Electronic Supplementary Material (ESI) for Molecular Omics.

Supplementary information

Systems-level analysis of transcriptome reorganization during liver regeneration

Manisri Porukala and P K Vinod*

Centre for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad-500032, India

*Correspondence: vinod.pk@iiit.ac.in

Supplementary methods:

We describe the DREM method proposed by Schulz *et al*., (2012) to integrate time-series gene expression data with TF-gene association data. DREM method is based on Input-Output Hidden Markov Model where the hidden states are used to group the genes into paths. Each of these paths pass through the hidden states over time. Transitions among the hidden states are constrained to obtain the tree-like structure resulting in the bifurcation points. Many possible tree-like structures are searched and scored to select the best one.

The DREM model M is a tuple (H, E, Ψ, Θ, n, γ), where n corresponds to the number of discrete time points, H is the set of hidden states 'h' such that every h is characterized by a Gaussian output distribution f_h and is associated with one time point. Θ is the set of parameters corresponding to (μ_h, σ_h) the Gaussian distribution of each hidden state. Every hidden state h_t with t < n can have at most γ child nodes (hidden states of h in the next time point). E is the set of directed edges between h_a and h_b s.t a+1=b reflecting the transitions among the hidden states (i.e hidden states of the consecutive time points are connected). Ψ consists of parameters controlling the transition probabilities between the hidden states (and their child nodes). Ψ_h for every hidden state h is a vector of parameters for logistic regression classifier which makes use of TF-gene association information. Let I_g be the vector defining TF-gene association for gene g and all its regulating TFs (1 for activating, -1 for inhibiting, 0 for no regulation). If gene g is in hidden state h_a at time t-1 and has h_b and h_c as child states at time t, then the probability that g undergoes a transition from h_a to h_b(or h_c) is given by logistic function with intercept INT as follows:

$$\frac{1}{1 + e^{-\Psi_h . INT - \sum_x \Psi_h x \,.\, I_g x}}$$

The expression level of TF is also incorporated to influence the learning of the classifier using the shifted version of logistic function as follows

$$f_w^*(x) = sign(x)\left(\frac{2}{1 + e^{-xw}} - 1\right)$$

where x is the expression ratio of a TF between two time points, w is the expression scaling weight that controls the steepness of the function. To incorporate the efficiency of TF, it is selected based on the minimum threshold

$$select(x) = \begin{cases} f_w^*(x), & if\ abs(f_w^*(x)) \geq minExpTF, \\ sign(x).minExpTF, & else \end{cases}$$

Therefore, pairwise binding information of a TF t to gene g ($B_g \in \{-1, 1, 0\}$) is now transformed to

$$B_g' = select(x_t) \,.\, B_g$$

If the vector $O_g = (o_g(1), o_g(2),\ldots,o_g(n-1))$ denotes the log ratio of expression values for a gene g at time points 1 to n-1 with respect to zeroth time point, with I_g as its TF-gene association mapping vector and H_i is its hidden state variable at time, then the probability it is in state h_b at time t given that it is in state h_a at time t-1 is given by $P(H_t = h_b | H_{t-1} = h_a, I_g)$. This probability is 0 if h_b is not

the child of h_a and 1 if it is the only child. If number of child nodes is greater than or equal to 2, then the transitions are probabilistic and are based on the vector I_g.

The likelihood density r, for a set of genes G for the model is given by

$$r(G|M) = \underbrace{\sum_{g \in G} log}_{all\ genes} \underbrace{\sum_{q \in Q}}_{all\ paths} \underbrace{\prod_{t=1}^{n-1} f_{q(t)}(o_g(t))}_{output\ densities} \underbrace{\prod_{t-1}^{n-1} P(H_t = q(t)|H_{t-1} = q(t-1), I_g)}_{transition\ probabilities}$$

where, Q is the set all paths of hidden states of length n beginning from the root with non-zero probability. For a path q Є Q, q(i) is the hidden state of the path at time point i. The first product corresponds to the product of output densities for the expression values and a given sequence of hidden states. Given a tree structure determined by H and E, the parameters that maximize the likelihood density r are estimated using Baum-Welch algorithm.

To build the dynamic regulatory map, the algorithm begins with a search over structures of trees. Parameters of Gaussian distribution and the classifiers corresponding to a tree structure are learnt by randomly selecting the subset of genes as training set using a version of Baum-Welch algorithm. The remaining genes are used as test set to assign scores to various tree structures considered. The best scoring tree structure is selected, and all genes are used to train the parameters and arrive at a final model.

Table S1: Significant KEGG pathways and biological processes associated with WGCNA modules. * represents uncorrected p-value.

Modules	KEGG pathways	Biological processes
Blue (M1)	DNA replication, Cell cycle, Homologous recombination, Mismatch repair, Cellular senescence, Spliceosome, Nucleotide excision repair, p53 signaling pathway, Pyrimidine metabolism, RNA transport, Glutathione metabolism, Steroid biosynthesis, Amino sugar and nucleotide sugar metabolism, Arginine and proline metabolism, Protein processing in endoplasmic reticulum	DNA metabolic process, mitotic cell cycle phase transition, DNA replication, G1/S transition of mitotic cell cycle, DNA repair, cellular response to DNA damage stimulus, tRNA export from nucleus, RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
Green (M2)	Complement and coagulation cascades*, Fatty acid degradation*, Protein export*, PPAR signaling pathway*	Platelet degranulation, acute inflammatory response positive regulation of protein secretion, cargo loading into COPII-coated vesicle, homotypic cell-cell adhesion, Golgi vesicle transport, fatty acid beta-oxidation
Red (M3)	Protein processing in endoplasmic reticulum, Protein export, N-Glycan biosynthesis, Amino sugar and nucleotide sugar metabolism	IRE1-mediated unfolded protein response, protein N-linked glycosylation, ERAD pathway, ubiquitin-dependent ERAD pathway, ATF6-mediated unfolded protein response
Pink (M4)	TNF signaling pathway, Ribosome biogenesis in eukaryotes, JAK-STAT signaling pathway*	Ribosome biogenesis, rRNA processing, maturation of 5.8S rRNA, protein localization to nucleus, ribosome assembly, response to cytokine, interleukin-6-mediated signaling pathway*
Purple (M5)	PPAR signaling pathway, Fatty acid degradation, Peroxisome, Biosynthesis of unsaturated fatty acids, TGF-beta signaling pathway*	Regulation of lipid metabolic process, fatty acid beta-oxidation, positive regulation of pathway-restricted SMAD protein phosphorylation*, peroxisomal membrane transport*
Black (M6)	Steroid hormone biosynthesis, PPAR signaling pathway, Biosynthesis of unsaturated fatty acids, Primary bile acid biosynthesis, Arachidonic acid metabolism, Cholesterol metabolism, Retinol metabolism, Tryptophan metabolism, Glutathione metabolism, Bile secretion, Fatty acid degradation, Complement and coagulation cascades, Arginine biosynthesis	Alpha-amino acid catabolic process, aromatic amino acid family catabolic process, tryptophan metabolic process, tryptophan catabolic process, cholesterol homeostasis, sterol metabolic process, valine metabolic process
Brown (M8)	Chemical carcinogenesis, Steroid hormone biosynthesis, Retinol metabolism, Glutathione metabolism, Pentose and glucuronate interconversions, Tryptophan metabolism, Glycine, serine and threonine metabolism, Valine, leucine and isoleucine degradation, Glyoxylate and dicarboxylate metabolism, Histidine metabolism, Cysteine and methionine metabolism, Arachidonic acid metabolism, Phenylalanine metabolism, Tyrosine metabolism	Alpha-amino acid catabolic process, glutathione metabolic process, branched-chain amino acid metabolic process, histidine metabolic process
Magenta (M9)	Retinol metabolism, Glyoxylate and dicarboxylate metabolism, Glycine, serine and threonine metabolism, One carbon pool by folate, Tryptophan metabolism, Complement and coagulation cascades, Terpenoid backbone biosynthesis, Folate biosynthesis, Pentose and glucuronate interconversions	Response to sterol, regulation of cholesterol biosynthetic process, serine family amino acid metabolic process, coenzyme metabolic process

Table S2: Significant KEGG pathways and biological processes (adj p-value < 0.05) associated with different paths of cluster 1,2 and 3 (validation dataset). * represent uncorrected p-value.

Path	KEGG pathway	Biological processes
A	Cell cycle, p53 signaling pathway, Cellular senescence, DNA replication, Homologous recombination, Base excision repair*, FoxO signaling pathway *	Mitotic cell cycle phase transition, DNA metabolic process, cellular response to DNA damage stimulus
B	Circadian rhythm*, Steroid hormone biosynthesis *	Acute-phase response*, neutrophil degranulation*
C	Cell cycle, DNA replication, Homologous recombination, Base excision repair	DNA metabolic process, regulation of transcription involved in G1/S and G2/M transition of mitotic cell cycle, signal transduction by p53 class mediator
D	DNA replication, Homologous recombination, Mismatch repair, Cell cycle, Nucleotide excision repair, Base excision repair, Cellular senescence*, One carbon pool by folate *	DNA repair, DNA replication, tRNA export from nucleus
E	Glycerophospholipid metabolism*, Complement and coagulation cascades*, TNF signaling pathway*, Glycerolipid metabolism*, NF-kappa B signaling pathway*	Cellular response to cytokine stimulus, inflammatory response
F	DNA replication*, RNA degradation*, RNA transport*, Pyrimidine metabolism*, Mismatch repair*, Spliceosome*, Nucleotide excision repair *	DNA metabolic process*, protein import*, nucleobase-containing small molecule interconversion*
G	p53 signaling pathway*, Non-homologous end-joining*, Homologous recombination*, One carbon pool by folate*, Cysteine and methionine metabolism *	DNA recombination*, DNA damage induced protein phosphorylation*
H	Spliceosome, Cell cycle, p53 signaling pathway, Protein processing in endoplasmic reticulum*, Amino sugar and nucleotide sugar metabolism*, Purine metabolism*, RNA transport*.	RNA splicing, protein N-linked glycosylation, RNA transport,
I	NF-kappa B signaling pathway, Platelet activation, TNF signaling pathway,	Cellular defense response, inflammatory response
J	Fatty acid degradation*, Wnt signaling pathway*, PPAR signaling pathway*, Bile secretion*, Valine, leucine and isoleucine degradation*, Taurine and hypotaurine metabolism *	Fatty acid beta-oxidation, regulation of canonical Wnt signaling pathway*
K	Steroid hormone biosynthesis*, ECM-receptor interaction*, Metabolism of xenobiotics by cytochrome P450*	Positive regulation of cytokine biosynthetic process*
L	Steroid hormone biosynthesis, Metabolism of xenobiotics by cytochrome P450, Chemical carcinogenesis, Drug metabolism, PPAR signaling pathway, Primary bile acid biosynthesis, Retinol metabolism, Cholesterol metabolism, Glutathione metabolism, Tryptophan metabolism, Arachidonic acid metabolism,	Steroid biosynthetic process, bile acid biosynthetic process*
M	TGF-beta signaling pathway*, Primary bile acid biosynthesis*, Glycine, serine and threonine metabolism*, Bile secretion *	Bile acid metabolic process*,
N	Drug metabolism, Steroid hormone biosynthesis*, Glycolysis / Gluconeogenesis*, Biosynthesis of unsaturated fatty acids*, PPAR signaling pathway*	Regulation of biosynthetic process*, regulation of gluconeogenesis*

Table S3: Parameter values used for the phase plane and bifurcation analyses. Model1 corresponds to figure 6A and model 2 corresponds to figure 8A. The input S (proliferative) and M (compensatory) are varied.

Parameter	Description	Model 1	Model 2
β_{HNF4A}	Time scale for rate of change of HNF4A	1	1
δ_{HNF4A}	Steepness of soft-Heaviside function for HNF4A	8	8
w_{HNF4A_0}	Basal coefficient for HNF4A	0.3	0.5
w_{HNF4A_CycD}	Coefficient for HNF4A inhibition by Cyclin D	-1.2	-1.2
w_{HNF4A_SNAIL}	Coefficient for HNF4A inhibition by SNAIL	-	-0.2
β_{CycD}	Time scale for rate of change of Cyclin D	1	1
δ_{CycD}	Steepness of soft-Heaviside function for Cyclin D	8	8
w_{CycD_0}	Basal coefficient for Cyclin D	-0.5	-0.5
w_{CycD_Myc}	Coefficient for activation of Cyclin D by Myc	1.1	1.1
δ_{Myc}	Steepness of soft-Heaviside function	8	8
w_{Myc_HNF4A}	Coefficient for Myc inhibition by HNF4A	-0.5	-0.5
β_{SNAIL}	Time scale for rate of change of SNAIL	-	1
δ_{SNAIL}	Steepness of soft-Heaviside function	-	8
w_{SNAIL_0}	Basal coefficient for SNAIL	-	-0.1
w_{SNAIL_HNF4A}	Coefficient for SNAIL inhibition by HNF4A	-	-1
w_{SNAIL_SNAIL}	Coefficient for self-activation of SNAIL	-	1.3
w_{S_HNF4A}	Coefficient for HNF4A inhibition by S (input)	-1	
w_{M_HNF4A}	Coefficient for HNF4A activation by M (input)	1	

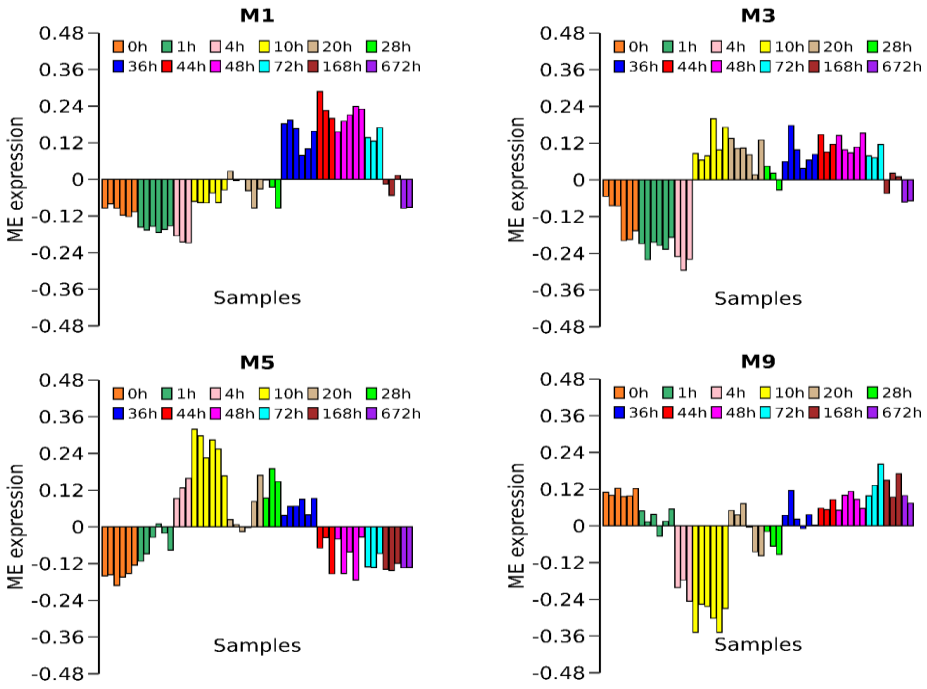

Figure S1: Eigen gene expression profile of individual modules with respect to different time points of liver regeneration.

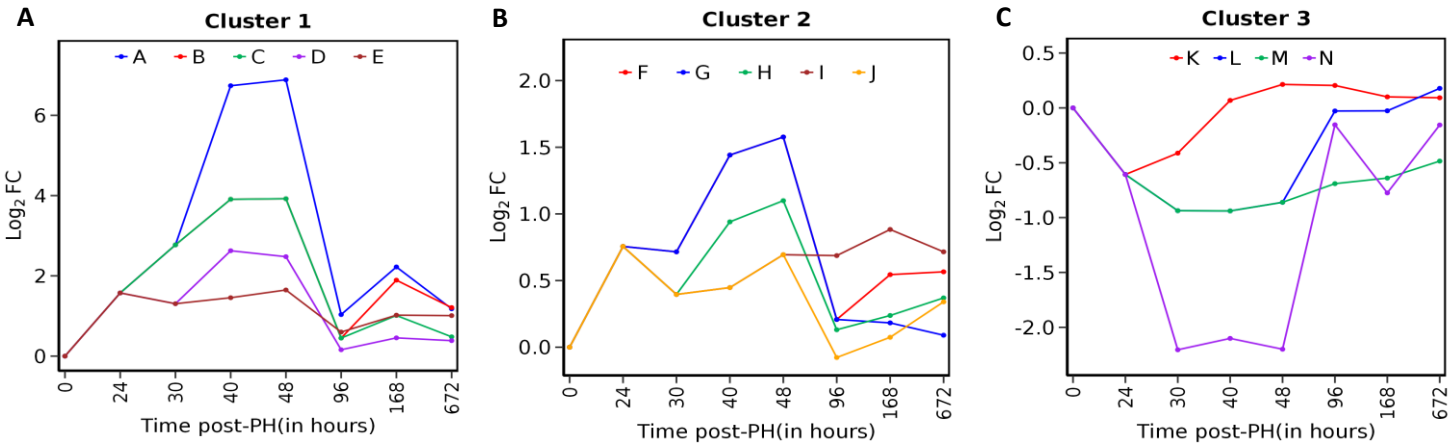

Figure S2: The regulatory paths of the set of co-expressed genes in the three core clusters of liver regeneration (validation dataset). The x-axis represents the time points of sample collection, and the y-axis represents the log2 fold change (log_2FC) in mRNA expression post-PHx. A path is split into multiple paths (split nodes) based on the divergence in gene expression.

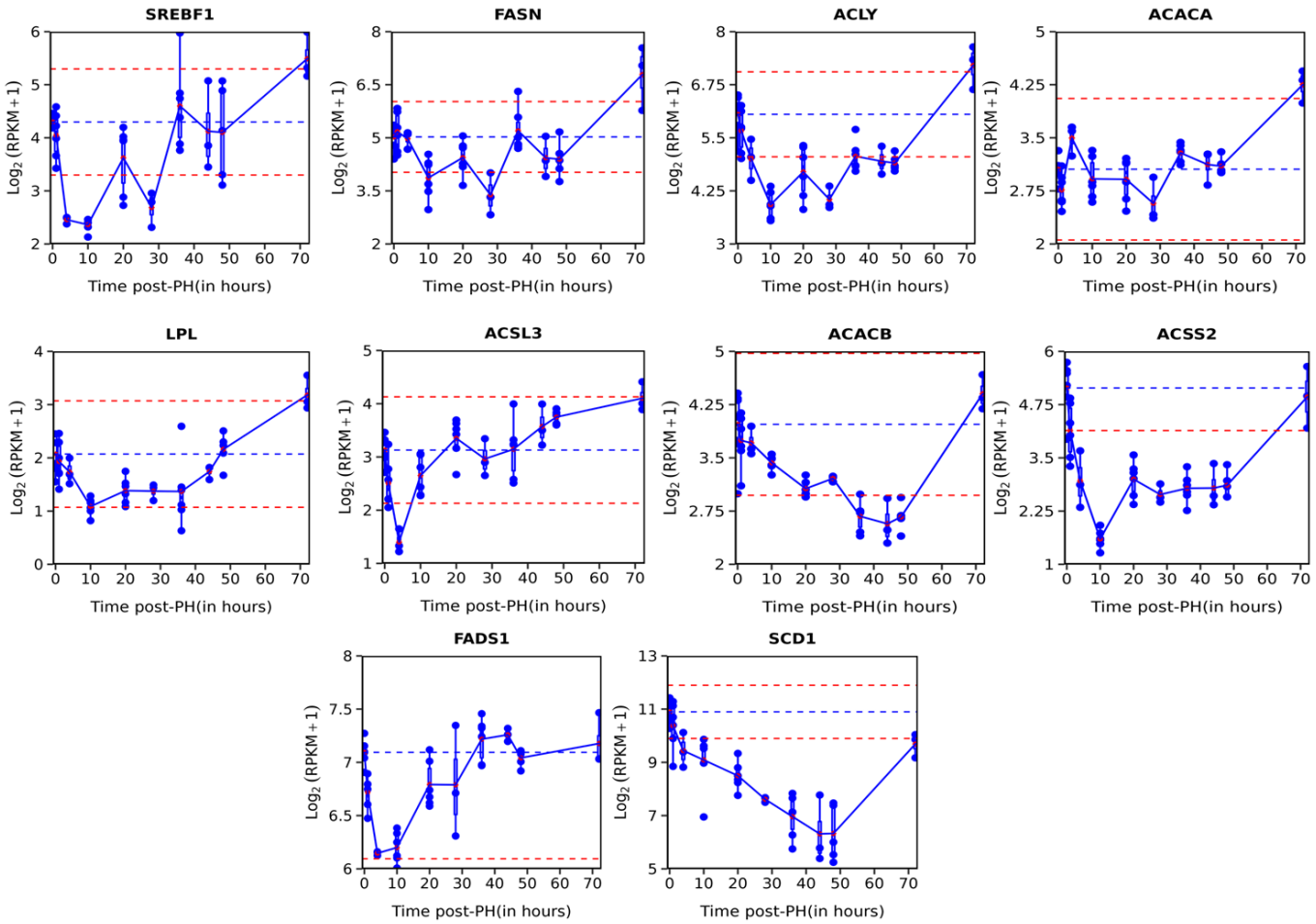

Figure S3: Expression profile of genes in lipid biosynthesis pathway. The blue line represents the baseline (0 hours), and the red line represents a two-fold change with respect to the baseline.

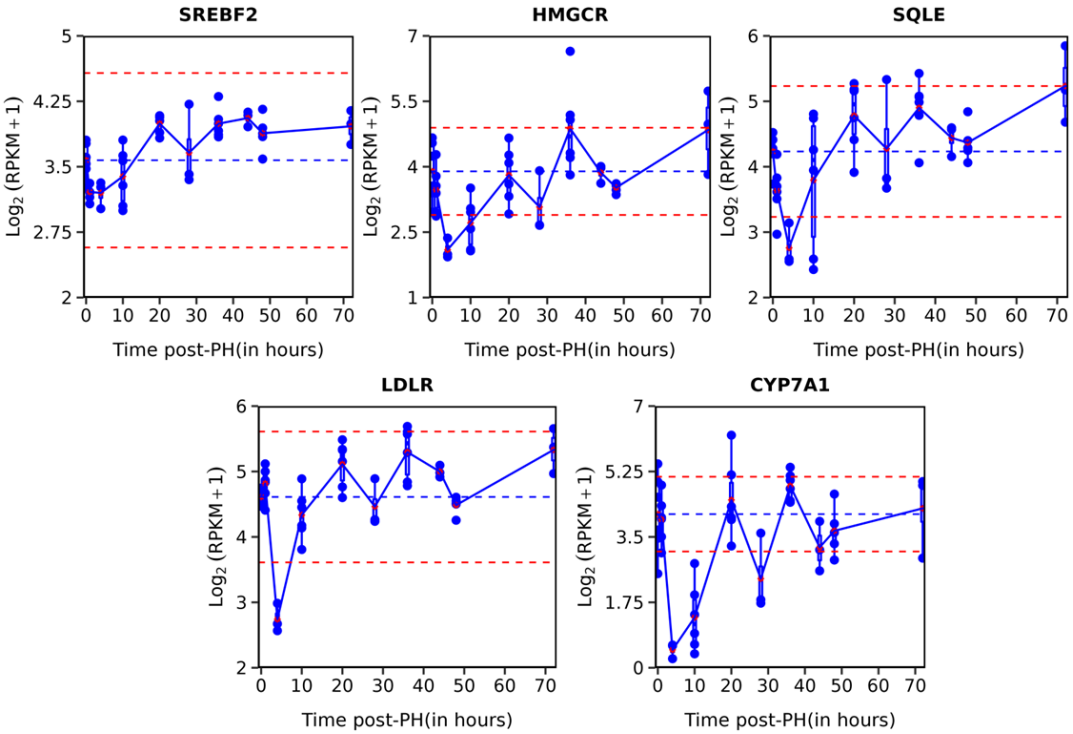

Figure S4: Expression profile of genes in cholesterol metabolism pathway. The blue line represents the baseline (0 hours), and the red line represents a two-fold change with respect to the baseline.

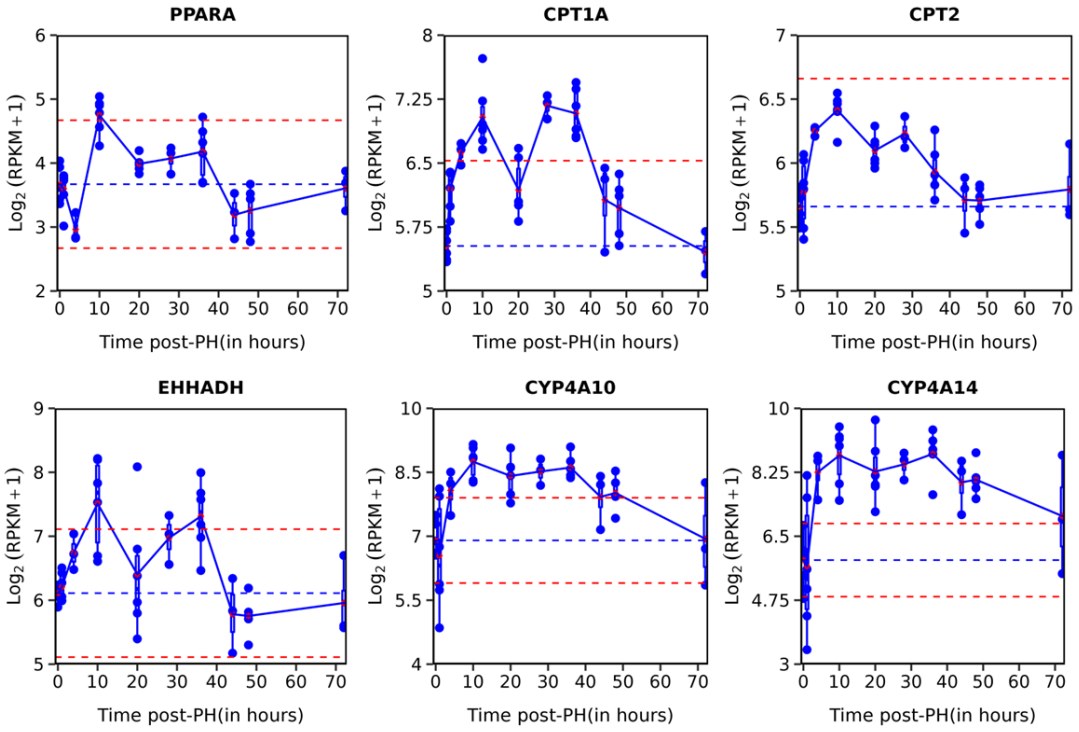

Figure S5: Expression profile of genes in β oxidation pathway. The blue line represents the baseline (0 hours), and the red line represents a two-fold change with respect to the baseline.

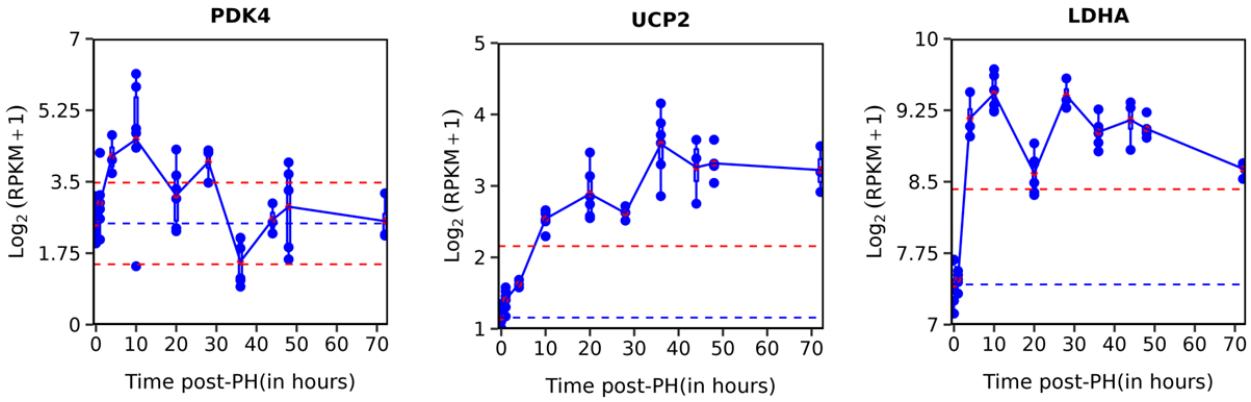

Figure S6: Expression profile of genes involved in the regulation of glucose and lipid metabolism and anaerobic glycolysis. The blue line represents the baseline (0 hours), and the red line represents a two-fold change with respect to the baseline.

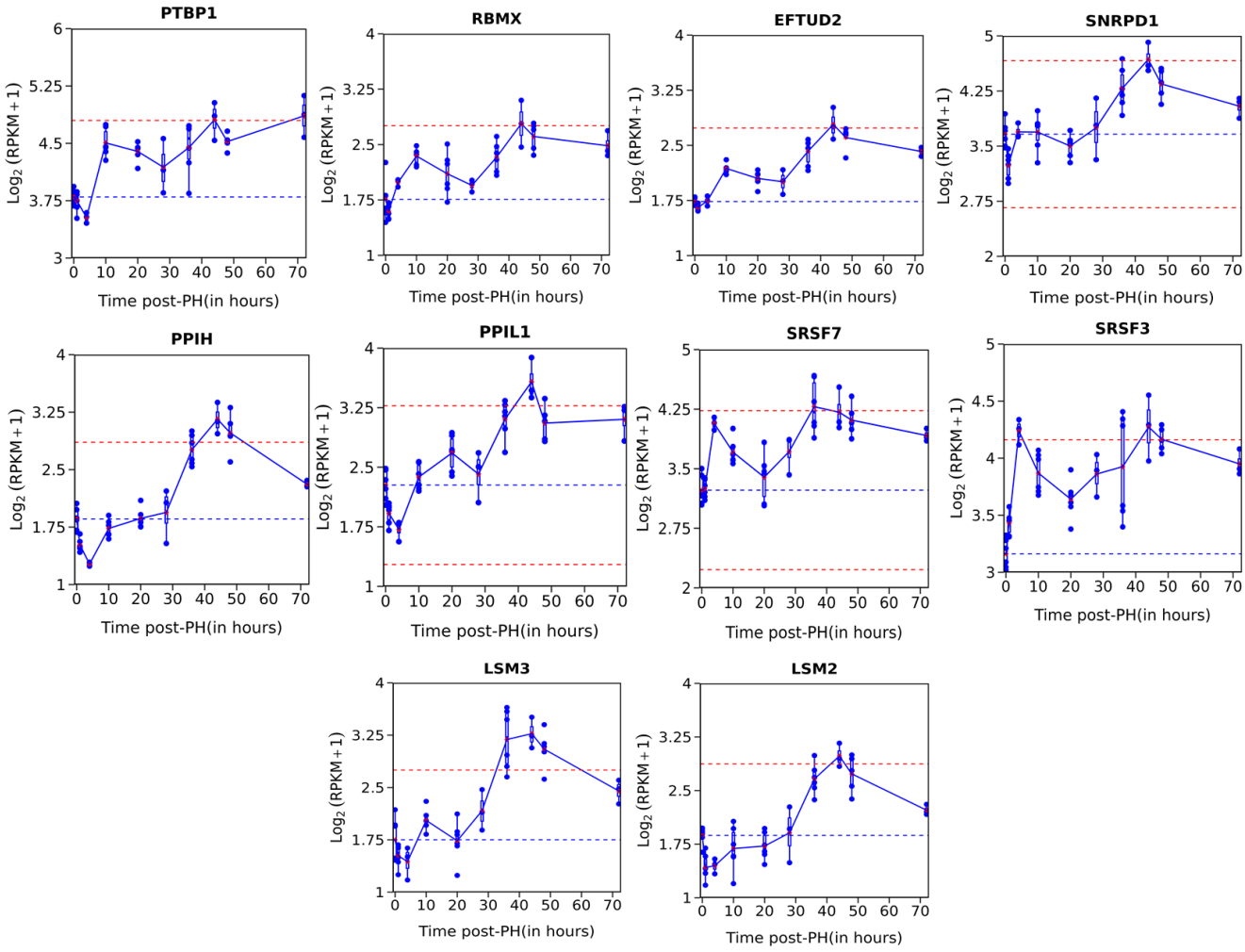

Figure S7: Expression profile of genes involved in the RNA splicing. The blue line represents the baseline (0 hours), and the red line represents a two-fold change with respect to the baseline.

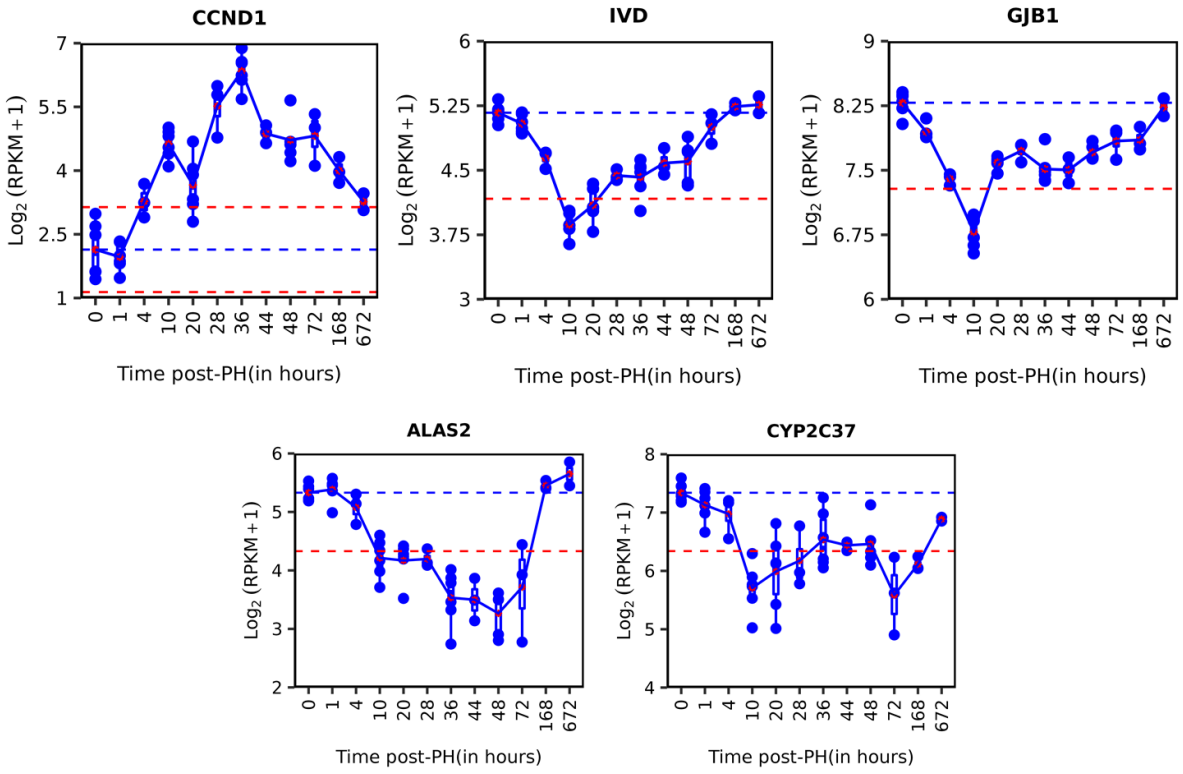

Figure S8: Transient activation of Cyclin D and inactivation of HNF4A (IVD, GJB1, ALAS2 and CYP2C37) targets during liver regeneration. The blue line represents the baseline (0 hours), and the red line represents a two-fold change with respect to the baseline.

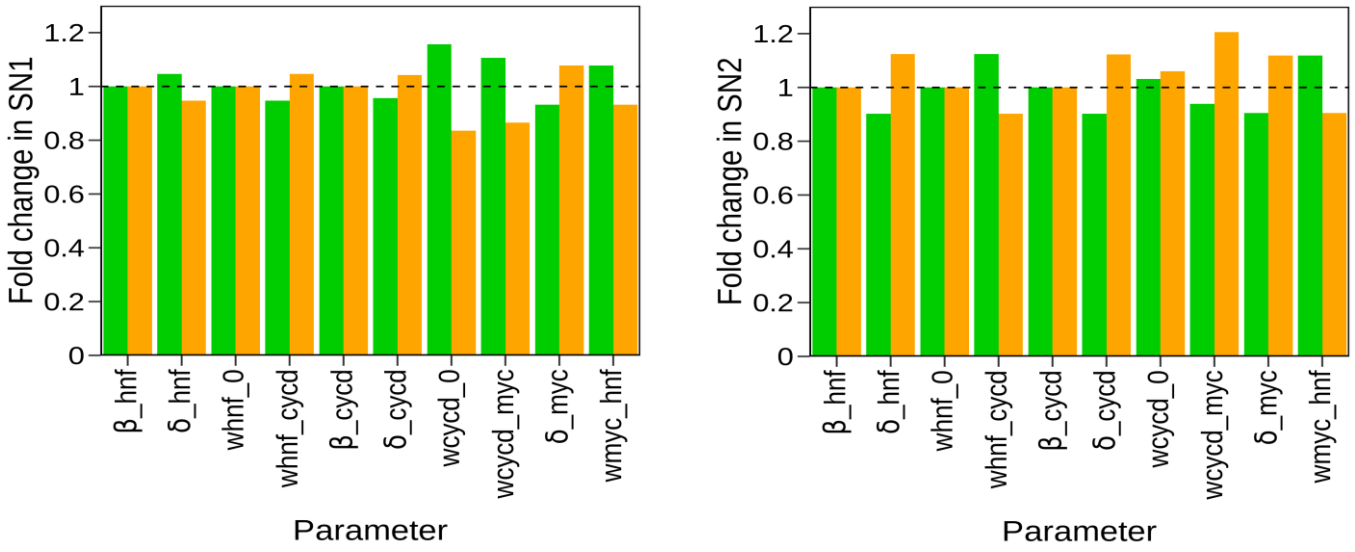

Figure S9: Effect of parameters on the HNF4A bistable response. The fold change in threshold for inactivation (saddle node, SN1) and re-activation (saddle node, SN2) of HNF4A with respect to variations in each parameter value is shown. The parameter values were varied in the range of 10% from default parameters value (as given in Table S3). Green and orange represent the increase and decrease in the corresponding parameter value. The fold change is SN (new)/SN (default).