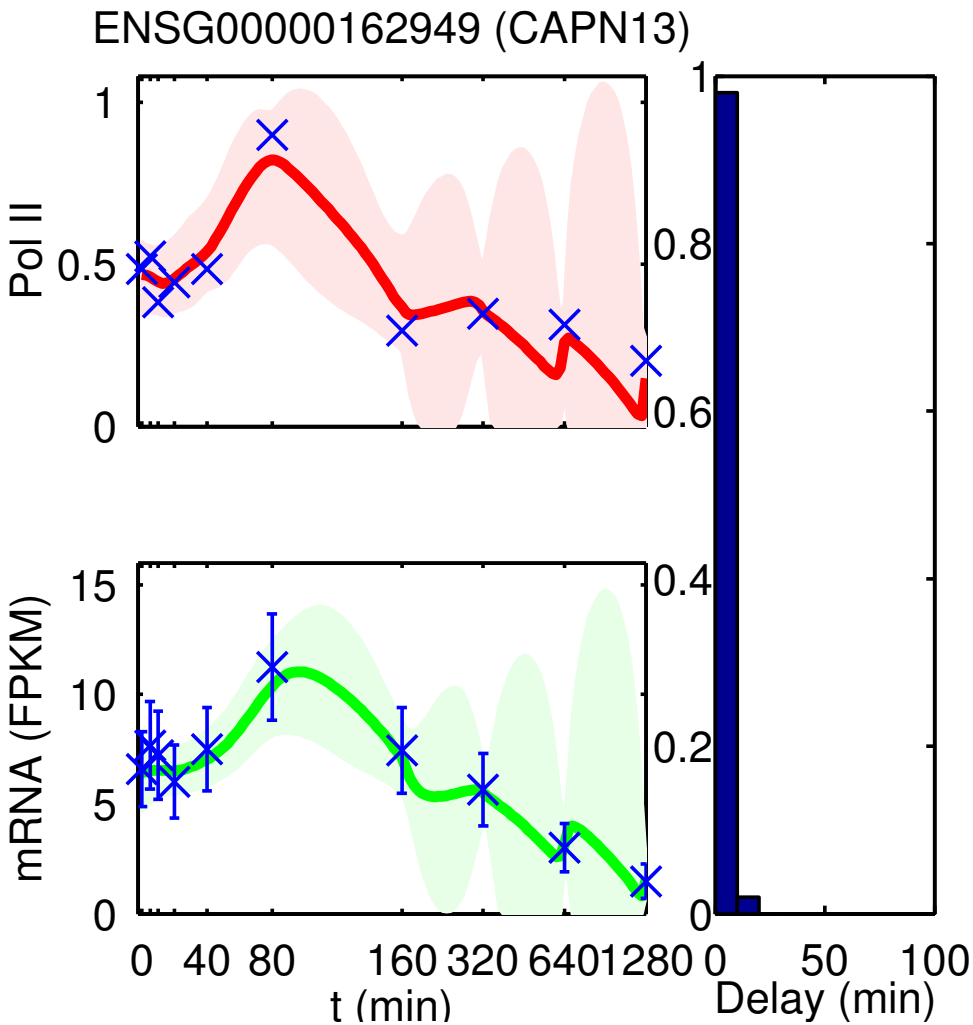
c: Later pol-II, delayed production



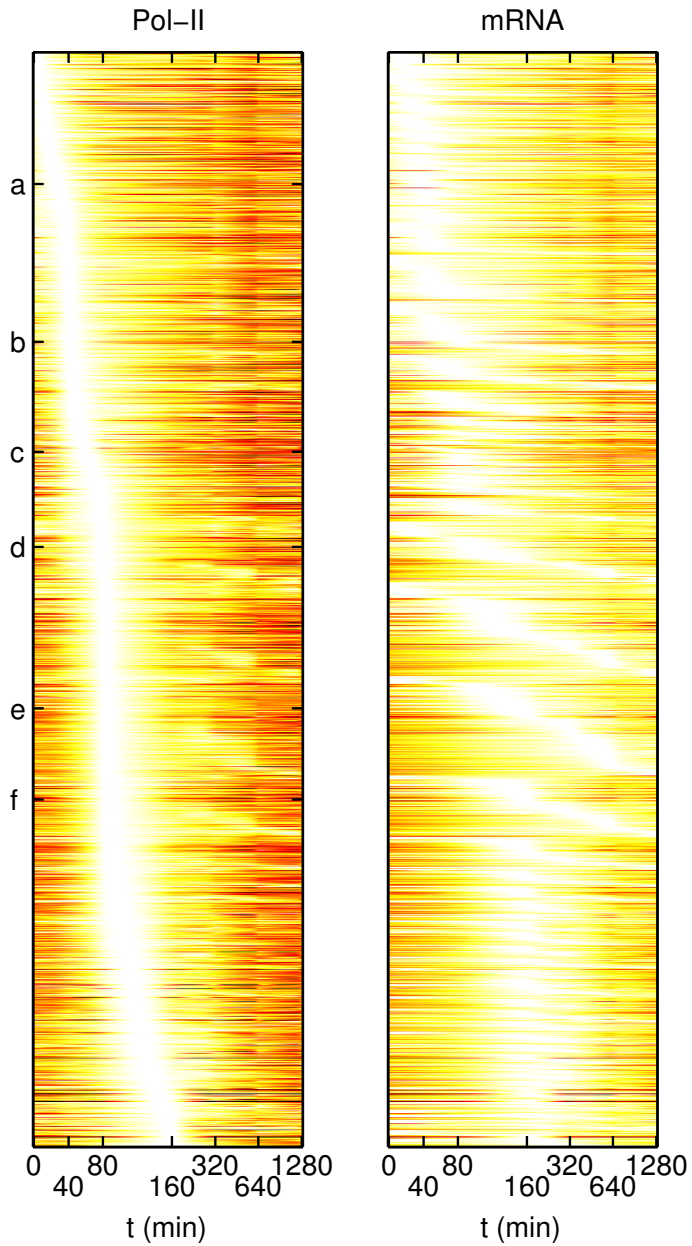
Example fits



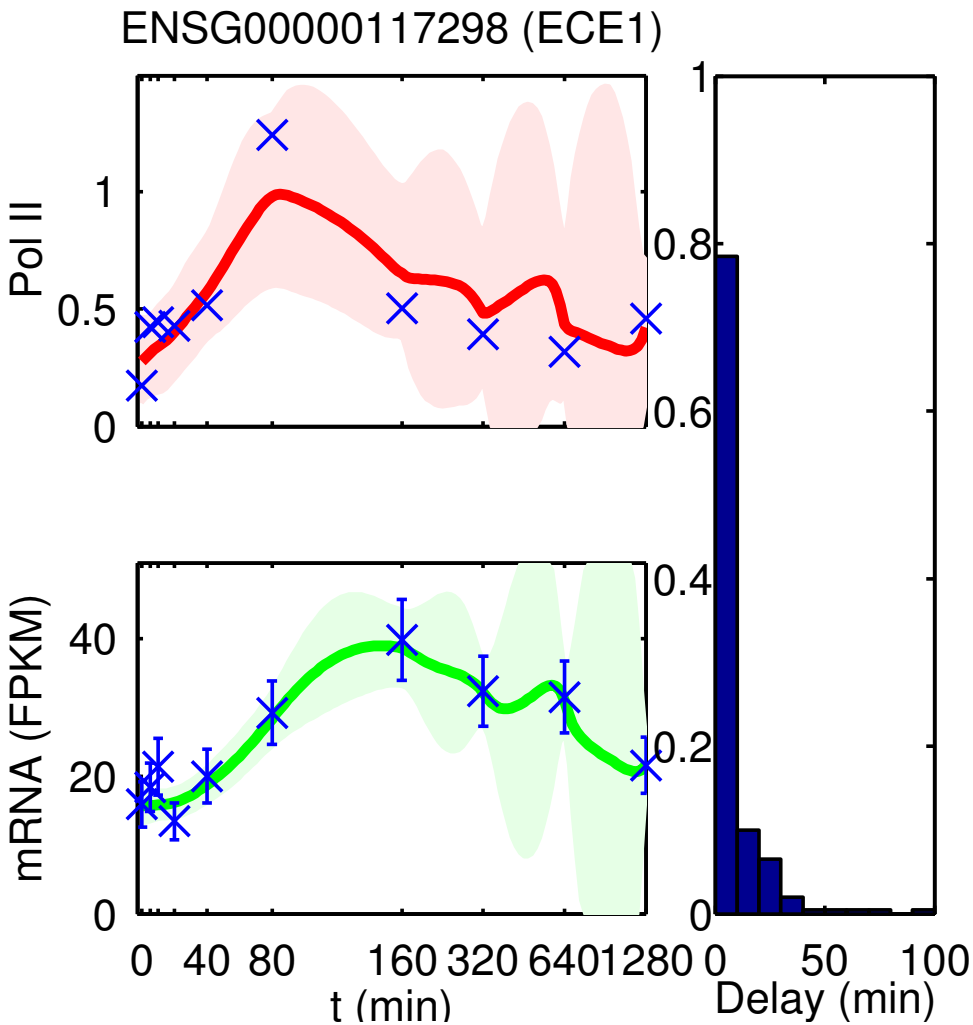
d: Late pol-II, no delay



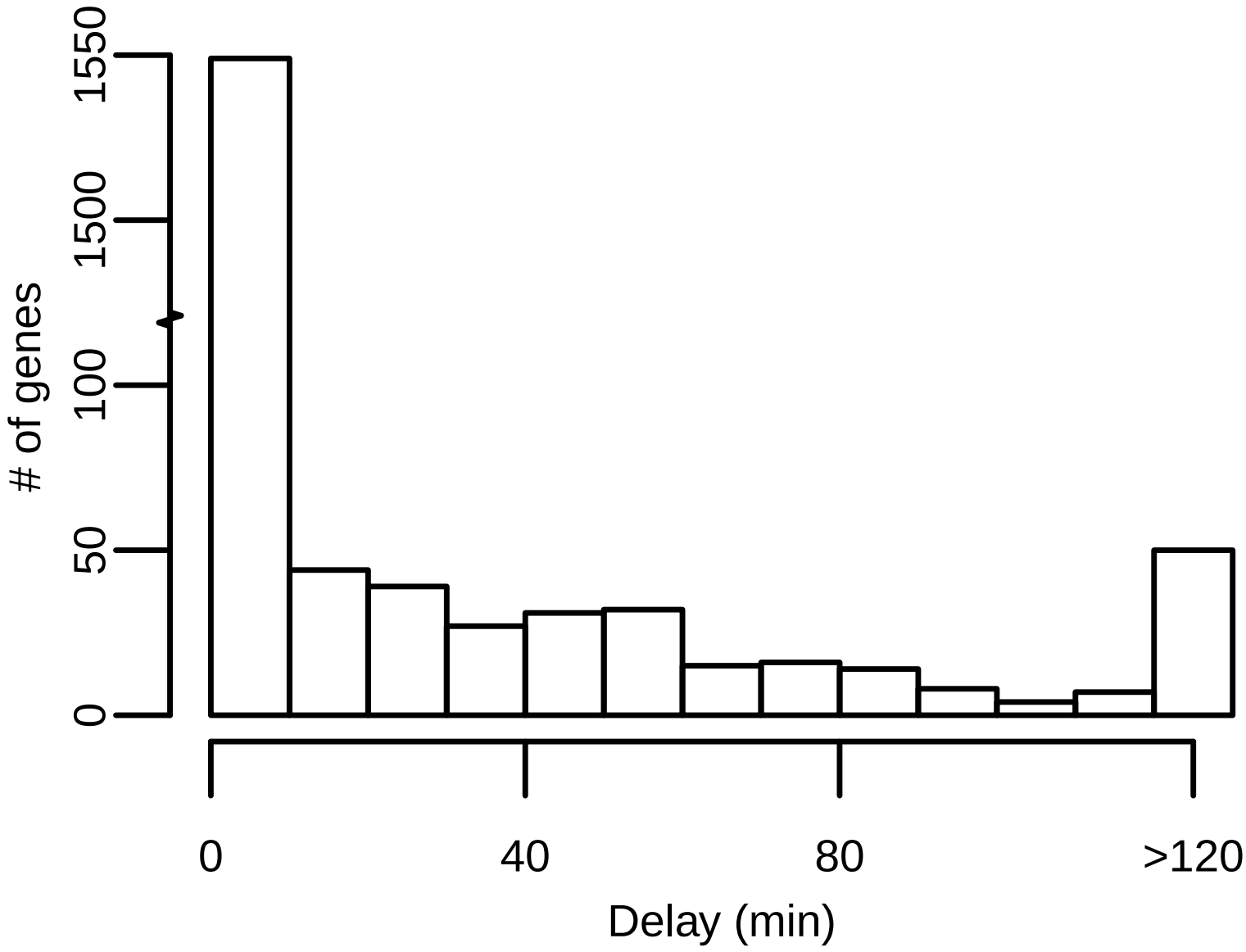
Example fits



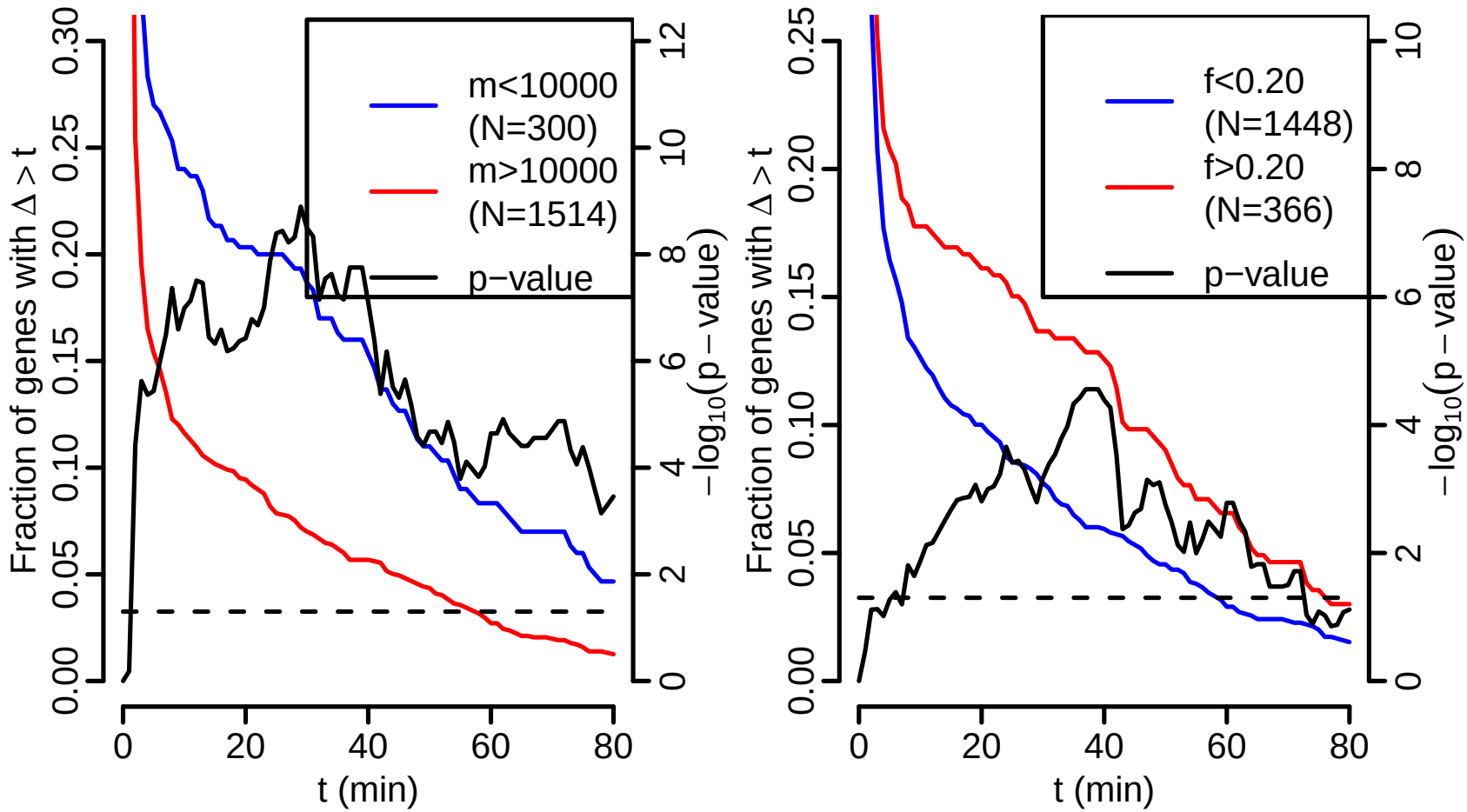
e: Late pol-II, no delay



Large processing delays observed in 11% of genes

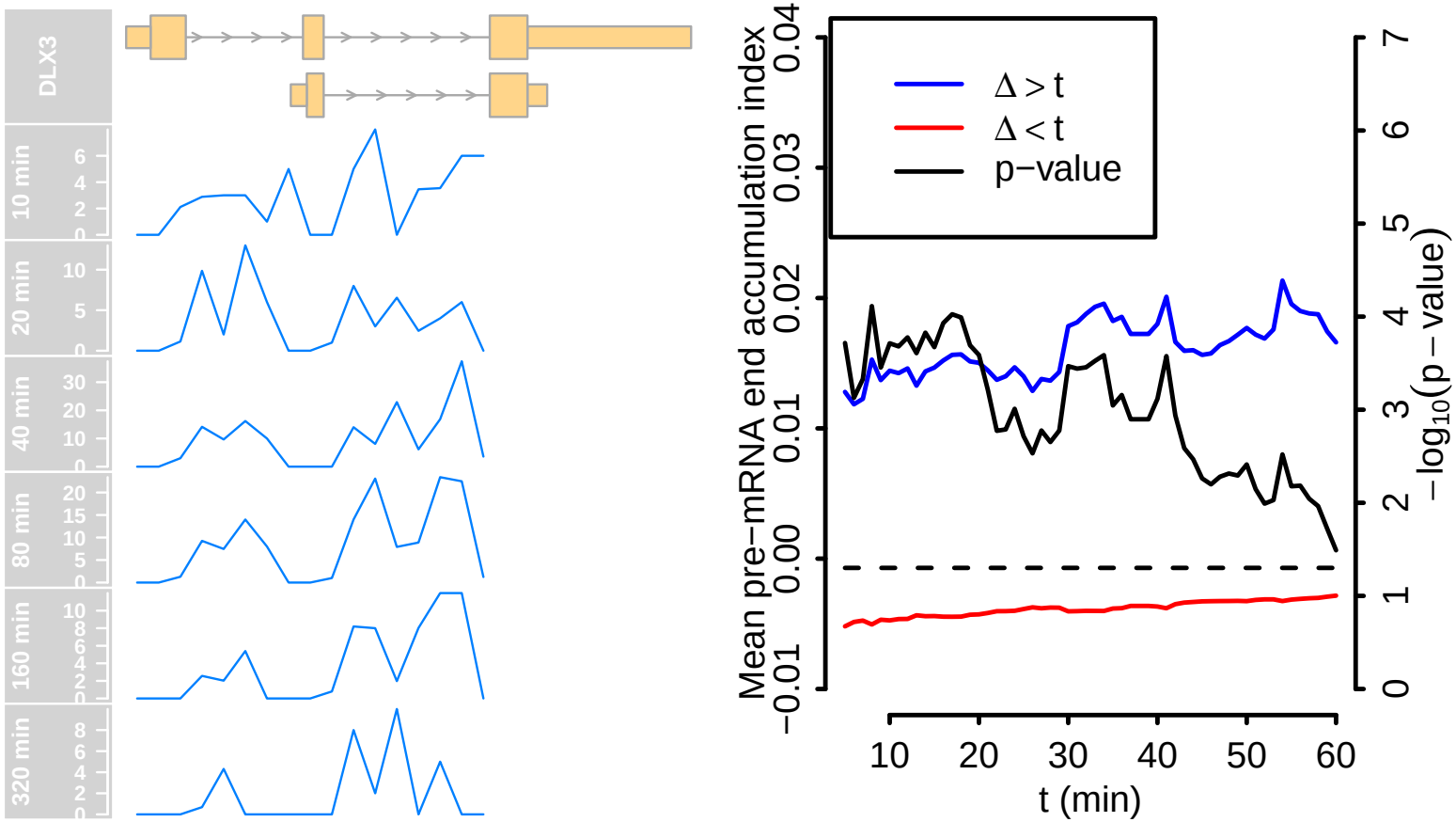


Delay linked with gene length and intron structure



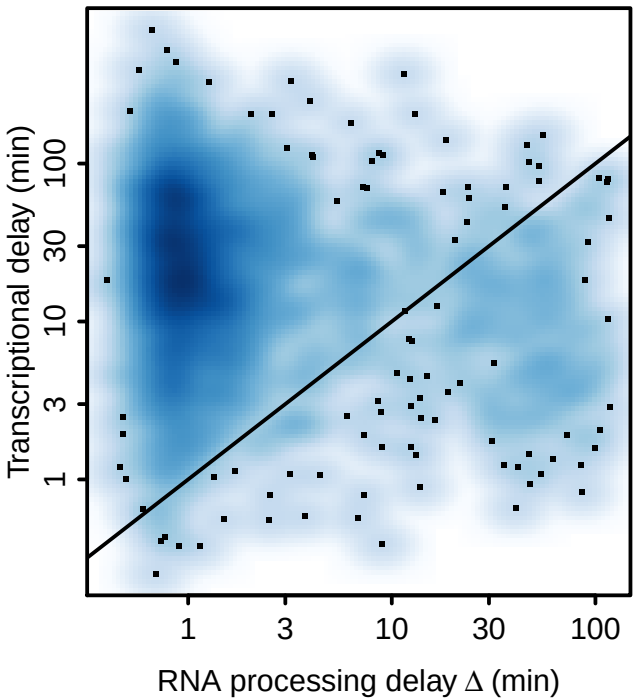
Δ: delay m: gene length f: final intron length / gene length

Delay link with splicing: evidence from intronic reads



Pre-mRNA accumulation index: ratio of intronic reads in last 50% of pre-mRNA to intronic reads in first 50% at late and early times.

Comparison of processing and elongation times



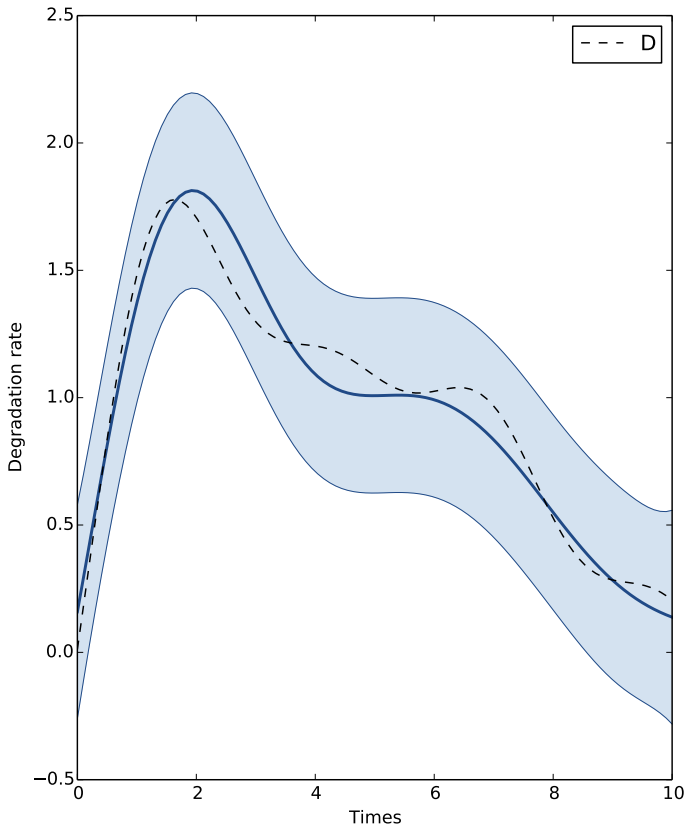
Elongation time = length/velocity estimate from Danko *Mol Cell* 2013

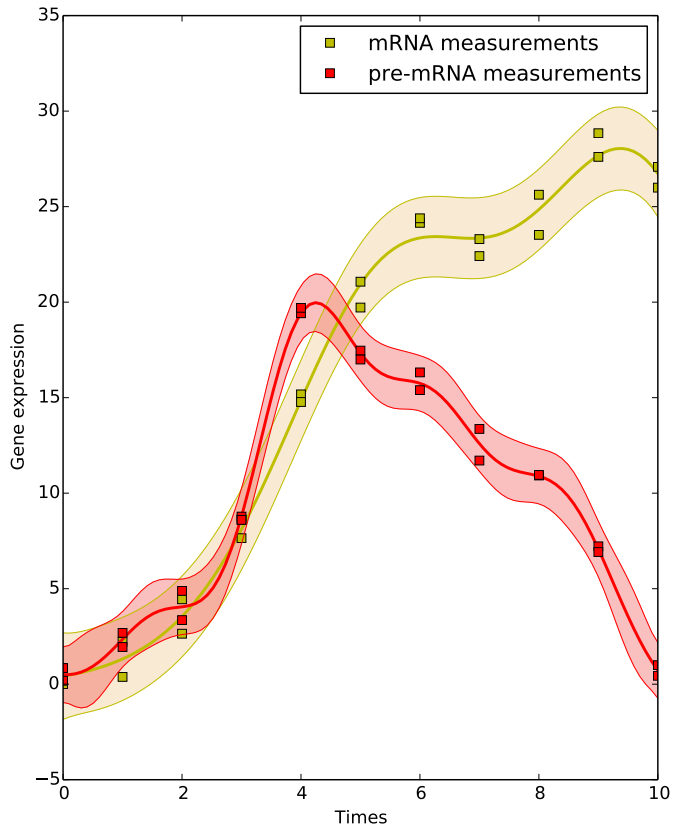
Elongation time > processing delay in 87% of genes

NB. limited to ∼1800 genes where data has enough signal to model

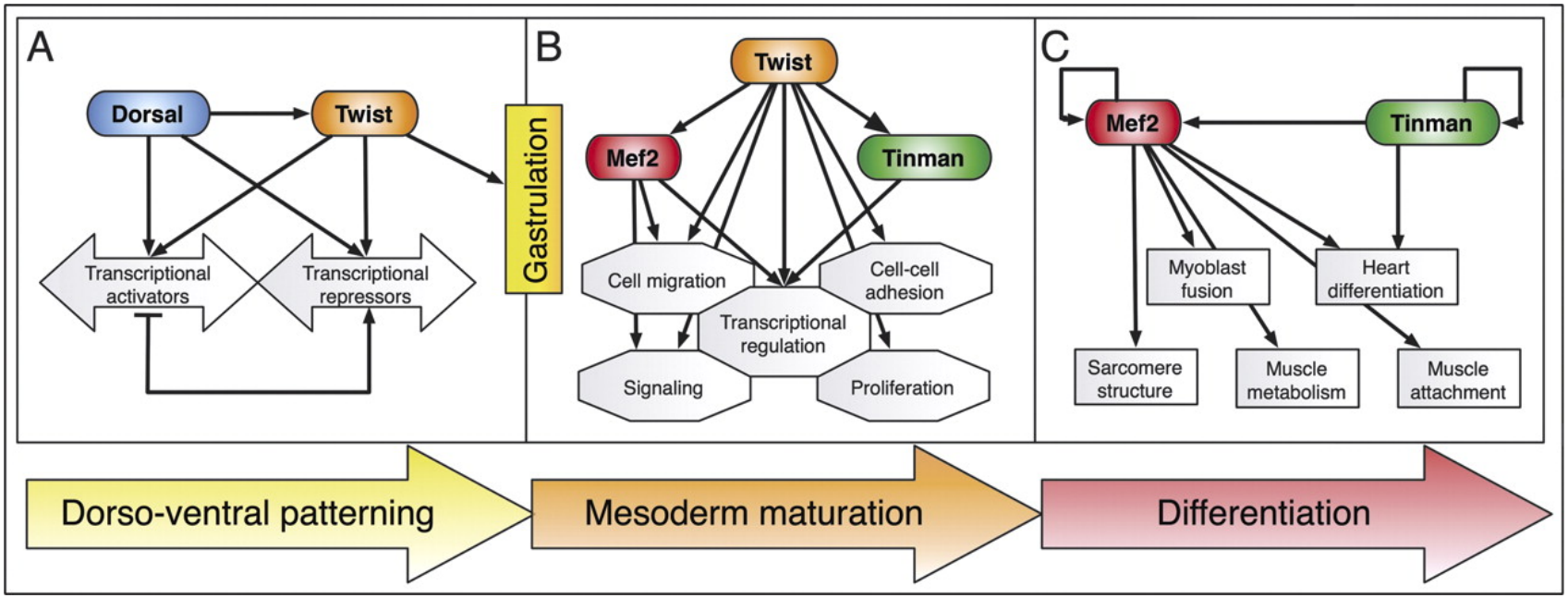
Current work - inferring time-varying degradation rates

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \beta p(t) - \alpha(t)m(t)$$





Part 3. Inferring transcriptional factor targets

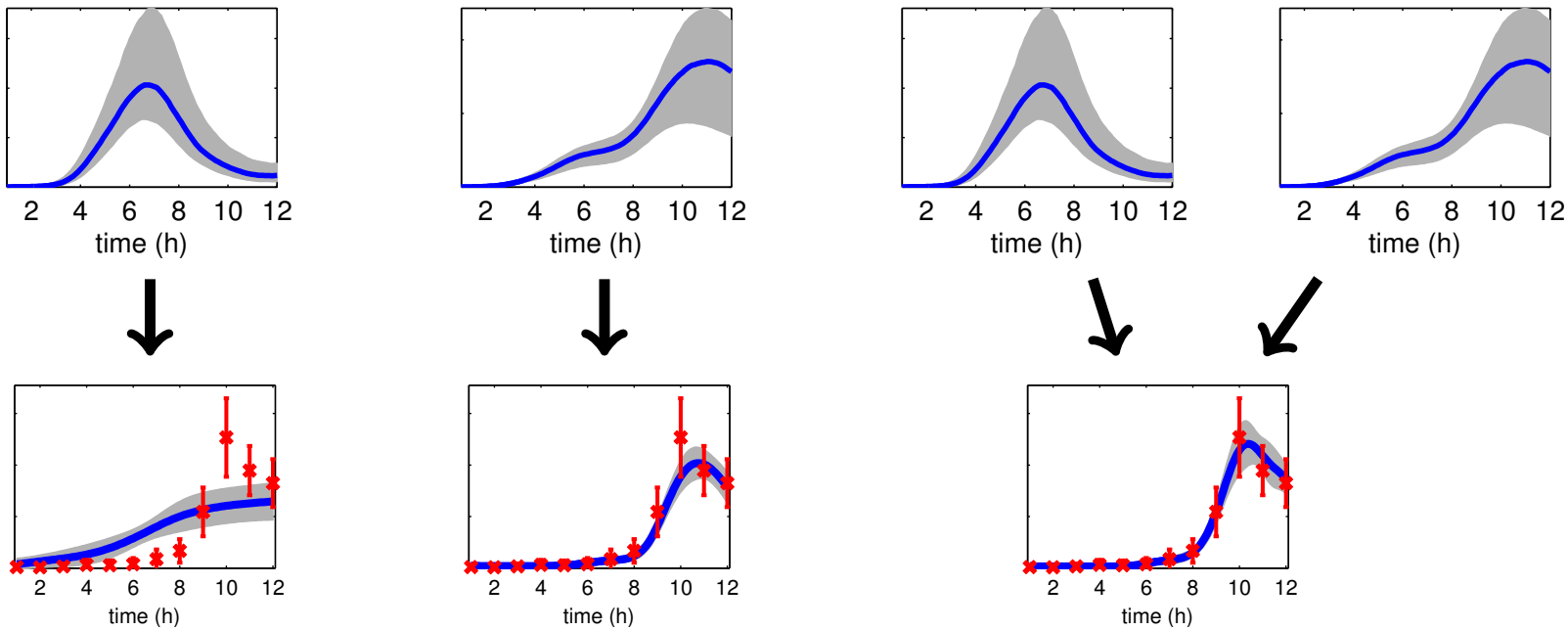


Sandmann *et al.* Genes and Development 2007

Joint work with Antti Honkela, Michalis Titsias and Neil Lawrence

Inferring targets by fitting regulation models

Which TF combinations are most likely given evidence from expression time-series data?



(a) Only BAP? (b) Only MEF2? (c) BAP & MEF2?

Bayesian model scoring trades off data fit and model complexity

Simplest case: Linear activation model

Consider a simple linear activation model

$$\begin{aligned} \frac{\mathrm{d}p(t)}{\mathrm{d}t} &= f(t) - \delta p(t) \\ \frac{\mathrm{d}m_i(t)}{\mathrm{d}t} &= B_i + S_i p(t) - D_i m_i(t) \end{aligned}$$

- $f(t)$ – concentration of transcription factor mRNA
- $p(t)$ – concentration of transcription factor protein
- $m_i(t)$ – concentration of target gene i's mRNA

Simplest case: Linear activation model

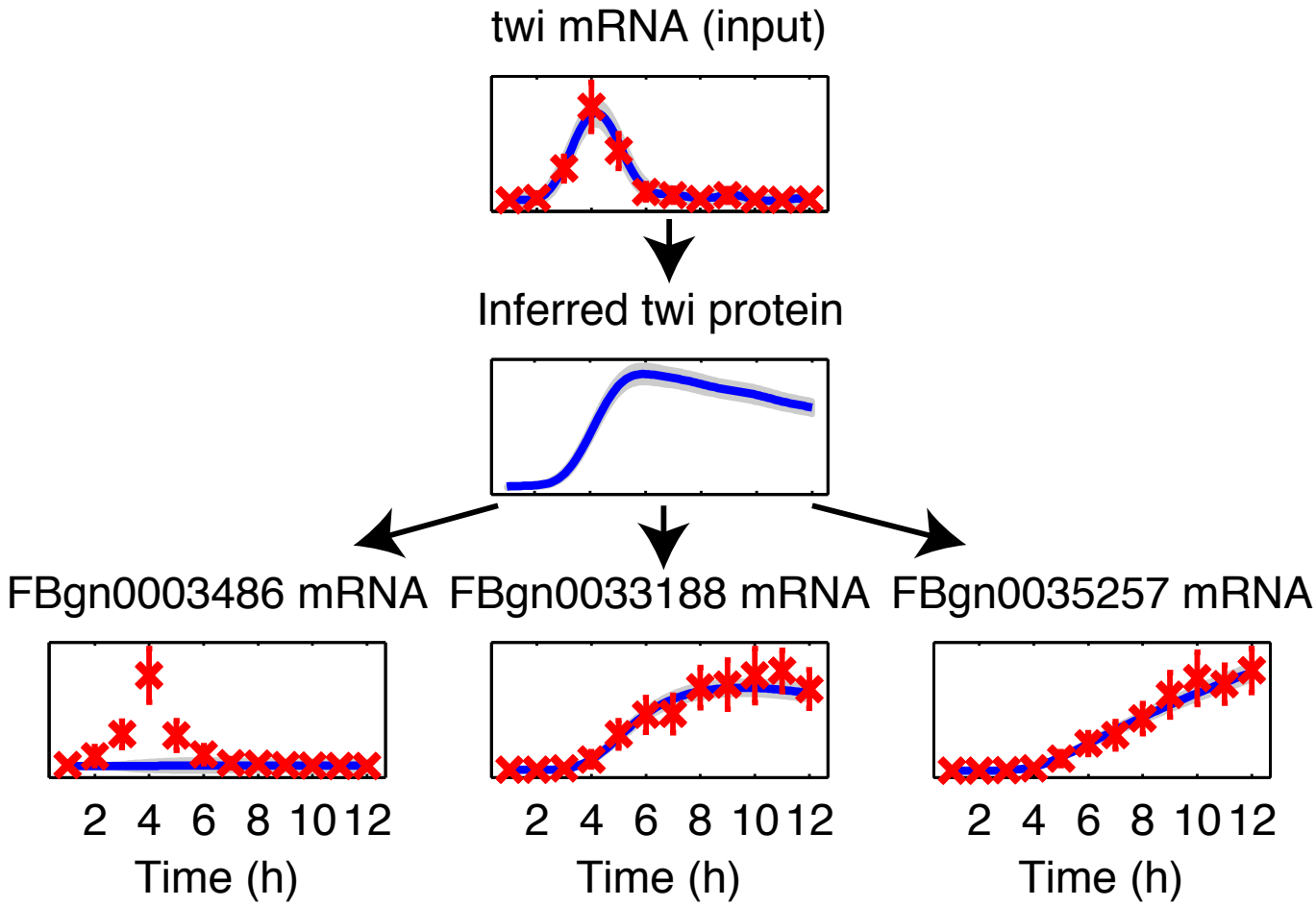
Consider a simple linear activation model

$$\begin{aligned} \frac{\mathrm{d}p(t)}{\mathrm{d}t} &= f(t) - \delta p(t) \\ \frac{\mathrm{d}m_i(t)}{\mathrm{d}t} &= B_i + S_i p(t) - D_i m_i(t) \end{aligned}$$

- $f(t)$ – concentration of transcription factor mRNA
- $p(t)$ – concentration of transcription factor protein
- $m_i(t)$ – concentration of target gene i's mRNA

We model $f(t) \sim \mathcal{GP}(0, k_f)$ as a Gaussian process

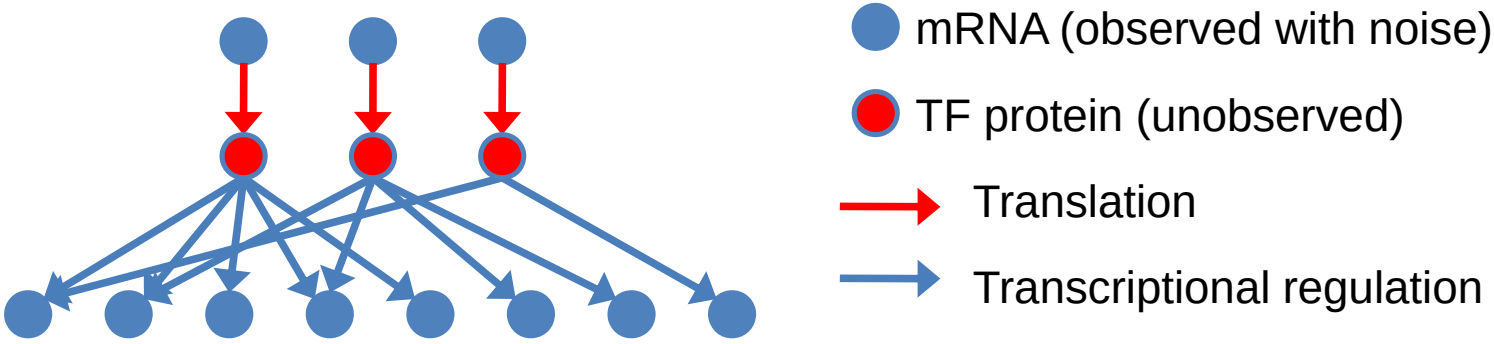
Simplest case: Linear activation model



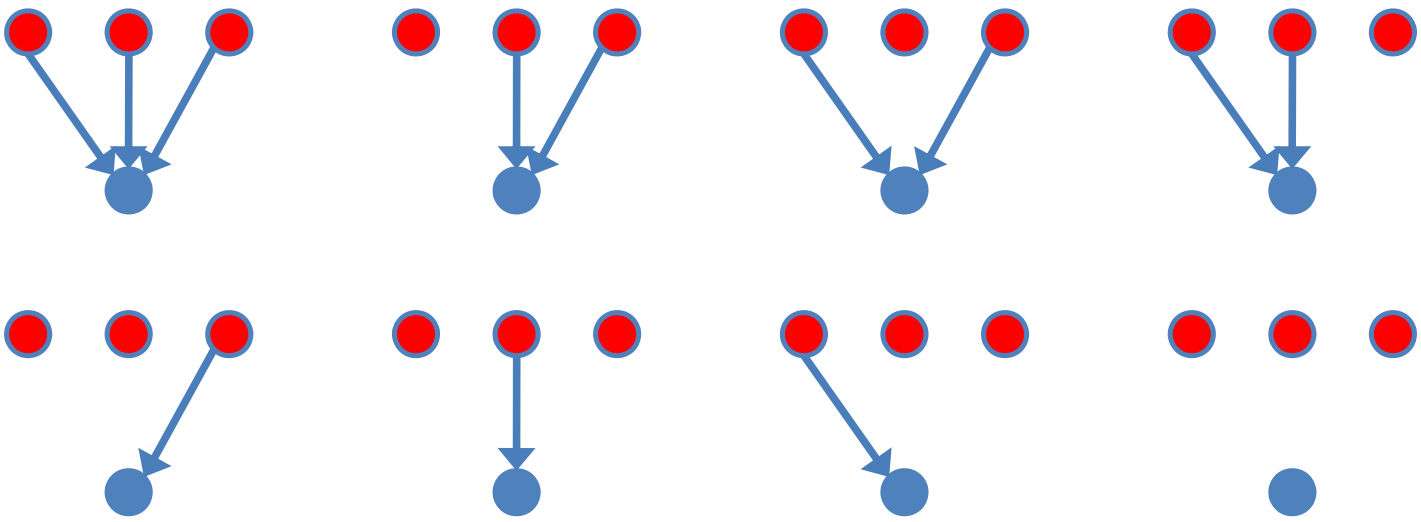
Honkela et al. *PNAS* 2010

Non-linear extension for multiple TFs

(a): Training phase



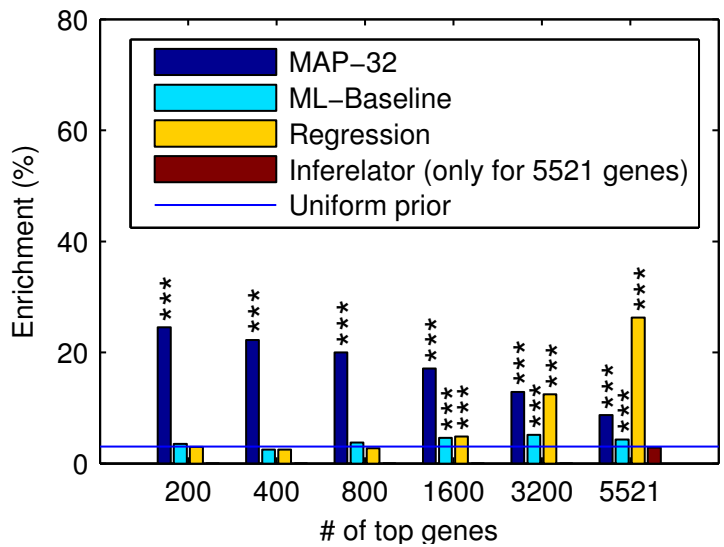
(b): Prediction phase



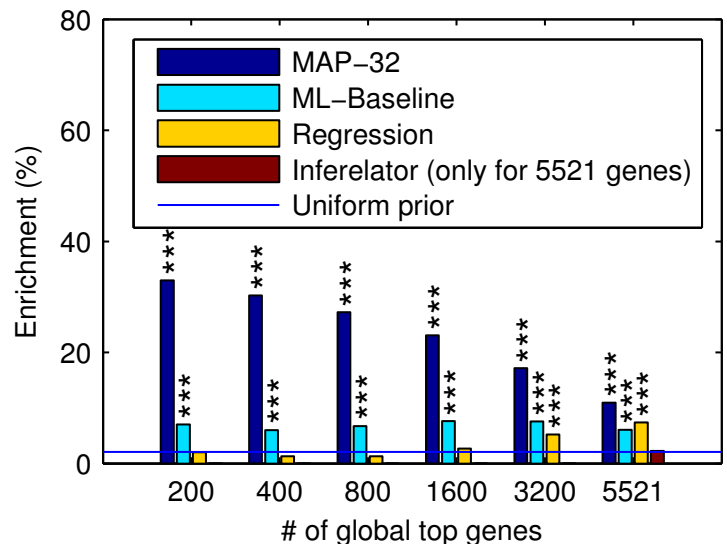
Titsias et al. *BMC Systems Biology* 2012

Referee requested validation on an external database

- Model developed for 5 TFs in early mesoderm development
- Genes ranked according to posterior probability of best model
- Validate predicted links using the DroID database
- Percentage of genes with the "correct" model



(a) Validating both positive and negative predictions



(b) Validating only positive predictions

Conclusions

- Gaussian processes provide a flexible model of temporal profiles and require estimation of only a few parameters
- We have used Gaussian processes to model transcription in several different models:
  1. elongation dynamics model was used to infer transcription time and promoter activity; model deals with "spreading-out" of signal in time; no assumption of constant velocity.
  2. mRNA production model was used to identify significant processing delays; found to associate with splicing.
  3. transcriptional regulation model was used to infer transcription factor activities and regulatory network links.
- Personalised medicine link - T-cell transcription dynamics, enhancer-mediated regulation, cytokine dynamics...

Funding: BBSRC (EraSysBio+ SYNERGY), EU FP7 (RADIANT)