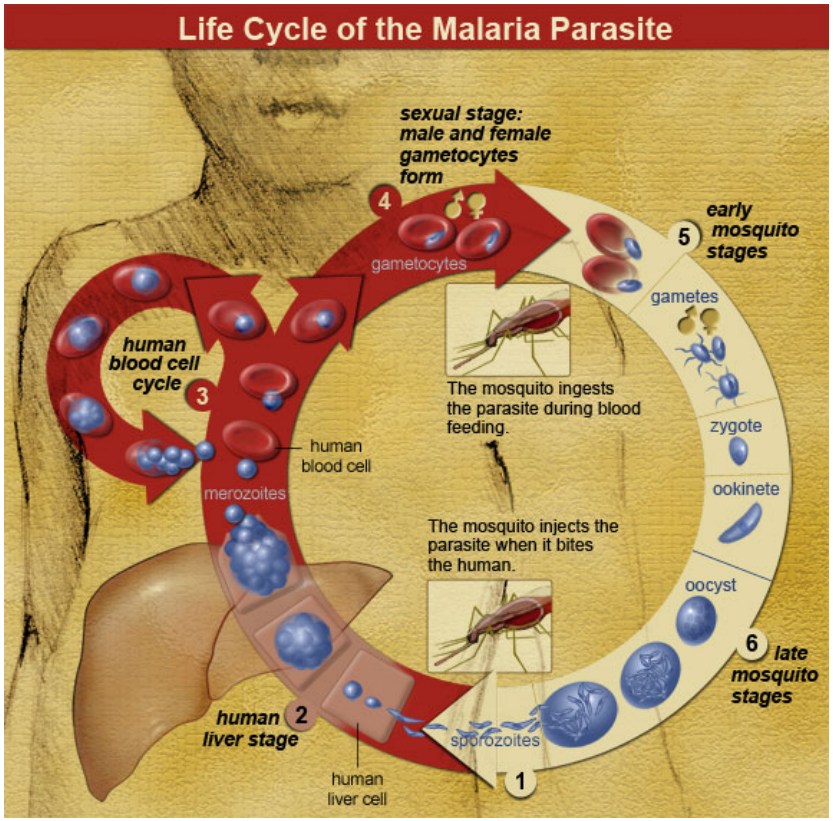
Transfer Learning of Genome Wide Transcription Dynamics during Malaria Infection

Venelin Mitov
ETH Zürich

October 2, 2013
Thesis presentation

Outline

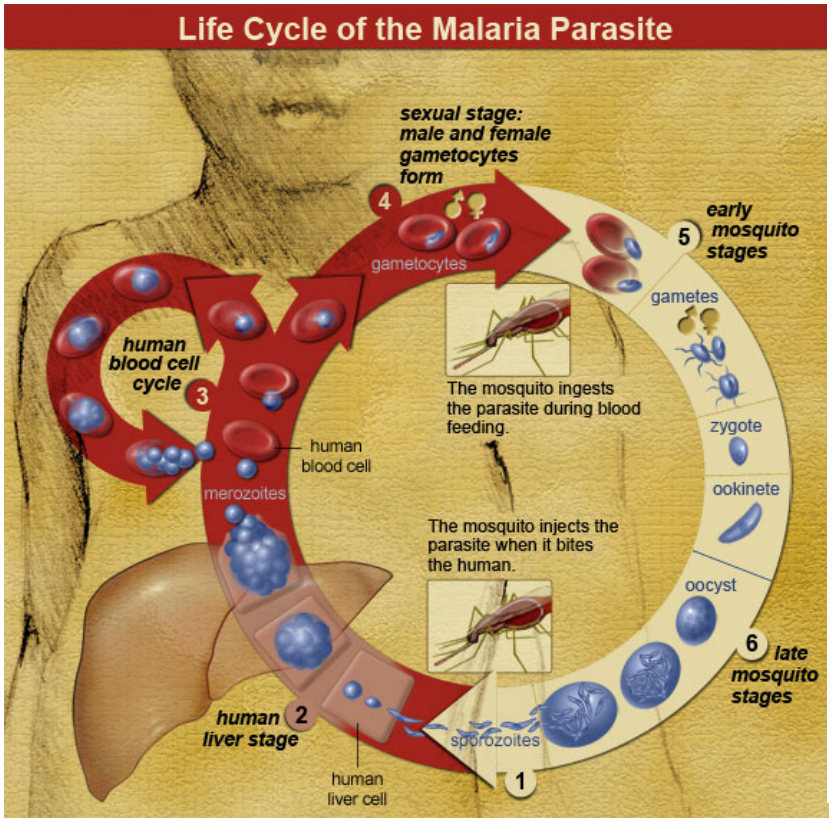
A Transfer Learning Approach

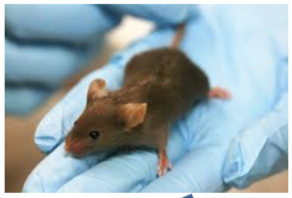


Courtesy: National Institute of Allergy and

Infectious Diseases

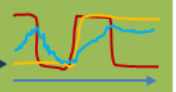
A Transfer Learning Approach

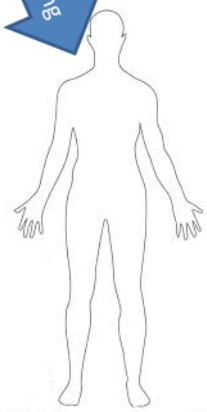




Transfer learning

1. Find genes in infected mice, which have informative time-course dynamics for the inference of post-infection time in mice





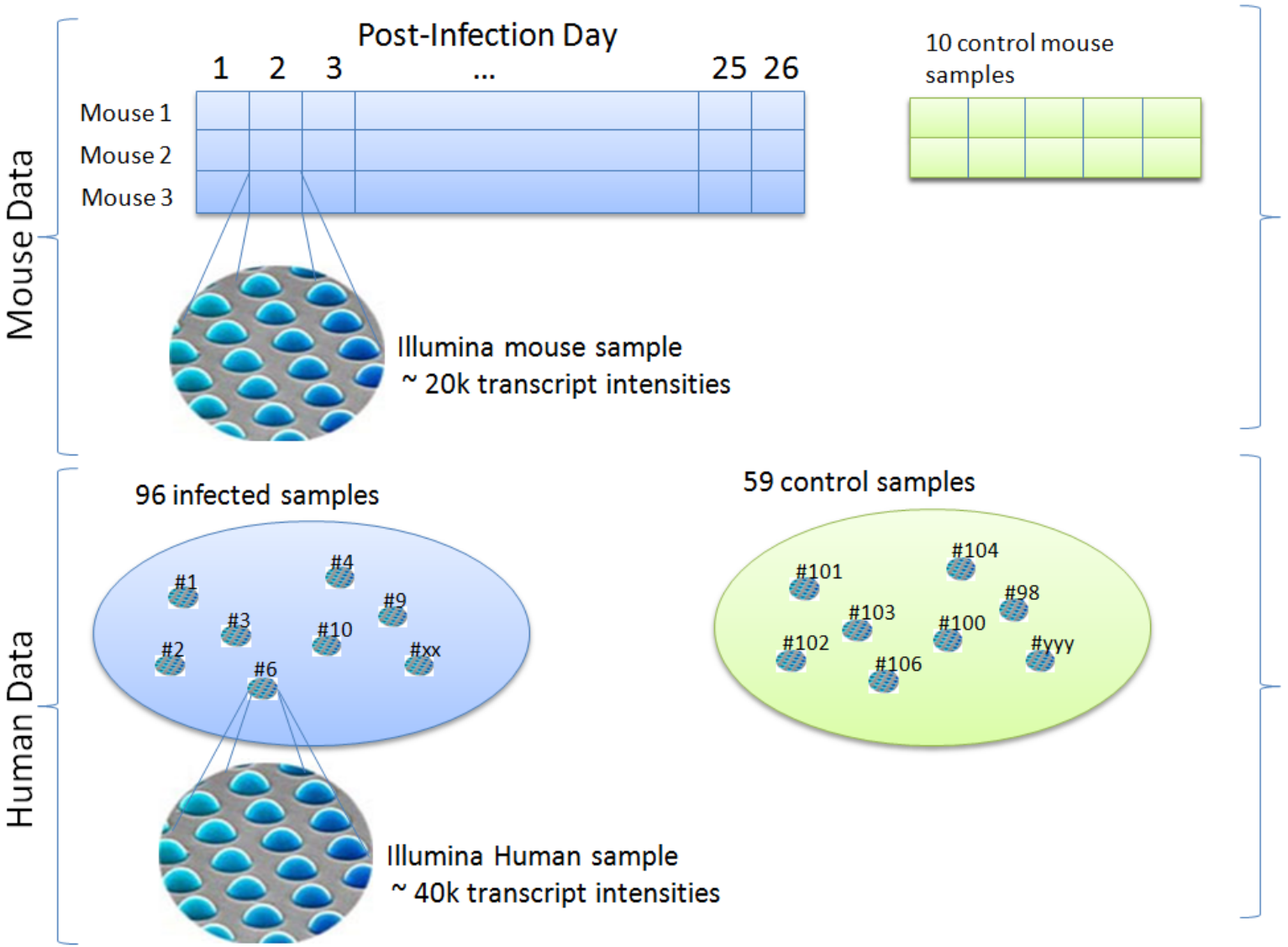
2. Map these genes to their human homologs and narrow down this set to genes that are relevant for malaria progression in the human context

M. musculus ⟷ H. sapiens

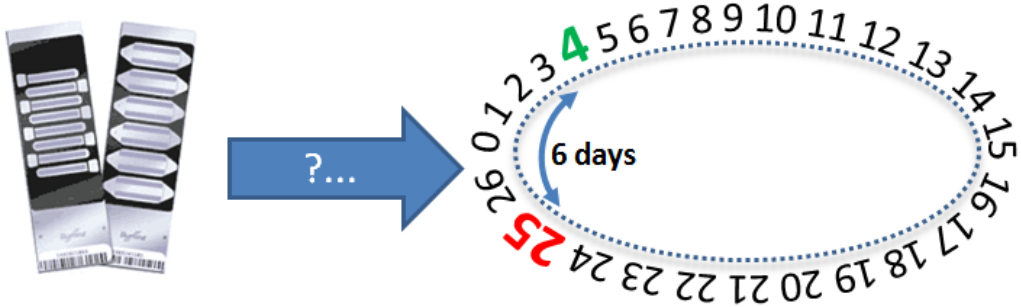
EIF6
PDCD2
ACHE
...

Courtesy: National Institute of Allergy and Infectious Diseases

Murine and Human GWAS Data



Approaches for Post-Infection Time Inference



Single-gene peaks

Such peaks need to be narrow and unique in time:

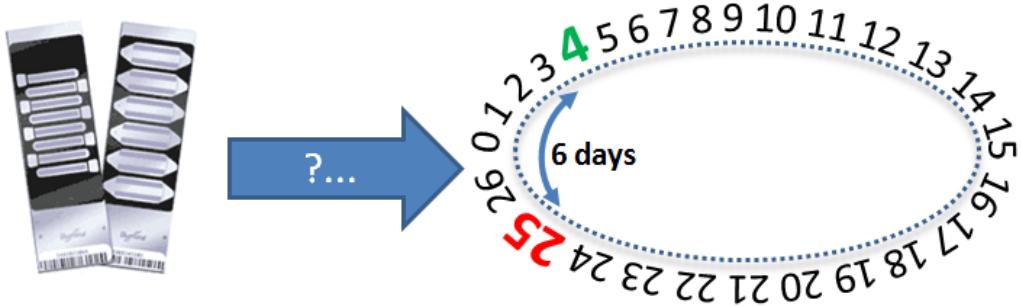
- Do such gene-markers exist for each day?
- Can narrow peaks be measured in all mice?

Multi-gene expression patterns

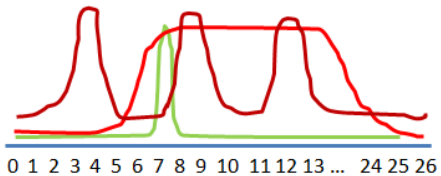
- more likely to be unique in time
- Possibly not narrow enough for 1-day precision
- Manually intractable

=> Supervised pattern recognition

Approaches for Post-Infection Time Inference



Single-gene peaks



Such peaks need to be narrow and unique in time:

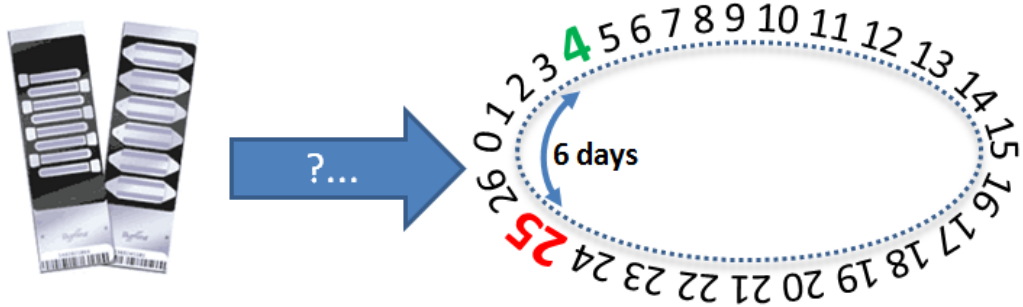
- Do such gene-markers exist for each day?
- Can narrow peaks be measured in all mice?

Multi-gene expression patterns

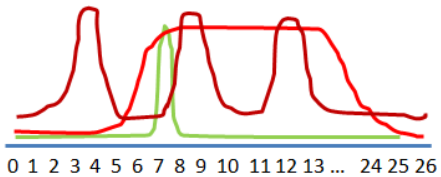
- more likely to be unique in time
- Possibly not narrow enough for 1-day precision
- Manually intractable

=> Supervised pattern recognition

Approaches for Post-Infection Time Inference



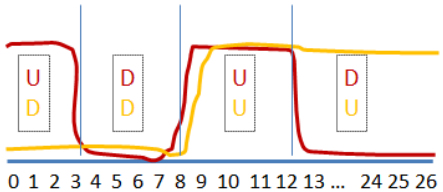
Single-gene peaks



Such peaks need to be narrow and unique in time:

- Do such gene-markers exist for each day?
- Can narrow peaks be measured in all mice?

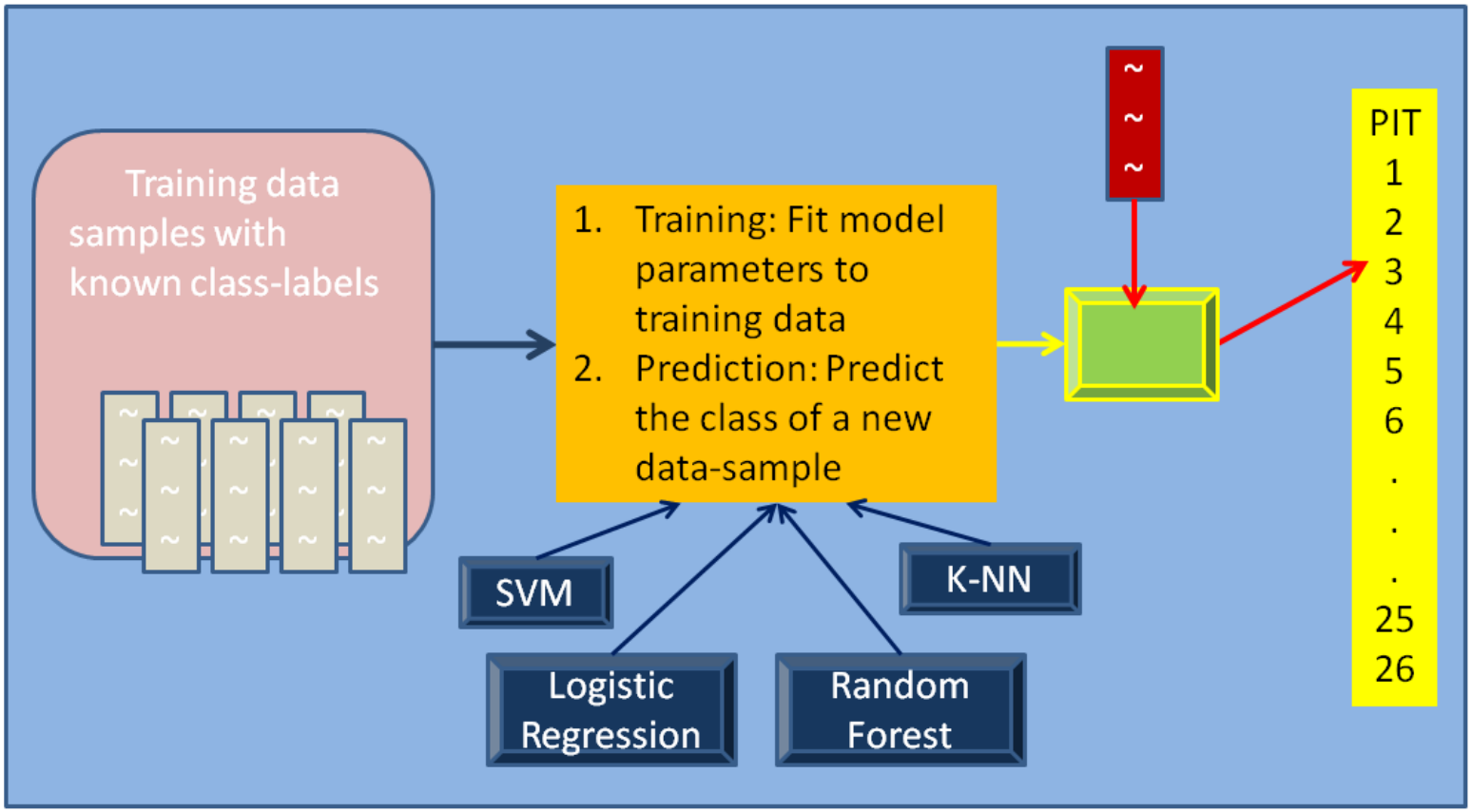
Multi-gene expression patterns



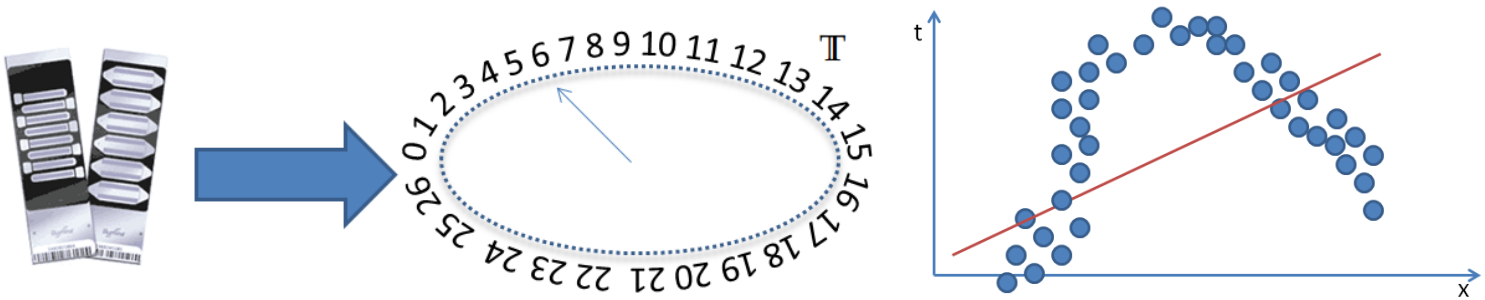
- more likely to be unique in time
- Possibly not narrow enough for 1-day precision
- Manually intractable

=> Supervised pattern recognition

Supervised Pattern Recognition



Linear Regression Formulation



Training data $[X|\mathbf{y}]$, $X \in \mathbb{R}^{n\times(1+d)}$ is the design matrix, $\mathbf{y} \in \mathbb{T}^n$ is the response vector. Model the post-infection time as a **real function** of the gene-expression profile:

$$f : \mathbb{R}^d \to [0, 26] \subset \mathbb{R}$$

Linear regression:

$$y_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, .., n,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

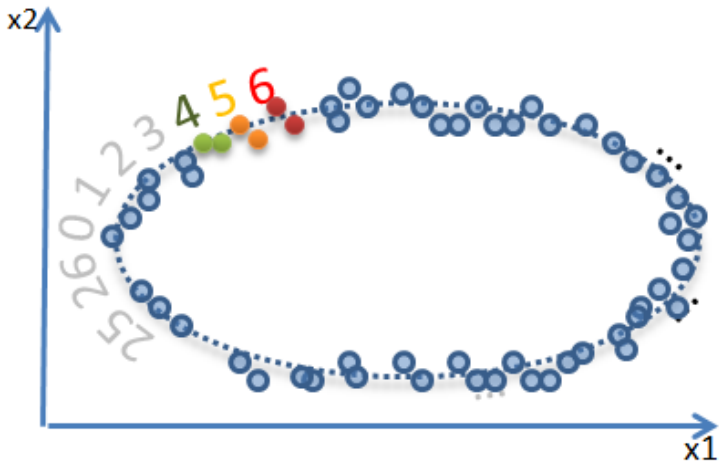
Analytical solution via Ordinary Least Squares (OLS):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 = (X^T X)^{-1} X^T \mathbf{y}$$

Leave-One-Mouse-Out Cross Validation

Can we do better?

Classification Formulation



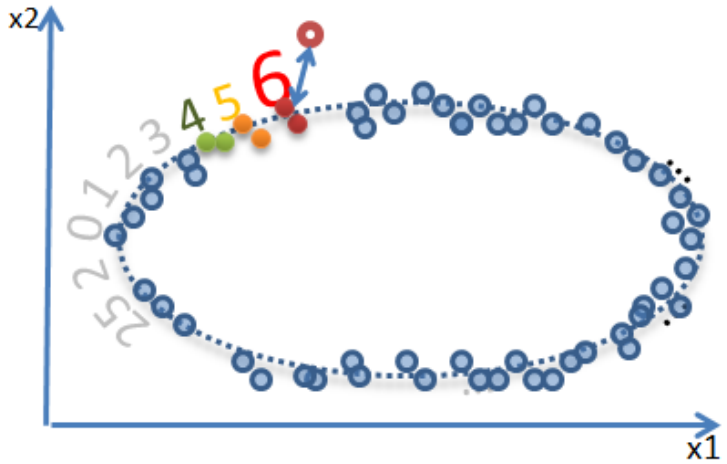
Consider the post-infection time as a **discrete variable**.
Learn a score function, $\mathcal{S}$, e.g. a probability, for each discrete value:

$$\mathcal{S}_j : \mathbb{R}^d \to \mathbb{R}$$

One-against-all "predictor" function:

$$\mathcal{P}(\mathbf{x}) := \arg\max_{j \in \mathbb{T}} \mathcal{S}_j(\mathbf{x})$$

First Nearest Neighbor



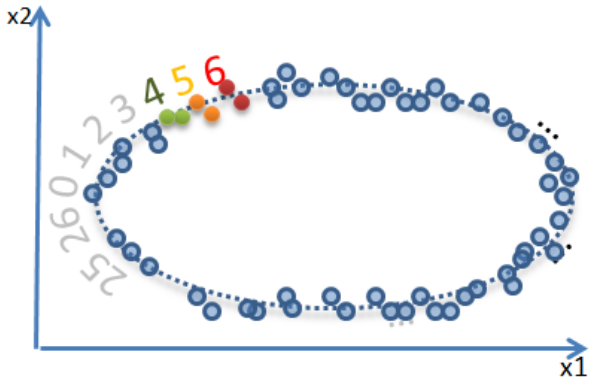
$$\mathcal{P}_{kNN}(\mathbf{x};\ [X|\mathbf{y}], \delta) := \arg\max_{y \in \mathbb{T}} \sum_{i \in N_k} 1[y = y_i]$$

Note: In the case of less than three training samples for every class, the only possible choise is $k = 1$ (First Nearest Neighbor)

Leave-One-Mouse-Out Cross Validation

Can we do better?

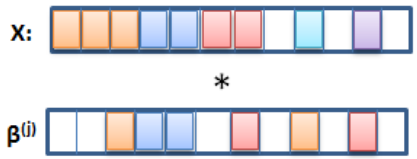
One-Against-All Binary Classification



Model the probabilities $\pi^{(j)}(\mathbf{x})$ of the transcriptome $\mathbf{x}$ to belong to the day j, $j \in \mathbb{T}$.
Training data for all days:$[X|Y]$ where $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^T \in \mathbb{R}^{n \times (1+d)}$ and
$Y \in \{-1, 1\}^{n \times t}$ is a binary representation of the post-infection time for each sample:

X	1	2	3	...	25	26
$\mathbf{x}_1$	1	0	0	...	0	0
$\mathbf{x}_2$	0	1	0	...	0	0
...	...	...	...	...	...	...
$\mathbf{x}_n$	0	0	0	...	0	1

One-Against-All Linear Logistic Regression



Model the logit function, $logit(\pi) := log(\pi/(1-\pi))$, as a linear function of $\mathbf{x}$:

$$logit(\pi^{(j)}(\mathbf{x})) \approx \mathbf{x}^T \beta^{(j)}.$$

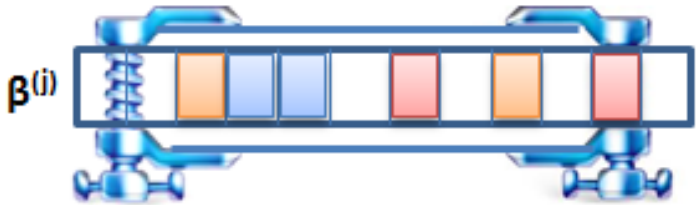
The negative log-likelihood is defined as:

$$-\ell^{(j)}(\beta^{(j)}; [X|\mathbf{y}_j]) = \sum \log\left(\mathbf{1} + \exp(-\mathbf{y}_j \odot X\beta^{(j)})\right), \quad j = 1, ..., t.$$

Maximum likelihood fit for $\beta^{(j)}$:

$$\beta^{(j)*} := \arg \min_{\beta^{(j)} \in \mathbb{R}^{(1+d)}} \left\{-\ell^{(j)}(\beta^{(j)}; [X|\mathbf{y}_j])\right\}$$

Regularization and Automatic Variable Selection

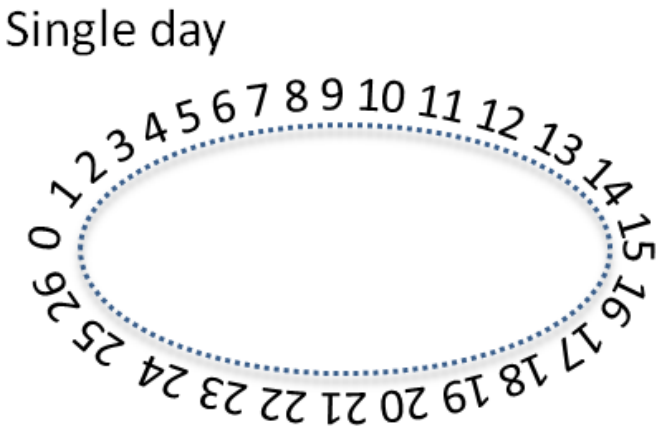


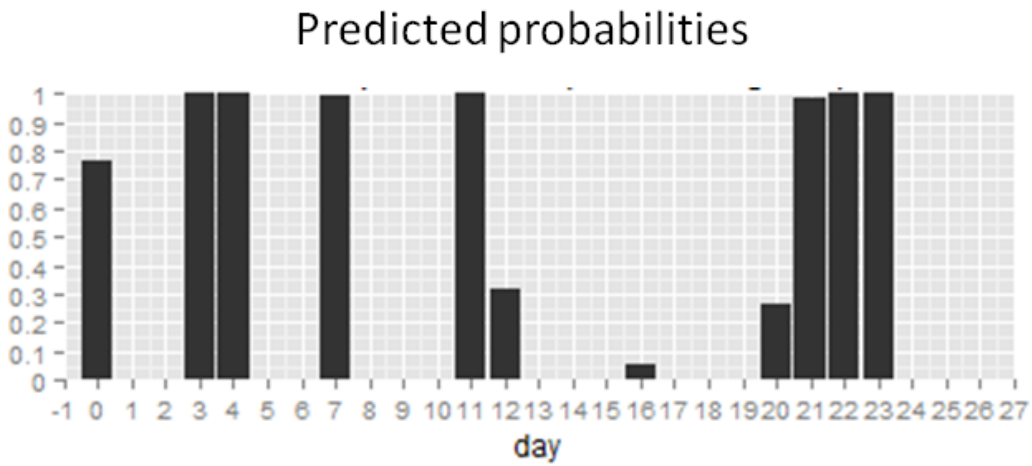
- L2-penalty (Ridge): $\frac{1}{2}\lambda_2||\boldsymbol{\beta}^{(j)}||_2^2 = \frac{1}{2}\lambda_2 \sum_{k=1}^{d} \beta_k^{(j)^2}$
- L1-penalty (Lasso): $\lambda_1||\boldsymbol{\beta}^{(j)}||_1 = \lambda_1 \sum_{k=1}^{d} |\beta_k|$
- Elastic Net penalty (Lasso+Ridge):$\lambda_1||\boldsymbol{\beta}^{(j)}||_1 + \frac{1}{2}\lambda_2||\boldsymbol{\beta}^{(j)}||_2^2$

Maximum A-Posteriori fit for $\boldsymbol{\beta}^{(j)}$:

$$\boldsymbol{\beta}^{(j)*} := \arg \min_{\boldsymbol{\beta}^{(j)} \in \mathbb{R}^{(1+d)}} \left\{ -\ell^{(j)}(\boldsymbol{\beta}^{(j)}; [X|\mathbf{y}_j]) + \lambda_1||\beta||_1 + \frac{1}{2}\lambda_2||\boldsymbol{\beta}^{(j)}||_2^2 \right\}$$

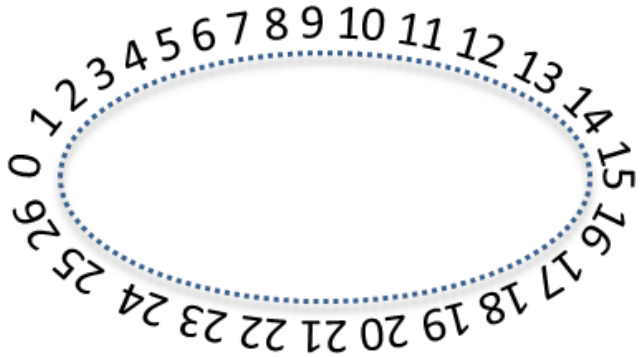
Single Day versus Time Window Prediction



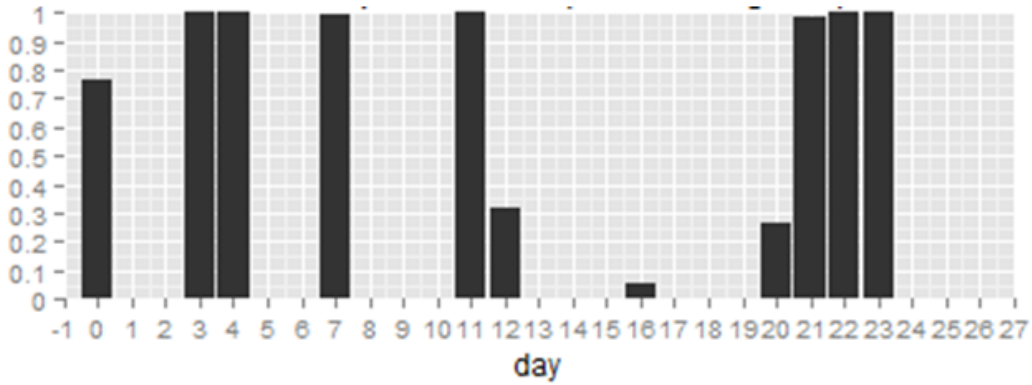


Single Day versus Time Window Prediction

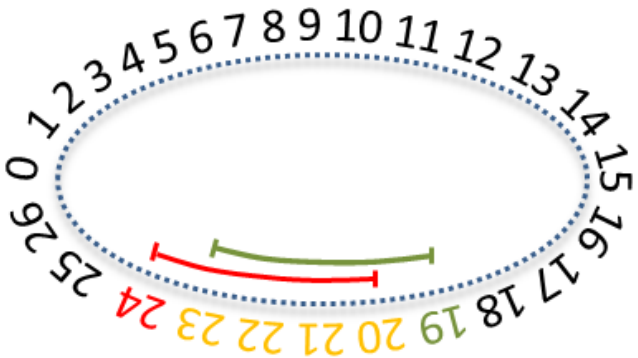
Single day

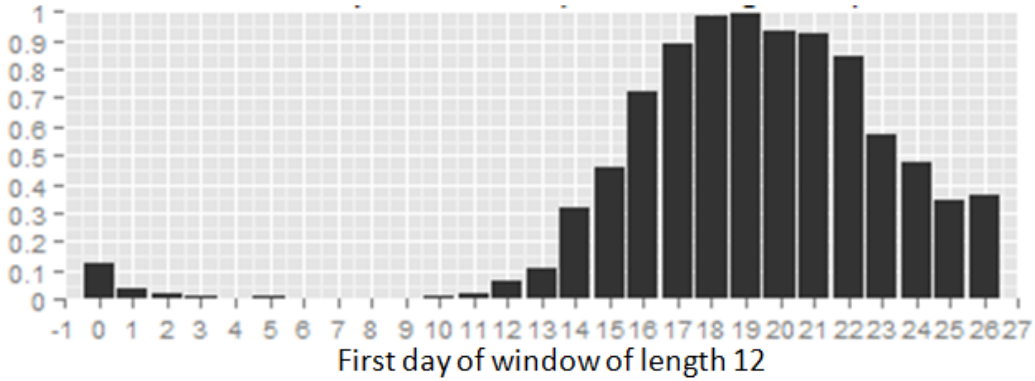


Predicted probabilities

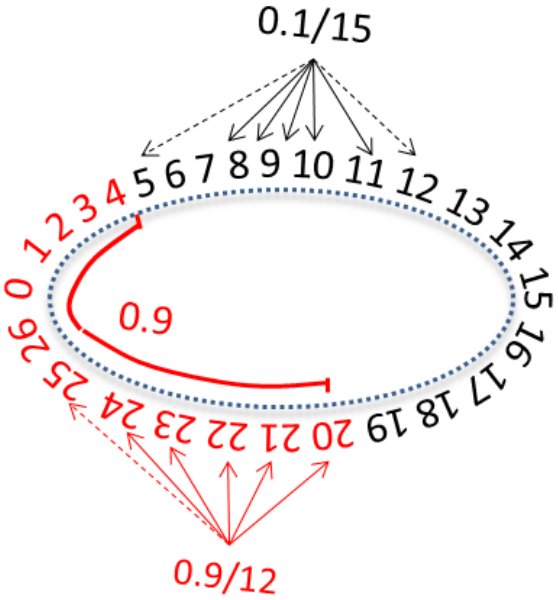


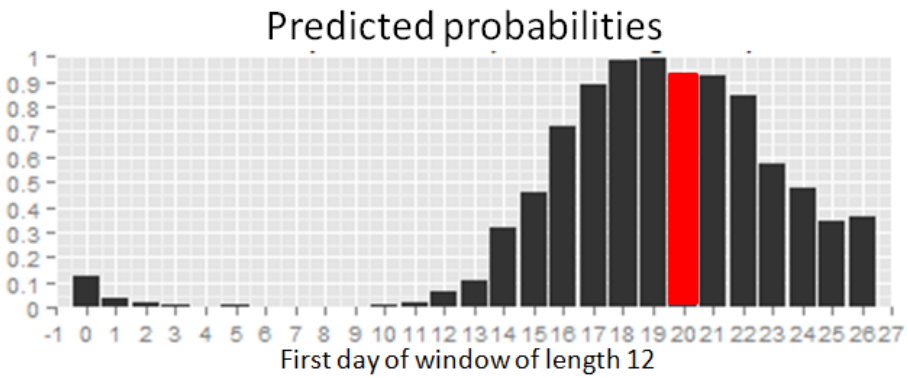
Time Window



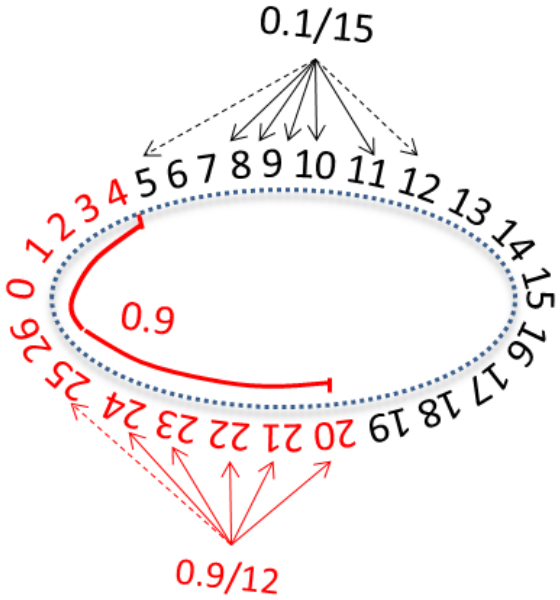


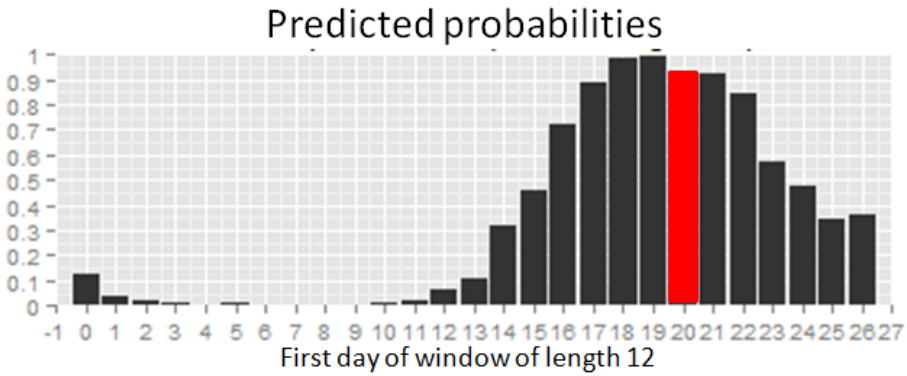
Aggregated Time Window Predictor (ATWINP)

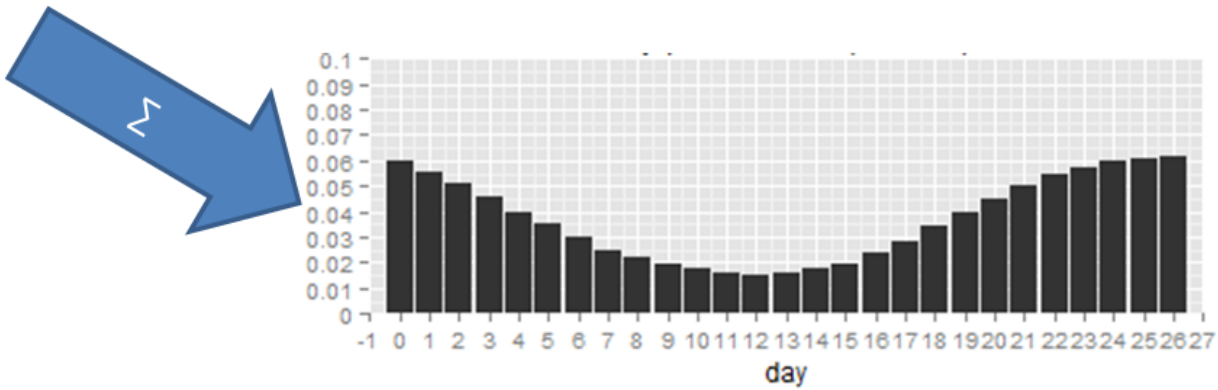




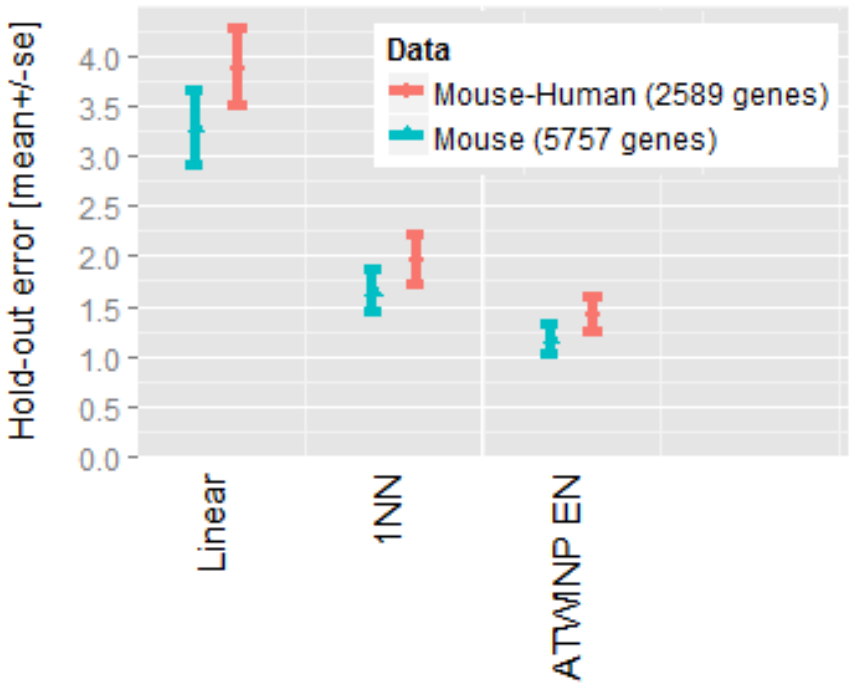
Aggregated Time Window Predictor (ATWINP)





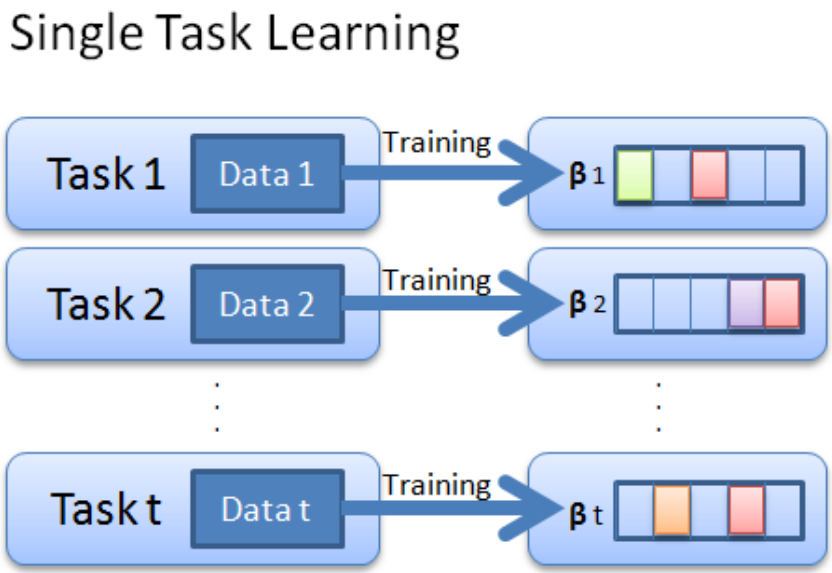


Leave-One-Mouse-Out Cross Validation

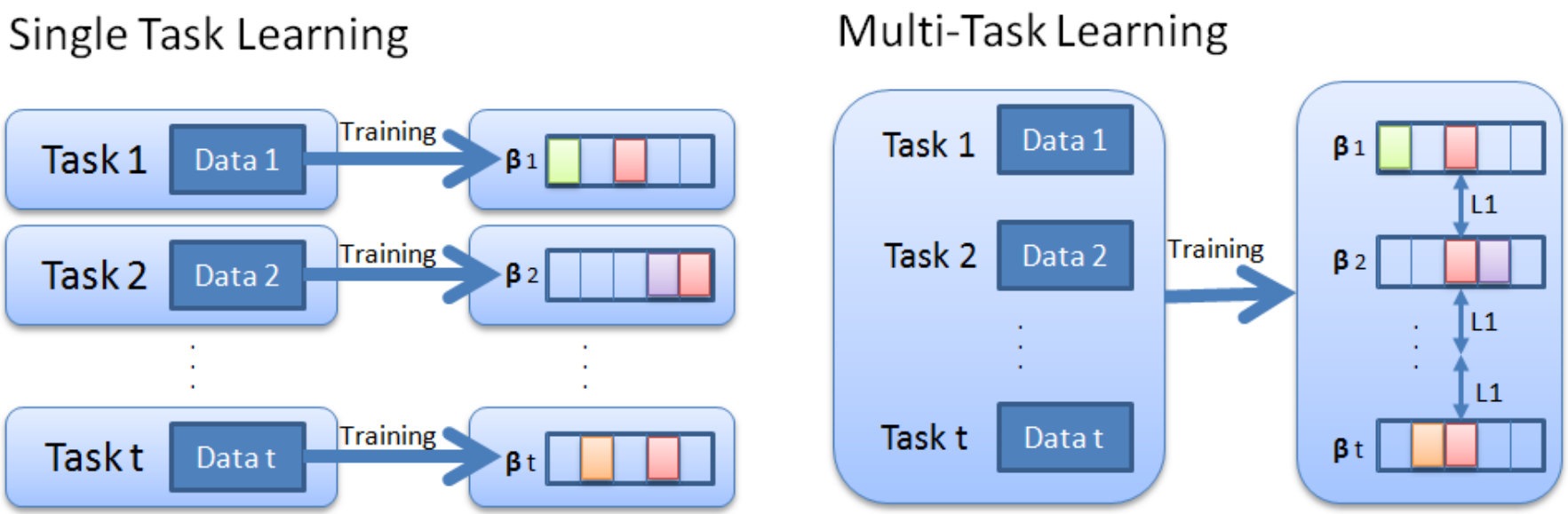


Can we do better?

The Idea of Multi-Task Learning



The Idea of Multi-Task Learning



Fused Elastic Net Logistic Regression (FLR)

Let $B := [\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(t)}] \in \mathbb{R}^{(1+d)\times t}$ be the coefficient matrix for all tasks and let $R \in \mathbb{R}^{t\times t}$ be a matrix defined in the following way:

$$R_{ij} := \begin{cases} 1 & \text{if } j = i-1 \text{ or } (i,j) = (1,t) \\ 0 & \text{otherwise} \end{cases}, \quad i,j = 1, ..., t.$$

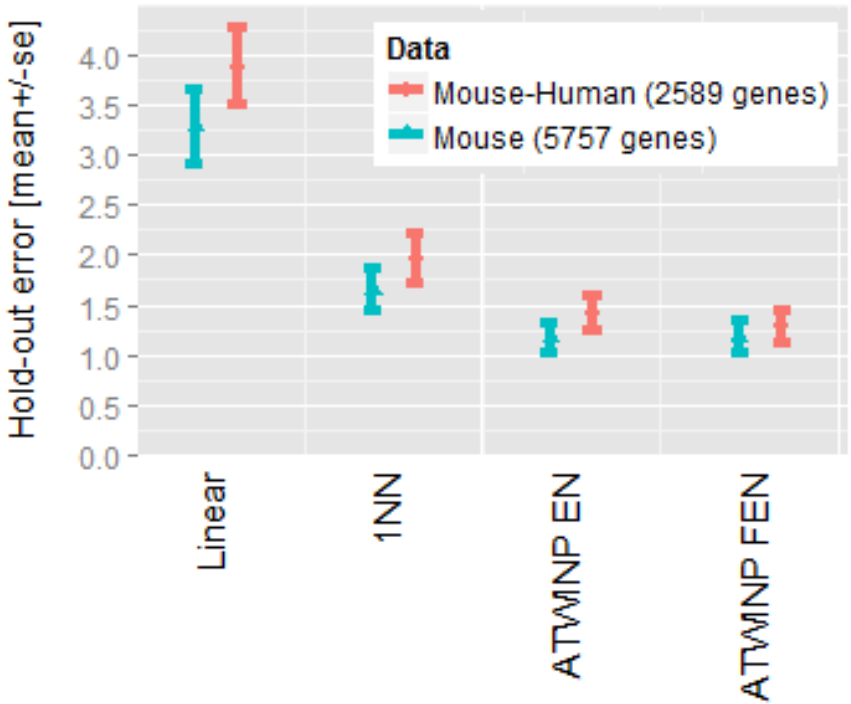
The multi-task fused elastic net negative log-likelihood is defined as:

$$\begin{aligned} -\ell^{MT}(B; [X|Y]) \quad := \quad & \sum \log\Big([1] + \exp(-Y \odot XB)\Big) \\ & + ||[\boldsymbol{\lambda}_1] \odot B||_1 + \frac{1}{2}||[\boldsymbol{\lambda}_2] \odot B||_2^2 \\ & + ||[\nu] \odot B(I - R)||_1 \end{aligned}$$

The Fused Elastic Neto Logistic Regression (FENLR) fit for B is obtained by solving

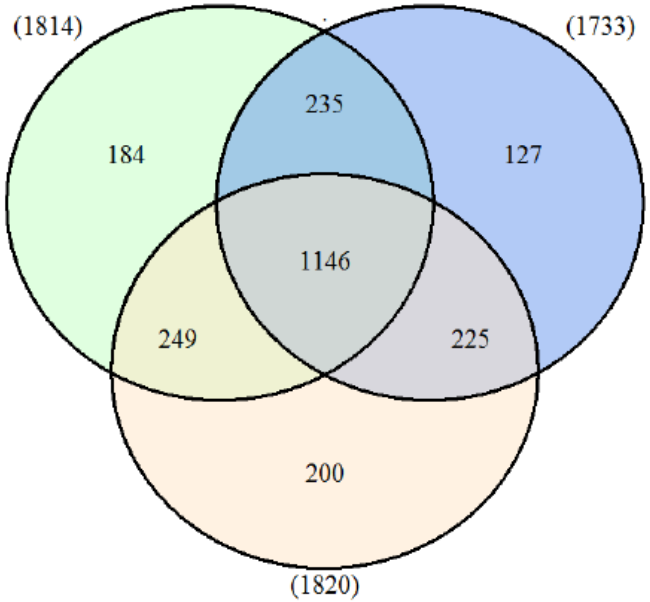
$$B^* = \arg \min_{B \in \mathbb{R}^{(1+d)\times t}} -\ell^{MT}(B; [X|Y]).$$

Leave-One-Mouse-Out Cross Validation

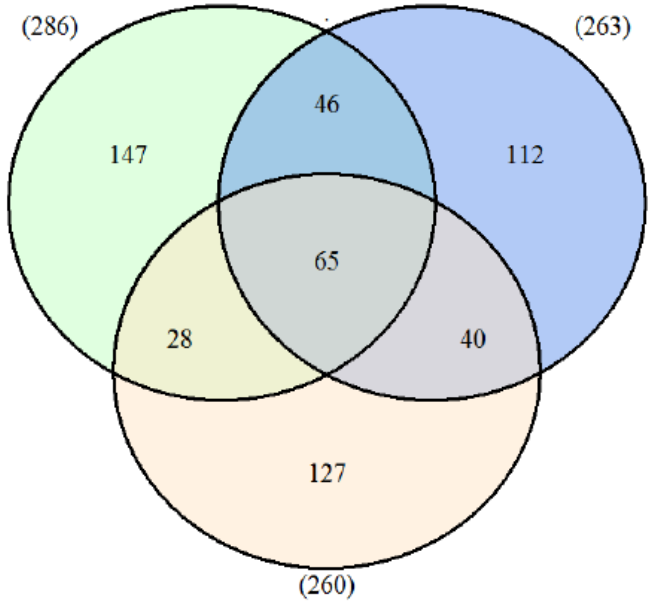


Selected Genes

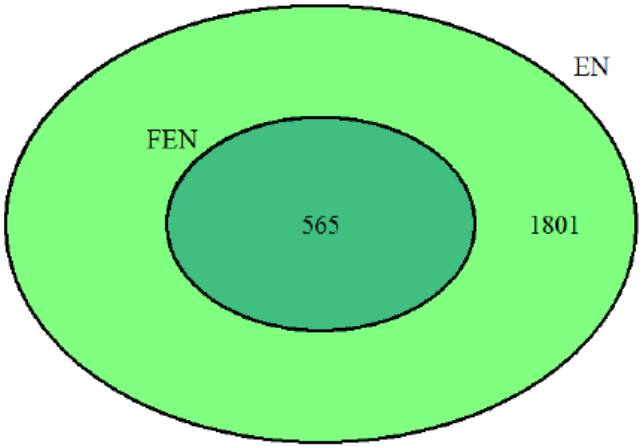
c) Mouse-Human, ATWINP EN (2366 out of 2589)



d) Mouse-Human, ATWINP FEN (565 out of 2589)

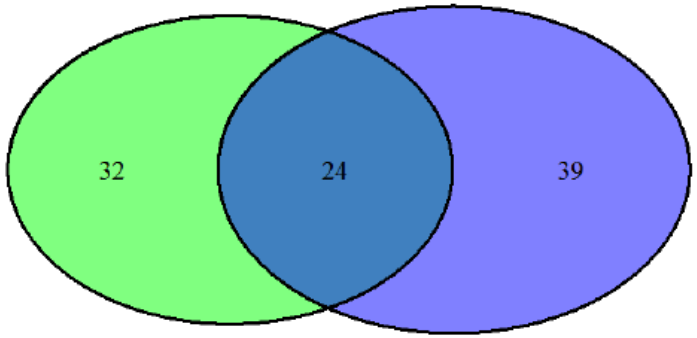


e) Mouse-Human: union selection of CV-folds:

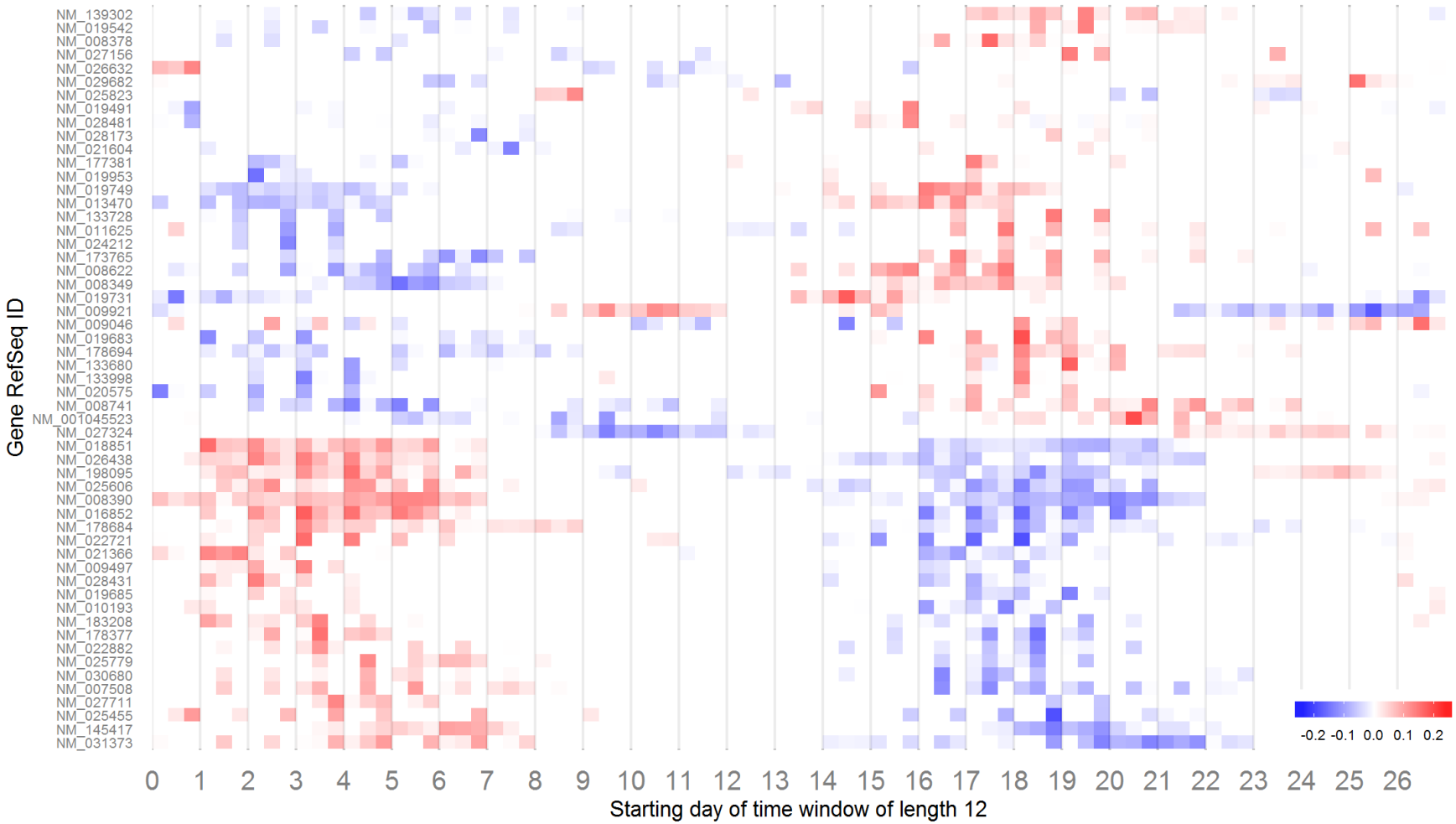


f) Mouse-Human (top ~60 out of 2589):

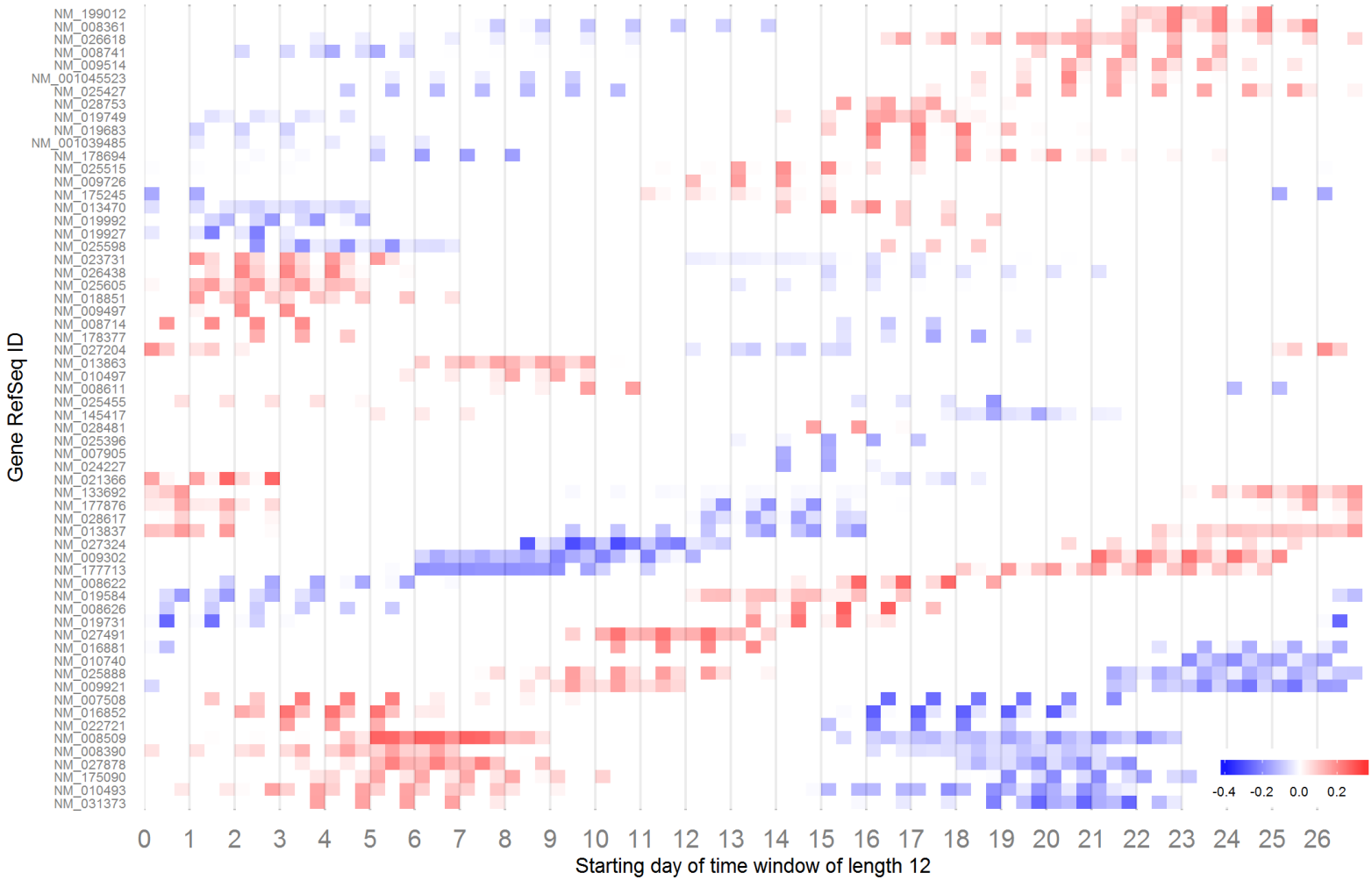
ATWINP EN (top 56) ATWINP FEN (top 63)

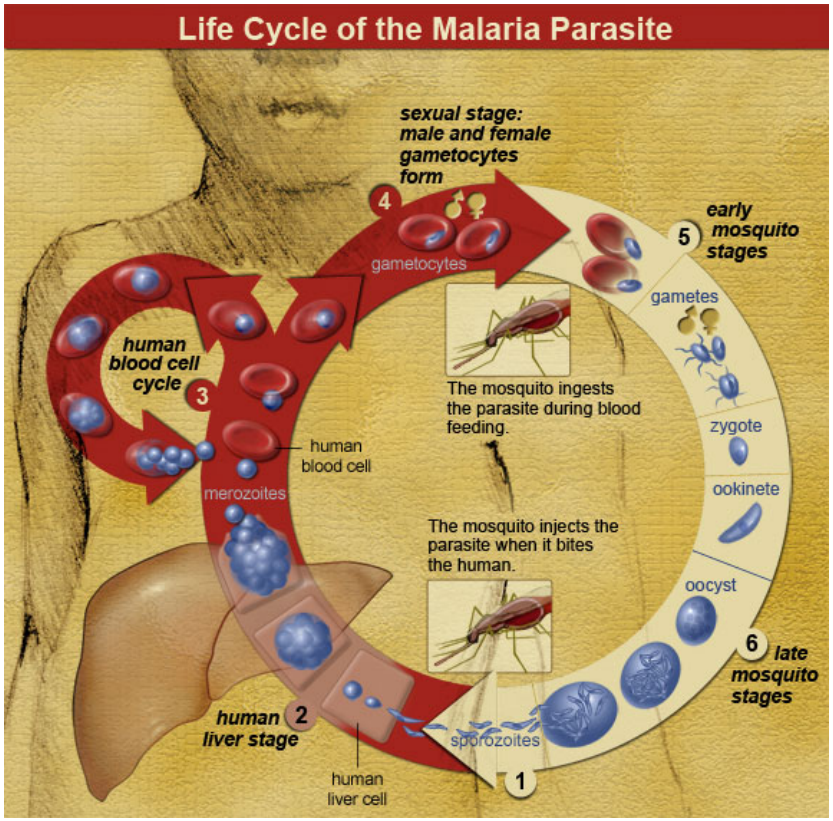


Top 60 Genes, ATWINP EN



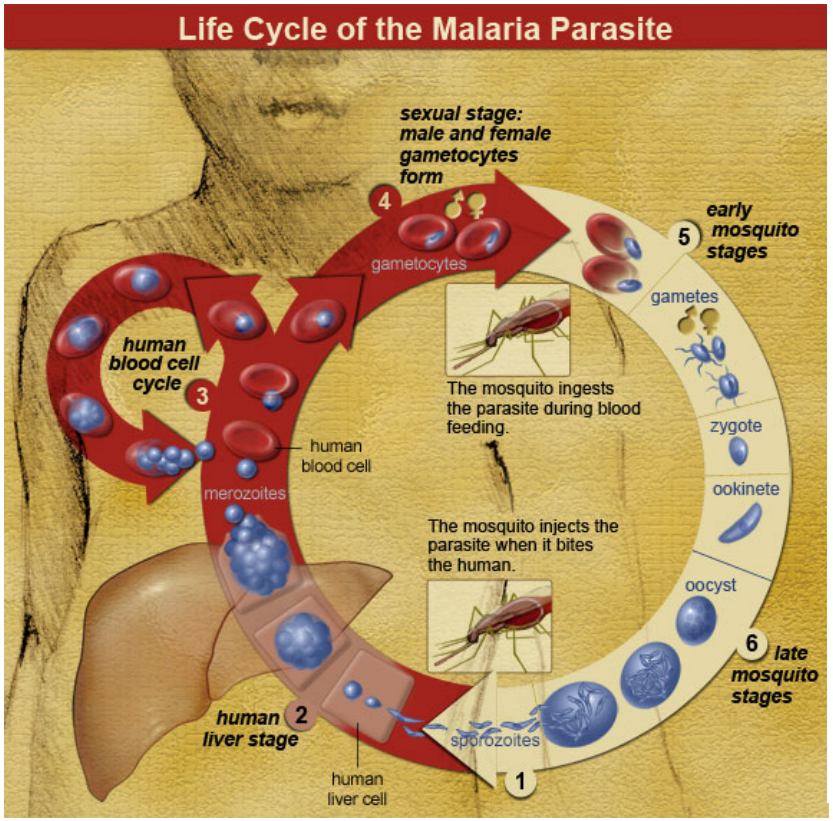
Top 60 Genes, ATWINP FEN

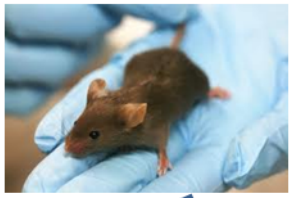




Courtesy: National Institute of Allergy and

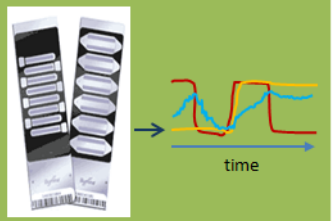
Infectious Diseases

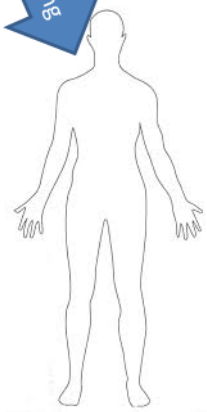




Transfer learning

1. Find genes in infected mice, which have informative time-course dynamics for the inference of post-infection time in mice





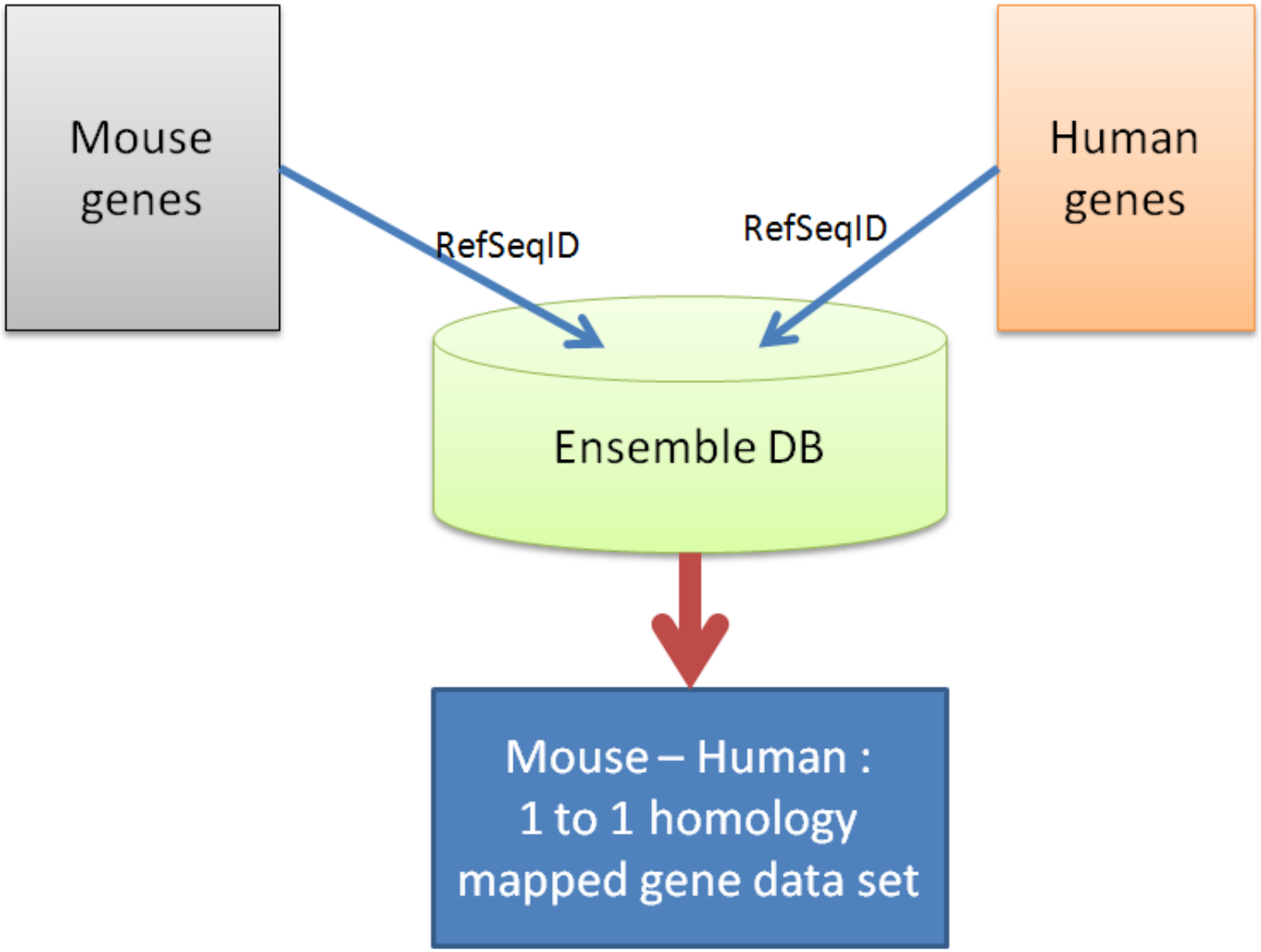
2. Map these genes to their human homologs and narrow down this set to genes that are relevant for malaria progression in the human context

M. musculus → H. sapiens

EIF6
PDCD2
ACHE
...

Courtesy: National Institute of Allergy and Infectious Diseases

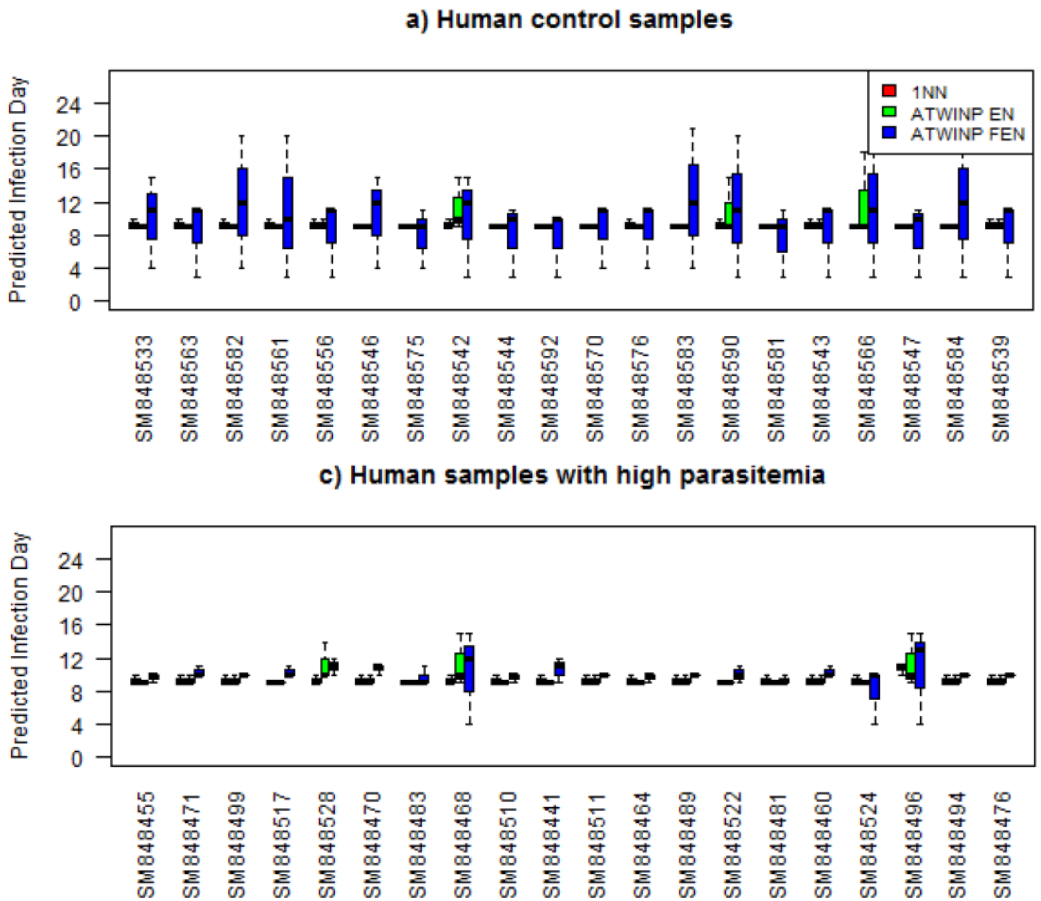
Homology Mapping Between Mouse and Human Genes



Homology Mapping Between Mouse and Human Genes

- Of 18744 mouse sequences:
  - 15587 have a homologous sequence found in human,
  - 15328 of which are available on the human BeadChip of which:
- 7683 mouse sequences point to a unique human sequence,
- 6832 mouse sequences point to more than one human sequence,
- 813 mouse sequences point to human sequences pointed by other mouse sequences

Post-Infection Time Prediction in Human Patients



Discussion

- Our model can predict the post-infection time of an unlabeled infected mouse-sample with expected deviation of 1.28 days from the true post-infection time.
- The gene-expression profile of an infected host-organism preserves information with respect to the beginning of the infection, and can be used to characterize the disease progression on a fine time-scale.
- We were able to identify a set of genes that are informative for the disease progression in mice and we could quantify the effect of each selected gene at all points in the time-course of the infection.
- At the current time knowledge transfer from mouse to human patients cannot provide a valuable estimation of the post-infection time in humans.

Acknowledgements

- prof. Manfred Claassen - Advisor of the master thesis project
- Stefan, Eirini, Anita, Ana - colleagues in the Claassen's Group
- David and Brenda - Stanford Microbiology and Immunology Lab