Bayesian Machine Learning approach for modelling gene expression time series

Joseph Lee

University of Manchester

Introduction

- **Time-series RNA-seq** requires modelling of **complex temporal patterns**
- Traditional models (e.g., **Poisson**, **linear regression**) assume **fixed variance**, limiting use due to overdispersion from biological and technical noise
- **Gaussian Processes** with a **negative binomial likelihood** offer a **flexible**, **non-parametric**, **Bayesian** approach to model dynamic expression
- **Bayesian models** quantify uncertainty and adapt to data-driven patterns
- This study compares **GPcounts** and **maSigPro** for liver gene expression in fasting vs *ad libitum* mice, highlighting **model-dependent** and **sex-specific** transcriptional responses

Methods

- **Data**: 80 mouse liver samples (40 fasting, 40 *ad libitum*; 40 males, 40 females) were collected at 10 time points (4-hour intervals) with 2 replicates per point, resulting in a dataset of 20,545 genes
- **Sample collection:** Mice were kept in 12:12 light-dark cycles for 10 days, then switched to darkness from day 11 to preserve endogenous circadian rhythms

maSigPro

- Two-step regression with dummy variables and 4th-degree polynomial [2]

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + b_4x_1^4$$

- Gene expression y is modelled as a quartic polynomial of time x_1, with coefficients b defining the magnitude and direction of temporal changes
- Significance assessed via F-statistics (min obs. 15, FDR < 0.05, $R^2 \geq 0.6$) with Benjamini-Hochberg correction

GPcounts

- Gaussian Processes with a negative binomial likelihood [3]
- Non-parametric modelling with no prior assumptions about data distribution
- Kernels define similarity between time points, assuming a zero-mean latent function:

$$f(x) \sim GP(0, k(x, x'))$$

- A log link function maps the latent function to expected gene expression, ensuring positivity and stabilising variance:

$$\mu = exp(f(x))$$

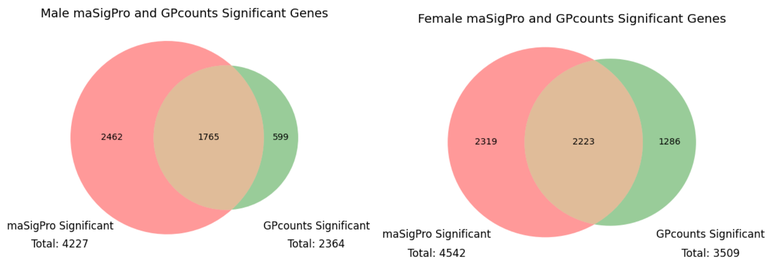
- Observed counts follow a negative binomial distribution capturing overdispersion:

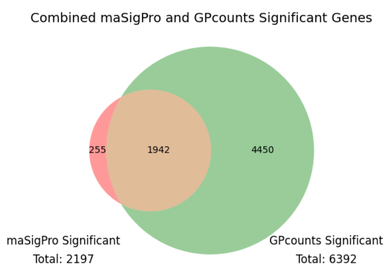
$$y \sim NB(\mu, \theta)$$

- Significance assessed using Log-Likelihood Ratio ((min obs. 15, FDR <0.05), LLR >0) with Benjamini-Hochberg correction

Model-Dependent Variation in Gene Detection

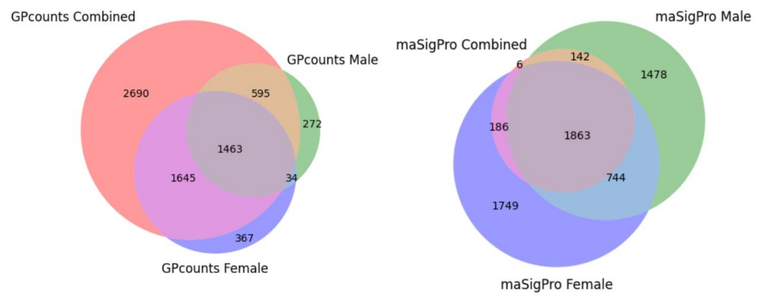
- **GPcounts** ((min obs. 15, FDR <0.05), LLR >0) and **maSigPro** (min obs. 15, FDR <0.05, R2 = 0.6) measured significantly differentially expressed genes (DEGs) in fasting and *ad libitum* single-cell liver mouse datasets





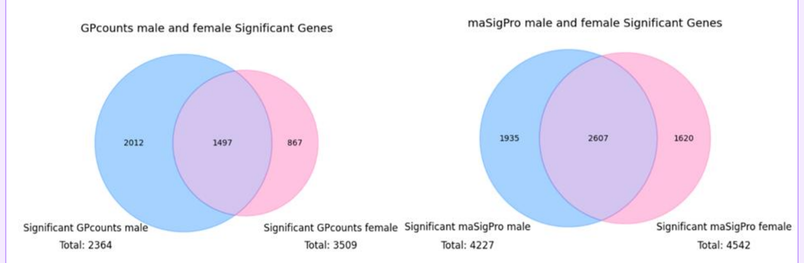
Significant Genes	Male	Female	Combined
Identified by maSigPro	4227	4542	2197
Identified by GPcounts	2364	3509	6392
Consistently Identified	1765	2223	1942
Unique to GPcounts	**599**	**1286**	**4450**
Unique to maSigPro	**2462**	**2319**	**255**

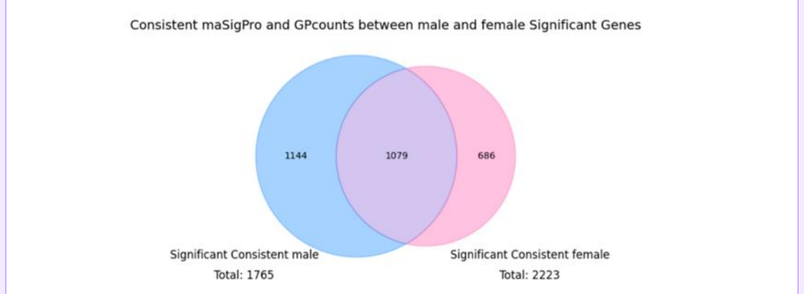
- **Top:** Significantly differentially expressed genes identified by GPcounts and maSigPro in male, female, and combined sex datasets



- **Bottom:** Combined comparison of significantly differentially expressed genes identified by GPcounts and maSigPro

Sex-Dependent Variation in Gene Detection

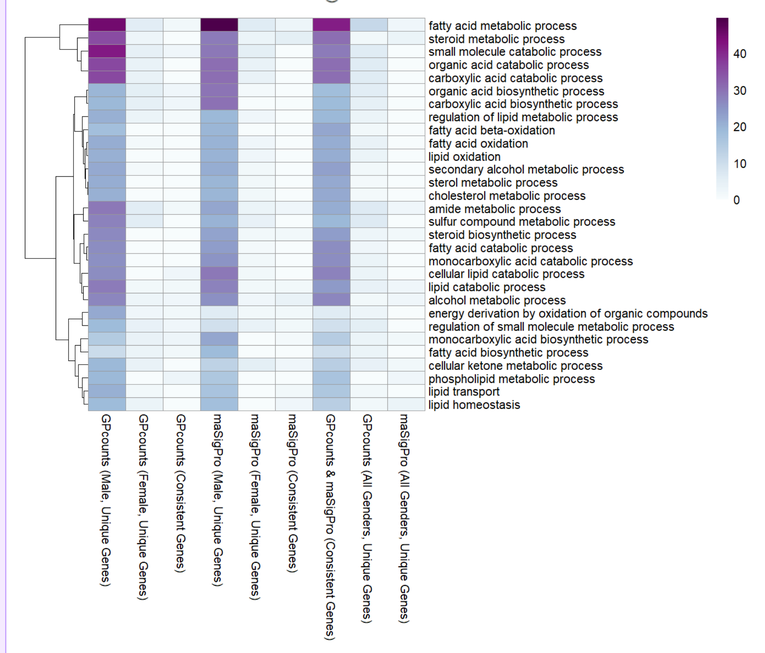




	Significant Male	Significant Female	Shared	Unique Male	Unique Female
maSigPro	4227	4542	2607	1935	1620
GPcounts	2364	3509	1497	2012	867
Combined	1765	2223	1079	1144	686

- **Sex-specific DEGs** were classified as male-unique, female-unique, or shared
- **maSigPro identified more shared DEGS**, suggesting better sensitivity to overlapping expression profiles
- Both models showed a greater number of male-unique DEGS
- Consistent trends indicate **robust sex differences** under fasting conditions, consistent across models

GO-Enrichment Heatmaps

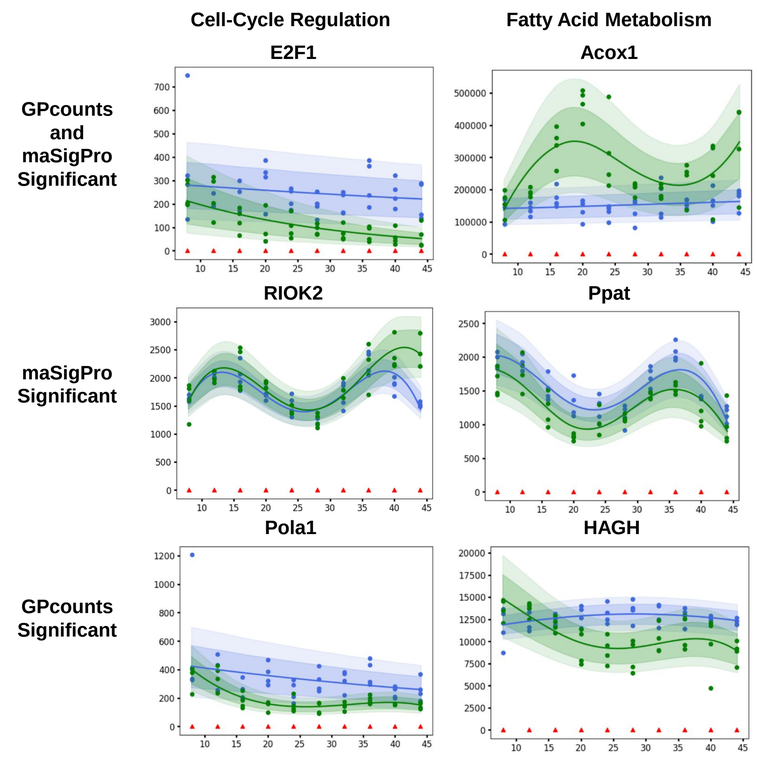


- **Core Metabolic Enrichment:** Shared DEGs identified by both models are strongly enriched for fatty acid metabolism ($-\log_{10} p > 15$)

- **Model-Specific Pathways (Top):**
- **GPcounts-unique** DEGs show high enrichment for cell-cycle processes ($-\log_{10} p > 15$)
- **maSigPro-unique** DEGs are broadly distributed over general catabolic pathways

- **Sex-Specific Enrichment (Bottom):**
- **Male-unique** DEGs (both models) are exceptionally enriched in fatty acid metabolism and small-molecule catabolism ($-\log_{10} p > 40$)
- **Female-unique** DEGs show moderate enrichment in fatty acid and sulphur compound metabolic processes ($-\log_{10} p > 20$)

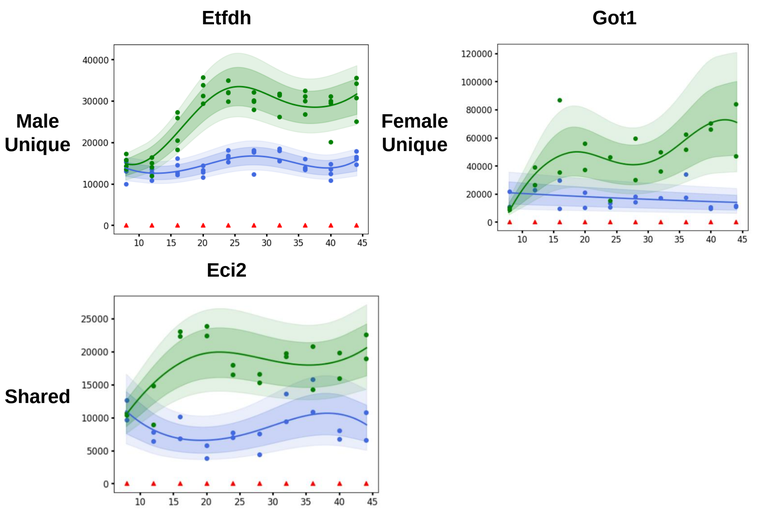
Comparison of maSigPro and GPcounts



Statistical model assumptions lead to the identification of distinct sets of differentially expressed genes, even with identical temporal trends

- **E2F1 –** S-Phase cyclin transcription and DNA replication regulation
- **RIOK2 –** ATPase preventing premature translation initiation
- **Pola1 –** encodes a catalytic subunit of DNA polymerase α
- **Acox1 –** acyl-CoA oxidase – the first enzyme in the fatty acid β-oxidation pathway
- **Ppat –** amidotransferase involved in purine biosynthesis
- **HAGH –** Mitochondrial enzyme involved in glutathione metabolism

Sex-dependent gene visualisation



- **Minor differences in upregulation may reflect sex-specific lipolysis and amino acid catabolism** [4]
- **Etfdh –** mitochondrial electron transfer system enzyme
- **GOT1 –** cytosolic enzyme involved in amino acid metabolism, energy production, glucocorticoid responses, and steroid metabolism
- **Eci2 –** Key mitochondrial enzyme involved in the β-oxidation of unsaturated fatty acids

Conclusion and Future Directions

Model choice profoundly influences the discovery of differentially expressed genes

- **Model Comparison:** maSigPro and GPcounts detect key metabolic DEGs but diverge on cell-cycle and sex-specific signals
- **GPcounts Strengths:** Flexible overdispersion modelling, captures subtle signals (e.g., cell-cycle regulation)
- **GPcounts Limitations**: Computationally intensive, environment-specific failures, and reproducibility challenges
- **maSigPro Strengths:** Fast, reproducible, ideal for rapid exploratory analysis in resource-limited settings
- **maSigPro Limitations:** Rigid polynomial model, Gaussian error assumptions; may miss complex dynamics

- **Future Directions: CRISPR-based validation** of key genes, migration of GPcounts to **GPyTorch** for improved scalability and stability; **periodic kernel** integration to model circadian rhythms, and application of **shrinkage** to reduce potential overfitting

Acknowledgements

I want to thank the Rattray Group, specifically Magnus and Sokratia, for their help with the GPCounts package. I'd also like to thank Jessica and the Fustin group for providing the data used in this study.

References

1. C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. c 2006 Massachusetts Institute of Technology.
2. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics. 2006 May;22(9):1096–102.
3. Nuha BinTayyash, Sokratia Georgaka, S T John, Sumon Ahmed, Alexis Boukouvalas, James Hensman, Magnus Rattray, Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments, Bioinformatics, Volume 37, Issue 21, November 2021, Pages 3788–3795
4. Bazhan, N., Jakovleva, T., Feofanova, N., Denisova, E., Dubinina, A., Sitnikova, N., & Makarova, E. (2019). Sex Differences in Liver, Adipose Tissue, and Muscle Transcriptional Response to Fasting and Refeeding in Mice. Cells, 8(12), 1529.