Dear Julio Vera,

We would like to thank you and the reviewers very much for your work and for the constructive criticism, which helped us to improve our manuscript considerably. Please find below a detailed account of our responses to the reviewer’s questions and remarks, and how we have dealt with them in the revised version of our manuscript.

**Reviewer #2:** The manuscript "Pathobiochemical signatures of cholestatic liver disease in bile duct ligated mice" by Abshagen et al. aims at identifying markers for chronic liver disease progression by measuring physiological parameters as well as the expression of selected genes in mice following disrupted bile ligation over time up to 14 days.

The authors performed extensive correlation analyses between the expression data and the various hematological measurements and combined them in a "consensus correlation" value to predict, which genes and factors explain best the different disease phases.

Finally, they propose a decision tree based on the main markers they found to predict the disease stages. The manuscript is a nice example of a collaborative effort in experiment, analysis and interpretation, yielding interesting new insights into the chronic liver disease. Overall, the results are well explained and compared to the literature. In particularly the decision tree approach is a good idea to translate the experimental data into new biological insight of chronic liver disease progression.

**Nevertheless, the manuscript needs a major restructuring of the data analysis and a clear focus in its presentation. The reader is flooded with gene lists, information and interpretation, while the motivation and explanation of the analysis is underrepresented in the main text. All information is there, but scattered throughout the main text, methods section and supplement.**

*Answer: we thank the reviewer for the objective estimation of the manuscript. We rewrote and restructured substantial parts of the manuscript with the focus of a clear presentation of the motivation and explanation of the analysis (see introduction & method section statistical analysis). The gene lists were substantially reduced, and scattered information removed or unified in one place. Additional sections providing information/results about the time course correlation analysis and our decision tree approach were added.*

*Our main focus in this study was to generate time resolved multi-level knowledge on cholestasis disease progression, aiming to define markers representative for the different disease stages from initiation to severe fibrosis as possible candidates for diagnosis. We made this motivation much clearer in the introduction and results.*

**The manuscript would win substantially, if the authors focused on selected genes and results, better explained their analysis approaches in the main text and redrew their figures in a more intelligible format with readable fonts and structured information.**

*Answer: Large parts of the statistical analysis were redone. A central aim was dimension reduction of the high-dimensional data set. This was achieved via ANOVA filtering and subsequent focus on clusters of parameters showing similar time courses (hierarchical clustering of correlation matrix). Consequently, we focused on much less genes. The method section describing the statistical analysis was completely rewritten better, now explaining the analysis approach in more detail. Additional sections were added discussing the time course clusters and the decision trees in more detail. All figures presenting the correlation information, hierarchical clustering results and the decision trees, i.e. Figure 7, 8, 9, and 10, were completely redrawn with the latest results. We focused on readability and clear presentation of the high density information. Unified coloring schemes were applied to presented results on time course clusters (color coding cluster c1-c6), and the correlation information (red-white-blue) throughout all figures.*

*A specific intention of this study was the presentation of the time-dependent disease progression following bile duct ligation in a comprehensive manner. Thus, we explicitly wanted to systemize existing and newly acquired knowledge on the molecular biomarkers of cholestasis resulting in a multitude of information, resulting in lists of genes associated with certain time courses of disease development. Like mentioned above, we restructured and substantially shortened the gene lists in the results section to not overwhelm the reader.*

**Major Points**

**- The authors should consider dimension reduction methods such as multidimensional scaling or principal component analysis, when discussing the samples' expression patterns, use statistical tests, when assessing the significance of gene regulation over time and display ROC curves, when discussing their decision tree model. This would reduce lengthy description of the data and provide better overview on the dynamic behavior of the system.**

*Answer: As suggested by the reviewer we applied dimension reduction methods to discuss the time course patterns: We filtered significant factors via ANOVA (reduction from 153 to 90 factors) followed by clustering analysis to find relevant time courses after BDL (reduction from 90 factors to 6 clusters).*

*Statistical tests used to test for significance of expression changes were: 1) ANOVA tests, corrected for multiple testing on the factors with Holms, to find significantly changed genes after BDL, 2) two-tailed unpaired t-tests (Welch correction for nonhomogeneity of variance) were used to test for up-/down-regulation between 0h and 6h.*

*The decision trees are now based on established methods for fitting regression trees (CART models), instead of the previous separator approach. To show the predictive performance of the decision tree we are now showing the prediction results for all time course classes (Figure 10B). ROC curves are not applicable for regression trees, so we selected this alternative approach to communicate the model performance. Prediction performance was calculated with multiple test data sets: mean cluster data, single factor combinations & best gene single factor combination, and random two factors from each cluster.*

*The robustness of the fitted decision tree was tested with a leave-one-out approach, which showed robustness of the predicted time classes and good predictive performance on the left out data (Supplement 2).*

*Changes in the manuscript: Additional sections were added in the manuscript (methods & results) describing the performed dimension reduction, the statistical analysis, regression trees, and calculation of the predictive performance of the regression tree. All analyses are presented in Supplement 2.*

**- For the qPCR data the authors used a single Gene, Gapdh, for normalization, which possibly results in noisy dCT (delta CT) values. There is no guarantee that the expression of this gene remains constant across the samples, given the severe impact of BDL and the measurement time of 14 days. Usually, dCT values are normalized to two control genes, also e.g. 18S and/or Hprt1. The authors need to check and show the behavior of Gapdh and that their normalization approach does not affect their results.**

*Answer:* *We totally agree with the reviewer. Parallelization of qPCR such as by the microfluidic Taqman Fluidigm Biomark Platform enables evaluation of multiple transcripts in the samples treated under various conditions. Despite of advanced technologies, correct evaluation of the measurements remains challenging. We wanted to apply ΔΔCt method and therefore assessed expression of two reference genes in order to find the stable one. We accurately controlled the expression levels of Gapdh and Actb over the time period and their variability between different Fluidigm runs and chose Gapdh for normalization due to its stability. We assure that the normalization approach does not affect the main results of our evaluation.*

*Changes in the manuscript: none*

**- The authors state a delta delta CT of approx. 17 between IL28b and Gapdh (Fig. 6K, 5d). This is a rather large difference and would mean that IL28b appears around the qPCR cycle 5 already. This should be checked. What is the primer efficiency for IL28b? Maybe it is unusually high.**

*Answer:* *Indeed the Ct value of the IL28b expression is very low. All gene expression assays used in this study were purchased commercially as validated assays by Applied Biosystems. We analyzed the amplification efficiency of the IL28b assay on our own using the dilution series of a control cDNA sample and could reproduce the data provided by the manufacturer (E ranged 100± 5%).*

*Changes in the manuscript: none*

**- The benefit of the consensus score needs to be better discussed, e.g. the choice of different weights seems rather arbitrary. In supplement2, 1.2 Consensus correlations the authors state that "the correlation of time average has 4 times higher weight". Where does this number come from? Why are the correlations separated in positive and negative parts and then the smaller part are ignored? A lot of information is lost this way. The score should be normalized to be able to compare in between factors.**

*Answer: The consensus score was completely removed from the analysis. All correlation calculations are performed based on an established correlation score for time-course gene expression data (YS3).*

*Changes in the manuscript: Methods, results, figures, supplement were rewritten to according to the results based on YS3 score.*

**- Why do the authors use a consensus measure to perform clustering (Fig. 7)? Why do they not use all correlation data for each factor as a matrix, and then apply clustering method on this matrix instead?**

*Answer: Clustering is now performed like suggested on the complete correlation matrix based on hierarchical clustering with complete linkage (See Figure 7)*

*Changes in the manuscript: Methods, results, figures, supplement were rewritten to describe the new approach and results.*

**- The authors used the Pearson Correlation, which can be easily influenced by outliers. In particular in combination with the noisy qPCR data this can lead to spurious correlations. The authors should check for consistency of their results by using either more robust correlation measures like Spearman correlation or by low-pass filtering their data before performing the analysis.**

*Answer: We checked for the influence of outliers, and indeed Spearman correlation proved much more robust against them (Supplement 2). Consequently, all components of the used correlation score (YS3) were based on Spearman correlation (S\*, A\*\*).*

*Changes in the manuscript: The Spearman bases YS3 approach is described in detail in the methods in Supplement 2.*

**- The description of the separator approach in the manuscript is very wordy and formulas are necessary to understand what has been done.**

*Answer: The separator approach was completely dropped. We now fit of decision tree model based on established statistical methods, i.e. CART models for regression trees, described in the method section.*

*Changes in the manuscript: All sections regarding separators were removed from the manuscript.*

**- Figures 8 and 9: Significance and consensus scales are not really readable.**

**It is hard to distinguish between the not significant ones (0.1) and the significant ones (<0.05). For example, few different colors would be more efficient than yellow gradient.**

**Also the correlation scale should be from -1 to 1.**

*Answer: The information of Figures 8 and 9 was completely restructured and is not presented in a much cleaner way (see also discussion above). Consensus score is no longer used.*

*Changes in the manuscript: All sections affected were rewritten to account for the new figures.*

**- Overall, figure legends need to be improved, as they lack sufficient annotation**

**to understand what is displayed.**

*Answer:* *agree*

*Changes in the manuscript: All figure legends have been carefully revised to improve readability and to supplement missing indications. Changes are marked with red font and underlined. For the newly generated Figures 7-10 the figure legends were completely rewritten.*

**Minor Points**

**- All gene symbols should be written consistently with small letters and a capital**

**first letter throughout the text and figures, e.g. use Gapdh instead of GAPDH.**

*Answer:* *As recommended, the manuscript text has been corrected for the consistently notation of all gene symbols with small letters and a capital first letter.*

*Comment: When talking about the gene we use Gapdh, talking about the protein GAPDH is used.*

**- in the derivation of the consensus correlation the authors use a rather unusual -log100 transformation. is this true or a typo?**

*Answer:* *Consensus correlation is no longer used.*

**- The relative expression of the genes in Fig. 6 should be displayed in log2 scale and the domain ranges should be the same for all genes to make the changes in gene expression comparable.**

*Answer:* *As suggested by the reviewer, we have modified figure 6 and now display gene expression in log2 scale.*

**- Most likely, the authors used a log2 scale in Fig. 5, but annotation of the color bars is missing and needs to be added.**

*Answer: Annotation of the color bars in Fig. 5 was already included in the original version of the manuscript, but has now been improved.*

**- Page 9 lines 17 and 19, Page 14 line 19: Figs. 7 and 8 are actually Figs. 8 and 9**

*Answer:* *We corrected these mistakes.*

**- Place Figure 7 after Figure 8, as the former is based on the latter**

*Answer: The clustering dendrogram is presented with the correlation matrix in Figure 7.*

**- The abbreviations in the leaf names of the circular tree in Fig. 7 are nowhere explained in the main text and need to be added.**

*Answer: We provide a table listing for all gene identifiers the gene name and annotation to UniProt in supplement 2. This information is added to the legend of Figure 7 containing the short names.*