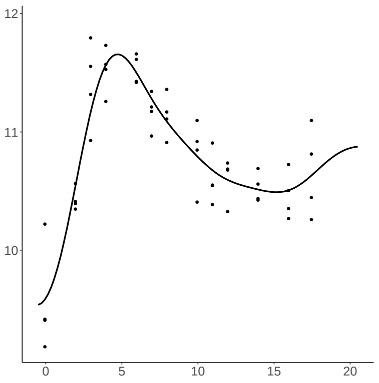
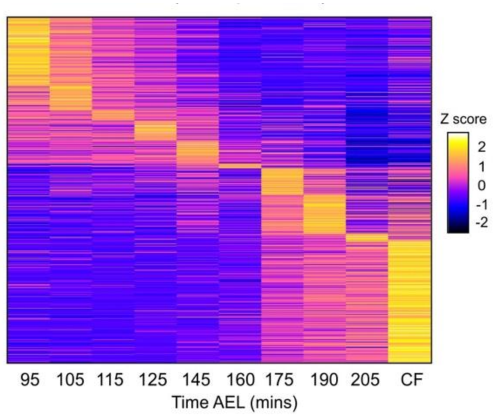


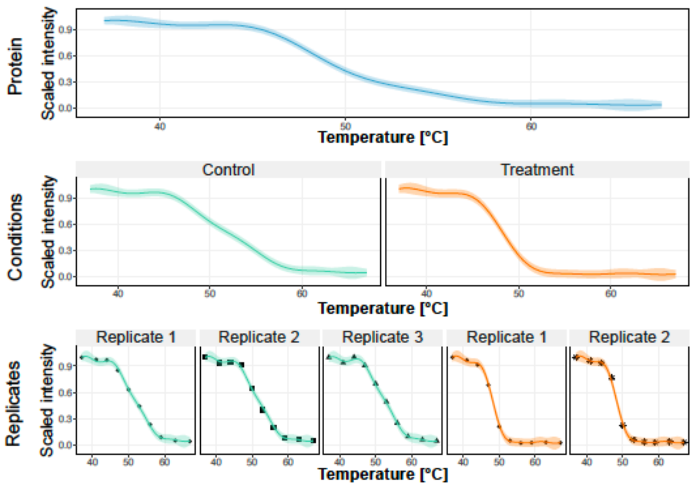


MANCHESTER
1824
The University of Manchester

Biological applications of Gaussian process modelling

Magnus Rattray, University of Manchester

Gaussian Process Summer School
10th September 2024, Manchester



wellcome

Talk outline

Biological applications:

(1) Differential gene expression

(2) Protein melting curves

(3) mRNA production and degradation

(4) Single-cell pseudotime and branching

Talk outline

Biological applications:

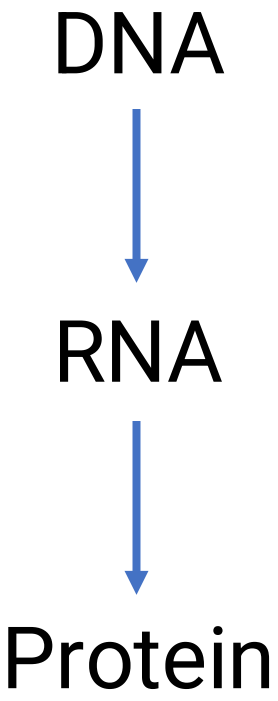
(1) Differential gene expression

(2) Protein melting curves

(3) mRNA production and degradation

(4) Single-cell pseudotime and branching

Gene expression



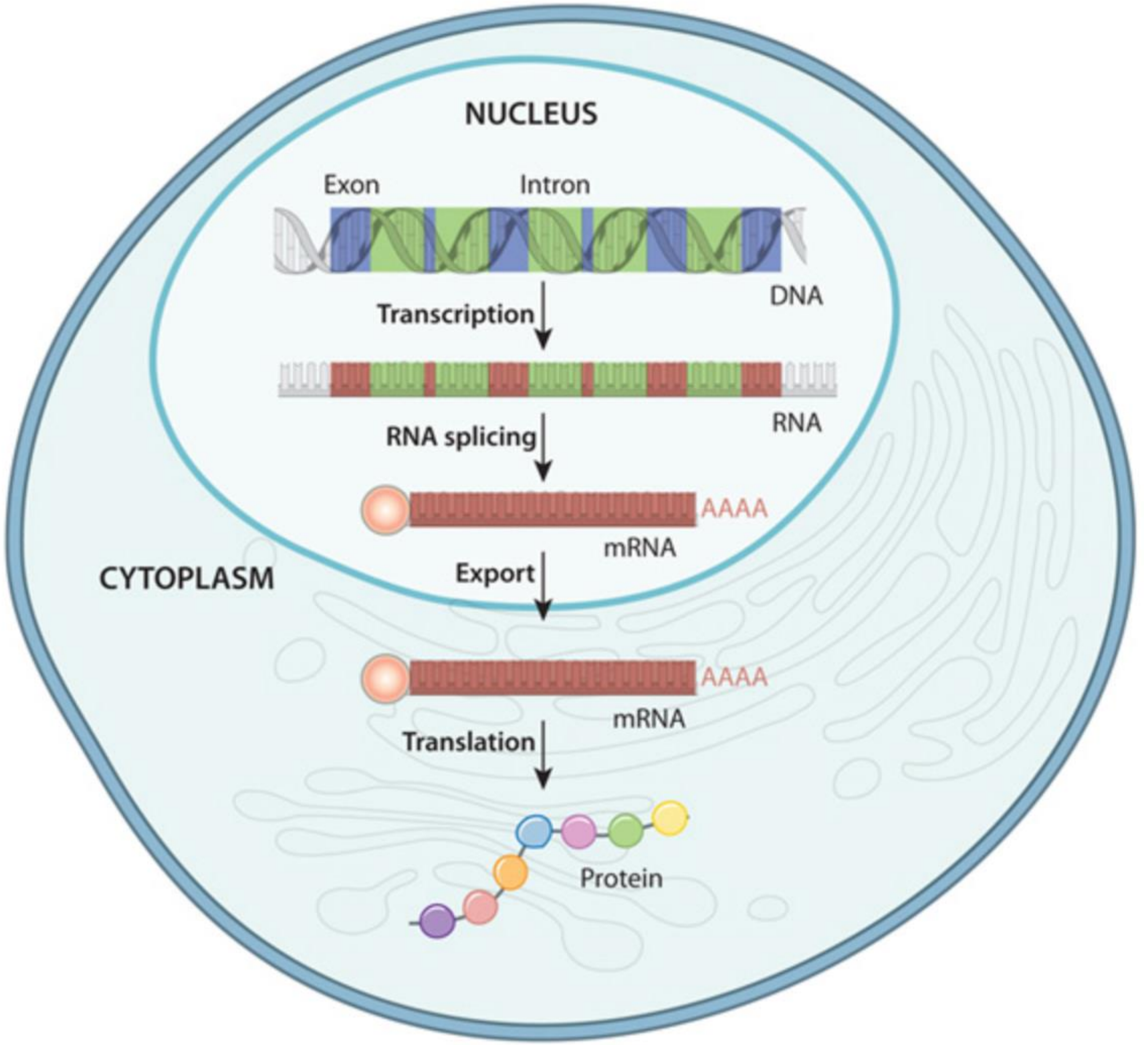
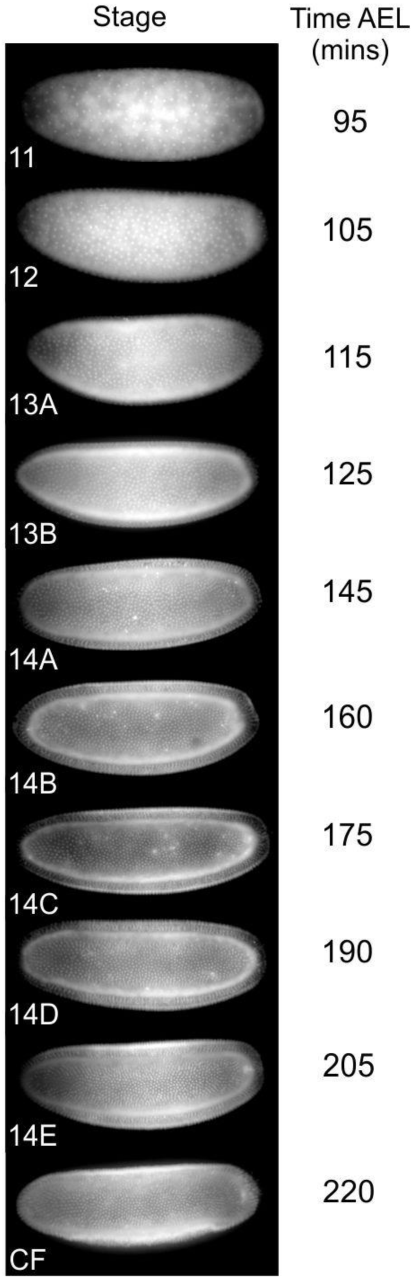
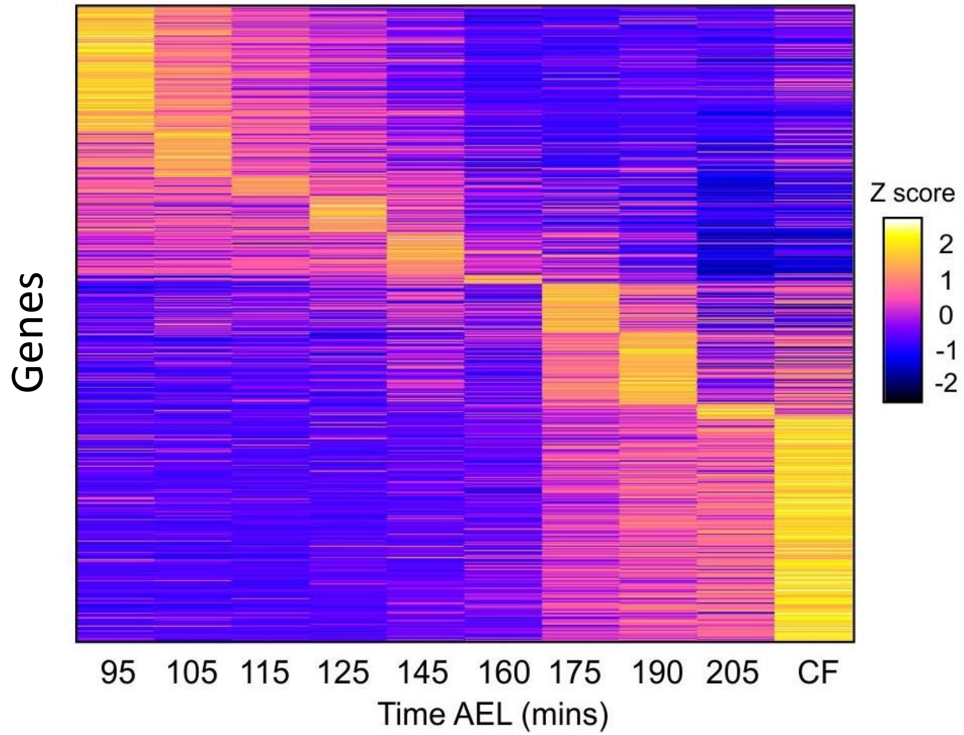
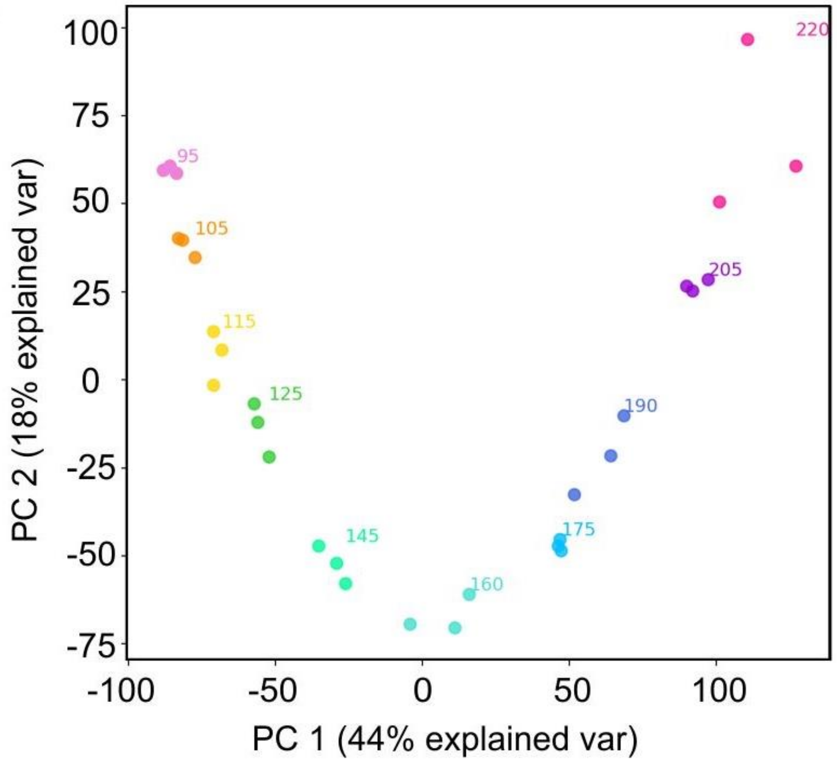


Figure 1: An overview of the flow of information from DNA to protein in a eukaryote
First, both coding and noncoding regions of DNA are transcribed into mRNA. Some regions are removed (introns) during initial mRNA processing. The remaining exons are then spliced together, and the spliced mRNA molecule (red) is prepared for export out of the nucleus through addition of an endcap (sphere) and a polyA tail. Once in the cytoplasm, the mRNA can be used to construct a protein.

Gene expression time course data help us understand how genes switch on and off during a biological process



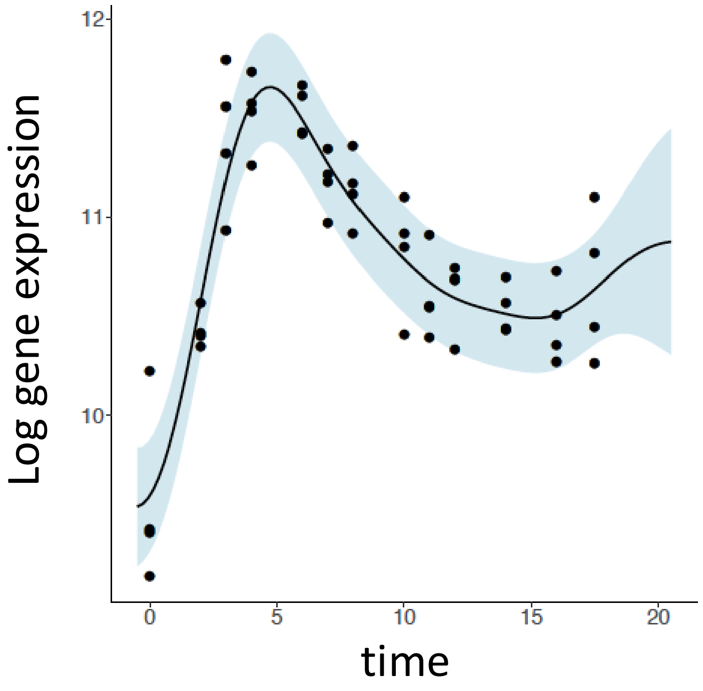


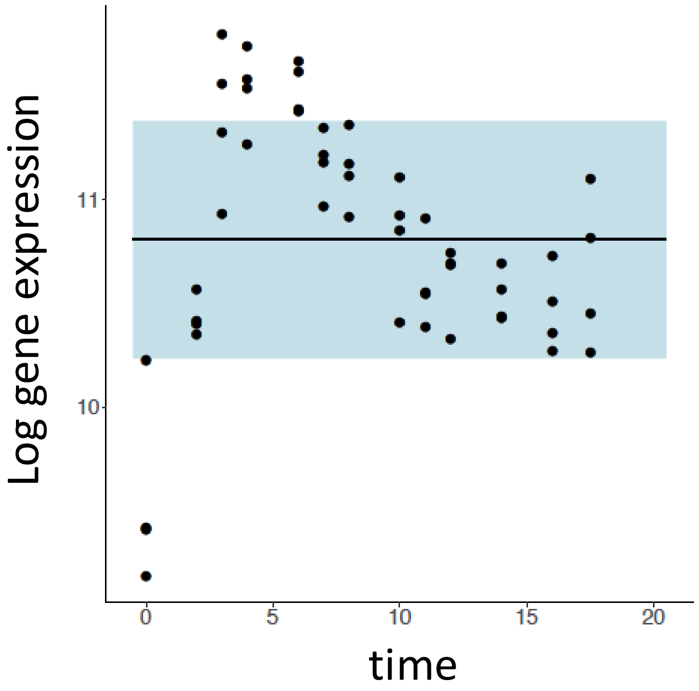


Differential gene expression – one sample test

Data are noisy and high-dimensional (e.g. 20K genes) with signal-to-noise varying by orders of magnitude

Gaussian processes are useful for identifying genes with evidence of differential expression





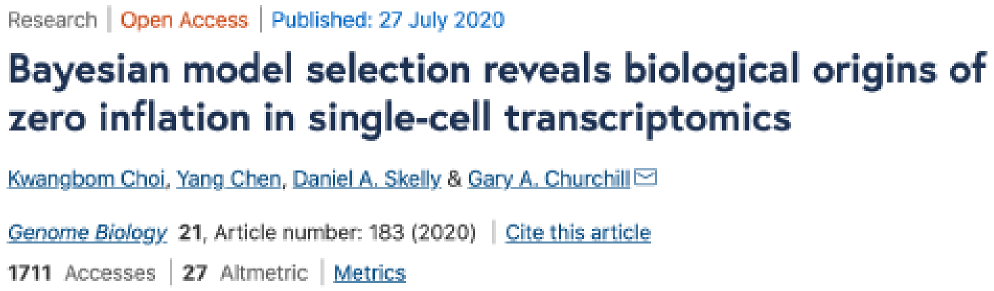
Test statistic: $\text{LLR} = \log P(Y|\text{dynamic}) - \log P(Y|\text{constant})$

RESEARCH ARTICLE Open Access

A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression

Alfredo A Kalaitzis* and Neil D Lawrence*

Modelling counts data from RNA-sequencing



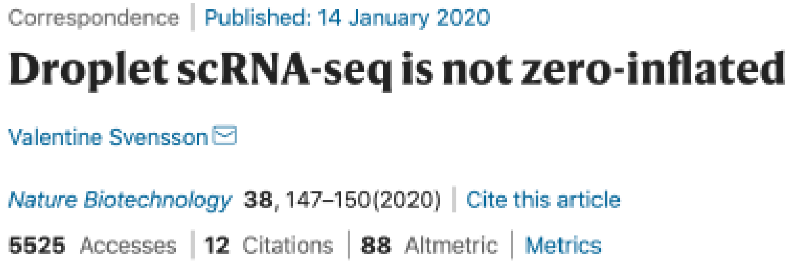
Research | Open Access | Published: 27 July 2020

Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

Kwangbom Choi, Yang Chen, Daniel A. Skelly & Gary A. Churchill

Genome Biology **21**, Article number: 183 (2020) | Cite this article

1711 Accesses | **27** Altmetric | Metrics



Correspondence | Published: 14 January 2020

Droplet scRNA-seq is not zero-inflated

Valentine Svensson

Nature Biotechnology **38**, 147–150(2020) | Cite this article

5525 Accesses | **12** Citations | **88** Altmetric | Metrics

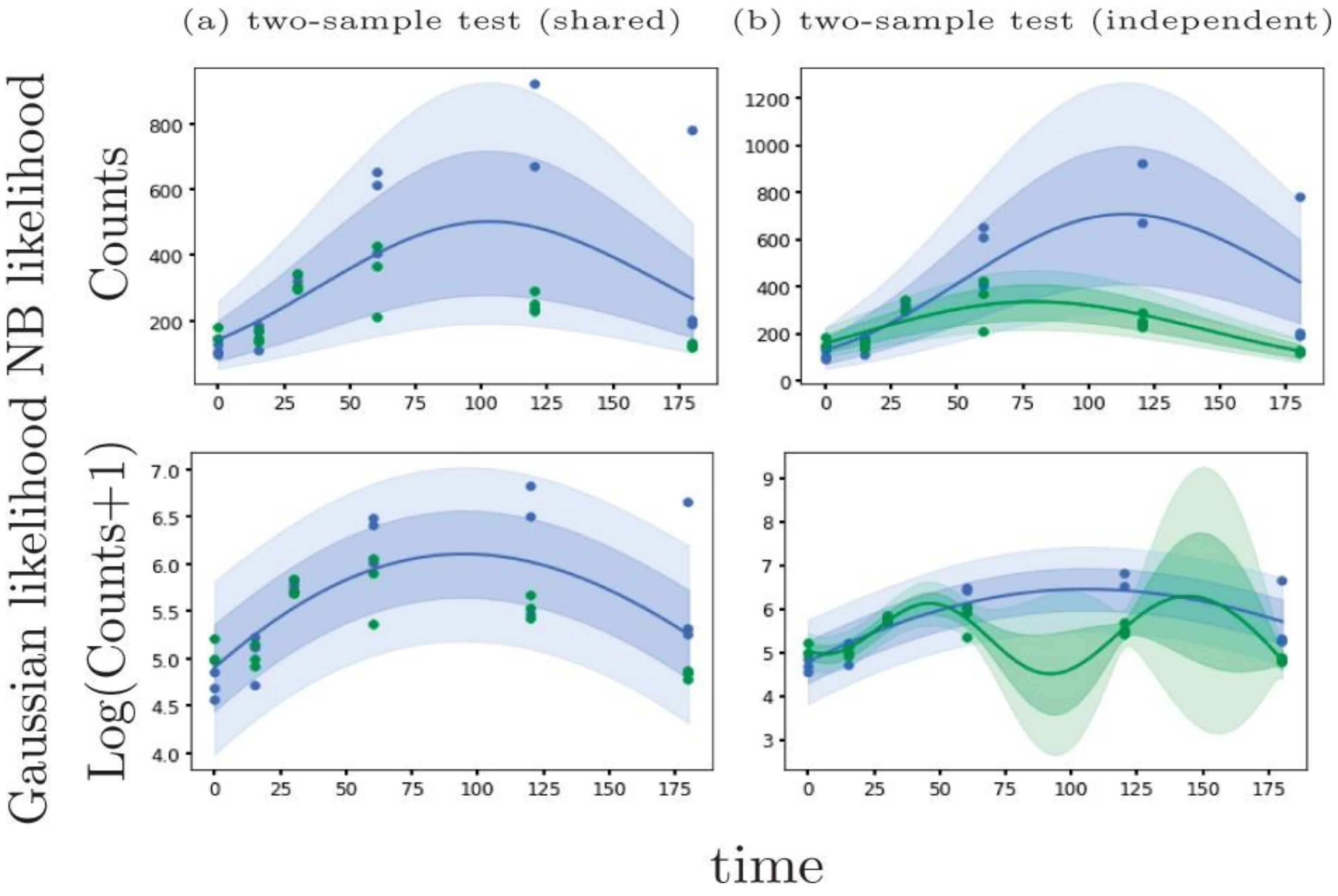
$$\mathrm{NB}(y;\mu,r) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\left(\frac{r}{r+\mu}\right)^{r}\left(\frac{\mu}{r+\mu}\right)^{y}, \quad \forall y \in \mathbb{N}$$

Dispersion $\alpha = r^{-1}$ captures excess variance relative to a Poisson

$$\mathrm{Var}[y] = \mu + \alpha\mu^2$$

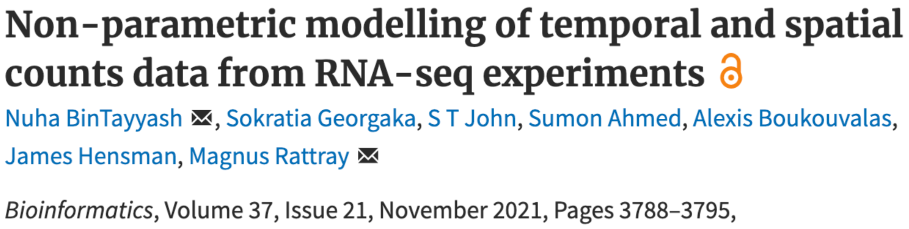
We use logarithmic link function $f(x) = \log \mu(x)$ and $f \sim \mathcal{GP}(0, k)$

Differential gene expression – two sample test



Counts likelihood
(negative binomial)

Gaussian likelihood

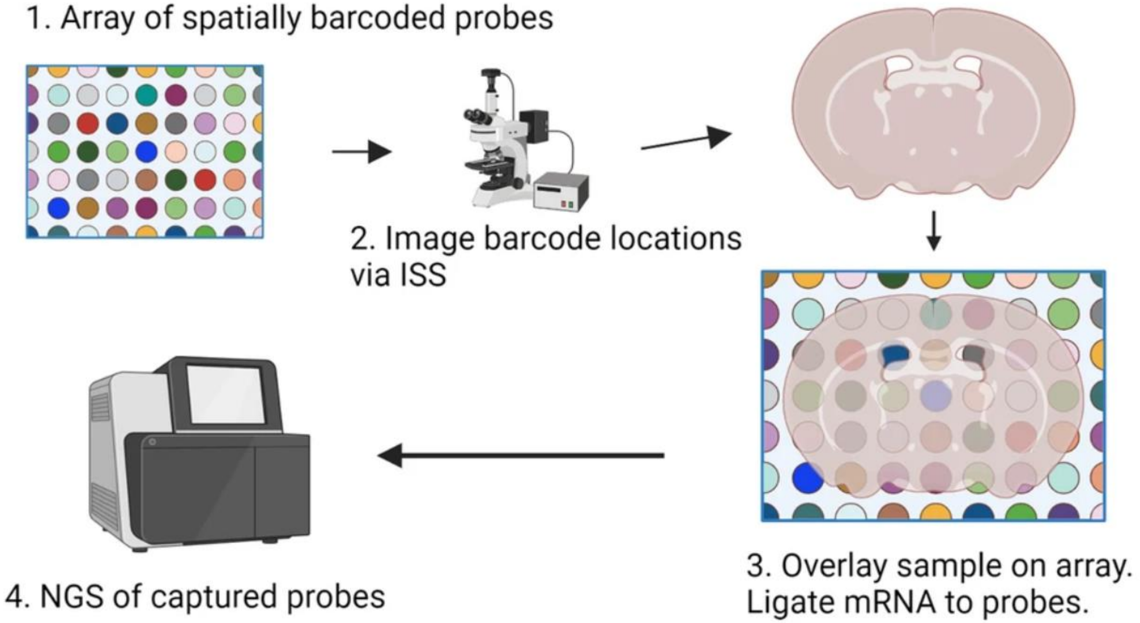


Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments

Nuha BinTayyash, Sokratia Georgaka, S T John, Sumon Ahmed, Alexis Boukouvalas, James Hensman, Magnus Rattray

Bioinformatics, Volume 37, Issue 21, November 2021, Pages 3788–3795,

Differential gene expression – spatial



1. Array of spatially barcoded probes

2. Image barcode locations via ISS

3. Overlay sample on array. Ligate mRNA to probes.

4. NGS of captured probes

An introduction to spatial transcriptomics for biomedical research

Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo & Ashraful Haque

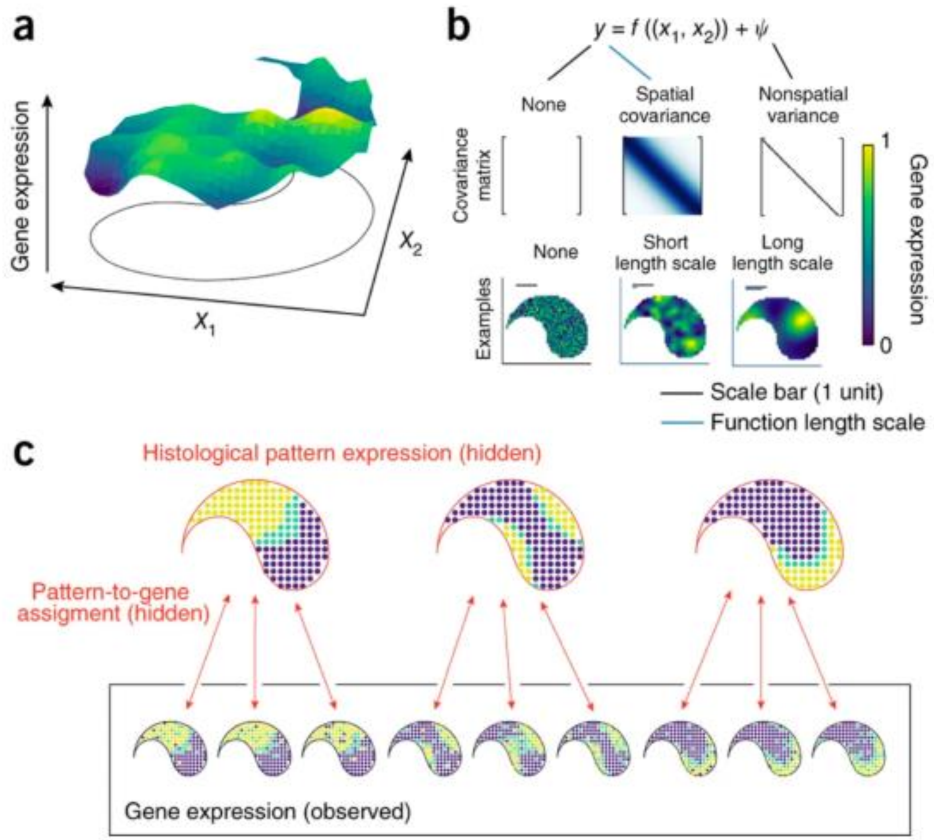
Genome Medicine **14**, Article number: 68 (2022) | Cite this article

Published: 19 March 2018

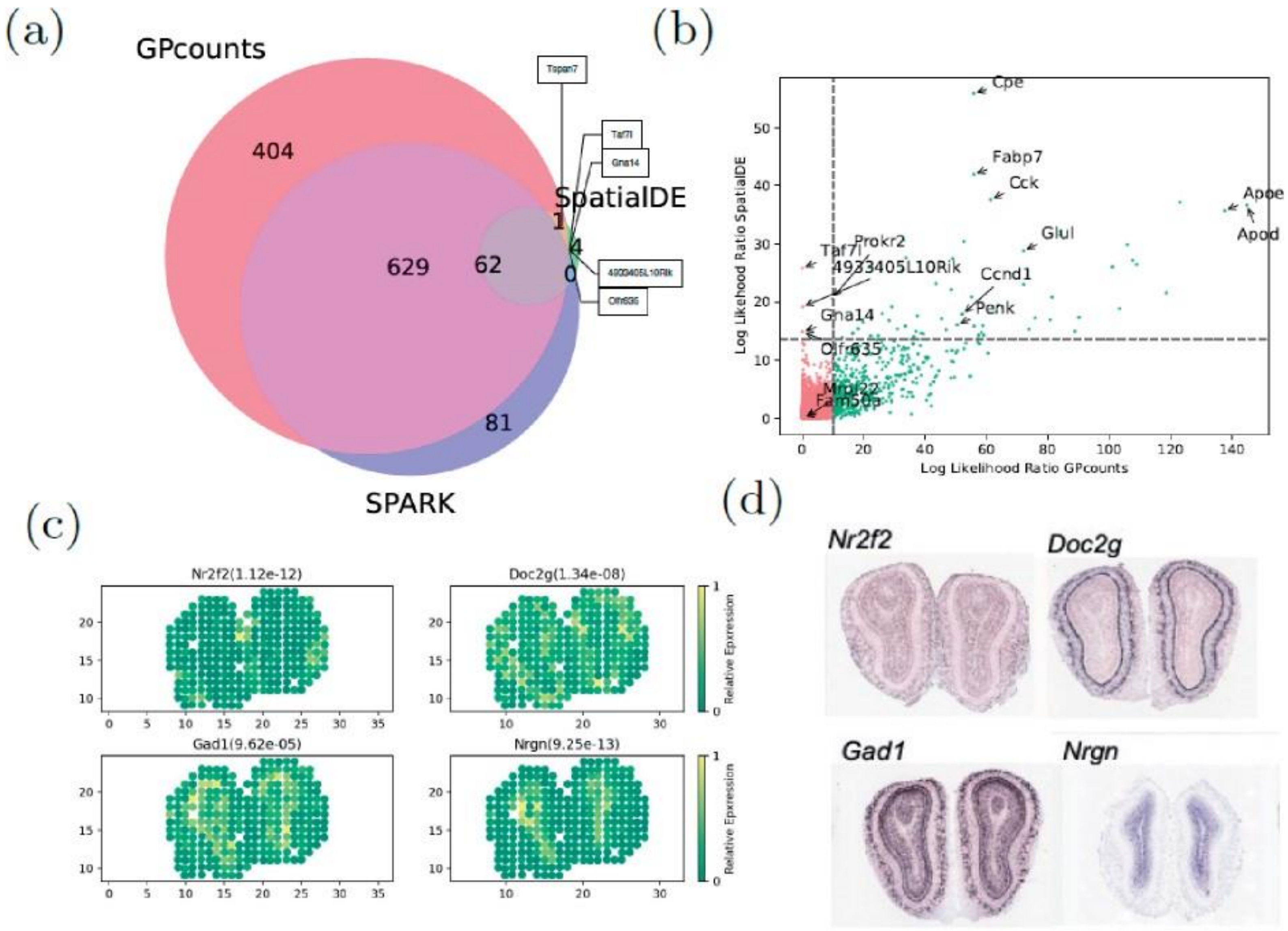
SpatialDE: identification of spatially variable genes

Valentine Svensson, Sarah A Teichmann & Oliver Stegle

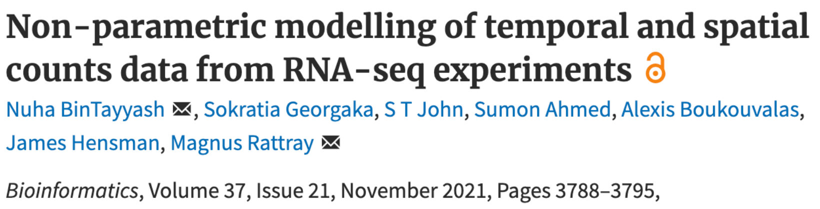
Nature Methods **15**, 343–346(2018) | Cite this article



Differential gene expression – spatial



Using a counts likelihood improves sensitivity to detect DE genes



Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments

Nuha BinTayyash, Sokratia Georgaka, S T John, Sumon Ahmed, Alexis Boukouvalas, James Hensman, Magnus Rattray

Bioinformatics, Volume 37, Issue 21, November 2021, Pages 3788–3795,

Code

https://github.com/ManchesterBioinference/GPcounts

Uses:

GPflow
Sparse variational inference
Non-Gaussian likelihoods (negative binomial)

Also implements branching kernel (discussed later)

Talk outline

Biological applications:

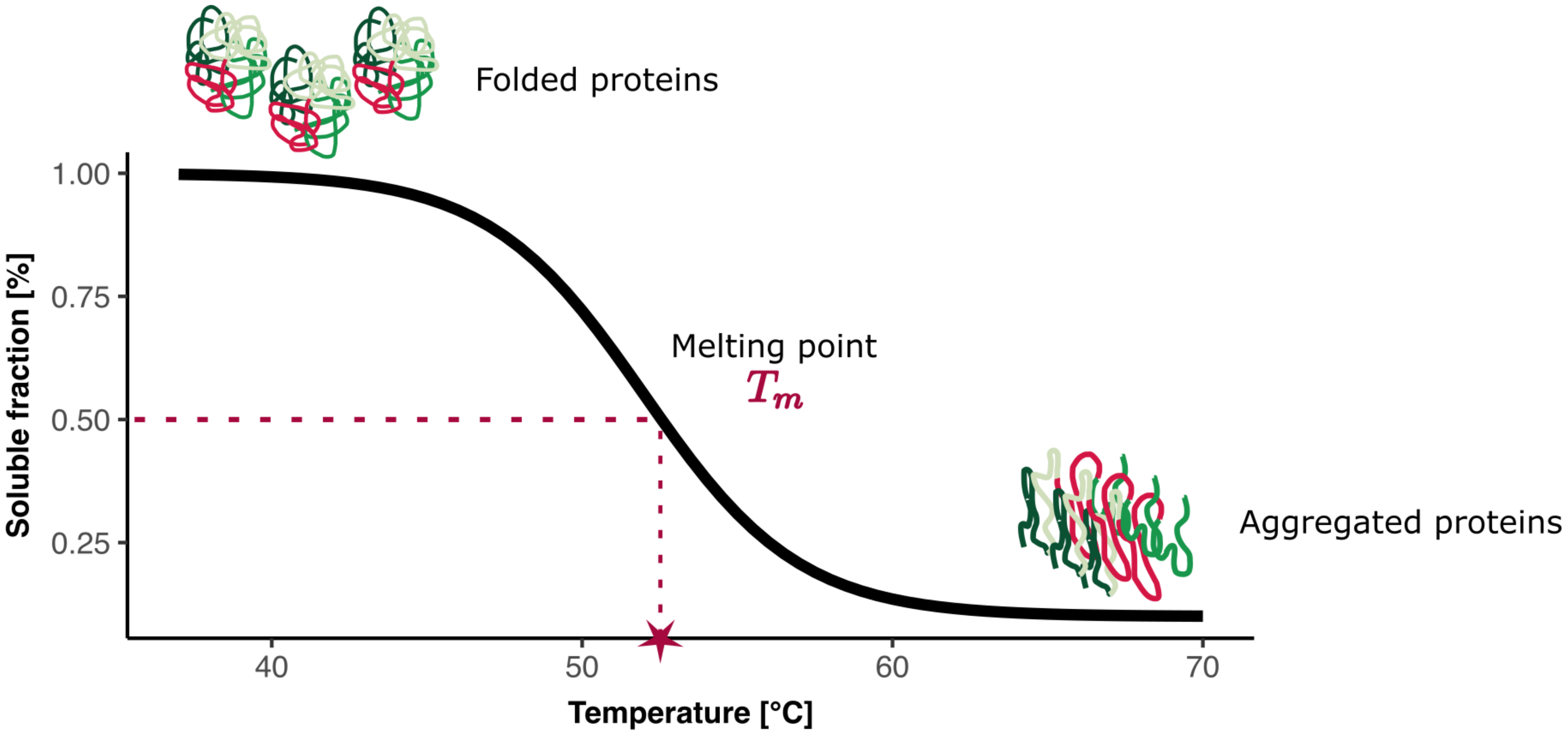
(1) Differential gene expression

(2) Protein melting curves

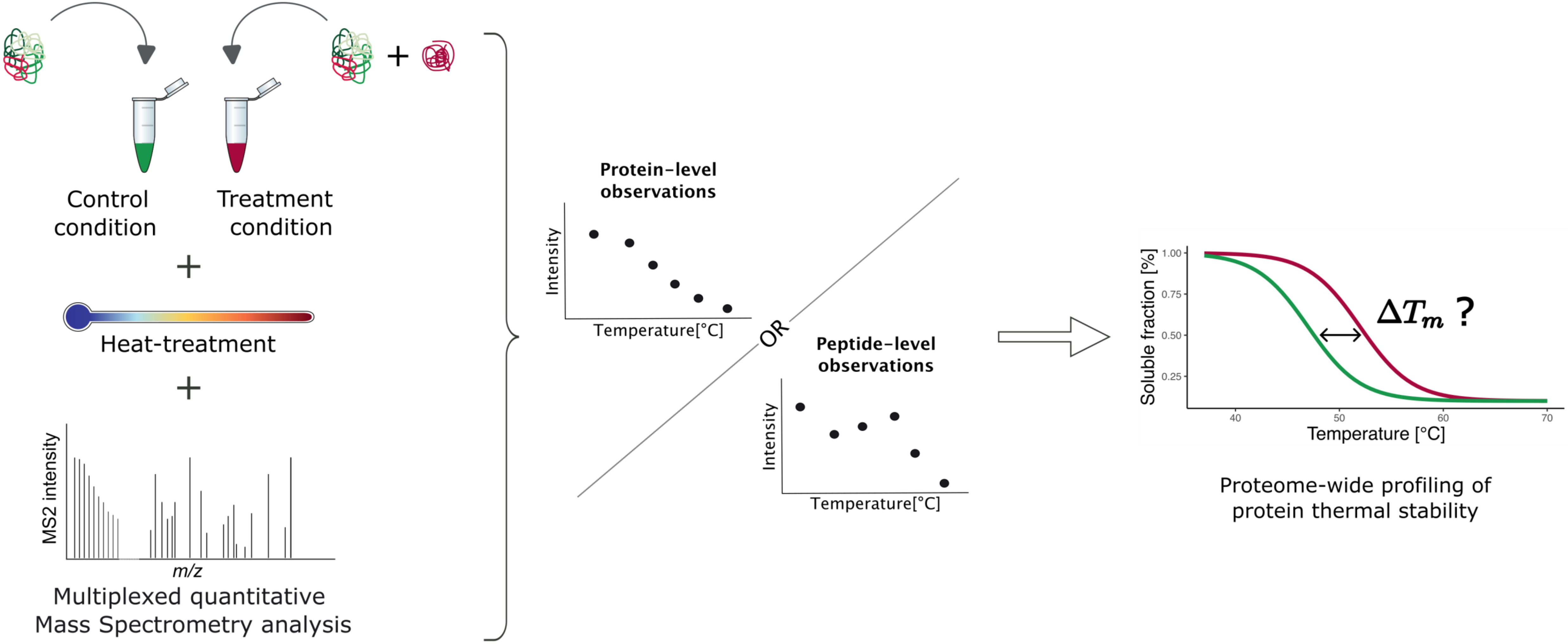
(3) mRNA production and degradation

(4) Single-cell pseudotime and branching

Heat-induced protein denaturation



Thermal Proteome Profiling



Control condition

Treatment condition

+

Heat-treatment

+

Multiplexed quantitative Mass Spectrometry analysis

Protein–level observations

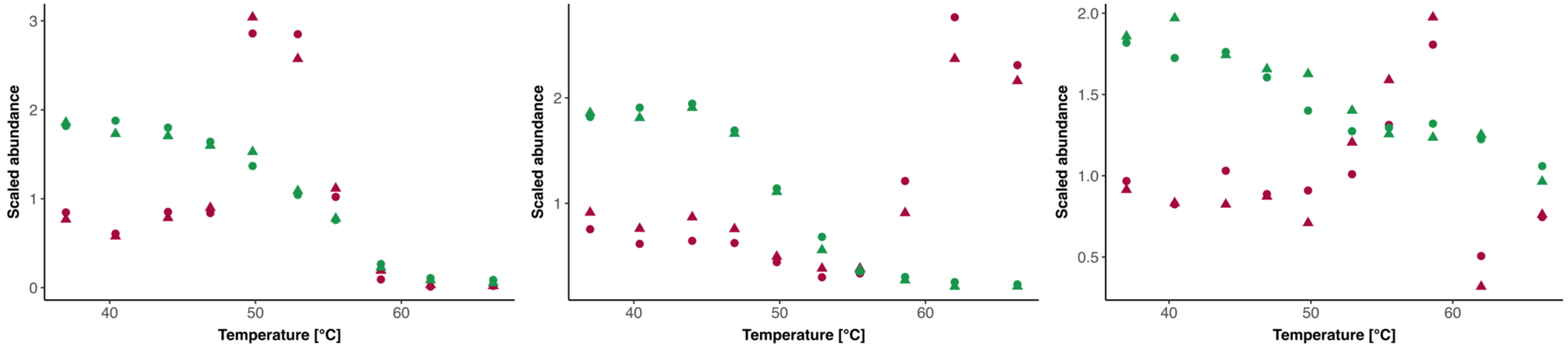
Peptide–level observations

ΔT_m ?

Proteome-wide profiling of protein thermal stability

Dark Meltome

Some melting curves present non-sigmoidal behaviours



Phase transitioning events

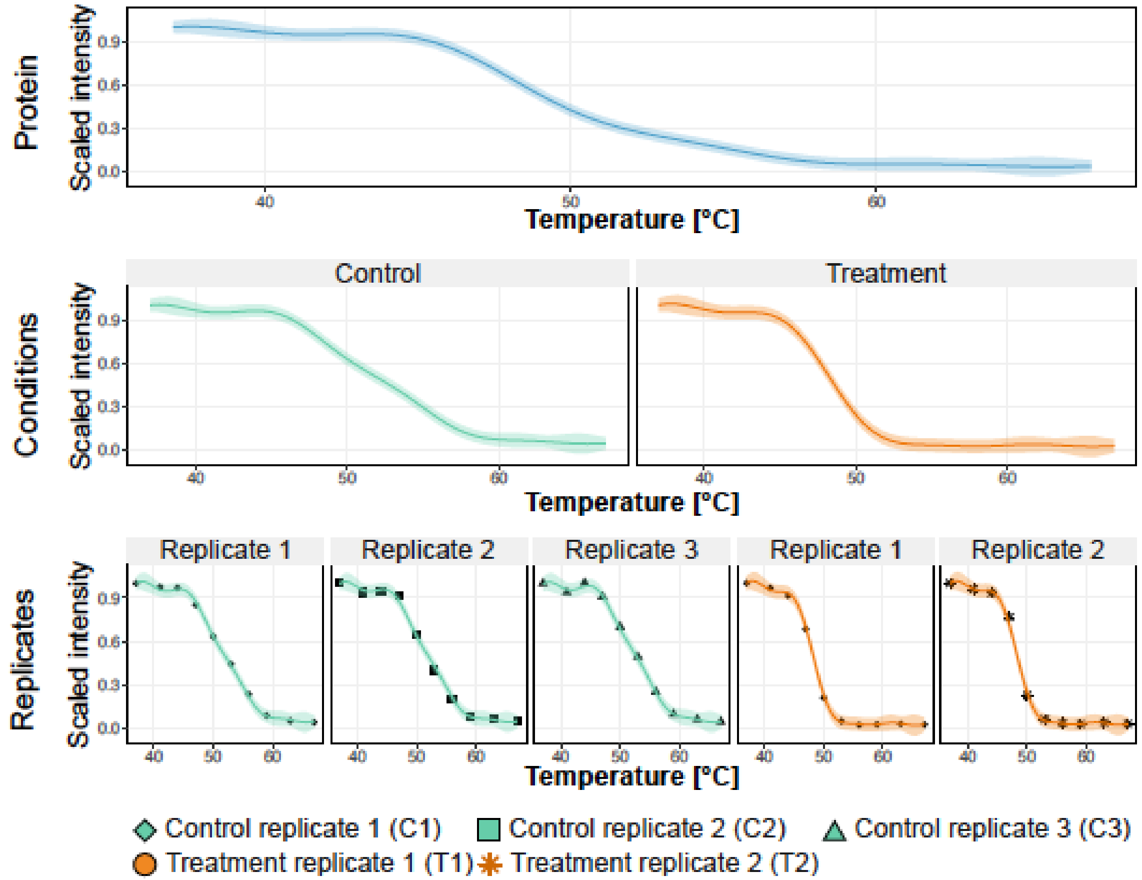
Nucleic acid binding proteins

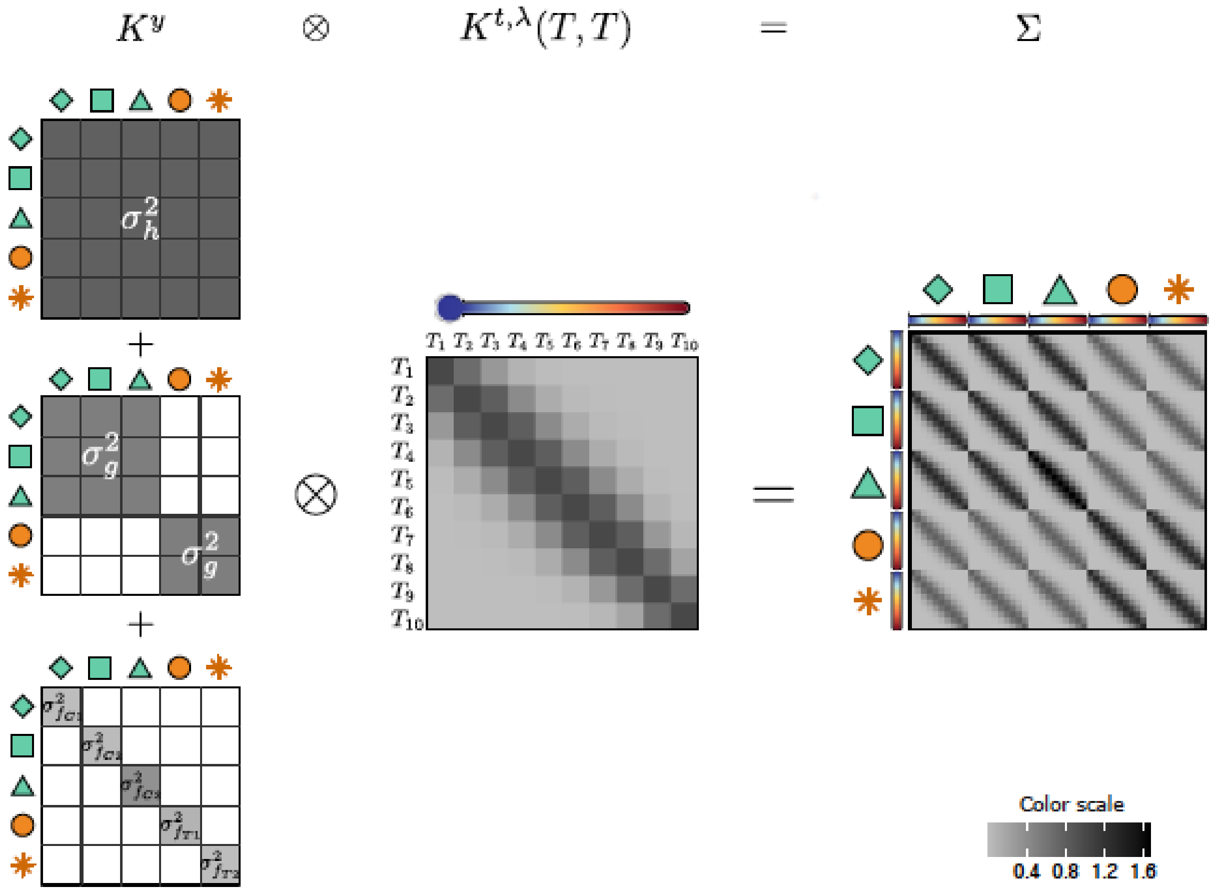
Simultaneous presence of multiple isoforms in the cell

Post-Translational Modifications

Different sub-cellular localisation of different sub-populations

Hierarchical Gaussian processes: GPmelt





Hierarchical Gaussian process models explore the dark meltome of thermal proteome profiling experiments.

Cecile Le Sueur[1], Magnus Rattray[2]*, Mikhail Savitski[1]*

Hierarchical Gaussian process

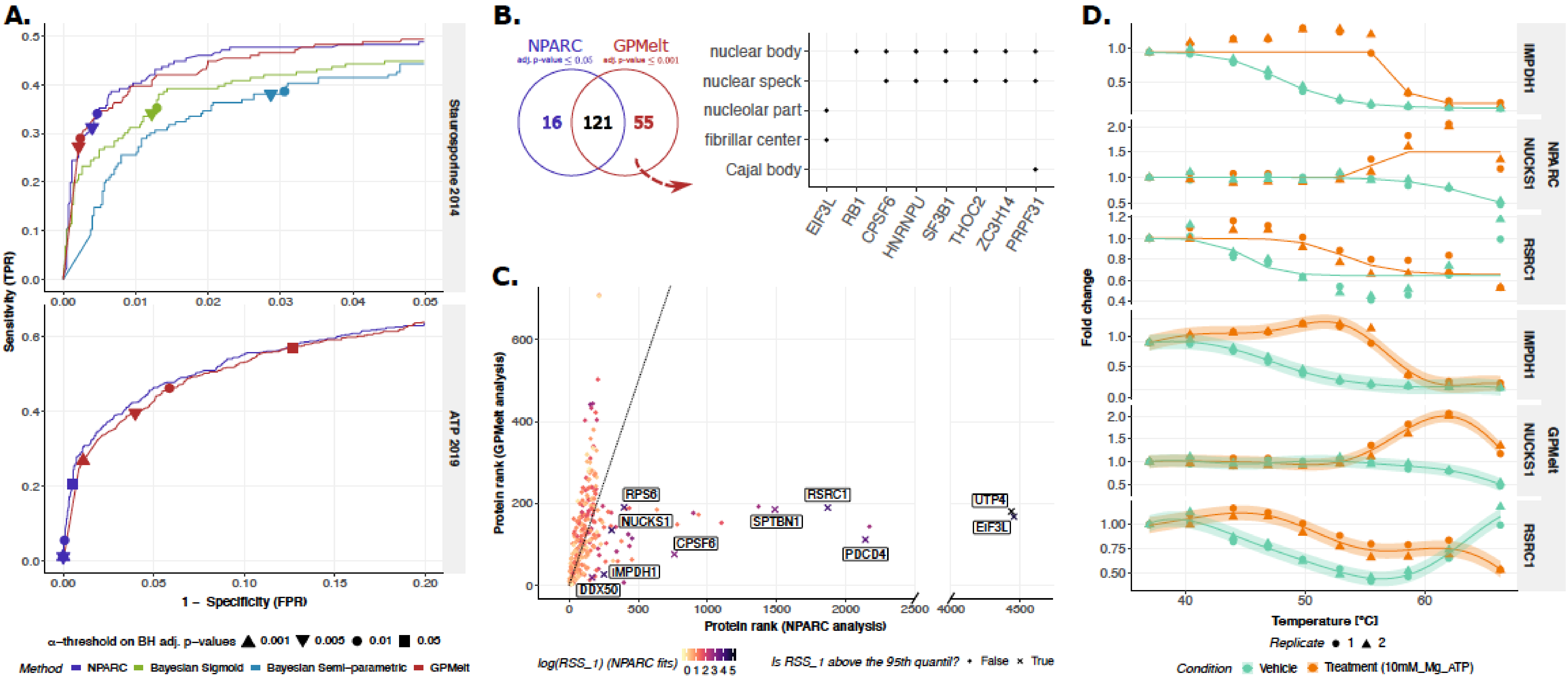
Three-level hierarchical model for protein-level TPP-TR datasets analysis:

$$
\begin{array}{l}
\forall\ i \in [\![1, N]\!] \\
\forall\ r \in [\![1, R]\!] \\
\forall\ c \in [\![1, C]\!]
\end{array}
\left\{
\begin{array}{ll}
h \sim GP(0, k_h(t, \cdot|\lambda_1)) & \textit{protein} \\
g_c \sim GP(h, k_g(t, \cdot|\lambda_1)) & \textit{conditions} \\
f_{cr} \sim GP(g_c, k_{f_{cr}}(t, \cdot|\lambda_2)) & \textit{replicates} \\
y_{cri} = f_{cr}(t_i) + \epsilon_i & \textit{observations} \\
\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \beta^2) &
\end{array}
\right.
$$

null hypothesis $g_{c_1} = g_{c_2} \equiv g_{c_0}$

$$LR = -2 \cdot \log\left(\frac{p_{\text{null}}(Y_p|T_p, \theta_p)}{p_{\text{alt}}(Y_p|T_p, \theta_p)}\right)$$

Hierarchical Gaussian process



Code

https://embl-community.io/grp-savitski/gpmelt

Uses:

GPyTorch
Hadamard multi-task GP regression
Nextflow for whole pipeline

Talk outline

Biological applications:

(1) Differential gene expression

(2) Protein melting curves

(3) mRNA production and degradation

(4) Single-cell pseudotime and branching

Embryonic development: transition from maternal to zygotic expression

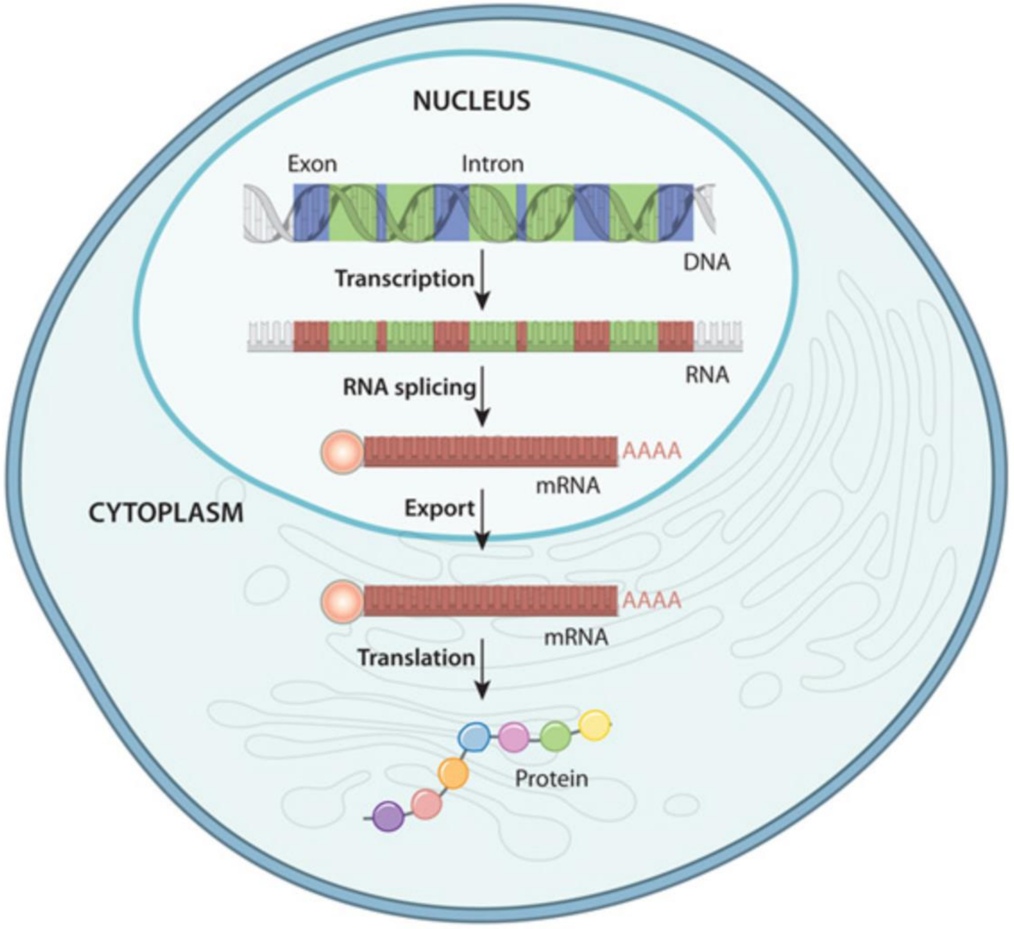


Figure 1: An overview of the flow of information from DNA to protein in a eukaryote
First, both coding and noncoding regions of DNA are transcribed into mRNA. Some regions are removed (introns) during initial mRNA processing. The remaining exons are then spliced together, and the spliced mRNA molecule (red) is prepared for export out of the nucleus through addition of an endcap (sphere) and a polyA tail. Once in the cytoplasm, the mRNA can be used to construct a protein.

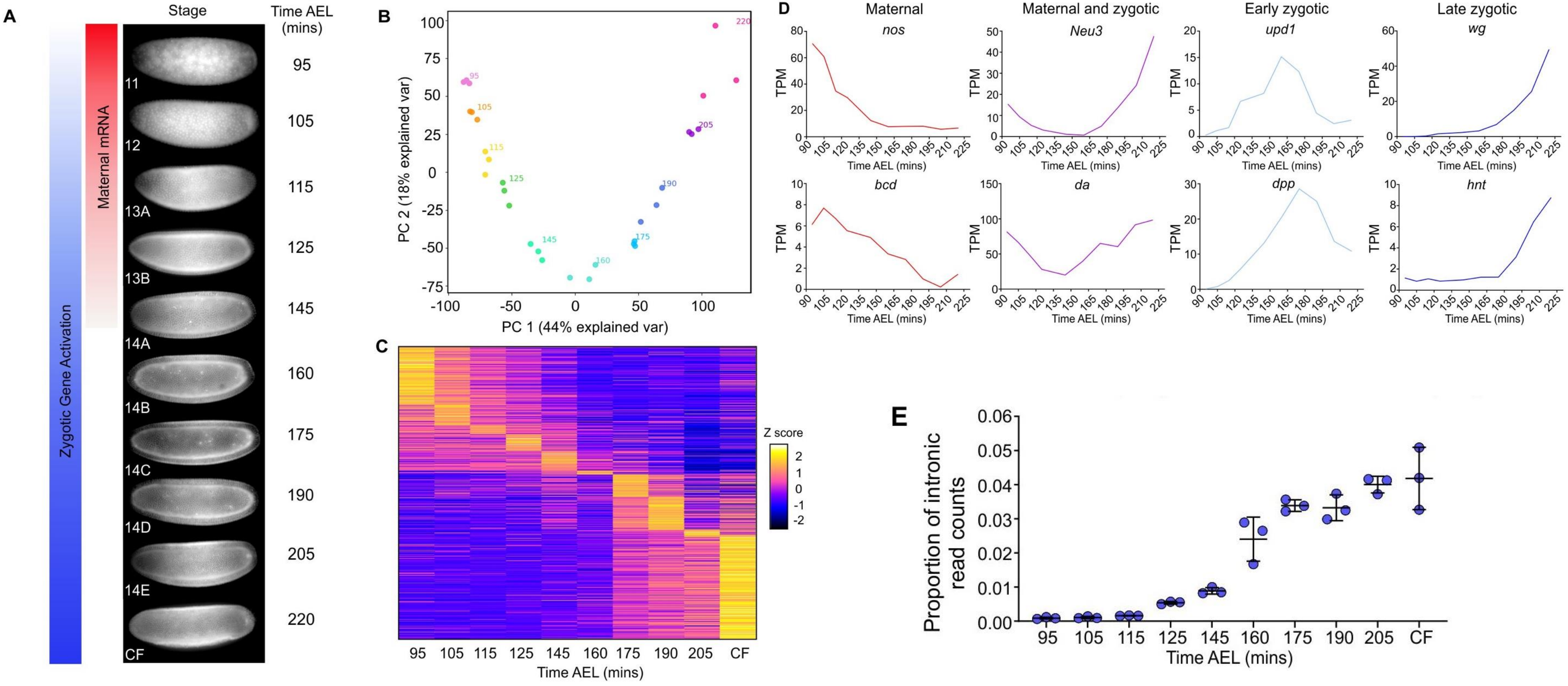
Newly transcribed “pre-mRNA” contains both **introns** and **exons**

The introns are **spliced out** to make mature mRNA containing only exons

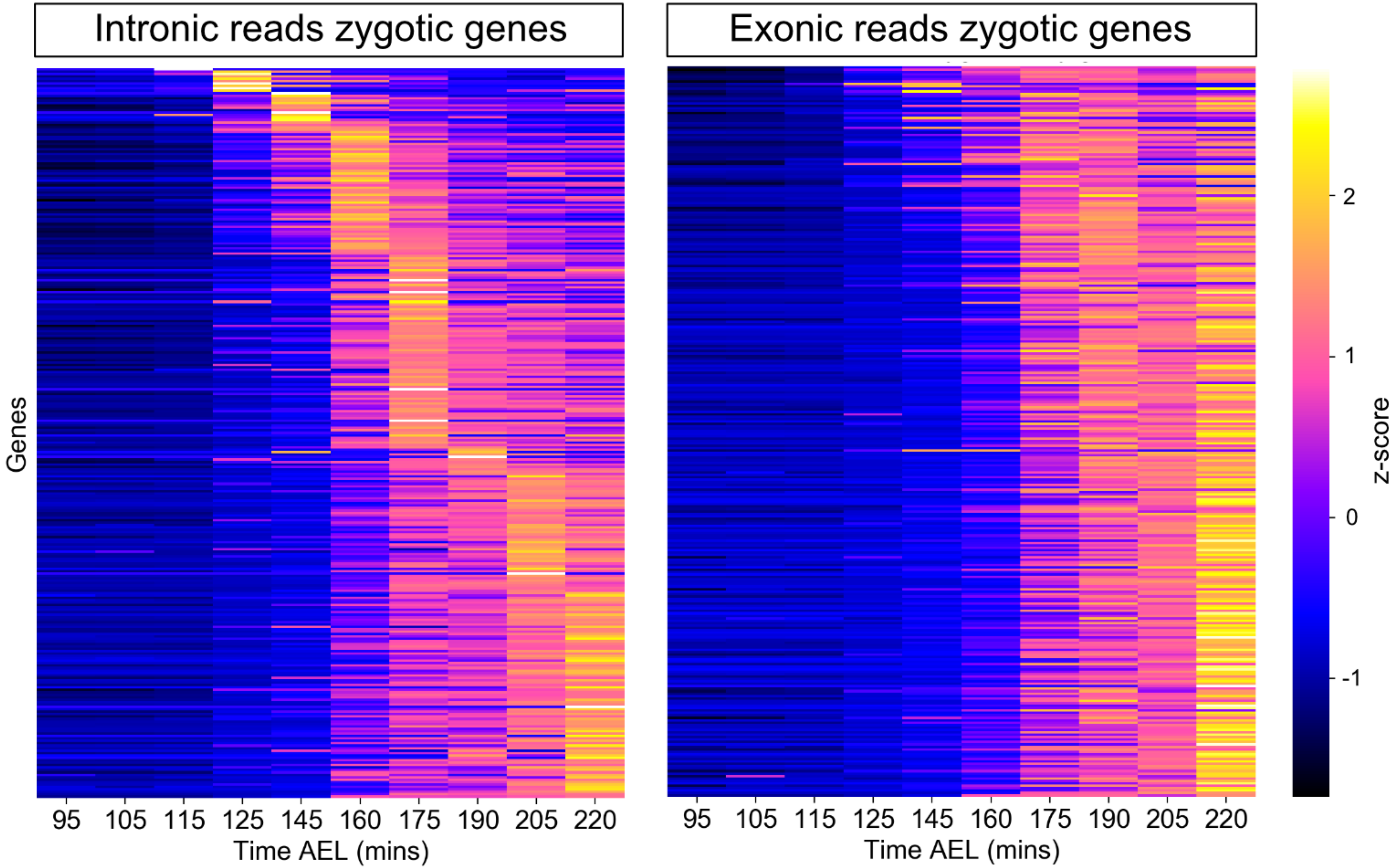
Embryos inherit some mature mRNA from their mothers (**maternal RNA**)

mRNA produced by the embryo is called **zygotic RNA**

Embryonic development: transition from maternal to zygotic expression



pre-mRNA expression precedes mature RNA production



Modelling mRNA production & degradation

pre-mRNA (introns) $\frac{\mathrm{d}p}{\mathrm{d}t} = T(t) - Sp(t)$

mRNA (exons) $\frac{\mathrm{d}m}{\mathrm{d}t} = T(t) - Dm(t)$

$T(t)$ transcriptional rate

S splicing rate

D mRNA degradation rate

Drosophila splicing half-lives are short (median 2 min) so we make large S approximation

$$p(t) = \frac{T(t)}{S} \quad \text{as} \quad S \to \infty$$

$$\frac{\mathrm{d}m}{\mathrm{d}t} = Sp(t) - Dm(t)$$

Modelling mRNA production & degradation

$$\frac{\mathrm{d}m}{\mathrm{d}t} = Sp(t) - Dm(t)$$

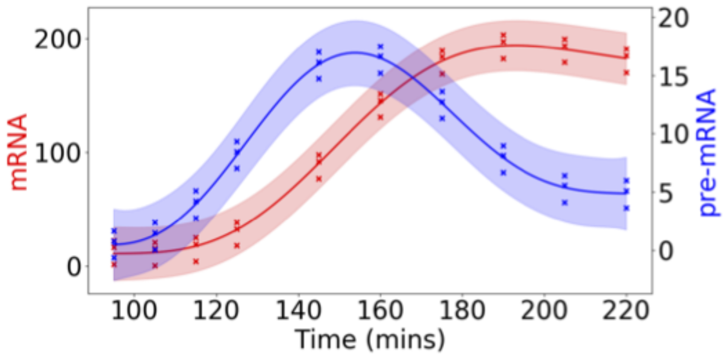
$m(t)$ mRNA (exonic reads)

$p(t)$ pre-mRNA (intronic reads)

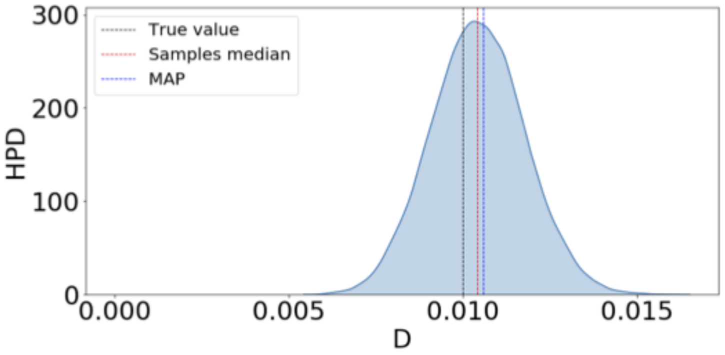
S splicing rate

D mRNA degradation rate

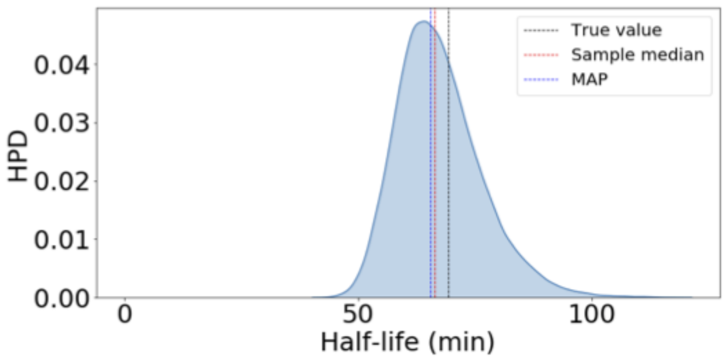
How can we model pre-mRNA dynamics $p(t)$ and infer parameters?



Simulated data D=0.01
half-life=69.3 min



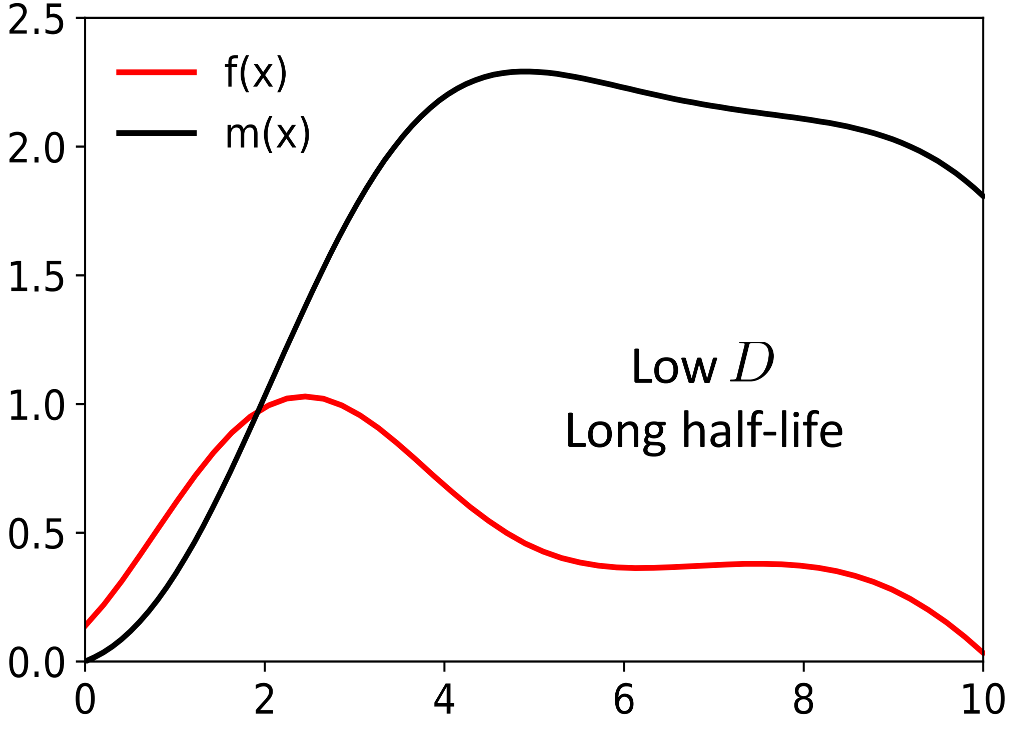
Posterior distribution of D

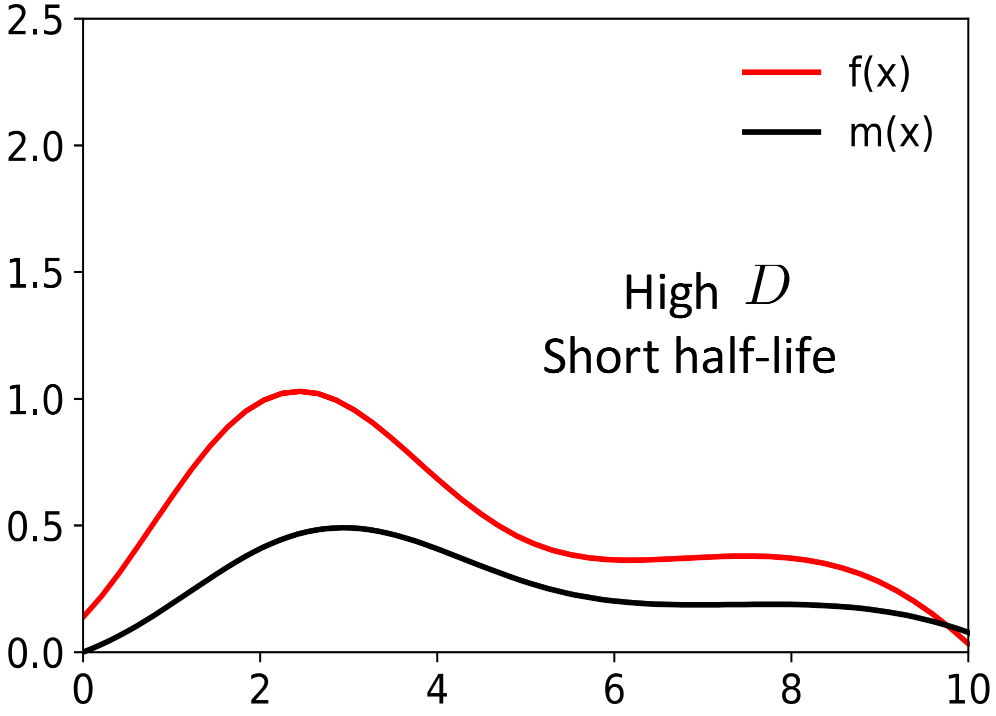


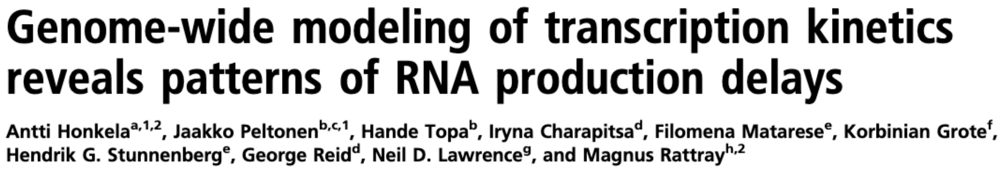
Posterior distribution of half-life

Modelling mRNA production & degradation

$$f(t) \sim GP(0,k) \qquad \frac{\mathrm{d}m}{\mathrm{d}t} = Sf(t) - Dm(t) \longrightarrow [f,m] \sim GP(0,k_{\mathrm{LFM}})$$



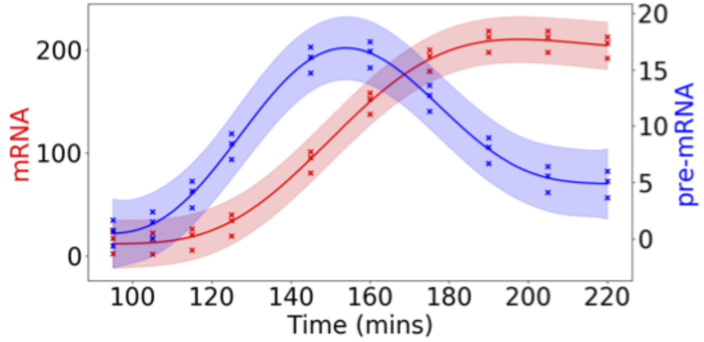




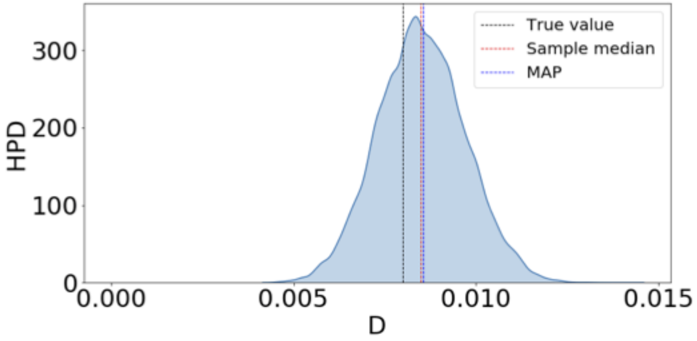
Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays

Antti Honkela[a,1,2], Jaakko Peltonen[b,c,1], Hande Topa[b], Iryna Charapitsa[d], Filomena Matarese[e], Korbinian Grote[f], Hendrik G. Stunnenberg[e], George Reid[d], Neil D. Lawrence[g], and Magnus Rattray[h,2]

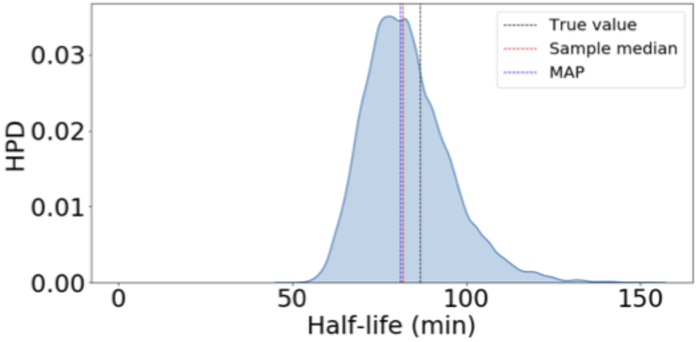
Gaussian process estimation of half-lives



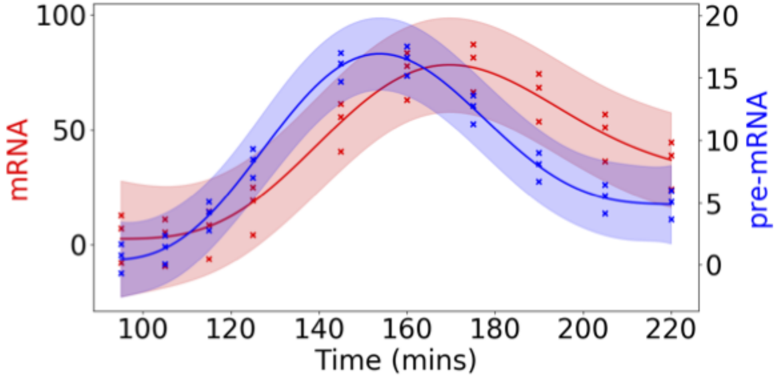
Simulated data $D=0.008$, half-life=86.6 min



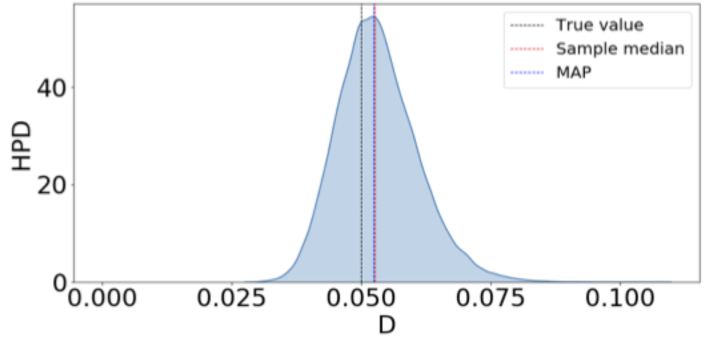
Posterior distribution of D



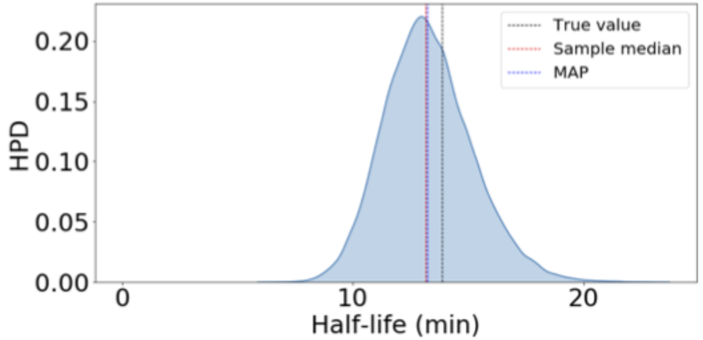
Posterior distribution of half-life



Simulated data $D=0.05$, half-life=13.8 min

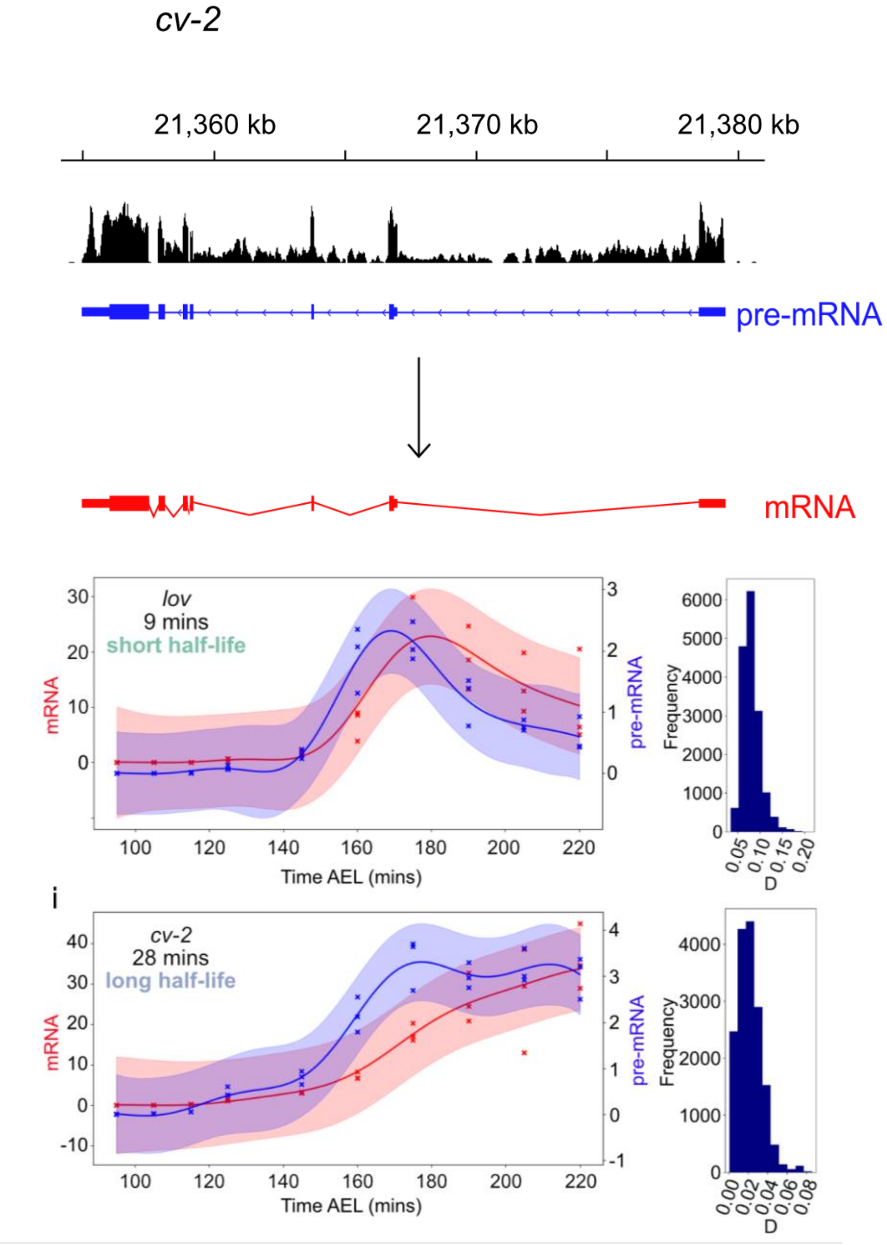


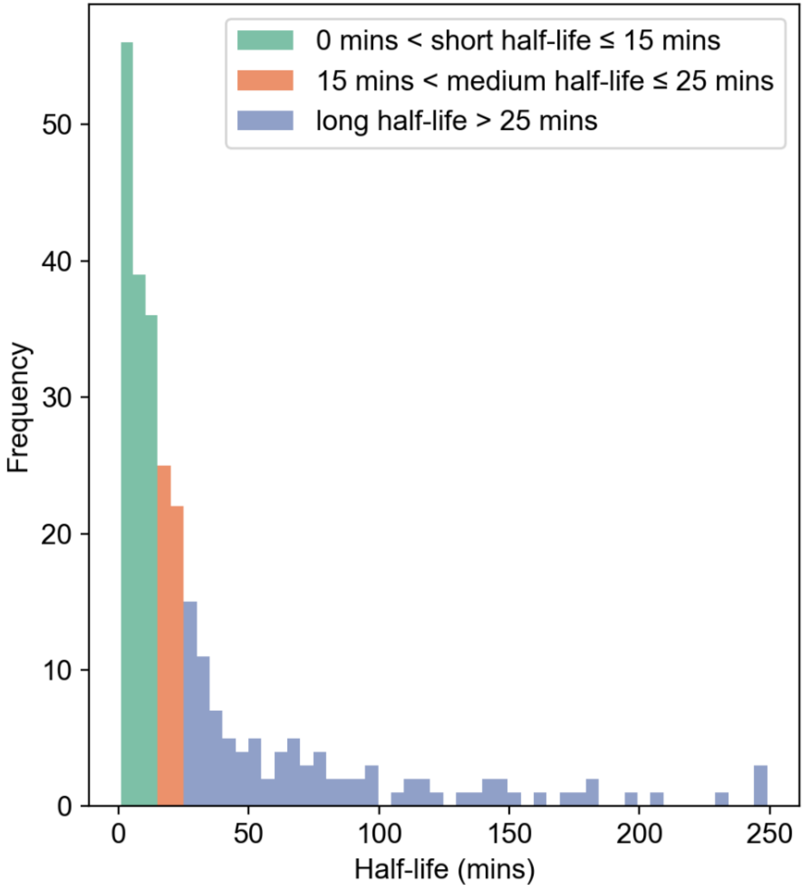
Posterior distribution of D



Posterior distribution of half-life

Zygotic transcripts exhibit a broad range of half-lives

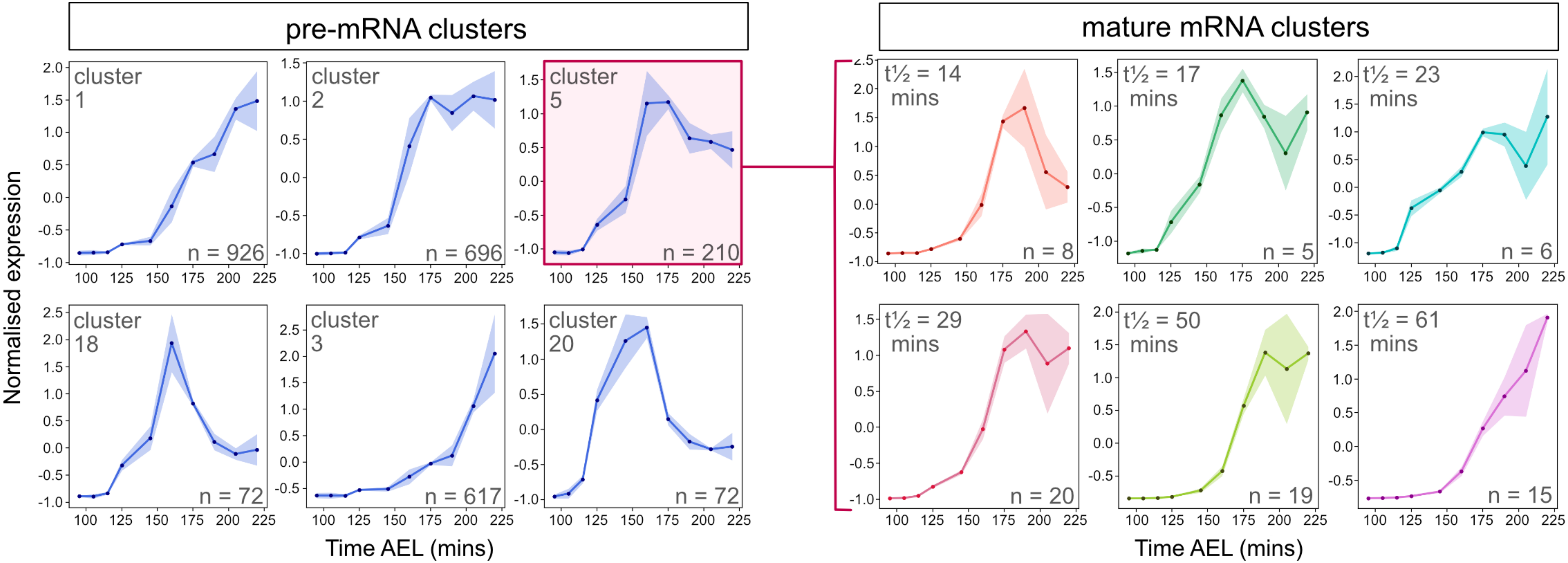




Short half-life: cell-adhesion proteins, transcription factors

Long half-life: signalling receptor binding

mRNA degradation shapes gene expression dynamics



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 37, NO. 2, FEBRUARY 2015

Fast Nonparametric Clustering of Structured Time-Series

James Hensman, Magnus Rattray, and Neil D. Lawrence

Code

https://github.com/ManchesterBioinference/GP_Transcription_Dynamics

Uses:

GPFlow (to implement latent force covariance)
Tensorflow probability (for MCMC over hyper-parameters)

Talk outline

Biological applications:

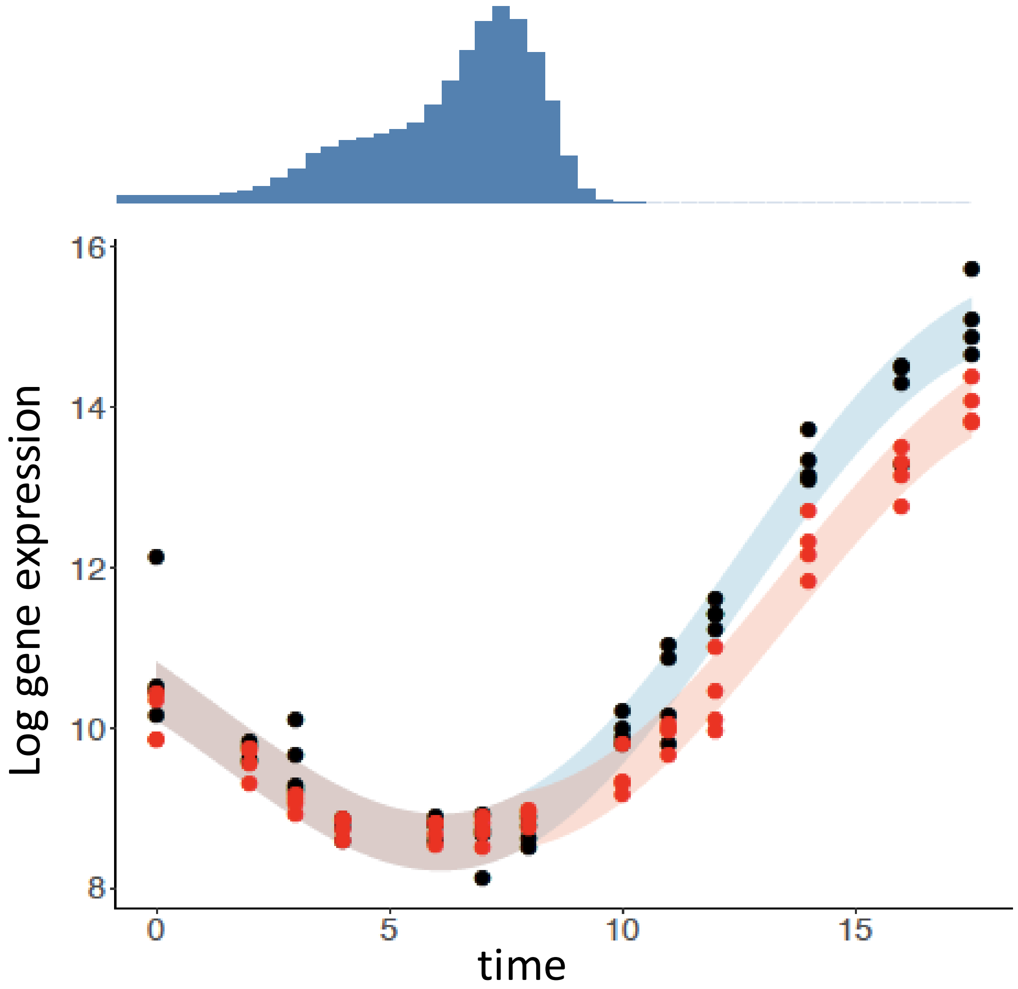
(1) Differential gene expression

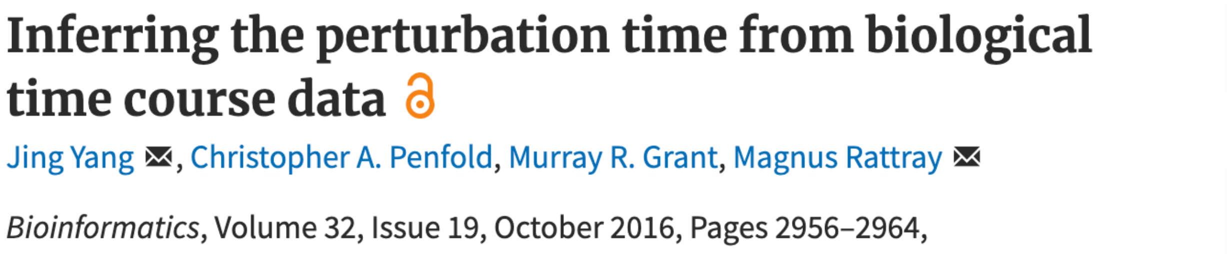
(2) Protein melting curves

(3) mRNA production and degradation

(4) Single-cell pseudotime and branching

Gaussian process model of branching dynamics



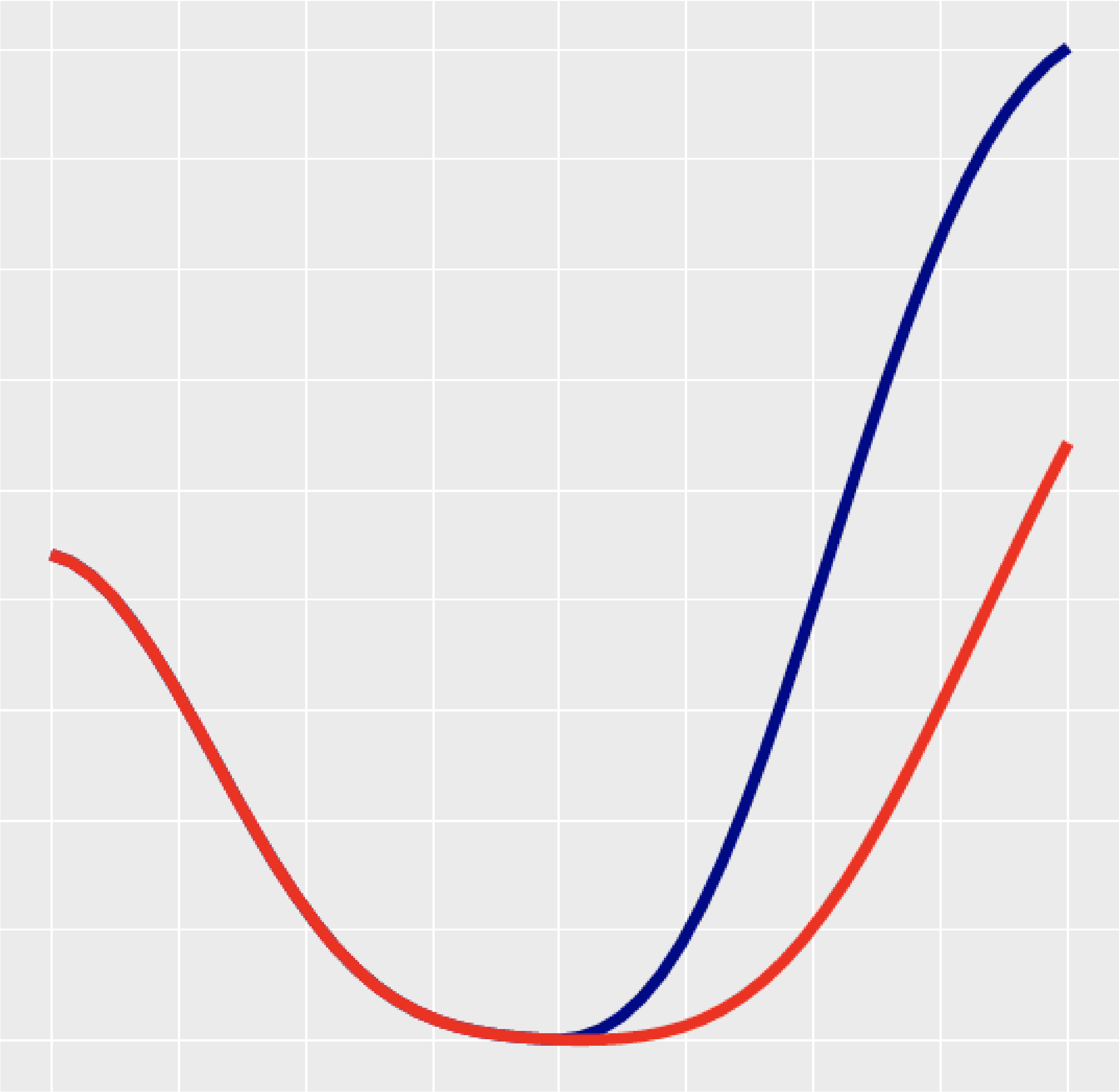


Inferring the perturbation time from biological time course data

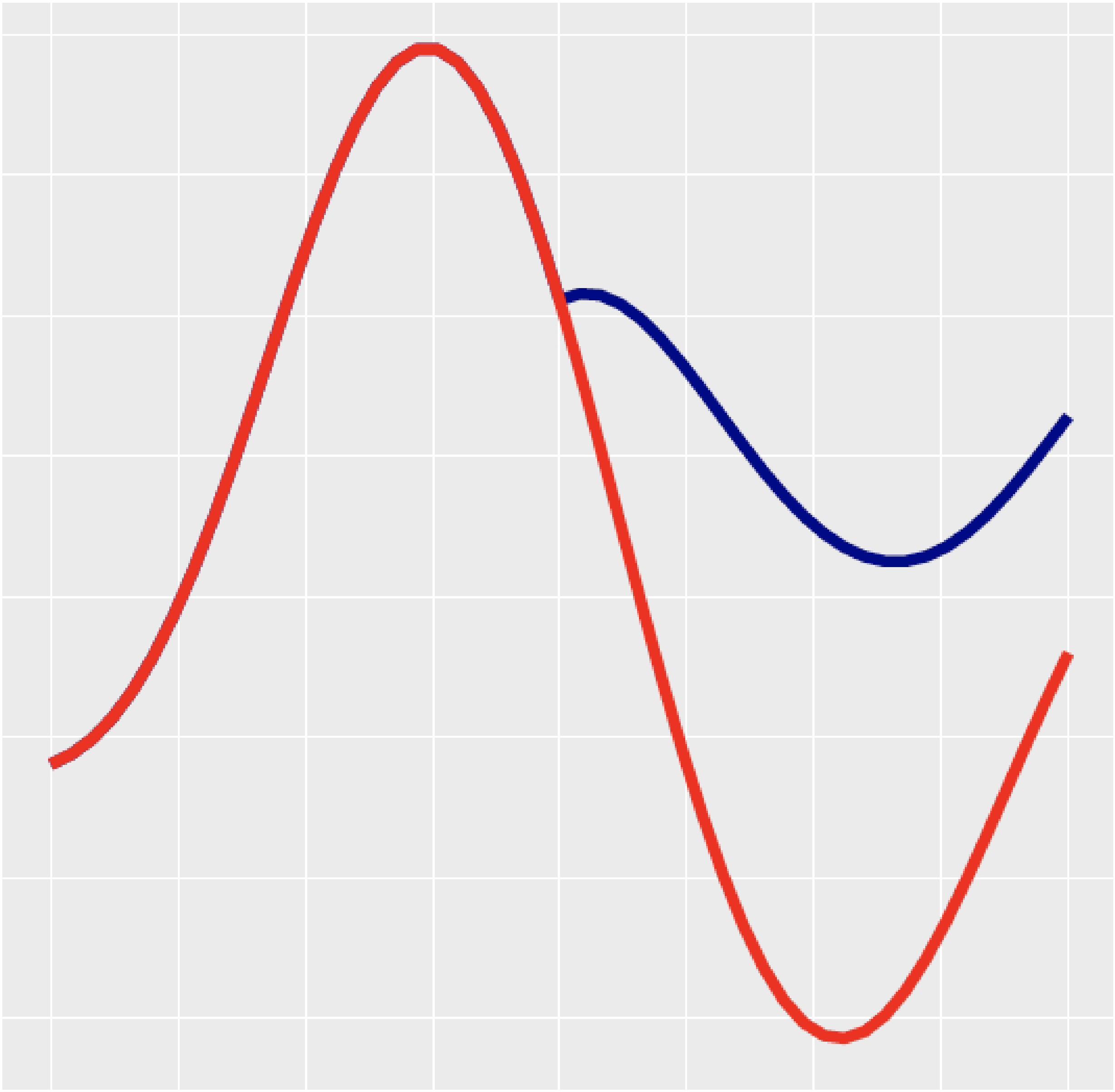
Jing Yang ✉, Christopher A. Penfold, Murray R. Grant, Magnus Rattray ✉

Bioinformatics, Volume 32, Issue 19, October 2016, Pages 2956–2964,

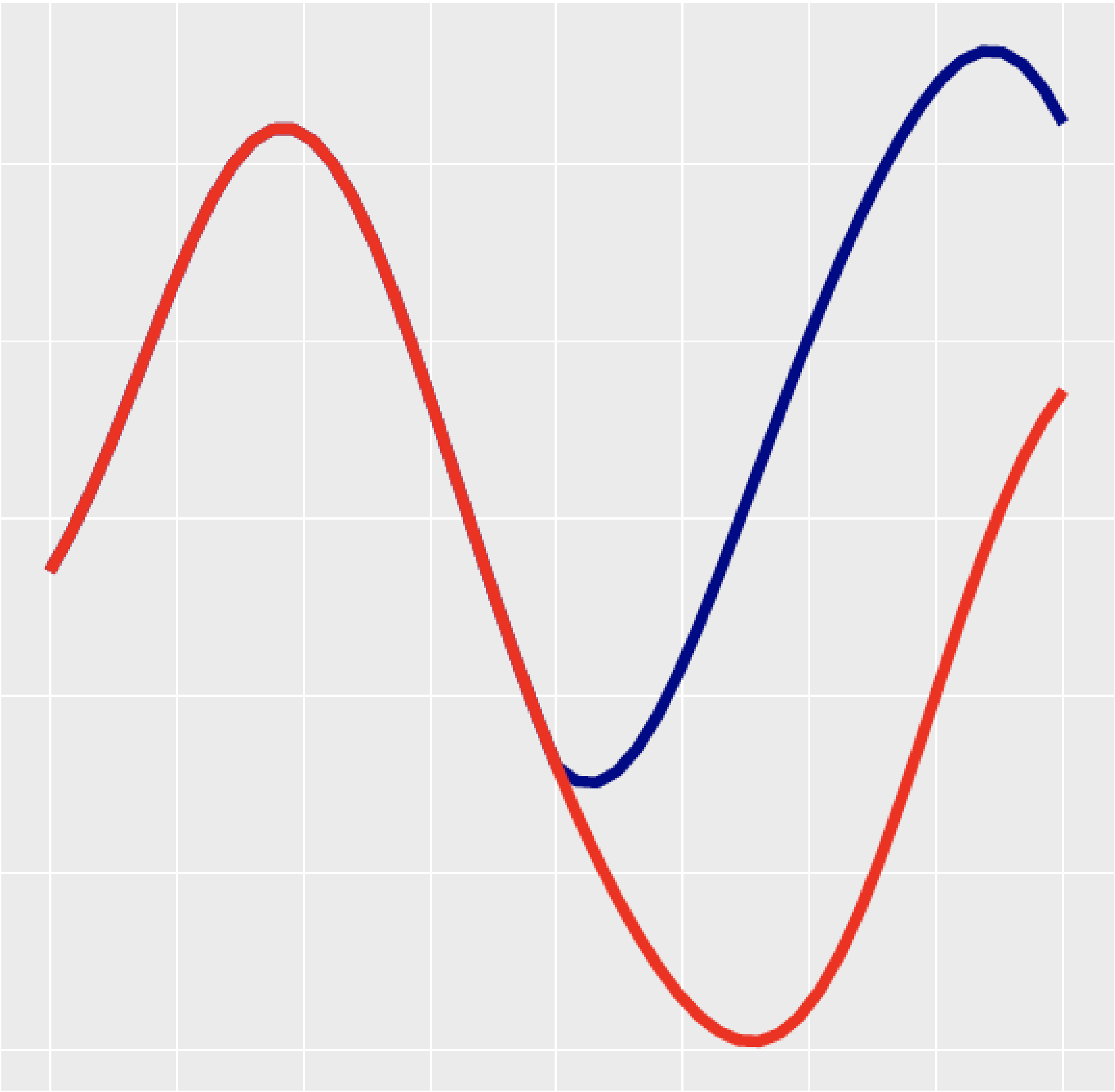
Gaussian process model of branching dynamics



Gaussian process model of branching dynamics

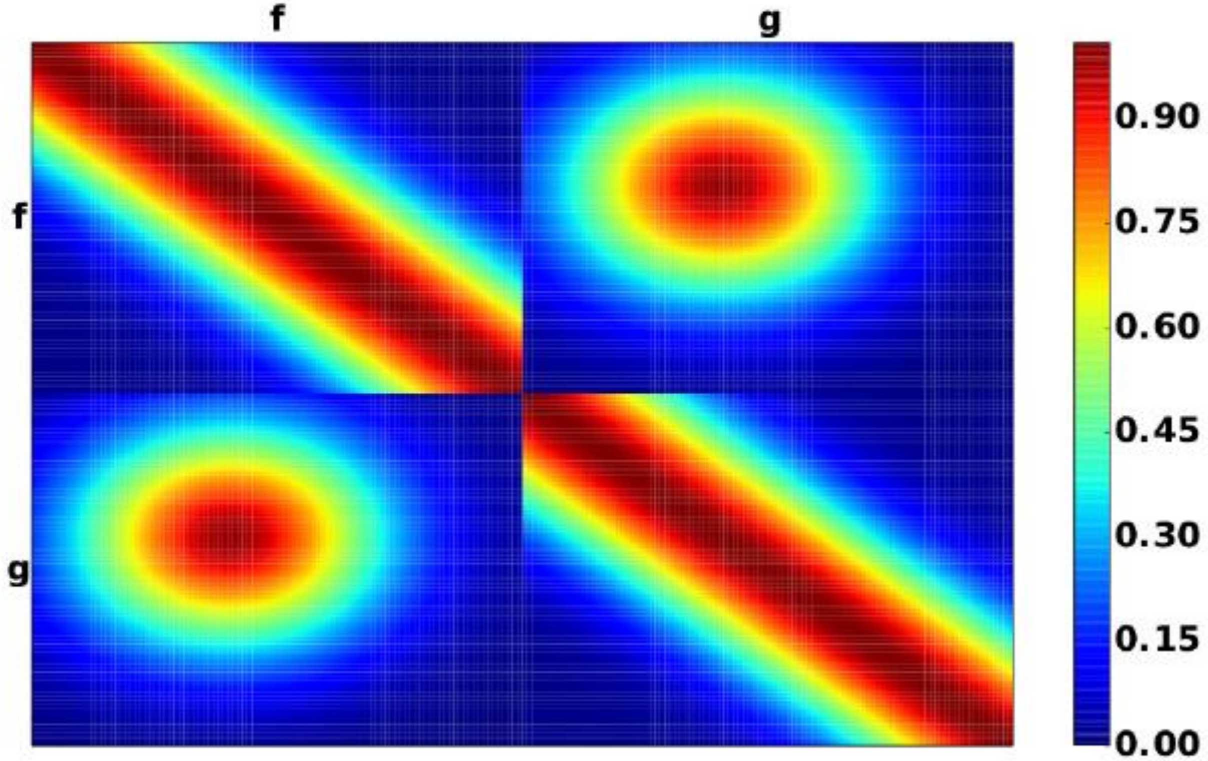


Gaussian process model of branching dynamics



Joint distribution to two functions crossing at t_p

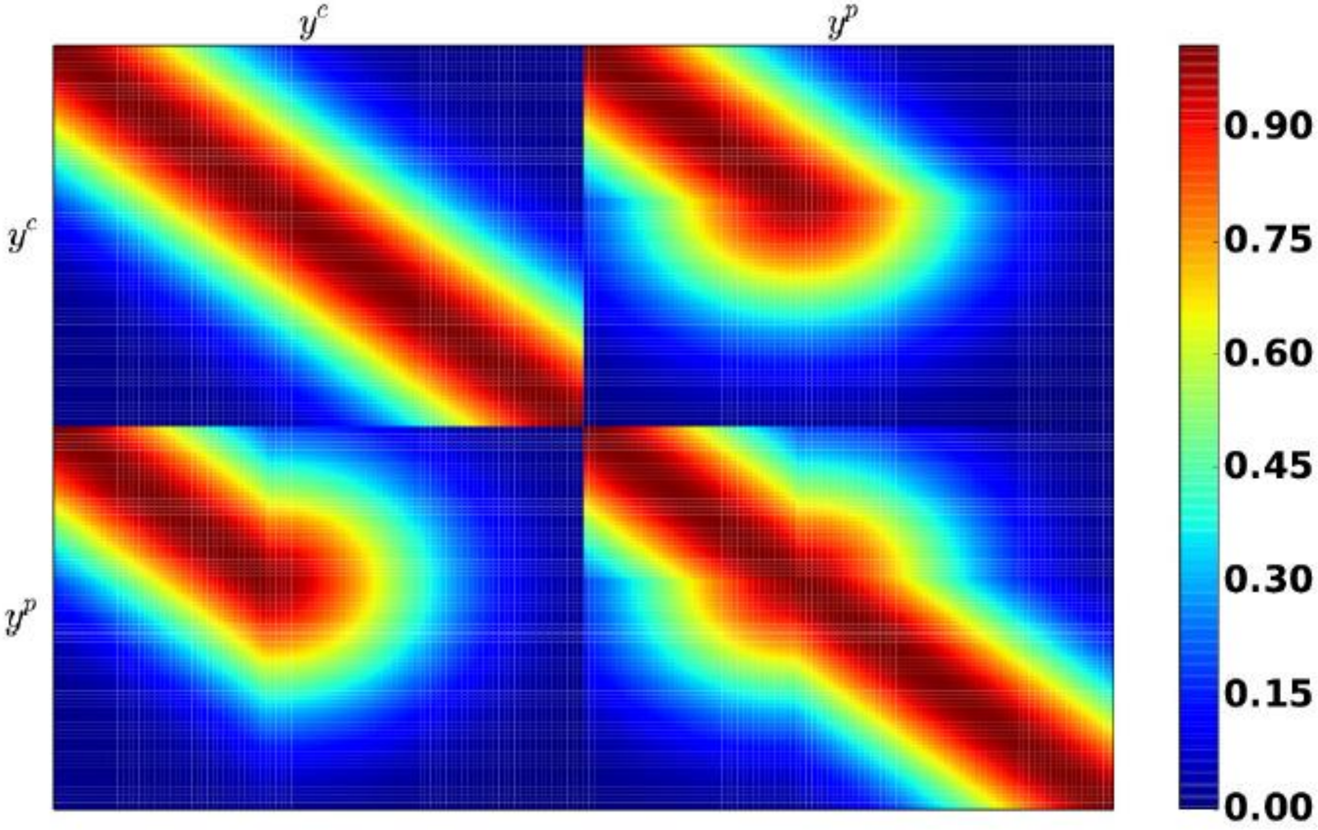
$$f \sim \mathcal{GP}(0, K)\ , \quad g \sim \mathcal{GP}(0, K)\ , \quad g(t_p) = f(t_p)$$



$$\Sigma = \begin{pmatrix} K_{ff} & K_{fg} \\ K_{gf} & K_{gg} \end{pmatrix} = \begin{pmatrix} K(\mathbf{T},\mathbf{T}) & \frac{K(\mathbf{T},t_p)K(t_p,\mathbf{T})}{k(t_p,t_p)} \\ \frac{K(\mathbf{T},t_p)K(t_p,\mathbf{T})}{k(t_p,t_p)} & K(\mathbf{T},\mathbf{T}) \end{pmatrix}$$

Joint distribution of two datasets diverging at t_p

$$
\begin{aligned}
y^c(t_n) &\sim \mathcal{N}(f(t_n), \sigma^2) \\
y^p(t_n) &\sim \mathcal{N}(f(t_n), \sigma^2) \quad \text{for } t_n \leq t_p \\
y^p(t_n) &\sim \mathcal{N}(g(t_n), \sigma^2) \quad \text{for } t_n > t_p
\end{aligned}
$$



Inference tasks

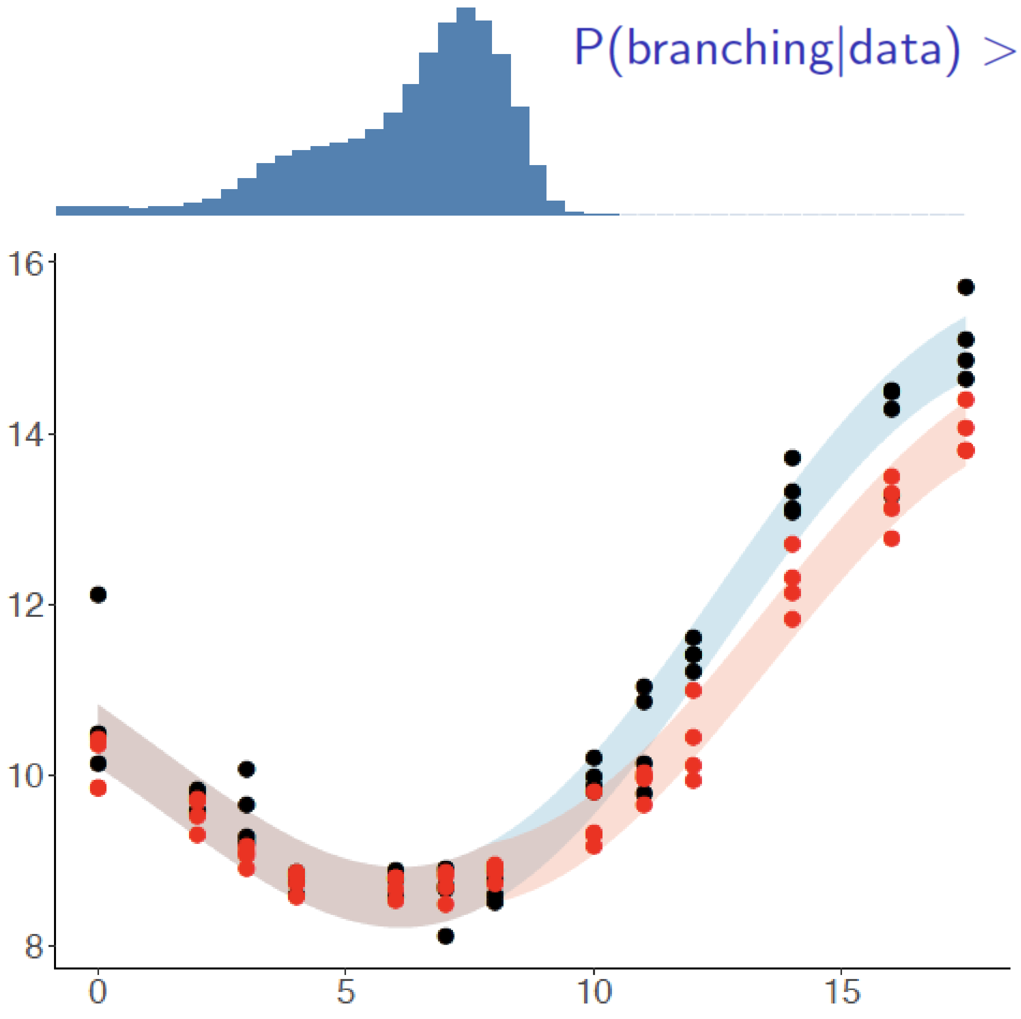
(1) Posterior probability of the branching time t_b:

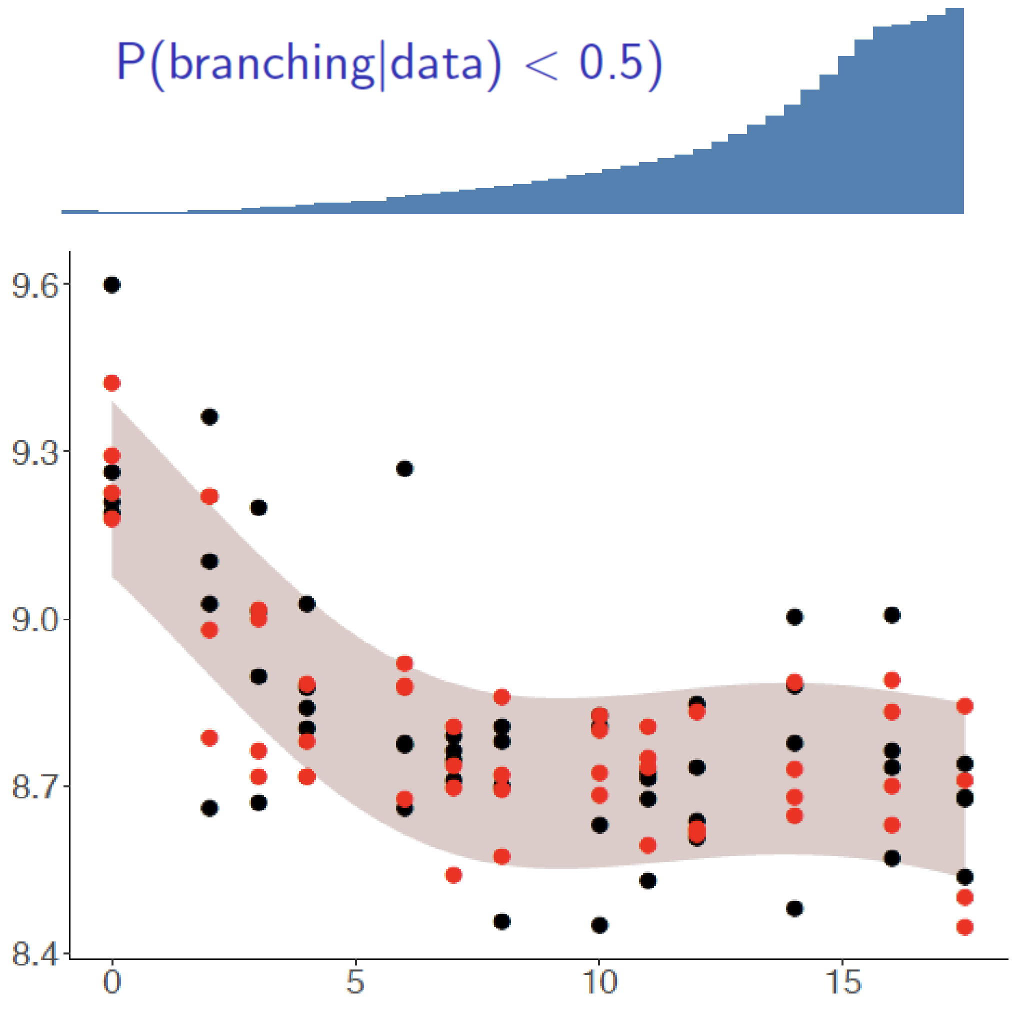
$$p(t_b|Y^c, Y^p) \simeq \frac{p(Y^c, Y^p|t_b)}{\sum_{t=t_{\min}}^{t=t_{\max}} p(Y^c, Y^p)|t)}$$

(2) Bayes factor for branching versus not branching

$$\frac{p(0 < t_b < t_{\max}|Y^c, Y^p)}{p(t_p = t_{\max}|Y^c, Y^p)} = \frac{\frac{1}{N_b}\sum_{t_b=t_{\min}}^{t_b=t_{\max}} p(Y^c, Y^p|t_b)}{p(Y^c, Y^p|t_{\max})}$$

Application to gene expression time-series data





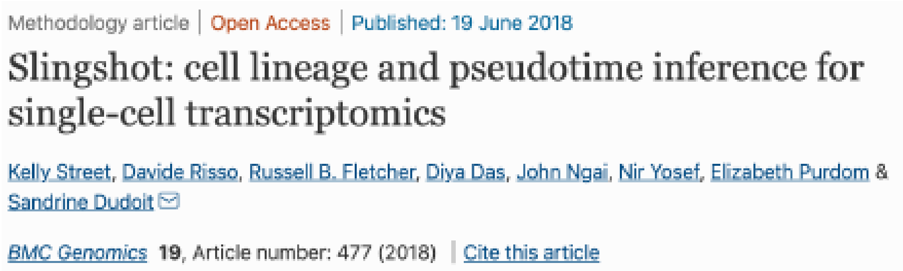
Modelling branching in single-cell gene expression data

Single-cell experiments destructive – can't follow cell through time

We can *infer* time in some dynamic process in the cell

Like re-discovering time from high-dimensional time-series data

Identifies a *pseudotemporal* ordering of cells

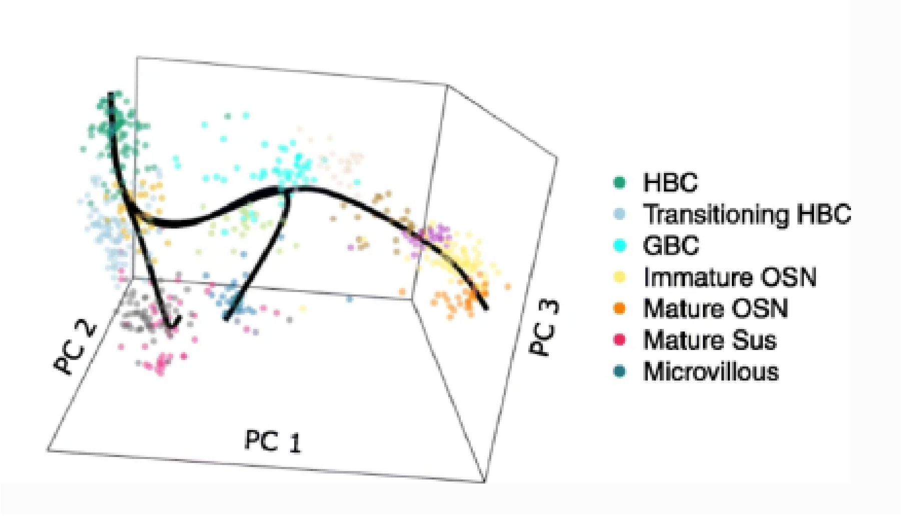


Methodology article | Open Access | Published: 19 June 2018

Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics

Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom & Sandrine Dudoit

BMC Genomics 19, Article number: 477 (2018) | Cite this article



Gaussian process models can be used for pseudotime inference

DeLorean package (Reid & Wernich 2016) uses Bayesian GPLVM for pseudotime inference with capture times τ_c

$$y_g(t) \sim \mathcal{GP}(0, k_t)\,\forall g \qquad t \sim \mathcal{N}(\tau_c, \sigma^2)$$

for gene g and inferred pseudotime t.

Re-implemented using GPflow and sparse variational inference:

$$y_g(t, x) \sim \mathcal{GP}(0, k_{xt})\,\forall g \qquad t \sim \mathcal{N}(\tau_c, \sigma^2)$$

allowing for other sources of variation $x \sim \mathcal{N}(0, \sigma_x^2)$, e.g. branching

GrandPrix: scaling up the Bayesian GPLVM for single-cell data

Sumon Ahmed ✉, Magnus Rattray, Alexis Boukouvalas

Bioinformatics, Volume 35, Issue 1, January 2019, Pages 47–54,

Modelling branching in single-cell snapshot data

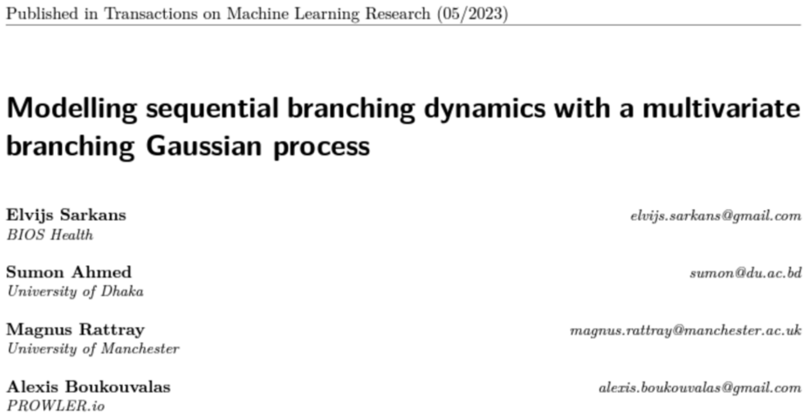
$F = \{f_1, f_2, f_3\}$ is a branching Gaussian Process

$Z \in \{0, 1\}^{N \times 3}$ indicates which branch each cell comes from

$$p(Y|F, Z) = \mathcal{N}(Y|ZF, \sigma^2 I)$$

The likelihood conditional on the branching process is,

$$p(Y|F) = \sum_Z p(Y|F, Z)\, p(Z)$$



Published in Transactions on Machine Learning Research (05/2023)

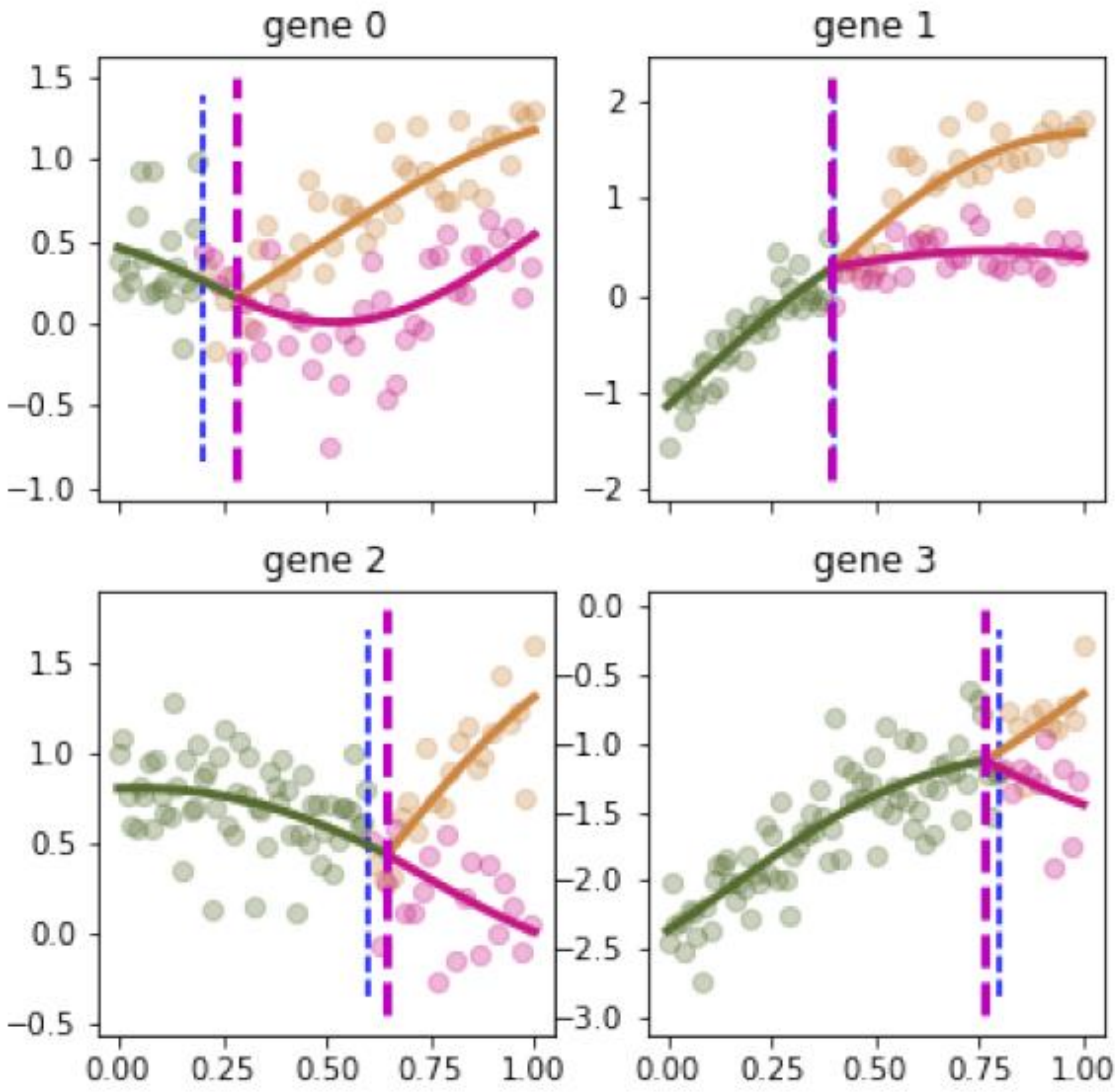
Modelling sequential branching dynamics with a multivariate branching Gaussian process

Elvijs Sarkans *elvijs.sarkans@gmail.com*
BIOS Health

Sumon Ahmed *sumon@du.ac.bd*
University of Dhaka

Magnus Rattray *magnus.rattray@manchester.ac.uk*
University of Manchester

Alexis Boukouvalas *alexis.boukouvalas@gmail.com*
PROWLER.io



Code

https://github.com/ManchesterBioinference/BranchedGP/

Uses:

GPflow and Tensorflow
Sparse variational inference
Mean-field variational inference of branch labels

Summary

Differential expression: Using GPs to model differential expression avoids assuming simple parametric forms (alternative to negative binomial GLMs)

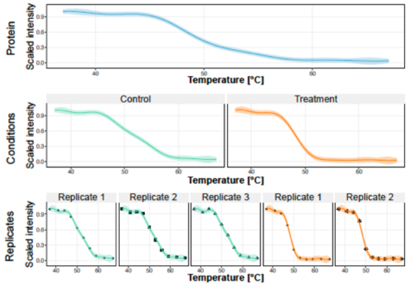
Protein melting: Hierarchical GP models can be used to share data across complex experimental designs (e.g. different protein isoforms/conditions).

mRNA degradation: GPs are tractable under linear operations, so we can use a simple linear ODE with a GP "force" term to model degradation.

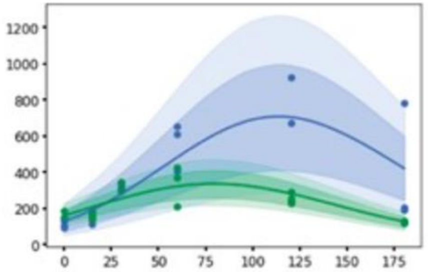
Branching: GPs are tractable under marginalization, so by marginalized out the point where two samples cross one can derive a branching model

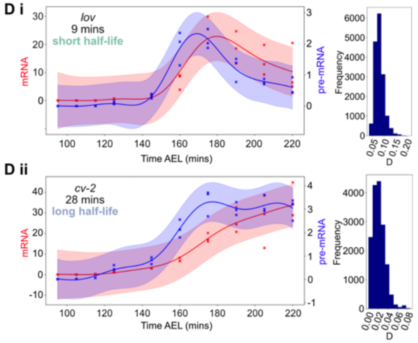
Acknowledgements

Mikhail Savitski, Cecile Le Sueur



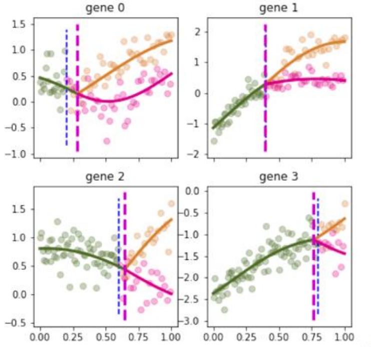
Ti John, Nuha BinTayyash, Sokratia Georgaka





Hilary Ashe, Lauren Forbes Beadle, Jennifer Love, Yuliya Shapovalova

Alexis Boukouvalis, Elvijs Sarkans, Sumon Ahmed



wellcome