scientific reports

Check for updates

OPEN

Identification and clustering analysis of drug-responsive temporally varying genes through high-frequency longitudinal RNA sequencing

Qi Jiang[1,4], Xiaoyu Weng[1,4], Yi Chai[1], Katherine B. Ragan[1], Zhengjin Liu[2], Zhongwu Li[3], Bonnie Wang[1], Olivia Jin[1], Ashwin Gopinath[1], David Yu Zhang[1] & Wei Chen[1]✉

High-frequency longitudinal RNA sequencing has emerged as a powerful approach for capturing dynamic transcriptional responses to therapeutic interventions, yet traditional differential expression analysis fails to identify genes with temporal variability that lack static expression differences. We used our previously developed computational framework for identifying Temporally Varying Genes (TVGs) from daily blood samples collected over 10–21 days in Sprague–Dawley rats treated with hepatotoxic compounds including tetracycline, isoniazid, carbon tetrachloride, and valproate. Our methodology employs variance-based scoring to detect genes exhibiting significant temporal fluctuations under treatment conditions. Unsupervised hierarchical clustering of TVGs identified three distinct temporal patterns: early-transient responses, sustained activation, and late-phase upregulation, each enriched for specific biological processes. Principal component analysis demonstrated clear treatment-induced transcriptomic shifts from baseline "Healthy Region" clusters to treatment-adapted "Response Region" states, with sample trajectories reflecting dose-dependent temporal dynamics. Cross-compound analysis revealed 186 commonly regulated genes across all treatments, representing conserved hepatotoxicity signatures, while compound-specific responses highlighted distinct mechanistic pathways. This approach enables kinetic-pharmacodynamic modeling that distinguishes primary drug targets from secondary adaptive responses, advancing precision medicine applications through dynamic molecular portraits of drug action and individual treatment variability.

RNA sequencing (RNAseq) has been a foundational approach to understanding the effects of pharmaceuticals and treatments on a transcriptomic level[1]. Due to the highly dynamic nature of gene expression, monitoring these fluctuations is imperative for a deeper understanding of how biological systems adapt to various stimuli over time. Historically, RNAseq has been a powerful method for decoding gene expression patterns, particularly in pre-clinical drug trials[2,3]. However, traditional approaches rely on a cross-sectional view, comparing gene expression prior to treatment to a single time-point after treatment. While this approach has merit, it can often fail to capture the real-time fluctuations of dynamic gene regulation in response to external stimuli[4].

With advancements in sequencing technologies, the capacity to perform high-resolution, longitudinal studies expanded significantly[4]. Repeated sampling of biological tissues over time offers a more detailed and dynamic view of gene expression, revealing changes in gene expression over weeks, days or even hours[5–7]. These temporal studies are especially valuable in pharmaceutical research, where a comprehensive understanding of drug induced gene expression changes is required to understand potential toxicities induced by treatment[8]. Longitudinal analyses can highlight transient or delayed gene responses that go otherwise unnoticed in traditional, static analyses[9].

Temporally Varying Genes (TVGs) are of particular interest in this regard. TVGs exhibit changes in expression over time, rendering them a critical component in the research of the effects of drug treatment.

[1]Biostate AI, Inc., Houston, USA. [2]Zhongshan Hospital of Xiamen University, Xiamen, China. [3]Peking University Cancer Hospital and Institute, Beijing, China. [4]Qi Jiang and Xiaoyu Weng contributed equally to this work. ✉email: wei.chen@biostate.ai

Identifying and understanding the behavior of these genes in response to external stimuli provides increased clarity into biological pathways and mechanisms driving treatment, as well as toxicity, responses. Traditional RNAseq analysis tools are not yet equipped to disseminate the complexities of longitudinal datasets, where temporal variability and biological noise can obscure significant gene expression changes.

To address these limitations, we have implemented previously developed methods designed to capture and quantify TVGs in a drug-dosing study designed to induce liver toxicities[10]. By employing these advanced and novel methods, we can more effectively distinguish routine gene expression fluctuations from biologically significant toxicity responses to treatment. The use of these methods grants deeper insight into temporal patterns of toxicity induced gene expression and highlights the potential of using these data sets to predict toxicity responses on an individual level. The applications of such technology are broad, such as providing increased insights into disease mechanisms to allow for the development of novel therapeutics or providing better understanding of individual response to therapeutics.

Results

Overview of longitudinal rat RNAseq studies

After acclimating the rats to their new environments for 3 days, we began daily blood draws, with Day 1 being defined as the first day of blood draws (Fig. 1a). Each blood draw was 200µL; this blood volume was selected as a balance between (1) obtaining sufficient biospecimen for analysis and banking and (2) not inducing undue harm to the rats from the bleeding process. The 200µL blood volume was decided considering of the potential accumulated effects of daily blood draws and followed IACUC approved protocols. To ensure reproducibility of results and minimization of the effects of potential confounding factors, all blood draws were performed at approximately 10:00am local time each day. The relevant rats dosed with the respective drug intravenously on their Day 3 or Day 8, about 30 min prior to the next blood draw. Consequently, the first blood sample after dosing

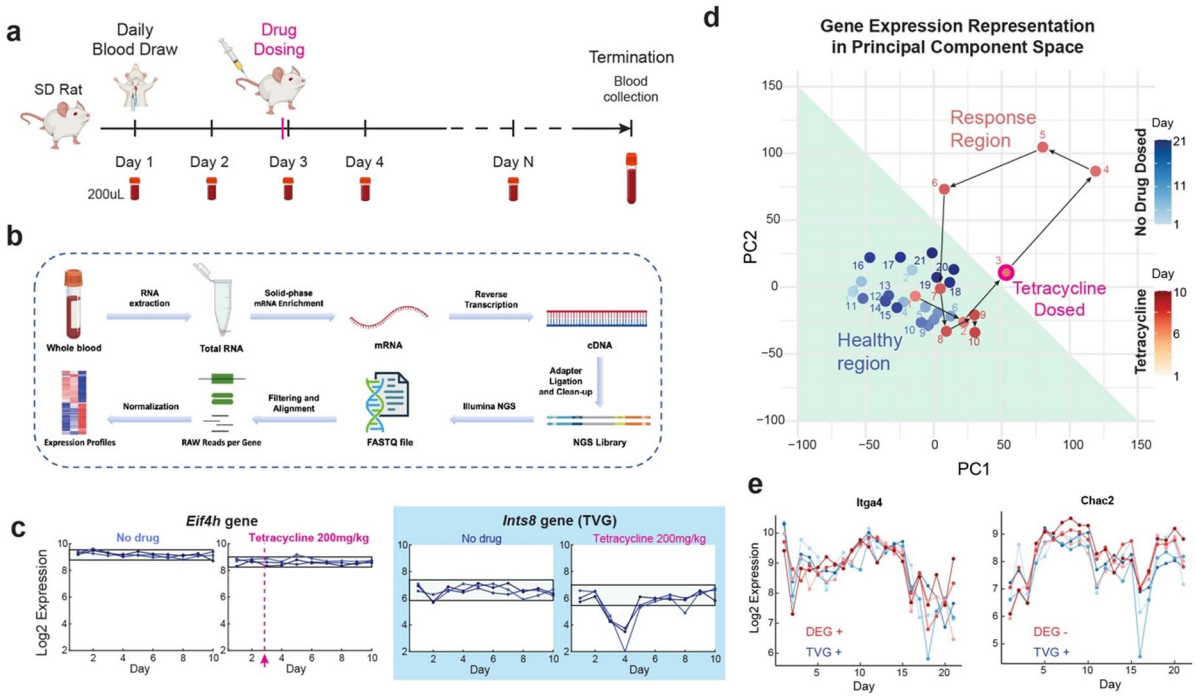


Fig. 1. Overview of longitudinal rat RNAseq studies. (**a**) Animal experiment design. Sprague Dawley rats are acclimated for 3 days at the contract research organization animal facilities before beginning daily blood draws (200 µL). Blood draws were performed at a consistent time of the day (around 10am) to minimize the impact of Circadian rhythm on gene expression. At least 2 daily blood draws were performed before dosing the animals with any drug molecules. (**b**) Workflow for mRNA sequencing and expression analysis. (**c**) Example longitudinal gene expression profiles for rats dosed with tetracycline just prior to Day 3 blood draw. (**d**) PCA trajectory of rat RNA-seq transcriptomic profiles. Each node represents the centroid of three biological replicates at a given time point. Blue dots: control group (no treatment); red dots: experimental group (tetracycline 200 mg/kg, single dose administered on Day 3). The light green shaded area indicates the "Healthy Region" where samples cluster near baseline; the white area indicates the "Response Region" representing acute drug response. Individual replicate data are shown in Supplementary Fig. 2. (**e**) Representative examples of TVG expression patterns over time in the control group (21-day observation, no drug treatment). Both genes exhibit temporal fluctuations potentially related to bleeding effects or periodic physiological responses. *Itga4* (left, TVG+/DEG+) shows significant temporal variation and is also identified as a DEG by cross-sectional comparison (Day 1 vs. Day 3, $|\log2FC| > 1$, adjusted p-value < 0.05). *Chac2* (right, TVG+/DEG−) exhibits significant temporal variation but does not meet DEG criteria.

likely do not reflect the full effects of each molecule, based on the pharmacokinetics properties of each molecule. Following the end of each experiment, rat liver necropsy samples were analyzed via standard histopathology.

RNA was extracted from whole blood within 6 h of collection. Subsequently, we used bead-based solid phase enrichment of mRNA from the total RNA sample using poly-T probes to remove noncoding RNA and ribosomal RNA. The enriched mRNA sample was then reverse transcribed using random hexamer primers, and the cDNA was sequenced using Illumina NGS at a depth of between 10 and 20 M reads per sample. After filtering and alignment, we obtained the raw RNA expression profile, which was subsequently normalized based on the number of reads in each NGS library. See Fig. 1b and the methods section for an overview and for additional details on experimental and bioinformatics preprocessing methods.

As a preview of the data we collected and analyzed, Fig. 1c shows the temporal expression dynamics for two example genes, *Eif4h* and *Ints8*. Both genes showed temporally stable expression for the control rats, but the *Ints8* gene shows a fourfold drop in expression on Day 4 for all 3 rats before returning on pre-dose levels on Day 5. Principal component analysis was performed to visualize the temporal dynamics of the transcriptomic response. Each node represents the centroid of three biological replicates at a given time point, with blue and red dots indicating the control and tetracycline-treated groups, respectively (Fig. 1d). The PCA trajectory revealed a clear spatial separation between a "Healthy Region" (light green shaded zone) and a "Response Region" (white zone). Following a single dose of tetracycline on Day 3, the experimental group exhibited pronounced deviation from the control cluster during Days 3–6, representing the acute drug response phase. By Day 7 and beyond, the treated samples returned to positions proximal to the control group, indicating transcriptomic recovery to a healthy baseline state. Sample trajectories were colored according to experimental day (blue gradient, Days 1–21) and tetracycline dosing intensity (red scale, 1–10), demonstrating treatment-induced transcriptomic shifts. The PCA plot showed PC1 values ranging from −100 to 150 and PC2 values from −100 to 150, with numbered points (1–21) indicating specific timepoints and connected trajectories illustrating temporal progression. The trajectory of the rats through this parameter space shows that it took the rats roughly 4 days to fully recover from the dose, suggesting that the Ints8 gene is an "early responder" gene. The dynamics of the rats' recovery back to full health follows similar trajectories for all the molecules we tested, though we also noticed important individual variabilities in the magnitude of TVG RNA perturbation and the speed at which the rats recovered.

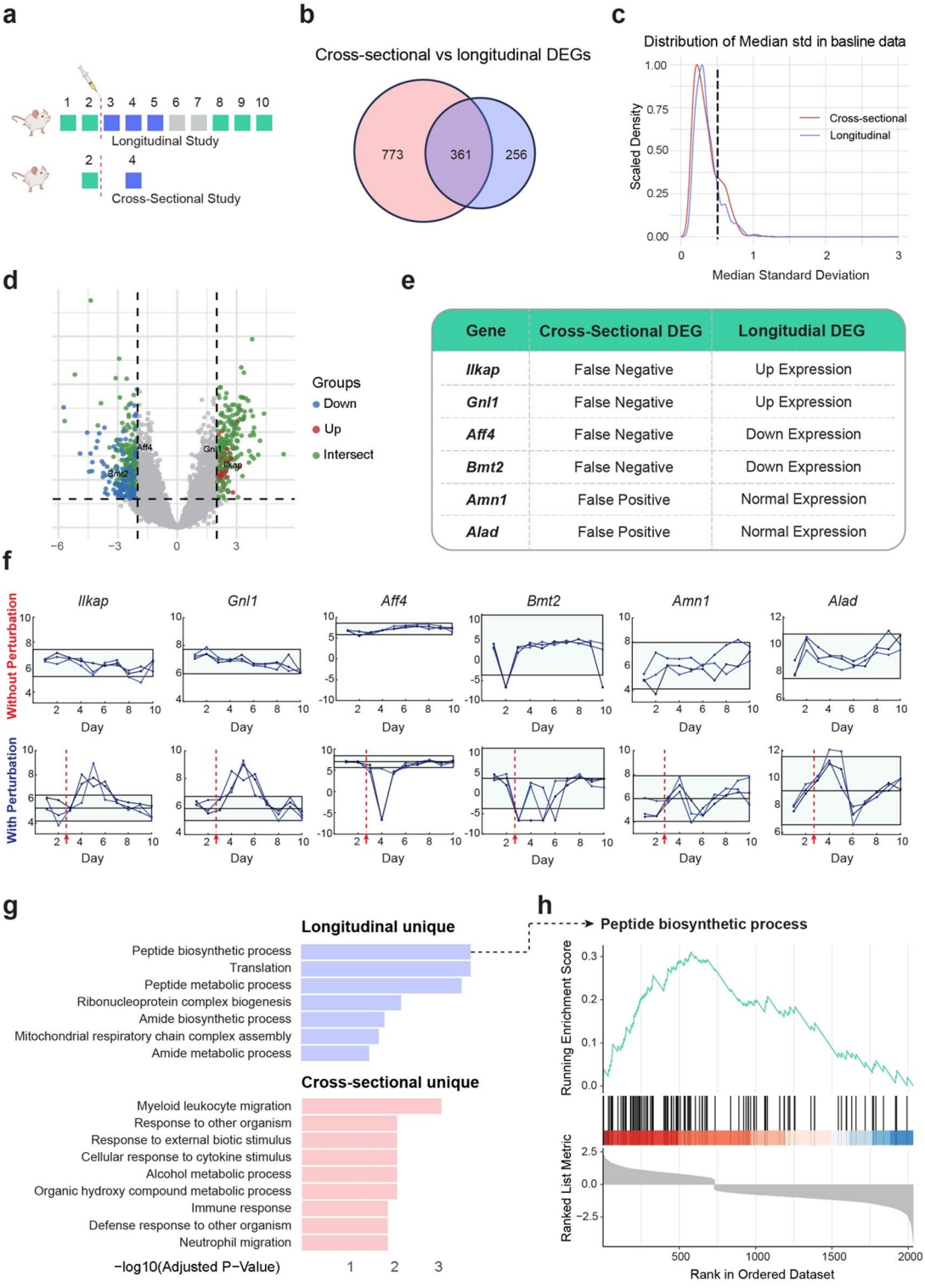
To illustrate the distinction between cross-sectional and longitudinal approaches, we categorized genes based on their TVG and DEG status using control group data (21-day observation without drug treatment). DEGs were identified by comparing Day 1 vs. Day 3 expression levels ($|\log2FC| > 1$, adjusted p-value < 0.05). Both *Itga4* and *Chac2* exhibited substantial temporal fluctuations in expression, potentially attributable to bleeding effects or periodic physiological responses (Fig. 1e). However, only *Itga4* (TVG+/DEG+) met the DEG criteria in cross-sectional analysis, while *Chac2* (TVG+/DEG−) was identified exclusively through longitudinal TVG analysis. This comparison underscores the unique capability of high-frequency longitudinal RNA sequencing to capture gene expression dynamics beyond the scope of traditional differential expression analysis. These profiles illustrated the complex temporal dynamics of drug-responsive genes, with multiple sample trajectories showing consistent directional changes in expression following tetracycline exposure. See Supplementary Excel 1 for the full list of called TVGs.

Comparative analysis of cross-sectional versus longitudinal differential gene expression

To compare the performance of cross-sectional and longitudinal designs in detecting drug-responsive genes, we analyzed DEGs from tetracycline-treated rats (200 mg/kg) using two approaches: a cross-sectional comparison between Days 2 and 4 and a longitudinal comparison contrasting the dosing window (Days 3–5) against pre-dose and recovery days (Days 1–2 and 8–10) (Fig. 2a). Venn analysis showed 361 overlapping DEGs, with 773 and 256 genes uniquely detected by cross-sectional and longitudinal methods, respectively (Fig. 2b). To evaluate the reliability of DEGs identified by each method, we assessed their expression stability in the baseline control group (21-day observation without drug treatment). A robust drug-responsive DEG should exhibit stable expression in the absence of drug perturbation. As shown in Fig. 2c, while DEGs identified by both methods showed similar peak positions in median standard deviation, the cross-sectional DEGs exhibited a longer tail extending into higher variability values (std > 0.5), whereas longitudinal DEGs showed markedly reduced density in this region. This suggests that longitudinal analysis identifies DEGs with more consistent baseline expression, potentially reducing false positives arising from genes with inherently high expression variability.

A volcano plot of longitudinal DEGs showed both upregulated (red) and downregulated (blue) genes, with overlapping DEGs from the cross-sectional analysis highlighted in green (Fig. 2d). Several genes identified only in the longitudinal design, including *Ilkap*, *Gnl1*, and *Aff4*, exhibited low baseline expression followed by consistent increases across multiple days post-dose. These genes were not detected in the cross-sectional Day 2 vs. 4 comparison due to their transient or delayed expression dynamics (Fig. 2e–f). Conversely, genes such as *Amn1* and *Alad*, identified as DEGs in the cross-sectional design, did not show consistent changes during the dosing window and appeared to be false positives when evaluated in the full time course.

To assess the functional relevance of these distinct gene sets, we performed GO enrichment and GSEA analysis on DEGs uniquely identified by each method[11,12]. Longitudinal-specific DEGs were enriched for peptide biosynthetic process, ribonucleoprotein complex biogenesis, and translation-related pathways, consistent with tetracycline's known mechanism of action on ribosomal inhibition (Fig. 2g)[13]. Cross-sectional-only DEGs were associated with immune-related and stress response pathways, including neutrophil migration and cytokine stimulus response. GSEA of the peptide biosynthetic process (GO:0043043) showed a strong enrichment signal in longitudinal DEGs but not in cross-sectional results (Fig. 2h). Together, these data demonstrate the improved resolution and understanding of biological systems that can be obtained when examining longitudinal gene expression rather than cross sectional.



Identification of TVGs across drug treatments

We dosed the rats with 4 different small molecules: tetracycline, isoniazid, valproate, and carbon tetrachloride. The first 3 are FDA approved drugs, and the last is a formerly used drug (Necatorina) that is now a standard molecule for testing liver damage[14–16]. In our studies, we found that the highest number of TVGs in response to tetracycline dosing and will primarily display tetracycline results in the main text and figures. See Supplementary Excel 2 for more details on the experimental design. We dosed N = 3 rats at each of 4 different concentrations of tetracycline (4, 15, 50, and 200 mg/kg), and unsurprisingly the highest dose elicited the strongest gene expression response. Consequently, we used the 200 mg/kg dataset to do TVG identification, following the same method as previously[10]. In this analysis, we excluded the previously called 300 bleeding/baseline TVGs and identified 3206 TVGs for associated with tetracycline response (Fig. 3a)[10].

◂**Fig. 2**. Comparative analysis of cross-sectional versus longitudinal DEGs reveals distinct biological pathway enrichments and dynamic expression patterns. (**a**) Schematic overview of sampling design for the longitudinal study (Days 1–10) and the cross-sectional comparisons (Days 2 vs. 4). (**b**) Venn Diagram of differentially expressed genes (DEGs) in tetracycline-treated (200 mg/mL) samples. Cross-sectional analysis (Day 2 vs. Day 4) is contrasted with longitudinal analysis (days 3, 4, 5 vs. days 1, 2, 8, 9, 10) (cutoff: adjusted p-value = 0.01, abs log2foldchange = 2). (**c**) Distribution of median standard deviation (std) for DEGs identified by cross-sectional (red) and longitudinal (blue) methods, measured in the baseline control group (21-day observation, no drug treatment). The scaled (0–1) density of median std across individual animals is displayed. Dashed line: 50th percentile. Reliable drug-responsive DEGs should exhibit stable baseline expression (low std). While both distributions show similar peak positions, longitudinal DEGs display reduced density in the high-variability tail region (std > 0.5), indicating that longitudinal analysis identifies DEGs with more consistent baseline expression. (**d**) Volcano plot of DEGs in longitudinal experimental design. Displaying the -log10 p-value vs. log2 fold change of DEGs (cutoff: p-adjust < 0.01, abs log2 fold change > 2) from the longitudinal experimental design. Upregulated genes are marked in red and downregulated in blue. DEGs also identified from cross-sectional design are highlighted in green. (**e**) Example of false positive discovery gene and false negative gene using only cross-sectional data (Day 2 vs. Day 4). (**f**) Gene expression profiles of 3 male rats across 10 days with and without perturbations (Tetracycline, 200 mg/kg). Genes listed in (**e**) are depicted. (**g**) GO enrichment bar plots of biological processes uniquely identified by longitudinal-only or cross-sectional-only DEGs (BH adjusted p-value < 0.05). Longitudinal DEGs are enriched for core pharmacodynamic targets such as peptide biosynthetic processes and ribonucleoprotein complex biogenesis, whereas cross-sectional DEGs are enriched for immune-related and metabolic stress pathways. (**h**) GSEA plot for peptide biosynthetic process (GO:0043043) showing positive enrichment specifically among TVG+/DEG− genes.

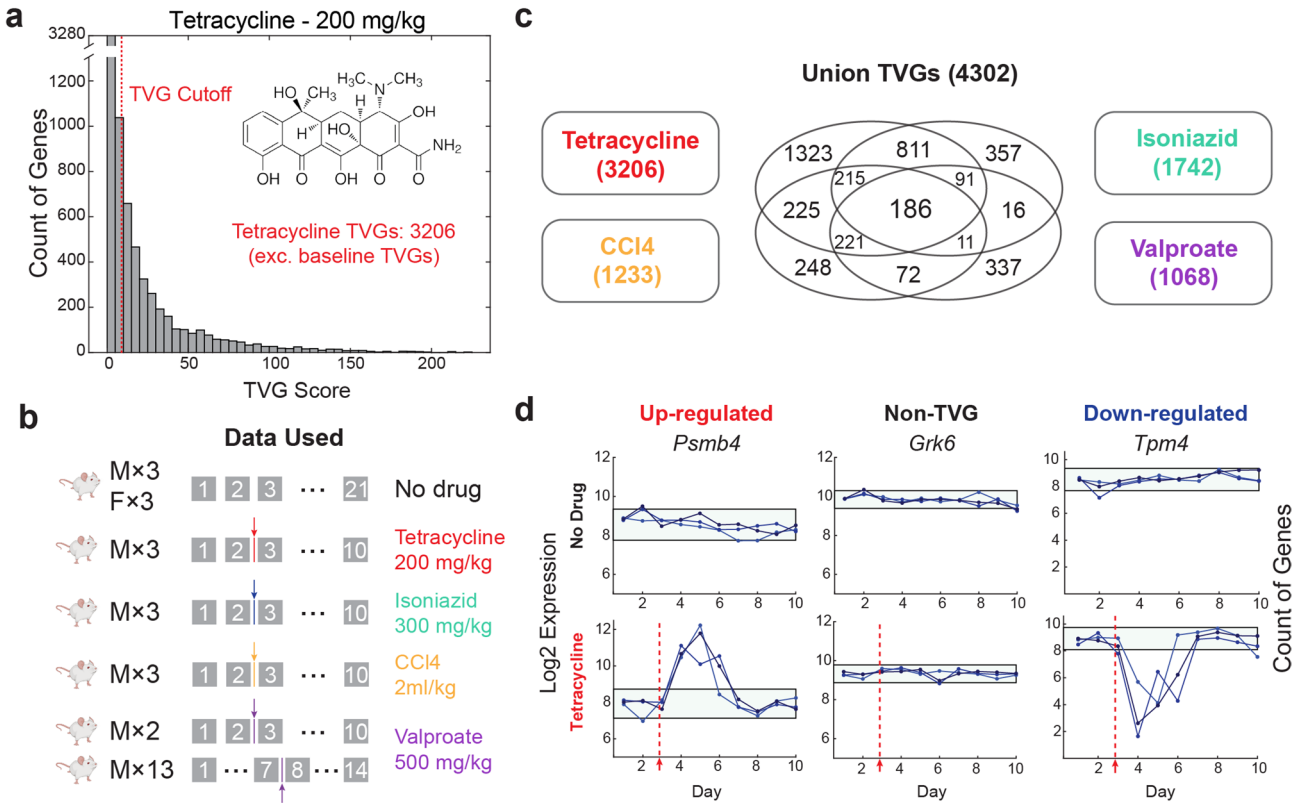


Fig. 3. Identification of TVGs across drug treatments. (**a**) Summary of TVG scores for 200 mg/kg tetracycline. 3206 TVGs were identified. (**b**) Schematic of the drug-dosing information for the rat experiments. (**c**) Venn diagram illustrating the 4302 TVGs identified for each drug using the data from 482 blood samples from 30 rats as shown in (**b**). The remaining 358 blood samples were not used for PCA because they used lower doses of drugs or resulted in animal death. In addition to the number of intersections TVGs shown, there were 134 TVGs for tetracycline and valproate only and 55 TVGs for isoniazid and carbon tetrachloride. (**d**) Examples of up-regulated TVG, non-TVG, and down-regulated TVG for tetracycline. Horizontal lines show + /− 2 standard deviations from mean log expression.

The multi-drug comparative study design incorporated five treatment groups with varying sample sizes and timepoints: no drug control (n = 2–3 mice, 21 timepoints), tetracycline 200 mg/kg (n = 2–3 mice, 10 timepoints), isoniazid 300 mg/kg (n = 2–3 mice, 10 timepoints), CCl_4 2 ml/kg (n = 2–3 mice, 10 timepoints), and valproate 500 mg/kg (n = 13 mice, 14 timepoints with specific sampling at days 7–8) (Fig. 3b). This design enabled comprehensive characterization of both drug-specific and shared transcriptional responses across multiple hepatotoxic compounds.

Using data from 482 blood samples across a total of 30 rats, we identified 4302 total unique TVGs across all treatments, with treatment-specific contributions varying substantially: tetracycline (3206 TVGs), isoniazid (1742 TVGs), CCl_4 (1233 TVGs), and valproate (1068 TVGs) (Fig. 3c). Notably, 186 genes were commonly regulated across all four compounds, indicating a core set of genes responsive to hepatotoxic insults regardless of the specific mechanism of action. To investigate the biological significance of the 186 shared TVGs, we performed pathway enrichment analysis (Supplementary Fig. 3). Ferroptosis[17], an iron-dependent cell death mechanism implicated in drug-induced liver injury, was the most significantly enriched KEGG pathway (p.adjust = 1.74×10^{-6}). GO analysis further identified ferric iron binding and iron ion homeostasis as top enriched terms, with multiple ferritin genes (*Ftl1*, *Ftl1l1*) and transferrin (*Tf*) among the shared TVGs. These results suggest that iron metabolism disruption and ferroptosis represent common molecular mechanisms underlying hepatotoxic drug responses. In addition to the number of intersections TVGs shown, there were 134 TVGs for tetracycline and valproate only, and 55 TVGs for isoniazid and carbon tetrachloride. The remaining 358 blood samples were not used for this because they used lower doses of drugs or resulted in animal death. Representative longitudinal expression profiles comparing control and tetracycline conditions demonstrated three distinct gene categories: upregulated TVGs (exemplified by *Psmb4*), non-TVG controls (*Grk6*), and downregulated TVGs (*Tpm4*), with clear temporal patterns emerging over the 10-day treatment period (Fig. 3d). The profiles showed that treatment initiation at Day 3 (marked by red dashed lines) consistently triggered measurable expression changes in TVGs while leaving non-TVG controls largely unaffected.

Temporal Clustering Reveals Distinct Response Patterns

To characterize the functional programs underlying drug response, we performed unsupervised soft clustering on the 3206 tetracycline-responsive TVGs using the Mfuzz algorithm (fuzzy c-means)[18]. The optimal number of clusters (k = 6) was determined using the elbow method implemented in the ClusterGVis R package (Supplementary Fig. 4)[19]. Clustering was performed on Z-score normalized expression values across the 10-day time course.

Six distinct temporal response patterns were identified (Fig. 4a). Cluster C3 (462 genes) showed early upregulation peaking at Day 3–4, followed by gradual decline. Cluster C1 (679 genes) exhibited moderate early upregulation with sustained elevation through Day 5–6. Cluster C2 (294 genes) displayed sharp upregulation at Day 3–4 followed by a transient downregulation at Day 6 and then a rapid return to baseline. In contrast, clusters C5, C6, and C4 showed downregulation patterns: C5 (324 genes) showed delayed response with downregulation beginning at Day 5–6; C6 (782 genes) exhibited downregulation during acute phase (Day 4) with recovery afterward; C4 (665 genes) showed sustained downregulation with peak at Day 5.

To investigate the biological significance of these temporal patterns, we performed Gene Ontology enrichment analysis for each cluster (Fig. 4b). Cluster 1 was significantly enriched in ribosome biogenesis, cytoplasmic translation, and mRNA processing, indicating activation of protein synthesis machinery during acute drug response. Clusters 3 and 4 were enriched in immune response regulation. Clusters 5 and 6 showed enrichment in autophagy and DNA metabolic processes, reflecting cellular stress responses and repair mechanisms.

These results demonstrate that temporal clustering of TVGs reveals biologically coherent gene modules with distinct functional roles in the progression of drug response, from acute metabolic disruption through immune activation to cellular recovery processes.

Dose-dependent transcriptomic responses

To investigate how drug dosage affects the magnitude and timing of transcriptomic responses, we analyzed tetracycline treatment data across five dosage groups (0, 4, 15, 50, and 200 mg/kg). PCA analysis revealed dose-dependent temporal dynamics, with PC1 scores showing increasingly larger deviations from baseline during the response period (Days 3–6) as dosage increased (Fig. 5a). ROC analysis confirmed that PC1 scores effectively distinguish the response period from baseline across all dosages (AUC = 0.896), with individual dose groups achieving AUCs ranging from 0.866 to 0.949 (Fig. 5b).

We systematically categorized TVGs based on their dose-sensitivity thresholds and identified four distinct response patterns (Fig. 5c, Supplementary Excel 4). *Guk1* (guanylate kinase 1), an essential enzyme in nucleotide metabolism associated with mitochondrial function and hepatopathy[20], responded to all four dosages, suggesting that nucleotide metabolism is highly sensitive to hepatotoxic perturbation. *Prpf39* (pre-mRNA processing factor 39), a spliceosome component involved in mRNA processing[21], responded to medium-to-high doses (15, 50, 200 mg/kg), indicating engagement of RNA processing machinery under more substantial toxic stress. *Fcgr1a* (Fc gamma receptor Ia/CD64), a high-affinity IgG receptor and established biomarker for infection and inflammation[22], responded only to higher doses (50, 200 mg/kg), suggesting that immune/inflammatory pathways require a threshold level of hepatotoxic injury. *Pcsk6* (proprotein convertase subtilisin/kexin type 6), which plays a critical role in acute liver injury through regulation of inflammatory responses[23], responded exclusively to the highest dose (200 mg/kg), indicating that severe hepatotoxicity is required to activate this metabolic remodeling pathway. This spectrum of dose-sensitivities suggests that different biological pathways have distinct activation thresholds, and such categorization may facilitate the identification of sensitive biomarkers for early detection of drug toxicity at low exposure levels.

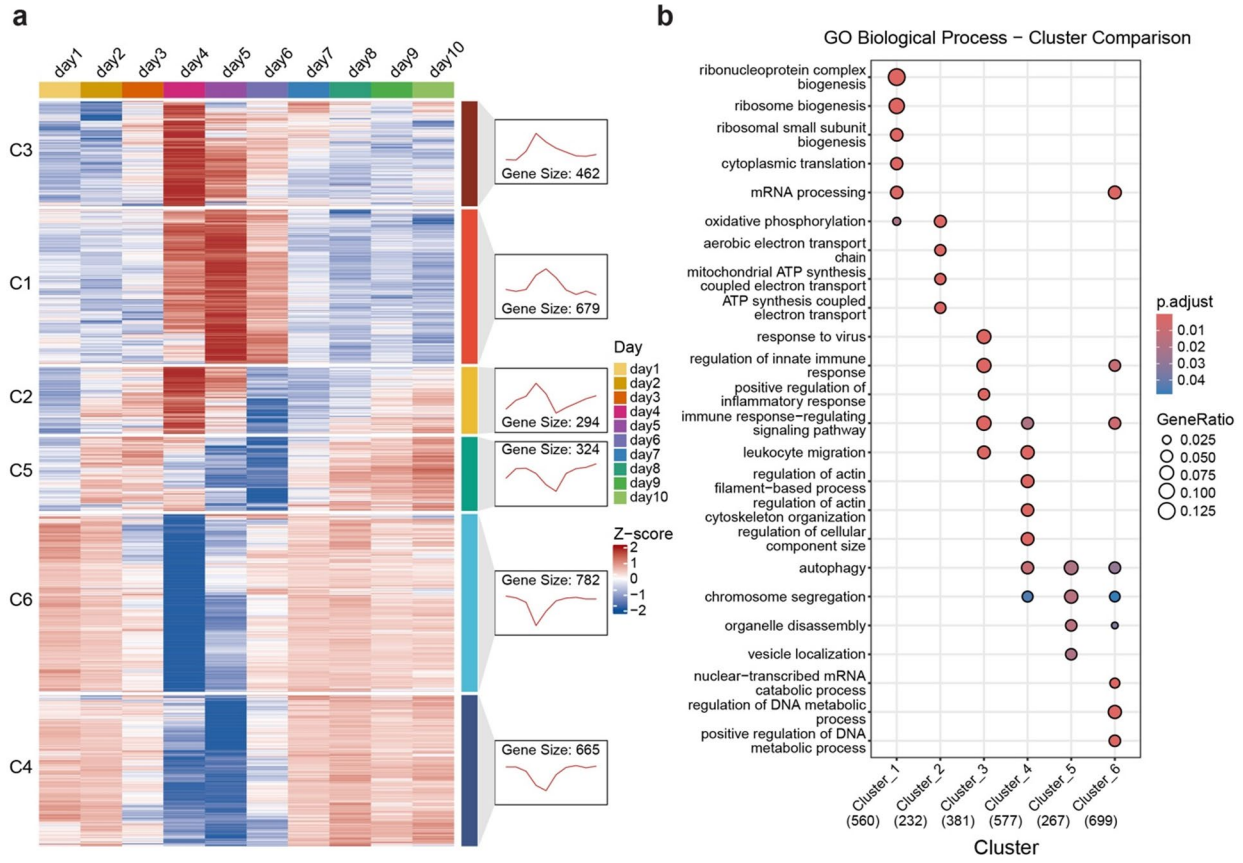


Fig. 4. Temporal clustering and functional enrichment of tetracycline-responsive TVGs. (**a**) Hierarchical clustering of 3206 TVGs identified six distinct temporal response patterns (C1-C6). Left: Heatmap showing Z-score normalized expression across 10 days. Right: Mean expression trajectory for each cluster with gene counts. C3 and C1 show early upregulation patterns; C2 exhibits transient upregulation and downregulation; C5, C6 and C4 display downregulation patterns with different kinetics. (**b**) GO Biological Process enrichment analysis comparing all six clusters. Dot size represents gene ratio; color indicates adjusted p-value. Distinct functional signatures were identified for each cluster.

Expansion of analysis to other hepatotoxic drug treatments

To expand our analysis, we performed unsupervised clustering on the temporal gene expression patterns across individual treatment conditions for the remaining 3 drug treatments and compared the results to the unsupervised clustering performed on the tetracycline 200 mg/kg analysis (Supplementaryl Fig. 5). Similar clustering patterns were observed for isoniazid 300 mg/kg (Supplementary Fig. 6), CCl_4 2 ml/kg (Supplementary Fig. 7), and valproate 500 mg/kg (Supplementary Fig. 5), with each treatment showing characteristic temporal dynamics and treatment-specific transcriptional signatures. The consistent clustering patterns across different compounds validated the robustness of the temporal gene expression analysis approach and confirmed that distinct drugs induce characteristic temporal transcriptional programs while sharing common response elements.

Comprehensive gene ontology enrichment analysis across all treatment conditions revealed commonality in activated pathways between the 3 unique treatments in terms of activation the immune response (Fig. 5a–c). However, each treatment, despite imparting hepatotoxic effects, activated different sets of pathways indicative of cellular stress. Given the large number of TVGs identified specific to different subsets of drugs, it is likely that these drugs activate multiple different pathways for recovery. This understanding of differential TVG profiles for similar drug treatments reveals an important facet important to understand when selecting therapeutics.

Analyzing the other 3 drug molecules, we identified 1742 TVGs for isoniazid, 1233 TVGs for carbon tetrachloride, and 1068 TVGs for valproate (Fig. 3c). The union of these TVGs with the 3206 tetracycline TVGs yielded a total of 4302 TVGs. Of these, 186 genes were identified as TVGs for all 4 drugs, including the *Lrp1* gene (Fig. 5d). Given the large number of TVGs identified specific to different subsets of drugs, it is likely that these drugs activate multiple different pathways for recovery. Detailed temporal gene expression profiles for eight representative genes (*Lrp1*, *Crip1*, *Hypk*, *Kat6a*, *Maf1*, *Slc16a10*, *Stat6*, *Tlr8*) across four treatment conditions revealed distinct treatment-specific and gene-specific temporal dynamics (Fig. 5d). The analysis compared tetracycline 200 mg/kg (red), isoniazid 300 mg/kg (blue), CCl_4 2 ml/kg (orange), and valproate 500 mg/kg (purple) over 10 days, with Log2 expression values ranging from approximately 4–12 depending on the gene.

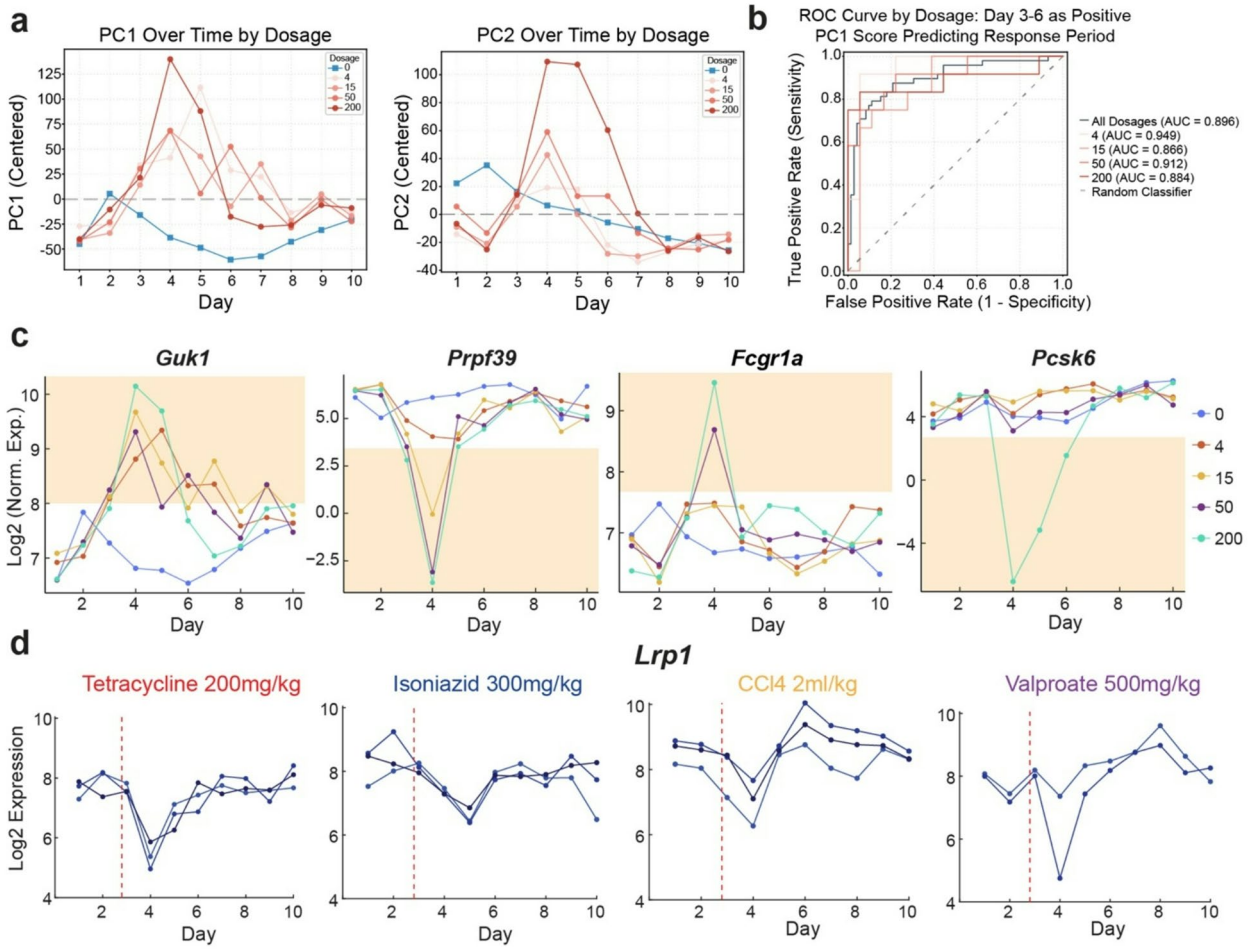


Fig. 5. Dose-dependent transcriptomic responses to tetracycline treatment. (**a**) PCA trajectory analysis showing PC1 (left) and PC2 (right) scores over time for different tetracycline dosages (0, 4, 15, 50, 200 mg/kg). Higher doses show larger deviations from baseline during the acute response period (Days 3–6). (**b**) ROC curves evaluating PC1 scores as predictors of the drug response period (Days 3–6 vs. other days). AUC values indicate classification performance for all dosages combined (0.896) and individual dose groups. (**c**) Representative genes demonstrating distinct dose-sensitivity patterns. *Guk1*: responds to all four dosages; *Prpf39*: responds to 15, 50, and 200 mg/kg; *Fcgr1a*: responds to 50 and 200 mg/kg; *Pcsk6*: responds only to 200 mg/kg. Shaded areas indicate the drug response window (Days 3–6). Complete gene lists for each dose-sensitivity category are provided in Supplementary Excel 4. (**d**) Temporal expression profile of *Lrp1* across four hepatotoxic drug treatments: tetracycline (200 mg/kg), isoniazid (300 mg/kg), CCl4 (2 ml/kg), and valproate (500 mg/kg). Red dashed lines indicate drug administration time (Day 3). *Lrp1* (low-density lipoprotein receptor-related protein 1) demonstrates consistent temporal variation across all four treatments, supporting its role as a marker of common recovery and adaptation processes following hepatotoxic perturbation. Additional gene expression profiles are provided in Supplementary Fig. 10.

Some genes (*Stat6*, *Tlr8*) showed relatively stable expression across treatments, while others (*Lrp1*, *Crip1*) displayed dynamic temporal changes with treatment-specific patterns.

Discussion

In this study, we demonstrate that high-frequency longitudinal RNA sequencing can fundamentally transform our understanding of drug-responsive transcriptional dynamics by revealing treatment-variable genes (TVGs) that remain undetectable through conventional static analyses. Our comprehensive analysis of 3206 tetracycline-responsive TVGs, compared to the limited scope of traditional differentially expressed gene approaches, establishes temporal transcriptomics as an important methodology for capturing the full spectrum of pharmacological responses. The identification of distinct temporal clusters provides mechanistic insights into the sequential activation of stress response, metabolic adaptation, and tissue remodeling pathways that define therapeutic intervention outcomes. To facilitate analysis of this large longitudinal RNA expression data, we developed and utilized several ad hoc bioinformatics methods to (1) minimize the number of false positive DEGs called from the large number of potential pairwise differential expression analyses, and (2) identify TVGs from longitudinal data. Although these tools have not be rigorously optimized for performance, spot check

evaluation results suggests that the TVG calling is reasonably accurate and catches a significant number of false positive DEGs called based on cross-sectional tools.

We made several qualitatively unexpected findings. First, many TVGs were activated or repressed at different drug dose thresholds, suggesting that rat biology include multiple layers of response mechanisms activated at different levels of toxicity severity. Second, the relatively small number of genes that were called as TVGs for all 4 drugs suggests that there may be multiple separate recovery/detoxication mechanism biologically in rats. Finally, the same drug dose can elicit very different phenotypic outcomes (including Death, Slow Recovery, and Fast Recovery) in different individuals, and these phenotypic outcomes could be reflected in the degree and duration of response in TVGs. Importantly, all of these observations required high frequency longitudinal data, and are not visible from a single before/after comparison. From the findings in this work, we advocate for broader adoption of high frequency longitudinal RNA expression profiling.

The biological significance of our findings extends beyond methodological advancement to reveal critical distinctions between immediate drug targets and secondary adaptive responses. TVG+/DEG− genes, exemplified by transient transcription factors and feedback regulators, capture pulse-like expression patterns that mediate therapeutic windows and resistance mechanisms, yet remain invisible to endpoint analyses. The dramatic threshold effect observed at 200 mg/kg tetracycline, producing a 3.3-fold increase in TVG activation compared to lower doses, demonstrates non-linear dose–response relationships that challenge conventional pharmacokinetic models and highlight the importance of temporal profiling for precision dosing strategies.

Cross-compound analysis revealing 186 commonly regulated genes across all hepatotoxic treatments establishes a core transcriptional signature for drug-induced liver injury, while the limited 4.3% overlap among 4302 total TVGs underscores the compound-specific nature of toxicity mechanisms. This dual pattern of conserved and divergent responses provides a framework for both universal safety screening and mechanism-specific risk assessment, advancing our ability to predict and prevent adverse drug reactions through temporal biomarker strategies.

The principal component analysis demonstrating clear transitions from "Healthy Region" to "Response Region" transcriptomic states offers a quantitative framework for monitoring therapeutic efficacy in real-time. These findings establish TVGs as dynamic biomarkers with superior sensitivity for detecting drug effects compared to static expression measures, enabling earlier intervention and personalized treatment optimization based on individual temporal response patterns.

This work demonstrates the potential of high-frequency longitudinal RNA sequencing as a valuable approach for characterizing dynamic transcriptomic responses to drug exposure. By capturing temporal gene expression changes, TVG analysis provides insights into the kinetic signatures of drug action that are not accessible through conventional cross-sectional approaches. Our findings suggest that this methodology may help distinguish between acute responses and adaptive processes and could potentially inform the identification of therapeutic windows in future studies. However, we acknowledge that further validation in additional drug classes, species, and clinical settings will be necessary to fully establish the translational utility of this approach. As computational frameworks continue to evolve and sampling technologies advance, longitudinal transcriptomics holds promise for improving our understanding of drug action dynamics and individual treatment variability, which may ultimately contribute to the development of more refined therapeutic strategies. In addition, matrix factorization approaches such as non-negative matrix factorization (NMF) offer a complementary framework for identifying coordinated gene programs from high-dimensional transcriptomic data[24]. Future incorporation of such data-driven module discovery methods may further strengthen longitudinal response modeling.

While our study focused on hepatotoxic compounds in rodent models, the methodological framework is broadly applicable across therapeutic areas and species. The temporal clustering approach successfully resolved biologically coherent response modules across multiple drugs, validating the robustness of longitudinal transcriptomic analysis for pharmacological research. However, translational applications will require validation in human clinical studies and development of standardized protocols for high-frequency sampling and RNA preservation.

Longitudinal RNA expression studies can be costly to conduct because the large number of RNAseq libraries that need to be constructed and sequenced. In the experiments presented here, we typically would take 10 or more daily blood samples per animal, corresponding to 5× more samples than a simple before/after dosing analysis. New technologies and methods for improving the total cost of longitudinal RNAseq studies would be greatly synergistic with and beneficial to the adoption of longitudinal RNAseq-based diagnostics. Cost reduction would need to be pursued in every step of the process, including animal handling, blood collection, RNA extraction, NGS library preparation, NGS chemistry, and data analysis.

The data analysis methods we used in this work are primarily human-driven, using traditional tools based on linear combinations of gene expression levels. Given the depth and complexity of RNA expression data, it is reasonable to assume that a well-trained deep neural network would be able to better utilize the "long tail" of information inherent within gene expression variations that are below our stringent cutoffs needed to avoid false positives. Deep neural networks[25], such as those based on transformers[26], require a large amount of data to achieve performance significantly superior to traditional machine learning models. In our dataset of 829 samples, roughly 9000 genes were expressed at non-zero levels in each sample, so the total size of this dataset is roughly 7.56 million tokens. 7.46 M tokens is large for biological datasets, but just a small early step in comparison to large language models (LLMs). For example, GPT-1[27] was trained on the BooksCorpus dataset[28], corresponding to approximately 250 M tokens. Given the remarkable emergent properties of LLMs as they are provided more parameters, floating point operations, and training data, we look forward to the impact of domain-specific biology AI[29].

Methods

Chemicals

Saline solution was purchased from Shandong Qidu Pharmaceutical (China). TRIzol was purchased from Aibixin Biotechnology (China). Chloroform was purchased from Sinopharm Chemical Reagent Company (China). Ficoll-Paque PLUS was purchased from Cytiva Life Sciences (USA). Isoniazid was purchased from MedChemExpress (China). Valproic acid and Tetracycline hydrochloride were purchased from Aladdin Scientific (China). Carbon tetrachloride were purchased from Sinopharm Chemical Reagent (China).

Animals and animal care

Animal experiments in this study were approved by the Animal Care Committee at Biostate.AI (approval number: 101) and HD Biosciences Co. Ltd (approval number: 118-8). Sprague–Dawley rats were purchased from certified provider—Charles River Laboratories International, Inc (Shanghai, China). Rats were acclimatized prior to the experiment and were housed in enriched and ventilated housing cages throughout the experimental phase. Cage litter was changed at least once a week. The rats were housed under specific-pathogen-free (SPF) condition and subjected to a normal 12 h light and night cycle, at 22 ± 2 °C and 50 ± 10% relative humidity. Chow and water were available ad libitum. The health conditions were examined and recorded daily by a veterinarian. All animal experiments were carried out in accordance with relevant guidelines and regulations, and all methods are reported in accordance with the ARRIVE guidelines (https://arriveguidelines.org). This study did not involve human participants or human tissue. Therefore, ethical approval and informed consent were not required.

Rat blood samples and liver samples collection and storage

250 μL of whole blood was collected daily from the same site of jugular vein of each rat between 10:00 am and 10:37 am. Whole blood samples were immediately processed for RNA extraction and liver function biomarker testing. On the final day of the study, rats were fasted overnight and then euthanized by overdosing with pentobarbital (120 mg/kg) through intraperitoneal injections. Whole blood and liver samples were immediately collected under RNase-free conditions. Whole blood samples were obtained via heart puncture. 1 mL of whole blood underwent the standard PBMC separation process, and the rest was preserved at − 80 °C for RNA extraction. Liver samples were preserved at − 80 °C and used for RNA extraction and histological analysis.

RNA extraction from whole blood samples

For RNA extraction from whole blood, 100 uL of whole blood was mixed with 700 uL of Trizol reagent and then underwent standard TRIzol and Chloroform RNA extraction method. Extracted RNA was further purified using Automatic Nucleotide Isolation Machine (Bioer Technology, China) with MagaBio Plus Total RNA Purification Kit (Bioer Technology, China).

For RNA extraction from liver samples, around 30 mg of chopped liver tissue was mixed with 700 uL of TriZol Reagent and Lysing MatrixD. The mixture was ground for 2 min and then mixed with 140 uL of Chloroform. After 2 min of incubation at room temperature and 10 min of centrifugation at 12,000 rpm and 4 °C, supernatant was transferred to a tube and was further purified with MagaBio Plus Total RNA Purification Kit.

Purified RNA was quantified by using Nanodrop 2000 (Thermo Fisher, USA). The quality of RNA was measured by using RNA ScreenTape Assay (Agilent, USA) and 4200 TapeStation System (Agilent, USA).

mRNA enrichment, library preparation and sequencing

Total RNA extracted from blood and liver samples was subjected to poly(A) mRNA enrichment using magnetic oligo(dT) beads. During this step, mRNA was also fragmented. cDNA libraries were prepared using a three-step protocol consisting of reverse transcription, adaptor ligation, and indexed PCR amplification, following the manufacturer's standard protocol (Vazyme, Universal V6 RNA-seq Kit). Cleanup was performed using magnetic bead-based purification after both ligation and PCR steps. Library concentration and fragment size were assessed using Qubit and Agilent TapeStation systems. Libraries were sequenced on an Illumina NovaSeq 6000 platform using 150 bp paired-end read.

Animal study—baseline study—bleeding related gene investigation

3 male and 3 female Sprague–Dawley rats (SD Rats) were acclimatized for 21 days with daily 0.25 mL blood extraction and without any substances administration. Detailed study design is shown in Supplementary Excel 3.

Animal study—toxic study

39 male SD rats were acclimatized for 10 days with daily 0.25 mL blood extraction. Rats in the Gcon group (control) were intraperitoneally injected with 0.9% sodium chloride on day 3. Rats in the GA group were administered a 1:1 mixture of corn or olive oil and carbon tetrachloride orally at different dosages (0.5, 1, 1.5, 2 ml/kg) on day 3. Rats in the GB group were intraperitoneally injected with Isoniazid dissolved in sterile saline at different dosages (400, 800, 1600, 3200 mg/kg) on day 3. Rats in the GC group were intraperitoneally injected with Valproic acid diluted in sterile saline at different dosages (250, 500, 1000, 2000 mg/kg) on day 3. Detailed study design is shown in Supplementary Excel 3.

Animal study—sub toxic study

24 male SD rats were acclimatized for 10 days with daily blood extraction. Rats in GB group were handled as described above but with different dosages (10, 30, 100, 300 mg/kg) on day 3. Rats in GD group were intraperitoneally injected with tetracycline hydrochloride dissolved in sterile saline at different dosages (4, 15, 50, 200 mg/kg) on day 3. Detailed study design is shown in Supplementary Excel 3.

Bioinformatics and data analysis

Read processing and gene expression quantification

Read mapping and quantification were performed as previously described[10]. Briefly, raw FASTQ files were trimmed to remove Illumina adaptor sequences and aligned to the *Rattus norvegicus* reference genome (NCBI GCF_015227675.2; mRatBN7.2) using HISAT2 with default parameters[30]. Gene-level expression counts were generated from the aligned reads. To account for differences in sequencing depth across samples, expression data were normalized using a multi-step median-based normalization approach, including log2 transformation and sample-level adjustment based on global expression shifts across genes.

Differential expression and identification of temporally varying genes

Differentially expressed genes (DEGs) and temporally varying genes (TVGs) were identified as previously described[10]. Briefly, paired differential expression analysis was performed using DESeq2 on raw count data after filtering out lowly expressed genes (total counts across all samples < 10)[31]. Genes were considered differentially expressed if they met the criteria $|\log2(\text{FoldChange})| > 1$ and adjusted $P < 0.05$. To further reduce false discoveries, an additional composite threshold incorporating both statistical significance and effect size was applied. TVGs were subsequently identified by aggregating significance scores across time-point comparisons. For each comparison, a score was calculated based on adjusted P values and capped to prevent dominance by a single comparison. Genes were ranked according to their cumulative scores to capture overall temporal variability, and an empirical cutoff was applied to define the final TVG set.

Temporal gene expression clustering

Log2-transformed expression data for 3206 tetracycline TVGs were obtained from the 200 mg treatment group across 10 days. Three biological replicates were averaged at each time point to generate a mean expression matrix (3206 genes × 10 time points), which was used for subsequent time-series clustering analysis. Soft clustering analysis was performed using the Mfuzz algorithm (fuzzy c-means clustering)[18] implemented in the ClusterGVis R package[19]. The optimal number of clusters was determined based on the elbow method. Prior to clustering, the expression matrix was standardized using Z-score transformation. Visualization of clustering results was conducted using ClusterGVis in combination with ComplexHeatmap[32] to generate cluster trend plots and heatmaps. GO enrichment analysis are performed using the R package clusterProfiler[11] based on rat genome annotation, and statistical cutoff of Benjamini–Hochberg (BH) corrected $P < 0.05$.

PCA analysis

We used the unions of TVGs identified from each drug, excluding 300 bleeding TVGs and low count (sum up < 10 across all samples) genes. Variance stabilizing transformation (VST) was applied to stabilize variance across the dataset for PCA analysis.

Date availability

Our dataset is publicly available for download from NCBI SRA and NCBI GEO at [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1026523] (PRJNA1026523). Because this dataset is the first high frequency longitudinal RNAseq dataset, encapsulating 4x as many samples as the previous largest rat RNAseq studies on the GEO database[33], we believe that deeper analysis than what we have performed in this work could yield additional qualitative and quantitative findings. We encourage interested readers to independently analyze and report findings from the data, and welcome academic collaborations as we continue to collect larger scale longitudinal omics data on animal models.

Code availability

Code are available upon request.

Received: 22 September 2025; Accepted: 12 March 2026
Published online: 19 March 2026

References

1. Kukurba, K. R. & Montgomery, S. B. RNA sequencing and analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).
2. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
3. Deshpande, D. et al. RNA-seq data science: From raw data to effective interpretation. *Front. Genet.* **14**, 1166 (2023).
4. Oh, S., Song, S., Grabowski, G., Zhao, H. & Noonan, J. P. Time series expression analyses using RNA-seq: A statistical approach. *Biomed. Res. Int.* **2013**, 203681 (2013).
5. Archer, S. N., Möller-Levet, C., Bonmatí-Carrión, M. -Á., Laing, E. E. & Dijk, D.-J. Extensive dynamic changes in the human transcriptome and its circadian organization during prolonged bed rest. *iScience* **27**, 109331 (2024).
6. McKibben, L. A. et al. Transcriptional changes across tissue and time provide molecular insights into a therapeutic window of opportunity following traumatic stress exposure. *Transl. Psychiatry* **15**, 244 (2025).
7. Zhou, P., Hirsch, C. N., Briggs, S. P. & Springer, N. M. Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. *Mol. Plant* **12**, 410–425 (2019).
8. Cui, Y. & Paules, R. S. Use of transcriptomics in understanding mechanisms of drug-induced toxicity. *Pharmacogenomics* **11**, 573–585 (2010).
9. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564 (2012).
10. Chen, W. et al. Novel methods for temporally varying gene identification in longitudinal studies reveal bleeding and clotting pathway activation caused by blood draws. *Brief. Bioinform.* **26**, bbaf527 (2025).

11. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
12. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
13. Chopra, I. & Roberts, M. Tetracycline antibiotics: Mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol. Mol. Biol. Rev.* **65**, 232–260 (2001).
14. Weber, L. W., Boll, M. & Stampfl, A. Hepatotoxicity and mechanism of action of haloalkanes: Carbon tetrachloride as a toxicological model. *Crit. Rev. Toxicol.* **33**, 105–136 (2003).
15. Tostmann, A. et al. Antituberculosis drug-induced hepatotoxicity: Concise up-to-date review. *J. Gastroenterol. Hepatol.* **23**, 192–202 (2008).
16. Silva, M. F. B. et al. Valproic acid metabolism and its effects on mitochondrial fatty acid oxidation: A review. *J. Inherit. Metab. Dis.* **31**, 205–216 (2008).
17. Dixon, S. J. et al. Ferroptosis: An iron-dependent form of nonapoptotic cell death. *Cell* **149**, 1060–1072 (2012).
18. Kumar, L. & Futschik, M. E. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* **2**, 5–7 (2007).
19. Zhang, J., Li, H., Tao, W. & Zhou, J. ClusterGVis: An advanced visualization and clustering tool for gene expression analysis. *Genomic Proteomics Bioinform.* qzag005 (2026).
20. Hidalgo-Gutierrez, A. et al. Guanylate kinase 1 deficiency: A novel and potentially treatable mitochondrial DNA depletion/deletions disease. *Ann. Neurol.* **96**, 1209–1224 (2024).
21. De Bortoli, F. et al. Increased versatility despite reduced molecular complexity: Evolution, structure and function of metazoan splicing factor PRPF39. *Nucleic Acids Res.* **47**, 5867–5879 (2019).
22. Hoffmann, J. J. M. L. Neutrophil CD64: A diagnostic marker for infection and sepsis. *Clin. Chem. Lab. Med.* **47**, 903–916 (2009).
23. Wu, X. et al. AKAP12 ameliorates liver injury via targeting PI3K/AKT/PCSK6 pathway. *Redox Biol.* **53**, 102328 (2022).
24. Stein-O'Brien, G. L. et al. Enter the matrix: Factorization uncovers knowledge from omics. *Trends Genet.* **34**, 790–805 (2018).
25. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *AI Open* **3**, 111–132 (2022).
26. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates Inc., 2017).
27. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. *OpenAI Tech. Rep.* (2018).
28. Zhu, Y. et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proc. IEEE Int. Conf. Comput. Vis.*19–27 (2015).
29. Orion, L., Vinayak, A., Summer, Z., Ashwin, G. & Timothy, K. K-dense analyst: Towards fully automated scientific analysis. *Preprint at* arXiv:2508.07043 (2025).
30. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
31. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
32. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
33. Barrett, T. et al. NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Res.* **41**, D991-995 (2013).

Acknowledgements

Funding for this project was provided by Biostate AI.

Author contributions

W.C. and D.Z. conceived the project. W.C., D.Z. and A.G. performed the sequences and experiments design. Q.J., X.W., Y.C. and W.C. conducted the experiments and data analysis. Z.Liu and Z.Li helped improve some experiments. O.J. helped with figures visualization improvement. K.B., B.W. and W.C. wrote the manuscript.Q.J., X.W. and W.C. revised the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

All authors listed except for Zhongwu Li and ZhengJin Liu are employees of Biostate AI Inc. David Zhang and Ashwin Gopinath are shareholders of Biostate AI Inc.

Ethics approval

Animal experiments in this study were approved by the Animal Care Committee at Biostate.AI (approval number: 101) and HD Biosciences Co. Ltd (approval number: 118-8). All methods were carried out in accordance with relevant guidelines and regulations, and all methods are reported in accordance with the ARRIVE guidelines (https://arriveguidelines.org). This study did not involve human participants or human tissue. Therefore, ethical approval and informed consent were not required.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-44560-y.

Correspondence and requests for materials should be addressed to W.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.