Gaussian Process Models for Time-Series Omics Analysis

Sahil Loomba[1], Diogo M. Camacho[1], James J. Collins[1,2]

[1]Wyss Institute for Biologically Inspired Engineering at Harvard University, [2]Massachusetts Institute of Technology

Abstract

Most biological systems have associated temporal dynamics, which are key to their complete mechanistic understanding. Regulation and expression of genes, their translation into proteins, and their consequent effects on biochemical reactions, are events unfolding in accordance with their respective kinetics. Current omics analyses focus on single snapshots of a system's state, either at a particular point in or averaged over time. Here we present two approaches to principally model time-series omics data using Gaussian process regression, and introduce techniques for comparative analysis of multiple phenotypes. The univariate conditioned model regresses (gene) expression on time given any categorical phenotype, while the bivariate joint model regresses on both time and a continuous phenotype. We applied these two modeling strategies to transcriptional data of frog embryos infected with four different initial doses of *Pseudomonas aeruginosa,* collected over the first 3 days of development. We identified key genes that have markedly different temporal responses under different infection regimens. Additionally, we were able to cluster genes based on their expression landscape given time. We validated these gene clusters by establishing their topological smoothness over the *Xenopus* gene regulatory network. Performing gene set enrichment we were able to characterize groups related to pathways and processes involved in host-pathogen interactions.

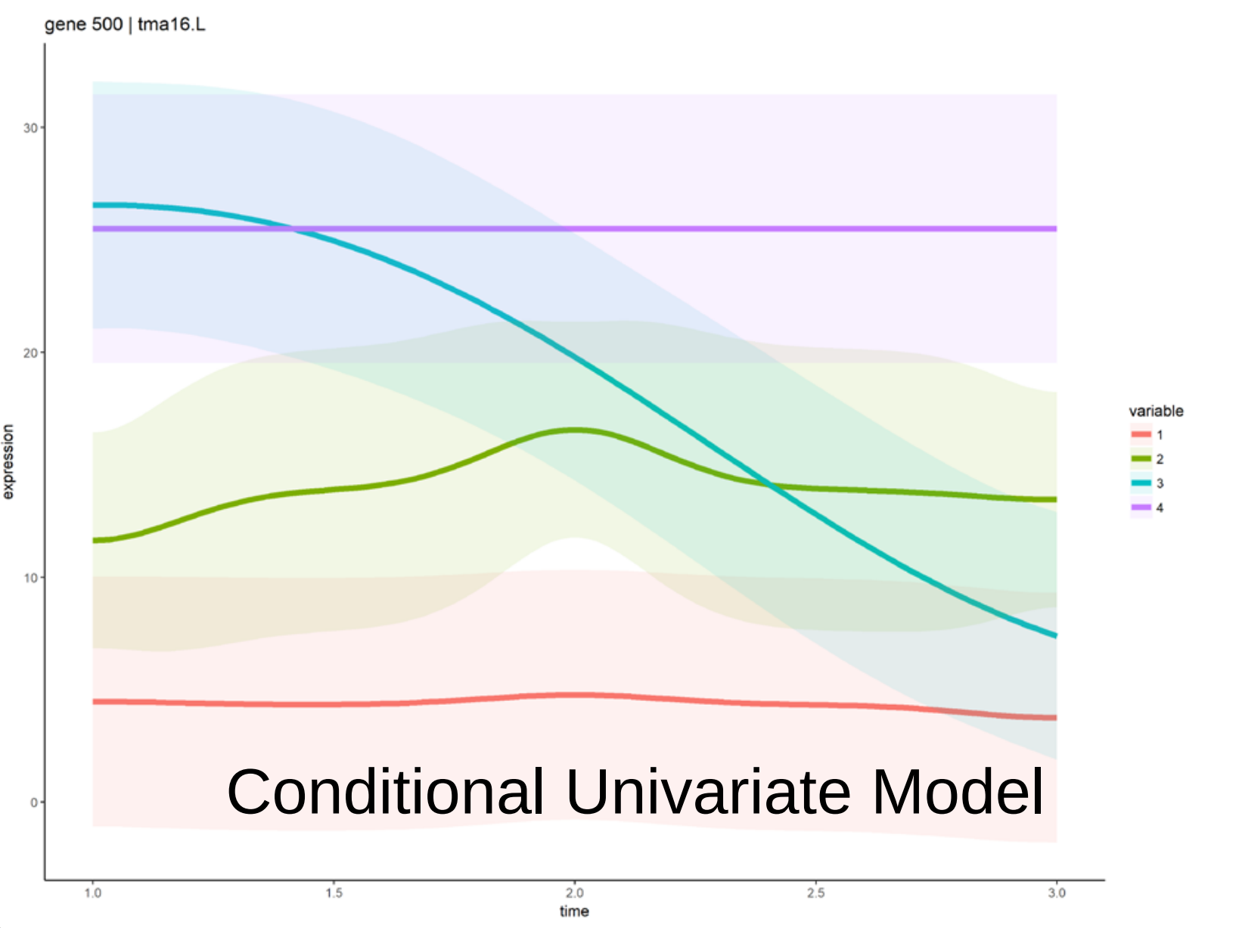
Data

We obtained transcriptomics data for frog embryos infected with *P. aeruginosa* at different infection levels ("conditions"), and at multiple stages of development. Probes were mapped to 8726 *Xenopus* genes

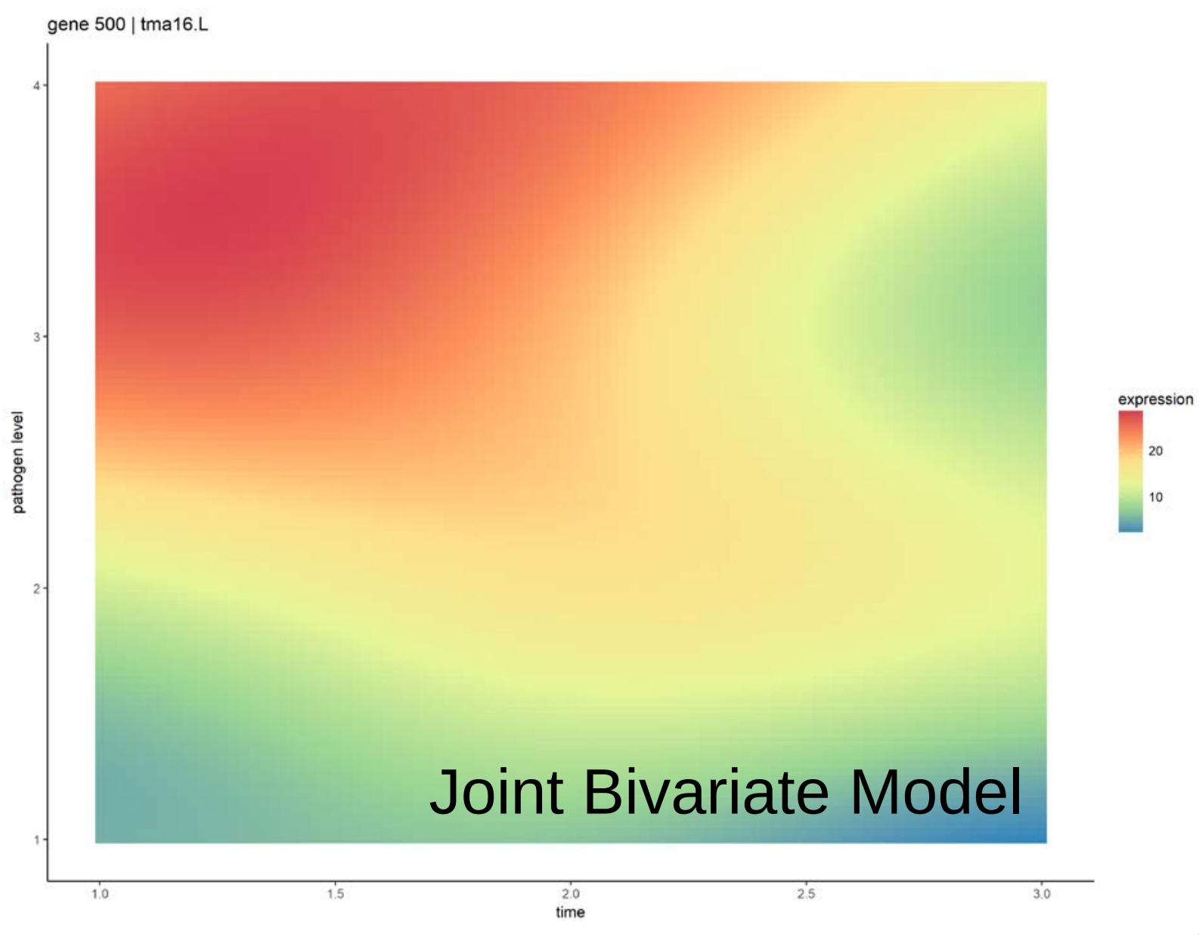
Infection load (cfu)	# Replicates	Time points (day)
0	2	1, 2, 3
100	3	1, 2, 3
1000	3	1, 2, 3
10000	3	1, 2

Method

We develop two methods of modelling omics using GPR: (1) the conditional univariate method, where every gene (for a *given* condition) is an independent GP mapping from "time" to "expression" space, and (2) the joint bivariate method, where every gene is an independent GP mapping from "time X condition" to "expression" space. We define the idea of "key genes" experiencing maximum "separation" across the 4 conditions. We plot one such gene, TMA16:



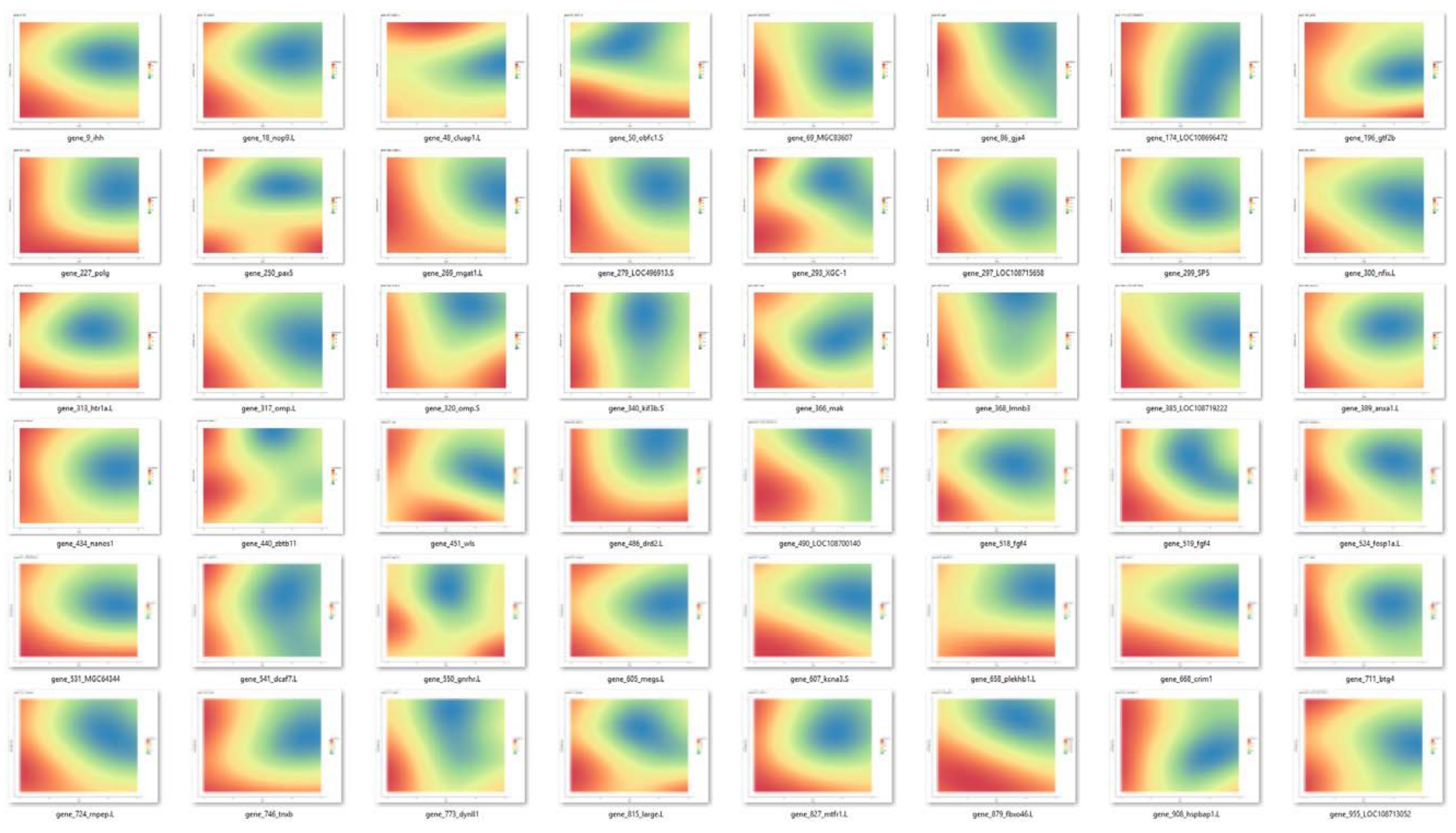
Conditional Univariate Model



Joint Bivariate Model

Conclusions

We developed model(s) for time-series data under multiple conditions, with limited number of samples; good tool for quick visual inspections (such as the heat-map below), outputs key genes and gene groups mapped to appropriate pathways; can be generalized to any temporal dataset with multiple conditions, say in proteomics or metabolomics.



Next steps include using actual CFU counts of samples in the joint bivariate model, and build a browser-based app for everyone to use.

Acknowledgements: We are very thankful to Richard Novak, Vishal Keshari and Alexandre Dinis for performing the embryo infections and microbiology experiments. This work is supported by DARPA THoR 15-21.

Objective

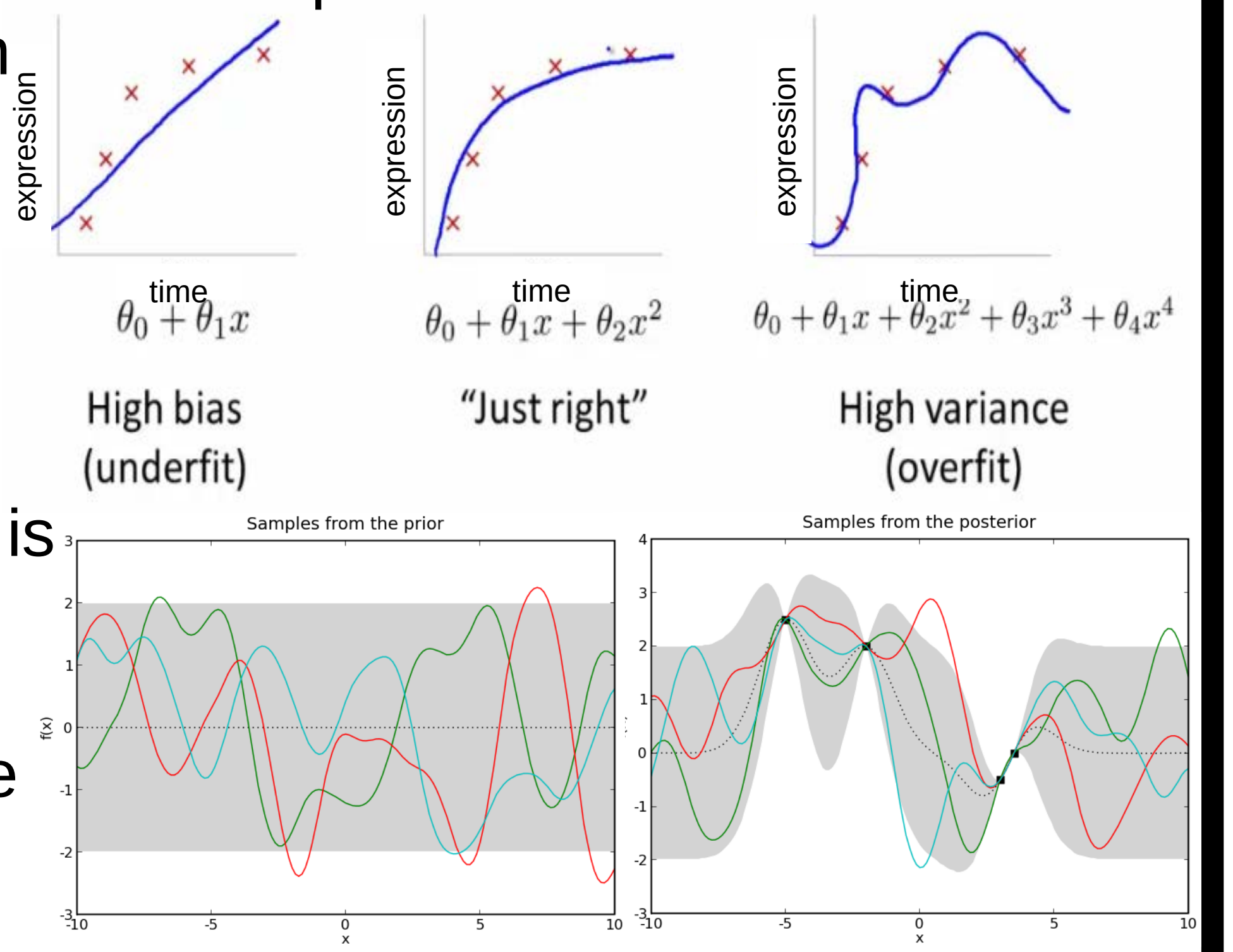
To develop a transcriptomics model that captures key differences across treatment conditions of a biological system. It should:

1. Be sensitive to (is a function of) time: usual differential expression analysis takes place at a particular time point, which can ignore dynamical information about the system's evolving phenotype
2. Solve key challenges of most omics studies:
   1. Few samples per condition
   2. Even fewer samples per condition per time point
   3. Sample variance
   4. Error of measurement

Concept

A regression model that finds a mapping from a domain to a range is parametric, in the sense that it describes a model with parameters which are then inferred given the data. However, this method of performing regression often suffers from the problem called overfitting, especially when few data samples are available.

On the other hand, a Gaussian Process regression considers the mapping between two feature spaces as a random variable itself, following a GP prior whose covariance is a function of the observations. It is therefore "non-parametric", whose complexity grows with more data. If the domain space is time, it can model a time-series.



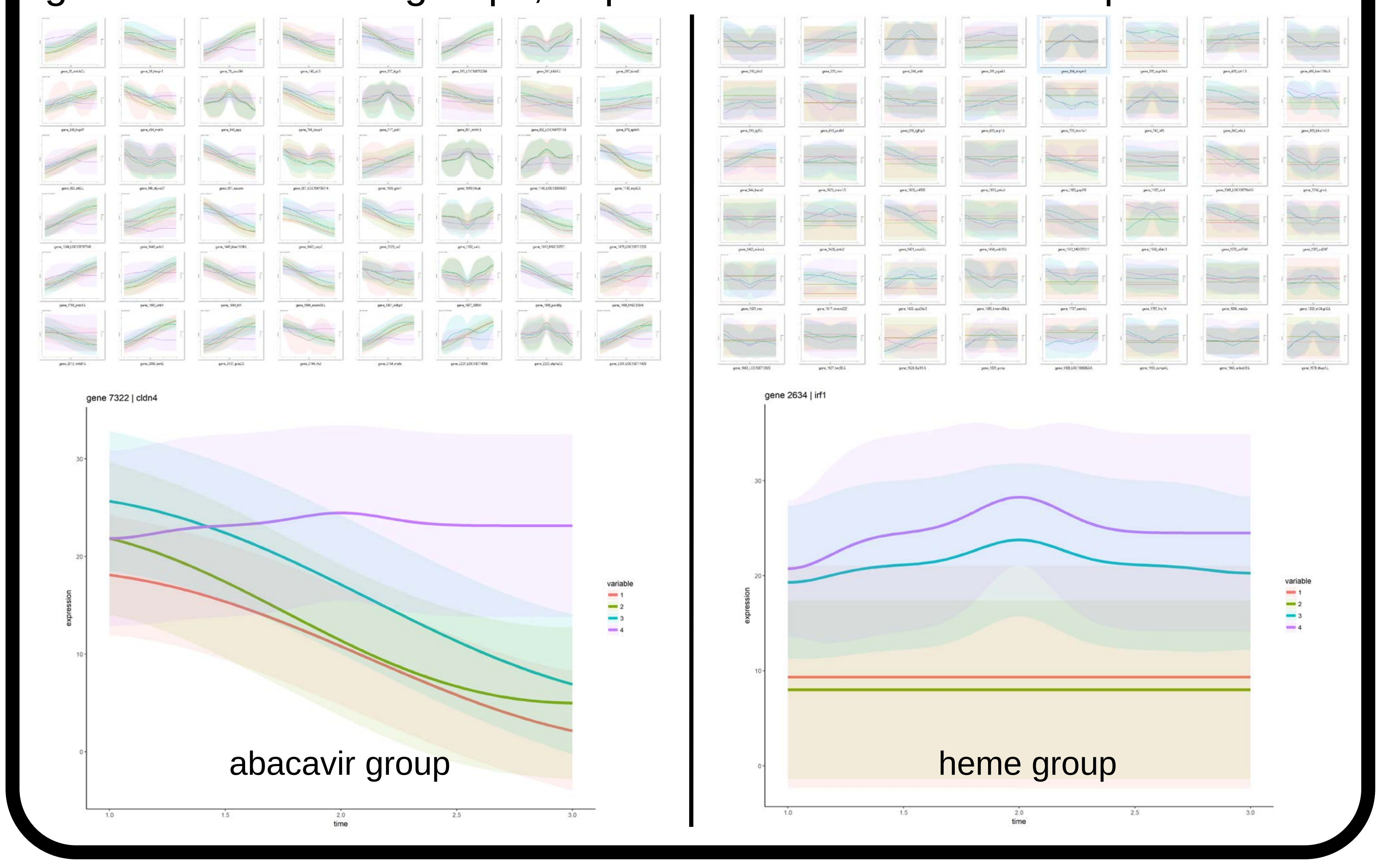
Other advantages include: (1) there is no need to specify a model of regression, (2) it only makes a "smoothness" assumption, (3) it makes predictions in unseen spaces with bounds on the uncertainty, and (4) it decouple various aspects of the data like the bias, scale, mean and covariance functions, and Gaussian noise terms.

Analysis

We want to "group" those genes that provide a "similar separation" of the 4 conditions. After defining such a criterion, we use hierarchical clustering of genes in the GP model space to find 42 such "gene groups". To make sense of these clusters, we map each one to appropriate biological pathways using gene-set enrichment analysis. Being a many-to-many map, we filter out the "significant clusters" by picking those which map to exactly one "key pathway".

Gene Group	Mapped Pathway
4	g2/m_dna_replication_checkpoint
11	rna_polymerase_ii_transcription
17	dcc_mediated_attractive_signaling
20	**abacavir_transport_and_metabolism**
25	norc_negatively_regulates_rrna_expression
28	glycogen_synthesis
37	**scavenging_of_heme_from_plasma**

Interestingly, while the abacavir pathway is associated with HIV, the heme scavenging pathway seems to hint at the role of iron metabolism in host-pathogen interactions. We can readily visualize all genes in these two groups, or pick a few out for visual inspection.



abacavir group

heme group