Inferring cell diversity in single cell data using consortium-scale epigenetic data as a biological anchor for cell identity

Yuliangzi Sun[1,5], Woo Jun Shim[1,5], Sophie Shen[1], Enakshi Sinniah[1], Duy Pham[1], Zezhuo Su[2,3], Melanie D. White[1], Joshua W.K. Ho[2,3], Quan Nguyen[1], Mikael Bodén[4, 6], Nathan J. Palpant[1, 6, 7]

[1] Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia

[2] School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

[3] Laboratory of Data Discovery for Health Limited ($D^2$4H), Hong Kong Science Park, Hong Kong SAR, China

[4] School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia

[5] These authors contributed equally to this work

[6] Senior author

[7] Lead Contact

Corresponding Author

Nathan J. Palpant

Institute for Molecular Bioscience

The University of Queensland

Brisbane, Australia, 4067

Email: n.palpant@uq.edu.au

Phone: +61 0439241069

Abstract

Methods to identify cell types from single-cell RNA sequencing (scRNA-seq) often rely on mathematically defined analysis protocols. This study we describe a new method called TRIAGE-Cluster which uses genome-wide repressive epigenetic data from diverse bio-samples to identify genes demarcating cell diversity in any scRNA-seq data set. TRIAGE-Cluster integrates patterns of H3K27me3 domains deposited across hundreds of cell types with weighted density estimation to determine cell clusters using an scRNA-seq UMAP. We couple this with an unsupervised machine learning method, TRIAGE-ParseR, which parses any input rank gene list to define gene groups governing the identity and function of cell types. We demonstrate the power of these methods by analysing atlases of *in vivo* and *in vitro* cell diversification and organogenesis. We also provide a web accessible dashboard for analysis and download of data and software. Collectively, genome-wide epigenetic repression provides a powerful biological reference point for identifying and studying genetic regulation in scRNA-seq expression data.

Keywords

H3K27me3, epigenetic, scRNA-seq, clustering, cell identity, genetic regulation

Issue Section: Problem solving protocol

Introduction

Single-cell RNA sequencing technologies provide unprecedented opportunities to study the diversity of cell processes underpinning development and disease. Standard analysis workflows involve stepwise analytic tools that reduce high dimensional data to reveal the underlying complexity of and relationships between heterogeneous cell types. However, the computational challenges involved in single cell analysis are significant at least in part because of common user-defined parameter settings that make interpretation of cell diversity a subjective exercise. Downstream analyses are similarly heavily reliant on prior knowledge including the use of known marker genes, subjective data subtraction using differential expression, and the use of reference annotation databases. These methods limit the power of single cell data for discovering novel biology because they lack unsupervised biological reference points for defining cellular diversity and interpreting genetic regulation of cell states.

Cell type clustering methods enable users to identify cell diversity represented in the data. This step underpins most downstream analyses to guide hypotheses and establish conclusions about cellular processes. Current clustering algorithms group cells based on gene expression similarity and relied on mathematically defined constraints for data analysis [1]. Agglomerative clustering methods such as SINCERA [2], pcaReduce [3], and CIDR [4] use distance measurements such as Euclidean, Manhattan, Jaccard, Minkowski, and Canberra [5] to merge clusters iteratively based on similarities [6]. K-means clustering assigns data points based on Euclidean distance measures and iteratively optimises the centroid of each group to minimize within-cluster distance [7]. K-means requires multiple runs to achieve high robustness and accuracy since the number of clusters is predetermined and optimization involves randomization. Model-based approaches including Gaussian Mixture Models (GMM) such as scGMAI [8] assume data are composed of numerous Gaussian distributions and cell-cell similarity decreases as data points locate further from the centre of the distribution [7]. Density-based clustering such as DBSCAN [9] and GiniClust [10] are non-parametric approaches that do not require a pre-defined number of clusters and instead group cells based on data point density, where low-density areas are considered as outliers. Lastly, graph-based clustering is an extension of density-based clustering [11] where relationships among cells are represented by a similarity graph where nodes are cells and weighted edges reflect similarities among cells [12]. For instance, Louvain and Leiden are popular clustering approaches implemented in Seurat [13] and SCANPY [14].

Owing to the complex nature of cell clustering there is no gold-standard for parameter setting or method selection [1]. Indeed, for the above methods, the ongoing absence of unsupervised biological reference points for determining cell populations remains a key drawback of available clustering methods. Furthermore, current methods assume all cells in a data set are meaningful and are therefore interpreted equally while it is well appreciated that sequencing depth and technical variation in scRNA-seq data are major variables influencing data variability and accuracy [15-17]. However, establishing justifiable and generalisable criteria for data exclusion remains challenging.

Similarly, downstream gene expression analysis of cell populations often depends on arbitrary selection of computational parameters and reference points. For instance, differential expression gene analysis (DE) between cell clusters, while commonly practiced, yields highly variable output depending on cell-cell comparison choices and statistical models [18].

To address these limitations, newer methods utilise cell-level abundance features to identify distinct cell clusters [19], or modelling distributional differences in expression of individual genes to extract biological meaning of the regulatory network pertinent to cell clusters [20-24]. However, these methods lack more complex gene regulatory information encoded in the epigenome to interpret gene expression networks and are significantly influenced by the data structure regardless of their biological relevance [23, 25].

Our recent study [18] demonstrates that histone modification deposition patterns embedded in consortium-scale genomic data can reveal unsupervised biological reference points for interpreting complex genomic data. We found that genes having broad H3K27me3 domains with high frequency across diverse cell types significantly enrich for cell-type specific regulatory genes [18]. This analysis approach, which we call TRIAGE (Transcriptional Regulatory Inference Analysis of Gene Expression), is unsupervised and does not depend on external reference data, statistical cutoffs, or prior knowledge. We used consortium data from the Human Cell Atlas, FANTOM, and a draft map of the human proteome to show that TRIAGE provides an independent, orthogonal model to interpret genomic data from any cell or tissue type. TRIAGE devises a genome-wide quantitative feature called a repressive tendency score (RTS) which can be used as an unsupervised independent reference point to infer cell-type regulatory potential for each protein-coding gene.

In this study, we outline a novel pipeline for scRNA-seq data clustering and gene regulatory network analysis using a biologically defined strategy for data interpretation. We integrate epigenetic data with dimensionality reduced data by using a genes' repressive tendency as an independent biological reference point for cell type identification in scRNA-seq data. Furthermore, we demonstrate that hidden orthogonal patterns of H3K27me3 deposition observed in diverse

cell and tissue types provides a basis for grouping genes with shared biological processes. Overall, we present a two-step computational pipeline drawing on the fundamental principles of TRIAGE to enable analysis of cell clustering and interpretation of cell populations that scales to any cell type and data type, including applications across species to enable analysis of biological diversity in scRNA-seq data.

Method

Please refer to the supplemental information.

Results

Determining cellular mechanisms of differentiation requires robust methods to distinguish the identity and gene regulatory basis of cell types. **Figure 1a-d** provides a stepwise workflow for using TRIAGE defined RTS metric as an independent, unsupervised biological reference point to assist in distinguishing cell populations and their regulatory gene programs in any scRNA-seq data set. This provides an integrated view of how our previous work in development of TRIAGE [18] is implemented in the current study. First, for all protein coding genes the frequency of broad H3K27me3 domains at each locus across 834 EpiMap biosamples is calculated to derive the RTS for all protein coding genes. Second, any orthogonal input gene expression data set (or any data set mapped to protein coding genes such as ATAC-seq data and methylation data) is transformed by a simple product of the gene expression values with the RTS to derive a discordance score (DS), showing repression discordant with expression. For every individual gene expression data set (e.g., single-cell transcriptome), data transformation by TRIAGE enables genes with high regulatory potential to be prioritized (highly ranked by DS value). In the current study (Step 3), we next implement TRIAGE-Cluster to identify cell types in any single cell data set by using the RTS rank of genes as a biological anchor point to identify cell types in UMAP space. Fourth, to assist in parsing gene networks from a single rank gene list (e.g., DS), we use epigenetic data in a machine learning model (Gaussian Mixture Model) to parse pseudo-bulk DS on PCA space for any cell cluster to separate genes into biologically functional groups. A glossary of terms is provided in **Table 1**.

Using EpiMap data to calculate genome-wide repressive tendency scores

While our first study [18] developed RTS using 111 NIH Epigenome Roadmap H3K27me3 samples, this study expands the reference data for calculating RTS by measuring H3K27me3 deposition across 834 cell samples and 27 tissue groups in EpiMap [26] (**Table S1, Figure 2a**). The H3K27me3 broad domains for each gene were used to calculate the RTS which quantifies cell-type-specificity for each gene. We identify a priority set of 993 genes using values above the

inflection point of the interpolated RTS curve (RTS > 0.013) (**Figure 2b**). Highly variably expressed transcription factors (TFs) represent a positive gene set of cell type specific regulatory genes [18] and are significantly enriched among these RTS priority genes (**Figure 2c**). Analysis of genes ranked by RTS demonstrates that highly ranked genes tend to be more cell type specific and lowly expressed (**Figure 2d-f**), a profile typical of regulatory factors like transcription factors. Indeed, gene ontology analysis of the 993 RTS priority genes demonstrates significant enrichment in DNA-binding factors such as homeobox genes (FDR $6.2e^{-114}$) controlling fundamental processes in cell differentiation, development, and tissue morphogenesis (**Figure 2g)**.

We next identified scRNA-seq data with robust cellular diversity to develop and test clustering parameters and performance. We selected an atlas of mouse embryonic development which profiles 116,312 cells collected at nine sequential time-points ranging from e6.5 to e8.5 days post-fertilisation [27] (**Figure S1, S2**). These data comprise all primary germ layers, consisting of 37 classified cell types from ectoderm, mesoderm, and endoderm (**Figure S2**). We used this data to demonstrate two critical features of RTS. First, genes in higher RTS ranking have higher cell-type specificity **(Figure 2d)** which is supported by their cell-specific expression patterns across the UMAP space (**Figure 2h**). Second, TRIAGE transformation of gene expression data using RTS enables prioritisation of cell-type specific regulatory genes. We demonstrate this by analysing pseudo-bulk gene expression of two annotated cell types, cardiomyocytes and endothelial cells. As expected, original gene expression values show enrichment in housekeeping and functional genes, while TRIAGE transformation results in robust enrichment of regulatory genes underpinning the mechanistic basis of heart (*NKX2-5*, *GATA4*, *TBX5*) and endothelial (*SOX17*, *GATA2*, *TAL1*, *LMO2*) development and morphogenesis (**Figure 2i**). We build on these principles of TRIAGE to implement a new two-step computational pipeline for clustering and gene network analysis as a strategy for studying cell diversity in scRNA-seq data.

TRIAGE RTS demarcates cell specific clusters in scRNA-seq data

As shown in **Figure 3a** the RTS assigned to genes by TRIAGE provides an unsupervised, genome-wide method to quantitatively distinguish between non-specific vs cell type specific genes. For example, *EOMES* represents a broad mesendoderm cell type (RTS 0.19), while more specific markers of lateral plate mesoderm (*HAND1*, RTS: 0.28) and cardiomyocytes (*NKX2-5*, RTS: 0.35) demonstrate that RTS values increase according to cell specificity. Similar examples are shown for paraxial mesoderm and endoderm (**Figure 3a**).

We designed a proof-of-concept analysis comparing Seurat clustering vs using high-RTS genes as a decision-making feature influencing clustering tree analysis over 10 resolutions (**Figure 3b**). Standard Seurat clustering progressively divides clusters into small groups with increasing resolution (**Figure 3c**). For TRIAGE-Cluster, we assumed enrichment

of RTS priority genes in a cell population demarcates unique cell types and evaluated this at each stepwise resolution of clustering. We found that TRIAGE genes characterising new clusters quickly plateaus demonstrating that increased clustering resolutions do not necessarily imply biologically distinct cell types. Furthermore, RTS priority genes trimmed the number of clusters, while still being sensitive enough to identify new clusters even at very high resolutions, representing rare cell types that cannot be detected in broad categorizations at low resolutions (**Figure 3c**). We therefore aimed to develop a method to consistently identify distinct cell types among diverse populations of cell types in any scRNA-seq data set by reference to a fixed, unsupervised, and quantitative genomic reference point encoded by TRIAGE.

Weighted density analysis of RTS priority genes determines cell clustering

We designed an approach drawing on the concept of a topographic map view where contour lines demarcate peaks of cell specificity in the UMAP space (**Figure 3d**). We hypothesised that high-RTS genes expressed in cells can demarcate cell-specific identity and the RTS assigned to genes can define contour lines demarcating peaks (specific cell types) from valleys (transitional or non-specific cell types) (**Figure 3d**). Using a weighted density estimation approach, RTS was used as the weighting parameter to adjust cell density in a whole transcriptome derived UMAP space.

We applied this method to the mouse gastrulation atlas data [27] (**Figure 3e-h**). Bandwidth resolutions are shown for the contour plot (**Figure 3e**) and pairwise Jaccard similarity between peaks (**Figure 3f**). **Figure 3g** shows the cell clustering result at bandwidth 0.3 identifying 77 peaks. For selected populations we show how high-RTS genes are enriched in clusters, revealing the clustering basis derived from identity-defining genes across the data set. For example, cardiomyocytes are anchored primarily by the mes-endoderm-associated gene *GATA4* and cardiac regulatory factor *NKX2-5*. Parietal endoderm is also anchored by *GATA4* but also the endoderm-associated transcription factor *SOX17*.

Since the clustering output is governed by cell distributions defined by the UMAP, we evaluated cell-cell relationships derived by k-means clustering of the pairwise Jaccard similarity analysis (**Figure 3f**). The output provides an integrated view of cell relationships determined by cell identity defining genes in each cell type across the UMAP space linked with the time point when the cell type appears during development, the original annotation, and the top ranked TRIAGE anchor gene defining each cell type (**Figure 3h** and **Figure S3**).

TRIAGE-Cluster captures granular cell diversity in scRNA-seq data

We next analysed performance of the underlying data structure based on parameters that define the clustering. Since RTS assigned to priority genes (**Figure 2b**) are predicted to infer cell-specific genetic features, we tested whether removal of RTS genes influences the clustering output using an adjusted rand index (ARI) analysis. These results

demonstrate that removal of high-RTS genes governs TRIAGE-Cluster accuracy (**Figure 4a**). Using the ARI to determine the impact on clustering across diverse single cell data sets, we removed RTS genes from bottom to top across thirteen independent *in vivo* and *in vitro* scRNA-seq data sets, which shows that high-RTS genes are the primary drivers of clustering across all data sets (**Figure 4b**).

We next compared TRIAGE-Cluster against Seurat as a benchmark clustering method and tested gene expression similarity between all pairwise peaks or clusters at different clustering resolutions using Spearman rank correlation. We assume that more distinct cell types should have gene expression differences reflected by lower cluster-cluster pairwise correlation. Surprisingly, Seurat clustering shows highly stable similarity between cell types at all clustering resolutions regardless of input data type (original expression or TRIAGE transformed DS matrix). Furthermore, Seurat shows stable correlation between clusters using either the whole transcriptome or highly variable genes for gene expression similarity comparison. This suggests that separation of distinct cell types over different cluster resolutions in Seurat is not driven by significant global gene expression differences (**Figure 4c**).

In contrast, TRIAGE-Cluster peaks consistently result in lower gene expression correlation compared to Seurat and this correlation increases at higher clustering resolutions as differences between cell types become less distinct but have more demarcated cell subtypes (**Figure 4c**). Remarkably, peaks from TRIAGE-Cluster using TRIAGE transformed matrix showing the lowest correlation with highly variable genes. This suggests that a combination of DS score with TRIAGE-Clustering efficiently prioritizes cell identity genes to demarcate distinct cell types in the data set. We next compared peak or cluster correlation of TRIAGE-Cluster and Seurat using the same cluster numbers for each method and compared these correlations to those measured using original annotations in the single cell atlas of mouse gastrulation [27]. Again, these data show that peaks defined by TRIAGE-Cluster are significantly less correlated than equivalent resolutions defined by Seurat or original annotations (**Figure 4d-e)**. To further validate that TRIAGE peaks capture functionally relevant gene sets in the data, we performed Gene Ontology analysis using the top 100 discordance score genes for each peak or cluster. These data show that TRIAGE peaks have higher enrichment of biological process diversity than Seurat clusters (**Figure 4f, Table S2-5**).

In our companion study, we demonstrate the utility of TRIAGE-Cluster by applying it on a multiplexed single cell atlas of iPSC differentiation data that integrates temporal data across eight time points and signalling data with nine developmental signalling perturbations (Shen *et al*., 2022 (preprint)). 48 peaks were identified from the *in vitro* differentiation atlas data, and clustering performance was evaluated using a trajectory approach VIA which evaluates cell states relationship [28]. Comparing analysis using all cells vs TRIAGE peaks, TRIAGE peaks give better peak-to-node assignment in the trajectory and significant correlation between pseudotime and actual time. Taken together, these

data demonstrate the value of TRIAGE-Cluster in defining the biological diversity in a data set including improved clarity in evaluating cell-cell trajectory analysis.

TRIAGE-ParseR enhances identification of cell gene regulatory networks

While clustering quantifies cell diversity of a data set, down-stream analysis of gene programs is critical for understanding the biological relevance of gene expression programs in individual cell types. Gene expression outputs underpin fundamental knowledge into mechanisms of cell diversity and differentiation. While TRIAGE and other gene expression analyses provide a single rank gene list based on relative abundance, deriving meaningful, biologically relevant information from these gene lists remains an ongoing challenge. To complement TRIAGE-Cluster, we developed a method to study gene regulatory networks of individual peaks in an unsupervised manner using a machine learning method that groups genes based on their shared epigenetic co-regulatory pattern. The primary goal of this approach was to enable separation of biologically meaningful gene groups within an individual gene expression data set without reliance on subjective external reference points (such as differential expression) or prior knowledge (such as GO, STRING, marker gene selection, or gene class biases).

We developed the method based on the rationale that expression of key developmental genes is modulated by both gain or loss of H3K27me3 during cell differentiation (Boyer, 2006; Mikkelsen, 2007). Across any cell state, gain or loss of repressive chromatin may provide a surrogate indication of functionally similar gene programs. For example, lineage specific transcription factors show high repression tendency whereas housekeeping genes rarely share repressive chromatin. We hypothesized that by evaluating patterns of repressive chromatin across diverse cell states and evaluating the most relevant eigenvectors of data influencing the epigenetic relationship of genes, we could separate functionally relevant gene-gene relationships in an unsupervised manner.

We first performed principal component analysis (PCA) on the breadth of H3K27me3 deposition at each gene observed in diverse human tissue and cell types. We then evaluated the input genes list for their shared H3K27me3 deposition patterns modelled by a Gaussian mixture model (GMM). To generate the input gene list, the original expression matrix of a cell type is converted to discordance score using TRIAGE (**Figure 1a-b**). To limit the gene list to those most likely to govern cell identity and function, we selected the top 100 genes ranked by the average discordance score for a given input cell type. This method, called TRIAGE-ParseR, is described in **Figure 5a**.

Extracted H3K27me3 patterns by PCA showed that (i) most of the data variance can be explained by a much smaller number of principal components than the original data (**Figure S4a**), and (ii) distinct tissue types can be characterised by these H3K27me3 patterns (**Figure S4c**). Consistent with these observations, genes enriched with a specific

H3K27me3 pattern demonstrate tissue-specific biological processes (**Figure S4b**). Furthermore, these data show that the first PC (first H3K27me3 pattern) identifies genes with strong tendency to deposit broad H3K27me3 across many tissue types (i.e., high-RTS genes) while subsequent PCs capture more tissue-type specific patterns (**Figure S4d**) and therefore justifies the importance of incorporating diverse PCs to reveal biologically informative signals that distinguish gene programs underpinning different cell types. We next used TRIAGE-ParseR to evaluate gene expression from heart tissue and blood (**Figure S4e-f**). In addition to the first PC which shows high concordance with the RTS score, analysis of diverse PCs reveals unique gene-gene TRIAGE-ParseR signatures that can be used to inform their biological relationships.

We next applied TRIAGE-ParseR on TRIAGE cell clustering data derived from the mouse gastrulation atlas data and focussed the initial demonstration of the method on cardiomyocytes (**Figure 5b**). We provide a data workflow pipeline (**Figure 5c**) whereby pseudo-bulk gene expression (enriched in housekeeping and structural genes) is transformed with TRIAGE to derive the discordance score (enriched in cardiac regulatory genes). The top 100 discordance score genes are then used as input for analysis resulting in gene groups (**Figure 5c**). Three gene clusters are highlighted because of significant enrichment by STRING analysis (protein-protein interaction FDR<0.001). **Figure 5d** shows the raw H3K27me3 deposition patterns and clearly reveal unique epigenetic patterns of each gene group.

We compared Gene Ontology (GO) enrichment at each step of our data analysis workflow outlined in **Figure 5c.** These results show that using the top 100 genes based on original gene expression lack any enrichment in GO terms except for structural genes like sarcomere genes. In contrast, TRIAGE discordance score provides robust enrichment in diverse pathways associated with cardiomyocyte morphogenesis and differentiation, TRIAGE-ParseR separates these genes into groups to suggest how genes are functionally related. For example, genes in cluster 3 (the most significantly repressed genes across diverse cell types, **Figure 5d**) are enriched with heart development terms including diverse transcription factors known to be critical for heart development (**Figure 5e**). Clusters 4 genes show intermediate repression (**Figure 5d**) and similarly have transcription factors that are more generic genes governing embryo development (**Figure 5e**). Lastly, gene cluster 1 has minimal repression across diverse cell types (**Figure 5d**) and is enriched in genes involved in cardiac muscle function (**Figure 5e**). Indeed, STRING analysis revealed that genes in these clusters formed significant protein-protein interactions and therefore revealed both known and unknown gene-gene relationships (PPIs, enrichment $p<0.01$) (**Figure 5f**). Taken together, these data demonstrate that H3K27me3 deposition patterns reveal shared gene regulatory networks relevant to interpret nuance of biological function between genes. We show that TRIAGE-ParseR is capable of meaningfully segregating top TRIAGE ranked genes to facilitate cell type identification from any input gene expression data set.

Integrating TRIAGE-Cluster and TRIAGE-ParseR to study single cell data

We integrated TRIAGE-Cluster and TRIAGE-ParseR to evaluate cell types in the mouse gastrulation atlas data. **Figure 6a** shows a holistic view of all TRIAGE peaks, peak relationship at bandwidth 0.3, as well as corresponding original annotation and timepoint for each TRIAGE peak. While this mouse gastrulation data set provides a powerful resource for studying the earliest stages of mammalian organogenesis, most cell classifications capture redundant or broad cell groups that lack the nuance of cell diversification during early development. We implement TRIAGE-Cluster as a complementary strategy for identifying more distinct cell types based on its ability to determine local maxima of RTS-ranked genes in the data set as a surrogate indicator of biologically distinct cell populations.

We selected cells annotated as forebrain/midbrain/hindbrain in the original data. TRIAGE analysis identifies five peaks (23, 48, 63, 67, and 76) mapped to this annotation **(Figure 6b)**. We provide a multi-scale data output to evaluate identity-defining features of each cell type including their relationships defined by TRIAGE-derived dendrogram (**Figure 6c**), discordance score genes enriched for each cell type by reference to all other cell types in the data **(Figure 6d)**, and TRIAGE-ParseR analysis of gene-gene groupings analysed by Gene Ontology analysis **(Figure 6e)**. These data reveal a separation of distinct cell relationships within this broad cell grouping that suggest more nuanced cell subtypes governing ectoderm development. Clusters 48 and 67 share high degrees of similarity based on their gene ontology enrichment for neural patterning (Neural tube, Dorsal/Ventral and Anterior/posterior pattern specification, **Figure 6e)** and DS with genes including the shared TRIAGE anchor gene *PAX6*, an organizer of hindbrain patterning. Detailed gene expression data suggest cluster 67 is boundary-defining cells of presumptive rhombomeres R1-R6 (*CRABP1* and DS for *FGF3*) while cluster 48 is more specifically associated with presumptive R2-R4 (*FST* and *HOXA2*, **Figure 6f)**. Cluster 76, anchored by *PAX2*, shares a broad range of identity defining genes related to establishment of the Isthmic organizer at the midbrain/hindbrain boundary (*HES3*, *FGF8*, *WNT1*, *EN1*). In contrast, clusters 23 and 63 share similarity based on the TRIAGE-dendrogram (**Figure 6c**). Cluster 23 shows signatures of Forebrain/midbrain marked by *RAX1*, *SIX3*, *LHX2*, *LHX5*, and *PCSK1N* **(Figure 6f)** and gene ontology enrichment for Forebrain development **(Figure 6e)**. Cluster 63 is associated with the dorsal midline of the midbrain marked by *PCSK1N*, *HES3*, *FGF8* and *DLX6*. We provide a second analysis of paraxial mesoderm lineage cell types based on detailed dissection using TRIAGE-associated outputs (**Figure S5**). Furthermore, the data interpretation pipeline outlined here is implemented at scale for analysis of *in vitro* differentiation from pluripotency outlined in the companion paper (Shen et al., (preprint)). The collective results of both studies including new data and data analysis dashboard can be accessed online (http://cellfateexplorer.d24h.hk/).

Discussion

Understanding biological relationships between cell types and the gene networks that govern these differences remains central to maximising the value of technologies capturing genomic data at single cell resolution. This study demonstrates that patterns of H3K27me3 deposition provide a biological anchor for defining cell clustering and interpretating the genetic basis of cell populations. To identify cell types without prior knowledge or computational parameter settings, the RTS acts as an independent biological reference point implemented downstream of dimensionality reduction algorithms to cluster cell types. We demonstrate that genes sharing patterns of H3K27me3 deposition across diverse cell types can provide an unsupervised approach for studying the gene expression basis of biological processes underpinning cell differentiation and function.

Regulation of the genome is in part mediated through histone modifications that control genome accessibility, acting as an on/off switch to regulate access to regions of the genome [29]. We reasoned that regions with the broadest histone modification domains (representing extreme levels of on/off control) could be used to quantify the probability that a locus has a role controlling cell decisions and functions. Based on this rationale, we developed TRIAGE [18] which uses patterns of histone methylation data (H3K27me3) to identify regions of the genome that are enriched in cell-specific identity defining genes. Genes with a high RTS have consistent repressive chromatin in most cell types, indicating that in most cell types they are inaccessible. In the rare instances when these genes get turned on, they likely have key roles controlling cell decisions or functions. Conversely, regions with a low RTS infrequently have repressive chromatin and perform housekeeping functions.

The collective suite of methods built around TRIAGE provide several major advantages: 1) TRIAGE is a customisable metric with a single value assigned to genes which makes it simple to implement in any genomic data set, 2) TRIAGE is versatile, providing a quantitative genomic model to weight gene prioritisation in existing or emerging genetic analysis tools, 3) the simplicity of TRIAGE enables broad implementation in data analysis pipelines for any data sample from any cell, tissue, disease, or individual thereby providing a method to enrich for factors most likely to cause phenotypic changes in cell identity, and therefore 4) helps prioritise genetic targets for drug development and identifying the primary determinants of disease and development.

TRIAGE-ParseR provides several critical advances in studying gene-gene relationships from input expression data. Namely, 1) it provides an unsupervised approach to study the gene regulatory basis of cells without drawing on external reference points (like DE) or prior knowledge, 2) it parses genes into groups to reveal biological mechanisms

underpinning cell states without imposing arbitrary categories, and 3) the use of an independent, consortium-scale reference data set for evaluating epigenetic patters provides a mechanism for applying this method to any input gene list.

This study provides a unique strategy for cell type identification using epigenetic information as biological reference point for cell clustering and deconstructing gene modules in single cell expression data. We provide examples of the workflow by analysing *in vivo* organogenesis and *in vitro* differentiation atlases. Collectively, TRIAGE highly ranks regions of the genome enriched in genetic features with profound effects on cell identity because these regions encode potent determinants of cell developmental processes. These ranking features and genome-wide predictions make TRIAGE unique among genomic analysis methods and position it to address areas of need in cell biology and genetics. In addition to our prior studies [18], TRIAGE-Cluster and TRIAGE-ParseR provide a growing portfolio of methods, outperforming benchmark methods like Seurat, to facilitate analysis of large-scale data sets to parse cell-cell and gene-gene relationships underpinning mechanisms of cell identity.

In future studies we aim to address three limitations in the current workflow. First, we demonstrate that the current clustering method provides a mechanism for restricting cell selection to capture clusters with the most distinct biological identity and difference. While this is a major advantage, it comes at the cost of trimming the data, focusing on a minority of cells in the data set. Currently, TRIAGE-Cluster provides a complementary approach to identify cell diversity in single cell data but should be weighed relative to the loss of statistical power provided by the scale of cell sampling. Future studies will aim to address this limitation by, for example, using TRIAGE-Cluster identified cell types as a reference point in combination with data imputation or trajectory inference methods to interpret cell-cell relationships using all cells in the data set. Second, the current method is reliant on assumptions built into dimensionality reduction representing cell relationships in UMAP space. While useful, major limitations exist for simplifying the data into these graphical presentations. To address this, we propose additional strategies such as comparison of cell types using TRIAGE-dendrograms (**Figure 3h**) which provide independent approaches to evaluate cell-cell relationships. We aim to develop approaches that provide the analysis value of TRIAGE-Cluster but avoid the limitations associated with topographical map representations. Third, TRIAGE-ParseR provides a unique approach to decipher gene-gene relationships but draws on only a single epigenetic reference point. Future developments of this method need to parameter test additional histone modifications representing active promoters (H3K4me3) or enhancers (H3K27ac and H3K4me1) as complementary to repressive chromatin marked by H3K27me3 as a basis for more nuanced analysis of gene-gene relationships using PCA-GMM.

Key Points

- TRIAGE-Cluster utilizes gene repressive tendency, RTS, as an unsupervised biological anchor point to facilitate cell type identification.
- TRIAGE-ParseR uses epigenetic regulatory pattern to study gene regulatory networks by separating genes into biologically functional groups without prior knowledge.
- TRIAGE-Cluster and TRIAGE-ParseR provide methods to integrate genome-wide epigenetic data with scRNA-seq data to parse cell-cell and gene-gene relationship governing mechanisms of cell identity.

Data Availability

The mouse gastrulation atlas data is available in *MouseGastrulationData* [30] at https://github.com/MarioniLab/MouseGastrulationData . Code for TRIAGE-Cluster and TRIAGE-ParseR are available online at https://github.com/palpant-comp/TRIAGE-Cluster and https://github.com/palpant-comp/TRIAGE-ParseR. The combined analysis of the mouse gastrulation atlas data and *in vitro* iPSC differential data can be explored at http://cellfateexplorer.d24h.hk/.

Author Contributions

Y.S. further developed, implemented, and benchmarked TRIAGE-Cluster, developed topographical visualisation approach, and wrote the manuscript. W.J.S. conceived and developed TRIAGE-ParseR and wrote the manuscript. S.S. helped develop TRIAGE-Cluster method. E.S. generated the overall pipeline Figure 1. D.P. and Z.S. constructed the website. M.D.W identified marker genes for each peak in forebrain midbrain hindbrain. Q.N. and J.W.K.H. supervised website development. M.B supervised TRIAGE-ParseR development. N.J.P conceptualised the study, supervised the project, raised funding, and wrote the manuscript.

Author Biographies

Yuliangzi Sun is a postgraduate student at the University of Queensland's Institute for Molecular Bioscience. Her current research focus on single cell analytics, particularly clustering and data integration.

Woo Jun Shim is a postdoctoral researcher at the University of Queensland's Institute for Molecular Bioscience. His research interests include bioinformatics, machine learning and personalised medicine.

Sophie Shen is a postgraduate student at the University of Queensland's Institute for Molecular Bioscience. Her research interests include understanding developmental biology using sing-cell genomics.

Enakshi Sinniah is a postdoctoral researcher at the University of Queensland's Institute for Molecular Bioscience.

Duy Pham is a postgraduate student in Genomics and Machine Learning Lab at the University of Queensland's Institute for Molecular Bioscience. His research interests include web development and trajectory inference using spatial technology.

Zezhuo Su carried out his doctoral work in cancer single cell biology and is a postdoctoral researcher in the School of Biomedical Sciences at the University of Hong Kong.

Melanie D. White heads the Dynamics of Morphogenesis Lab at the University of Queensland's Institute for Molecular Bioscience. Her research aims to understand how cellular properties and transcriptional regulators interact with mechanical forces in real time to direct vertebrate neural tube formation and neural cell fate specification.

Joshua W.K. Ho is an Associate Professor in the School of Biomedical Sciences at the University of Hong Kong (HKU). He is the Academic Lead of Bioinformatics Core of the Centre for PanorOmic Sciences (CPOS) at HKU. His research interests include bioinformatics, single cell analytics, gut microbiome metagenomics, artificial intelligence and IoT in medicine, and digital health.

Mikael Bodén is an Associate Professor at the University of Queensland's School of Chemistry and Molecular Biosciences. He is the director of UQ's postgraduate program in bioinformatics. His research interests focus on analysis, modelling and integration of biological data.

Quan Nguyen is the head of Genomics and Machine Learning Lab at the University of Queensland's Institute for Molecular Bioscience. With spatial technology approaches, he investigates cell types, their spatial organisation and cell-cell interactions underlying differential responses to treatment and risks of cancer metastasis, revealing potentially new gene targets to modulate cancer-immune cell interactions.

Nathan J. Palpant is an Associate Professor and a group leader at the University of Queensland's Institute for Molecular Bioscience. He is on the leadership of regional and national research programs including as co-chair of the Queensland Cardiovascular Research Network. His laboratory uses multi-scale experimental approaches including genomics, stem cell biology, and animal physiology to study mechanisms of cardiovascular development and disease.

Declarations of Interests

The authors declare no competing interests.

Reference

1. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data, Nat Rev Genet 2019;20:273-282.
2. Guo M, Wang H, Potter SS et al. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis, PLoS Comput Biol 2015;11:e1004575-e1004575.

3. žurauskiene J, Yau C. pcaReduce: Hierarchical clustering of single cell transcriptional profiles, BMC bioinformatics 2016;17:140-140.
4. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data, Genome Biology 2017;18:59-59.
5. Parasa NA, Namgiri JV, Mohanty SN et al. Introduction to Unsupervised Learning in Bioinformatics. Data Analytics in Bioinformatics. 2021, 35-49.
6. Costa IG, de Carvalho FdAT, de Souto MCP. Comparative analysis of clustering methods for gene expression time course data, Genet. Mol. Biol 2004;27:623-631.
7. Jelili O, Itunuoluwa I, Funke O et al. Clustering Algorithms: Their Application to Gene Expression Data, Bioinform Biol Insights 2016;2016:237-253.
8. Yu B, Chen C, Qi R et al. scGMAI: a Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder, Briefings in bioinformatics 2020.
9. Ester M, Kriegel HP, Sander J et al. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, Menlo Park, CA (United States), 1996.
10. Jiang L, Chen H, Pinello L et al. GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index, Genome Biol 2016;17:144-144.
11. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq, Mol Aspects Med 2018;59:114-122.
12. Wu H, Mao D, Zhang Y et al. A new graph-based clustering method with application to single-cell RNA-seq data from human pancreatic islets, NAR Genom Bioinform 2021;3:lqaa087-lqaa087.
13. Stuart T, Butler A, Hoffman P et al. Comprehensive Integration of Single-Cell Data, Cell 2019;177:1888-1902.e1821.
14. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis, Genome Biology 2018;19:15.
15. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis, Front Genet 2019;10:317-317.
16. Haque A, Engel J, Teichmann SA et al. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications, Genome Med 2017;9:75-75.
17. Ziegenhain C, Vieth B, Parekh S et al. Comparative Analysis of Single-Cell RNA Sequencing Methods, Mol Cell 2017;65:631-643.e634.
18. Shim WJ, Sinniah E, Xu J et al. Conserved Epigenetic Regulatory Logic Infers Genes Governing Cell Identity, Cell Systems 2020;11:625-639.e613.
19. Zhao J, Jaffe A, Li H et al. Detection of differentially abundant cell subpopulations in scrna-seq data, Proc Natl Acad Sci U S A 2021;118:1.
20. Vandenbon A, Diez D. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data, Nature communications 2020;11:4318-4318.
21. Thalia EC, Michael S, Ann CB. Gene regulatory network inference from single-cell data using multivariate information measures, bioRxiv 2017.
22. Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity, Genome Biology 2019;20:110-110.
23. Jin S, Zhang L, Nie Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles, Genome Biol 2020;21:25-25.
24. Zhang M, Liu S, Miao Z et al. IDEAS: individual level differential expression analysis for single-cell RNA-seq data, Genome Biol 2022;23:33-33.
25. Fatemeh Behjati A, Kathrin K, Karl N et al. Integrative analysis of single cell expression data reveals distinct regulatory states in bidirectional promoters, bioRxiv 2018.
26. Boix CA, James BT, Park YP et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics, Nature 2021;590:300-307.
27. Pijuan-Sala B, Griffiths JA, Guibentif C et al. A single-cell molecular map of mouse gastrulation and early organogenesis, Nature 2019;566:490-495.
28. Stassen SV, Yip GGK, Wong KKY et al. Generalized and scalable trajectory inference in single-cell omics data with VIA, Nature communications 2021;12:5528.
29. Kundaje A, Meuleman W, Ernst J et al. Integrative analysis of 111 reference human epigenomes, Nature 2015;518:317-330.
30. Griffiths J LA. MouseGastrulationData: Single-Cell -omics Data across Mouse Gastrulation and Early Organogenesis. R package version 1.10.0. 2022.

Tables

Table 1. Glossary of terms

RTS	Repressive tendency score. Calculated from broad H3K27me3 domains across 834 cell types. High RTS genes are lowly expressed and have higher cell-type-specific regulatory potential.
Discordance score	Discordance score is calculated by multiplying log-transformed normalised gene expression values of any input data set with TRIAGE-derived RTS values for each gene.
TRIAGE Peak	Cell populations identified by high RTS anchor genes that are well-defined cell types, representing regions with local RTS maxima.
Anchor genes	Genes with RTS above the inflection point of the interpolated RTS curve are used to identify peak regions in the TRIAGE clustering method.

Figures

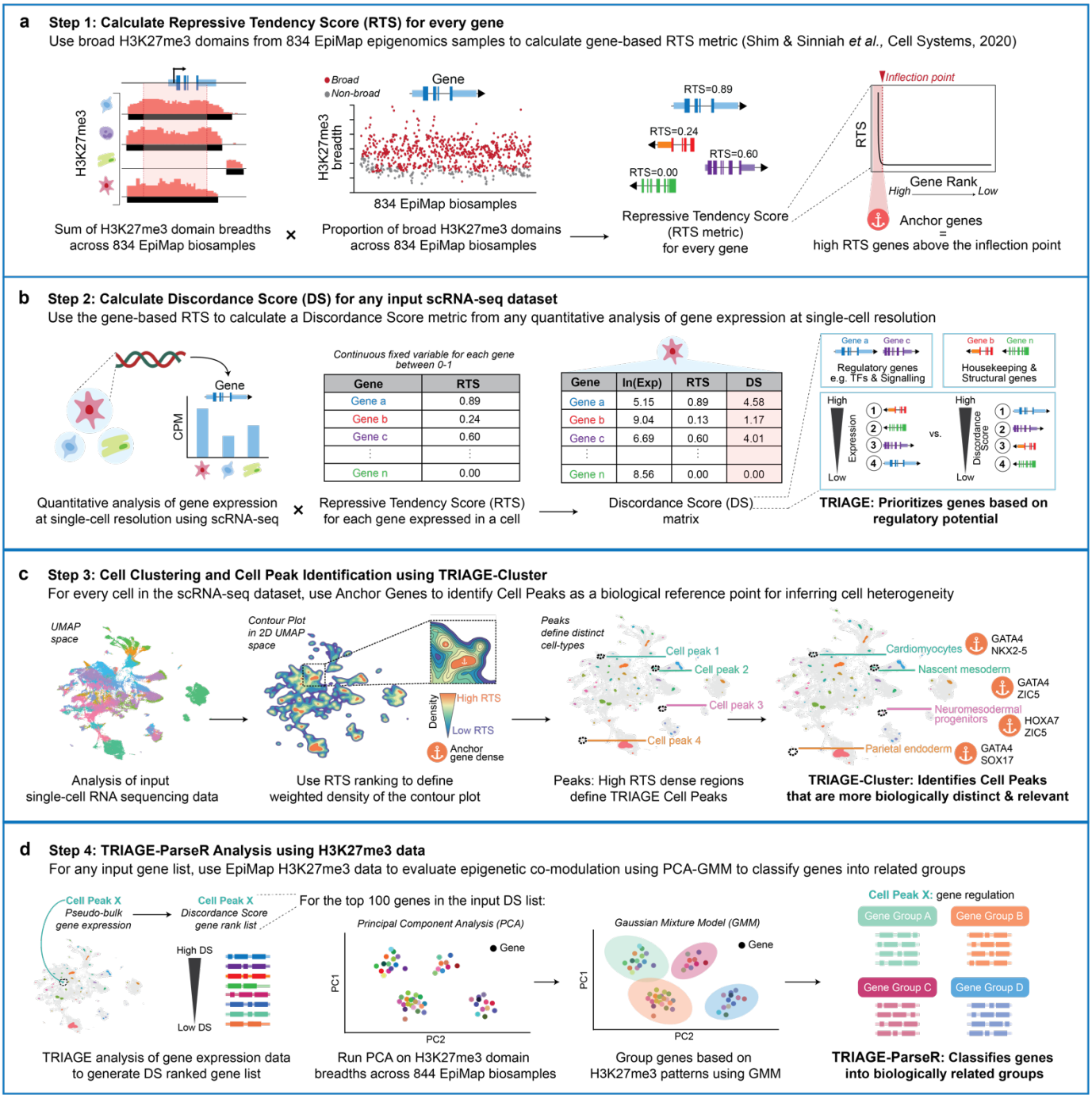


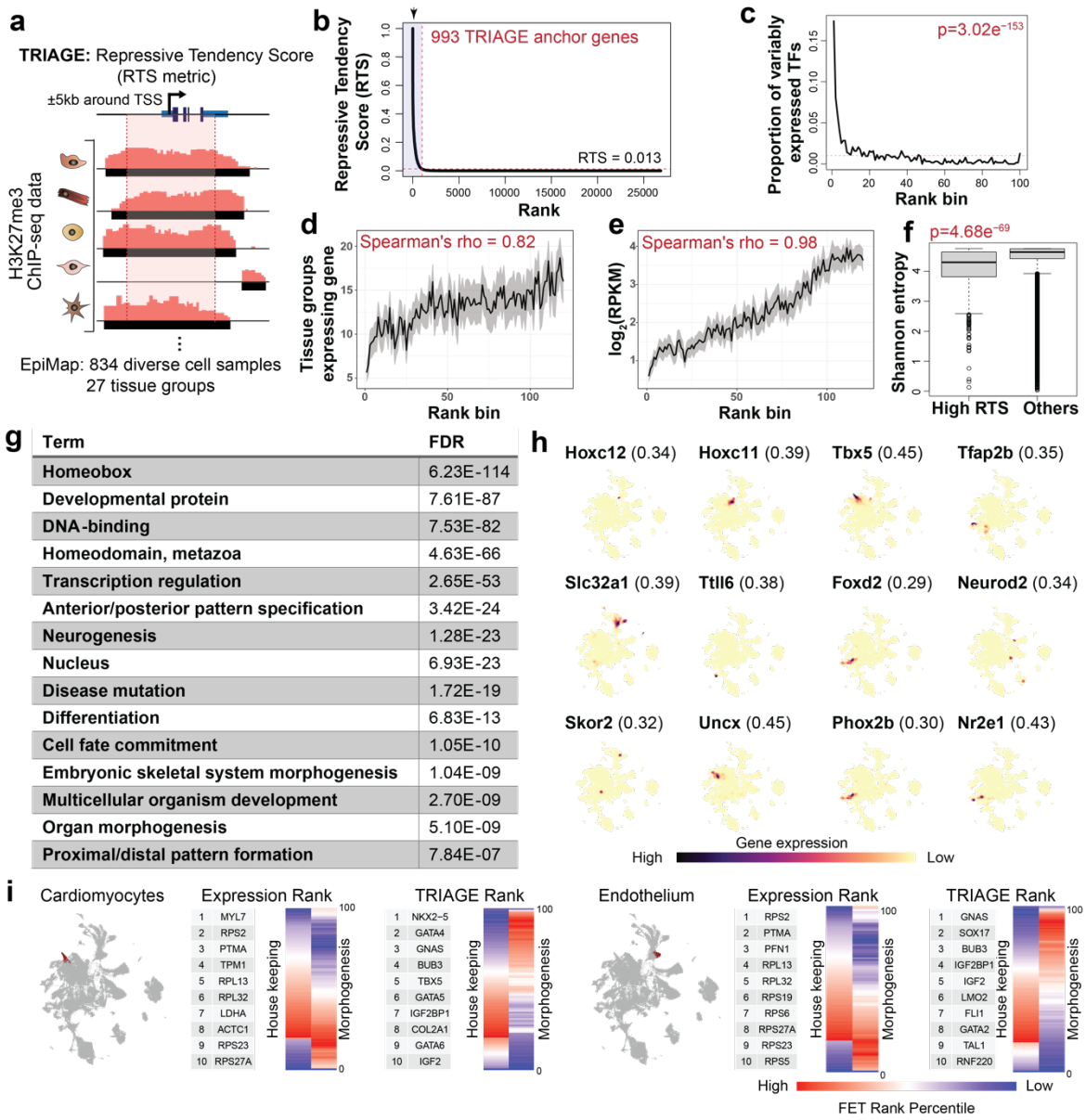
Figure 1. Overview of unsupervised pipeline for analysis single cell data to identify cell type.

a) TRIAGE calculates a repressive tendency score (RTS) providing a single quantitative value for every gene based on its association with broad H3K27me3 domains across 834 EpiMap biosamples [1] and generates unsupervised reference point called anchor genes which is the priority gene set based on RTS ranking.

b) Input single cell expression matrix is transformed to discordance (DS) matrix where expression value is converted to discordance score by RTS metric then high cell-type regulatory potential genes can be prioritized.

c) Anchor genes corresponding RTS are implemented in density-based clustering to position cells at UMAP space and cell populations locate in high RTS dense area are defined as peaks which represent transcriptionally distinct cell groups.

d) For each peak, genes are ranked based on pseudo-bulk discordance score are called discordance (DS) genes, then unsupervised TRIAGE-ParseR analyses of the top 100 DS genes with PCA and Gaussian mixture model (GMM) group genes according to H3K27me3 patterns, thus genes are parsed into functional groups to help cell type identification.



Term	FDR
Homeobox	6.23E-114
Developmental protein	7.61E-87
DNA-binding	7.53E-82
Homeodomain, metazoa	4.63E-66
Transcription regulation	2.65E-53
Anterior/posterior pattern specification	3.42E-24
Neurogenesis	1.28E-23
Nucleus	6.93E-23
Disease mutation	1.72E-19
Differentiation	6.83E-13
Cell fate commitment	1.05E-10
Embryonic skeletal system morphogenesis	1.04E-09
Multicellular organism development	2.70E-09
Organ morphogenesis	5.10E-09
Proximal/distal pattern formation	7.84E-07

Figure 2. TRIAGE is a gene quantification and selection strategy providing a biological reference point for cell-specific genetic features.

a) TRIAGE calculates a repressive tendency score (RTS), providing a single quantitative value for every gene based on its association with broad H3K27me3 domains across hundreds of EpiMap biosamples [1].

b) 993 genes have RTS values above the inflection point of the interpolated curve (red dashed line is RTS = 0.013).

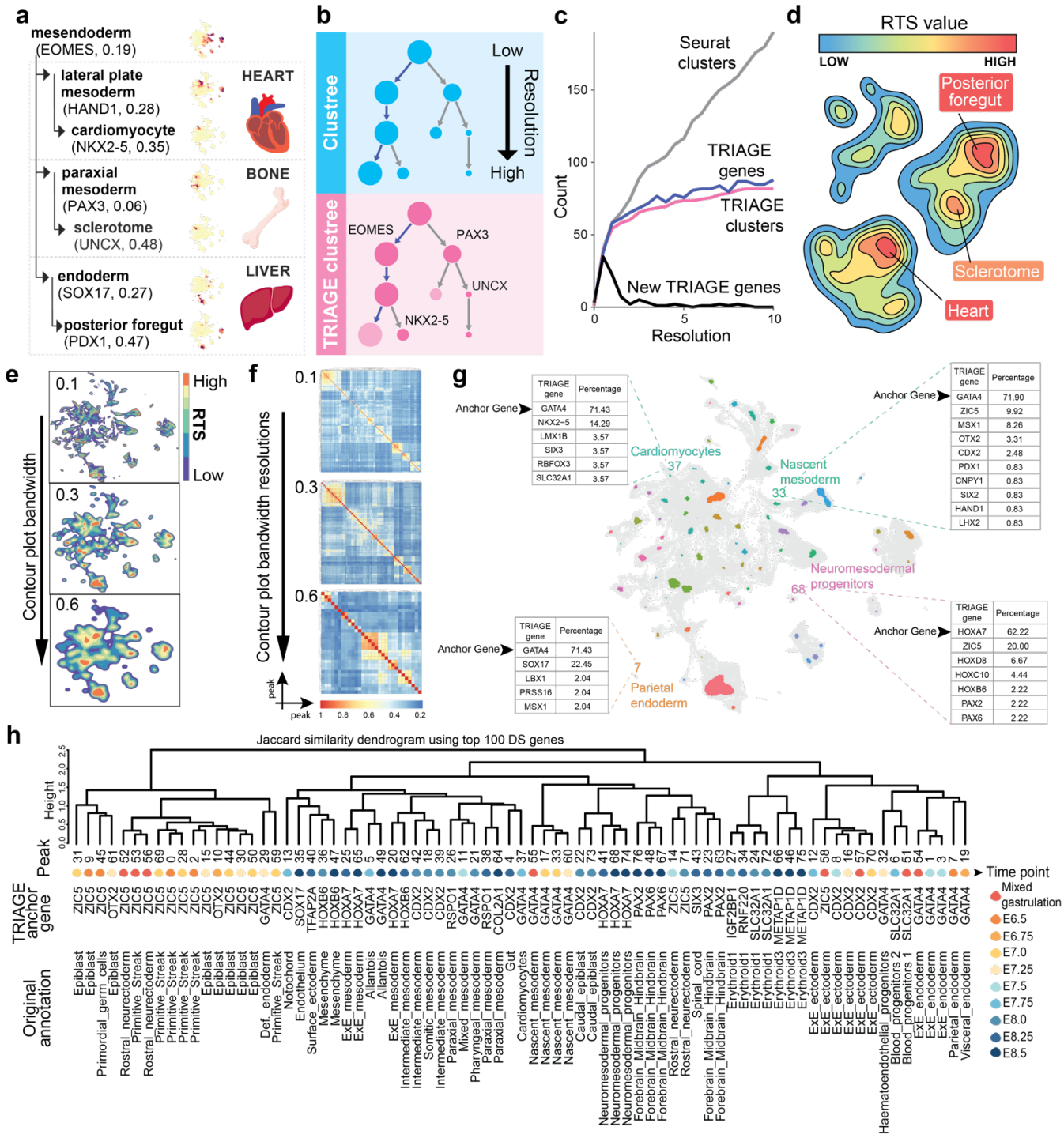
c) High RTS genes are significantly associated with variably expressed transcription factors (VETFs). One-tailed Fisher's exact test was performed and showing top 5% which represents 1,340 RTS genes were enriched with VETFs with p = 3.023e-153. Red dashed line represents the uniform distribution (proportion = 0.01). VETFs are classified from TRIAGE [2] based on previous study shows that developmentally regulated gene expression is highly variable temporally and spatially [3], in total 634 transcription factors (TFs) show substantial variable expression across 46 Roadmap cell types.

d-f) Genes with high RTS values are associated with cell/tissue specificity **(d),** are lowly expressed **(e)** and have higher entropy showing high-RTS genes are significantly cell-type-specific **(f)**.

g) Gene Ontology terms enriched in the top 993 RTS priority genes (see panel a) are significantly enriched in gene programs controlling cell identity and development.

h) UMAP plots of single cell atlas of mouse gastrulation (Pijuan-Sala *et al.,* 2019) showing selected RTS priority genes show sequestered expression in the UMAP space, suggesting cell-specific expression. RTS priority genes and their respective RTS value are shown in parentheses.

i) TRIAGE transformation of gene expression re-ranks genes on the basis of their regulatory potential. This is shown for pseudo-bulk gene expression analysis of annotated cardiomyocytes (left) and endothelial cells (right) from [4]. Data show UMAP of annotated cell types (left) followed by comparison of original expression vs TRIAGE gene ranking using top 10 ranked genes. The heatmaps showing distribution of housekeeping and morphogenesis genes ranked by original expression and TRIAGE. Enrichment is calculated for each rank bin relative to all genes (one-tailed Fisher's exact test).



TRIAGE gene	Percentage
GATA4	71.43
NKX2−5	14.29
LMX1B	3.57
SIX3	3.57
RBFOX3	3.57
SLC32A1	3.57

TRIAGE gene	Percentage
GATA4	71.90
ZIC5	9.92
MSX1	8.26
OTX2	3.31
CDX2	2.48
PDX1	0.83
CNPY1	0.83
SIX2	0.83
HAND1	0.83
LHX2	0.83

TRIAGE gene	Percentage
GATA4	71.43
SOX17	22.45
LBX1	2.04
PRSS16	2.04
MSX1	2.04

TRIAGE gene	Percentage
HOXA7	62.22
ZIC5	20.00
HOXD8	6.67
HOXC10	4.44
HOXB6	2.22
PAX2	2.22
PAX6	2.22

Figure 3. TRIAGE-Cluster is an unsupervised approach that uses cell-type-specific regulatory factors as biological anchor points to enable cell type identification.

a-b) Concept examples **(a)** and schematic **(b)** for using RTS as an orthogonal reference point for cell-type identification. The quantitative value of RTS priority genes enables identification of specific cell types where increasing RTS values corelate with increasing cell type specificity **(a)**. Strategy for implementing decision making criteria when splitting clusters from low to high resolution in normal clustering and TRIAGE-Cluster **(b)**.

c) Evaluation of TRIAGE decision-making criteria using Seurat clustering resolutions. Seurat cluster: number of clusters at each Seurat resolution. RTS priority gene: total number of TRIAGE priority genes expressed in any cluster. TRIAGE cluster: number of clusters with at least one RTS priority genes expressed at each Seurat resolution. New TRIAGE gene: number of new RTS priority genes at each resolution compared to all previous resolutions.

d) Schematic concept for using RTS genes as biological anchoring approach to analyse single cell data sets. We use a weighted density analysis of TRIAGE-determined cell identity genes to derive a topographic landscape view of the data with defined cell types identified as peaks in the contour plot.

e) Contour plot analysis applied to single cell atlas of mouse gastrulation [4] with granularity from high to low in visualisation using parameter bandwidth setting.

f) Heatmaps with same bandwidth resolutions (**panel e)** showing pairwise Jaccard similarity score between peaks calculated from top 100 genes ranked by TRIAGE discordance score.

g) UMAP of single cell atlas of mouse gastrulation using TRIAGE-Cluster with bandwidth setting of 0.3. There are four examples showing top TRIAGE genes expressed in the majority cells within the peak, where the gene that expressed in most cells within a peak is identified as anchor gene, and peaks are annotated by the original classification.

h) Jaccard similarity dendrogram with bandwidth 0.3 (**panel f**) showing peak relationship (first row), mapped timepoint (second row), anchor gene (third row), and mapped original annotation (forth row).

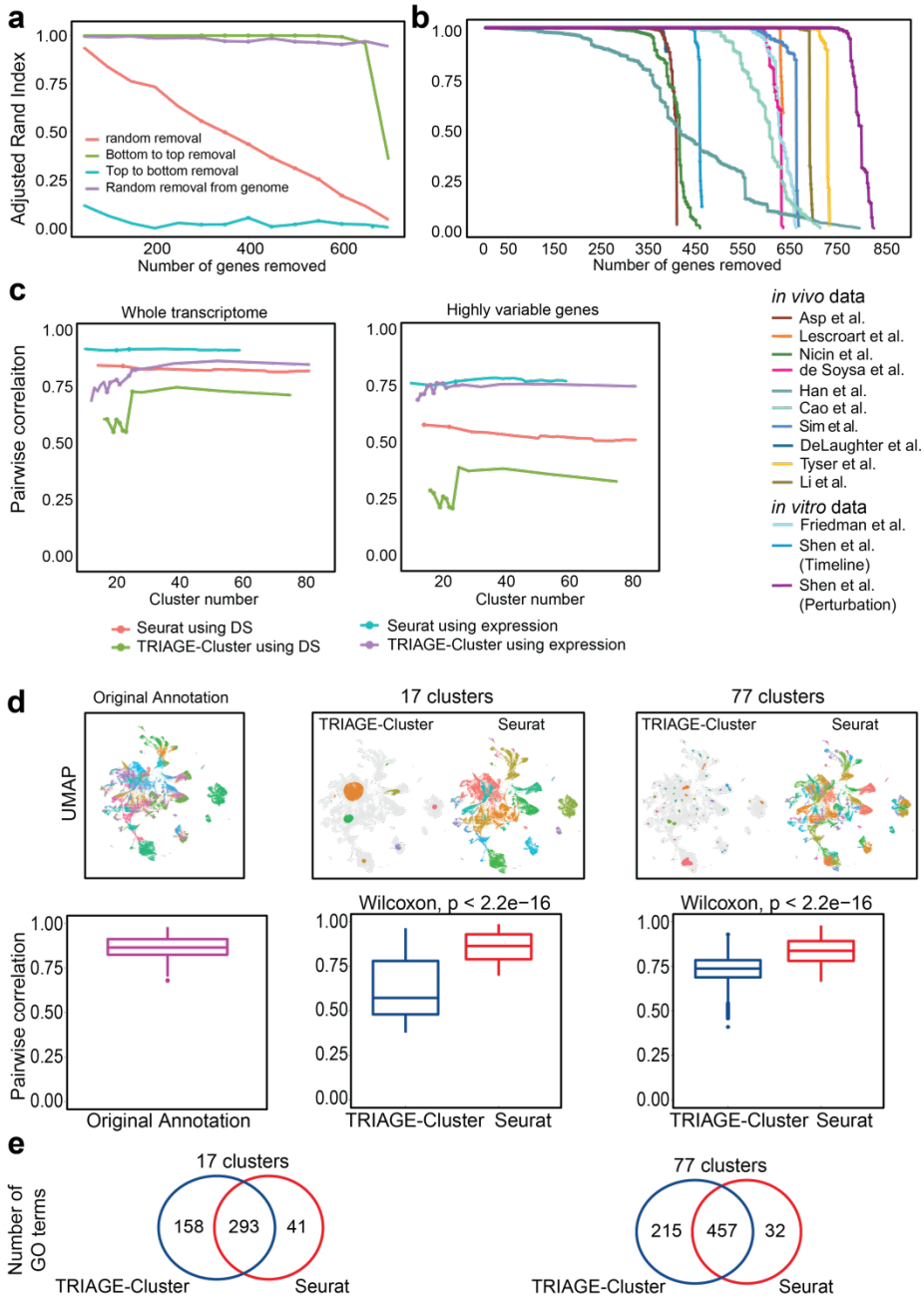


Figure 4. TRIAGE peaks identify unique cell biology.

a) Adjusted Rand Index analysis of single cell atlas of mouse gastrulation [4] demonstrating that the quantitative RTS gene ranking order is required for accurate cell clustering. Using peaks from clustering using original RTS gene rank as gold standard, these data show impact on accuracy of data structure using random removal of genes from RTS gene list, bottom to top removal of RTS genes, removal of RTS genes from top to bottom, and random removal of genes from the whole genome. Each step is 50 genes.

b) Thirteen data sets from in vivo and in vitro cell atlases analysed for clustering performance by removal of RTS genes from bottom to top demonstrating similar results that emphasize a critical role for RTS rank order in clustering accuracy.

c) Comparison of peak-peak similarity using Spearman rank correlation of whole transcriptome (left) and highly variable genes (right). We evaluate the correlation with the mouse gastrulation atlas data [4] of original expression and TRIAGE transformed matrix (DS), showing Seurat and TRIAGE-Cluster with increased number of clusters at 20 resolution/bandwidths (0.1-2 with 0.1 increments). Overall, peaks from TRIAGE-Cluster using TRIAGE transformed matrix showing the lowest correlation with highly variable genes.

d) UMAPs of the mouse gastrulation atlas data coloured by original annotations showing TRIAGE and Seurat results when cell diversity is fixed at 17 or 77 cell types.

e-f) Spearman rank correlation **(e)** and Gene Ontology analysis **(f)** demonstrate that TRIAGE-Cluster gives significantly lower correlation and therefore captures greater differences in cell diversity **(e)** and supported by increased gene ontology enrichment compared to Seurat or original annotations (**f**).

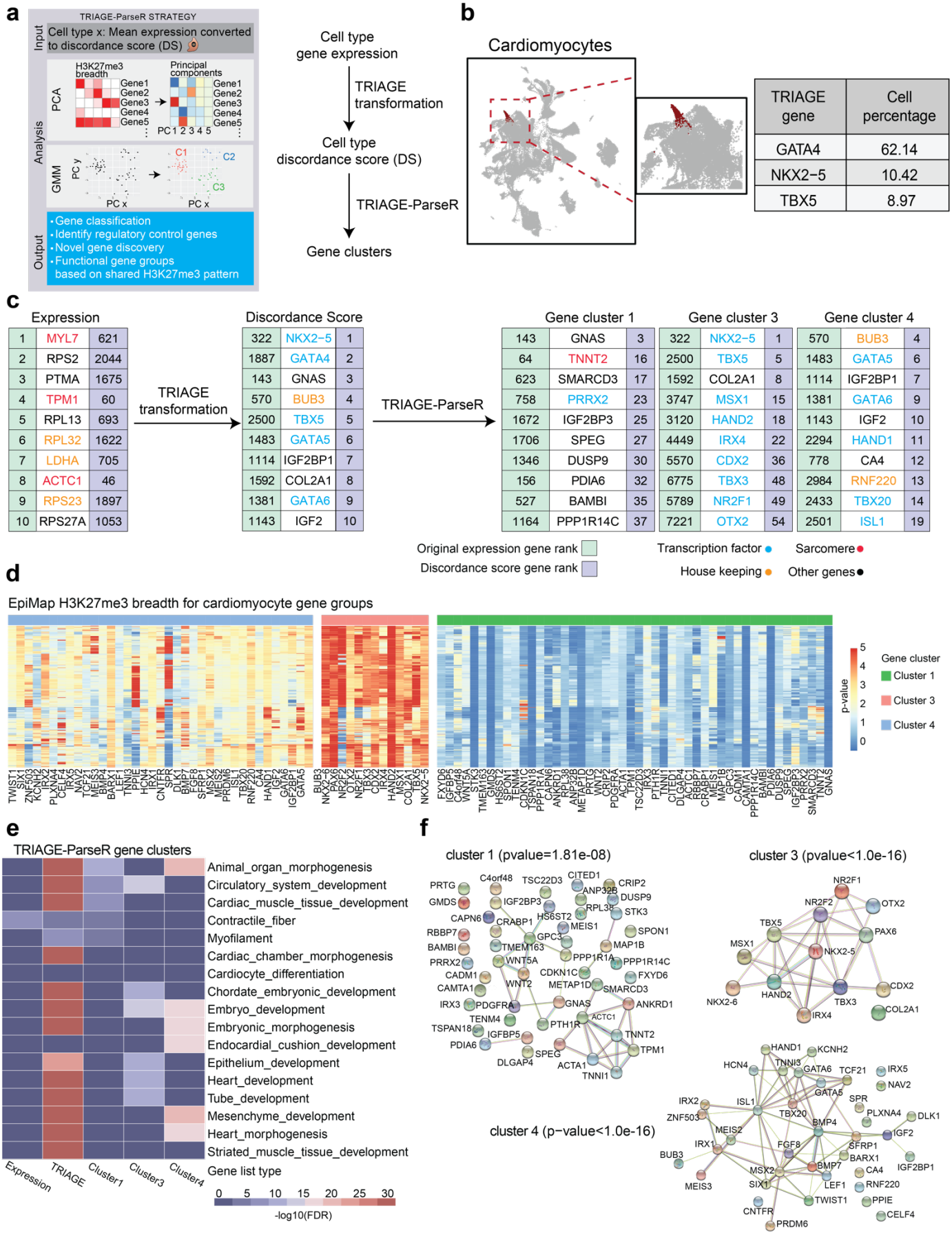


Figure 5. TRIAGE-ParseR clusters genes using unsupervised analysis of epigenetic gene TRIAGE-ParseR.

a) Schematic workflow describing TRIAGE-ParseR which uses H3K27me3 across hundreds of bio-samples to cluster genes using GMM on PCA space to improve analysis of cell gene regulatory networks.

b) UMAP of cardiomyocytes from mouse gastrulation atlas data showing top 3 TRIAGE genes expressed in 81.53% of cardiomyocytes.

c) Using cardiomyocyte pseudo-bulk RNA expression, tables show the top 10 genes ranked by original expression (left), discordance score (middle), and top 10 genes in three gene clusters resulting from TRIAGE-ParseR. Gene cluster 2 and gene cluster 5 are removed because only one gene in each cluster.

d) Abundance of H3K27me3 deposited across each gene locus measured from hundreds of EpiMap biosamples demonstrate unique epigenetic patterns demarcating specific gene groups shown in panel c.

e) Heatmap showing gene ontology enrichment (-log10 (FDR)) across top 100 genes from cardiomyocytes gene expression ranked by original expression (first column), TRIAGE discordance score (second column), and PCA-GMM TRIAGE-ParseR gene groups (last three columns).

f) String gene network analysis showing protein-protein interactions of the three cardiomyocyte gene clusters.

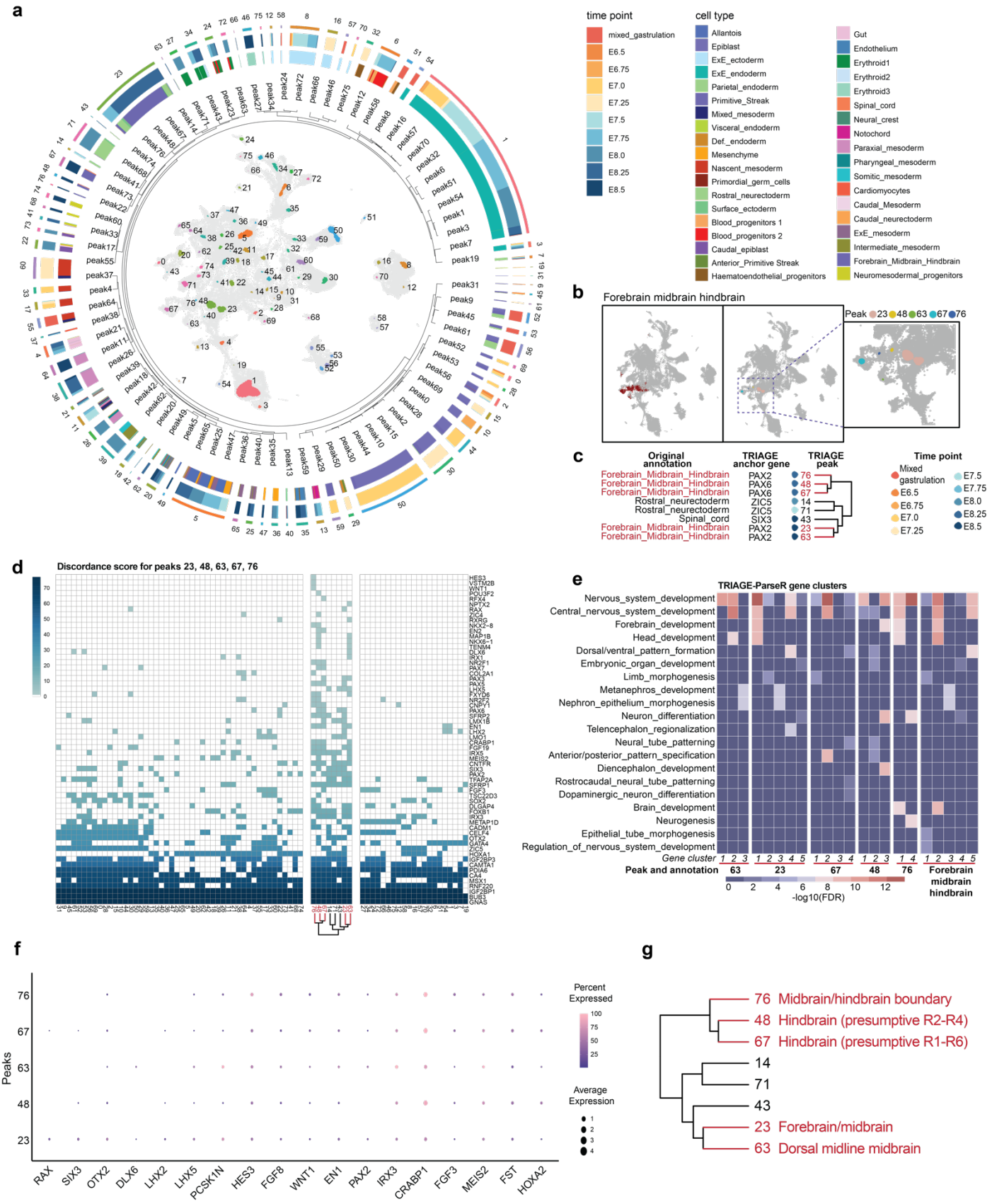


Figure 6. The TRIAGE-Cluster analysis pipeline enables efficient and unsupervised identification of cell subtype diversity in single cell data sets.

a) Integrated metadata for all peaks identified by TRIAGE-Cluster analysis in the single cell atlas of mouse gastrulation. From inside to out: data include UMAP identifying TRIAGE-Cluster peaks, peak relationships based on 0.3 bandwidth

dendrogram (**Figure 3h**), TRIAGE-Cluster peak number, proportion of cells in TRIAGE-Cluster peaks based on original annotation and timepoints. Example forebrain midbrain hindbrain is further evaluated in panels b-e.

b) UMAP of original annotation forebrain midbrain hindbrain (left) and TRIAGE-Cluster peaks within the original forebrain_midbrain_hindbrain cells (right and inset).

c) Jaccard similarity dendrogram with bandwidth 0.3 showing five distinct "forebrain midbrain hindbrain" TRIAGE-Cluster peaks with their respective mapped timepoint and TRIAGE anchor genes (extracted from **Figure 2h**).

d) Genes from TRIAGE-ParseR of the five "forebrain midbrain hindbrain" TRIAGE-Cluster peaks ordered by discordance score of the forebrain midbrain hindbrain peaks (red) and measured across 77 peaks. Output focuses on discordance score genes enriched in forebrain midbrain hindbrain samples (row) for each TRIAGE-Cluster peak (column) to facilitate cell subtype classification.

e) Heatmap showing different neural cell subtype gene ontology enrichment (-log10 (FDR)) from TRIAGE-ParseR across gene clusters in the five forebrain midbrain hindbrain peaks (the first five blocks) and original forebrain midbrain hindbrain cells (the last block).

f) Dot plot showing marker genes associated with neural lineage subtypes in each forebrain midbrain hindbrain peak.

g) Jaccard similarity dendrogram showing revised cell group classifications for the original forebrain_midbrain_hindbrain cell type using integrated analysis of TRIAGE-Cluster and TRIAGE-ParseR.

1. Boix CA, James BT, Park YP et al. Regulatory genomic circuitry of human disease loci by integrative epigenomics, Nature 2021;590:300-307.
2. Shim WJ, Sinniah E, Xu J et al. Conserved Epigenetic Regulatory Logic Infers Genes Governing Cell Identity, Cell Systems 2020;11:625-639.e613.
3. Pérez-Lluch S, Blanco E, Tilgner H et al. Absence of canonical marks of active chromatin in developmentally regulated genes, Nature Genetics 2015;47:1158-1167.
4. Pijuan-Sala B, Griffiths JA, Guibentif C et al. A single-cell molecular map of mouse gastrulation and early organogenesis, Nature 2019;566:490-495.