WYSS INSTITUTE

Time Series Transcriptomics

On Frog Embryos infected with *P. Aeruginosa*

Extracting Key Genes

and Gene Groups

SAHIL LOOMBA

Time Series Data

- Transcriptomics data for frog embryos infected with *P. aeruginosa*
  - Different infection levels
  - Different stages of development
  - Replicates

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

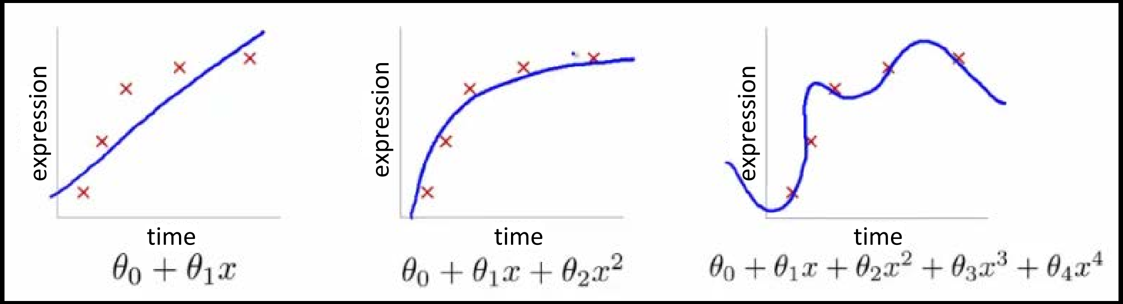
- Probes mapped to 8726 frog genes from Xenbase

Objective

- To develop a transcriptomics model that captures key differences across treatment conditions
- Sensitive to (is a function of) time
  - Usual differential expression analysis takes place at a particular time point
- Challenges
  - Few samples per condition (6-9)
  - Even fewer samples per condition per time point (0-3)
  - Sample variance and error of measurement
- Can we still capture temporal and condition resolution?
  - While principally handling these issues

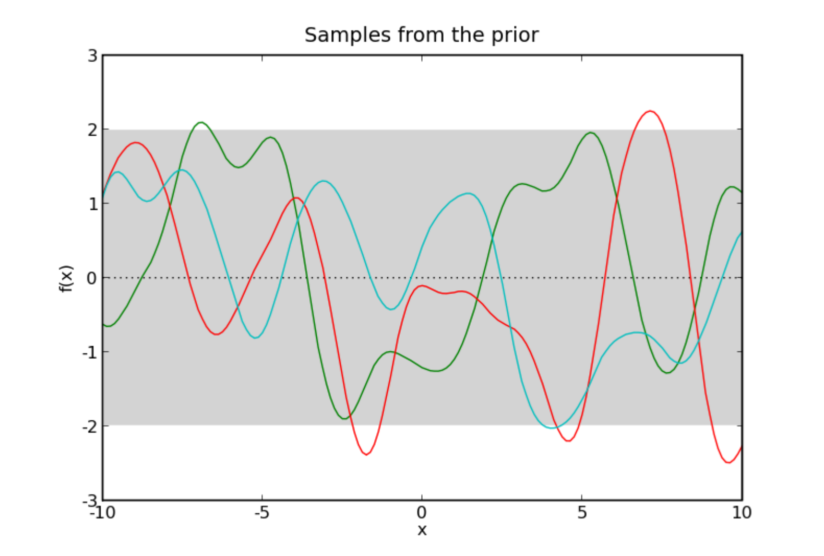
Gaussian Process Regression

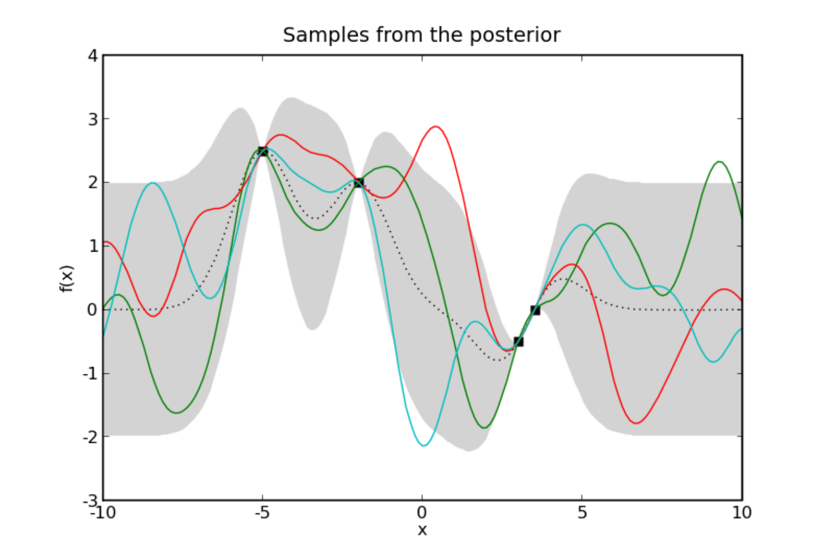
- Regression: Given a dataset, finding a mapping from a domain to a range; usually “parametric”

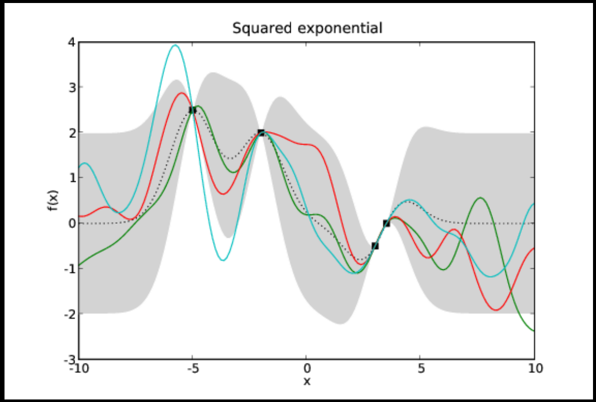


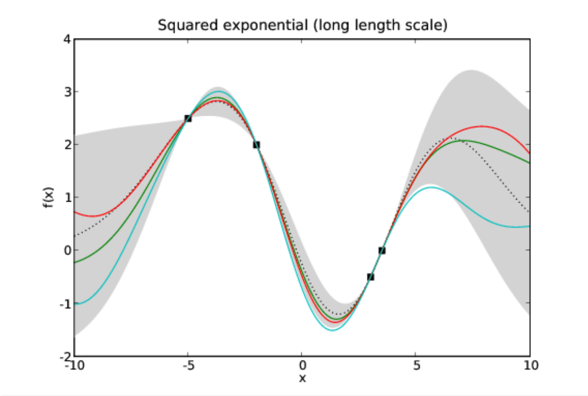
- GPR: Considers the mapping between two feature spaces as a random variable itself, following a Gaussian Process prior whose covariance is a function of the observations (“non-parametric”)
- If the domain space is time, it can model a time-series

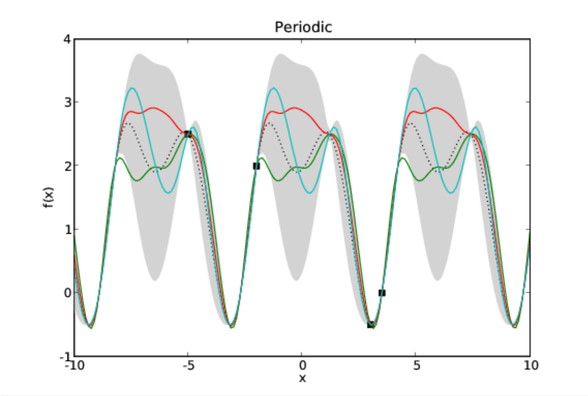
Gaussian Process Regression


Samples from the prior


Samples from the posterior


Squared exponential


Squared exponential (long length scale)

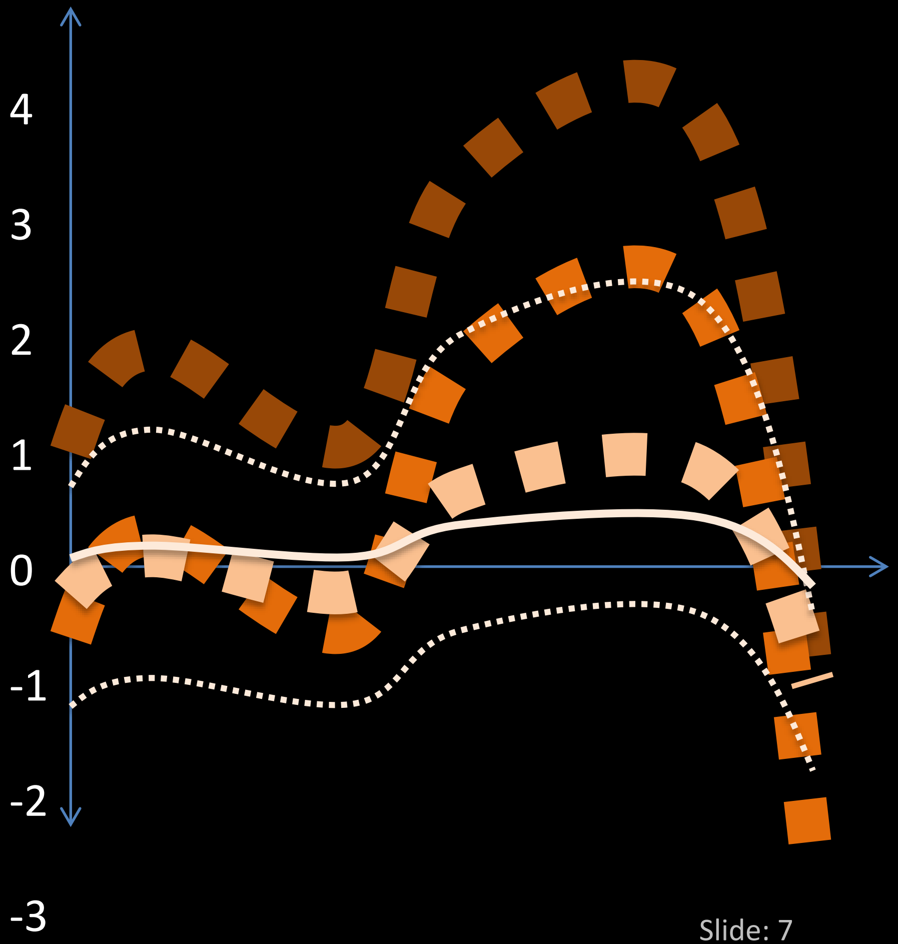

Periodic

Gaussian Process Regression

- No need to specify a model of regression
  - Assumption of “smoothness”
- Automatic Occam’s Razor
- Makes predictions in previously unseen spaces with bounds on uncertainty of the prediction

Gaussian Process Regression

- Decouples various aspects of the data:
  - Bias: *b*
  - Scale: *s*
  - Mean function: *m(x)*
  - Variance around mean function: *k(x,x')*
  - Error term: *e*

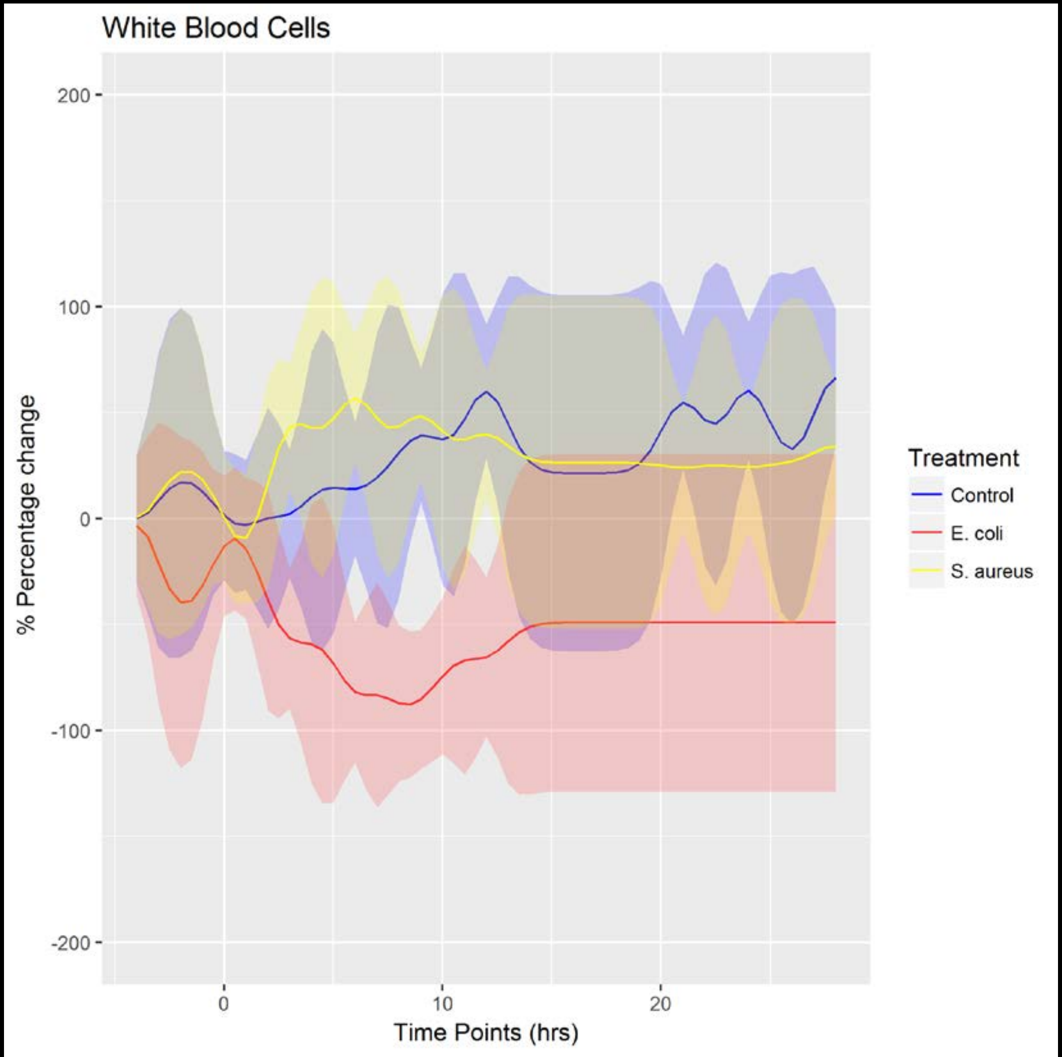


Gaussian Process Regression

For the pig studies:

Biomarkers/Physiological data for

- 15 pigs across
- 3 conditions



GPR for Omics, Method 1

- Treat every gene (for a given condition) as an independent GP mapping from time space to expression space

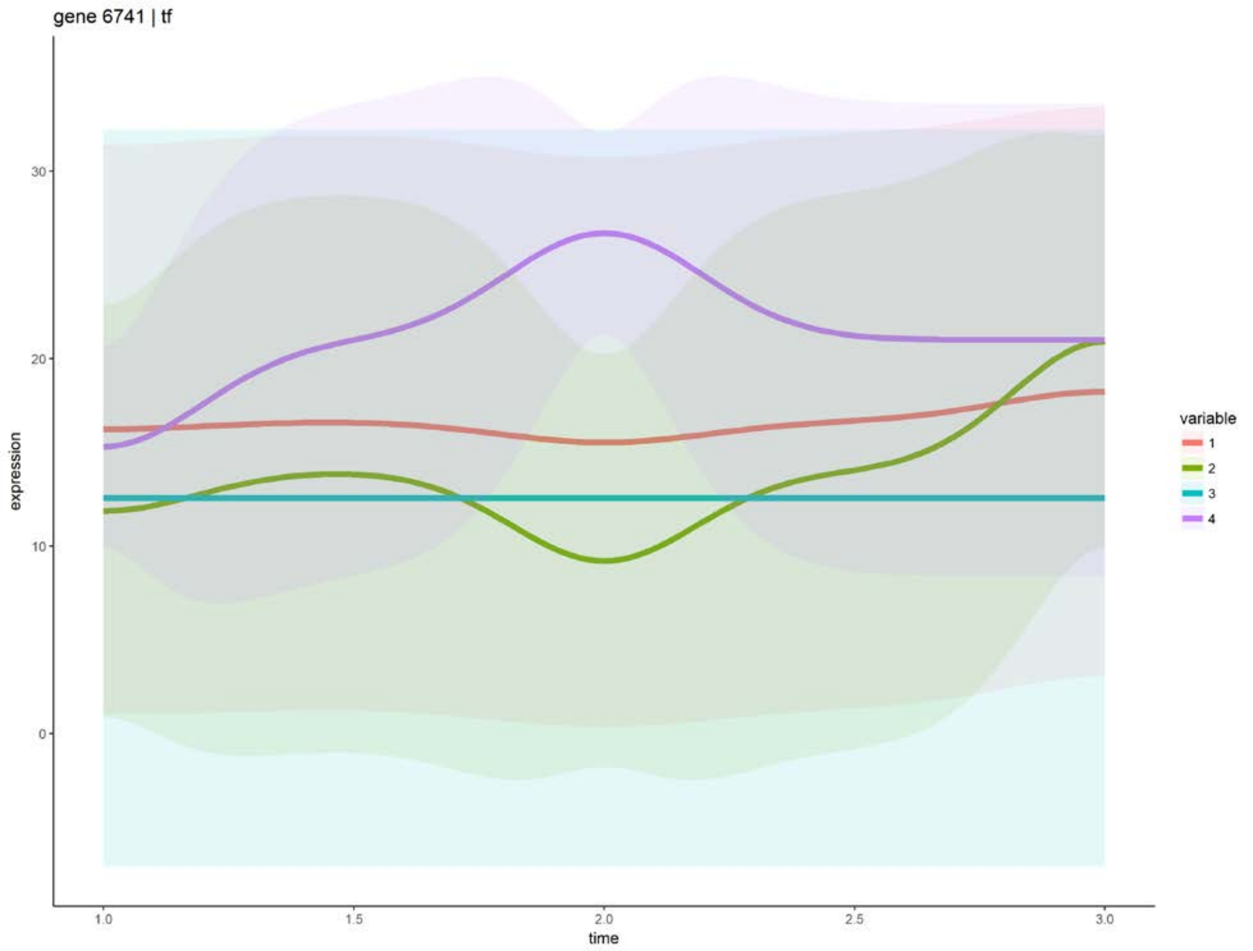
$$f: t \rightarrow g$$

$$f(t') \sim GP\big(m(t), k(t, t')\big) + \beta$$

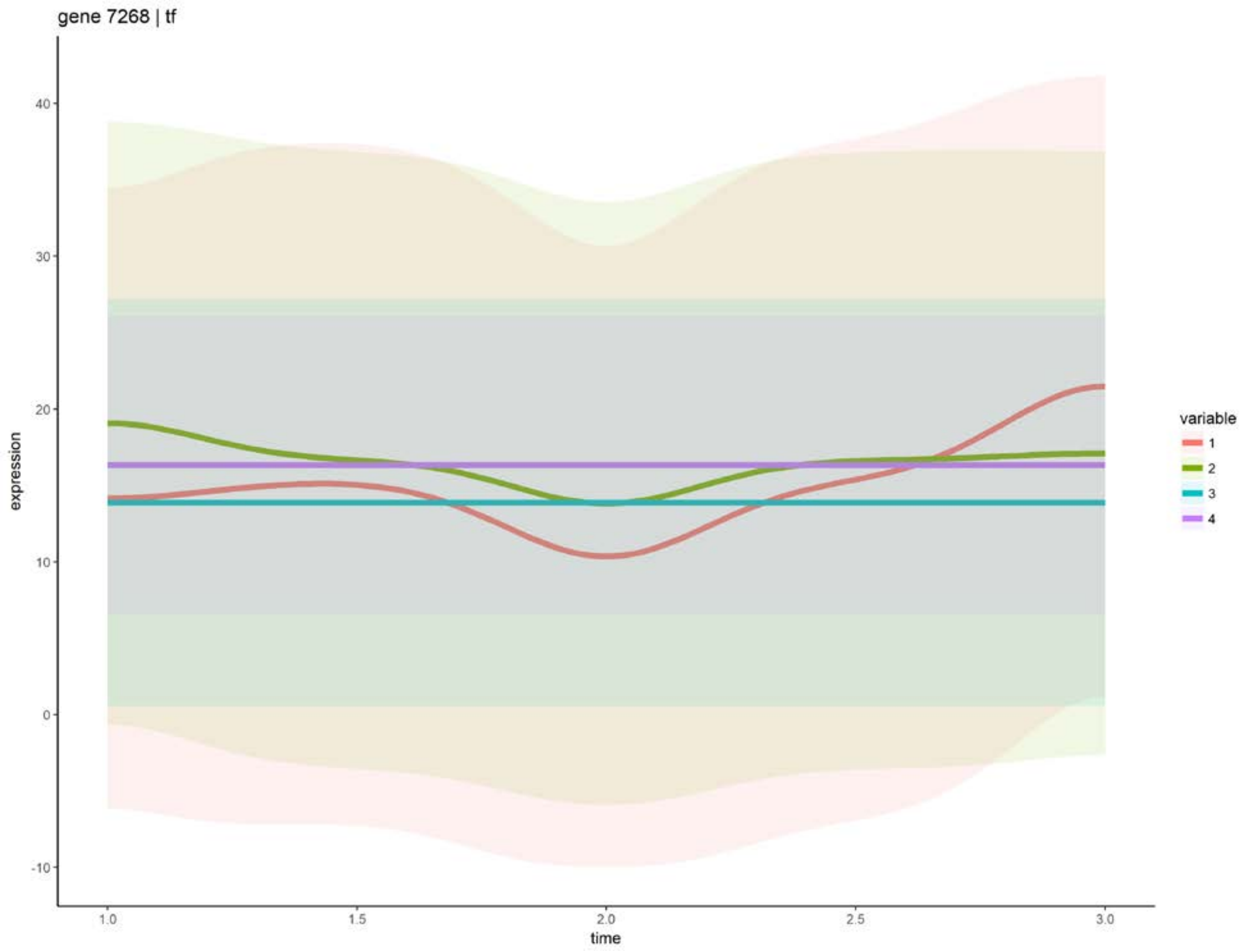
- For every gene. learn 4 GP models corresponding to the 4 conditions
- We can plot the output of a learnt GP model by plotting its posterior mean estimate and posterior covariance estimate

WYSS INSTITUTE

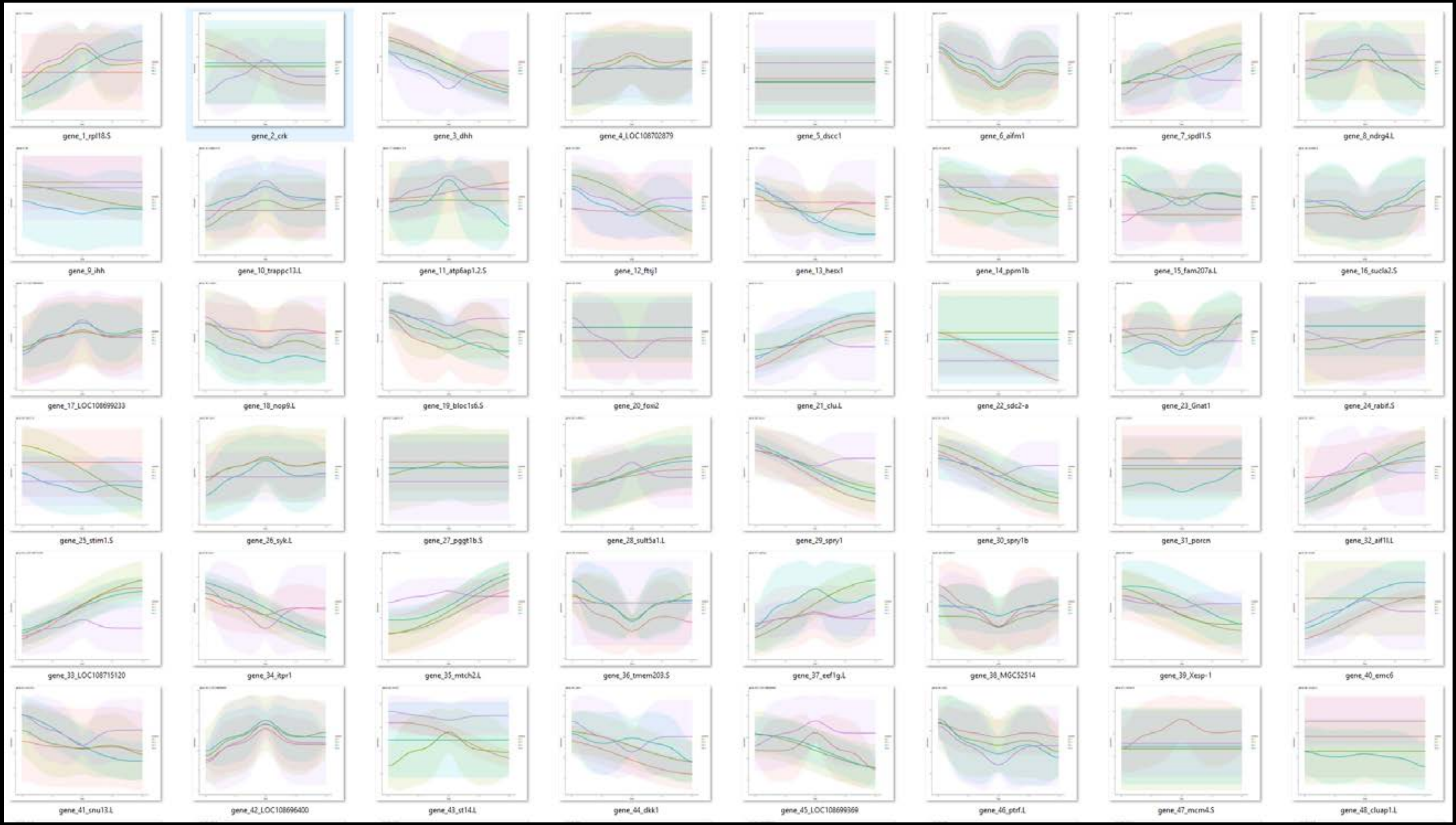
Example: Transferrin (probe1)



Example: Transferrin (probe2)



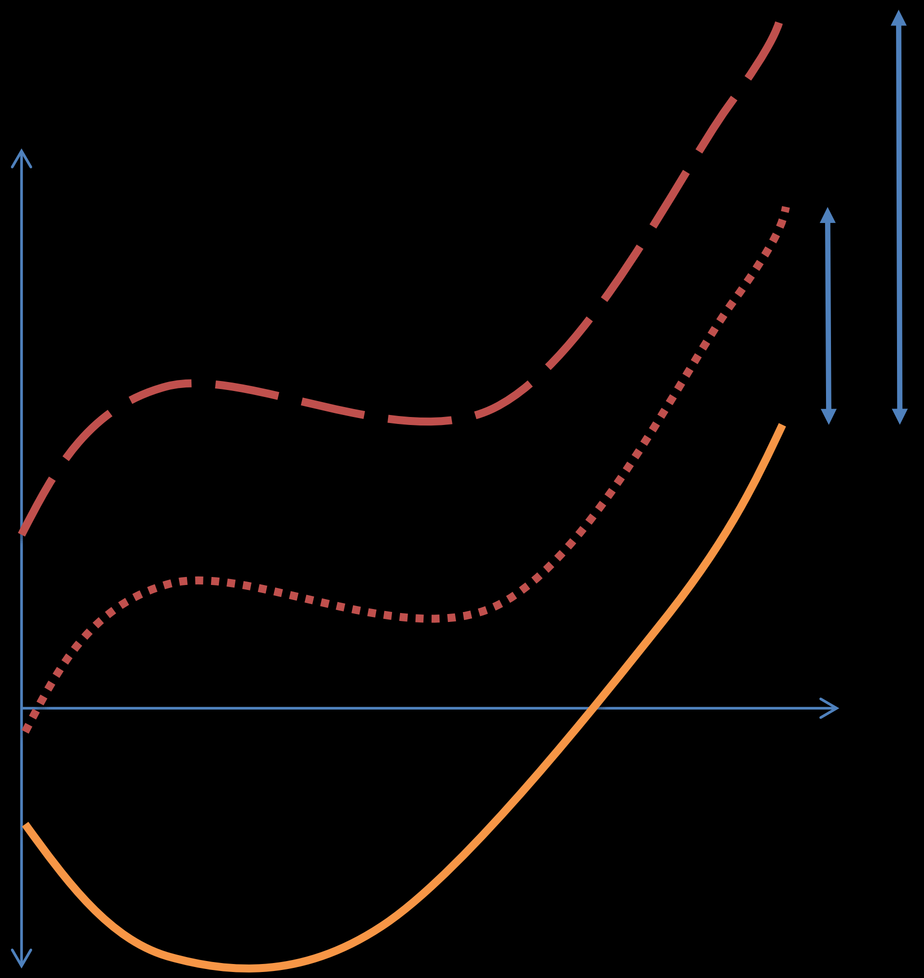
For all 8726 genes…



To extract key genes and gene groups…

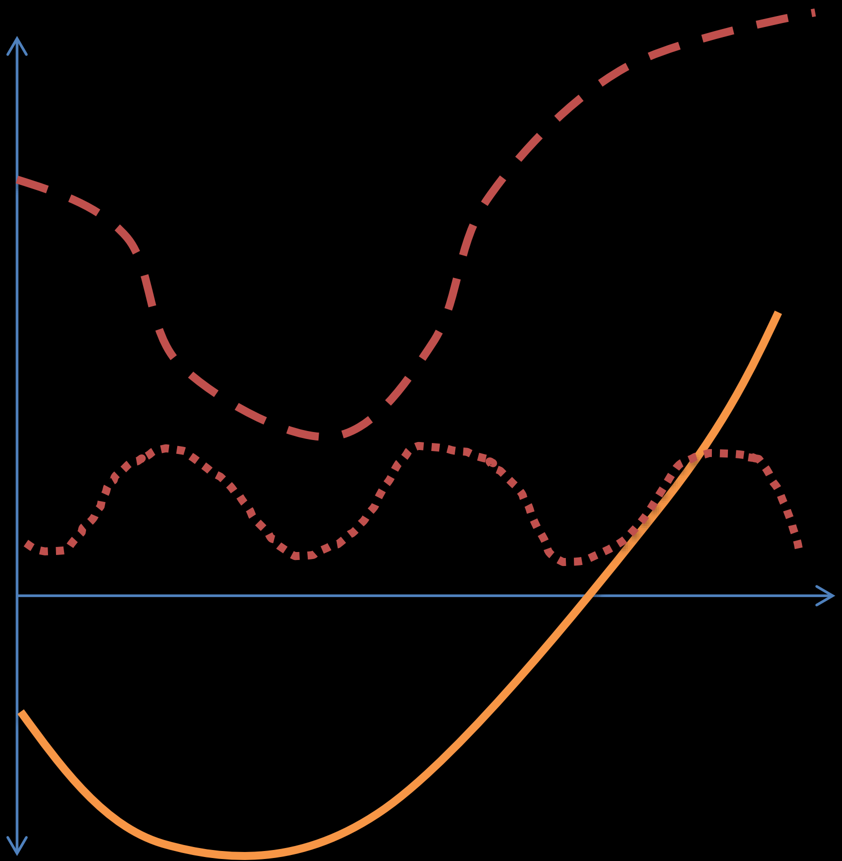
- Visually!
- Question: How do we compare two GP models?
  - Can't compare in model space because GP is non-parametric and we have different samples to every GP
- Solution: use ideas from information theory to recharacterize every model
  - Treat output (posterior mean estimate) of learnt GP as the new "smoothed" time-series
  - Define metrics to compare those smooth series

Metric 1: (Euclidean) Distance



The further away two conditions are, the better a gene is in discriminating between them

Metric 2: (Mutual) Information



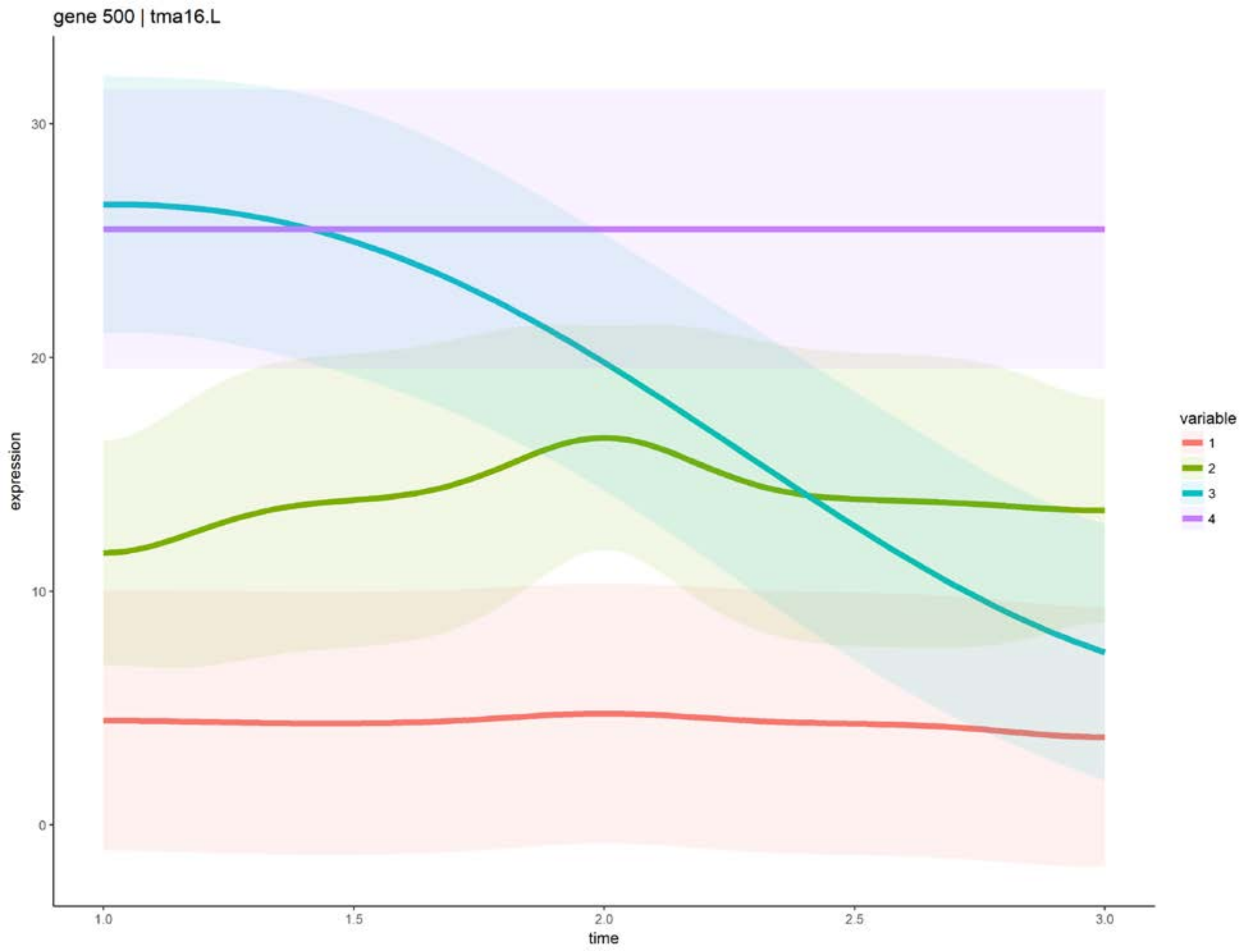
The more informative (~correlated*) one condition is about the other, the more it matters

*Unlike correlation, mutual information doesn’t assume linearity; more so a notion of predictability of one time-series from the other

Definition of Key Genes

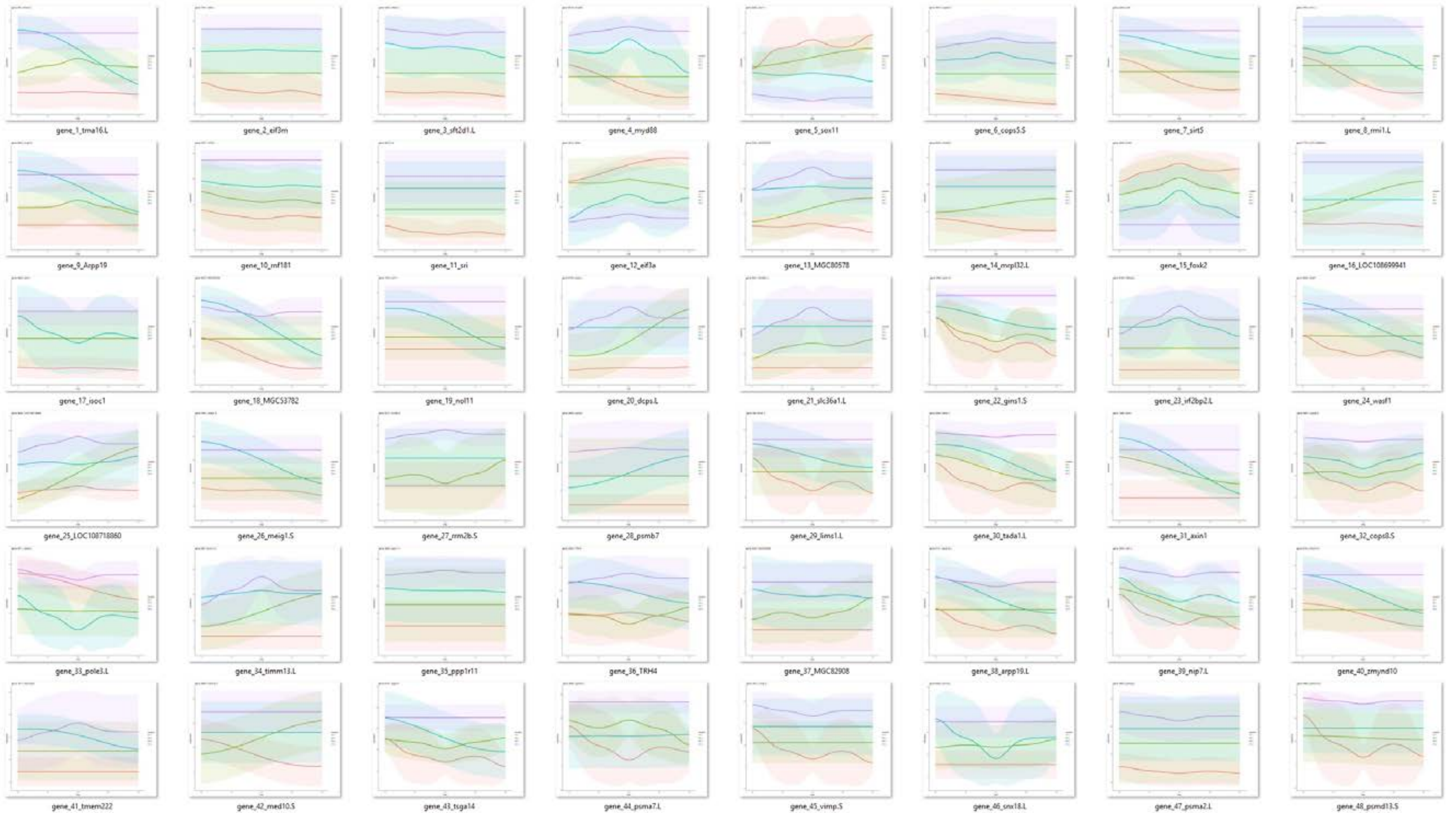
- Which induce maximum sum of pairwise distances between all informative conditions
- We rank genes in decreasing order of the combined score
- (But we present results of Euclidean too, separately…)

Example: Euclidean Only



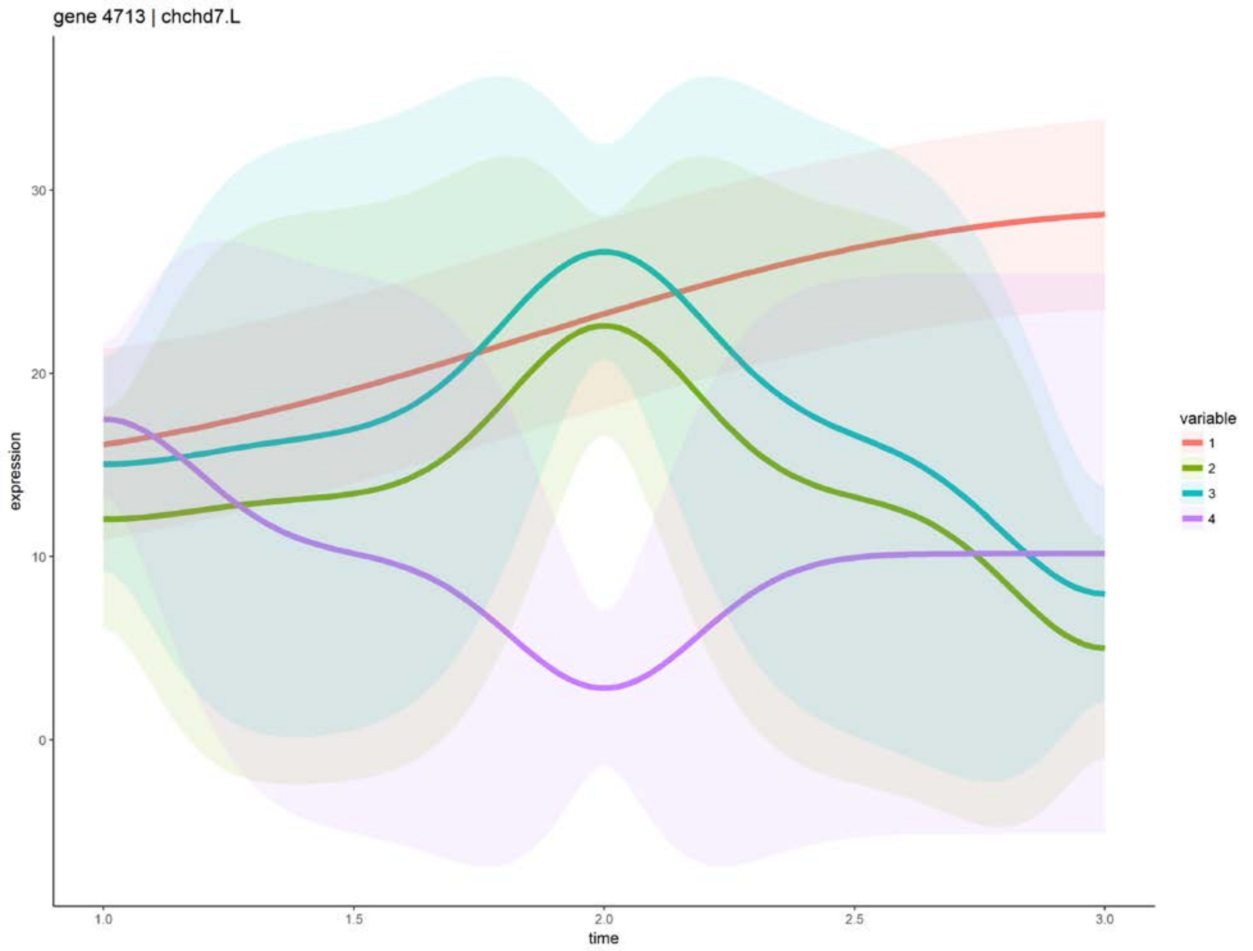
WYSS INSTITUTE

Examples: Euclidean Only

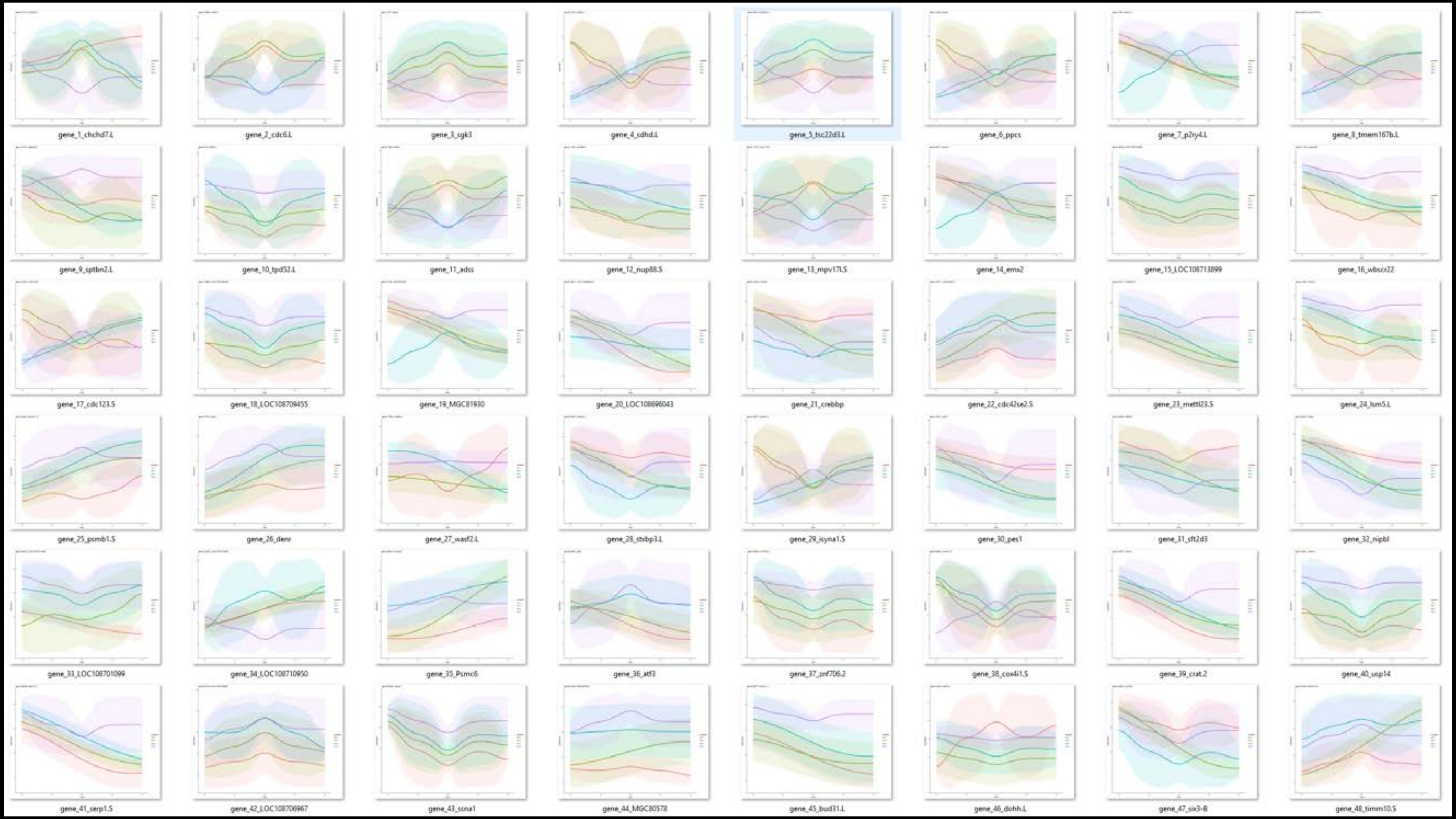


WYSS INSTITUTE

Example: Combo



Examples: Combo



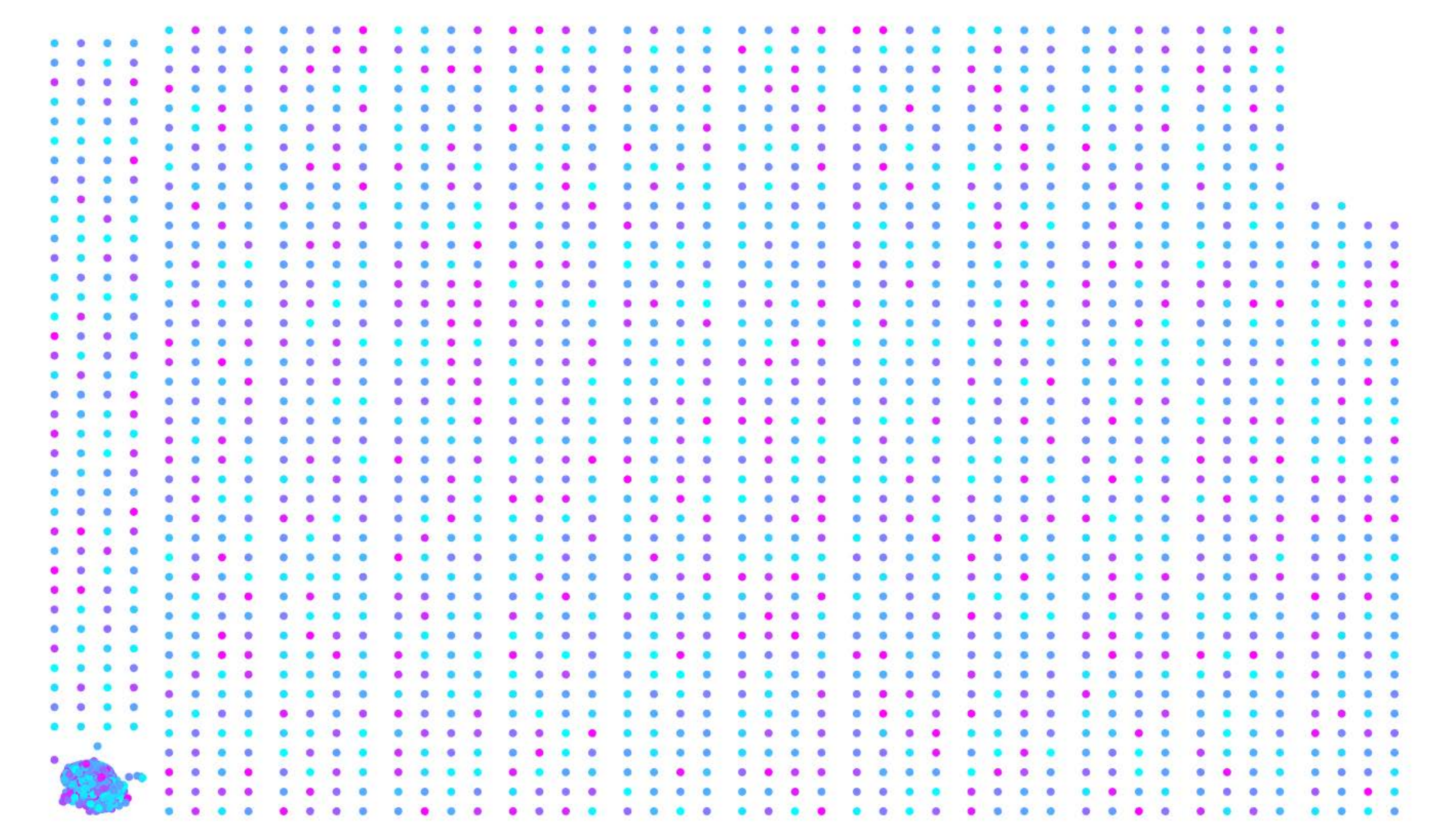
Definition of Gene Groups

- We want to “group” those genes that provide a “similar separation” of the four conditions
  - Conditions have similar pairwise distances
  - Conditions are similarly informative of one another
- We use the two metrics to define a “feature space” of genes, and turn this into a clustering problem
  - “Distance over distances”
  - Can do regular hierarchical clustering on this!
  - Scale issues?
    - MI is scale-invariant
    - Distance is normalized such that maximum distance within a gene is unity

Hierarchical Clustering of Genes

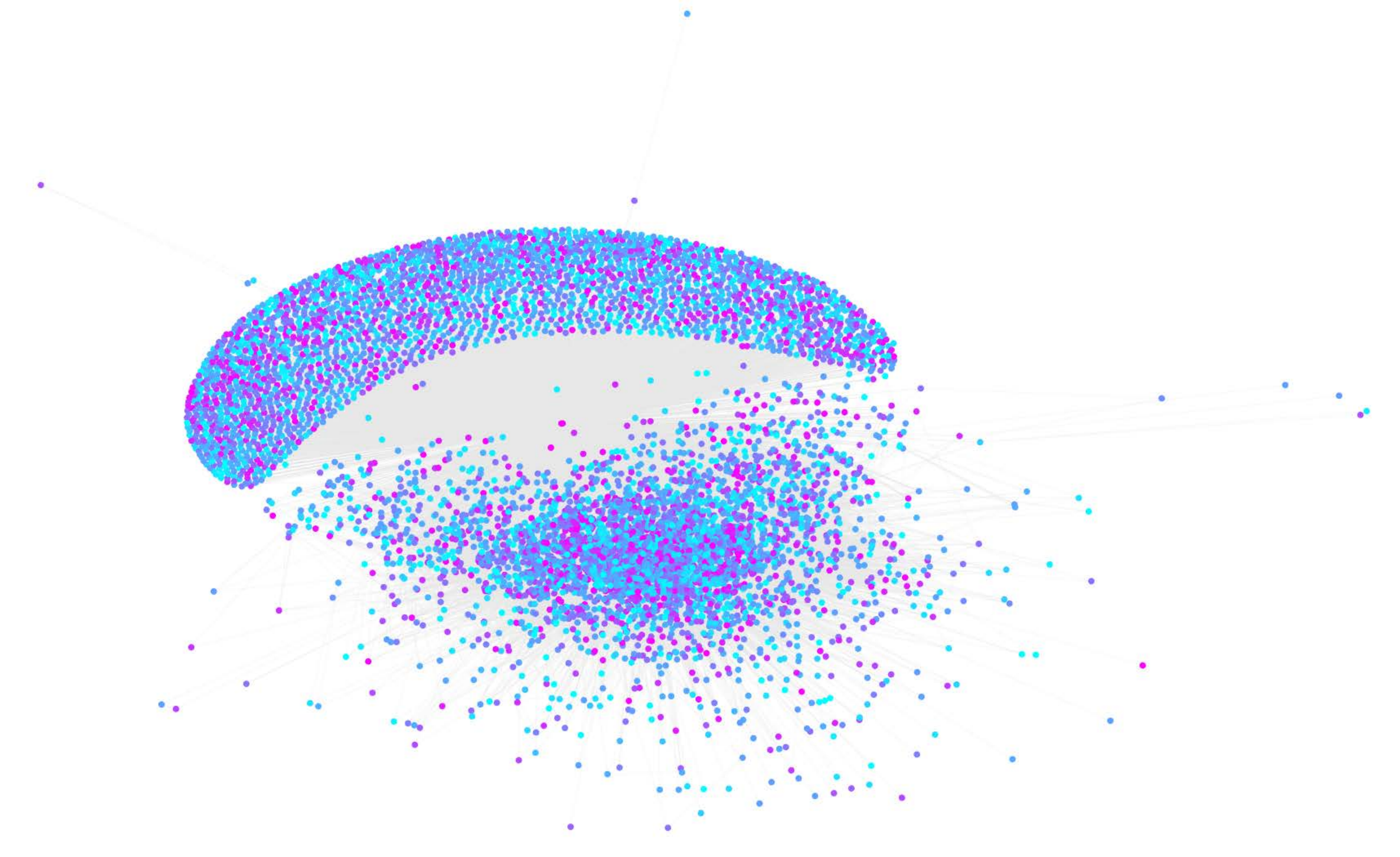
- We experiment with different cluster sizes: 2, 4, 42, …
- How do we diagnose the goodness of this clustering?
  - – Use the GRN of Xenopus to define ground truth relationships between genes
  - – Do we observe a topological smoothness of gene labels?
    - Intuition: genes closer in regulation would be in the same group

Xenopus GRN Visualized

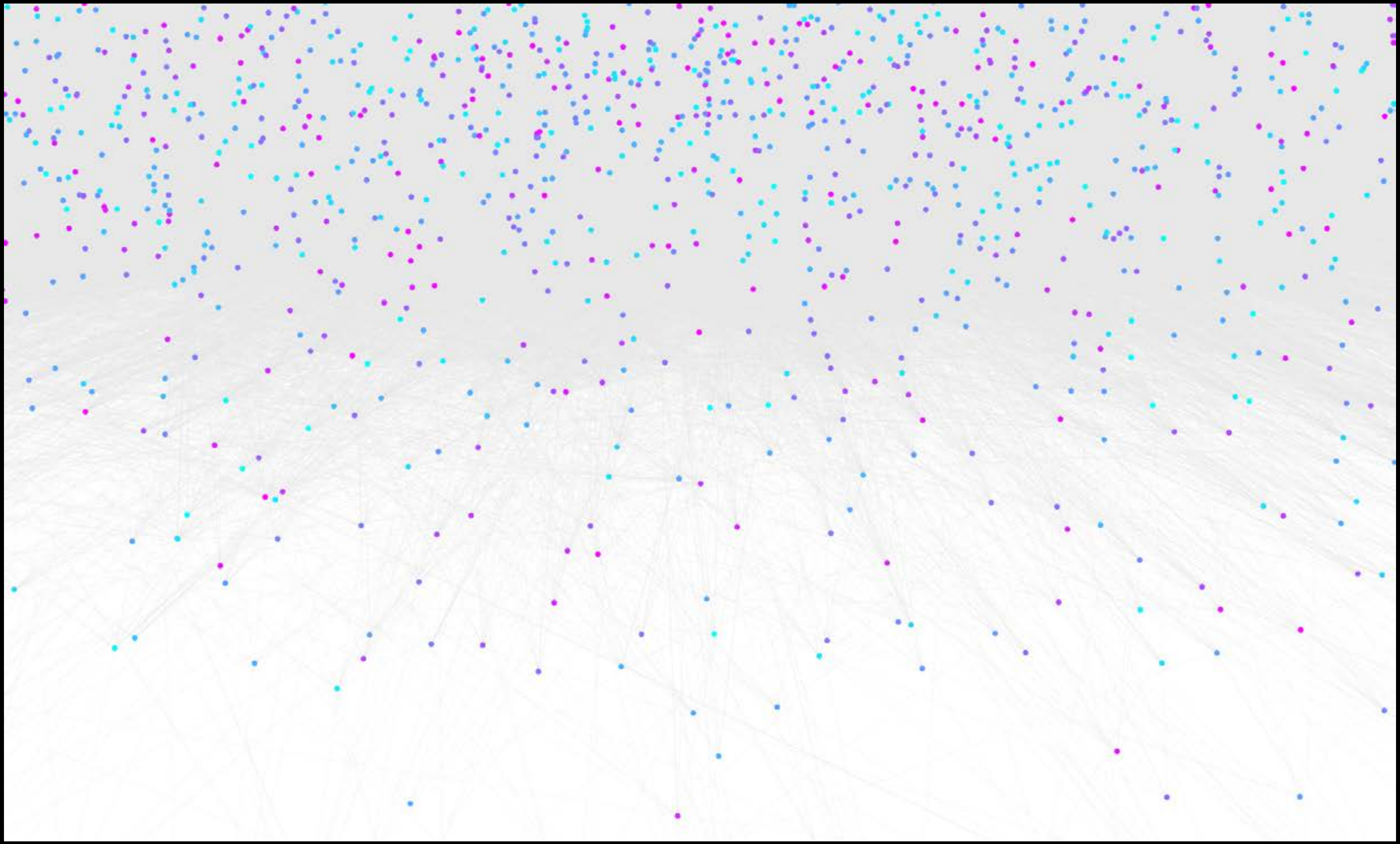


WYSS INSTITUTE

Xenopus GRN Visualized



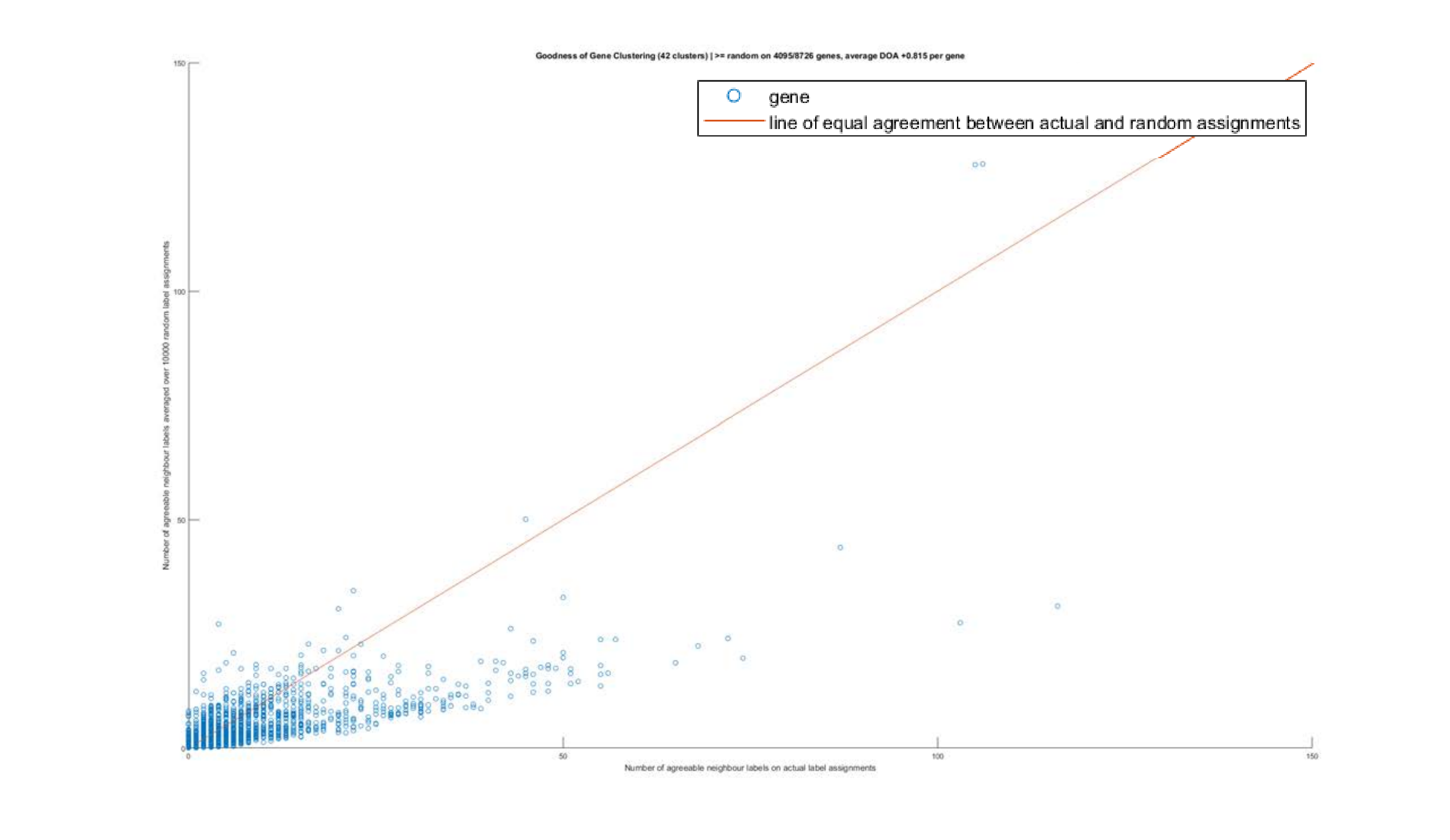
Xenopus GRN Visualized



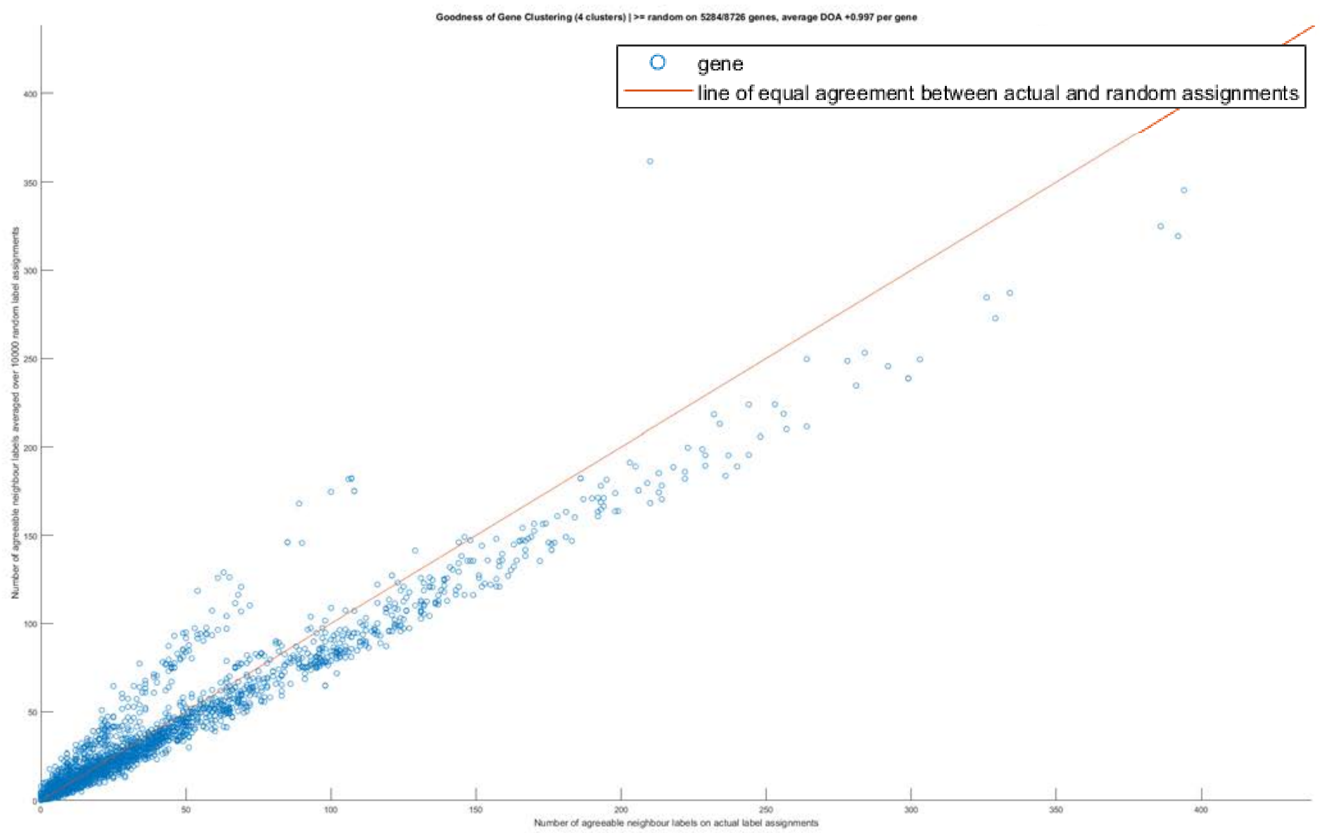
Clustering Goodness Metric

- = Number of agreements (NOA) in the neighborhood of a gene
- We should be better than chance (random)
- We plot the NOA for every gene in
  - (1) actual, versus
  - (2) mean of 10K random assignments
- Difference of Agreements: the more positive, the better

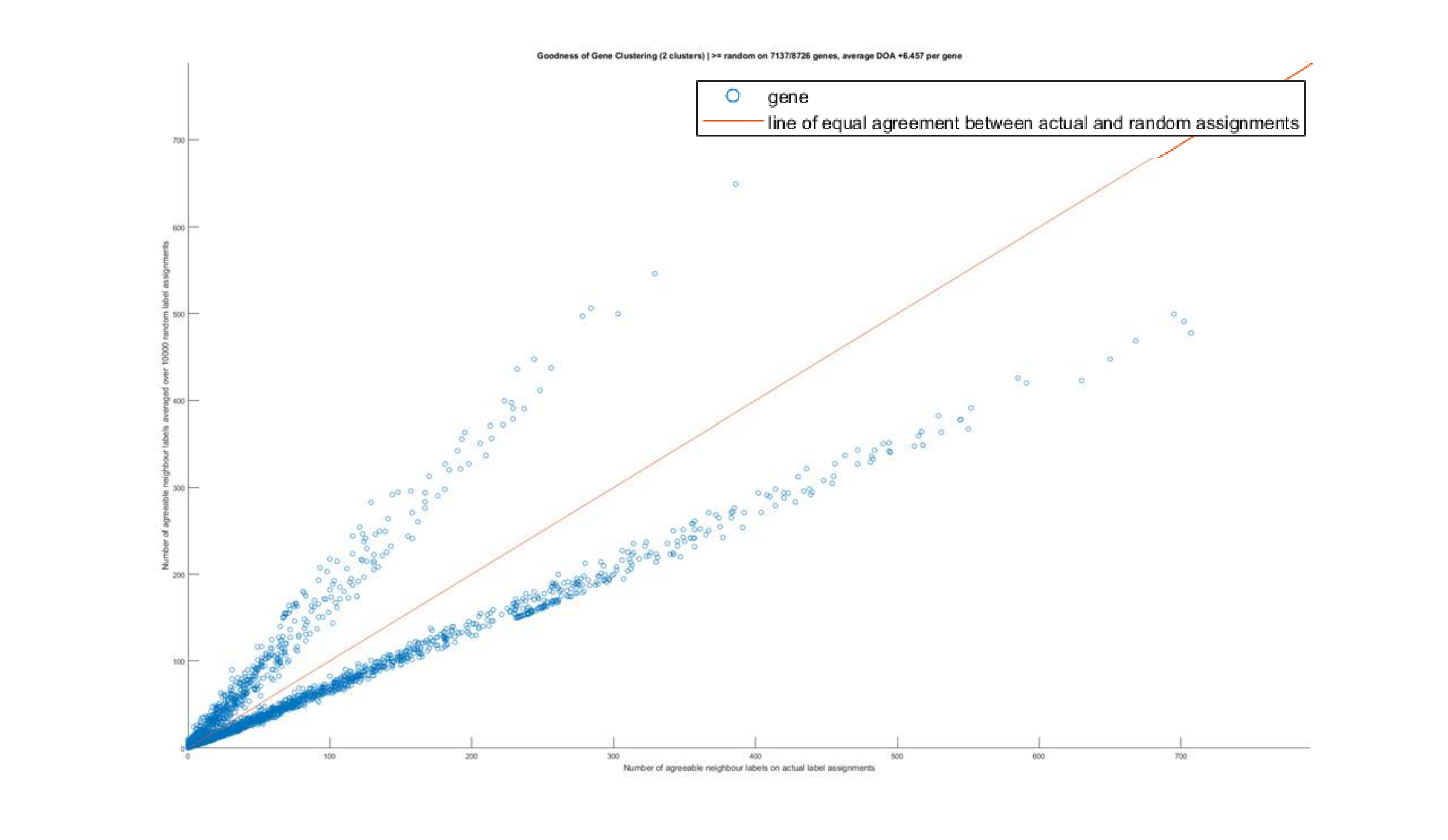
Clustering Goodness k=42 (DOA +0.815 per gene)



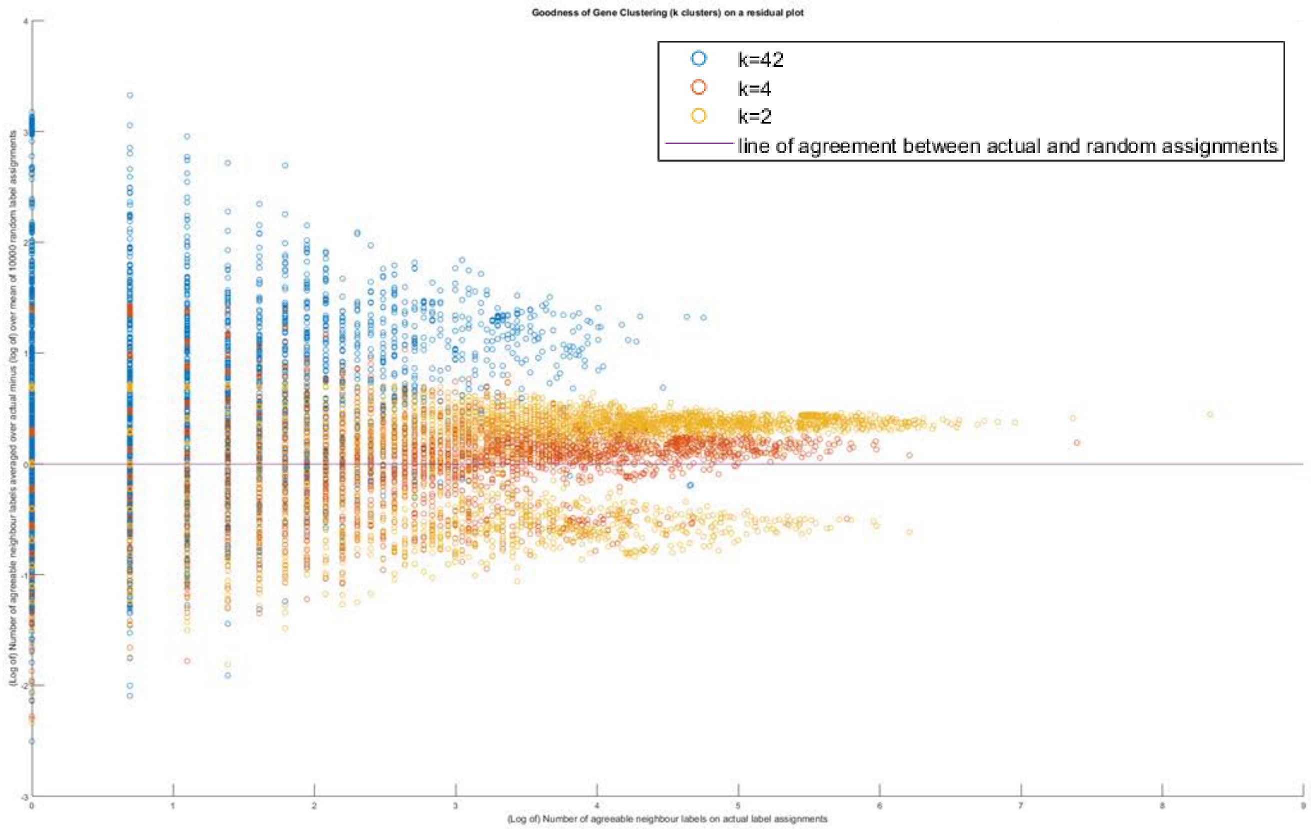
Clustering Goodness k=4 (DOA +0.997 per gene)



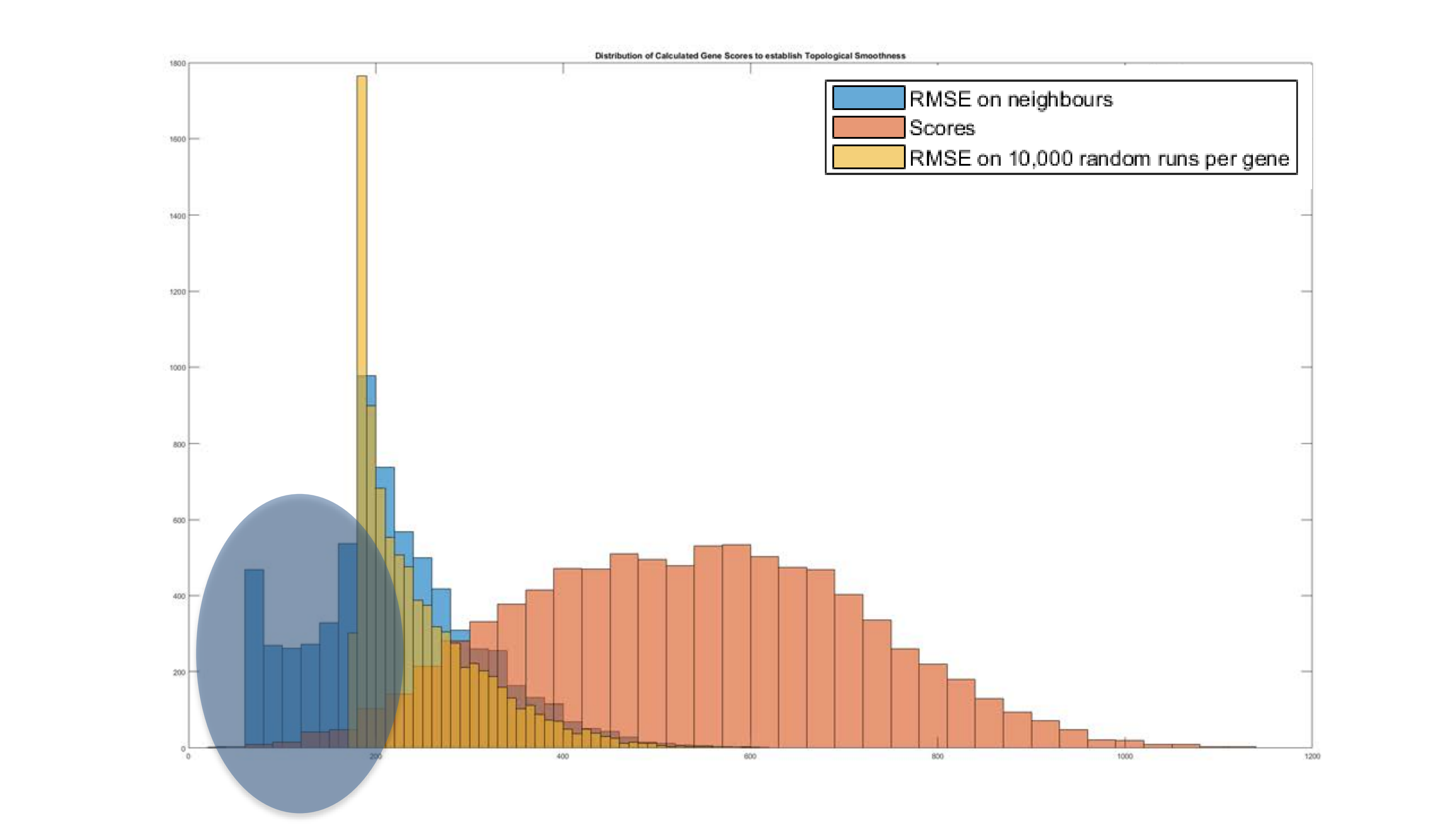
Clustering Goodness k=2 (DOA +6.457 per gene)



Clustering Goodness (Log-Log plot of DOA)

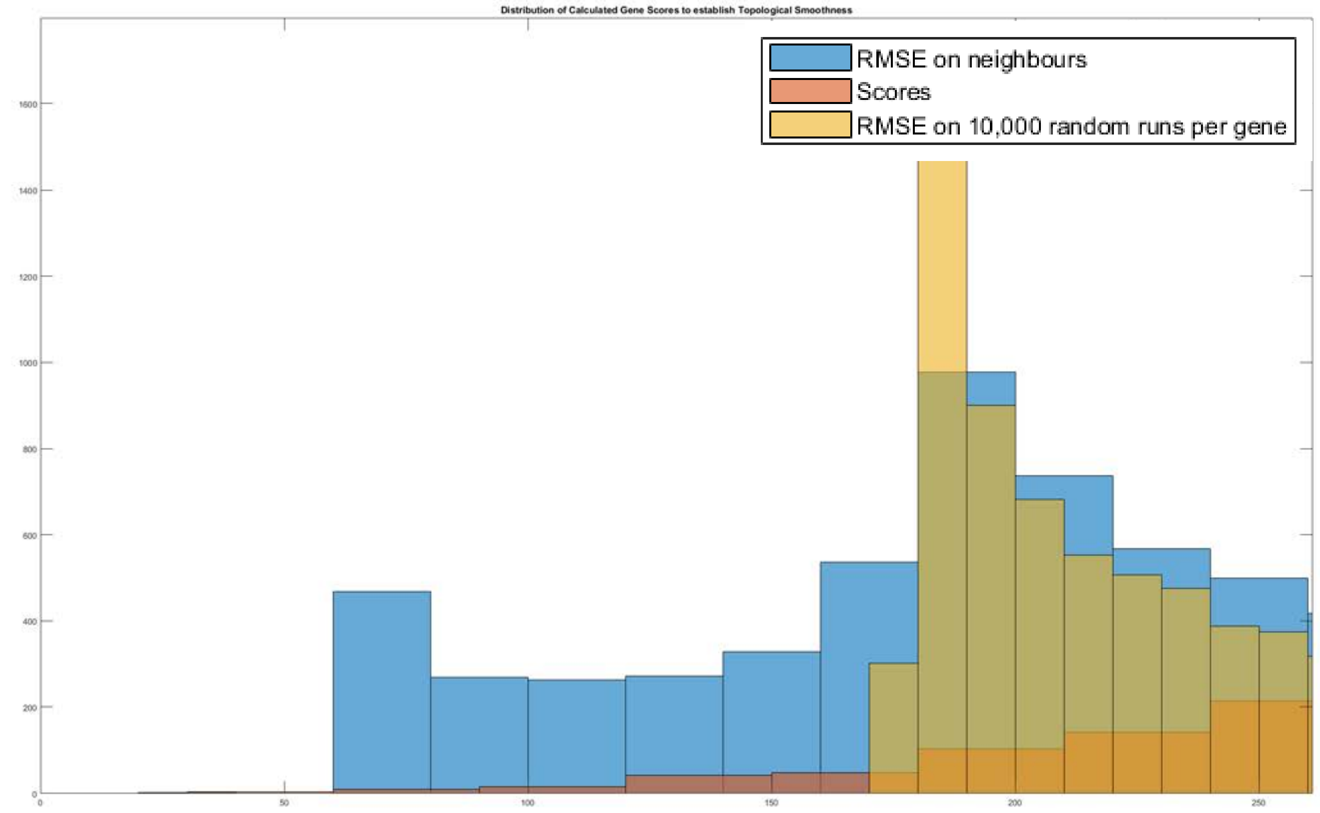


Are the Gene Scores also topologically smooth?



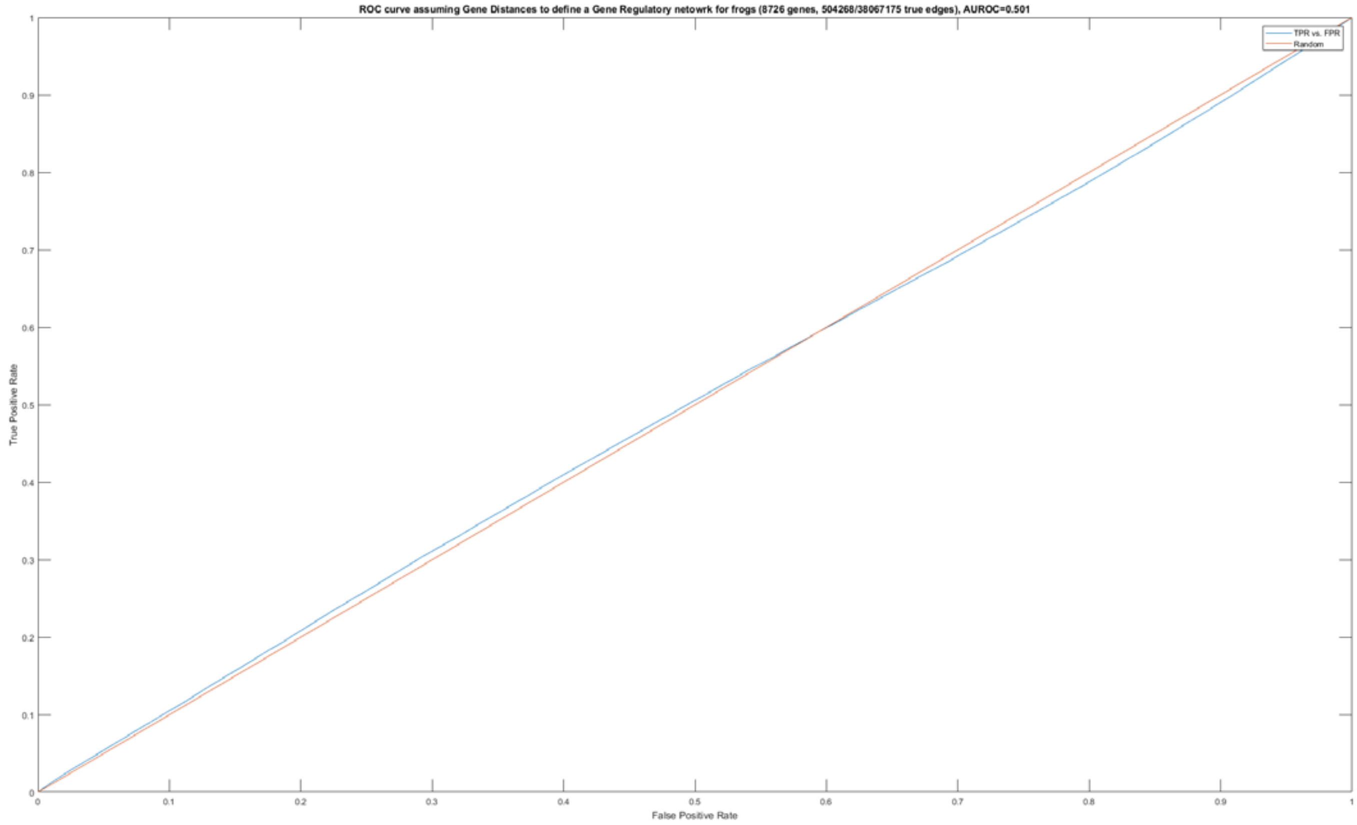
WYSS INSTITUTE

Are the Gene Scores also topologically smooth?

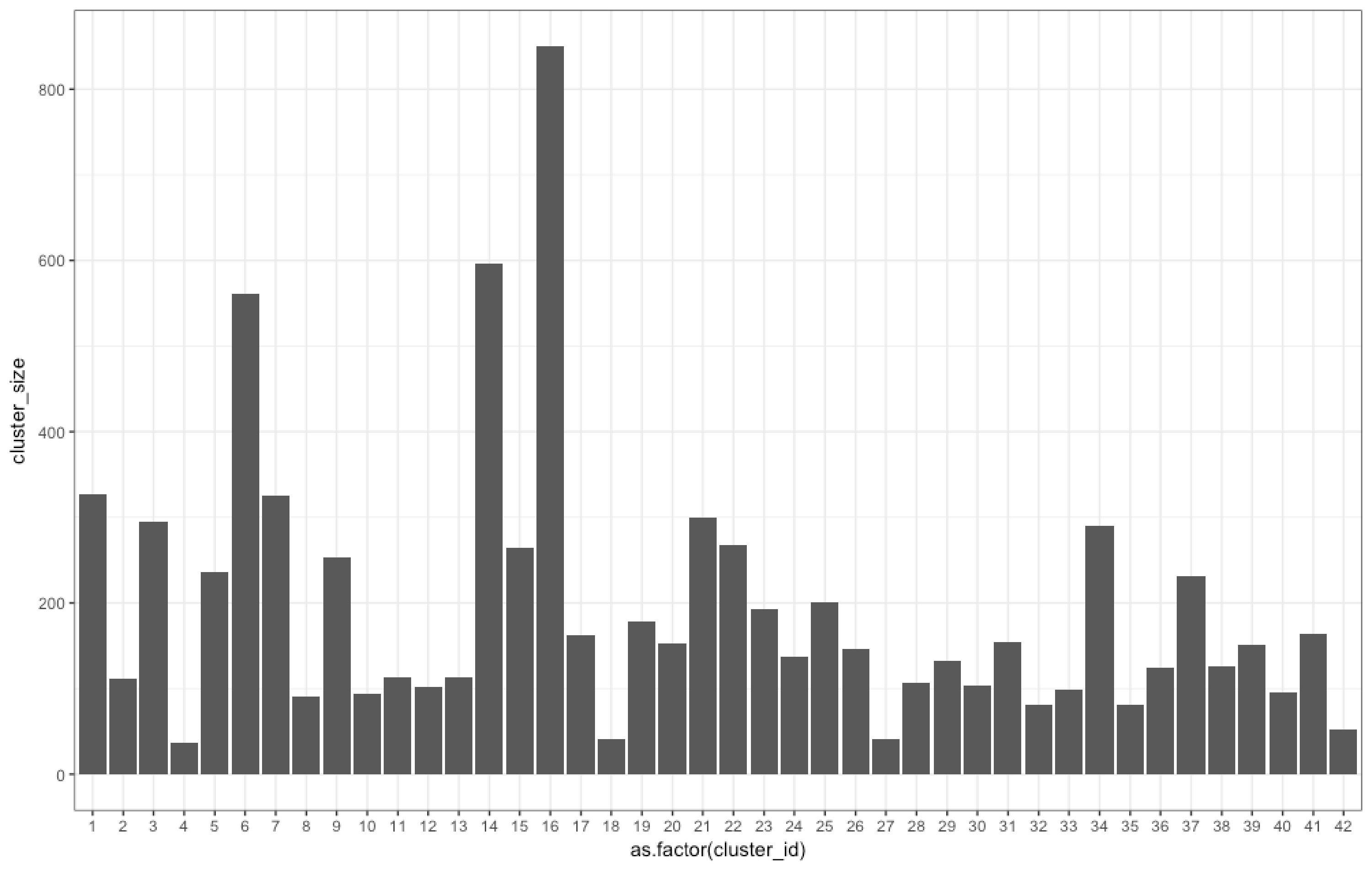


Are the Gene Scores also topologically smooth?

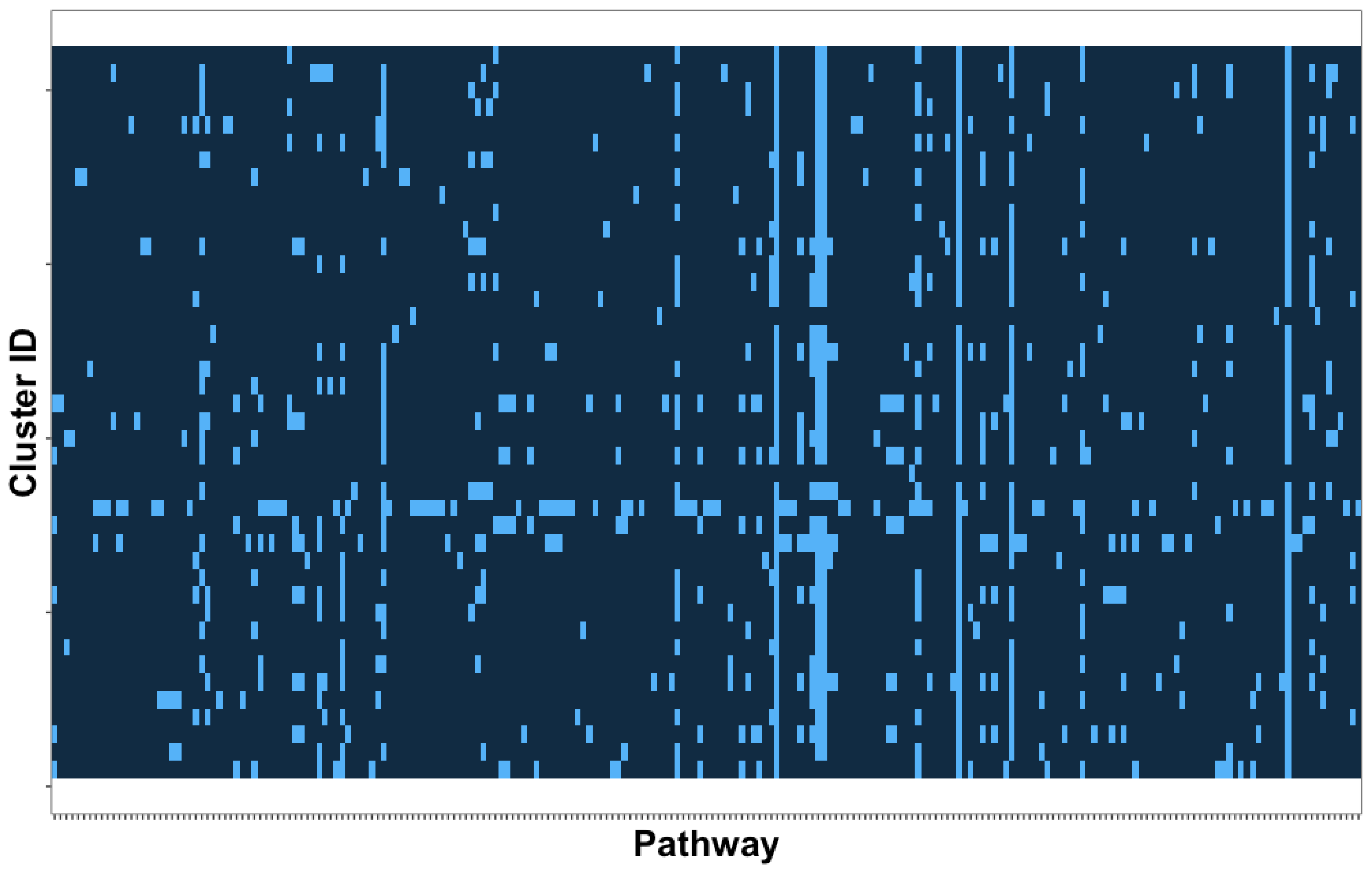
(GRN has 504268 true edges out of 38067175; AUROC=0.501)



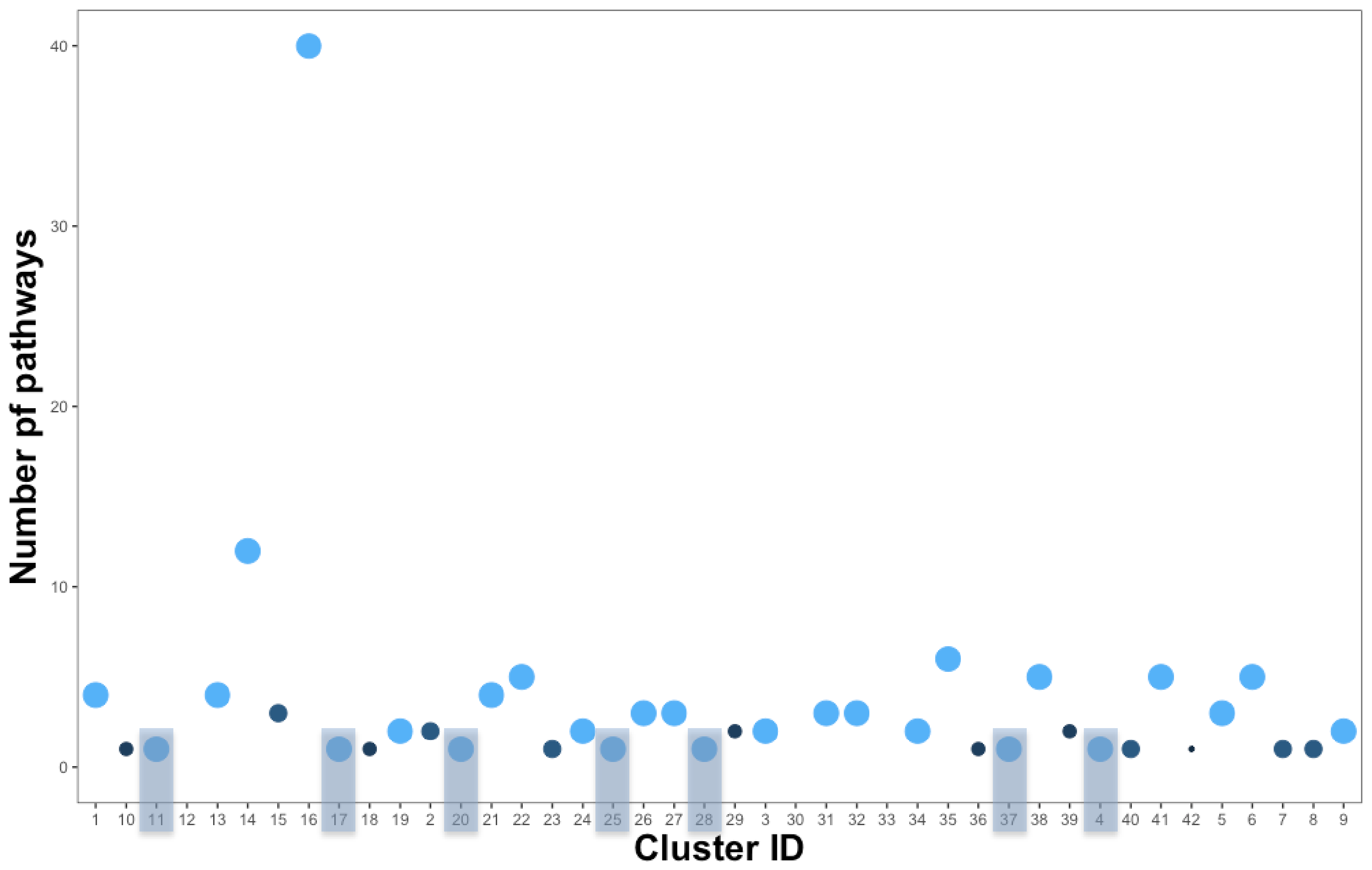
Mapping Gene Groups to Pathways



Mapping Gene Groups to Pathways



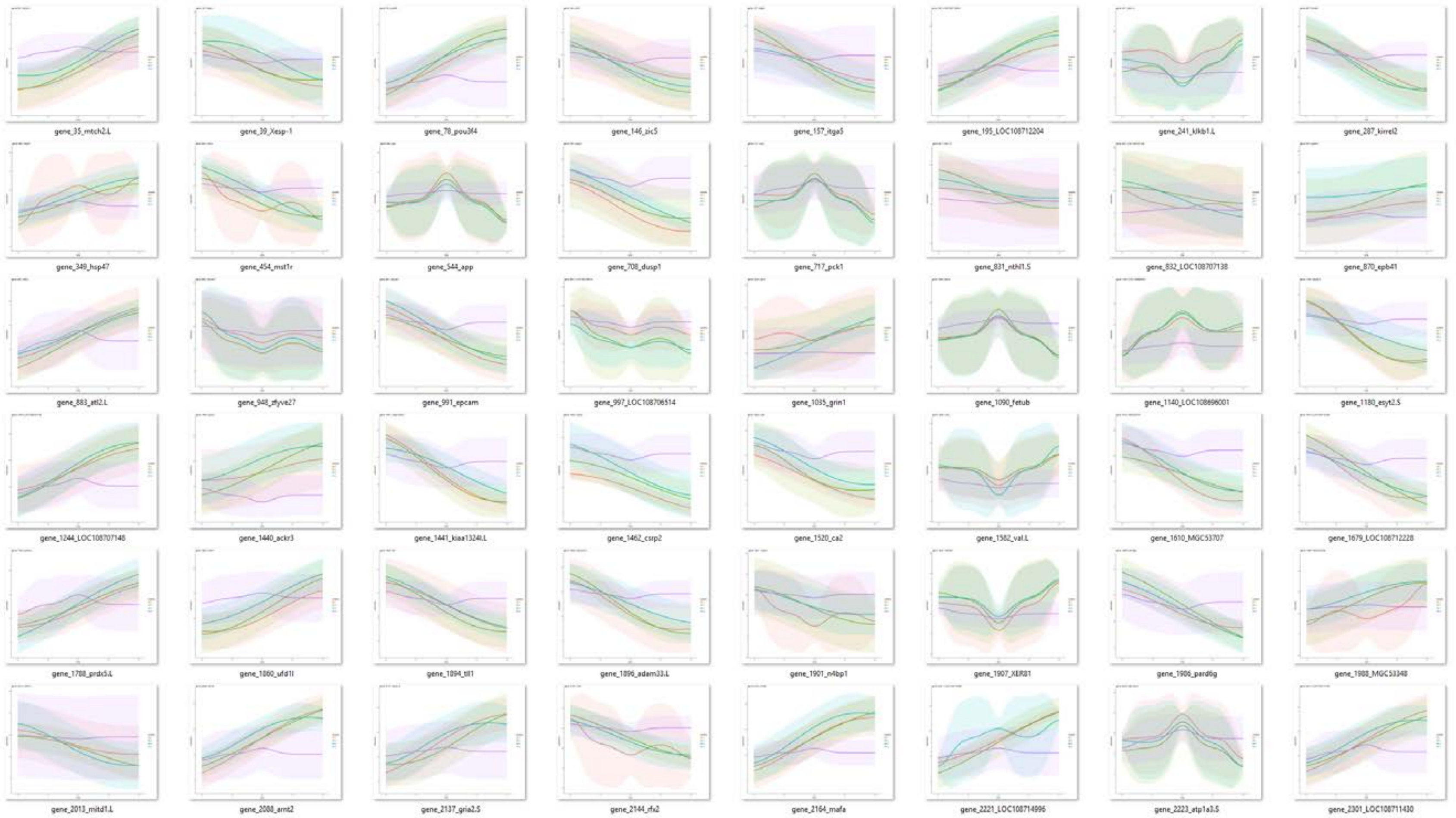
Mapping Gene Groups to Pathways



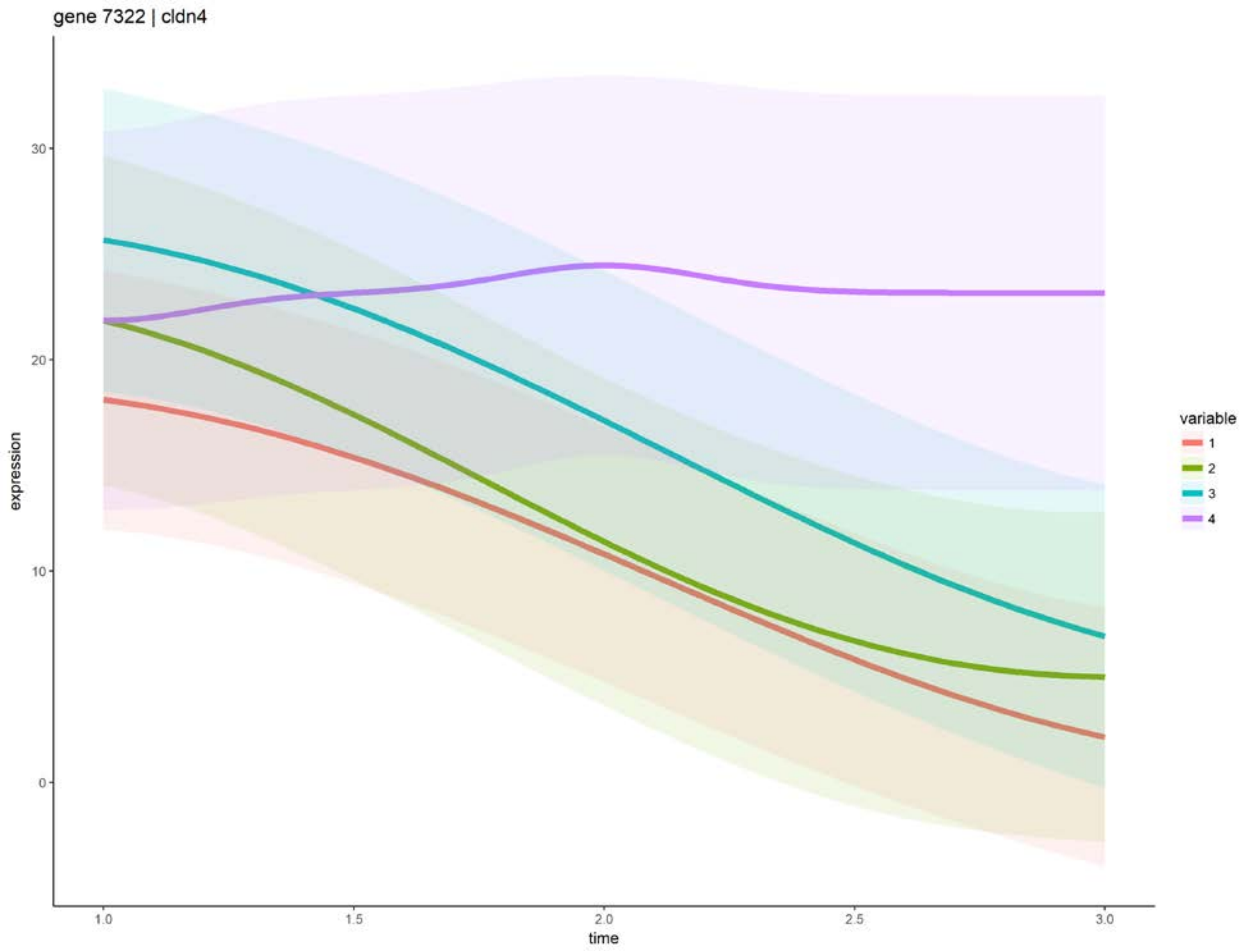
Mapping Gene Groups to Pathways

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**

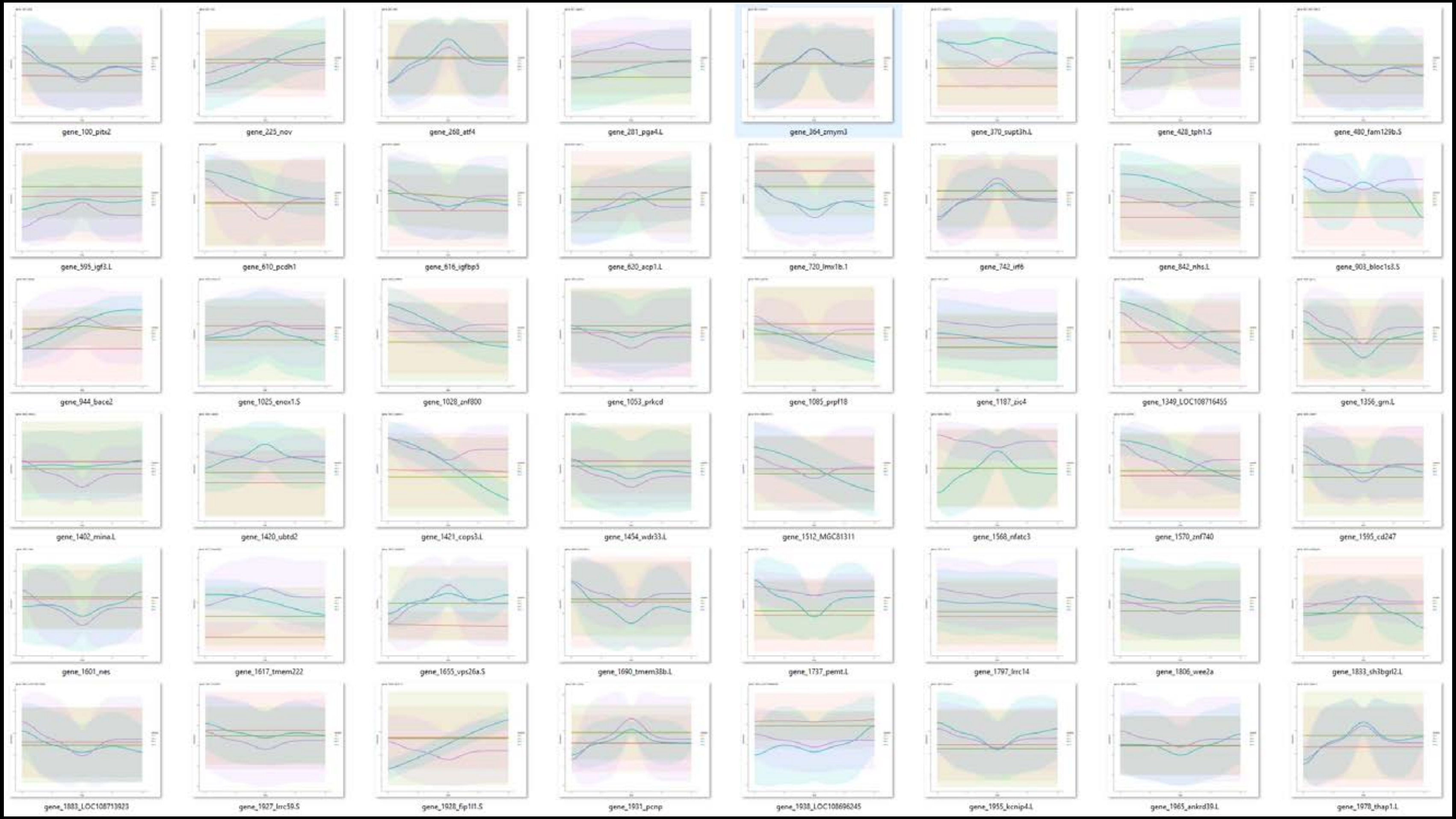
Gene Group 20 (abacavir)



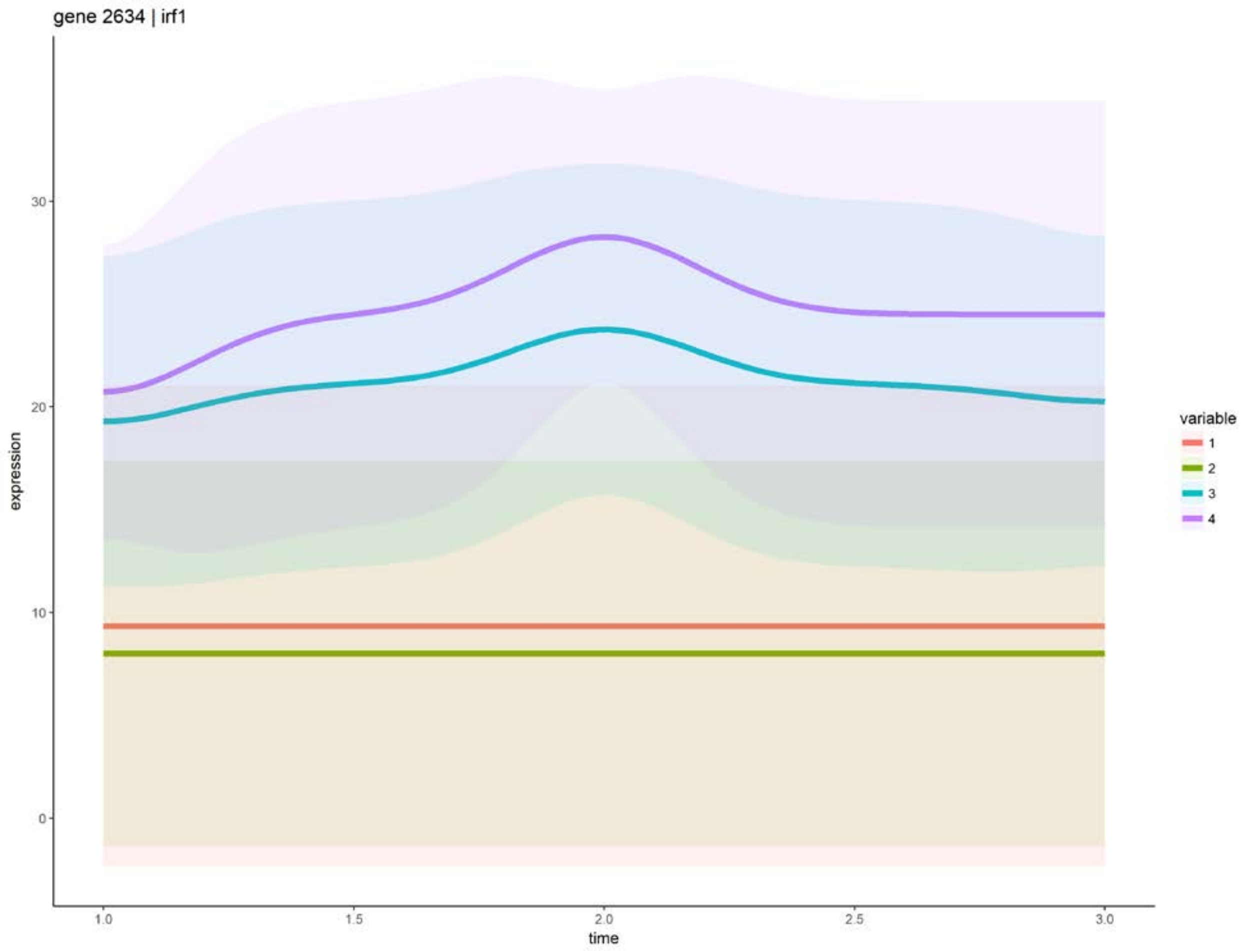
Example Gene from Group 20



Gene Group 37 (heme)



Example Gene from Group 37



GPR for Omics, Method 2

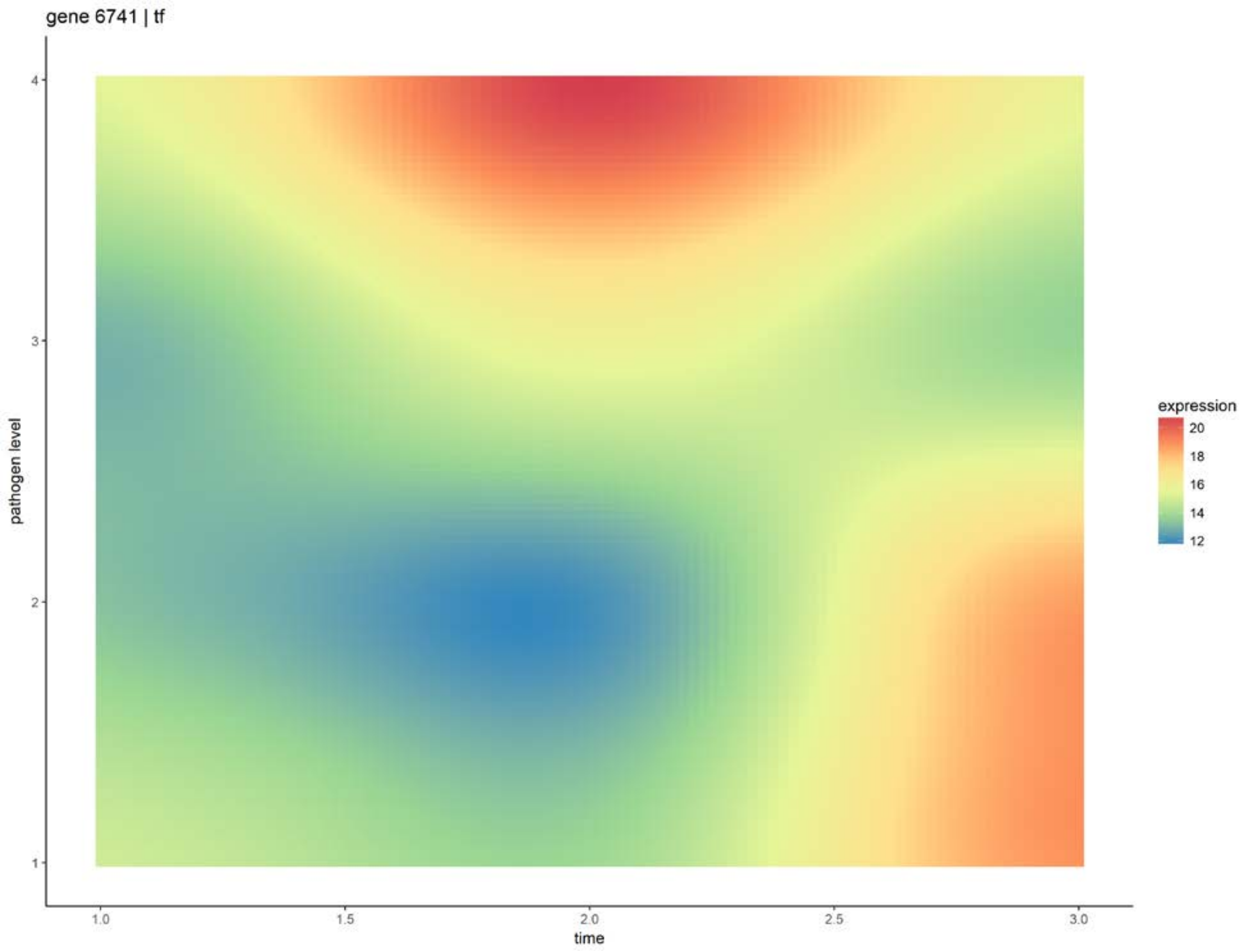
- Treat every gene as an independent GP mapping from time space to expression space

$$f: t \to g$$

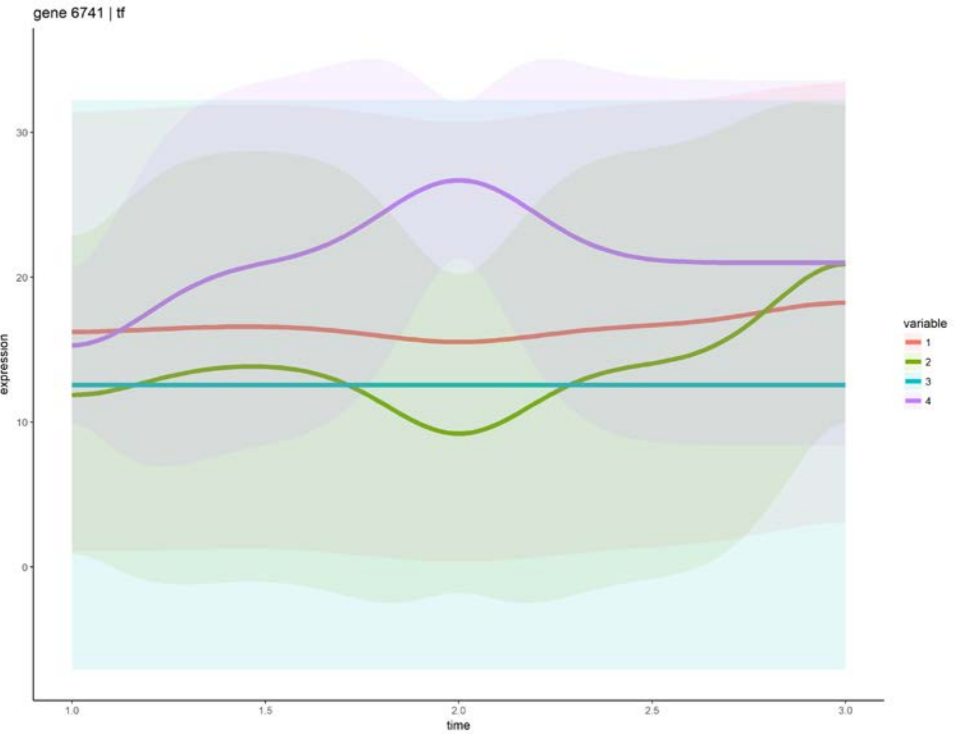
$$f(t') \sim GP\big(m(t), k(t, t')\big) + \beta$$

- Every gene has only 1 GP model to be learnt
- Allows regression over both time and pathogen levels, together

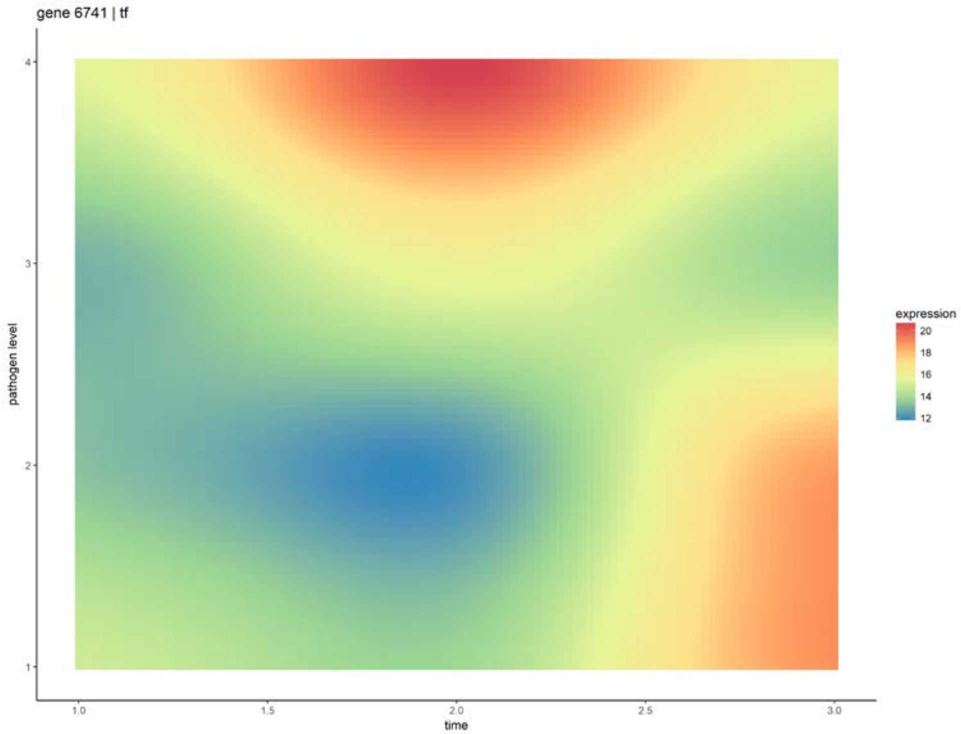
Example: Transferrin (probe1)



Example: Transferrin (probe1)



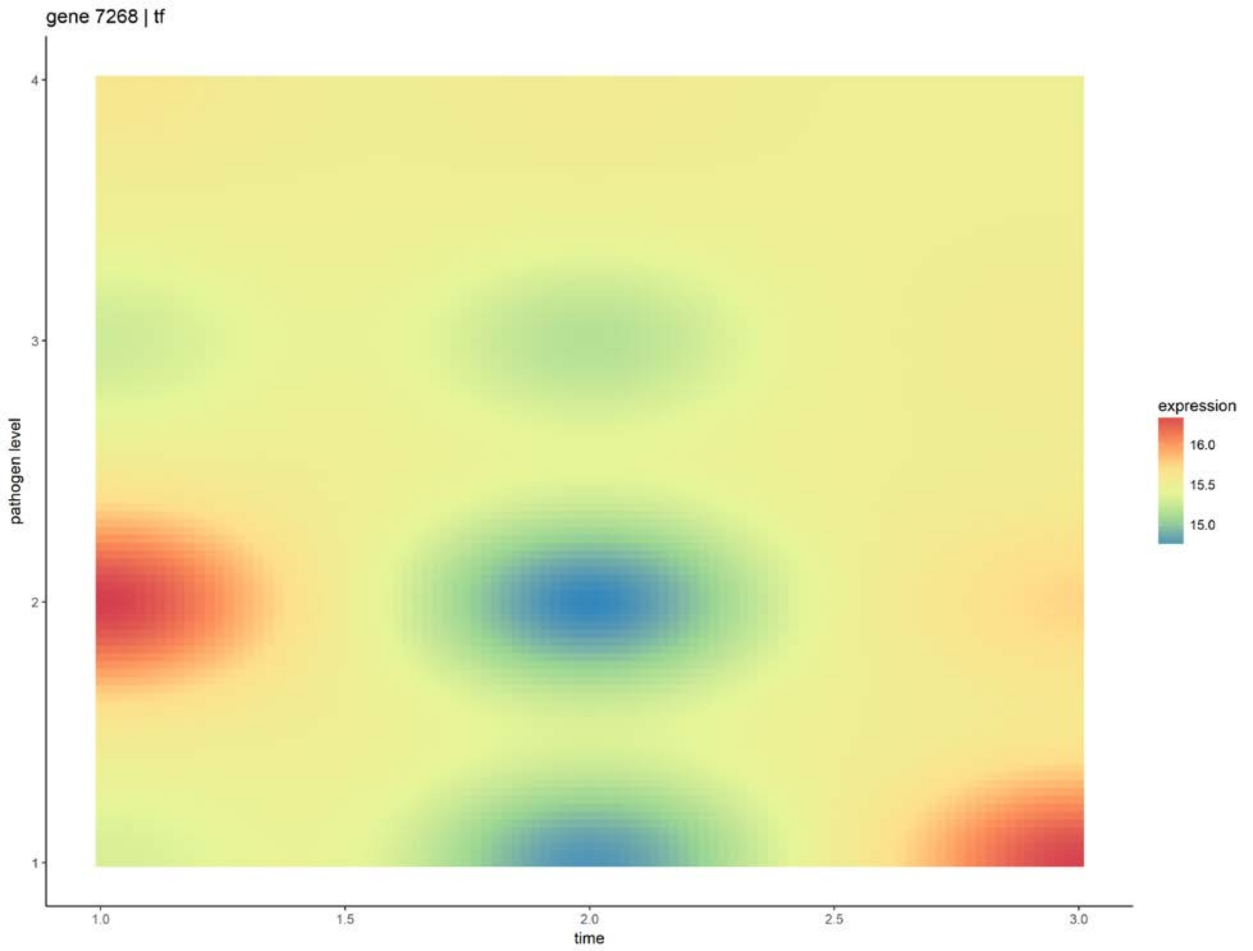
Univariate method



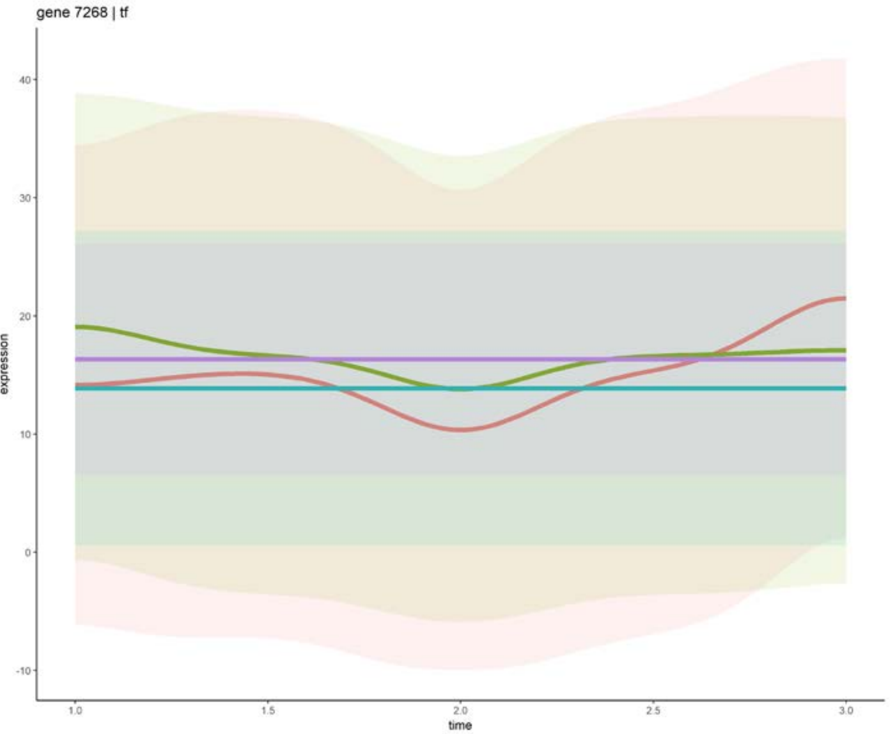
Bivariate method

WYSS INSTITUTE

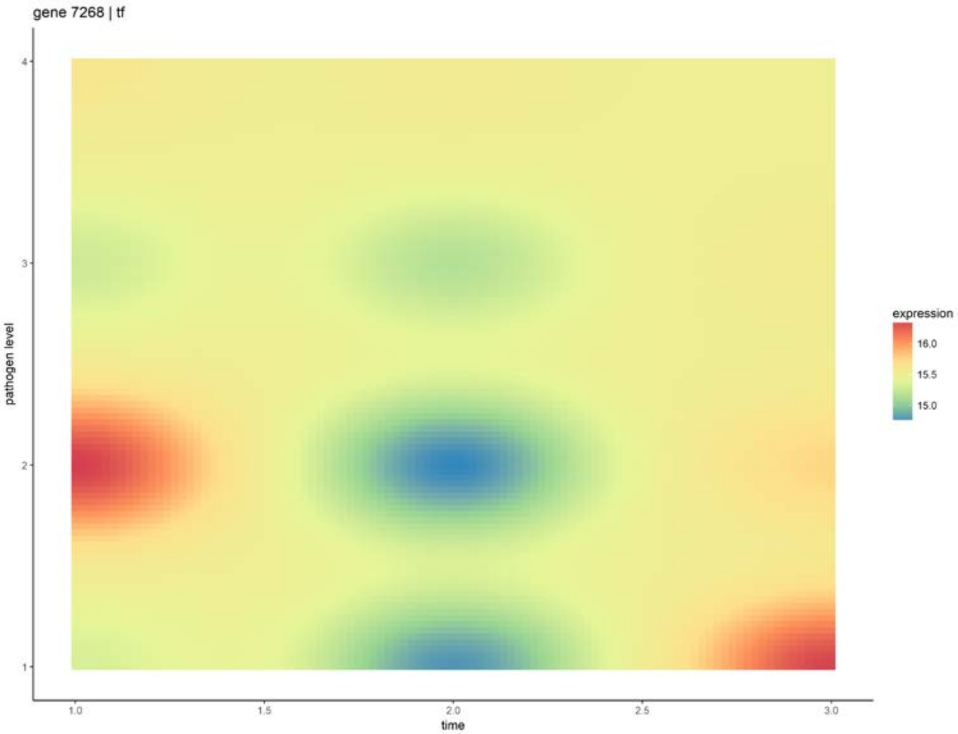
Example: Transferrin (probe2)



Example: Transferrin (probe2)

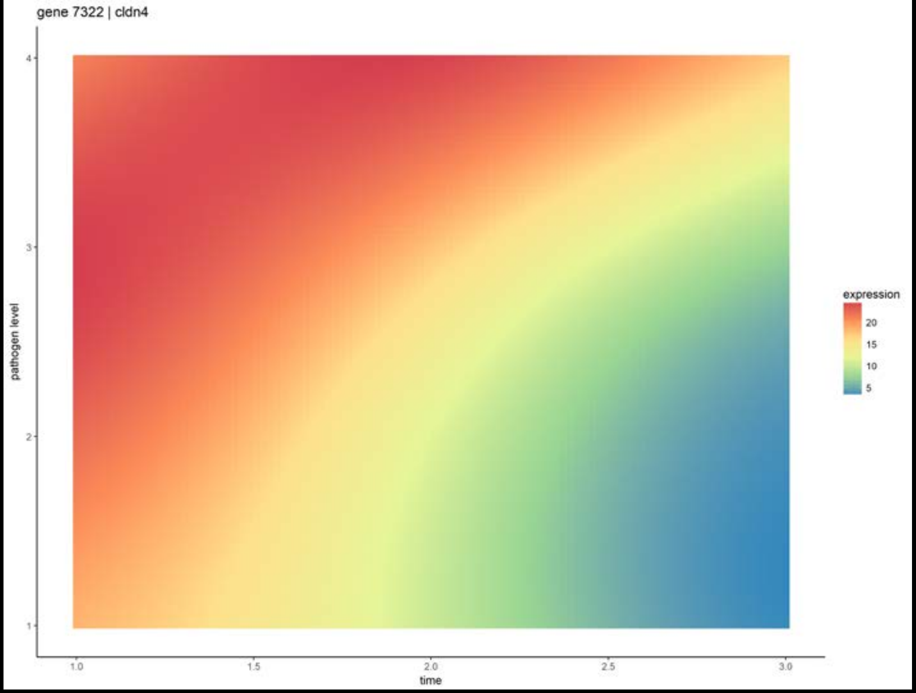


Univariate method

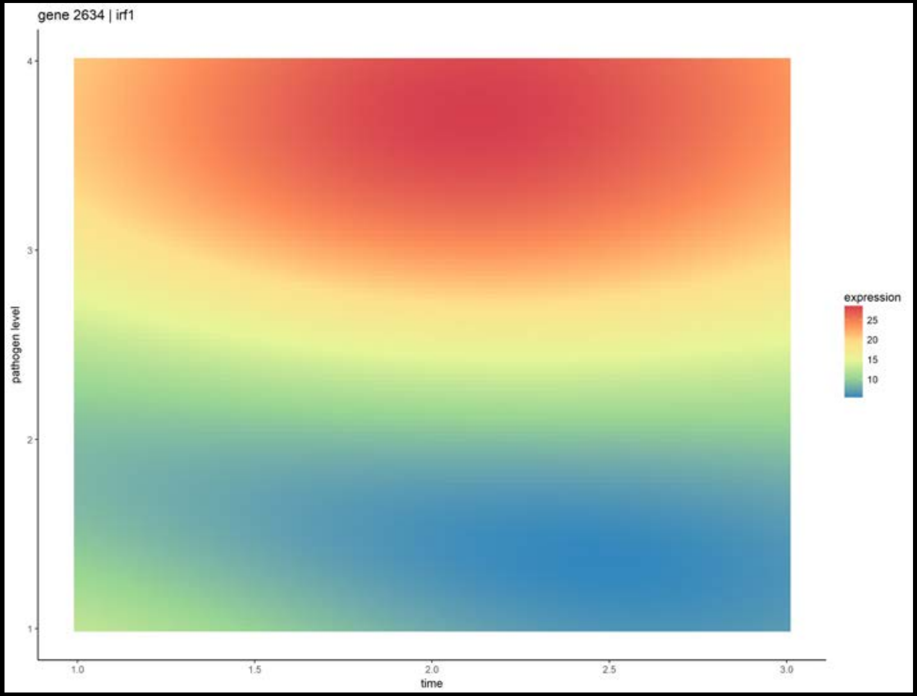


Bivariate method

Examples: from abacavir and heme Gene Groups

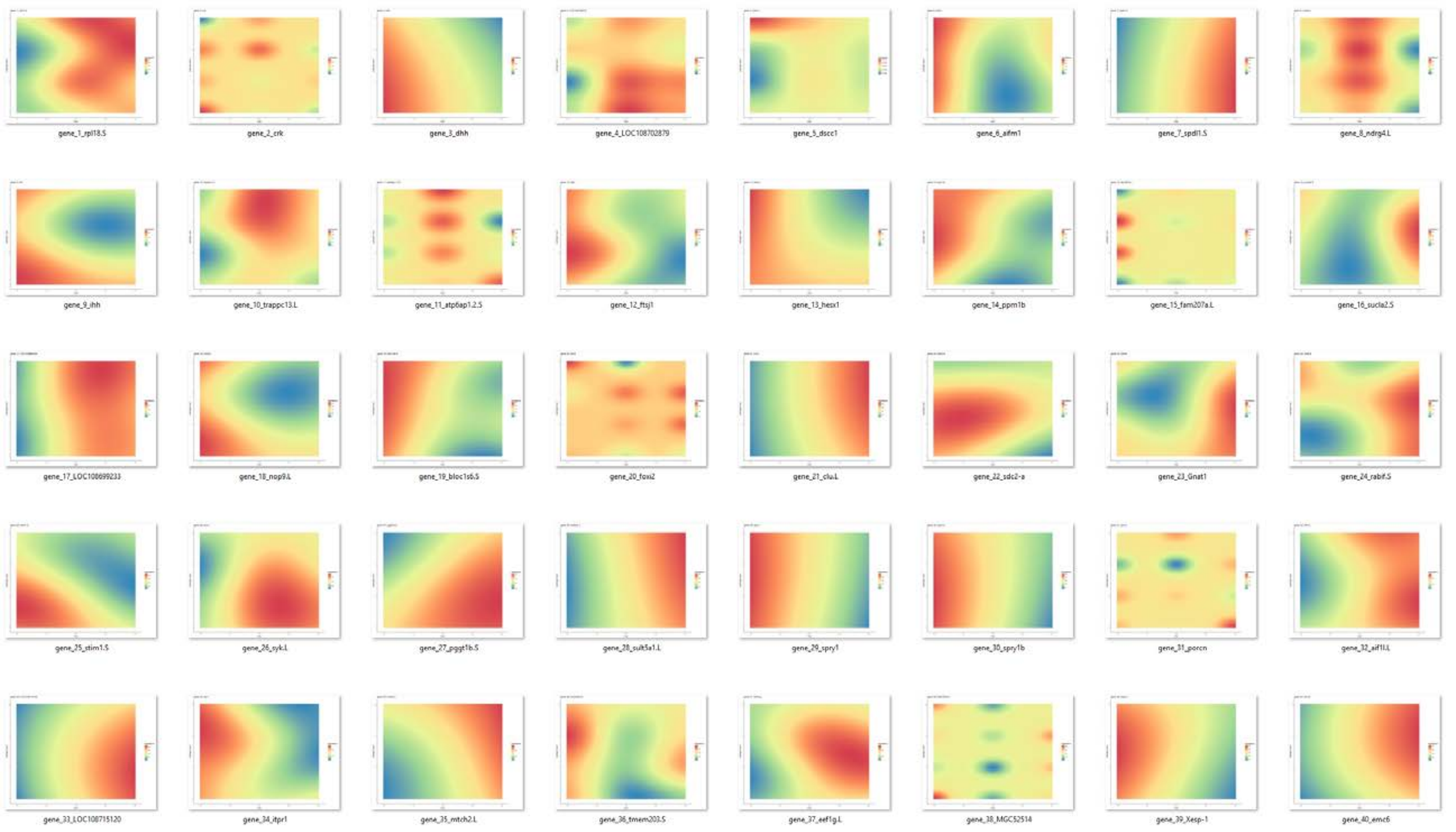


abacavir gene group



heme gene group

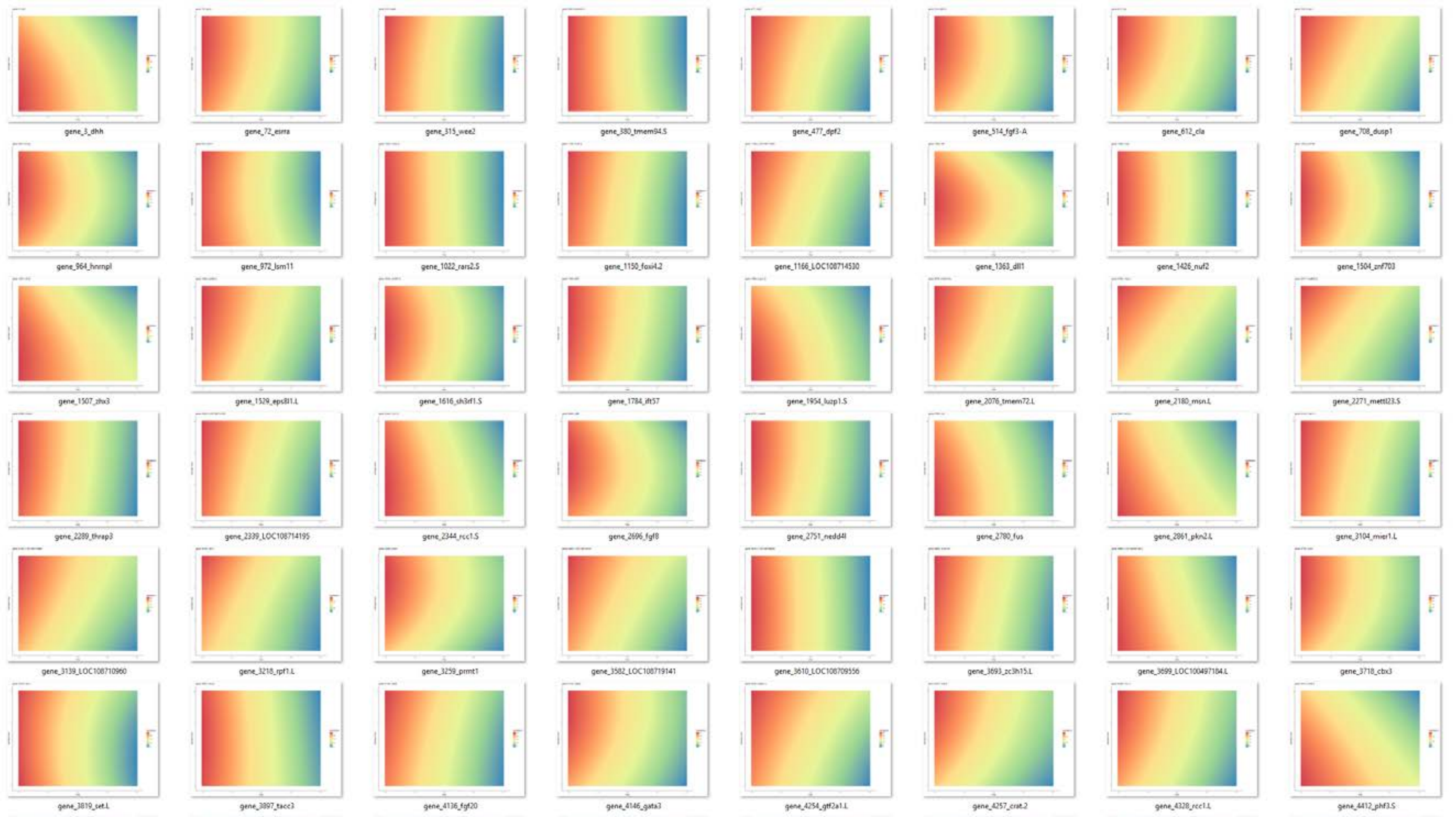
For all 8726 genes…



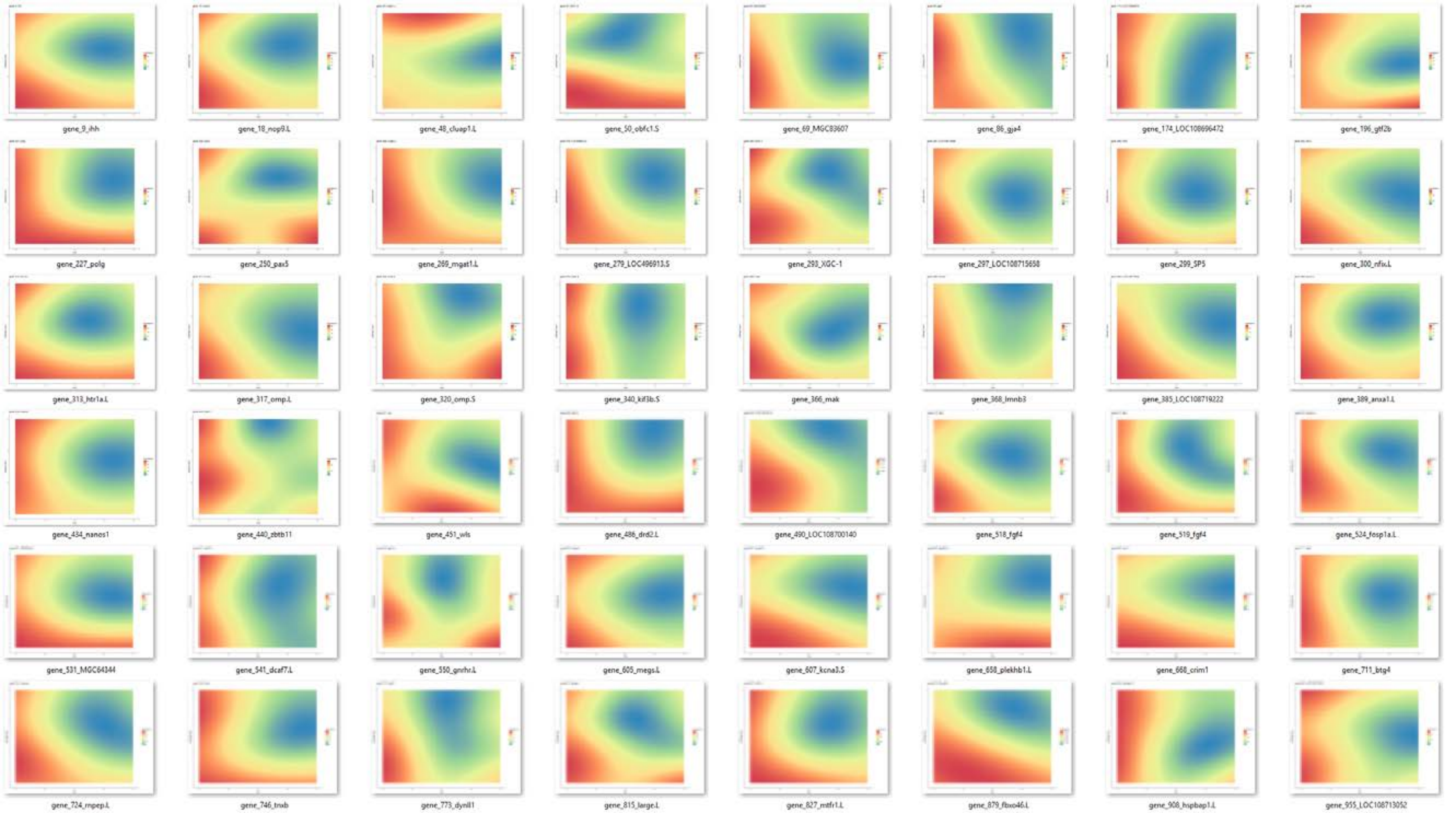
To extract key genes and gene groups…

- Visually!
- Question: How do we compare two GP models?
  - Can compare in model space itself because we have same inputs to all GP models!
- Solution: we use the mean function and covariance function of learnt GP model to define a “feature space” of genes
  - We conveniently manage to ignore scale and bias in comparing genes

Examples: Gene Group

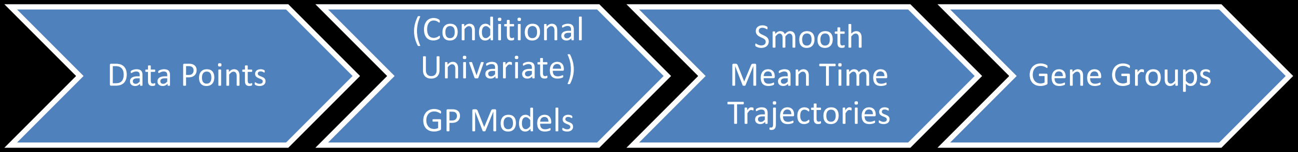


Examples: Gene Group



Summary

- Model(s) for time-series data under multiple conditions, with limited number of samples



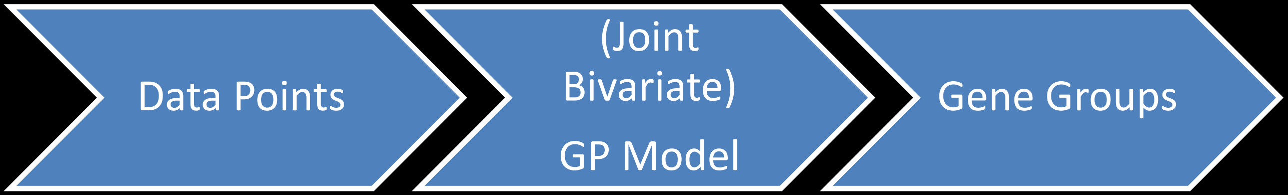
Data Points

(Conditional Univariate)

GP Models

Smooth Mean Time Trajectories

Gene Groups



Data Points

(Joint Bivariate)

GP Model

Gene Groups

Summary

- Model(s) for time-series data under multiple conditions, with limited number of samples
- Good tool for quick visual inspections
- Outputs key genes and gene groups to look at, mapped to appropriate pathways
- Generalized to any temporal dataset with multiple conditions where “smooth” assumption can be applied
- Next steps:
  - Use actual CFU counts of every sample in the joint bivariate regression model
  - Build a simple browser-based app for everyone to use