

Western Diet induces MASLD and major transcriptional reprogramming in the liver-gut axis of early zebrafish larvae

Busra Korkmaz^{1, #}, Thomas Darde^{2, #}, Rana Acar³, Helin Ozaktas-Kececi^{1,4}, Serap Birincioglu¹, Yulia A. Nevzorova^{5,6,7}, Ozlen Konu^{3,*}, Gulcin Cakan-Akdogan^{1,4,*}

¹ Izmir Biomedicine and Genome Center, Izmir, 35340, Türkiye

² SciLicism, 35 boulevard du Portugal, 35200, Rennes, France

³ Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, Ankara, Türkiye

⁴ Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, 35340, Izmir, Türkiye

⁵ Department of Immunology, Ophthalmology and ENT, Complutense University School of Medicine, Madrid, Spain

⁶ Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Madrid, Spain

⁷ Instituto de Investigación Sanitaria Gregorio Marañón (IiSGM), Madrid, Spain

co-first authors

*Co-correspondence: konu@fen.bilkent.edu.tr, gulcin.cakan@ibg.edu.tr

Supplementary Methods

BRB-seq Data Analysis and Quality Control

Libraries were prepared by bulk RNA barcoding and sequencing (BRB-seq) and multiplexed on a single plate to avoid batch effects. Raw BRB-seq counts were filtered on low-expressed genes and log-transformed. Sample distributions were inspected for outliers, a batch-effect correction was applied, and the count matrix was then normalized. The same expression matrix is displayed at the four successive steps of the workflow in Fig S3 and Fig S11 for the two experiments. No library had to be discarded in the BEY dataset. Distributions were already comparable between samples before normalization and became fully aligned after normalization (median 2.585 to 2.301; mean 2.959 to 2.473), indicating that the normalized matrix is free of sample-level technical bias and suitable for differential expression analysis. Three biological replicates (each comprising ~25 pooled larvae) were analyzed per condition and time point: RD 9 dpf (n = 3), BEY 9 dpf (n = 3), control 12 dpf (n = 3) and BEY 12 dpf (n = 3), for a total of twelve libraries. Features were filtered using a minimum count threshold of 3 and a background

quantile cutoff of 0.5. Sample quality was assessed by principal component analysis (PCA), UMAP and inter-replicate correlation (Fig. S4). In the PCA, the first two components explained 52.7% and 13.5% of the variance, and samples separated primarily by developmental time (9 vs 12 dpf) and by diet (control vs BEY), with no clustering attributable to technical batch, as all samples were processed on a single plate. Pairwise correlations between all samples were computed on the normalized expression matrix and the resulting matrix was clustered hierarchically in order to detect discordant replicates (Fig. S5). All pairwise correlations exceed 0.992, indicating a high global reproducibility between libraries. No sample behaves as an isolated outlier in the clustering; residual variability between replicates is small compared with the differences between conditions, and all 12 libraries were retained for differential expression analysis.

In the WD/CD experiment three biological replicates (each comprising ~25 pooled larvae) were analyzed per condition and time point: RD 9 dpf (n = 3), CD 9 dpf (n = 3), WD 9 dpf (n = 3), RD 12 dpf (n = 3), CD 12 dpf (n = 3), WD 12 dpf (n = 3) for a total of eighteen libraries. One CD library at 2 dpf (BK_73) showed a collapsed expression distribution, was identified as an outlier and was excluded; 17 libraries were retained. After its removal the mean value rose from 2.842 to 3.004 and the remaining distributions were homogeneous across diets and time points, and fully aligned after normalization (median 2.365, mean 2.539). Features were filtered using a minimum count threshold of 3 and a background quantile cutoff of 0.5. Sample quality was assessed by principal component analysis (PCA), UMAP and inter-replicate correlation (Fig. S12). Samples are distributed along dimension 1 (49.7% of the variance) according to experiment/time point, while the three diets (cont, WD and CD) form partially overlapping groups on both the PCA and the UMAP, consistent with the partial overlap of their transcriptomic responses. Replicates of a given condition remain grouped, supporting the quality of the dataset. Pairwise correlations between all samples were computed on the normalized expression matrix and the resulting matrix was clustered hierarchically in order to detect discordant replicates after exclusion of the outlier library (Fig. S13). All pairwise correlations exceed 0.99. The clustering splits the samples into two main blocks, separating control libraries from WD- and CD-fed libraries, in agreement with the PCA/UMAP projections. No remaining library shows an aberrant correlation profile, confirming that the retained samples are of homogeneous quality.

Differential gene expression analysis

Statistical filtering and differential expression analysis were conducted in R v4.4.1. Principal component analysis (PCA) was performed using the *FactoMineR* package¹. Differentially expressed genes (DEGs) were identified using the *LIMMA* package², with multiple testing correction via the Benjamini–Hochberg method.³ A global expression background cut-off was defined as the median gene expression across all samples. Adjusted *p*-values were used to identify genes commonly modulated across

comparisons, whereas raw *p*-values highlighted uniquely regulated genes within specific conditions. Heatmaps of DEGs were generated using the *pheatmap* package in R. Because the BEY and WD/CD experiments were generated and processed as independent datasets, differential-expression thresholds were optimized separately for each. For every dataset, the optimal fold-change cutoff was determined using the elbow method, yielding $FC \geq 1.2$ (adjusted $p < 0.05$) for the BEY dataset and $FC \geq 1.3$ (adjusted $p < 0.01$) for the WD/CD dataset; the reported qualitative contrasts therefore rest on dataset-specific optimal thresholds rather than a single arbitrary cutoff.

K-means clustering was performed independently on the row-scaled (z-score) expression matrices differentially expressed gene (DEG) sets, in experiments. For each value of K, k-means clustering was run with $\text{iter.max} = 10,000$, and the resulting cluster centroids were hierarchically ordered using Pearson correlation-based hierarchical clustering to standardize cluster labels. Then K4 solution heatmap showing 4 clusters of genes that behave differently according to diet and/or timepoint were selected. Metrics supporting the four-cluster choice are shown in Fig. S7.

Functional analysis

An enrichment analysis was performed as described in Chalmel & Primig, 2008⁴. A specific annotation term was considered enriched in a gene cluster when the False Discovery Rate (FDR)-adjusted *p*-value (Fisher's exact probability) was ≤ 0.05 and the number of genes associated.

Liver and Intestine Specific Expression Analysis

To evaluate liver and intestine geneset, genes featured in the heatmaps for liver (derived from Gao et al., 2022 Supplementary Table 4), intestinal regions, and intestinal cell types (derived from Gao et al., 2022 Supplementary Table 23) were crosschecked against the Daniocell single-cell transcriptomic atlas for early zebrafish larvae up to 5 dpf (Sur et al., 2023) and provided in Supplementary Table S18. The table details the enrichment of these genes across endoderm-specific cell types with database hyperlinks, followed by their enrichment across all documented cell types. To categorize tissue enrichment based on the Daniocell database, a binary flagging system was implemented: liver genes were flagged as "1" if they demonstrated liver enriched expression in Daniocell, and "0" if not (Figure 7B; 44/75 flagged as liver enriched). A parallel approach was applied to the intestinal regions gene set, where genes were flagged if exhibited gut enriched expression in the Daniocell dataset (Figure 7C; 70/100 flagged as gut enriched). Although 100% concordance in enrichment flagging was not observed, the concordance ratios from cross-checking with the Daniocell single-cell atlas strengthened our tissue-localized analysis approach. Enrichment details across all Daniocell cell types were incorporated for genes associated with intestinal cell types (Figure 7D). Genes previously reported as intestine-enriched (derived from Gao et al., 2022 Supplementary Table 3) were denoted with an asterisk (*) throughout the tables.

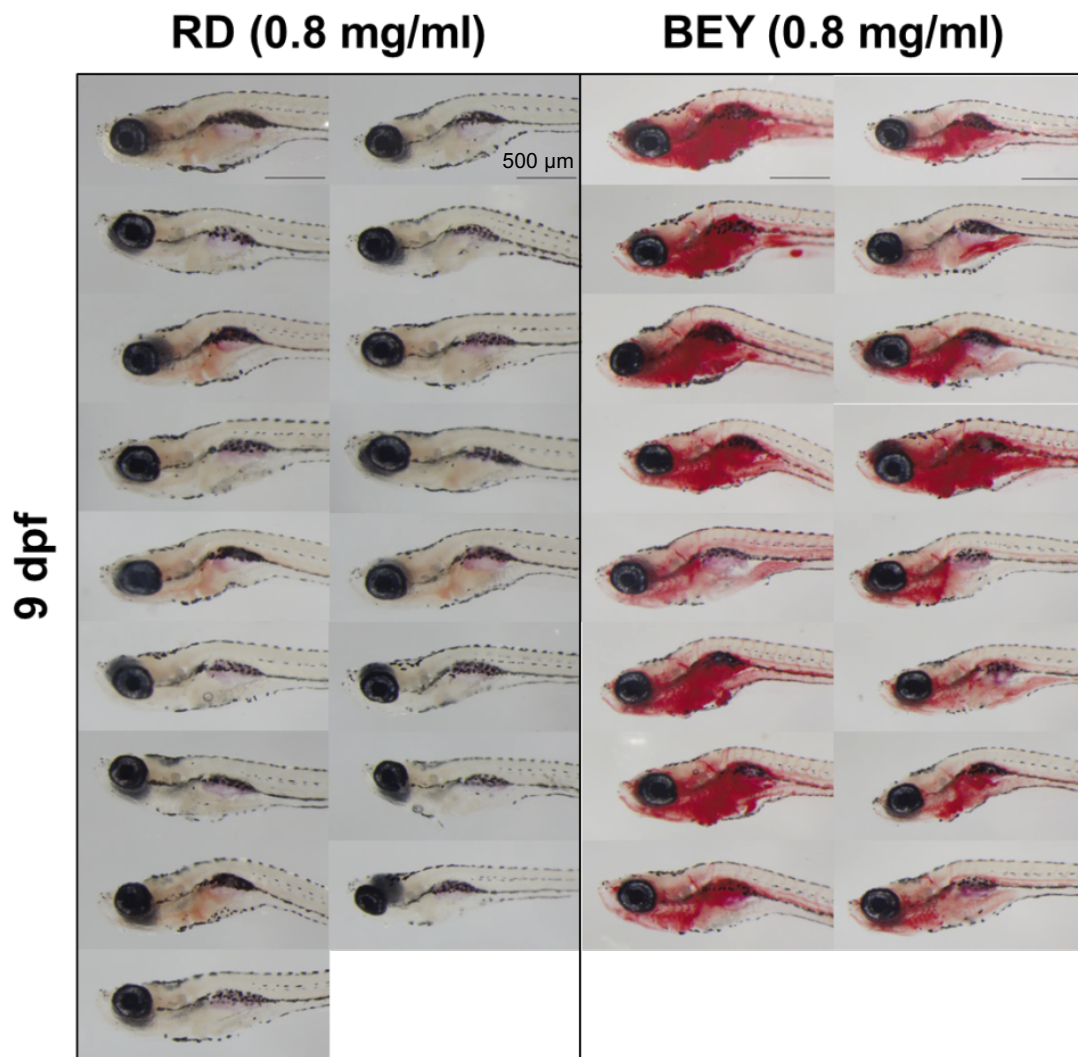


Fig. S1. Images of all Oil Red O (ORO) stained 9 dpf BEY experiment. AB strain zebrafish larvae were fed with regular diet (RD) or boiled egg yolk (BEY) from 5 days post fertilization (dpf) to 9 dpf. The larvae were fasted overnight (O/N) before staining to ensure digestion of ingested food.

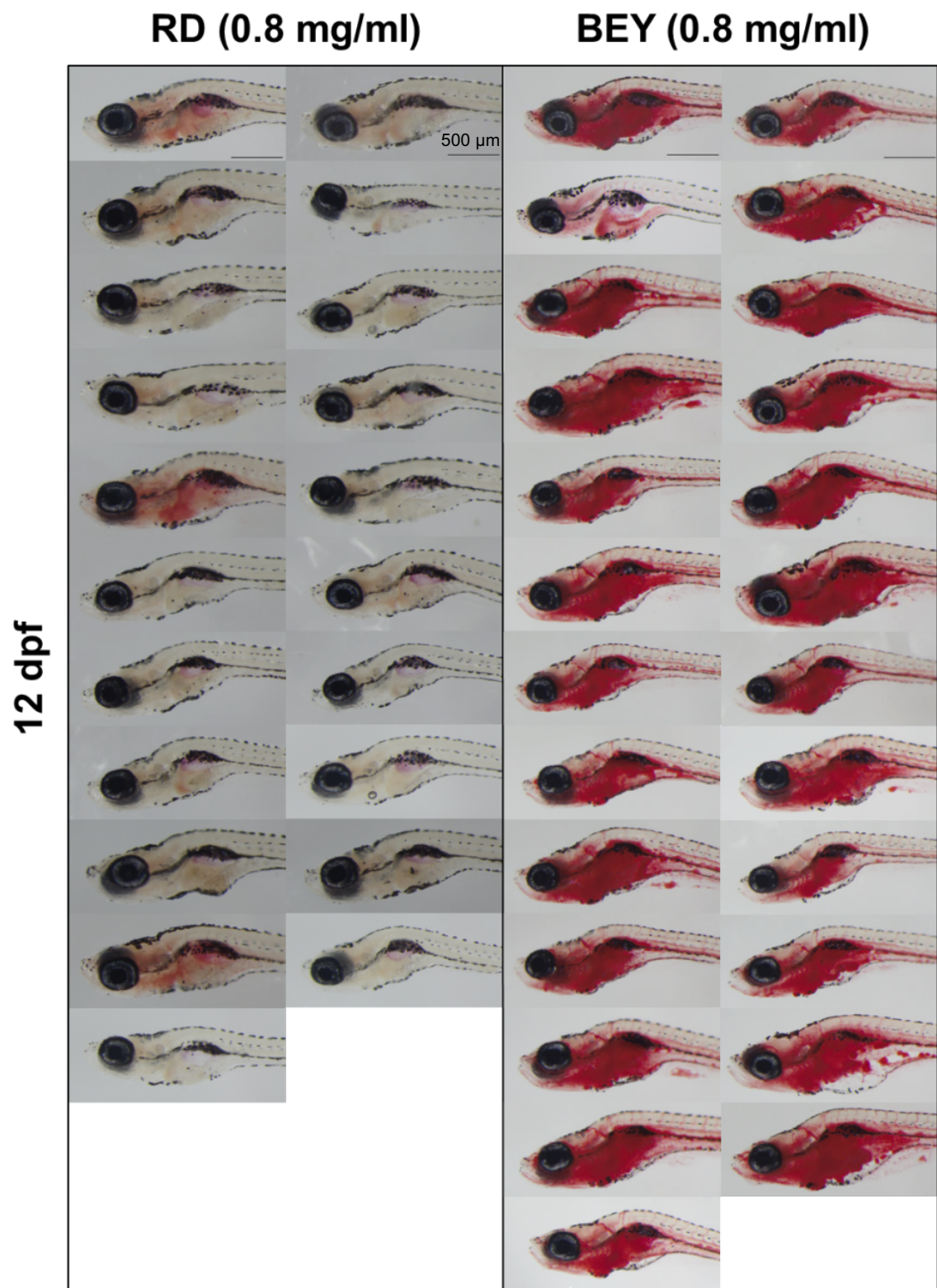


Fig. S2. Images of all Oil Red O (ORO) stained 12 dpf BEY experiment. AB strain zebrafish larvae were fed with regular diet (RD) or boiled egg yolk (BEY) from 5 days post fertilization (dpf) to 12 dpf. The larvae were fasted overnight (O/N) before staining to ensure digestion of ingested food.

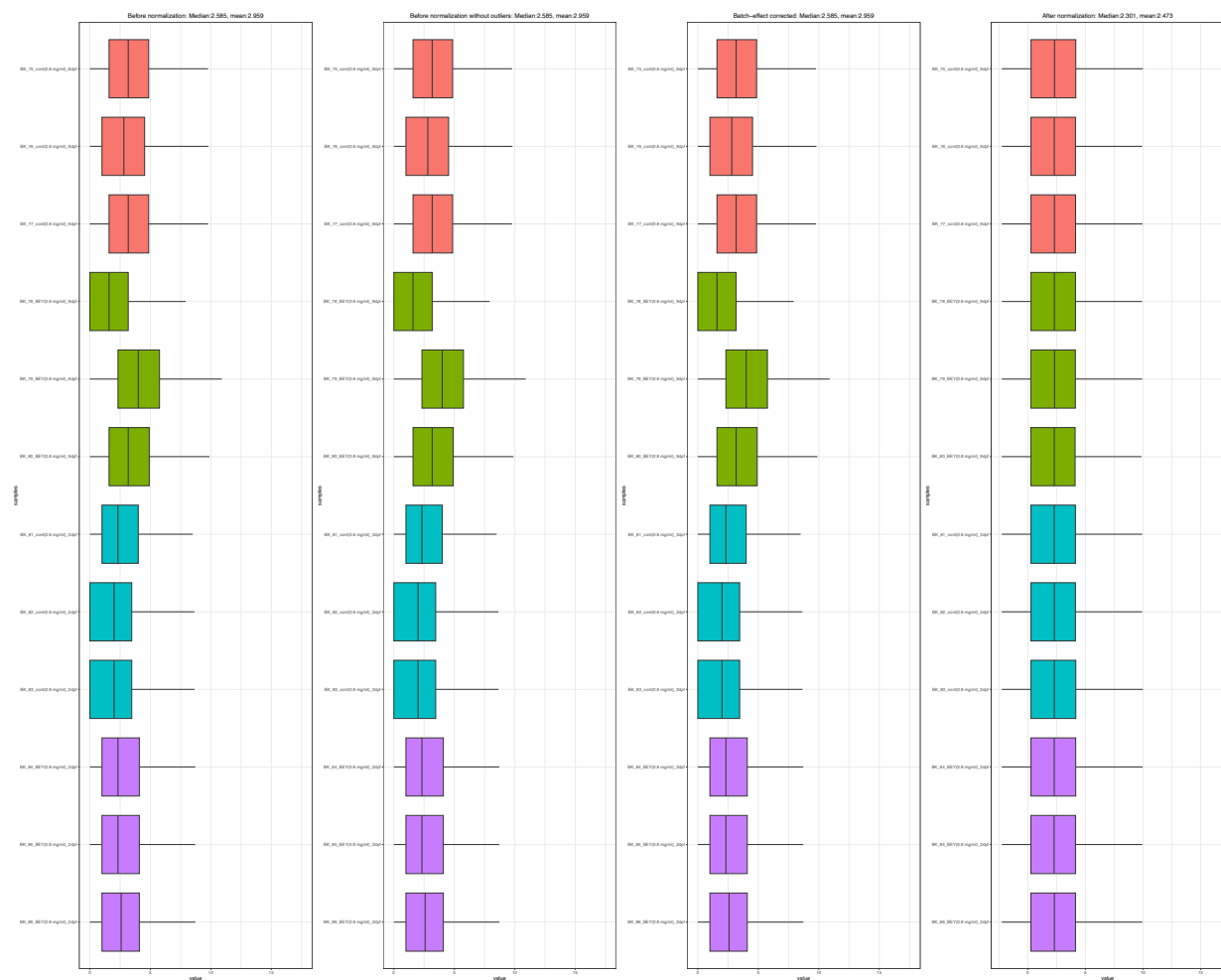


Fig. S3. Distribution of BRB-seq expression values across samples at each pre-processing step in the BEY dataset. Box plots of log-transformed expression values for the 12 libraries of the BEY experiment (BK_75–BK_86; RD and BEY 0.8 mg/ml, 9 dpf and 12 dpf). *From left to right:* before normalization (median 2.585, mean 2.959), before normalization after outlier removal, after batch-effect correction, and after normalization (median 2.301, mean 2.473). Colors indicate diet × time point groups: red (RD 9 dpf), green (BEY 9 dpf), turquoise (RD 12 dpf), violet (BEY 12 dpf).

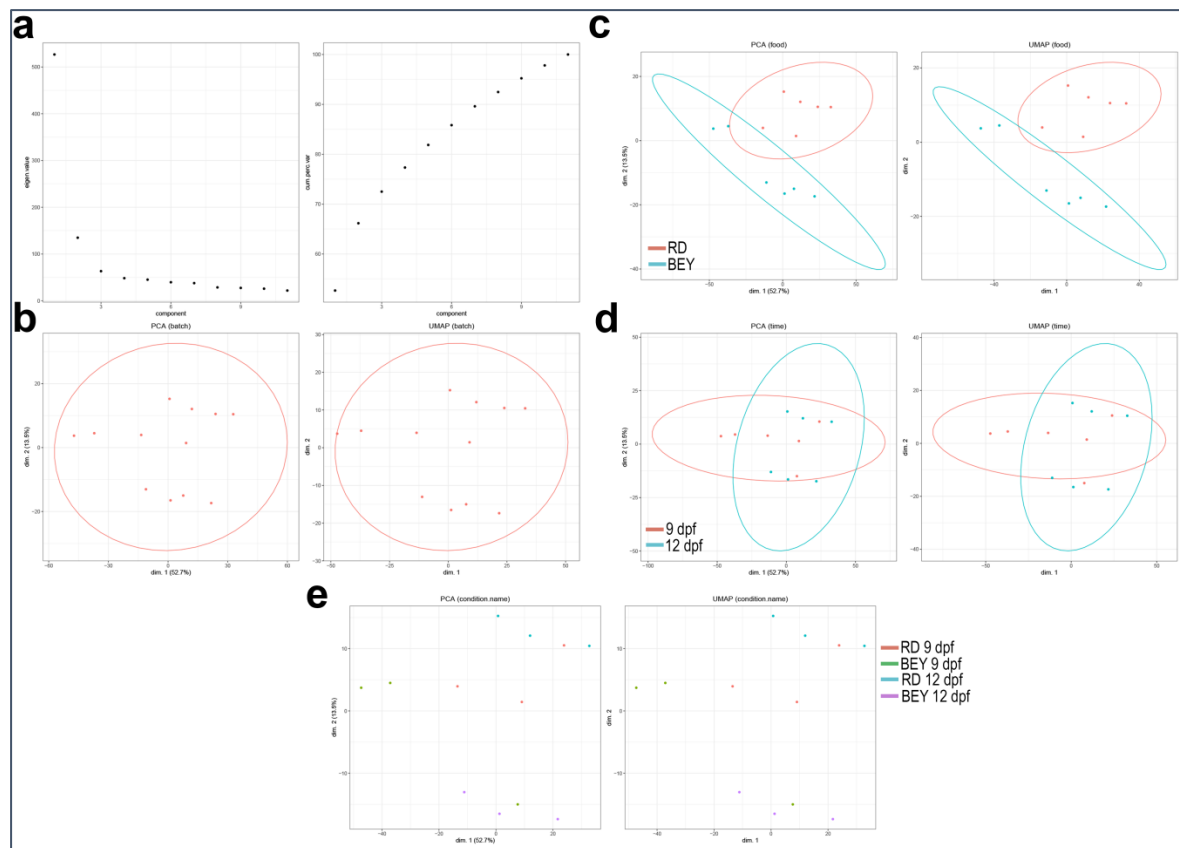


Fig. S4. PCA and UMAP plots of BEY vs RD experiment. Unsupervised sample projections of the BEY dataset. **(a)** scree plot of eigenvalues and cumulative percentage of explained variance per principal component. **(b-e)** PCA (left column) and UMAP (right column) projections of the 12 BEY libraries, colored successively by **(b)**sequencing batch, **(c)** diet (RD vs BEY 0.8 mg/ml), **(d)** time point (9 dpf vs 12 dpf) and **E)** condition. Ellipses represent the 95% confidence region of each group. The first two principal components account for 52.7% and 13.5% of the total variance.

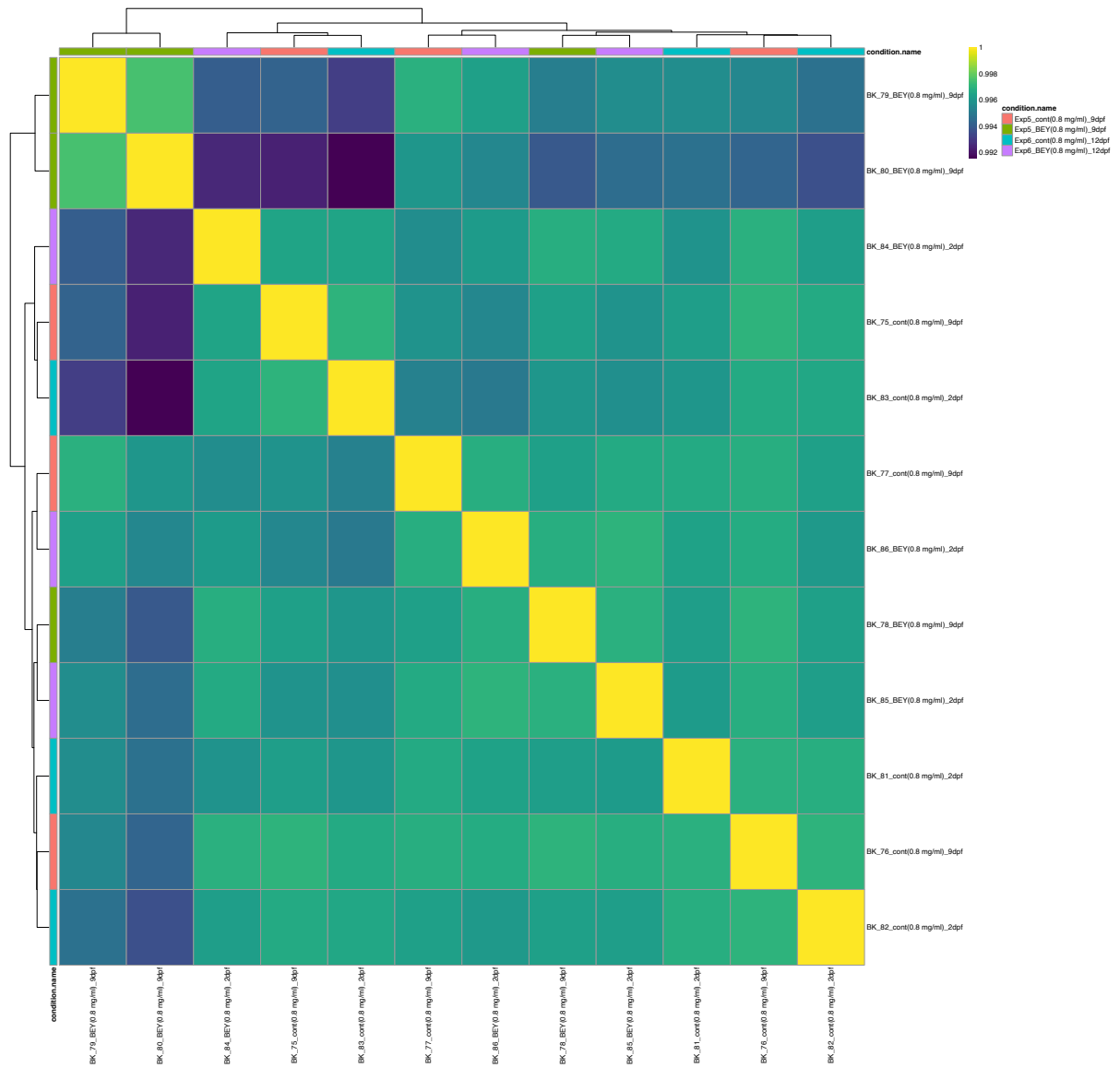


Fig. S5. Sample-to-sample correlation heat map of the BEY dataset. Pairwise correlation coefficients between the 12 BEY libraries, computed on the normalized expression matrix and displayed with hierarchical clustering on rows and columns. The color scale ranges from 0.992 to 1; the annotation bars indicate the condition (Exp5_cont(RD)_9dpf, Exp5_BEY_9dpf, Exp6_cont(RD)_12dpf, Exp6_BEY_12dpf).

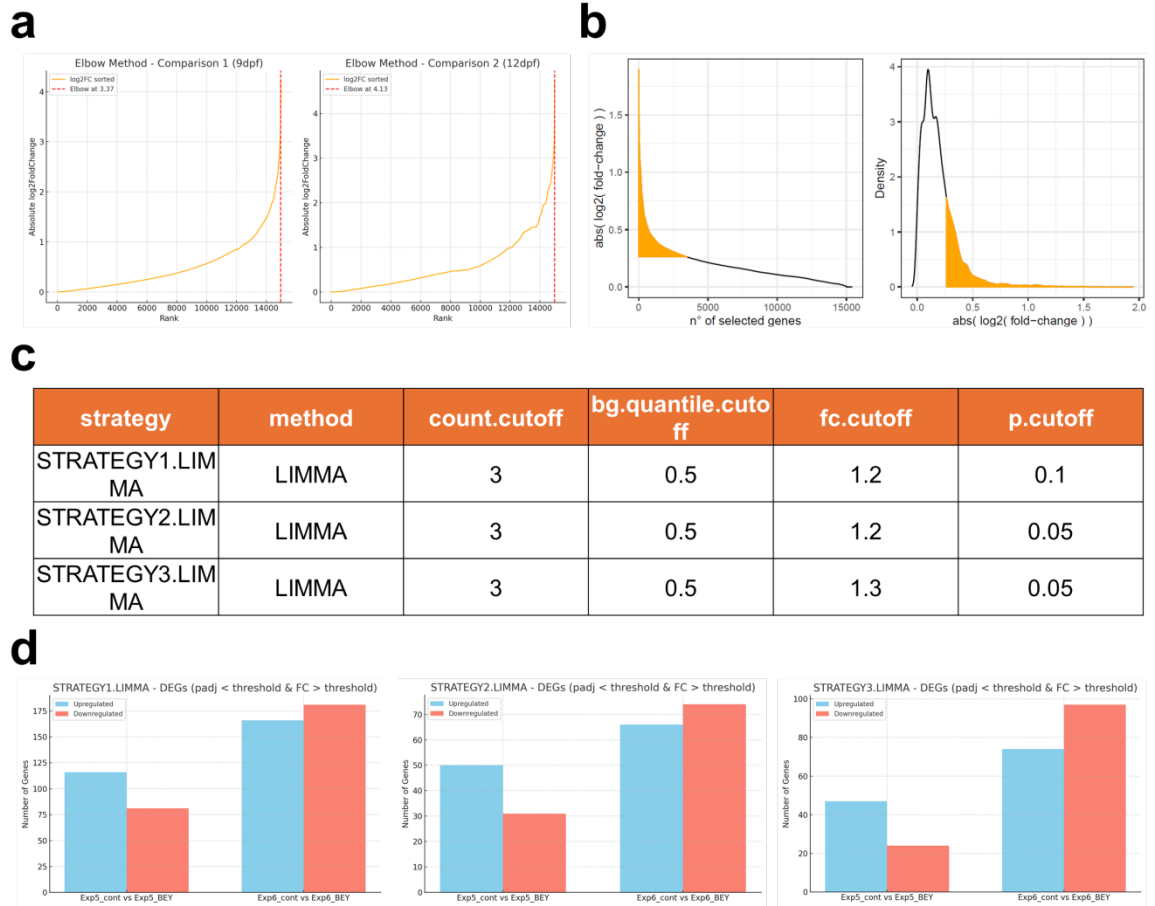


Fig. S6. (a) The plots obtained based on elbow method approach to determine an optimal log₂ fold-change (FC) threshold that works reasonably well for both 9 days post fertilization (dpf) and 12 dpf comparisons for boiled egg yolk (BEY). x-axis shows rank, while y-axis shows absolute log₂FC. (b) The plots that show elbow points in two representations: absolute log₂FC with respect to number of selected genes and density with respect to absolute log₂FC. (c) Different strategies for statistical filtration while analyzing differentially expressed genes (DEGs), each with different cutoff values for log₂FC and adjusted *p* value (*p*-adj). (d) The graphs showing the number of upregulated and downregulated genes based on these strategies when BEY was compared to regular diet (RD) at 9 and 12 dpf. Blue and coral represent upregulated and downregulated genes, respectively.

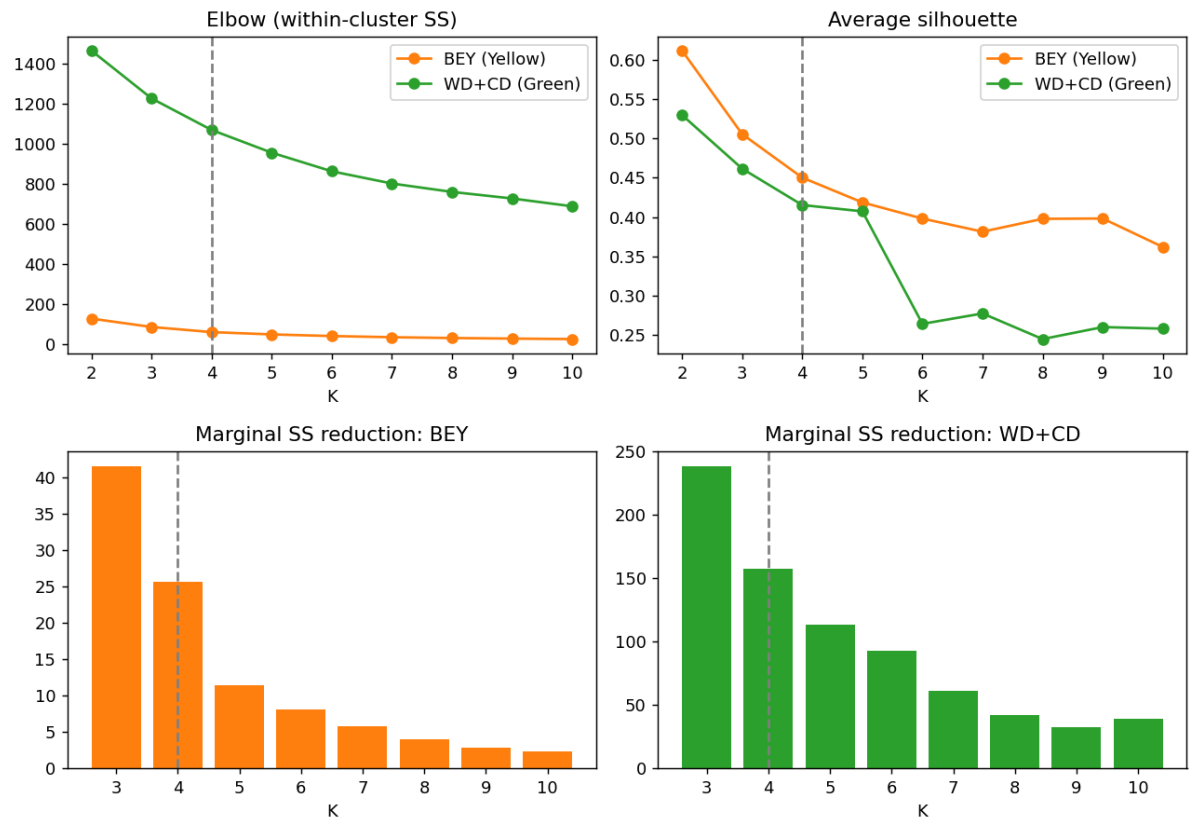


Fig. S7. Metrics supporting the four-cluster ($K = 4$) solution used for K-means clustering of differentially expressed genes. (A) Total within-cluster sum of squares (SS) versus K , (B) average silhouette width versus K (Euclidean distance on the scaled matrix), and (C, D) marginal within-cluster SS reduction from $K-1$ to K for the BEY (C) and WD+CD (D) datasets; dashed lines mark the retained solution, $K = 4$. Because the SS curves show no single sharp break, K was selected by combining marginal variance gain, which flattens after $K = 4$, with average silhouette width, which is maximal at $K = 2$ as expected for DEG sets dominated by the direction of regulation, remains above 0.40 in both datasets up to $K = 4$, and collapses for WD+CD at $K \geq 6$ (about 0.26); $K = 4$ was therefore retained for both datasets to allow direct comparison of the BEY and WD response architectures. (orange, $n = [X]$ DEGs; Fig. 3A) and western diet (WD) plus control diet (CD) versus RD at the same time points (green, $n = [Y]$ DEGs; Fig. 5A), with the number of clusters K varied from 2 to 10 (**Instart = 25**), fixed seed).

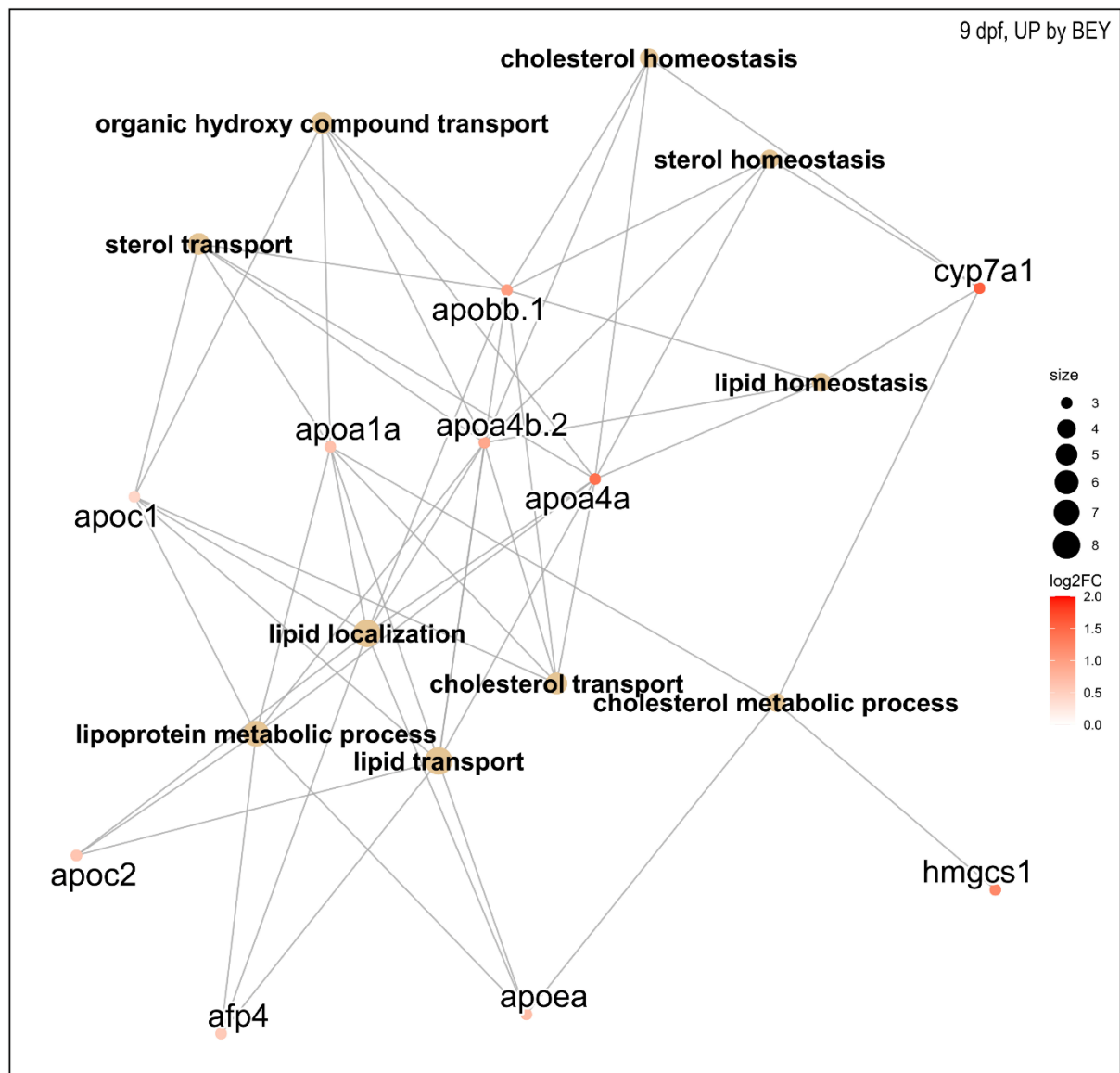


Fig. S8. BEY 9 dpf upregulated processes. Context-specific gene regulatory network (cNet) representation for the biological processes and related genes upregulated by boiled egg yolk (BEY) at 9 days post fertilization (dpf) ($p\text{-adj} < 0.05$). Fold change was shown by red for upregulation. The size of the circles for each gene refers to the gene count. 11

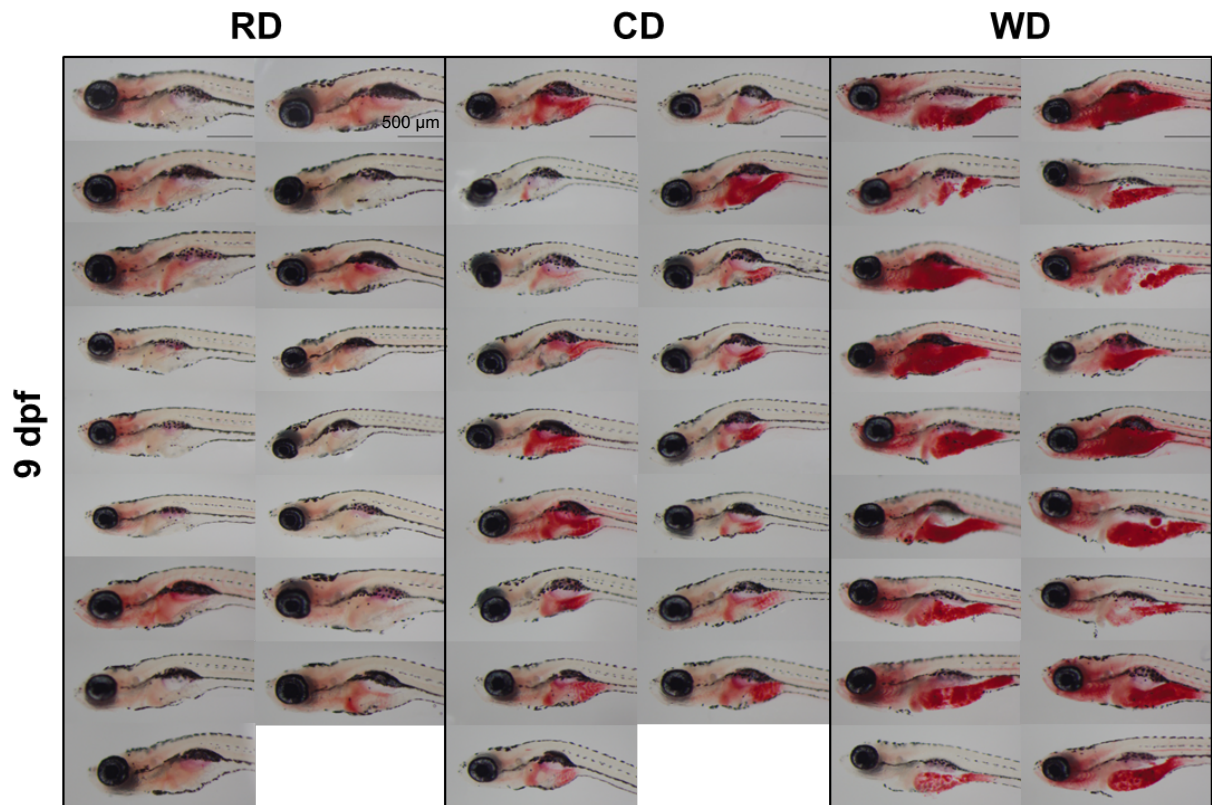


Fig. S9. Images of all Oil Red O (ORO) stained 9 dpf WD/CD experiment. AB strain zebrafish larvae were fed with regular diet (RD), high fiber control diet (CD) or western diet (WD) from 5 days post fertilization (dpf) to 9 dpf. The larvae were fasted overnight (O/N) before staining to ensure digestion of ingested food.

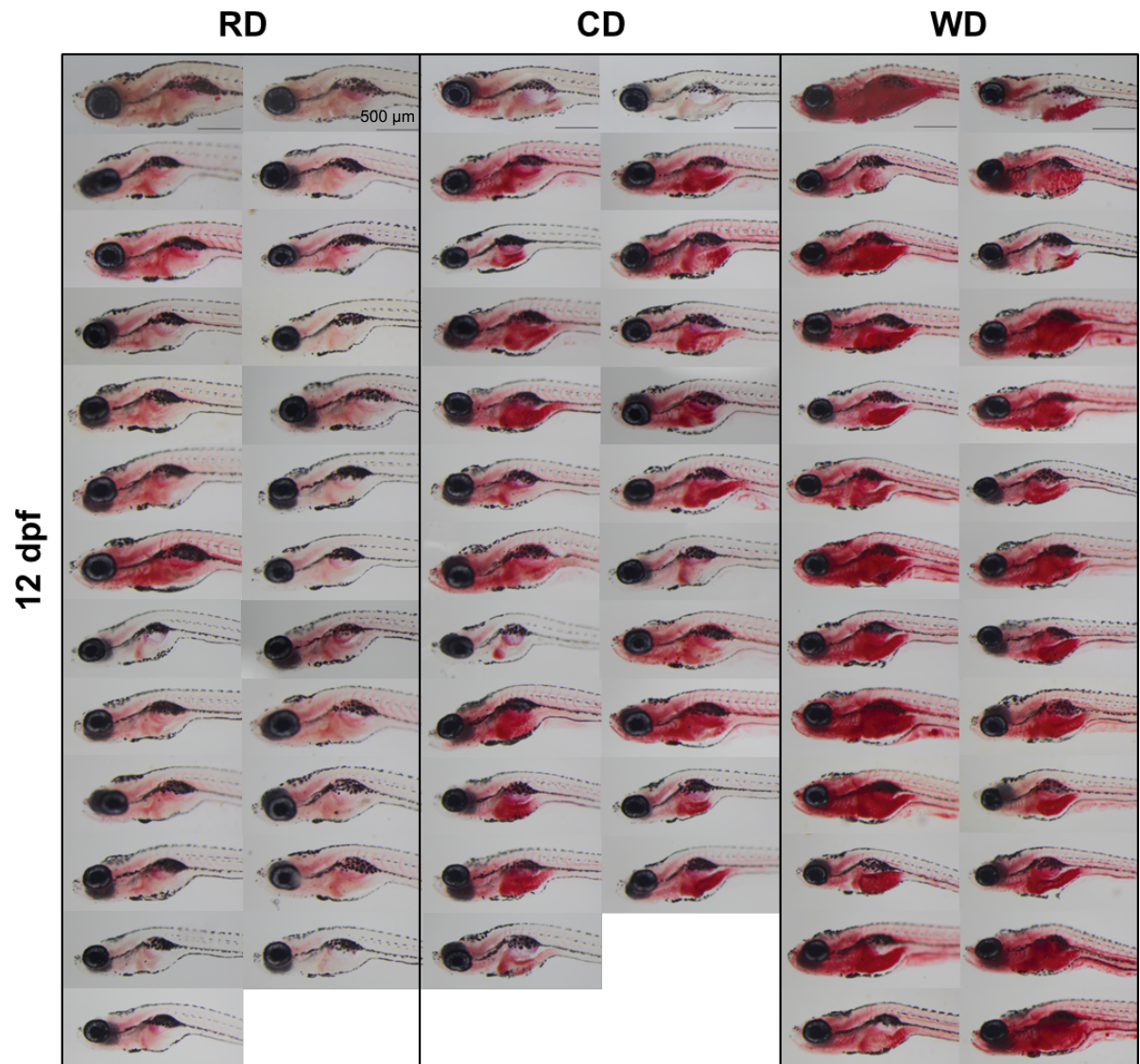


Fig. S10. Images of all Oil Red O (ORO) stained 12 dpf WD/CD experiment. AB strain zebrafish larvae were fed with regular diet (RD), high fiber control diet (CD) or western diet (WD) from 5 days post fertilization (dpf) to 12 dpf. The larvae were fasted overnight (O/N) before staining to ensure digestion of ingested food.

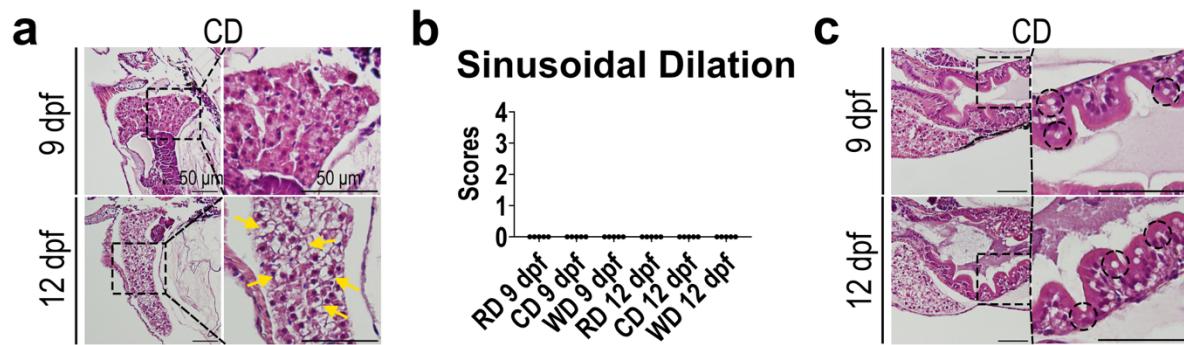


Fig. S11. Hematoxylin & eosin (H&E) staining of larvae sections that belong to CD group. (a) The images of liver sections, (b) scoring of sinusoidal dilation and (c) the images of gut sections are shown. The regions marked with dashed lines are shown more closely on the right. Arrows and dotted circles in the close-up images show ballooning and lipid droplets, respectively. For scoring, non-parametric Kruskal-Wallis test was applied (* p value < 0.05, ** p value < 0.01, *** p -adj < 0.001, **** p -adj < 0.0001). Scores: 0 = none, 1 = mild, 2 = medium and 3= high. The comparison of CD or WD with RD was performed separately for different time points.

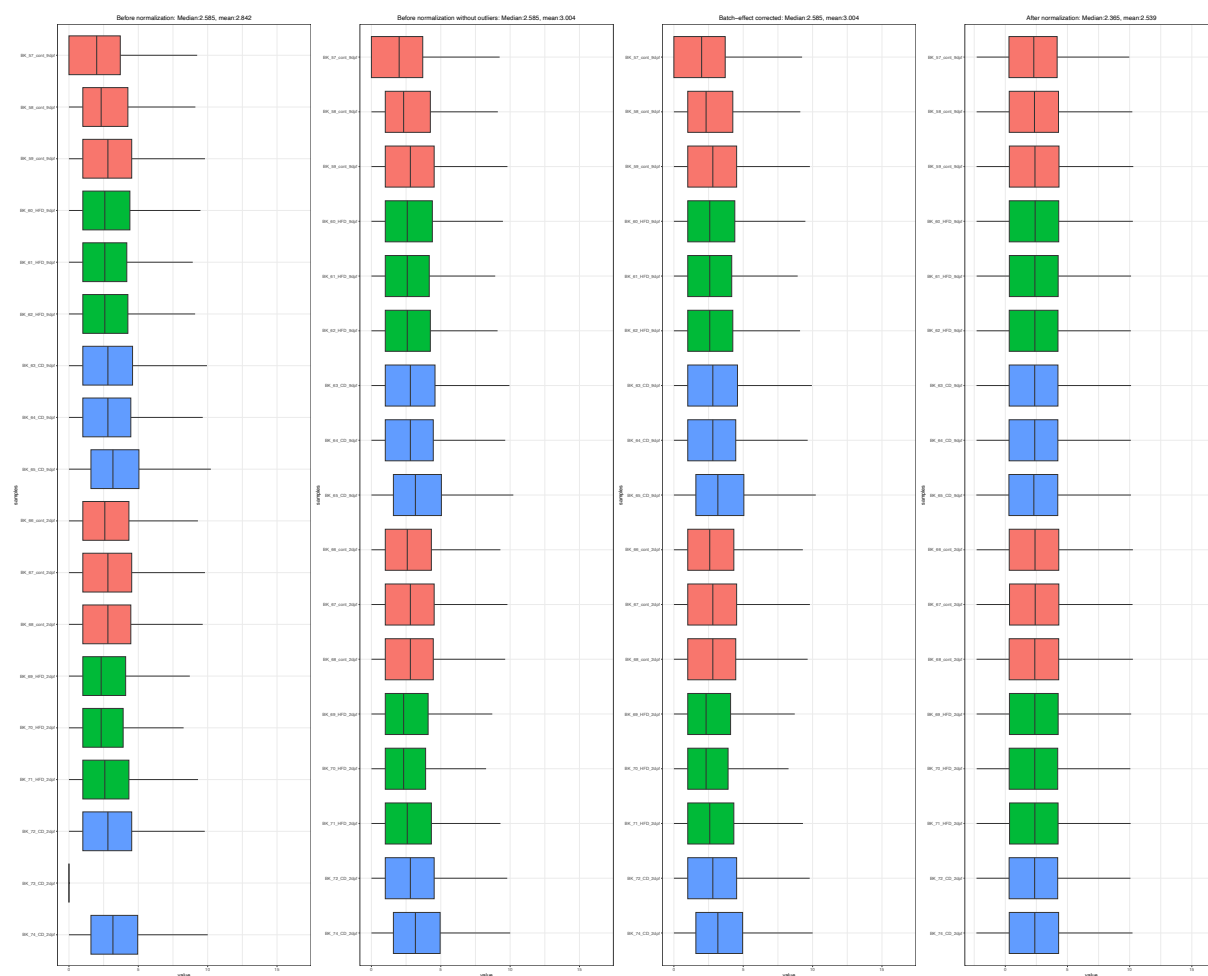


Fig. S12. Distribution of BRB-seq expression values across samples at each pre-processing step in the WD/CD dataset. Box plots of log-transformed expression values for the libraries of the WD/CD experiment (BK_57–BK_74; cont, HFD/WD and CD, 9 dpf and 12 dpf). From left to right: before normalization (median 2.585, mean 2.842), before normalization after outlier removal (mean 3.004), after batch-effect correction, and after normalization (median 2.365, mean 2.539). Colors indicate diet groups: red (RD), green (WD), blue (CD).

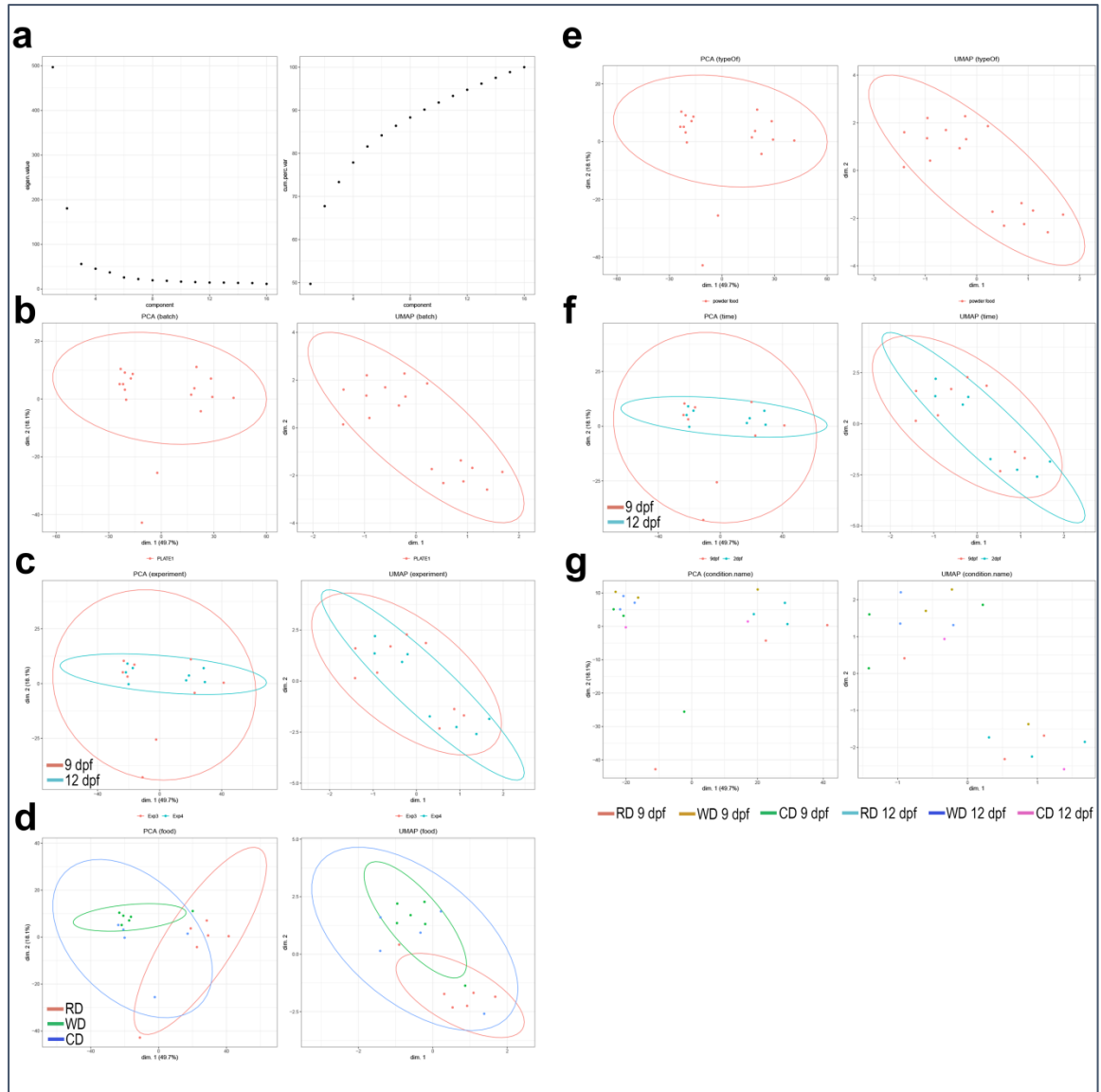


Fig. S13. PCA and UMAP plots of WD vs CD and RD experiment. Unsupervised sample projections of the WD/CD dataset. (a) scree plot of eigenvalues and cumulative percentage of explained variance per principal component. PCA (left column) and UMAP (right column) projections of the retained WD/CD libraries, colored successively by (b) sequencing batch, (c) experiment (Exp3, Exp4), (d) diet (RD, WD, CD), (e) type of food, (f) time point (9 dpf vs 12 dpf) and (g) condition. Ellipses represent the 95% confidence region of each group. The first two principal components account for 49.7% and 18.1% of the total variance.

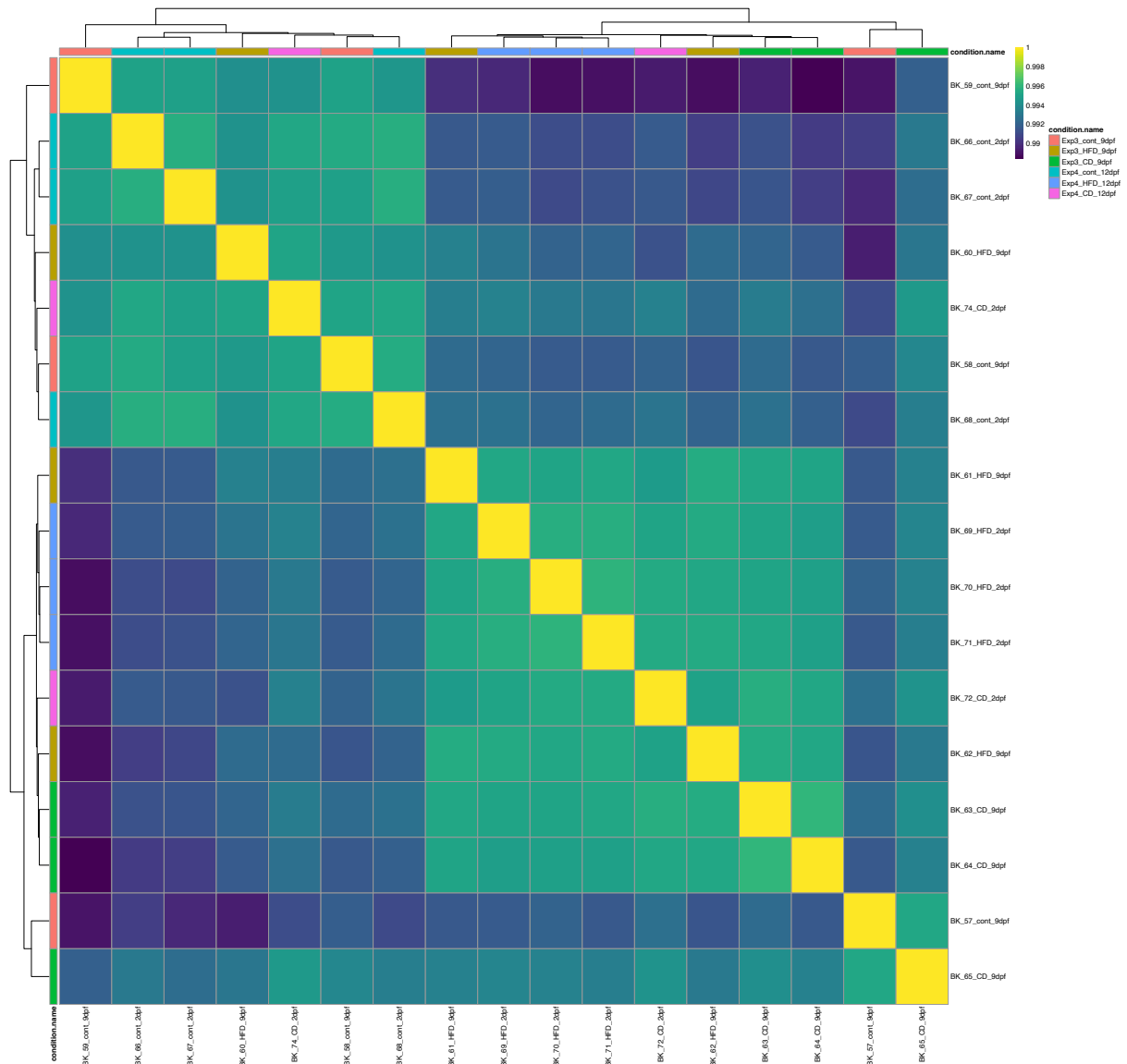


Fig. S14. Sample-to-sample correlation heat map of the WD/CD dataset. Pairwise correlation coefficients between the 17 retained WD/CD libraries, computed on the normalized expression matrix and displayed with hierarchical clustering on rows and columns. The colour scale ranges from 0.99 to 1; annotation bars indicate the condition (Exp3_cont(RD)_9dpf, Exp3_HFD(WD)_9dpf, Exp3_CD_9dpf, Exp4_cont(RD)_12dpf, Exp4_HFD(WD)_12dpf, Exp4_CD_12dpf).

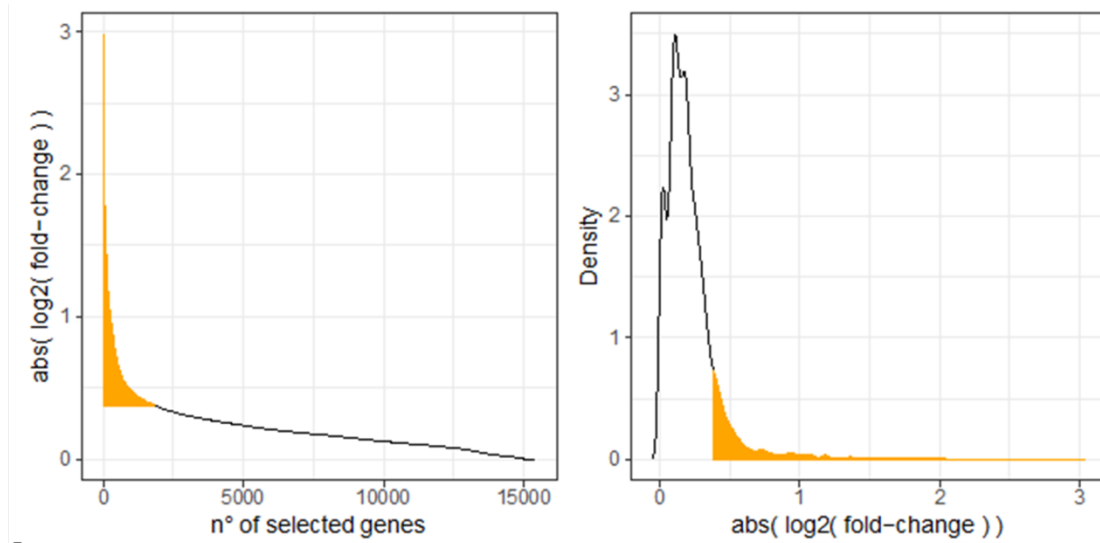
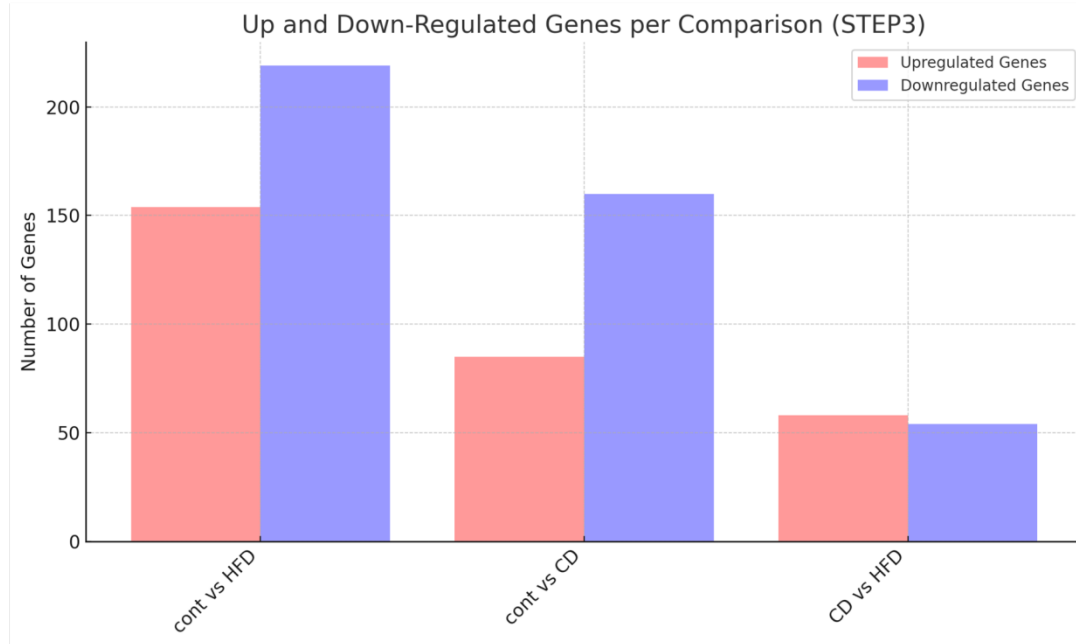
a**b**

Fig. S15: (a) Elbow method approach for western diet (WD) and high fiber control diet (CD) comparisons. The plots show elbow points in two representations: absolute log2 fold change (FC) with respect to number of selected genes and density with respect to absolute log2FC. (b) The graphs showing the number of upregulated and downregulated genes based on the selected statistical filtration strategy in WD vs regular diet (RD), CD vs RD or WD vs CD comparisons. Pink and violet represent upregulated and downregulated genes, respectively.

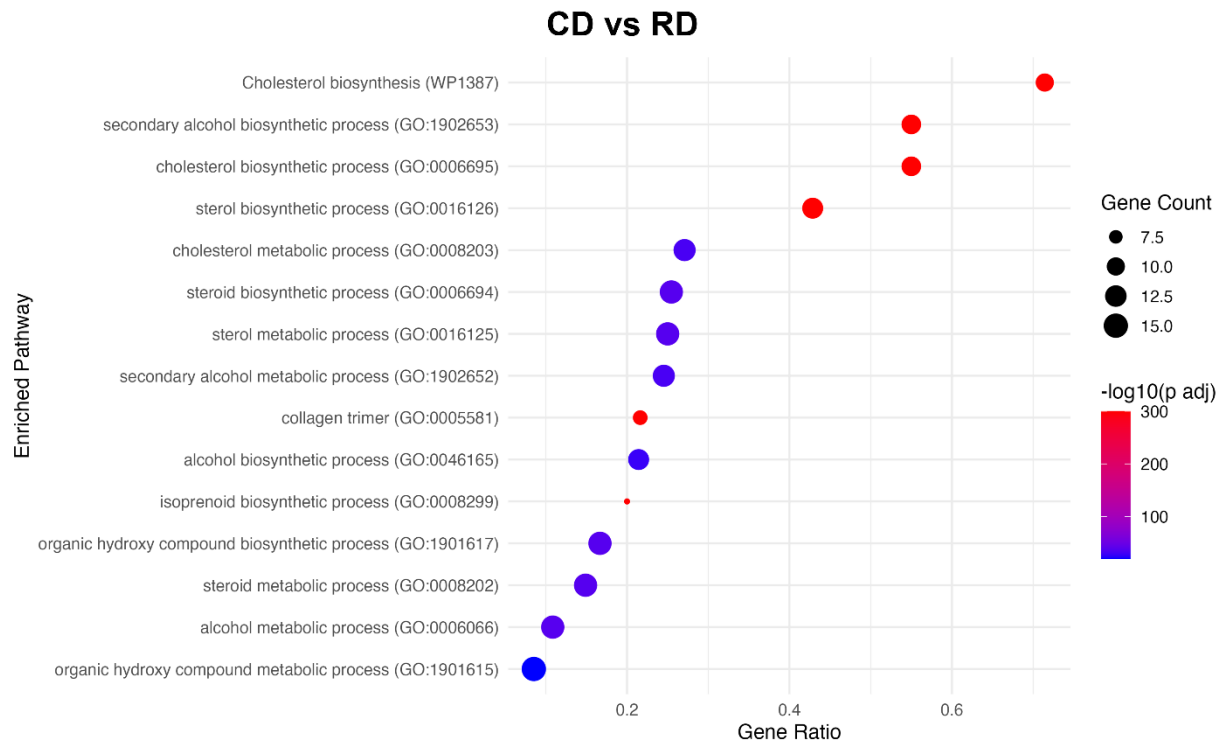


Fig. S16: Enrichment analysis showing the processes enriched and downregulated with high fiber control diet (CD) in comparison to regular diet (RD). The comparison is regardless of the time effect.

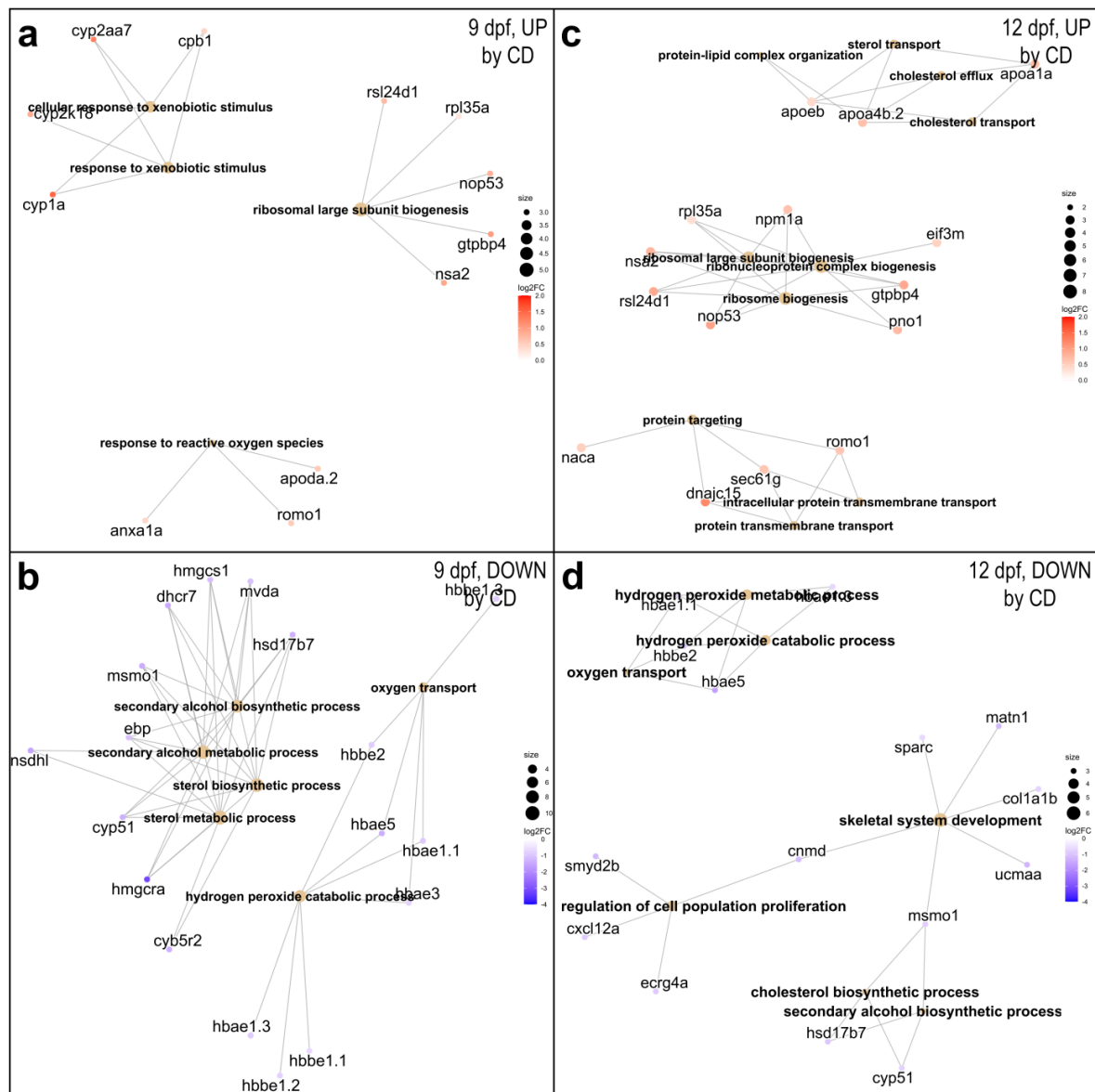


Fig. S17: Context-specific gene regulatory network (cNet) representations for the biological processes and related genes **(a)** upregulated and **(b)** downregulated by high fiber control diet (CD) at 9 days post fertilization (dpf); **(c)** upregulated and **(d)** downregulated by CD at 12 dpf ($p\text{-adj} < 0.05$). Fold change was shown by red for upregulation and blue for downregulation. The size of the circles for each gene refers to the gene count.

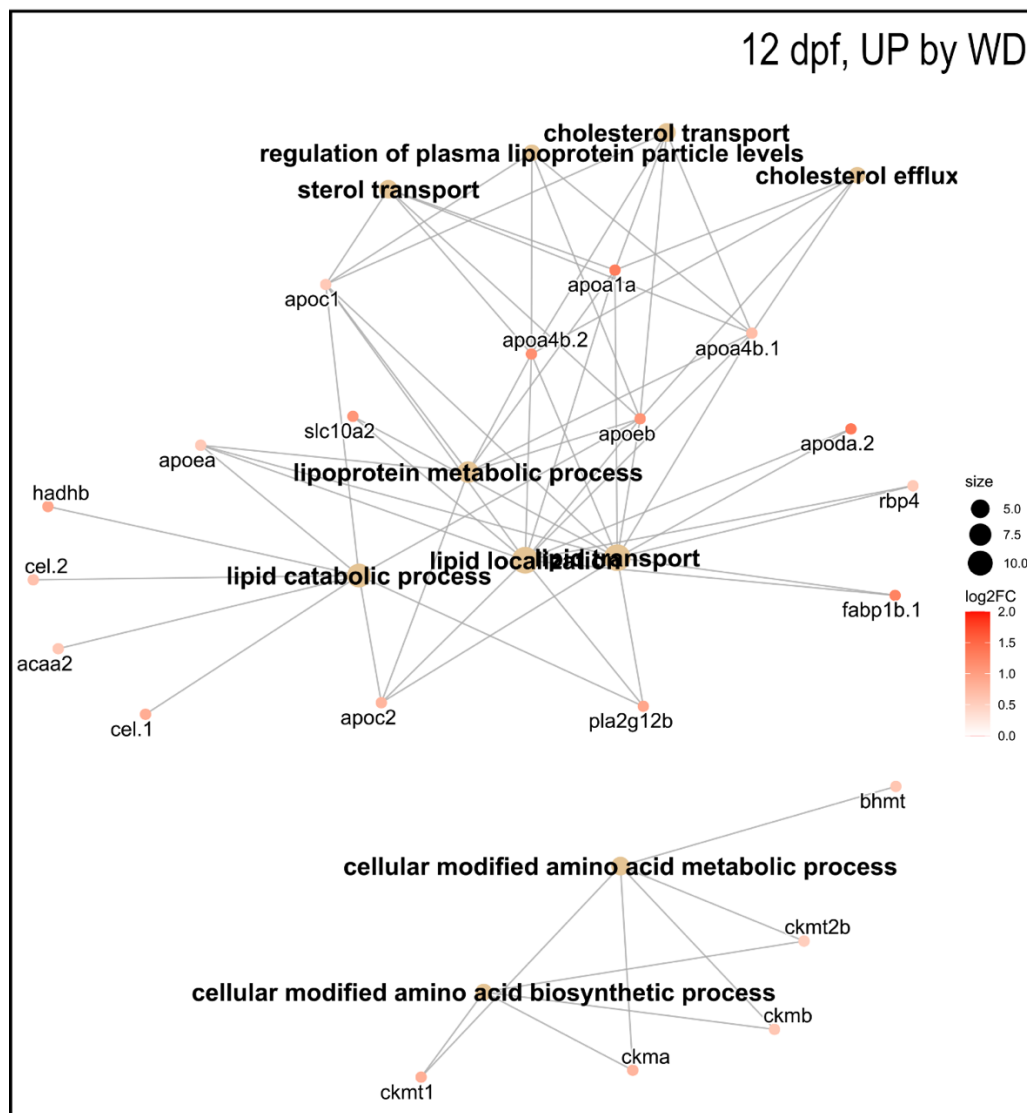


Fig. S18: Context-specific gene regulatory network (cNet) representation for the biological processes and related genes upregulated by western diet (WD) at 12 days post fertilization (dpf) ($p\text{-adj} < 0.05$). The processes were selected with an emphasis on lipid-related pathways. Fold change was shown by red for upregulation. The size of the circles for each gene refers to the gene count.

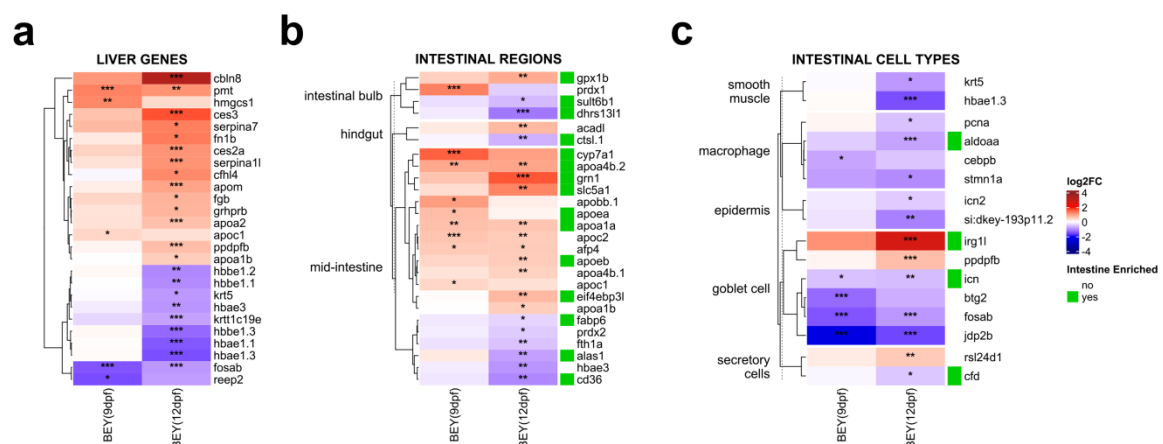


Fig. S19: Significantly modulated genes among liver and intestinal gene sets from Gao et al., 2022 in boiled egg

yolk (BEY) group larvae at 9 or 12 days post fertilization (dpf). Heatmaps show the clustering of (a) liver genes, (b) intestinal regions and (c) intestinal cell types based on the intestinal genes divided into single cell RNA sequencing (scRNA-seq) gene clusters with annotated intestine enriched genes in BEY group larvae (*absolute log2FC* > 0.2630; **p*-adj < 0.05, ***p*-adj < 0.01, ****p*-adj < 0.001).

- 1 Lê, S., Josse, J., & Husson, F. FactoMineR: An R Package for Multivariate Analysis. . *Journal of Statistical Software* **25**, 1–18 (2008). <https://doi.org/https://doi.org/10.18637/jss.v025.i01>
- 2 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015). <https://doi.org/10.1093/nar/gkv007>
- 3 Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**, 279-284 (2001). [https://doi.org/10.1016/s0166-4328\(01\)00297-2](https://doi.org/10.1016/s0166-4328(01)00297-2)
- 4 Chalmel, F. & Primig, M. The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. *BMC Bioinformatics* **9**, 86 (2008). <https://doi.org/10.1186/1471-2105-9-86>
- 5 Gao, Y. *et al.* Unraveling Differential Transcriptomes and Cell Types in Zebrafish Larvae Intestine and Liver. *Cells* **11** (2022). <https://doi.org/10.3390/cells11203290>