



# TRIAGE: an R package for regulatory gene analysis

Qiongyi Zhao<sup>1</sup>, Woo Jun Shim<sup>1</sup>, Yuliangzi Sun<sup>1</sup>, Enakshi Sinniah<sup>1</sup>, Sophie Shen<sup>1</sup>, Mikael Boden <sup>2</sup>, Nathan J. Palpant <sup>1,\*</sup>

<sup>1</sup>Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, Brisbane, QLD 4072, Australia

<sup>2</sup>School of Chemistry and Molecular Biosciences, The University of Queensland, 68 Cooper Rd, St Lucia, Brisbane, QLD 4072, Australia

\*Corresponding author. N.J.P., Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, Brisbane, QLD 4072, Australia.

Tel: +61 0733462054; E-mail: n.palpant@uq.edu.au

## Abstract

Regulatory genes are critical determinants of cellular responses in development and disease, but standard RNA sequencing (RNA-seq) analysis workflows, such as differential expression analysis, have significant limitations in revealing the regulatory basis of cell identity and function. To address this challenge, we present the TRIAGE R package, a toolkit specifically designed to analyze regulatory elements in both bulk and single-cell RNA-seq datasets. The package is built upon TRIAGE methods, which leverage consortium-level H3K27me3 data to enrich for cell-type-specific regulatory regions. It facilitates the construction of efficient and adaptable pipelines for transcriptomic data analysis and visualization, with a focus on revealing regulatory gene networks. We demonstrate the utility of the TRIAGE R package using three independent transcriptomic datasets, showcasing its integration into standard analysis workflows for examining regulatory mechanisms across diverse biological contexts. The TRIAGE R package is available on GitHub at [https://github.com/palpant-comp/TRIAGE\\_R\\_Package](https://github.com/palpant-comp/TRIAGE_R_Package).

**Keywords:** TRIAGE R package; regulatory elements; regulatory gene analysis; single-cell RNA sequencing data analysis; RNA sequencing data analysis

## Introduction

Recent advances in transcriptomics technologies have revolutionized our ability to study genome-wide expression profiles in tissues, single cells, and spatial transcriptomes. Gene expression analyses enable unsupervised discovery of gene programs governing cell biological processes. Unfortunately, highly abundant transcripts captured from gene expression analysis are often enriched for housekeeping or structural genes, which reveal fundamental aspects of cellular function [1]. In contrast, transcription factors (TFs) and other regulatory elements that control cell state identities are frequently expressed at lower levels, making them more challenging to identify. Regulatory genes play a pivotal role in shaping cellular responses during development and disease [2], yet conventional RNA sequencing (RNA-seq) analysis methods such as differential expression often overlook changes in lower-expressed regulatory elements, such as TFs and other regulatory genes, and cannot effectively prioritize them [3].

In recent years, many efforts have been made to advance regulatory gene analysis. GENIE3, e.g. is an algorithm that infers gene regulatory networks (GRNs) by predicting the expression of target genes based on the expression patterns of input genes using tree-based ensemble methods [4]. A faster implementation of GENIE3, GRNBoost2, uses gradient boosting to infer GRNs and assigns an importance score to each TF [5]. However, GRNBoost2 relies on a predefined list of TFs to guide inference, limiting its application when known TF information is unavailable or sparse. In addition, inference from transcriptomic data alone can introduce false positives by overlooking other mechanisms involved in gene

regulation. More advanced approaches, such as Lisa [6], infer GRNs from transcriptomic data by using public ChIP-seq data and chromatin accessibility profiles. Applied to gene sets from targeted TF perturbation experiments, Lisa has demonstrated improved accuracy in identifying transcriptional regulators compared to alternative methods. More recently, SCENIC+ has expanded GRN inference to single-cell RNA-seq (scRNA-seq) data by joint profiling of chromatin accessibility and gene expression in individual cells [7]. However, these tools primarily focus on TFs and their targets, leaving a gap for broader regulatory gene analysis, which should include not only TFs but also non-coding RNAs, signaling pathway components, RNA-binding proteins, and other regulatory elements.

To bridge this analytical gap, we previously developed TRIAGE (Transcriptional Regulatory Inference Analysis of Gene Expression) as a computational approach that efficiently predicts the regulatory potential of genes controlling cell identity [3]. The approach draws on consortium-level deposition data of broad H3K27me3 domains across diverse cell types to calculate genome-wide repressive tendency scores (RTS) that provide a fixed quantitative metric applied as a weight for each gene. When used as a weight to evaluate orthogonal input gene expression data, the quantitative value assigned to each gene is referred to as a TRIAGE-weighted value, also known as a discordance score (DS), which reflects the gene's potential regulatory role [3]. Building upon the foundational TRIAGE approach, we developed TRIAGE-Cluster and TRIAGE-Parser to broaden the application in more diverse analysis workflows [8]. TRIAGE-Cluster uses RTS values to

Received: September 11, 2024. Revised: December 4, 2024. Accepted: January 3, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

refine clustering from scRNA-seq data, improving the identification of cellular diversity in complex samples. TRIAGE-Parser categorizes gene–gene relationships based on shared epigenetic patterns to aid in classification of genes into groups with functional similarity. It performs principal component analysis to extract orthogonal patterns of H3K27me3 depositions from consortium-level epigenomic data [9, 10] and uses Bayesian information criterion [11] to optimally determine gene clusters. TRIAGE-Parser then assesses each gene cluster by searching the protein–protein interaction (PPI) networks from the STRING database [12] and conducts GO enrichment analysis for genes with direct PPI interactions. These methods collectively represent unique approaches to study the regulatory networks defining cellular differentiation and identity.

Despite the demonstrated utility of TRIAGE methods [13–18], the adoption has been hindered by complex interfaces and requirements. In this study, we develop an R package that integrates these methods into a user-friendly suite, providing a suite of streamlined functions that allow for seamless integration of regulatory mechanism analysis into standard workflows, thereby making these capabilities accessible to a broader range of researchers.

## Materials and methods

### Repressive tendency score and TRIAGE-prioritized genes

The EpiMap dataset [10] was downloaded from the EpiMap Repository (<https://compbio.mit.edu/epimap/>), which contains H3K27me3 signal data for 833 human biosamples. The repressive tendency score for each gene (2.5 kb upstream plus gene body) was calculated using the H3K27me3 broad domains from the EpiMap dataset, following the method described in previous study [3]. To identify TRIAGE-prioritized genes, we used a custom Python script to detect the elbow point [3]. The elbow point was determined as the point with the maximum perpendicular distance from the straight line connecting the first and last points on the curve. Genes with RTS values above this elbow point were classified as TRIAGE-prioritized genes. These genes were then annotated using several sources, including gene symbols and gene aliases from NCBI gene sources (2024.05.09), gene-disease associations from the DisGeNET platform (v7.0) [19], gene descriptions and GO Slim annotations from BioMart databases via the biomaRt R package (v2.54.1) [20, 21], gene summaries from NCBI Entrez Programming Utilities (<https://www.ncbi.nlm.nih.gov/books/NBK25500/>), and GWAS catalog data from the FUMA platform [22]. GO enrichment analysis was performed using the clusterProfiler R package (v4.6.2) [23]. SuperPath and disease category enrichment analyses were performed using the GeneAnalytics tool on the GeneCards website (<https://www.genecards.org/>) [24].

### In vivo mouse RNA-seq data analysis

The mouse bulk RNA-seq dataset was downloaded from the GEO database, with the accession ID GSE95755. It contains cardiomyocytes, fibroblasts, leukocytes and endothelial cells from infarcted and non-infarcted neonatal (P1) and adult (P56) hearts [25]. In this study, we used the data from adult mouse cardiomyocytes cell population to demonstrate the application of the TRIAGE R package on non-human species. The normalized gene expression table downloaded from the GEO database was used as the input. TRIAGEgene was then applied to transform the normalized gene expression data to TRIAGE-weighted DS values, with the ‘species’

parameter set to ‘mouse’. The ‘pvalue’ option was enabled, allowing a rank-based Z-score method to assess whether the DS value assigned to each gene was statistically higher than expected. The Jaccard index heatmap was generated using the ‘plotJaccard’ function, based on TRIAGE-weighted DS values of all DE genes. GO enrichment analysis was performed using clusterProfiler on the top 100 TRIAGE-weighted values and the top 100 genes with the smallest adjusted p-values. Enriched GO terms were then compared using custom R scripts and visualized using ggplot2 (v3.4.0). To facilitate a direct comparison between TRIAGEgene and Lisa [6], we restricted our analysis to 1611 mouse TFs based on AnimalTFDB v4.0 [26] and evaluated TF prioritization in the context of heart development. Specifically, we focused on 56 TFs within the ‘heart development’ Gene Ontology term (GO:0007507) and evaluated performance using the ROC curve. Lisa was run with default settings, with ‘—species’ set to mm10 and ‘—epigenome’ set to ‘[“DNase”, “H3K27ac”]’. The TF rankings from Lisa were based on combined p-values from its ChIP-seq and motif-based analyses, while TRIAGEgene rankings used TRIAGE-weighted DS values. The ROC curve compared true positive and false positive rates across Lisa’s two outputs and TRIAGEgene’s output. True positives were defined as TFs within the ‘heart development’ GO term, and performance was quantified using the AUC score. Additionally, GO enrichment analysis was performed on the top 136 genes identified in the TRIAGEgene analysis with  $P < .01$ . These genes were then annotated with TFs and cofactors from AnimalTFDB v4.0 and heart development-related GO terms. For genes without these specific annotations, manual literature curation was performed. All gene annotations are provided in [Supplementary Data 2](#).

### In vivo human single-cell RNA-seq data analysis

The single-cell RNA-seq dataset of peripheral blood mononuclear cells (PBMCs) was downloaded from SeuratData (v3.0.0, <https://github.com/satijalab/seurat-data>). In this experiment, PBMCs were split into an IFN- $\beta$ -treated (stimulated) group and an untreated control group [27]. To perform the joint analysis of the two groups of data, we followed the standard scRNA-seq integration strategy to perform data integration using the Seurat R package (v4.3.0) [28]. Briefly, the dataset was split into a list of two Seurat objects using the ‘SplitObject’ function. Each dataset was normalized and variable features were identified independently using the ‘NormalizeData’ and ‘FindVariableFeatures’ functions. Anchors were identified using the ‘FindIntegrationAnchors’ function with the default canonical correlation analysis method, and the data was then integrated using the ‘IntegrateData’ function. UMAP grouped by pre-defined cell type annotations was generated using the standard approach via the ‘ScaleData’, ‘RunPCA’, ‘RunUMAP’, and ‘FindNeighbors’ functions. TRIAGEcluster was then applied to this data to identify TRIAGE peaks using a bandwidth of 0.4. The ‘byPeak’ function in the TRIAGE R package was used to calculate the average gene expression for each TRIAGE peak. Subsequently, TRIAGEgene was applied to generate the TRIAGE-weighted values for each TRIAGE peak, followed by TRIAGEparser to group the top 100 DS genes into gene clusters. The ‘plotGO’ function was used to visualize the enriched STRING GO pathways in these gene clusters for TRIAGE peak0, peak1, and peak11, respectively. To compare TRIAGEcluster and Seurat clustering, we performed the Seurat clustering using the ‘FindClusters’ function and TRIAGEcluster across 50 resolutions/bandwidths (ranging from 0.1 to 5 with 0.1 increments) and applied the clustering analysis to the data obtained from both the integrated low-dimensional PCA space, as well as the integrated high-dimensional gene expression levels.

Within each TRIAGE peak and Seurat cluster, we assessed the similarity of gene expression profiles between each pair of cells using Spearman rank correlation and cosine similarity methods separately.

### In vitro human RNA-seq data analysis

This human bulk RNA-seq dataset was downloaded from the GEO database, with the accession ID GSE246079. It contains WTC-11 cells differentiated under various conditions to investigate how tranilast, a small-molecule drug, affects the regulation of cardiomyocyte differentiation [29]. For this study, RNA-seq data from cells treated under two conditions were used: 1) 1  $\mu$ M CHIR-9902 and 2) 1  $\mu$ M CHIR-9902 + 50  $\mu$ M tranilast. For the application of the TRIAGE R package, we used 1431 differentially expressed genes between the two conditions, with a fold-change of at least 2 and an adjusted p-value <0.05. TRIAGEgene was used to transform the normalized gene expression data into TRIAGE-weighted DS values. The 'topGenes' function then extracted the top 20, 50, and 100 DS genes separately, and GO enrichment analyses were performed on these gene sets using clusterProfiler. Furthermore, TRIAGEparser was applied to these 1431 DE genes for the identification of gene clusters with distinct biological functions. The 'getClusterGenes' function was used to extract genes for each gene cluster. GO enrichment analysis was performed using clusterProfiler on each gene cluster and on all DE genes separately. To prioritize GO terms related to the early stages of cardiac cell differentiation and development, we used the following keywords: 'wnt', 'embryonic organ', 'embryonic heart', 'mesoderm', 'mesenchyme', and 'pattern specification'. These keywords were used to select GO terms from the enrichment analysis results of each gene cluster and all DE genes. The comparison of enriched GO terms in these gene clusters was performed using custom R scripts and visualized with ggplot2. Gene networks for the GO terms 'mesoderm development', 'Wnt signaling pathway', and 'embryonic heart tube morphogenesis' were visualized using the 'cnetplot' function in the clusterProfiler R package.

### Building the TRIAGE R package

TRIAGE R package is written in the R programming language (version 4.2.2 and above), with the implementation of the reticulate framework to enable seamless and high-performance interoperability between Python and R.

#### TRIAGEgene

TRIAGEgene, formerly known as TRIAGE [3], has been rebranded to emphasize its focus on gene-level analysis. We have rewritten TRIAGEgene's R codebase, building upon the original TRIAGE method [3]. Key features and improvements include: (i) multi-species support: We used Ensemble BioMart to retrieve orthologous genes between humans and other species [30], allowing RTS values from human genes to be applied to corresponding orthologs in non-human datasets; (ii) a more efficient data retrieval method through specialized matrix and vector operations, accelerating processing speed; (iii) pre-set default values for 'species', 'log', and 'data\_source' parameters to simplify usage; (iv) the addition of statistics on gene numbers and percentages in the RTS table, offering insights into potential issues with gene name mapping or species selection; (v) an automated detection system for input data assessment. If the 'log' parameter is not provided, the program will evaluate data scaling to determine if a natural logarithm transformation is necessary, and will inform users of the decision made; and (vi) a rank-based Z-Score method was employed to assess whether a DS assigned to a gene is

statistically higher than expected. For each query gene, a subset of comparable genes with similar expression values is selected (by default, 0.1 percentile, though this threshold is customizable). The DS values of these comparable genes are ranked, and the ranks are transformed into Z-scores using the quantile function of the standard normal distribution. The Z-score of the query gene is then compared to those of the comparable genes, and a p-value is derived from the normal cumulative distribution function to determine if its DS value is significantly higher than that of the comparable genes.

#### TRIAGEcluster

The core functions of TRIAGEcluster, initially developed in Python, have been redeveloped to include adjustable parameters, eliminating the need for source code modification. Improvements include the integration of argparse parameter controls for increased user flexibility. Key additions and parameters are: (i) '—expr': Specifies the path to the DS or gene expression matrix; (ii) '—metadata': Requires the path to the metadata file; (iii) '—outdir': Sets the output directory, with the default being 'TRIAGEcluster\_results'; (iv) '—output\_prefix': Allows specification of the output file prefix, with the default being 'TRIAGEcluster'; (v) '—cell\_column': Defaults to 'Barcode', with an option to specify an alternative column name for cell identification; (vi) '—umap\_column': Defaults to 'UMAP\_' for UMAP coordinate columns, with the option for alternative column names; (vii) '—priority\_rts': Specifies the path to the priority RTS gene list file, with the default being 'Priority\_epimap\_rts.csv'; and (viii) '—min\_cells\_per\_peak': Ensures reporting of TRIAGE peaks containing a minimum number of cells, with the default being 5, which helps reduce noise from TRIAGE peaks with very few cells.

#### TRIAGEparser

Expanding beyond its original Python framework [8], TRIAGEparser has been seamlessly integrated into the R environment with several key improvements: (i) support for a wide range of input formats, including gene lists, tab-delimited tables, and .csv files; (ii) automatic updates to utilize the latest version of the STRING database for PPI analyses; (iii) the implementation of time delays and a requests session with built-in retry logic to effectively manage and handle potential overloading and connection issues with the STRING database server; (iv) the replacement of the 'optparse' module with 'argparse' to ensure forward compatibility; and (v) the incorporation of additional parameters, including tolerance ('—EM\_tol') and maximum iterations ('—EM\_max\_iter'), providing advanced users with greater control over the convergence criteria for the Gaussian Mixture Model fitting procedure.

#### User-friendly functions

In addition to its three primary components, the TRIAGE R package provides a variety of user-friendly functions to support and extend data analysis capabilities. These include: (i) 'plotJaccard': Generates heatmaps based on Jaccard similarity index from TRIAGEgene outputs or any gene expression data, facilitating intuitive data comparisons; (ii) 'byPeak': Generates average gene expression data at the TRIAGE peak level, which can be connected with subsequent analysis with TRIAGEparser; (iii) 'getClusterGenes': Extracts genes associated with each gene cluster from TRIAGEparser outputs; (iv) 'topGenes': Identifies the top genes with the highest values for each cell group; (v) 'plotGO': Produces heatmaps of gene ontology enrichment

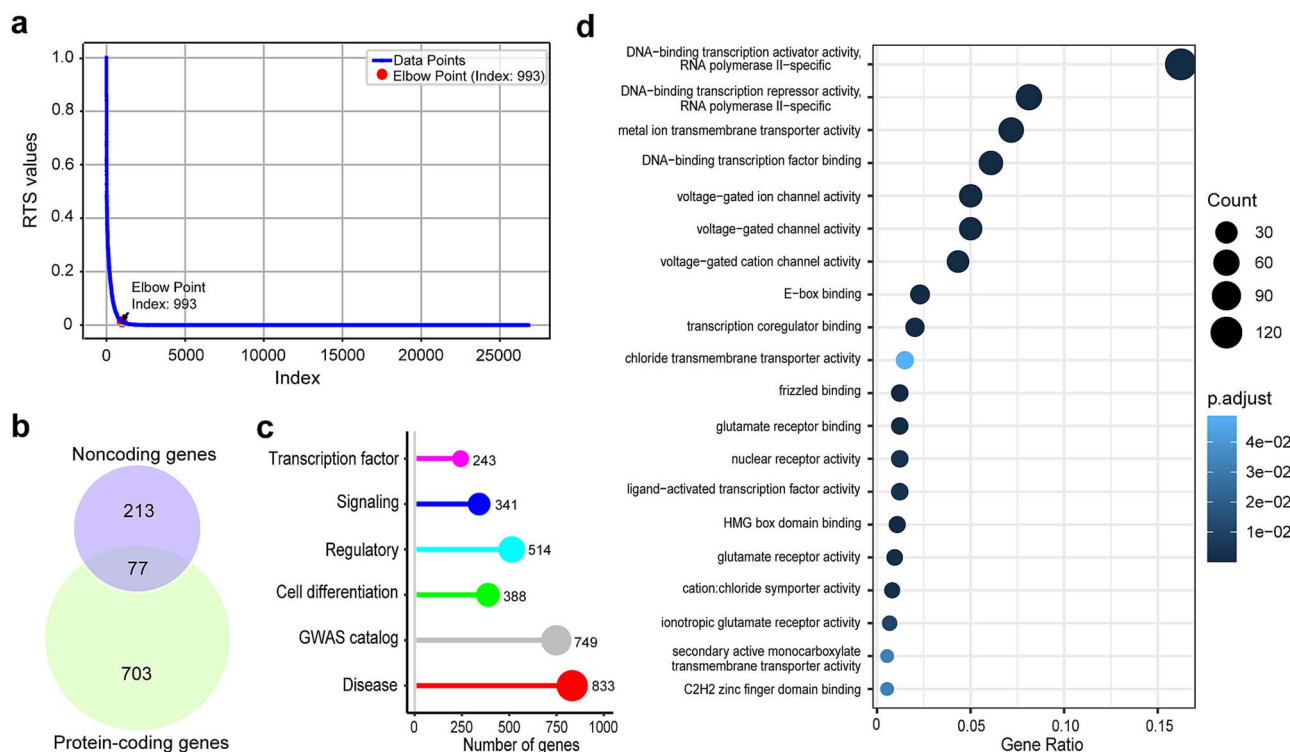


Figure 1. TRIAGE-prioritized genes are enriched in regulatory and disease-associated categories. (a) TRIAGE-prioritized genes were identified using the elbow point detection approach. (b) Venn diagram showing the distribution of coding and non-coding genes among TRIAGE-prioritized genes. (c) Functional categorization of TRIAGE-prioritized genes using a mix of databases (see methods). (d) GO enrichment analysis of TRIAGE-prioritized genes in molecular functions.

based on TRIAGEparser outputs, aiding in the visualization and interpretation of gene function distributions; (vi) ‘compareGO’: Compares GO enrichment across different gene sets, producing dot plots to visualize enrichment patterns for selected GO terms.

## Results and discussion

### TRIAGE-prioritized genes are enriched in regulatory and disease-associated categories

Leveraging 833 biological samples from diverse cell and tissue types in the EpiMap dataset [10], we calculated gene-specific RTS values. RTS captures two important features of H3K27me3 associated with genes governing cell identity: (i) the breadth of the H3K27me3 domain and (ii) the consistency of domains observed across diverse cell types. Genes with consistently broad H3K27me3 domains tend to play important regulatory roles in defining cell identity, while housekeeping genes typically lack this association. By ranking genes according to their RTS values from high to low, we identified 993 TRIAGE-prioritized genes with RTS values above the elbow point (Fig. 1a). Among these, 703 are protein-coding genes, 213 are noncoding genes, and an additional 77 genes contain both protein-coding and noncoding transcripts (Fig. 1b).

Functional categorization of these 993 genes shows that most are disease-associated (833/993, 83.89%) and/or play regulatory roles (514/993, 51.76%), with 243 being TFs, 341 involved in signaling pathways, 388 in cell differentiation, and 749 listed in the GWAS catalog (Fig. 1c). Gene ontology (GO) enrichment analysis in molecular function highlights that DNA-binding transcription activator and repressor activity (GO:0001228 and GO:0001227) are the top two significantly enriched terms, with

many other binding activity GO terms showing significant enrichment in TRIAGE-prioritized genes (Fig. 1d). SuperPath [24] and disease category enrichment analyses further support that TRIAGE-prioritized genes are enriched in pathways related to development, differentiation, and various disease categories (Supplementary Fig. 1). Building upon the RTS values and TRIAGE-prioritized genes’ framework, TRIAGE methods identify regulatory, developmental, and disease-related genes, as well as demarcate identity-defining genes in heterogeneous cellular transcriptomics data.

### Overview of the TRIAGE R package

By incorporating the TRIAGE methods, we developed the TRIAGE R package which comprises three core components: (i) TRIAGEgene, (ii) TRIAGEcluster, and (iii) TRIAGEparser, each serving distinct yet interconnected roles in processing and interpreting transcriptomic data, particularly in identifying regulatory elements. In addition, the package offers a suite of functions that integrate regulatory analysis into standard RNA-seq workflows, enabling efficient data processing and interpretation. These functions also support advanced visualization capabilities, streamlining the creation of publication-ready figures to convey complex biological insights.

TRIAGEgene utilizes pre-calculated RTS from consortium-level H3K27me3 data, integrating it with gene expression data to generate TRIAGE-weighted matrices. In complement to classic differential expression analysis, TRIAGEgene offers a novel system to rank regulatory and disease-related genes, with the ‘plotJaccard’ function for visualizing expression pattern similarities and the ‘topGenes’ function for extracting the top-ranked genes based on TRIAGE-weighted values (Fig. 2a).



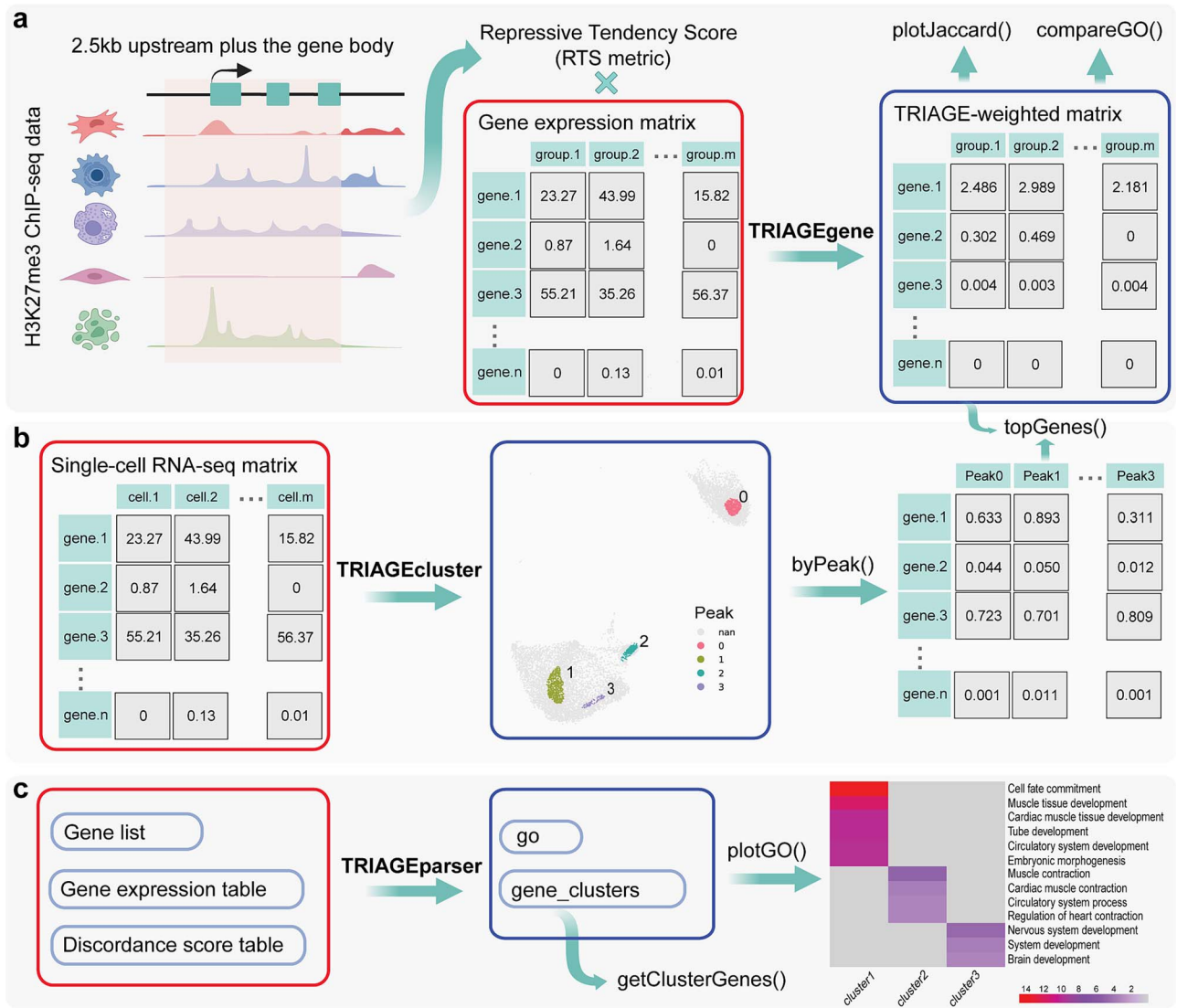


Figure 2. Overview of functions in the TRIAGE R package. (a) Repressive tendency scores were calculated by analyzing broad H3K27me3 domains in the gene region, with TRIAGEgene integrating this data with gene expression data to generate TRIAGE-weighted matrices. The 'plotJaccard' function visualizes the Jaccard similarity index between groups as a heatmap, the 'compareGO' function compares GO enrichment across different gene sets, producing dot plots to visualize enrichment patterns for selected GO terms, and the 'topGenes' function identifies the top genes with the highest TRIAGE-weighted values for each group. (b) TRIAGEcluster refines cell clustering by identifying more specific 'TRIAGE peaks' demarcating biologically distinct cell populations from scRNA-seq matrices. The 'byPeak' function interfaces with TRIAGEcluster's output to produce peak-level gene expression data and the 'topGenes' function facilitates the identification of top genes with the highest values for each TRIAGE peak. (c) TRIAGEparser, taking a gene list or table as input, identifies gene clusters along with their gene ontologies. The 'getClusterGenes' and 'plotGO' functions interface with TRIAGEparser's output for the extraction of genes from each cluster and for the visualization of gene ontology enrichment, respectively. Part of the visuals in this figure were created with [BioRender.com](https://www.biorender.com).

TRIAGEcluster, applied to scRNA-seq data, leverages the RTS framework to demarcate identity-defining genes in heterogeneous cellular transcriptomics data. It employs weighted kernel density estimation to identify distinct cell types, referred to as TRIAGE peaks, in a 2D space. TRIAGEcluster accepts a normalized gene expression matrix and corresponding metadata file as input, allowing seamless integration into other scRNA-seq analysis workflows. The output includes a set of Uniform Manifold Approximation and Projections (UMAPs) with various bandwidth resolutions, along with a set of metadata files for each bandwidth. Each TRIAGE peak in the UMAP represents a predicted cell population. Similar to cluster analysis using Seurat 'FindClusters' where users need to select a suitable resolution, users can visualize the TRIAGE peaks generated by TRIAGEcluster

and choose a suitable bandwidth resolution for their study. The 'byPeak' function generates peak-level data, as well as any desired subsets of the data, for downstream analyses (Fig. 2b).

TRIAGEparser parses gene lists, such as those enriched in TRIAGE peaks or obtained from differential expression analysis, into gene clusters with distinct biological functions. It also explores protein-protein interaction networks using data from the STRING database. TRIAGEparser accepts either a gene list or a table as input. The output includes gene clusters and associated GO enrichments for each gene cluster, organized into separate folders named 'gene\_cluster' and 'go'. The 'getClusterGenes' function helps to extract genes in each gene cluster for further analysis, while the 'plotGO' function visualizes STRING GO enrichment outcomes for gene clusters (Fig. 2c).

These functions facilitate the use of the TRIAGE R package as an efficient and versatile toolkit, enabling comprehensive analysis of regulatory elements and making it easy to adapt to other R environment tools such as DESeq2 [31] and edgeR [32] for bulk RNA-seq data analysis, as well as the Seurat workflow [28] for scRNA-seq data analysis across various biological contexts. We demonstrated its adaptability to standard RNA-seq analysis workflows using three different datasets: an *in vivo* mouse RNA-seq dataset, an *in vitro* human RNA-seq dataset, and an *in vivo* human scRNA-seq dataset, showcasing the package's broad utility for analyzing regulatory mechanisms across both bulk and single-cell transcriptomics modalities. TRIAGE R package is designed to accommodate various data scales, with runtime and memory usage manageable on personal computers for typical studies, including the following case studies (see details in [Supplementary Data 1](#)). Furthermore, the package's scalability was tested across larger dataset sizes, proving its suitability even for consortium-level data volumes ([Supplementary Fig. 2](#) and [Supplementary Data 1](#)).

### Case study 1: Apply the TRIAGE R package to an *in vivo* mouse RNA-seq dataset

A key feature of TRIAGEgene in the TRIAGE R package is its simple application to transcriptomic data from other chordate species. Given the conservation of H3K27me3 patterns across eukaryotes [33], the RTS calculation for a gene can be applied to other species by orthology despite being generated using human H3K27me3 data, as demonstrated previously [3]. In this R package, TRIAGEgene incorporates the inter-species gene-mapping process to find orthologous genes between the human and the species of interest, enabling users to easily extend the TRIAGEgene analysis to other model animals like mice, zebrafish, pigs, etc., facilitating broader research applications. Notably, while the direct application to more distantly related species, such as plants, poses challenges due to limited orthology with humans, the underlying TRIAGE concept could be adapted. Specifically, plant-specific H3K27me3 data could be leveraged for regulatory gene analysis in plants and related pathologies. To support such adaptations, the TRIAGE R package allows users to input custom RTS files, facilitating the extension of the analysis framework to more distant species.

Here, we applied TRIAGEgene to an *in vivo* mouse RNA-seq dataset [25] to demonstrate its utility in multi-species support and regulatory gene prioritization. This RNA-seq dataset contains cardiomyocyte cell populations extracted from adult mice with myocardial infarction (MI) and from sham-operated (Sham) controls. The study identified a total of 658 differentially expressed (DE) genes between the Sham and MI groups (adjusted *p*-value < 0.05) using the edgeR analysis pipeline [25]. TRIAGEgene transformed normalized gene expression values (counts per million) into TRIAGE-weighted DS values. The 'plotJaccard' function in the TRIAGE R package was used to visualize sample similarities. All four biological replicates were well-grouped in the Jaccard similarity analysis ([Fig. 3a](#)). Next, we investigated the biological functions represented by the top-ranked DS genes in the MI group. Manual examination of the top ten DS genes revealed that all these genes are crucial regulators involved in cardiac development, differentiation, and disease ([Table 1](#)), demonstrating that TRIAGEgene can effectively prioritize regulatory genes.

We then selected the top 100 genes with the highest DS values and observed that only three of these genes were differentially expressed between the Sham and MI groups. Notably, there was no overlap between the top 100 DS genes and the top 100 DE genes ranked by adjusted *p*-value ([Fig. 3b](#)). GO enrichment analysis of

these two gene sets indicated that genes ranked by DE *p*-values were more likely involved in wound healing, immune system processes, and immune response, whereas genes with the highest DS values were significantly enriched in cardiac development, cell fate commitment, and cell differentiation terms ([Fig. 3c](#)). It is expected that after MI, genes involved in wound healing and immune system processes change, and these are indeed shown in the top DE gene list. However, studies focusing on the regulatory basis of cell responses to MI will benefit from the use of TRIAGEgene which uses a unique ranking system to prioritize these genes for further downstream analysis.

Next, we sought to evaluate the performance of TRIAGEgene against other regulatory analysis tools. Since there are currently no other tools like the TRIAGE R package that prioritize regulatory genes beyond TFs, we selected Lisa for comparison, as it also uses a gene expression matrix as input, similar to TRIAGEgene, and leverages public chromatin accessibility and ChIP-seq data to enhance performance over alternative methods. However, since Lisa and other tools, such as GENIE3 and GRNBoost2, are limited to prioritizing TFs, we restricted the comparison to mouse TFs. Lisa produced two sets of results using ChIP-seq and motif-based methods, ranking TFs based on a combined *p*-value [6]. We assessed how Lisa and TRIAGEgene prioritized heart development-related TFs within the set of all mouse TFs. TRIAGEgene demonstrated a higher capacity to prioritize TFs relevant to the given biological context (AUC = 0.73), surpassing both Lisa's ChIP-seq and motif-based methods ([Fig. 3d](#)), as it effectively prioritizes heart developmentally relevant TFs in cardiomyocyte cell populations. Furthermore, the top 136 genes identified in the TRIAGEgene analysis (*P* < .01) are significantly enriched in heart development and signaling-related regulatory categories beyond TFs ([Fig. 3e](#) and [Supplementary Data 2](#)). As shown here, TRIAGEgene can be used as a standalone tool for regulatory analysis, making it broadly applicable in various biological contexts. It is worth noting that TRIAGEgene is not designed to replace classic differential expression analysis, but rather to complement it by prioritizing regulatory, developmental, and disease-related genes, thereby providing a gene regulatory perspective into the RNA-seq data analysis workflow.

### Case study 2: Apply the TRIAGE R package to an *in vitro* human scRNA-seq dataset

For the regulatory analysis in scRNA-seq data, we demonstrate the application of the TRIAGE R package on a dataset containing two groups of human peripheral blood mononuclear cells (PBMCs): IFN- $\beta$ -stimulated cells and control cells [27]. We processed and integrated the PBMCs dataset using the standard Seurat integration pipeline, visualizing the data in UMAPs grouped by pre-defined cell type annotations ([Fig. 4a](#)). TRIAGEcluster was then applied to this integrated PBMC dataset, resulting in the identification of 14 TRIAGE peaks ([Fig. 4b](#)), each representing a biologically related cell population characterized by a unique set of regulatory genes that define cell identity and function. To assess the effectiveness of TRIAGEcluster in identifying different cell populations, we further evaluated the clustering efficacy based on the assumption that cells within each cluster should have similar gene expression profiles, reflected by cell-cell correlation within each cluster of the integrated data. We compared the performance of TRIAGEcluster with Seurat 'FindClusters' by assessing the similarity of gene expression profiles between each pair of cells within TRIAGE peaks or Seurat clusters across a range of clustering resolutions with different numbers of clusters.

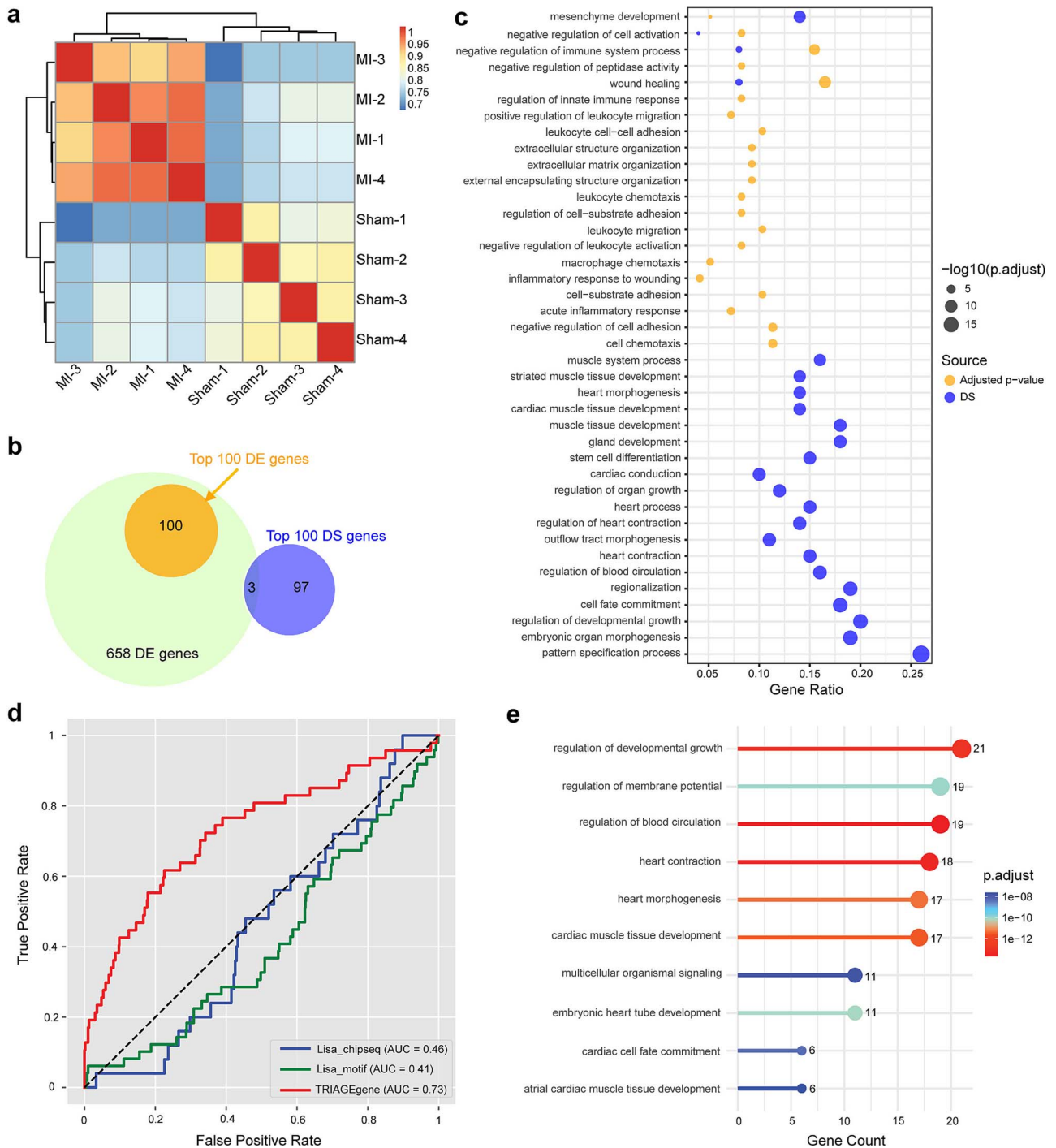


Figure 3. TRIAGE R package offers a novel system to rank regulatory genes. (a) Jaccard index heatmap showing sample similarities using the 'plotJaccard' function in the TRIAGE R package. (b) Venn diagram showing the distribution of the top 100 DS genes, top 100 DE genes, and all DE genes. (c) Comparisons of the top 20 enriched GO terms between the top 100 DS genes and the top 100 DE genes. (d) Comparison of TRIAGEgene and Lisa in prioritizing transcription factors relevant to heart development. Both ChIP-seq and motif-based methods in Lisa were used for comparison. (e) Enrichment of regulation, heart development, and signaling-related GO terms among top-ranked genes identified by TRIAGEgene. Genes with  $P$ -value  $< 0.01$  were used for GO enrichment analysis.

Comparisons between Seurat and TRIAGE revealed higher intra-cluster correlations of TRIAGE peaks over Seurat clusters in both low-dimensional PCA and high-dimensional integrated data, as determined by the Wilcoxon rank-sum test, irrespective of the correlation methods applied (Fig. 4c). These results demonstrate that TRIAGEcluster can effectively distinguish biologically relevant cell populations in scRNA-seq datasets, offering more accurate

representation of cell populations in both PCA and integrated datasets compared to Seurat.

To explore the regulatory basis of cells in TRIAGE peaks, we first used the 'byPeak' function to generate an average gene expression matrix for each TRIAGE peak (Supplementary Data 3). This matrix was subsequently transformed into DS values using TRIAGEgene. We then used the 'topGenes' function to extract

Table 1. Top 10 genes ranked by TRIAGE-weighted values (discordance scores)

Symbol	Annotation	Reference
Gata4	Zinc-finger transcription factor, a critical regulator of cardiac gene expression	Oka et al. [34]
Nkx2-5	Homeobox-containing transcription factor, a crucial regulator of cardiac development, regeneration and diseases	Cao et al. [35]
Tbx5	T-box transcription factor, a key regulator of heart development	Steimle et al. [36]
Irx4	Iroquois homeobox transcription factor, involved in heart development and function	Nelson et al. [37]
Ntn1	Prevents the development of cardiac hypertrophy and heart failure through the regulation of MEK-ERK1/2 and JNK1/2 signaling pathways.	Wang et al. [38]
Scn4b	Encodes a $\beta$ -subunit for the voltage-gated cardiac sodium channel complex, involved in generation and conduction of the cardiac action potential	Yang et al. [39]
Rnf220	E3 ubiquitin ligase, involved in cardiac development and disease	van de Vegte et al. [40]
Metap1d	A mitochondrial protein involved in adaptive responses of the heart	Arumugam et al. [41]
Hand2	A transcription factor involved in right ventricular remodelling	Videira et al. [42]
Tbx20	T-box transcription factor, a vital regulator of direct human cardiac reprogramming	Tang et al. [43]

the genes with the highest DS values, prioritizing key regulatory genes within each TRIAGE peak. As a demonstration, we identified the top ten DS genes in TRIAGE peaks 0, 1, and 11, corresponding to CD14<sup>+</sup> monocytes, CD4<sup>+</sup> memory T cells, and activated T cells, respectively, almost all of them were involved in regulatory or signaling pathways (Fig. 4d and Supplementary Fig. 3). Next, TRIAGEparser was used to group the top 100 DS genes into biologically relevant gene clusters and to identify enriched GO terms through the STRING protein–protein interaction network for each gene cluster. The ‘plotGO’ function was used to visualize enriched STRING GO pathways (Fig. 4e and Supplementary Fig. 3). The ‘getClusterGenes’ function can also be used here to extract gene lists from any gene cluster, facilitating downstream analyses such as GO and KEGG enrichment analyses to further investigate the enriched pathways involved in each gene cluster. These showcase the interconnected roles of functions in the TRIAGE R package for regulatory analysis.

These functionalities illustrate the integrated roles of the TRIAGE R package in conducting comprehensive regulatory analyses, including the identification of cell clusters (designated as TRIAGE peaks), characterization of regulatory components within these clusters, pathway enrichment studies, and the elucidation of complex regulatory networks involved in the gene clusters identified within each cell cluster. It is worth noting that the ‘byPeak’ function allows for the extraction of any cell group from scRNA-seq datasets and thus can also be applied to Seurat clusters or any group of cells of interest. In addition, the ‘topGenes’ function can also be applied to identify the top highly expressed genes when the input is a gene expression matrix. These features facilitate a range of downstream applications, including the identification of marker genes in scRNA-seq data and regulatory analysis by integrating other functions in the TRIAGE R package.

### Case study 3: Apply the TRIAGE R package to an *in vitro* human RNA-seq dataset

We further applied the TRIAGE R package to an *in vitro* human RNA-seq dataset, which includes human induced pluripotent stem cells (hiPSCs) differentiated into the cardiac lineage under two different conditions: (i) cells treated under standard protocol conditions induced with the small molecule CHIR-99021 (CHIR), and (ii) cells treated with an experimental protocol involving CHIR supplemented with a Wnt-agonist tranilast [29]. A total of 1431 DE genes were identified in the comparison between these

two conditions using the DESeq2 workflow [29]. To explore the regulatory genes orchestrating the influence of Wnt activity on the fate of hiPSCs during differentiation, we performed regulatory analysis on these DE genes. TRIAGEgene was used to transform the normalized gene expression data into DS values, prioritizing regulatory genes by ranking them in descending order. Subsequently, GO enrichment analyses were conducted separately for the top 20, 50, and 100 DS genes. By examining GO terms related to Wnt signaling and the differentiation of hiPSCs into the cardiac lineage, we identified five enriched GO terms across all three sets of analyses. Notably, the analysis based on the top 20 DS genes yielded the most significant results (Fig. 5a). Among these, the three most significant GO terms: ‘cardiac cell fate commitment,’ ‘cardiac muscle cell fate commitment,’ and ‘Wnt signaling pathway involved in heart development,’ were present in all three sets of analyses. Three key regulatory genes, SOX17, TBX3, and WNT3A, were identified from the top 20 DS genes in relation to these three GO terms. SOX17 is a key transcriptional regulator involved in myocardial development [44] and interacts with the canonical Wnt pathway to specify and pattern the endoderm [45]. TBX3, a TF, modulates the expression of Wnt target genes in a context-dependent manner [46]. WNT3A, one of the Wnt family members, regulates cardiac progenitor self-renewal [47]. This demonstrates that the TRIAGE R package effectively prioritizes regulatory genes and reveals the regulatory basis of cell identity and function.

Furthermore, TRIAGEparser can also function as a standalone tool to group genes into biologically relevant gene clusters and pinpoint regulatory components in the gene clusters. To illustrate this, we applied TRIAGEparser directly to the 1431 DE genes. TRIAGEparser identified three major gene clusters (Supplementary Data 4), and we used the ‘getClusterGenes’ function to extract the genes from each gene cluster. GO enrichment analysis was then performed for the genes within each cluster as well as for the full set of DE genes. The analysis revealed distinct pathways associated with each gene cluster (Supplementary Fig. 4), demonstrating that TRIAGEparser effectively categorized the genes in a biologically meaningful manner. Importantly, the enriched GO terms identified in the three gene clusters included nine of the top 10 GO terms that were enriched when analyzing all DE genes, underscoring the robustness of TRIAGEparser in capturing key biological processes. Additionally, gene cluster 2 highlighted several additional GO terms, such as ‘embryonic organ morphogenesis,’ ‘mesenchyme development,’ ‘anterior/posterior pattern specification,’ ‘heart



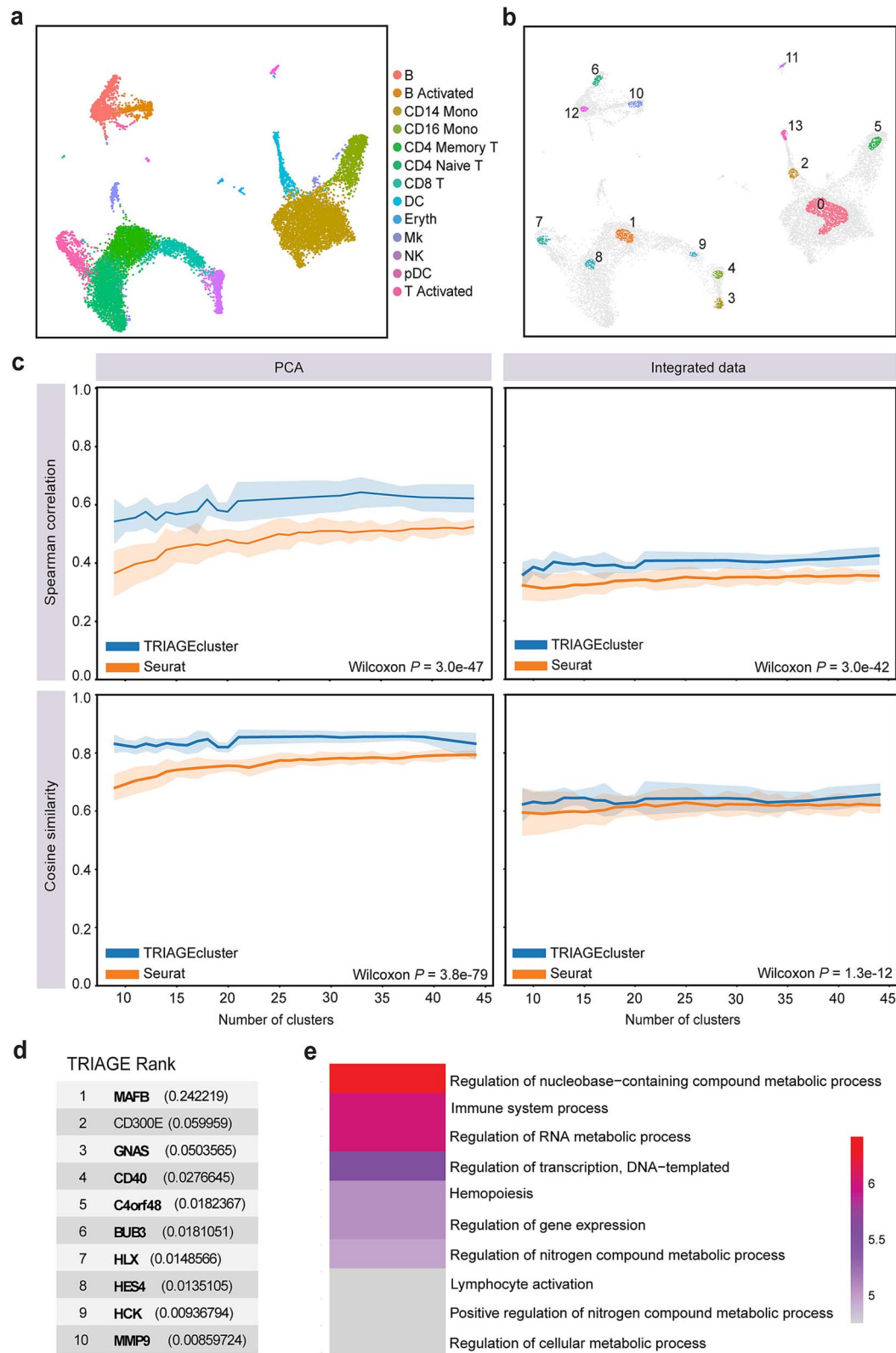


Figure 4. TRIAGE R package effectively distinguishes biologically relevant cell populations and uncovers the regulatory basis of these populations in single-cell RNA-seq datasets. (a) UMAP plots showing the integrated PBMC datasets coloured by pre-defined cell type annotations. (b) UMAP representation of TRIAGE peaks identified by TRIAGEcluster in the integrated PBMC datasets. (c) Line plots showing intra-cluster correlation assessments using Spearman rank correlation (upper) and cosine similarity (bottom) in both low-dimensional PCA (left) and high-dimensional integrated data (right) spaces. (d) The top 10 DS genes within TRIAGE peak 0, corresponding to CD14<sup>+</sup> monocytes, were identified using the 'topGenes' function. DS values are listed in parentheses. Regulatory and/or signaling genes are highlighted in bold. (e) Enriched STRING GO terms are displayed for the TRIAGEparser gene cluster derived from the top 100 DS genes within TRIAGE peak 0.

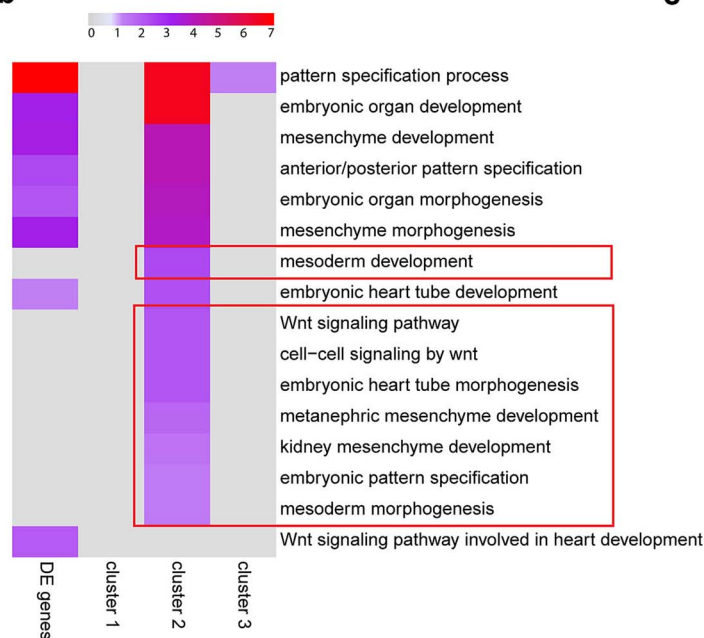
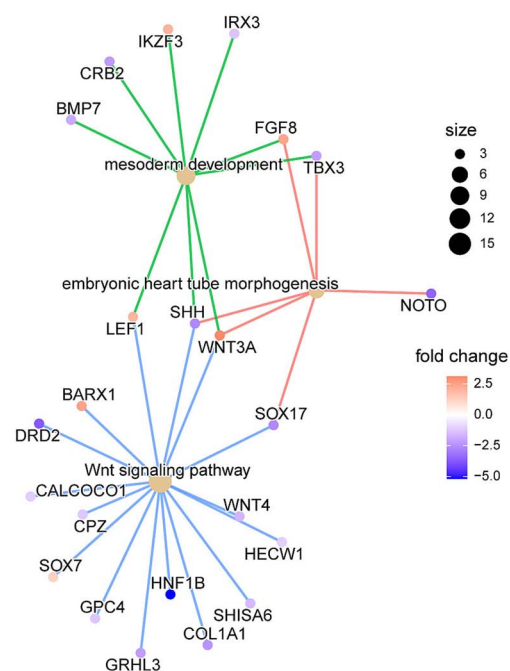
**a****b****c**

Figure 5. TRIAGE R package identifies key regulatory genes and groups them into biologically relevant clusters, facilitating the identification of regulatory gene networks. (a) Shown are enriched GO terms related to Wnt signaling and the differentiation of hiPSCs into the cardiac lineage, identified from GO enrichment analyses using the top 20, 50, and 100 DS genes from the *in vitro* human RNA-seq dataset. (b) GO enrichment analysis of TRIAGEparser gene clusters compared to all DE genes, focusing on terms related to Wnt signaling and embryonic organ/heart development. Development-related GO terms, exclusively identified in gene cluster 2, are highlighted. (c) Gene networks representing the top three enriched GO terms unique to gene cluster 2.

morphogenesis', and 'embryonic organ development', all of which are crucial for cell differentiation and heart development (Supplementary Fig. 4). These findings underscore the ability for TRIAGEparser to reveal gene subsets with distinct biological functions. Since this study focuses on early stages of cardiac cell differentiation and the role of the Wnt signaling pathway, we selected GO terms related to Wnt signaling and embryonic organ/heart development for further comparisons among these gene clusters. Notably, gene cluster 2 captured all but one of the GO terms enriched in the analysis of all DE genes, except for 'Wnt signaling pathways involved in heart development'. However, gene cluster 2 did capture two other Wnt signaling pathway-related terms: 'Wnt signaling pathway' and 'cell-cell signaling by wnt'. In addition, gene cluster 2 successfully identified several development-related terms that were not evident in the analysis of all DE genes (Fig. 5b). To explore the underlying regulatory gene networks, we examined the top three enriched GO terms unique to gene cluster 2—'mesoderm development,' 'Wnt signaling pathway,' and 'embryonic heart tube morphogenesis' - which were not found in the analysis of all DE genes. We identified four key node genes: *LEF1* and *WNT3A* upregulated following tranilast treatment, while *SHH* and *SOX17* were downregulated (Fig. 5c). Interestingly, *WNT3A* and *SOX17* were also identified in the above regulatory analysis from the top 20 DS genes, both playing key regulatory roles in cardiac development. *LEF1* is a TF that regulates endothelial-to-mesenchymal transition, proliferation, and differentiation [48]. *SHH* is crucial in morphogenesis, organogenesis, left-right asymmetry, and a potential cardiac therapeutic target [49]. These data illustrate the ability of TRIAGEparser to identify biologically relevant gene clusters and their underlying regulatory components.

## Limitations and future development plans

Currently, the TRIAGE R package relies on existing gene annotations, limiting its ability to prioritize regulatory genes or regions outside of these annotations. It is not yet capable of prioritizing novel transcripts, such as novel long non-coding RNAs (lncRNAs). To address this limitation, one key direction for future development will be extending the TRIAGE analysis to single-base resolution on a genome-wide scale [50]. This will enable the identification of novel regulatory transcripts not included in current annotations, such as lncRNAs, as well as the prioritization of any genomic regions with regulatory potential. In future releases, TRIAGE will enable users to input genomic regions of interest, such as genomic coordinates of novel lncRNAs, and rank them based on their potential regulatory roles. Additionally, we will continue to develop visualization functions specifically designed for regulatory gene analysis, enhancing users' ability to interpret and present their findings effectively. Although the TRIAGE R package efficiently handles various data scales and runs on personal computers for typical studies, consortium-level single-cell projects may require high-performance computing systems to meet memory demands. We will continue optimizing performance and resource management to better support large-scale datasets in future updates.

## Conclusion

We developed the TRIAGE R package as a toolkit for regulatory gene analysis in both bulk and single-cell RNA-seq datasets. TRIAGEgene prioritizes potential regulatory genes, TRIAGEcluster identifies biologically related cell types in scRNA-seq data, and

TRIAGEparser facilitates discovery of cell type regulatory pathways. Together, these tools provide insights into the regulatory networks underlying development and disease. By consolidating these capabilities into a user-friendly package, along with a suite of functions for efficient data processing, interpretation, and visualization, the TRIAGE R package makes these analyses accessible to researchers with basic R knowledge. Its seamless integration into existing RNA-seq workflows positions it as a useful resource for exploring regulatory elements and mechanisms across diverse biological systems, with potential for novel discoveries in both biological and medical research.

### Key Points

- We present the TRIAGE R package, for analyzing regulatory elements in bulk and single-cell RNA sequencing datasets.
- TRIAGEgene introduces a novel ranking system to prioritize regulatory genes.
- TRIAGEcluster demarcates regulatory identity-defining genes and refines cell clustering from single-cell RNA sequencing data.
- TRIAGEparser parses gene lists into gene clusters with distinct biological functions and pinpoints regulatory components in gene clusters.
- The package facilitates efficient and adaptable pipelines for regulatory gene analysis, seamlessly integrating into standard RNA-seq workflows.

## Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

## Acknowledgements

We would like to thank Sumedha Negi for testing the TRIAGE R package on multiple platforms.

## Data availability

All data used in this study are publicly available. To facilitate the reproducibility of our analyses and to ensure that users can fully utilize the TRIAGE R package in various scenarios, we have provided detailed analysis steps in Supplementary Data 5. The TRIAGE R package, along with all source data and custom scripts used in this study are publicly accessible on GitHub at [https://github.com/palant-comp/TRIAGE\\_R\\_Package](https://github.com/palant-comp/TRIAGE_R_Package). Complete documentation is available at <https://triage-r-package.readthedocs.io/en/latest/index.html>.

## Funding

This work was supported by the National Heart Foundation of Australia (106721 to N.J.P.) and the Medical Research Future Fund (APP2016033 to N.J.P.).

Conflict of interest: None declared.

## Author contributions

Q.Z. conceived the study, developed the TRIAGE R package, performed the analyses, and wrote the original draft. W.J.S., Y.S., E.S., S.S., and M.B. assisted in the development of the TRIAGE R package, performed the analyses, and reviewed and edited

the manuscript. N.J.P. conceived the study, provided resources, supervised the research, and reviewed and edited the manuscript.

## References

- Ramskold D, Wang ET, Burge CB. et al. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 2009;**5**:e1000598. <https://doi.org/10.1371/journal.pcbi.1000598>.
- Morris SA, Daley GQ. A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell Res* 2013;**23**: 33–48. <https://doi.org/10.1038/cr.2013.1>.
- Shim WJ, Sinniah E, Xu J. et al. Conserved epigenetic regulatory logic infers genes governing cell identity. *Cell Syst* 2020;**11**:625–639.e13. <https://doi.org/10.1016/j.cels.2020.11.001>.
- Huynh-Thu VA, Irrthum A, Wehenkel L. et al. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;**5**:e12776. <https://doi.org/10.1371/journal.pone.0012776>.
- Moerman T, Aibar Santos S, Bravo González-Blas C. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 2019;**35**:2159–61. <https://doi.org/10.1093/bioinformatics/bty916>.
- Qin Q, Fan J, Zheng R. et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol* 2020;**21**:32. <https://doi.org/10.1186/s13059-020-1934-6>.
- Bravo Gonzalez-Blas C, De Winter S, Hulselmans G. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;**20**:1355–67. <https://doi.org/10.1038/s41592-023-01938-4>.
- Sun Y, Shim WJ, Shen S. et al. Inferring cell diversity in single cell data using consortium-scale epigenetic data as a biological anchor for cell identity. *Nucleic Acids Res* 2023;**51**:e62. <https://doi.org/10.1093/nar/gkad307>.
- Roadmap Epigenomics Consortium, Meuleman W, Ernst J. et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30. <https://doi.org/10.1038/nature14248>.
- Boix CA, James BT, Park YP. et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;**590**: 300–7. <https://doi.org/10.1038/s41586-020-03145-z>.
- Konishi S, Kitagawa G. *Information Criteria and Statistical Modeling*. Springer Science & Business Media, New York, 2008.
- Szklarczyk D, Gable AL, Lyon D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13. <https://doi.org/10.1093/nar/gky1131>.
- Kojic M, Gawda T, Gaik M. et al. Elp2 mutations perturb the epitranscriptome and lead to a complex neurodevelopmental phenotype. *Nat Commun* 2021;**12**:2678. <https://doi.org/10.1038/s41467-021-22888-5>.
- Wehrens M, de Leeuw AE, Wright-Clark M. et al. Single-cell transcriptomics provides insights into hypertrophic cardiomyopathy. *Cell Rep* 2022;**39**:110809. <https://doi.org/10.1016/j.celrep.2022.110809>.
- Afonso J, Shim WJ, Boden M. et al. Repressive epigenetic mechanisms, such as the H3K27me3 histone modification, were predicted to affect muscle gene expression and its mineral content in Nelore cattle. *Biochem Biophys Res* 2023;**33**:101420. <https://doi.org/10.1016/j.bbrep.2023.101420>.
- Plaisance I, Chouvardas P, Sun Y. et al. A transposable element into the human long noncoding RNA CARMEN is a switch for cardiac precursor cell specification. *Cardiovasc Res* 2023;**119**: 1361–76. <https://doi.org/10.1093/cvr/cvac191>.
- Friedman CE, Cheetham SW, Negi S. et al. HOPX-associated molecular programs control cardiomyocyte cell states underpinning cardiac structure and function. *Dev Cell* 2024;**59**:91–107.e6.
- Qiu C, Cao J, Martin BK. et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat Genet* 2022;**54**: 328–41. <https://doi.org/10.1038/s41588-022-01018-x>.
- Pinero J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–9. <https://doi.org/10.1093/nar/gkw943>.
- Durinck S, Moreau Y, Kasprzyk A. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005;**21**:3439–40. <https://doi.org/10.1093/bioinformatics/bti525>.
- Durinck S, Spellman PT, Birney E. et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009;**4**:1184–91. <https://doi.org/10.1038/nprot.2009.97>.
- Watanabe K, Taskesen E, van Bochoven A. et al. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826. <https://doi.org/10.1038/s41467-017-01261-5>.
- Wu T, Hu E, Xu S. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;**2**:100141. <https://doi.org/10.1016/j.xinn.2021.100141>.
- Belinky F, Nativ N, Stelzer G. et al. PathCards: multi-source consolidation of human biological pathways. *Database (Oxford)* 2015;**2015**:bav006. <https://doi.org/10.1093/database/bav006>.
- Quaife-Ryan GA, Sim CB, Ziemann M. et al. Multicellular transcriptional analysis of mammalian heart regeneration. *Circulation* 2017;**136**:1123–39. <https://doi.org/10.1161/CIRCULATIONAHA.117.028252>.
- Shen WK, Chen SY, Gan ZQ. et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res* 2023;**51**: D39–45. <https://doi.org/10.1093/nar/gkac907>.
- Kang HM, Subramaniam M, Targ S. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* 2018;**36**:89–94. <https://doi.org/10.1038/nbt.4042>.
- Stuart T, Butler A, Hoffman P. et al. Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>.
- Wu Z, Shen S, Mizikovskiy D. et al. Wnt dose escalation during the exit from pluripotency identifies tranilast as a regulator of cardiac mesoderm. *Dev Cell* 2024;**59**:705–722.e8. <https://doi.org/10.1016/j.devcel.2024.01.019>.
- Kinsella RJ, Kahari A, Haider S. et al. Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database (Oxford)* 2011;**2011**:bar030. <https://doi.org/10.1093/database/bar030>.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. <https://doi.org/10.1186/s13059-014-0550-8>.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
- Arthur RK, Ma L, Slattery M. et al. Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. *Genome Res* 2014;**24**:1115–24. <https://doi.org/10.1101/gr.162008.113>.



34. Oka T, Maillet M, Watt AJ. et al. Cardiac-specific deletion of Gata4 reveals its requirement for hypertrophy, compensation, and myocyte viability. *Circ Res* 2006;**98**:837–45. <https://doi.org/10.1161/01.RES.0000215985.18538.c4>.
35. Cao C, Li L, Zhang Q. et al. Nkx2.5: a crucial regulator of cardiac development, regeneration and diseases. *Front Cardiovasc Med* 2023;**10**:1270951. <https://doi.org/10.3389/fcvm.2023.1270951>.
36. Steimle JD, Moskowitz IP. TBX5: a key regulator of heart development. *Curr Top Dev Biol* 2017;**122**:195–221. <https://doi.org/10.1016/bs.ctdb.2016.08.008>.
37. Nelson DO, Jin DX, Downs KM. et al. Irx4 identifies a chamber-specific cell population that contributes to ventricular myocardium development. *Dev Dyn* 2014;**243**:381–92. <https://doi.org/10.1002/dvdy.24078>.
38. Wang N, Cao Y, Zhu Y. Netrin-1 prevents the development of cardiac hypertrophy and heart failure. *Mol Med Rep* 2016;**13**: 2175–81. <https://doi.org/10.3892/mmr.2016.4755>.
39. Yang Q, Xiong H, Xu C. et al. Identification of rare variants in cardiac sodium channel beta4-subunit gene SCN4B associated with ventricular tachycardia. *Mol Genet Genomics* 2019;**294**:1059–71. <https://doi.org/10.1007/s00438-019-01567-7>.
40. van de Vegte YJ, Teegene BS, Verweij N. et al. Genetics and the heart rate response to exercise. *Cell Mol Life Sci* 2019;**76**:2391–409. <https://doi.org/10.1007/s00018-019-03079-4>.
41. Arumugam TV, Alli-Shaik A, Liehn EA. et al. Multiomics analyses reveal dynamic bioenergetic pathways and functional remodeling of the heart during intermittent fasting. *Elife* 2023;**12**:RP89214. <https://doi.org/10.7554/eLife.89214.2>.
42. Videira RF, Koop AMC, Ottaviani L. et al. The adult heart requires baseline expression of the transcription factor Hand2 to withstand right ventricular pressure overload. *Cardiovasc Res* 2022;**118**:2688–702. <https://doi.org/10.1093/cvr/cvab299>.
43. Tang Y, Aryal S, Geng X. et al. TBX20 improves contractility and mitochondrial function during direct human cardiac reprogramming. *Circulation* 2022;**146**:1518–36. <https://doi.org/10.1161/CIRCULATIONAHA.122.059713>.
44. Saba R, Kitajima K, Rainbow L. et al. Endocardium differentiation through Sox17 expression in endocardium precursor cells regulates heart development in mice. *Sci Rep* 2019;**9**:11953. <https://doi.org/10.1038/s41598-019-48321-y>.
45. Mukherjee S, Chaturvedi P, Rankin SA. et al. Sox17 and beta-catenin co-occupy Wnt-responsive enhancers to govern the endoderm gene regulatory network. *Elife* 2020;**9**:e58029. <https://doi.org/10.7554/eLife.58029>.
46. Zimmerli D, Borrelli C, Jauregi-Miguel A. et al. TBX3 acts as tissue-specific component of the Wnt/beta-catenin transcriptional complex. *Elife* 2020;**9**:e58123. <https://doi.org/10.7554/eLife.58123>.
47. Pahnke A, Conant G, Huyer LD. et al. The role of Wnt regulation in heart development, cardiac repair and disease: a tissue engineering perspective. *Biochem Biophys Res Commun* 2016;**473**:698–703. <https://doi.org/10.1016/j.bbrc.2015.11.060>.
48. Guo L, Glover J, Risner A. et al. Dynamic expression profiles of beta-catenin during murine cardiac valve development. *J Cardiovasc Dev Dis* 2020;**7**:31. <https://doi.org/10.3390/jcdd7030031>.
49. Paulis L, Fauconnier J, Cazorla O. et al. Activation of sonic hedgehog signaling in ventricular cardiomyocytes exerts cardioprotection against ischemia reperfusion injuries. *Sci Rep* 2015;**5**:7983. <https://doi.org/10.1038/srep07983>.
50. Sinniah E, Mizikovskiy D, Shim WJ. et al. Epigenetic constraint of cellular genomes evolutionarily links genetic variation to function. *bioRxiv* 2024.