# Conserved epigenetic regulatory logic infers genes governing cell identity

Woo Jun Shim[1,13], Enakshi Sinniah[2,13], Jun Xu[2], Burcu Vitrinel[3], Michael Alexanian[4], Gaia Andreoletti[5], Sophie Shen[2], Brad Balderson[1], Guangdun Peng[6,7], Naihe Jing[6,7], Yuliangzi Sun[2], Yuliang Wang[9], Patrick P L Tam[10], Aaron Smith[8], Michael Piper[11,12], Lionel Christiaen[3], Quan Nguyen[2], Mikael Bodén[1,*], Nathan J. Palpant[2,11,*,#]

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

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

[3] Center for Developmental Genetics, Department of Biology, New York University, New York, NY, USA

[4] The Gladstone Institute, University of California San Francisco, San Francisco, CA, USA

[5] Institute for Computational Health Sciences, University of California, San Francisco, CA 94158, USA

[6] CAS Key Laboratory of Regenerative Biology, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Guangzhou 510530, China; and Guangzhou Regenerative Medicine and Health GuangDong Laboratory (GRMH-GDL), Guangzhou 510005, China.

[7] State Key Laboratory of Cell Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, 200031, China

[8] Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology, Brisbane, Australia

[9] University of Washington, Department of Computer Science, Seattle, WA, USA

[10] The University of Sydney, Children's Medical Research Institute, and School of Medical Sciences, Faculty of Medicine and Health, Westmead NSW 2145 Australia

[11] School of Biomedical Sciences, The University of Queensland, Brisbane, Australia

[12] Translational Research Institute, Woolloongabba, Brisbane, Australia

[13] These authors contributed equally to this work

[*] Co-senior and co-corresponding authors

[#] Lead contact

## Contact information:

Nathan Palpant
Institute for Molecular Bioscience
The University of Queensland
Brisbane, Australia
T: 61 0439 241 069
E: n.palpant@uq.edu.au

Mikael Bodén
School of Chemistry and Molecular Biology
The University of Queensland
Brisbane, Australia
T: 61 07 336 51307
E: m.boden@uq.edu.au

**SUMMARY**

Determining genes orchestrating cell identity and function in development and disease remains a fundamental goal of cell biology. This study establishes a genome-wide metric based on the gene-repressive tri-methylation of histone 3 lysine 27 as deposited in over 100 human cell states from representative tissues and cell lines. On its own, the tendency of broad H3K27me3 occupancy at promoters strongly enriches for genes that drive cell diversification and fates. We show that the discordance between this repressive tendency and the abundance of expressed transcripts of any somatic cell type prioritizes cell type-specific regulatory genes in health and disease. We implement this repression-based regulatory logic to identify genetic drivers of cell identity across millions of genome-wide single cell transcriptomes, diverse omics platforms, and eukaryotic cells and tissue types. Its potential for novel gene discovery is demonstrated by experimentally validated predictions of previously unknown drivers of organ differentiation in two eukaryotic species, humans and *Ciona*. This simple and quantifiable regulatory inference analysis provides a novel and scalable computational approach to determine drivers of cell diversification and fates of any cell type from gene output alone.

## INTRODUCTION

Identifying regulatory factors underlying mechanisms of cell-specific fate and function remains a challenge in cell and developmental biology, reprogramming and regenerative medicine. Emerging technologies are enabling the capture of genome-wide information at the level of a single cell, providing unprecedented access to data describing cellular diversity in development and disease (Cao et al., 2019; Mulqueen et al., 2018; Ramani et al., 2017; Satija et al., 2015). However, efforts to computationally reveal the underlying causative factors of cell subpopulations and their relationships and pathways are challenged by the abundance of data that reliably recover structural and house-keeping genes at the cost of lowly abundant regulatory factors.

A cell's identity is determined and controlled by an interplay of regulatory molecules such as transcription factors (TFs), which cause changes in gene expression across the genome (Morris and Daley, 2013). While often multiple TFs act coordinately, a single transcription factor can be sufficient to reprogram cell states (Lattanzi et al., 1998). Changes in histone modifications (HMs) selectively activate or repress transcription of regulatory genes to govern cell decisions during differentiation and homeostasis, by mechanisms that are partially conserved across species (Boyer et al., 2006; Margueron and Reinberg, 2011; Nakamura et al., 2014).

HMs are structurally and functionally linked to cell type specific genome architecture and gene regulation (Cahan et al., 2014; Rehimi et al., 2016). H3K27me3 is a chromatin mark deposited by the polycomb repressive complex (PRC2) to suppress expression of genes, particularly those governing development and differentiation (Margueron and Reinberg, 2011). H3K27me3 plays a juxtaposing role in regulating gene expression relative to activating marks such as H3K4me3 and H3K27ac. The interplay of epigenomic control of gene expression is critical to cell lineage decisions, protecting the integrity of gene networks to safeguard cell differentiation pathways for derivation of specific functional states (Van Handel et al., 2012). Computational approaches incorporating genome-wide chromatin and transcriptional data have been used

to predict and prioritize genes with a capacity to drive cell fates (Benayoun et al., 2014; Rehimi et al., 2016; Whyte et al., 2013). While effective, a major limitation of these approaches is the requirement for input chromatin data. To collate histone modification data for all cell types is prohibitive, and currently not feasible in analyses at single cell level. To address this problem, we use available histone modification data across a diverse sampling of cellular states to infer genes that drive and determine cell identity of *any* cell based on transcript abundance. Here, we demonstrate that a computational method formulated around the repressive tendency *via* H3K27me3 is strongly predictive of regulatory genes governing mechanisms of cell differentiation and organ morphogenesis. We show that the method can analyze millions of individually heterogeneous cell transcriptomes simultaneously and infer cell type-specific regulatory genes from diverse species in the animal kingdom.

## DESIGN

Studies have demonstrated the capacity to infer genes that determine cell fate by analysis of chromatin or transcriptional data from a cell type of interest (Benayoun et al., 2014; Cahan et al., 2014; D'Alessio et al., 2015; Davis and Eddy, 2013; Heinaniemi et al., 2013; Lang et al., 2014; Rackham et al., 2016; Rehimi et al., 2016; Roost et al., 2015). However, approaches to date lack generalizable principles, thus limiting their utility and scalability. We define the *a priori* probability of broad H3K27me3 domains at a promoter, representing the repressive tendency of a gene across diverse tissues and conditions (Kundaje et al., 2015). TRIAGE (**T**ranscriptional **R**egulatory **I**nference **A**nalysis of **G**ene **E**xpression) capitalizes on the discordance between measured gene expression output in transcriptomic data and the correlated repressive tendency of detected genes. The resulting output ranks genes based on their regulatory potential. We test the versatility of TRIAGE in identifying genetic drivers of cell identity from single cell transcriptomes and other diverse quantitative measures of gene expression including ChIP-seq, CAGE-seq, bulk RNA-seq, and LC MS/MS.

Epigenetic features of open chromatin such as H3K4me3 and super-enhancers are known to mark genes associated with specific cell states (Benayoun et al., 2014; Whyte et al., 2013). By definition, this limits their application to assays where these features can be measured, with the information relevant only to cell types of interest. In contrast, TRIAGE relies on the selective *absence* of repressive marks. To demonstrate the distinctive advantage of TRIAGE, we evaluate the complementarity and redundancy of HMs and demonstrate its superior performance against other epigenetic inference methods. With new technologies and capabilities in studying the genetic state of individual cells, analytical tools like TRIAGE that leverage generalizable principles from genomic data will enhance our capacity to understand the mechanistic basis of cellular heterogeneity in development, homeostasis and disease conditions. The capacity to forego cell type specific HM data enables the analysis of millions of heterogeneous cell types, including analyses across species by reference to genetic orthologs.

## RESULTS

### Broad H3K27me3 domains demarcate cell type-specific regulatory genes

We took the approach that the genome is equivalent to an information source that can exist in a continuum to derive a theoretically infinite number of specific cell-states. On this basis, we hypothesized that analyzing the genome of diverse cell states could be used to determine variations in epigenetic control of regulatory genes underpinning biological complexity. We used NIH Epigenome Roadmap data (Kundaje et al., 2015), which contains ChIP-seq data for H3K4me3, H3K36me3, H3K27me3, H3K4me1, H3K27ac and H3K9me3 from 111 tissue or cell types **(Table S1)**. To associate HMs with proximal gene regulatory functions, we linked the broadest HM domain within 2.5 kb to known transcriptional start sites of RefSeq genes. For each HM, we found that the top 100 genes with the broadest domain were remarkably consistent between cell types **(Figure S1A)**, however, broad domains of different HMs marked distinct sets of genes **(Figure S1C)**.

We next assessed the breadth of histone domains in each of 111 samples as they correlate with genes governing cell identity. To establish a positive gene set for cell type-specific regulatory genes, we populated a list of 634 variably expressed transcription factors (VETFs) having a coefficient of variation greater than 1 across 46 NIH Epigenome RNA-seq data sets (Kundaje et al., 2015; Perez-Lluch et al., 2015) (**Table S2**). We used Shannon entropy to quantify cell type specificity (Schug et al., 2005) and demonstrate that VETFs are significantly more cell type-specific, compared to non-VETFs or protein coding genes (**Figure 1A inset, Figure S1B)**. This demonstrates that VETFs provide a positive gene set where their enrichment is a performance metric for identifying cell type-specific regulatory genes.

To calculate the enrichment of cell-specific regulatory genes as they relate to HM domains, all NIH Epigenome histone ChIP-seq data were ranked by breadth, comprising greater than thirteen million peaks, and analyzed using Fisher's exact test to assess enrichment of VETFs. These data show that H3K27me3 uniquely enriches for VETFs within the top 5% of broad domains **(Figure 1A)**, consistent with a strong correlation between repression of VETFs and breadth of H3K27me3 domains **(Figure S1D)**. Further supporting this, unsupervised analysis of genes marked by the broadest H3K27me3 domains are uniquely enriched in morphogenic and developmental regulators **(Figure 1B)**. Taken together, this demonstrates that H3K27me3 broad domains from diverse cell and tissue types provides a powerful strategy to reproducibly enrich regulatory genes governing the biological complexity of diverse cell states.

To illustrate the distinctive enrichment of H3K27me3 in regulatory genes as opposed to structural or housekeeping genes, we extracted expression and chromatin data from cardiomyocytes **(Figure 1C-D)**. We show that the transcript abundance of cardiac regulatory genes (i.e. *GATA4*, *GATA6*, *NKX2-5*, *TBX5* and *TBX20*) and structural sarcomere genes (i.e. *MYH6*, *MYH7*, *MYL2*, *MYL3* and *TNNI3*) are all significantly elevated in cardiac cells compared to other cell types, but cannot be distinguished as regulatory or structural genes except by differential expression **(Figure 1C).** In contrast, in all cell types *except* the heart, H3K27me3 domains broader than 30kb consistently identify cardiac regulatory genes from structural genes **(Figure 1C)**. No other HM analyzed demarcates cell type-specific regulatory genes from structural genes

in this manner **(Figure 1D, Figure S1E).** This establishes the rationale that genes having cell type-specific regulatory functions can efficiently be identified based on the frequency of H3K27me3 across the locus in heterogeneous cell types.

**The repressive tendency of a gene defines a genome-wide metric predictive of cell type-specific regulatory genes**

We established a simple, quantitative logic that leverages the significance of broad H3K27me3 domains for distinguishing regulatory genes. Deposition of broad H3K27me3 domains allows for setting the default gene activity state to "off" such that cell type-specific activity occurs by rare and selective removal of H3K27me3 while all other loci remain functionally repressed (Boyer et al., 2006; Lee et al., 2006). Conversely, genes with housekeeping or non-regulatory roles rarely host broad H3K27me3 domains.

We therefore calculated for each gene in the genome across 111 NIH epigenome cell and tissue types **(i)** the sum of breadths of H3K27me3 domains in base-pairs and multiplied this by **(ii)** the proportion of cell types in which the gene's H3K27me3 breadth is within the top 5% of broad domains **(Figure 2A).** This approach quantifies a single value for every gene that defines its association with broad H3K27me3 domains which we call its repressive tendency score (RTS) **(Table S3).** Using the NIH Epigenome Roadmap data, the RTS is calculated for 99.3% (or 26,833 genes) of all RefSeq genes. To demonstrate that this formulation is agnostic to the composition of cell types, we note that for all genes, the RTS is within one standard deviation of the mean of bootstrapping empirical distribution derived from 10,000 re-samplings of cell types. We note that the 111 cell types provided sufficient sample size to calculate a stable RTS independent of the peak calling method to define H3K27me3 domains, with over 85% of assigned H3K27me3 domains overlapping only a single protein-coding gene **(Figure S2A-C)**. The RTS only requires sufficient subsampling of H3K27me3 from any diverse collection of cell states to establish a stable metric.

Using RTS values above the inflection point (RTS>0.03022) of the interpolated RTS curve, we identified a priority set of 1,359 genes. These genes show significant enrichment for cell type specific, lowly expressed regulators of cellular diversification including organ development, pattern specification and multicellular organismal processes **(Figure 2B-D)**. Among the 1,359 priority genes, we identified significant enrichment of VETFs (**Figure 2E,** *p*=3.49e-162, Fisher's exact test, one-tailed), homeobox proteins (Zhong and Holland, 2011) (*p*=6.43e-123), signaling factors (Kanehisa and Goto, 2000), and non-coding RNAs including known regulators of development such as FENDRR and HOTAIR (Grote and Herrmann, 2013; Rinn et al., 2007). Furthermore, genes with high RTS values are enriched in regulators of fundamental biological processes including gastrulation and organ morphogenesis, comprise members of major signaling pathways, and genetic determinants of diverse pathologies including cardiovascular disease, diabetes, neurological disorders and cancer **(Figure 2F and Table S4).** Taken together, these data indicate that ranking based on a gene's repressive tendency is a simple and effective strategy to enrich for fundamental genetic determinants of biological complexity.

**RTS values provide a reference for predicting regulatory genes of a cell**

The transcriptome of a cell derives from the portion of the genome comprising cell-specific expression of structural, housekeeping and regulatory genes collectively representing the cell's state. Regulatory genes, such as transcription factors (TFs), controlling the identity of a cell are rare and difficult to identify due to their typical low abundance. We tested whether RTS values provide a mechanism for re-calculating gene expression data from any cell to rank the relative abundance of genes based on their regulatory potential, a method we call TRIAGE. For any gene *i* the product between a gene's expression ($Y_i$) and repressive tendency ($R_i$) gives rise to its discordance score ($D_i$) as defined by:

$$D_i = \ln (Y_i + 1) \cdot R_i$$

The discordance score (DS) reflects the juxtaposition of a gene's association with being epigenetically repressed and the observed transcriptional abundance of that gene in the input data. TRIAGE therefore introduces a non-linear, gene-specific weight designed to amplify the relative abundance measure of cell type-specific regulatory genes from expression data derived from any cellular state **(Figure 3A, Figures S2D, S3A-B)**.

As expected, analysis of cell-specific genes across 111 NIH Epigenome Roadmap samples shows that H3K27me3 broad domains reproducibly mark regulatory genes as opposed to structural or housekeeping genes **(Figure 3B)**. Using cell-specific RNA-seq data transformed with RTS values as described above, we show that TRIAGE efficiently reduces the relative abundance of structural and housekeeping genes, while enriching for regulatory genes in a cell type-specific manner **(Figure 3C, Figure S3C).** Taken to scale, TRIAGE transformation of all 46 Roadmap samples results in enrichment of tissue or cell type-specific TFs among the top 1% in every cell type while it reproducibly reduces the relative abundance of housekeeping genes **(Figure 3D).** Analysis of the Pearson distances between Roadmap tissue types (Scornavacca et al., 2011) shows that TRIAGE increases the similarity between samples from the same tissue by ~29% when compared to distances calculated using absolute expression levels **(Figure S3D)**.

**Identifying cell identity genes from diverse tissues and species**

Regulatory genes underlying cell identity are predominantly evolutionarily conserved. Using inter-species gene mapping, we tested whether TRIAGE could identify regulatory drivers of heart development across diverse chordate species including mammals (i.e. *Homo sapiens*, *Mus musculus*, and *Sus scrofa*), bird (*Gallus gallus*), fish (*Danio rerio*) and invertebrate tunicate (*Ciona robusta*) **(Figure 3E)**. In contrast to expression alone, TRIAGE recovered cardiac regulatory genes with high efficiency across all species. More broadly, we used TRIAGE to enrich for relevant tissue morphogenesis biological processes from diverse cell types and species including arthropods **(Figure 3F).** While TRIAGE is currently devised using human

epigenetic data, this suggests that TRIAGE can be used to identify regulatory genes from cell types that are conserved across the animal kingdom**.**

**Determining the regulatory control points of disease**

To assess the potential for TRIAGE for informing disease mechanisms, we analyzed genetic data from melanoma and heart failure (HF) pathogenesis. We used TRIAGE to analyze single cell RNA-seq data from 1,252 cells capturing a transition from proliferative to invasive melanoma (Tirosh et al., 2016). Among the top ranked genes, TRIAGE consistently outperforms expression in prioritizing genes with known involvement in melanoma proliferation and invasion **(Figure S4A)**. Using independently derived positive gene sets for proliferative versus invasive melanoma (Tirosh et al., 2016; Verfaillie et al., 2015), TRIAGE recovers with high sensitivity the genetic signatures of these two cancer states **(Figure S4B)**.

We next aimed to assess whether TRIAGE could identify transcriptional signatures of therapeutic interventions in heart failure (HF). Previous studies have shown that BRD4, a member of the BET (Bromodomain and Extra Terminal) family of acetyl-lysine reader proteins, functions as a critical chromatin co-activator during HF pathogenesis that can be pharmacologically targeted *in vivo* (Anand et al., 2013; Duan et al., 2017) to prevent and treat HF by targeting gene programs linked to cardiac hypertrophy and fibrosis (Duan et al., 2017). We analyzed RNA-seq data from adult mouse hearts where pre-established HF (transverse aortic constriction, TAC) was treated with JQ1. TRIAGE enriched for TFs and regulatory genes with known roles in HF pathogenesis **(Figure S4C)** and outperformed expression ranked genes based on enrichment for stress-associated positive gene sets **(Figure S4D)**. Importantly, comparing Sham, TAC and TAC+JQ1, TRIAGE-based ranked genes highlighted a potent anti-fibrotic effect of JQ1 without the use of a canonical differential expression analysis **(Figure S4D)**. Collectively, these data demonstrate the versatility of TRIAGE as a scalable strategy that identifies mechanistic insights in disease aetiology and therapy.

## TRIAGE performance against benchmark prediction algorithms

We evaluated performance of TRIAGE against previous epigenetic inference methods including functional heterogeneity (FH) analysis (Rehimi et al., 2016), H3K4me3 broad domains (Benayoun et al., 2014) and super-enhancers (Jiang et al., 2019; Whyte et al., 2013) **(Figure 4A)**. TRIAGE is unique among these analysis approaches, requiring only gene expression data without the need for epigenetic sequencing input. Comparing the overlap of genes prioritized by these approaches across diverse tissue types reveals largely non-overlapping genes sets (**Figure 4B**). For example, among cardiac transcription factors identified within each method, TRIAGE distinctively detects the largest number of cardiac transcription factors, many known to be central regulators of cardiac morphogenesis including *IRX4*, *GATA5*, *TBX5*, *MSX1*, and *EOMES* (**Figure 4B**) (Russ et al., 2000; Takeuchi and Bruneau, 2009; Waardenberg et al., 2014). This is confirmed across diverse tissues showing TRIAGE surpassing these methods in terms of sensitivity and precision at identifying cell type-specific regulatory genes **(Figure 4C-D, Figure S5A-B)**. Lastly, we show that TRIAGE predictions are not explained merely by prioritising TFs **(Figure S5C-D)** and does not simply recover lineage-specific genes based on the loss of H3K27me3 during development **(Figure S5E-F)**. These data demonstrate that TRIAGE is a gene enrichment strategy distinctive from prior benchmark approaches and significantly outperforms them in detecting regulatory drivers of cell states.

## Using TRIAGE to dissect the mechanistic basis of cell heterogeneity at single-cell resolution

Recent developments in barcoding and multiplexing have enabled scalable analysis of thousands to millions of cells (Cao et al., 2019). Determining mechanistic information from diverse cell states captured using single-cell analytics remains a challenge. TRIAGE is scalable for studies of cell heterogeneity because it is a simple metric, customizable to any bulk or single cell level transcriptomic data input.

We analyzed 43,168 cells captured across a 30 day time-course of *in vitro* cardiac-directed differentiation from human pluripotent stem cells (hPSCs) (Friedman et al., 2018). Analysis of day-30 cardiomyocytes using standard expression data show high abundance genes dominated by housekeeping and sarcomere genes, whereas TRIAGE efficiently identifies regulatory genes governing cardiomyocyte identity including *NKX2-5*, *HAND1*, *GATA4*, *IRX4* **(Figure 5A-B).** Importantly, TRIAGE retains highly expressed cell-specific structural genes providing an integrated readout of genes involved in cell regulation and function **(Figure 5C)**. We used TRIAGE to convert the genes-by-cells matrix comprising ten different subpopulations spanning gastrulation stage, progenitor and definitive cell types **(Figure 5D)**. In contrast to expression data, which significantly enriches for structural and housekeeping genes, TRIAGE consistently identifies gene sets associated with developmental regulation of diverse and biologically distinct subpopulation through differentiation **(Figure 5E, Figure S6A-B)**. We show that differential expression results in variable outcomes depending on the comparison cell type and consistently under-performs against TRIAGE, which identifies population-specific regulatory genes without external reference comparisons **(Figure 5F).**

**Scalable application of TRIAGE across diverse omics data types**

The simplicity of TRIAGE facilitates its use as a scalable application. We used VETFs as a positive gene set to test enrichment of regulatory genes across millions of diverse cell types by plotting the rank position of the peak significance ($-\log_{10}p$) value in a Fisher's exact test. Using tabula muris data of nearly 100,000 cells from 20 different mouse tissues at single-cell resolution (Schaum et al., 2018), TRIAGE consistently enriches for cell type-specific regulatory genes compared to original expression with no difference between droplet and smartseq2 data sets **(Figure 5g and Table S5).** Using the mouse organogenesis cell atlas (MOCA), which is among one of the largest single cell data sets generated to date (Cao et al., 2019), we demonstrated that TRIAGE outperformed expression values alone in prioritizing cell type-specific regulatory genes across more than 1.3 million mouse single-cell transcriptomes **(**Wilcoxon rank-sum test

for both median significance and rank, $p<2.2e-16$**,** one-tailed, **Figure 5H).** Lastly, we used benchmarking data to evaluate clustering accuracy based on ground truth (Tian et al., 2019) to assess the performance of TRIAGE using three independent algorithms (i.e. CORE, sc3, and Seurat) and show no difference in accurately assigning cells to the reference (ARI > 0.98) using original expression or TRIAGE transformed expression **(Figure 5I).**

We hypothesized that TRIAGE could be used to study any genome-wide quantitative measurement of gene expression. To test this, TRIAGE was applied to diverse quantitative data across hundreds of different cell types. TRIAGE outperforms original abundance metrics when measuring chromatin methylation for H3K36me3, a surrogate of RNA polymerase II activity deposited across gene bodies (Barski et al., 2007) collected from the 111 Roadmap samples **(Figure 5J)**. Similarly, cap analysis of gene expression (CAGE), which measures genome-wide 5' transcription activity, showed significant enrichment of VETFs using TRIAGE from 329 selected FANTOM5 CAGE samples **(Figure 5J and Table S1)** (Forrest et al., 2014). Lastly, analysis of a draft map of the human proteome shows that TRIAGE enriches for regulatory drivers of 30 different tissue types from high resolution Fourier transform mass spectrometry data (Kim et al., 2014) **(Figure 5j)**. Taken together, these data illustrate the power of utilizing TRIAGE to predict regulatory drivers of cell states using diverse genome-wide multi-omic endpoints.

**Discovering novel regulatory drivers of cardiac development in human and Ciona**

We set out to demonstrate TRIAGE as an engine for novel gene discovery. Using data from single cell analysis of cardiac differentiation (Friedman et al., 2018) we analyzed day 2 subpopulations and identified diverse, known regulatory genes governing mesendoderm cell differentiation **(Figure 6A)**. Among the TRIAGE identified genes was *SIX3*, a member of the sine oculis homeobox transcription factor family (RTS=0.54) **(Figure 6A-B)**. All pairwise differential expression analyses failed to enrich for *SIX3* **(Figure S7A).** Though the role of *SIX3* in neuroectoderm specification has been studied extensively, little is known

about its role in other germ layer derivatives (Carl et al., 2002; Lagutin et al., 2003; Steinmetz et al., 2010). Analysis of *SIX3* in hPSC *in vitro* cardiac differentiation shows robust expression in day 2 definitive endoderm (DE) (28.7%) and mesoderm (37.5%) cell populations **(Figure 6C)** with enrichment of *SIX3*$^+$ cells associated with definitive endoderm **(Figure S7B-C**). We analyzed the spatiotemporal transcriptional data from germ layer cells of mid-gastrula stage (E7.0) embryos (Peng et al., 2016), with an expanded analysis to include pre- (E5.5-E.6.0), early- (E6.5) and late-gastrulation (E7.5) mouse embryos (**Figure S7D**). Spatio-temporal expression of *SIX3* and other family members is observed in the epiblast and neuroectoderm, (**Figure 6D, Figure S7E)** consistent with its known role in these lineages (Carl et al., 2002; Lagutin et al., 2003; Steinmetz et al., 2010), as well as early endoderm lineages (**Figure 6D**).

We established CRISPRi loss-of-function hPSCs in which *SIX3* transcription is blocked at its CAGE-defined transcription start site (TSS) in a dox-dependent manner **(Figure 6E-F)**. Cells were differentiated using monolayer cardiac differentiation and analyzed at day 2 **(Figure 6G).** *SIX3* loss-of-function depleted endoderm and mesendoderm genes **(Figure 6H**) consistent with FACs analysis showing depletion of CXCR4$^+$/EPCAM$^+$ endoderm cells **(Figure 6i-K, Figure S7F, Figure S8)**. In contrast, FACs analysis of alpha-actinin$^+$ cardiomyocytes showed no difference between *SIX3*-knockdown cells compared to dox-treated controls indicating that loss of *SIX3* does not impact mesodermal fates **(Figure 6L-N, Figure S7G).** Taken together, these data demonstrate a novel role of *SIX3* in governing endoderm differentiation.

We also used TRIAGE to identify novel developmental regulators in a distant chordate species, *Ciona robusta.* Single cell RNA-seq data comprising cell subpopulations captured across a time-course of cardiac development were analyzed with TRIAGE using a customized gene mapping tool to link human to *Ciona* genes **(Figure 6O)** (Wang et al., 2019). The top ranked genes based on TRIAGE were analyzed **(Figure 6P)**. *RNF220* (RTS=0.30, **Figure 6Q**), an E3 ubiquitin ligase governing Wnt signaling pathway activity through β-catenin degradation (Ma et al., 2014), was identified as a novel regulatory gene not previously implicated in cardiopharyngeal development. Utilizing CRISPR control vs. *RNF220*-knockout, we demonstrate that *Mesp* lineage progenitors of control animals form the expected ring of pharyngeal muscle

progenitors around the atrial siphon placode, whereas *RNF220*-knockout embryos showed significant morphogenetic defects **(Figure 6R-S)**. Collectively, these data illustrate that TRIAGE efficiently identifies novel functional regulatory determinants as a demonstration for discovering novel biology underlying mechanisms of development.

## DISCUSSION

TRIAGE builds on the prevailing view that epigenetic states guide regulatory processes. The selective absence of broad H3K27me3 domains at promoters is first established from diverse cell types and epigenetic states hosted by NIH Roadmap Epigenomics (Kundaje et al., 2015) and represented by a simple gene-specific repressive tendency score. The RTS demarcates 5% of Refseq genes that are consistently regulated by broad H3K27me3 domains across major human cell lineages; we show this priority set is highly enriched in genes that govern cell differentiation and lineage diversification.

Secondly, TRIAGE exploits the discordance between any gene's expression and its RTS to predict cell-specific genes regulating cell identity with sensitivity and precision. We demonstrate a capacity of TRIAGE to prioritize genes involved in cell differentiation decisions during developmental lineage diversification from pluripotency as well as disease state transitions as observed, for example, in proliferative and invasive states of melanoma. These inferences establish a genetic basis for gaining mechanistic insights into cell decisions, including the discovery of novel genetic regulators as demonstrated by gene loss of function studies.

The evolutionary conservation of epigenetic regulation suggests that the repressive tendency can be applied across eukaryotic cell types where gene expression is governed by the polycomb group complex-2. PRC2 and its regulation of histone methylation are known to govern genes in protists, animals, plants as well as fungi (Margueron and Reinberg, 2011). Indeed, zebrafish (Wu et al., 2011) and medaka (Nakamura et al., 2014) genes with broad H3K27me3 deposition at promoter sites encode master developmental regulators overlapping with those found in our study. This illustrates the conservation of PRC2-mediated H3K27me3

regulation and repression of genomic loci across species and its role in safeguarding cell identity (Boyer et al., 2006; Fujikura et al., 2002).

In contrast to previous strategies to identify cell regulatory genes (Cahan et al., 2014; Rackham et al., 2016), TRIAGE has several key advantages: it does not require comparison against external reference states (akin to differential gene expression analysis); it does not require external metadata (e.g. gene and network databases), the selection of which are arbitrary, and subject to the limitations of prior knowledge; its tested performance is more sensitive in detecting known regulatory genes compared to earlier powerful approaches such as FH Score (Rehimi et al., 2016), H3K4me3 broad domains (Benayoun et al., 2014), or super-enhancers (Creyghton et al., 2010; Jiang et al., 2019; Whyte et al., 2013); it provides a seamless interface with a range of input data types measuring gene outputs of a cell; and importantly, TRIAGE scales easily with cell input number as seen in emerging single cell experiments, with application by genetic orthologs for the study of any cell type across the animal kingdom.

This study establishes a new computational strategy for the study of genomes and the transcriptional control mechanisms governing cell diversification. With expanded analysis of epigenomic states of diverse cell and tissue types across eukaryotic species, the analysis principles described here can be adapted or implemented to study mechanisms underlying cell biology, reprogramming and regenerative medicine, the genetic basis of complex traits, as well as gaining insights into evolutionary biology and genetic adaption across eukaryotic genomes governed by PRC2.

## LIMITATIONS

We acknowledge the complexity of regulatory control of cell states and the scope of insight provided by TRIAGE in revealing underlying drivers of cell identity. Promoters with high repressive tendency capture a broad diversity of molecules such as transcription factors, signaling molecules, cell adhesion and matrix molecules and non-coding RNAs. Genes identified by TRIAGE provide a complementary and largely non-

overlapping enrichment of genes when compared against to other chromatin features such as broad H3K4me3 and super-enhancers that also provide powerful methods for assessing regulatory determinants of cells. Due to the parameters of analysis, TRIAGE focuses on genes governing cell type-specific functions as opposed to homeostatic, non-specific regulatory factors, and therefore does not aim to enrich for all potential regulatory genes.

TRIAGE involves a simple calculation that does not attempt to integrate additional data beyond H3K27me3 deposition at promoters. We view the simplicity of TRIAGE as fundamental to its ability to uniquely scale across input data types from diverse cell states and species. Routine pipeline approaches for gene set analysis can be used in parallel with TRIAGE.

We used RTS values mapped specifically to protein coding genes. This limits TRIAGE inference predictions to the 2% of the genome encoding proteins and precludes analysis of regulatory intergenic domains that also encode information critical for governing cell identity (Andersson et al., 2014). Furthermore, we note limitations on cross-species applications based on gene mapping tools and the profound differences in composition of cellular diversity that distinguish animal phyla. Despite this limitation, we demonstrate cross-species predictions and functional evidence of novel gene discovery in distantly related chordates using human RTS values.

While robustly tested for consistency across millions of data sets, the repressive tendency scores are based on the biological diversity represented in the 111 NIH Roadmap Epigenomics samples and will require further assessment based on availability of new H3K27me3 data. We anticipate that the generation of new reference epigenome databases (Adsera et al., 2019) covering a broader range of cell and tissue-types will refine the precision of RTS revealing stress-sensitive loci and novel disease drivers.

## METHODS

**Data sets.** We used broad peak representations for 6 different histone modifications, namely H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K27ac and H3K36me3, for 111 human cell types from the NIH Roadmap database (Kundaje et al., 2015). We used identifiers from Animal TFDB and DBD to identify 1,605 human TFs (Wilson et al., 2008; Zhang et al., 2015). We also identified 3,804 housekeeping genes from the published literature (Eisenberg and Levanon, 2013).

For performance evaluation of TRIAGE, we selected 5 distinct tissue-groups; (i) Brain, (ii) Lung, (iii) Pancreas, (iv) Skeletal muscle and (v) Heart. We used transcriptomic abundance as well as HM ChIP-seq data from the Roadmap; Brain germinal matrix (E070) for the brain, Lung (E096) for the lung, Pancreas (E098) for the pancreas, Psoas muscle (E100) for the skeletal muscle and a published data set for cardiac progenitor cells for the heart (GSE97080) (Palpant et al., 2017b). We also used GTEx expression profiles of samples that belongs to the 5 tissue-groups (Lonsdale et al., 2013). The GTEx database provides a large number of sample transcriptomes covering a wide range of tissue-groups. We downloaded expression profiles of all available samples for the 5 selected tissue groups; 417 samples from the left ventricle (Heart), 195 samples from the cortex (Brain), 268 samples from the pancreas (Pancreas), 607 samples from the lung (Lung) and 718 samples from the skeletal muscle (Skeletal muscle). We averaged gene expression values of samples for each tissue-group.

To define a gene set regulating a given cell fate, we used gene ontology (GO) annotation data (Ashburner et al., 2000). We identified TFs with a relevant GO biological process (BP) term as the regulatory gene set. We chose the GO term to not only represent the identity of the tissue type but also to capture a spectrum of relevant regulatory genes. For instance, we used ‘heart development’ term (GO:0007507) to select a set of TFs specific for the heart development. Similarly, ‘brain development’ (GO:0007420), ‘lung development’ (GO:0030324), ‘pancreas development’ (GO:0031016) and ‘muscle structure development’ (GO:0061061) GO terms were used to collect regulatory gene sets for the brain, lung, pancreas and skeletal muscle

samples, respectively (**Table S2**). We identified expressed (i.e. RPKM>1) regulatory genes in a given input data set as the positive gene set.

**Identifying the gene set as a proxy for cell type-specific regulatory genes.** Along with TFs with a tissue-type-specific GO BP term, we identified a gene set that universally represents cell type-specific regulatory genes. We identified 634 TFs whose expression values shows substantial variation (i.e. coefficient of variation>1) across the 46 Roadmap cell types and labelled them as variably expressed TFs (VETFs) (**Table S2**). This classification is based on previous observations that expression of actively regulated genes is highly variable both temporally and spatially (Perez-Lluch et al., 2015). While these 634 VETFs encompass regulatory genes for a broad range of cell and tissue types, we used a subset of these TFs that were expressed (RPKM>1 or equivalent) in a given input transcriptome as a positive gene set specific for the input. We show the functional specificity of positive gene sets in different cellular contexts **(Figure 1B)** as well as the expressional specificity **(Figure 1C inset)** which is also quantified based on the Shannon's entropy. We used these 634 TFs as an effective proxy for cell type-specific regulatory genes.

**Quantifying cell type specificity of VETFs.** We use Shannon entropy to quantify the specificity of expression for VETFs, as observed across *m*=46 Roadmap cell types (Kundaje et al., 2015; Schug et al., 2005). The *relative* expression is calculated as

$p_{i,j} = w_{i,j} / \sum_{k=1}^{m} w_{i,k}$ where $w_{i,j}$ is the expression value of gene *i* in cell type *j*

and its cell type specificity is

$$H_i = -\sum_{k=1}^{m} p_{i,k} \cdot \log_2 p_{i,k}$$

The cell type specificity ranges between 0 (when the gene is uniquely expressed in a single cell type) and $\log_2 m$ (when the gene is expressed uniformly across all cell types). VETFs had significantly lower entropies (mean=3.64), compared to non-VETFs (mean=5.25, *p*=4.31e-230, Wilcoxon rank-sum test, one-tailed) and all protein-coding genes (mean=4.48, *p*=1.55e-108).

**Association of genes with a histone modification (HM) domain.** We used broad peak representation of consolidated histone modification files obtained from the Roadmap database. To assign peaks (referred to as domains hereafter) to genes, we used following steps.

1. Defining the proximal region of the gene. We defined a proximal region for each gene. The proximal region for a gene is the RefSeq gene body plus a region extended by 2.5kb from the TSS in the upstream direction.

2. Provisional assignment of domains to genes. For each gene, we first identified HM domains with their center position overlapping the proximal region. These domains were provisionally assigned to the corresponding gene. Domains that were broad, with their center position outside of the proximal region were still included if the domain overlapped with any proximal regions of genes, in which case, the domain was provisionally assigned to all overlapping genes. Suppose that gene *i* in cell type *j* have a set of provisionally assigned domains, $D_{i,j} = \{d_1, d_2, \dots, d_l, \dots\}$ where $d_l$ is the breadth (in base-pairs) of the *l*-th domain provisionally assigned to the gene.

3. Final assignment of domains to genes. If multiple domains were assigned to a gene *i* in cell type *j*, it is represented by the breadth of the broadest domain

$$b_{i,j} = max(D_{i,j})$$

We used BEDTOOLS intersect program to assign HM domains to RefSeq genes (Quinlan and Hall, 2010). We removed any intergenic domains that did not overlap any proximal regions of genes to reduce potential bias of including putative enhancers (Rada-Iglesias et al., 2011).

Our approach to annotate genes with breadth of the broadest H3K27me3 domain did not result in loss of relevant functional information. To illustrate this point, we calculated correlation between expression level of a gene and (i) the breadth of the broadest H3K27me3 domain or (ii) the sum of breadths of all H3K27me3 domains assigned to the gene. When we performed this analysis across all the cell types in Roadmap, we found no significant difference in the correlation between these two approaches, with the mean Spearman's *rho* of -0.364 and -0.337 respectively. Furthermore, the majority of domains finally assigned to protein-coding genes (i.e. approximately 85% of 1,537,514 assigned H3K27me3 domains across the 111 cell types) overlapped a single gene (**Figure S2C**). Approximately 66.9% of all H3K27me3 domains assigned to genes were identified in the RefSeq gene region while remaining 22.6% and 10.5% were in the promoter (+2.5kb from the TSS) and intergenic regions respectively, indicating that the majority (89.5%) of the assigned H3K27me3 domains were proximal to the gene.

For cell type $j$, we have a set of HM domain breadth values $B_j = \{b_{1,j}, b_{2,j}, \dots, b_{i,j}, \dots, b_{n,j}\}$ for $n$ genes. Subsequently, we normalized breadth values to yield the breadth score ($h$) for all genes across the cell types, $h_{i,j} = \frac{b_{i,j} - \overline{B_j}}{s_j}$ where $\overline{B_j}$ and $s_j$ are the sample mean and the standard deviation of HM domain breadths in cell type $j$.

**Genomic locations of H3K27me3 domain.** We investigated genomic locations of 1,537,514 H3K27me3 domains used for the RTS calculation. First, we identified a center position of each domain and used BEDTOOLS intersect program to identify if they overlap with (i) known RefSeq genes or (ii) promoters (defined as +2.5kb from the RefSeq TSS). If the domain does not overlap any of these, we labelled the domain as (iii) intergenic.

**Defining the broad H3K27me3 domain.** We observed that VETFs were strongly associated with broad H3K27me3 domains across various cell types. When genes were ranked by the breadth of their assigned H3K27me3 domain, VETFs were most significantly ($p$=6.66e-16, Fisher's exact test, one-tailed) enriched in the top 5% across all cell types (**Figure 1C**). Based on this observation, we defined the top 5% broadest domains for a given cell type as *nominally broad*. Let $X_{i,j}$ indicate whether gene *i* has a broad H3K27me3 domain in cell type *j*

$$X_{i,j} = \begin{cases} 1, if\ the\ assigned\ H3K27me3\ domain\ is\ in\ the\ broadest\ 5\% \\ 0, otherwise \end{cases}$$

**Similarity analysis on genes with broad HM domains.** To understand similarity of genes with broad HM domains across cell types, we first assigned genes to HM domains of 6 different types. For each HM type and each cell type, genes were ranked by the breadth of the assigned domain and grouped into bins of 100 genes. Jaccard similarity index was calculated at each rank bin between all pairs of cell types.

**Identification of genes frequently associated with broad HM domains.** To identify genes that are frequently associated with the broad HM domain across the Roadmap cell types, we counted the number of cell types where a given gene was associated with the broad HM. We ranked genes by the count and identified the top 200 genes for each HM type for downstream analysis.

**Estimating epigenetic repressive tendency of the gene.** We hypothesized that the regulatory importance of a gene in any cell type can be determined from evidence of (i) expression level of that gene in the same cell type, and (ii) breadth of H3K27me3 domains collected from a diverse range of cell types. To test this, we proposed a method that first quantifies the association of a gene with the H3K27me3 domain. For each gene, the method considers (i) a sum of H3K27me3 breadth scores for the gene calculated from *m* cell types

(e.g. 111 Roadmap cell types) and (ii) the number of cell types where the gene is associated with a broad H3K27me3 domain. For gene *i*, sum of the breadth scores ($v_i$) is defined as follows.

$$v_i = \sum_{k=1}^{m} h_{i,k}$$

The sum of breadth scores was then re-scaled into a range from 0 to 1 ($v_i{}'$) as follows, where $v_{max}$ and $v_{min}$ are maximum and minimum sums of the breadth scores from all genes respectively.

$$v_i{}' = \frac{v_i - v_{min}}{v_{max} - v_{min}}$$

The association score of gene *i* with the H3K27me3 domain ($a_i$) was then calculated as the product of the scaled sum of breadth scores ($v_i{}'$) and a proportion of cell types where the gene was associated with the broad H3K27me3 domain. To include genes without any broad H3K27me3 domains in any cell types, we added a pseudo-count of 1.

$$a_i = v_i{}' \cdot \sum_{k=1}^{m} \frac{X_{i,k} + 1}{m + 1}$$

Finally, we re-scaled the association score into a range of 0 (min) and 1 (max) to obtain the repressive tendency score (RTS) for the gene. For gene *i*, RTS is defined as follows, where $a_{max}$ and $a_{min}$ are maximum and minimum association scores for all genes, respectively.

$$R_i = \frac{a_i - a_{min}}{a_{max} - a_{min}}$$

**Performance analysis of TRIAGE against existing methods.** We extensively analyzed performance of TRIAGE against existing computational methods using various metrics. These include metrics based on (i) breadth of H3K4me3 peaks proximal to the gene (Benayoun et al., 2014), (ii) breadth of H3K27me3 peaks at the gene locus multiplied by the corresponding expression level in a spatially heterogenous cell population (Rehimi et al., 2016) and (iii) super-enhancer (SE) (Jiang et al., 2019; Whyte et al., 2013) as well as common practices of differentially expressed gene (DEG) analysis.

a. H3K4me3 breadth

Benayoun et al. postulated that broad deposition of H3K4me3 proximal to the gene is an indicator of cell identity gene by ensuring transcriptional consistency (Benayoun et al., 2014). As a HM of the gene activation, the study demonstrated efficacy of breadth of H3K4me3 peaks as a single determinant to identify cell identity genes across different tissue and cell types. A major limitation of this approach is requirement of input H3K4me3 ChIP-seq data, which can be especially problematic at the single-cell resolution (Clark et al., 2016). TRIAGE requires no input ChIP-seq data as it effectively utilizes epigenetic information collected from more than 100 diverse tissue and cell types represented by the Roadmap project (Kundaje et al., 2015).

To compare performance of TRIAGE against the H3K4me3 breadth-based metric, we used the 5 distinct tissue groups (See Datasets). H3K4me3 peaks were assigned to nearest RefSeq genes using an in-house Python script. Peaks located further than 2.5kb from any RefSeq TSSs were excluded. For genes with multiple assigned peaks, the broadest H3K4me3 peak was used to annotate the gene. Subsequently, genes were ranked by the breadth of the H3K4me3 peak.

b. Functional heterogeneity (FH) score

Previously, Rehimi et al. conceptualized that a broad H3K27me3 domain is often associated with regulatory genes to spatially restrict their expression in developing chicken embryo (Rehimi et al., 2016). This

observation led to a metric, called FH score, which integrates information from the gene expression and a breadth of H3K27me3 at the gene locus in a given cell state. While TRIAGE uses a similar quantification, the RTS captures a much more generalizable metric by interrogating H3K27me3 depositions from diverse tissue and cell types. This renders TRIAGE as a broadly applicable method to any tissue or cell states, without requiring input H3K27me3 data.

To test performance of TRIAGE on predicting developmental regulators against FH score, we used chicken embryo dataset for enriched H3K27me3 domains identified in Rehimi's study and the gene expression data (GSE89606) (Rehimi et al., 2016). These datasets satisfy the assumption of spatially heterogenous gene expression to calculate the FH score. We also downloaded genome annotation file (GFF3 format) for galGal4 chicken genome assembly (https://asia.ensembl.org/info/data/ftp/index.html) to quantify breadth of H3K27me3 peaks at the gene using BEDTOOLS intersect program (Quinlan and Hall, 2010). Furthermore, while the assumption of spatially heterogeneous gene expression is not strictly met, we also included FH score to evaluate its performance and applicability on the selected 5 distinct tissue groups (See Datasets).

Genes were ranked at both developmental time-points (HH14 and HH19) based on 3 metrics; (i) normalized gene expression value (TPM), (ii) FH score from Rehimi's study, and (iii) DS from TRIAGE. A total of 13,214 genes with an expression value (TPM>0) were included. We analyzed enrichment of 19 related GO BP terms used in Rehimi's study using Fisher's exact test (one-tailed) **(Figure S4D)**.

c. Super-enhancer (SE) based approach

Super-enhancers are defined as a cluster of enhancers characterized by exceptionally high level of Mediator signal or enhancer-associated HMs such as H3K27ac (Creyghton et al., 2010; Hnisz et al., 2013; Loven et al., 2013; Whyte et al., 2013). SEs have been studied as a genomic hot-spot to understand causal determinant of cell type-specific transcriptions in development and disease (Hnisz et al., 2013; Loven et al., 2013).

While accurately assigning target genes to SEs is not trivial, we aimed to test efficacy of TRIAGE to capture cell type-specific regulatory information compared to a simple but commonly used SE-based approach.

One way to functionally annotate genomic regions is by associating with the nearest gene (e.g. GREAT) (McLean et al., 2010). We downloaded lists of SEs for 5 selected tissue-types from a published SE database SEdb (http://www.licpathway.net/sedb/index.php) (Jiang et al., 2019); (i) Heart left ventricle, (ii) Lung_30y, (iii) Pancreas, (iv) Psoas muscle_30y, and (v) Brain astrocyte which were linked to Roadmap epigenomes E095 (Left ventricle), E096 (Lung), E098 (Pancreas), E100 (Psoas muscle) and ENCODE astrocyte (Brain) respectively. The algorithm to define SEs, ROSE (Rank-Order of Super-Enhancers) gives a binary outcome (i.e. SE or not) for each enhancer element (Whyte et al., 2013). As such, we extracted all nearest active genes of SEs and compared their functional enrichment against the same number of highly ranked genes by TRIAGE (Left ventricle ($n$=557), Lung ($n$=955), Pancreas ($n$=382), Psoas muscle ($n$=409) and Brain ($n$=689)) using Fisher's exact test (one-tailed).

d. Differentially expressed gene (DEG) analysis

DEG analysis is a commonly used method to identify cell type-specific genes by comparing expression values of genes between biologically different conditions (e.g. diseased vs. healthy tissues). Due to its comparative nature, success of the DEG analysis is highly dependent on careful selection of the control set. On the other hand, TRIAGE requires neither any prior knowledge on the reference point nor an arbitrary statistical threshold. We aimed to test performance of TRIAGE to identify genes governing cell fate against various approaches of the DEG analysis. To this end, we used published single-cell transcriptomes for *in vitro* cardiac-directed differentiation (Friedman et al., 2018) as well as selected bulk RNA-seq data for 3 distinct tissue groups (i.e. Blood, Brain and Heart) from the Roadmap project (Kundaje et al., 2015).

We obtained cardiac single-cell transcriptomes for differentiation days 2 and 30 from the ArrayExpress database (E-MTAB-6268). The data were processed and cells were clustered as previously described

(Friedman et al., 2018). For the Roadmap data, we extracted RNA-seq data for 3 representative samples for each tissue group (Blood; E037, E038, E047, Brain; E070, E071, E082 and Heart; E095, E104, E105). We calculated mean expression values of genes within each group or cell cluster. Subsequently, genes were ranked by fold change (FC) of the average expression value between different groups or clusters. We also tested performance of TRIAGE against DEG analysis with the input gene set restricted to only TFs. To this end, we first excluded all non-TF genes from the input data. While DEG analysis often focuses on differentially expressed TFs as a candidate regulatory gene set, TRIAGE offers an unsupervised approach to prioritize cell type-specific regulatory genes, without requiring any prior knowledge on the gene set.

e. H3K27me3 gain/loss function

Gain or loss of H3K27me3 dynamically regulates transcription of developmental genes in a cell type-specific fashion (Boyer et al., 2006; Mikkelsen et al., 2007). Therefore, any noticeable difference in the H3K27me3 deposition between distinct temporal points would indicate transcriptional change of the gene. Here we tested how TRIAGE performs against a simple gain/loss H3K27me3 function using a published dataset for induced *in vitro* differentiation of human cardiovascular cells between two different time-points; before differentiation (day0) and definitive cardiovascular cell stage (day14) (Paige et al., 2012).

To this end, we first obtained AffyExon microarray expression (GSE19090) and H3K27me3 ChIP-seq (GSE35583) data. We averaged probeset values to obtain the gene expression value and merged H3K27me3 peaks from replicates. To quantify H3K27me3 depositions, we mapped peaks to regions spanning from 2.5kb upstream of RefSeq TSSs to the entire gene body, which is represented by the number of overlapped base-pairs. We normalized depositions by the size of the region to get a value ranging between 0 (complete absence of H3K27me3) and 1 (completely covered by H3K27me3). Finally, change in the H3K27me3 deposition between the two time-points was calculated for all genes (i.e. $\Delta H3K27me3 =$

$H3K27me3_{day14} - H3K27me3_{day0}$). We ranked genes by loss of the H3K27me3 signal, DS and the expression values at day14.

**Statistical properties of the DS.** We defined a statistical test to gauge if a DS assigned to a gene is higher than expected. The expected DS of a gene is based on two components: the expected RTS given the observed level of expression, and the expected level of expression given the observed RTS of the gene. The test first estimates *p*-values for the two test statistics, which are then combined by Fisher's method to yield a single *p*-value based on the chi-squared distribution. In practical terms, an empirical null distribution is generated from repeated, random permutation of each test statistic of *n* comparable genes. Comparable genes are defined as those with the closest value to the gene of interest, in terms of the parameter that is *not* the test statistic (i.e. expression level or RTS). The *p*-value is then the probability that the rank of the gene of interest improves as a result of the random permutation.

**Functional enrichment analysis.** To compare efficacy of TRIAGE against other existing methods in prioritizing genes functionally important to a given tissue or cell state, we used a simple systematic approach to analyze enrichment of a selected GO term. For annotation purposes, we ranked only protein-coding genes. Ranked genes by a given metric (e.g. DS, gene expression value etc.) were first binned into a percentile bin (i.e. each bin includes 1% of the total gene populations in a given dataset). At each rank bin, we quantified enrichment of a selected GO biological process (BP) term using Fisher's exact test (one-tailed). Essentially, we tested how significantly the GO term was enriched among genes above a rank bin $x$ (i.e. above top $x$% of the gene population) compared to the rest using a sliding window. The resultant significance ($-log_{10}p$) was plotted against the rank bin, allowing visualization of how the enrichment changes as more lowly ranked genes were included

in the analysis. To identify significantly (FDR<0.05) enriched GO BP terms among top ranked genes, we used a R library topGO (Alexa, 2019).

**Consistency of the RTS between different peak callers.** To assess how the RTS changes with different peak calling methods, we independently calculated the RTS using peaks identified by 3 different peak callers, namely MACS2, SPP and Homer (Feng et al., 2012; Heinz et al., 2010; Kharchenko et al., 2008). Briefly, we first downloaded mapped ChIP-seq reads in tagAlign format for the 111 Roadmap cell types. Peaks were identified by comparing H3K27me3 tag signals with the input for each peak caller across the cell types. For MACS2 peak caller, 'callpeak' program was used with 'broad' and 'broad-cutoff' of 0.1 options to capture broad deposition of H3K27me3 (Feng et al., 2012). For SPP peak caller, 'get.broad.enrichment.clusters' function (available under 'spp' R library) with window.sizes=1000 and z.thr=3 was used as recommended for capturing broad HMs (Kharchenko et al., 2008). For Homer peak caller, 'findPeaks' program was used with '-style histone'. This parameter ensures the peak caller to initially find peaks of size 500 bp and subsequently stitch into regions with 1000bp; a suitable approach to identify broad regions of histone modifications (Kharchenko et al., 2008). From outputs from each peak caller, we calculated RTS values as described above. Subsequently, DSs were computed for 3 distinct Roadmap tissue groups; Left ventricle (E095), Germinal matrix (E070) and T helper naïve cells (E038), using these 3 different versions of the RTS. TFs with a selected GO term (Heart development GO:0007507, Brain development GO:0007420 and T cell differentiation GO:0030217 respectively) were used as the positive gene set for the enrichment analysis.

**Accuracy of estimated RTSs.** We estimated the RTS from 111 human cell types in Roadmap. To understand the presence of sampling bias, we performed a bootstrapping analysis by randomly re-sampling cell types 10,000 times. For each re-sampling, we calculated the RTS for each gene. We collected the

empirical bootstrap distribution of RTSs for each gene. Estimated RTSs of all 26,833 genes were within 1 standard deviation of their respective mean, which supports consistency of the estimated RTS.

**Saturation of H3K27me3 signals.** To understand whether the 111 cell types in Roadmap provide sufficient data to estimate stable RTSs, we developed an iterative process to quantify stability of the RTS with a differing number of cell type samples. Suppose we have *n* number of genes each of which has a RTS calculated based on H3K27me3 data observed in *k* number of cell types. We defined *saturation state* as a state where any change in the RTS for a given gene as a result of an addition of *l* number of cell types is within an arbitrarily defined range.

If the signal is in the saturation state, adding *l* number of different cell types would result in insignificant change to the resultant RTS. To help quantification of the RTS change, we define a term *stably ranked gene*. Suppose gene *i* has an estimated RTS derived from *k* number of cell types and is ranked at a certain position ($u_{i,k}$). We say the gene is *stably ranked* if a resultant RTS re-calculated with an addition of *l* number of cell types put the gene at a rank position ($u_{i,k+l}$) that is within a certain range of rank positions ($\theta$) from the previous rank position ($u_{i,k}$). In other words, gene *i* is stably ranked if the resultant RTS change is not large enough to shift its rank position more than $\theta$. Formally, a set of *stably ranked genes* with an addition of *l* number of cell types to *k* number of cell types ($G_{k,l}$) is a subset of all genes ($G_{all}$) and can be written as the follow.

$$G_{k,l} = \{i \in G_{all} \mid |u_{i,k+l} - u_{i,k}| < \theta\}$$

For instance, if gene A is ranked at *q* position by the RTS calculated using *k* number of cell types, the gene A is a *stably ranked gene* if the rank does not change more than $\theta$ positions (e.g. 1% of the total gene number) from *q* when an additional *l* number of cell types is included for the subsequent calculation.

We started with *k*=3 randomly selected cell types and iteratively calculated the proportion of *stably ranked genes* at an increment of *l*=3 randomly selected cell types without replacement until all 111 cell types were used. To address potential sampling bias, we repeated this process 1,000 times and obtained mean values and the 95% confidence interval of the estimation (**Figure S2E-F**). We used a range of different thresholds (i.e. 1~5% of the total gene number) for the stably ranked gene.

**Correlation between the expression and the H3K27me3 domain breadth.** H3K27me3 represses transcription of the target gene(Barski et al., 2007) but it is not known whether cell type-specific regulatory genes have a distinct functional relationship with this repressive HM. Based on strong association of cell type-specific TFs with broad H3K27me3 domains, we questioned whether the repressive effect of H3K27me3 was more prominent among the cell type-specific TFs. To this end, we first identified five classes of genes (i.e. (i) 634 VETFs to represent cell type-specific regulatory genes, (ii) 7,445 variably expressed non-TFs, (iii) 18,708 protein-coding genes, (iv) 793 non-VETFs and (v) 3,805 housekeeping genes). For each gene, we calculated a Pearsons's correlation coefficient between the gene expression value and the breadth of the corresponding H3K27me3 domain across the 46 Roadmap cell types.

**Functional relationship between the repressive tendency and gene transcription.** To understand how the RTS is associated with the transcriptional outcome, we first ranked genes in a descending order of the RTS. We only considered coding genes and assigned them into a bin of 100 genes. For the gene expression level, an average expression value of the gene set at each rank bin was calculated across the 46 cell types **(Figure 2D).** For the expressional specificity, a proportion of cell types where a given gene was expressed (RPKM>1 or equivalent) out of the 46 cell types was calculated **(Figure 2C).** As there were 100 genes in each rank bin, we calculated the average proportion of the 100 genes in each bin.

**Comparison of clustering accuracy.** We compared clustering accuracy based on expression values and DSs using the scRNA-seq mixology benchmarking data set (Tian et al., 2019). The data set we used (*sc_10x*) was generated on the 10X platform and provided expression values as counts for each of three distinct, labelled, human lung adenocarcinoma cell lines. We assessed the clustering of this data across three methods: *SC3* (1.10.0), *CORE* (ascend 0.9.6) and *Seurat* (2.3.0) (Butler et al., 2018; Kiselev et al., 2017; Senabouth et al., 2019). Each of the clustering methods were used as detailed by their authors in their documentation or tutorials, including any filtering and scaling steps. In the SC3 clustering, the k parameter (ks) was set to 3, and the remainder of the parameters across all three algorithms were chosen as specified in their documentation or left at default.

To quantify the performance of the clustering methods, we used the Adjusted Rand Index as calculated by the *mclust* package (5.4.2) (mclust::adjustedRandIndex) (Scrucca et al., 2016), comparing the cluster assignment in each clustering method to the three cell line labels from the mixology data set. The PCA plots in **Figure 3** were generated using the ascend package (ascend::plotPCA) (Senabouth et al., 2019)

**Application of TRIAGE to multi-omics platforms.** We hypothesized that TRIAGE would be applicable to any quantifiable genomic data that reasonably reflects the expression level of genes. To demonstrate applicability of TRIAGE to various multi-omics platforms, we incorporated single-cell transcriptomes (Cao et al., 2019; Friedman et al., 2018; Schaum et al., 2018), FANTOM5 cap analysis of gene expression (CAGE) peaks (Forrest et al., 2014), human proteomes (Kim et al., 2014) and Roadmap H3K36me3 ChIP-seq data sets (Kundaje et al., 2015).

We applied TRIAGE to tabula muris mouse single-cell RNA-seq data sets for nearly 100,000 cells from 20 different tissue types (Schaum et al., 2018). We averaged expression values of genes for each tissue type to calculate corresponding DSs. We also utilized the MOCA data set representing approximately 1.3 million mouse single-cell transcriptomes capturing a broad spectrum of organ development or organogenesis (Cao et al., 2019). For each transcriptome sample, we first ranked genes using either the expression value or the

DS. We then analyzed enrichment of VETFs across rank positions. We identified the best *p*-value (i.e. the lowest *p*-value) and the corresponding rank position for each sample.

For the CAGE data set, we used the normalized CAGE tag density for the expression value of the corresponding gene. The highest CAGE tag density assigned to the gene was used so that each gene was annotated with a value from a TSS with the highest value. We selected 329 FANTOM5 samples that covers 25 distinct cell types without a disease annotation **(Table S1).** For each sample, we applied TRIAGE to calculate the DS of genes. Genes were then ranked by either the expression value or the DS and the enrichment of VETFs was analyzed at each rank position using Fisher's exact test. Peak significance value (defined as the highest $-\log_{10}p$ value) and the corresponding rank position were identified for each sample.

The human proteomic data provides expression levels of genes, peptides and proteins across 30 different tissue groups. We used the protein expression value to link to the corresponding gene for the quantification. Again, TRIAGE was applied to the expression data to convert them into the DS. We identified peak significance value and the corresponding rank position for each tissue group.

For H3K36me3 data set, we collected mapped reads for H3K36me3 ChIP-seq for the 111 Roadmap cell types. Earlier findings showed that the gene expression is positively correlated with the H3K36me3 tag density over the transcribed gene body regions (Barski et al., 2007). For each cell type, we quantified the read density for each gene by calculating the number of reads per base-pair mapped to the RefSeq gene body. We then used the tag density as a proxy for the transcriptional abundance of the gene to calculate the DS.

**Visualization of multi-omics datasets.** To visualize performance of TRIAGE on multiple samples on a single plot simultaneously, we summarized the performance for each sample by (i) the most significant *p*-value (*y*-axis, Fisher's exact test for enrichment of VETFs, one-sided) and (ii) the corresponding rank

position where (i) is observed (*x*-axis, rank ranging from 1 the highest to 100 the lowest). Each sample was represented as a single point on the plot.

**Inter-species application of the TRIAGE.** The RTS for the gene was derived from evidence of the association with H3K27me3 domains across the 111 Roadmap human cell types. Given a high level of evolutionary conservation for the PRC2 (Margueron and Reinberg, 2011), we hypothesized that the RTS calculated from the human data could be effectively applied to equivalent genes of other species. To test this, we first downloaded a range of transcriptomic data sets from different species covering the 5 selected tissue-groups (**Table S1**). We then performed inter-species gene mapping using online Ensembl bioMart (http://asia.ensembl.org/biomart/martview/) by identifying human orthologues(Haider et al., 2009). Only genes mappable to human equivalent genes were included in the analysis.

**Melanoma gene set enrichment.** We used published pre-processed single cell RNA-seq data from melanoma tumors (Tirosh et al., 2016).1,252 melanoma cells were isolated from the set of approximately 4,000 cells based on the authors annotations. Melanoma proliferative and invasive gene sets were obtained from the same source. Based on the ratio of the average gene expression of the proliferative to invasive genes; the top 50 most proliferative and the top 50 most invasive cells were identified. TRIAGE was applied to the melanoma gene expression profiles to produce DS profiles. Averages of both expression and the discordance profiles of the top 50 proliferative and invasive cells were taken. Genes were subsequently ranked by the expression and discordance values for these representative profiles. For the ranked genes, fishers exact test was iteratively performed from the top ranked genes down the list as described above; adding 1% of the genes at each iteration. Gene sets tested for enrichment were obtained from published melanoma data sets (Tirosh et al., 2016). False discovery rate was used to correct for multiple hypothesis testing.

**Heart Failure pathogenesis dataset**. To determine the utility of TRIAGE in identifying regulatory elements and processes of disease in heart failure pathogenesis we used published pre-processed bulk RNA-seq data from adult mouse ventricles (GSE58453 and GSE68509) (Duan et al., 2017). Briefly, heart failure was induced using transverse aortic constriction (TAC) and the small molecule pan-BET inhibitor JQ1 was used to treat the TAC condition. TRIAGE was applied to the gene set of each condition (SHAM, TAC, TAC+JQ1) to produce DS profiles. Genes were subsequently ranked by the expression and discordance values for these representative profiles. For the ranked genes, Fishers exact test was iteratively performed from the top ranked genes down the list for GO terms of interest.

**Gene Ontology Visualization.** Gene ontology analysis was performed using DAVID with significance threshold set at FDR<0.05. The *p*-values from gene ontology analysis were visualized with the radius of the circle proportional to the negative natural log of the input *p*-value.

**iTranscriptome Sample Preparation and Data Analysis.** Samples were generated according to the methodology previously published(Peng et al., 2016). E5.5, E6.0, E6.5, E7.0 and E7.5 embryos were cryo-sectioned along the proximal-distal axis. Populations of approximately 20 cells were collected from different regions of the cross-section by laser microdissection and processed for RNA sequencing (see **Figure S7D**). From the RNA-sequencing dataset, differentially expressed genes (DEGs) were screened first by unsupervised hierarchical clustering method to group samples in the respective regions. Genes with an expression level FPKM>1 and a variance in transcript level across all samples greater than 0.05 were selected. To identify inter-region specific DEGs, each of these selected genes was submitted to a t-test against the level of expression in the other regions. Genes with a *p*-value<0.01 and a fold change>2.0 or <0.5 were defined as DEGs. The gene expression pattern (region and level of expression by transcript reads) of the gene of interest was mapped on the corn plots, where each kernel represents the cell population

sampled at a defined position in the tissue layers of the embryo, to generate a digital 2D rendition of the expression domain that emulated the display of the result of whole mount *in situ* hybridization.

**Generation and Maintenance of Human ESC/iPSC Lines.** All human pluripotent stem cell studies were carried out in accordance with consent from the University of Queensland's Institutional Human Research Ethics approval (HREC#: 2015001434). WTC CRISPRi GCaMP hiPSCs (Karyotype: 46, XY; RRID CVCL_VM38) were generated using a previously described protocol(Mandegar et al., 2016) and were generously provided by M. Mandegar and B. Conklin (UCSF, Gladstone Institute). WTC CRISPRi SIX3-g2 hiPSCs were generated in this study (see below). All cells were maintained as previously described(Palpant et al., 2017a). Briefly, cells were maintained in mTeSR media with supplement (Stem Cell Technologies, Cat.#05850) at 37º C with 5% $CO_2$. WTC CRISPRi GCaMP and WTC CRISPRi SIX3-g2 hiPSC lines were maintained on Vitronectin XF (Stem Cell Technologies, Cat.#07180) coated plates.

**WTC CRISPRi SIX3-g2 hiPSCs**. 3 separate guide RNAs (gRNA) targeting the CAGE-defined transcriptional start sites of the human SIX3 sequence were designed and cloned into the pQM-u6g-CNKB doxycycline-inducible construct and transfected into WTC CRISPRi GCaMP hiPSCs using GeneJuice Transfection Reagent (Merck, Cat.#70967). Stable clones were selected using successive rounds of re-plating with blasticidine at 10µg/ml (Sigma, Cat.#15205). Populations were tested for knockdown efficiency by qPCR following doxycycline addition at 1 µg/ml (Sigma, Cat.#D9891) continuously from day 0 of cardiac-directed differentiation. WTC CRISPRi SIX3-g2 line displayed high knockdown efficiency and therefore was chosen.

*Guide RNAs designed:*

| **gRNA Name** | **Oligo Sequences**<br>**5' – Forward Primer – 3'**<br>**5' – Reverse Primer – 3'** |
|---|---|
| | F 5' TTGGGCTGAATCTTGACTCGGCGG 3' |

| | |
|---|---|
| SIX3 gRNA1 | R 5’ AAACCCGCCGAGTCAAGATTCAGC 3’ |
| SIX3 gRNA2 | F: 5' TTGGTGTCATTAGGGCGATTGCGG 3' |
| | R: 5' AAACCCGCAATCGCCCTAATGACA 3' |
| SIX3 gRNA3 | F 5’ TTGGCTCTATGTGGCTGGCGGGTG 3’ |
| | R 5’ AAACCACCCGCCAGCCACATAGAG 3’ |

**Cell Culture.** All human pluripotent stem cell studies were carried out in accordance with consent from the University of Queensland’s Institutional Human Research Ethics approval (HREC#: 2015001434). hiPSCs were maintained in mTeSR media (Stem Cell Technologies, Cat.#05850). Unless otherwise specified, cardiomyocyte directed differentiation using a monolayer platform was performed with a modified protocol based on previous reports(Burridge et al., 2014). On day -1 of differentiation, hPSCs were dissociated using 0.5% EDTA, plated into vitronectin coated plates at a density of 1.8 x $10^5$ cells/$cm^2$, and cultured overnight in mTeSR media. Differentiation was induced on day 0 by first washing with PBS, then changing the culture media to RPMI (ThermoFisher, Cat.#11875119) containing 3µM CHIR99021 (Stem Cell Technologies, Cat.#72054), 500µg/mL BSA (Sigma Aldrich, Cat.#A9418), and 213µg/mL ascorbic acid (Sigma Aldrich, Cat.#A8960). After 3 days of culture, the media was replaced with RPMI containing 500µg/mL BSA, 213µg/mL ascorbic acid, and 5µM Xav-939 (Stem Cell Technologies, Cat.#72674). On day 5, the media was exchanged for RPMI containing 500µg/mL BSA, and 213µg/mL ascorbic acid without supplemental cytokines. From day 7 onwards, the cultures were fed every 2 days with RPMI plus 1x B27 supplement plus insulin (Life Technologies Australia, Cat.#17504001).

**Quantitative RT-PCR.** For quantitative RT-PCR, total RNA was isolated using the RNeasy Mini kit (Qiagen, Cat.#74106). First-strand cDNA synthesis was generated using the Superscript III First Strand Synthesis System (ThermoFisher, Cat.#18080051). Quantitative RT-PCR was performed using SYBR Green PCR Master Mix (ThermoFisher, Cat.#4312704) on a ViiA 7 Real-Time PCR System (Applied

Biosystems). The copy number for each transcript is expressed relative to that of housekeeping gene *HPRT1*. Samples were run in biological triplicate. FC was calculated on a gene by gene basis as gene expression divided by control gene expression. The following are qRT-PCR primers utilized in this study:

| **Gene Name** | **5'-Forward Primer-3'** | **5'-Reverse Primer-3'** |
|---|---|---|
| *HPRT* | TGACACTGGCAAAACAATGCA | GGTCCTTTTCACCAGCAAGCT |
| *SIX3* | GCAGAAGACGCATTGCTTCAA | CCCAGCAAGAAACGCGAAC |
| *SOX2* | TGGACAGTTACGCGCACAT | CGAGTAGGACATGCTGTAGGT |
| *HHEX* | AATGCTGGATGATGACCACT | TAATTGAGCAGTGCACCAAA |
| *GATA6* | TGCAATGCTTGTGGACTCTA | GTGGGGGAAGTATTTTTGCT |
| *SOX17* | ACGCCGAGTTGAGCAAGA | TCTGCCTCCTCCACGAAG |
| *FOXA2* | TGCACTCGGCTTCCAGTATG | CATGTTGCTCACGGAGGAGT |
| *NODAL* | TGGAGGTGGGATGAAGTCACCTAT | AACCCAGCCTGAGGCAATGAGATT |
| *GSC* | GAGGAGAAAGTGGAGGTCTGGTT | CTCTGATGAGGACCGCTTCTG |
| *EOMES* | CACATTGTAGTGGGCAGTGG | CGCCACCAAACTGAGATGAT |
| *SOX7* | TGACAACTTGTTGCCAACTCCCTG | TTCAGCAGTGGAGGAAGAGCAGAA |
| *GATA4* | GACCTGGGACTTGGAGGATA | ACAGGAGAGATGCAGTGTGC |
| *DKK1* | AACAGCTATCCAAATGCAG | TCACAGGGGAGTTCCATAAA |
| *MEST* | CTGTGGGTGTGGTTGGAAGT | TGTCACTGAAGCCAAAGCCT |
| *T (Bry)* | GTCAGAATAGGTTGGAGAATTG | CAAATCCTCATCCTCAGTTTG |
| *MESP1* | TCGAAGTGGTTCCTTGGCAGAC | CCTCCTGCTTGCCTCAAAGTGTC |
| *WNT8A* | GCAGAGGCGGAACTGATCTT | CGACCCTCTGTGCCATAGATG |
| *WNT3A* | AACTACGTGGAGATCATGCCC | GACTCCCTGGTAGCTTTGTC |

**Flow Cytometry.** On day 2 of cardiac differentiation cells were dissociated using 0.5% EDTA and put into blocking buffer of 50% fetal bovine serum (FBS) in Dulbecco's Modified Eagle Medium (DMEM)/F12 (ThermoFisher, Cat.#11320033). Cells were then pelleted and resuspended in 10% FBS in DMEM media. Cells were labeled live for flow cytometry using CD184 (Becton Dickinson, Cat.#555974), EpCAM (Becton Dickinson, Cat.#347199) and corresponding isotype controls were used to gate the cells. On day 15 of cardiac differentiation cells were fixed with 4% paraformaldehyde (Sigma, Cat.#158127) and permeabilized in 0.75% saponin (Sigma, Cat.#S7900). On day 15 of cardiac differentiation fixed cells were labeled for flow cytometry using alpha-actinin (Miltenyi Biotec, Cat.#130106937) and corresponding isotype control. Cells were analyzed using a BD FACSCANTO II (Becton Dickinson, San Jose, CA) with FACSDiva software (BD Biosciences). Data analysis was performed using FlowJo (Tree Star, Ashland, Oregon).

***Ciona robusta* CRISPR/Cas9 gene editing** For Rnf220 (KH2012:KH.C8.791) loss of function, 3 sgRNAs were designed to avoid genomic off-targets and tested as described(Stolfi et al., 2014). sgRNA expressing cassettes (U6 > sgRNA) were assembled by single step overlap PCR. Individual PCR products (~25 µg) were electroporated with EF1a > NLS::Cas9::NLS (20 µg), Myod905 > Venus (50 µg), driving ubiquitous expression of Cas9 and a widely expressed fluorescent reporter construct, respectively, as described(Christiaen et al., 2009). Efficient electroporation was confirmed by observation of fluorescence before genomic DNA extraction around 16 hpf (18°C) using QIAamp DNA Micro kit (Qiagen, German Town, MD). Mutagenesis efficacy of individual sgRNAs, as a linear function of Cas9-induced indel frequency, was estimated from electrophoregrams following Sanger sequencing of the targeted regions amplified from extracted genomic DNA by PCR. Results of the relative quantification of the indel frequency ('corrected peakshift' of 22% and 16%) for sgRNAs 2 and 3 was considered high enough for both sgRNAs targeting Rnf220, which were finally selected. The corresponding cassettes were cloned into plasmid for repeated electroporations to study the loss of function of Rnf220. In order to control the specificity of the CRISPR/Cas9 system, sgRNAs targeting *Tyrosinase*, a gene not expressed in the

cardiopharyngeal lineage, was electroporated in parallel. For imaging experiments, sgRNAs (25 µg) were electroporated with Mesp > NLS::Cas9::NLS (20 µg), Mesp > H2B:GFP (50 µg) and Mesp > mCherry (50 µg) . Sequences of the DNA targets and oligonucleotides used for the sgRNAs:

| **sgRNA name** | **Universal sgRNA name** | **Protospacer + PAM Sequence** | **Doench 16' score** |
|---|---|---|---|
| RNF220_sg1 | RNF220_p.A | GCGATGAACGGATGCGCTGG CGG | 64 |
| RNF220_sg2 | RNF220.e1.A | GGGTCGGGTTGATTGCACTT GGG | 63 |
| RNF220_sg3 | RNF220.e1.B | CCCCCACCAGACTTCAGCAG CGG | 65 |
| TyrC | sgTYR_e5.B | TCGATACTACCTGCTTAAGT GGG | 54 |

| **sgRNA name** | **OSO Primer Forward** |
|---|---|
| RNF220_sg1 | gCGATGAACGGATGCGCTGGgtttaagagctatgctggaaacag |
| RNF220_sg2 | gGGTCGGGTTGATTGCACTTgtttaagagctatgctggaaacag |
| RNF220_sg3 | gCCCCACCAGACTTCAGCAGgtttaagagctatgctggaaacag |
| TyrC | gCGATACTACCTGCTTAAGTgtttaagagctatgctggaaacag |
| | **OSO Primer Reverse** |
| RNF220_sg1 | CCAGCGCATCCGTTCATCGcatctataccatcggatgccttc |
| RNF220_sg2 | AAGTGCAATCAACCCGACCcatctataccatcggatgccttc |
| RNF220_sg3 | CTGCTGAAGTCTGGTGGGGcatctataccatcggatgccttc |
| TyrC | ACTTAAGCAGGTAGTATCGcatctataccatcggatgccttc |

Embryos were fixed in 4% MEM-FA for 30 minutes, incubated with an $NH_4Cl$ solution, and imaged using Leica SP8 X Confocal microscope.

**Code Availability statement.** We provide the source code written in Python and R (https://github.com/woojunshim/TRIAGE). Users can also run the TRIAGE analysis using a web accessible interface (http://bioinf.scmb.uq.edu.au/adhoc/).

**Quantification and statistical analysis.** Unless otherwise noted, all data are represented as mean ± standard error of mean (SEM). Indicated sample sizes (n) represent biological replicates including independent cell culture replicates and individual tissue samples. No methods were used to determine whether data met assumptions of the statistical approach or not. Due to the nature of the experiments, randomization was not performed and the investigators were not blinded. Statistical significance was determined in GraphPad Prism 7 software by using student's t test (unpaired, two-tailed) or ordinary one-way ANOVA unless otherwise noted. Results were considered to be significant at $p < 0.01$(*). Statistical parameters are reported in the respective figures and figure legends. All statistical data are represented as mean ± SEM.

## ACKNOWLEDGEMENTS

E.S acknowledges funding by Children's Hospital Foundation Queensland (Award Reference Number: 50268). B.V. acknowledges funding by American Heart Association grant #18PRE33990254. The *Ciona* work was supported by NIH/NHLBI award R01 HL108643 to L.C. M.A. was supported by the Swiss National Science Foundation (project P2LAP3_178056), P.P.L.T. is supported by the National Health and Medical Research Council of Australia (Grant 1110751). N.P is supported by the National Health and Medical Research Council of Australia (Grant 1143163) and the Australian Research Council (Grant SR1101002).

## AUTHOR CONTRIBUTIONS

**WJS**: Developed the computational basis for the study, performed data analysis and wrote the manuscript

**ES**: Contributed to experimental and computational design for the study, performed data analysis, carried out functional genetic studies in hPSCs and wrote the manuscript

**JX**: Assisted with computational analysis and developed web interactive interface

**MA**: Performed computational analysis on HF pathogenesis data

**GA**: Performed computational analysis on HF pathogenesis data

**SS:** Assisted the computational analysis on different single-cell data platforms

**BB**: Performed computational analysis on melanoma studies

**YS**: Performed computational analysis on Mouse Organogenesis Cell Atlas data

**BV**: Performed functional analysis on *ciona* and validated the findings

**GP**: Assisted with spatiotemporal transcriptomic profiling of mouse gastrulation

**NJ**: Assisted with spatiotemporal transcriptomic profiling of mouse gastrulation

**YW:** Helped with computational analysis of epigenetic data

**MP**: Assisted with analysis and interpretation of melanoma data

**AS:** Carried out experiments involving melanoma analysis

**PT**: Supervised work on spatiotemporal transcriptomic profiling of mouse gastrulation

**LC**: Performed functional analysis on *ciona* and validated the findings

**QN:** Provided assistance to implement TRIAGE on single-cell data sets

**MB** and **NJP**: Supervised the project, raised funding, and wrote the manuscript

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## FIGURE LEGENDS

**Figure 1: Broad H3K27me3 domains are associated with cell type-specific regulatory genes.**

**(A)** VETFs are strongly associated with broad H3K27me3 domains. Genes are ranked by the breadth of the associated HM domain across 111 Roadmap cell types and grouped into percentile bins (*x*-axis). Mean enrichment of VETFs for all pair-wise cell types with the 95% confidence interval is shown (*y*-axis) ($p$=6.66e-16 for top 5% broadest H3K27me3 domains, Fisher's exact test, one-tailed). **(Inset)** Expression levels of the 634 VETFs (row) across the 46 Roadmap samples (column).

**(B)** Functional enrichment of top 200 genes most frequently associated with top 5% broadest HM domains across the 111 Roadmap cell types (Fisher's exact test, one-tailed).

**(C)** Gene expression level (left) and the H3K27me3 breadth (right) for cardiac-specific regulatory genes and structural genes across 18 Roadmap samples; Heart (E095, E104, E105), Brain (E070, E071, E082), Epithelial (E057, E058, E059), Blood (E037, E038, E047), ES cell (E003, E016, E024) and ES-deriv. (E004, E005, E006).

**(D)** HM domains in the proximal region (-2.5kb upstream of the RefSeq TSS to +25kb downstream) of a cardiac regulatory gene *NKX2-5* and a structural gene *MYH7*, collected from the 111 Roadmap cell types.

**Figure 2: Genes associated with broad H3K27me3 domains regulate developmental and disease processes.**

**(A)** Schematic diagram showing steps involved in calculation of the repressive tendency score (RTS).

**(B) (Top)** Distribution of RTS values. Red dashed line is the inflection point on the interpolated curve (RTS=0.03022) above which genes exhibit a substantially higher RTS than the rest ($n$=1,359, priority genes). **(Bottom)** Functional enrichment of gene sets sorted by the RTS (Fisher's exact test, one-tailed) (Table S4 for the full list).

**(C)** Expressional specificity of protein-coding genes sorted by the RTS. Each rank bin includes 100 genes. Average proportion of cell types where the gene is expressed (RPKM>1) is calculated for each rank bin (*y*-axis). Shades mark the 95% confidence interval.

**(D)** Gene expression level and the RTS. Each rank bin includes 100 genes. Average expression value for each bin, with 95% confidence interval (shade) is shown (*y*-axis).

**(E)** VETFs are significantly associated with a high RTS ($p$<2.2e-16 at the first rank position, Fisher's exact test, one-tailed). Each rank bin includes 1% of all RefSeq genes with an RTS ($n$=26,833). Red dashed line represents a uniform distribution (proportion=0.01)

**(F)** Genes with a high RTS are enriched with various development and disease processes as well as signaling pathways (Table S4 for the full list). Each rank bin includes 1% of all RefSeq genes with an RTS ($n$=26,833). Asterisks mark significant enrichment of a given term among top 10% genes (Benjamini-Hochberg FDR<0.05, Fisher's exact test, one-tailed).

**Figure 3: TRIAGE infers cell type-specific regulatory genes of somatic cell types across diverse species.**

**(A)** TRIAGE transformation of gene expression input data to generate a discordance score (DS) enabling prioritization of genes based on regulatory potential.

**(B)** Breadths of H3K27me3 domains (in base-pairs) associated with selected tissue type-specific regulatory, structural and housekeeping genes.

**(C)** TRIAGE prioritizes tissue-type specific regulatory genes. Average expression values (Exp.) from GTEx samples for each tissue type is transformed into the DS (Dis.).

**(D)** TRIAGE shifts distributions of VETFs and housekeeping genes from transcriptomes of the 46 Roadmap cell types. Each rank bin includes 1% of all expressed genes in a cell type. VETFs are significantly associated with a high DS (top 1%) in all cell types ($p$<2.2e-16 for all cell types, Fisher's exact test, one-tailed) while proportions of housekeeping genes are substantially reduced ($p$<2.2e-16 for top 1% genes, Wilcoxon rank-sum test, one-tailed). Scale bars show the 95% confidence interval.

**(E)** Enrichment of TFs annotated with 'heart development' term (GO:0007507) in cardiac samples across different animal species (Fisher's exact test, one-tailed). Genes are sorted in a descending order by the DS (red) or the expression value (cyan) and grouped into a percentile bin. Each rank bin includes 1% of all expressed (RPKM>1 or equivalent) genes in a sample.

**(F)** Enrichment of TFs with a tissue-type-specific development GO term among top 100 genes from the expression value (blue) or the DS (red) across different animal species (Fisher's exact test, one-sided). Hyphen (-) indicates no data set available.

**Figure 4: TRIAGE performance against benchmark epigenetic inference methods.**

**(A)** Schematic diagrams of computational approaches to infer cell identity genes utilizing epigenomic and/or transcriptomic data with TRIAGE, FH Score, H3K4me3, or super-enhancers.

**(B)** Venn diagrams showing numbers of overlapping or uniquely identified genes by different methods. Cardiac transcription factors are shown for heart demonstrating unique TFs detected in each method.

**(C)** Receiver operating characteristic (ROC) plots comparing performance of different approaches to recover key regulatory genes. Area under curve (AUC) values are shown on the right bottom corner of the plot.

**(D)** Biological processes enriched in genes prioritized by TRIAGE versus super-enhancers.

**Figure 5: TRIAGE is scalable for multi-omics datasets.**

**(A)** Top 10 genes identified by expression or TRIAGE from hiPSC derived cardiomyocytes (Friedman et al., 2018).

**(B)** scRNA-seq analysis of *in vitro* differentiated cardiomyocytes at day 30 showing rank order based on original expression (left) versus the same rank order with TRIAGE transformation (right) showing changes in relative abundance for each gene. TRIAGE reveals known cardiac regulators lowly ranked based on original expression.

**(C)** TRIAGE-transformation shifts distributions of different gene categories including: 1. housekeeping, 2. heart signaling (genes with heart development term GO:0007507 and at least one KEGG signaling pathway

term), 3. sarcomere (genes with sarcomere term GO:0030017) or 4. heart TFs (TFs with heart development term GO:0007507). Genes are sorted by either the expression value (left) or TRIAGE (right) and grouped into a percentile bin. Each rank bin includes 1% of all expressed genes in the dataset. Enrichment of a given gene set is calculated at each rank bin (Fisher's exact test, one-tailed).

**(D)** UMAP representation of cell clustering using transcriptomic expression (left) or TRIAGE (right).

**(E)** Enrichment of developmental GO BP terms primarily associated with stage-specific regulatory developmental processes during *in vitro* cardiac-directed differentiation (*y*-axis) (Fisher's exact test, one-tailed) that are consistently and uniquely identified by TRIAGE. In contrast, expression data strongly detect structural and housekeeping genes.

**(F)** Enrichment of developmental GO terms among genes ranked by TRIAGE (red), expression (blue) or fold-change of gene expression (green or purple) in sub-cell populations (i.e. cluster) from *in vitro* cardiac-directed differentiation datasets (days 2 and 30). Each rank bin includes 1% of all genes.

**(G, H)** TRIAGE is applicable to **(G)** different scRNA-seq sequencing platforms (i.e. Smart-seq2 or Droplet 10X chromium), **(H)** and scalable for analysis of millions of cells as demonstrated using the mouse organogenesis cell atlas (MOCA). Genes are sorted by either expression value or TRIAGE and grouped into a percentile bin. Enrichment of VETFs for each sample is summarized by the most significant *p*-value (*y*-axis) at the corresponding rank bin position (*x*-axis).

**(I)** TRIAGE accurately clusters scRNA-seq data compared to original expression data based on ground truth analysis captured using Mixology data sets (Tian et al., 2019).

**(J)** TRAIGE analysis of diverse omics data analysis representing quantitative abundance of gene expression including CAGE-seq, proteomics, and H3K36me3 tag density. Analysis is as described in panels G-H.

**Figure 6: TRIAGE as an engine for novel gene discovery.**

**(A)** Top 10 genes ranked by expression value (left) or TRIAGE (right) from day 2 populations (germ layer specification) of hiPSC cardiac differentiation, highlighting *SIX3* as a candidate novel gene identified by TRIAGE.

**(B)** Breadths of H3K27me3 domains (in base-pairs) associated with *SIX3* gene across the 111 NIH Epigenome data sets.

**(C)** Analysis of *SIX3* expression during hiPSC-cardiac differentiation represented by t-SNE plots (left), percentage of cells expressing *SIX3* (top right) and gene expression level of *SIX3* (bottom right) in each subpopulation on days 0, 2 and 5.

**(D)** Corn plots showing the spatial domain of *SIX3* expression in the germ layers of E5.5-E7.5 mouse embryos. Positions of the cell populations (‘‘kernels’’ in the 2D plot of RNA-seq data) in the embryo: the proximal-distal location in descending numerical order (1 = most distal site) and in the transverse plane of the germ layers: endoderm, anterior half (EA) and posterior half (EP); mesoderm, anterior half (MA) and posterior half (MP); epiblast/ectoderm, anterior (A), posterior (P) containing the primitive streak, right (R)-anterior (R1) and posterior (R2), left (L) – anterior (L1) and posterior (L2).

**(E)** Schematic overview of *SIX3* gene targeting showing position of gRNAs blocking CAGE-defined TSS to achieve conditional KD of *SIX3*.

**(F)** qPCR analysis of *SIX3* transcript abundance in control vs *SIX3* CRISPRi KD iPSCs.

**(G)** Schematic of *in vitro* hiPSC cardiac-directed differentiation protocol.

**(H)** qPCR analysis showing significant decreases in endoderm and mesendoderm markers and increases in mesoderm markers, respectively, in *SIX3* CRISPRi KD iPSCs compared to control ($n$=6-14 technical replicates per condition from 3-6 experiments).

**(I-K)** Day 2 FACS analysis of endoderm markers EPCAM/CXCR4. **(I)** Changes in $EPCAM^+/CXCR4^+$ cells between control and dox-treated conditions in *SIX3* CRISPRi KD iPSCs and WTC GCaMP CRISPRi iPSCs are shown ($n$=12-16 technical replicates per condition from 4-5 experiments). **(J)** *SIX3* CRISPRi KD

iPSCs show significant ($p$<0.001) reduction in EPCAM$^+$/CXCR4$^+$ cells compared to dox-treated control iPSCs (WTC GCaMP CRISPRi). **(K)** Raw FACS plots of EPCAM/CXCR4 analysis.

**(L-N)** Analysis of cardiomyocytes by FACs for α-actinin at day 15 of *in vitro* differentiation. **(L)** Changes in α-actinin$^+$ cells between control and dox-treated conditions in *SIX3* CRISPRi KD iPSCs and WTC GCaMP CRISPRi iPSCs are shown ($n$=6 technical replicates per condition from 3 experiments). **(M)** *SIX3* CRISPRi KD iPSCs show no change in α-actinin$^+$ cells compared to dox-treated control iPSCs (WTC GCaMP CRISPRi). **(N)** Raw FACS plots of α-actinin analysis.

**(O)** Schematic overview of cardiac development in *Ciona* from 4 to 72 hours post fertilization (hpf) at 18$^o$C. Adapted from (Evans Anderson and Christiaen, 2016). TVC: trunk ventral cells; STVC: second TVC; FHP: first heart precursor; SHP: second heart precursor; ASMF: atrial siphon muscle founder cells; iASMP: inner atrial siphon muscle precursor; oASMP: outer atrial siphon muscle precursor.

**(P)** Top 10 genes ranked by expression value (left) or TRIAGE (right) from populations found during *Ciona* heart development *in vivo*, highlighting *RNF220* as a candidate novel gene.

**(Q)** Breadths of H3K27me3 domains (in base-pairs) associated with *RNF220* gene across the 111 NIH Epigenomes data sets.

**(R-S)** Mosaic plots **(R)** and images **(S)** showing ASM precursor phenotypes at 26 hpf labeled with Mesp>H2B:GFP and Mesp>mCherry in control knockout and *RNF220*-knockout animals ($n$=100). $p$-value represents the chi-sq test between two experimental conditions. Images in (S) derived from *Ciona robusta* cardiopharyngeal mesoderm.

## References


Adsera, C.B., Park, Y.P., Meuleman, W., and Kellis, M. (2019). Integrative analysis of 10,000 epigenomic maps across 800 samples for regulatory genomics and disease dissection. BioRxiv *810291*.

Alexa, A. (2019). topGO: Enrichment Analysis for Gene Ontology. R package version 2.38.1.

Anand, P., Brown, J.D., Lin, C.Y., Qi, J., Zhang, R., Artero, P.C., Alaiti, M.A., Bullard, J., Alazem, K., Margulies, K.B., *et al.* (2013). BET bromodomains mediate transcriptional pause release in heart failure. Cell *154*, 569-582.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Benayoun, B.A., Pollina, E.A., Ucar, D., Mahmoudi, S., Karra, K., Wong, E.D., Devarajan, K., Daugherty, A.C., Kundaje, A.B., Mancini, E., *et al.* (2014). H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell *158*, 673-688.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Burridge, P.W., Matsa, E., Shukla, P., Lin, Z.C., Churko, J.M., Ebert, A.D., Lan, F., Diecke, S., Huber, B., Mordwinkin, N.M., *et al.* (2014). Chemically defined generation of human cardiomyocytes. Nat Methods *11*, 855-860.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology *36*, 411.

Cahan, P., Li, H., Morris, S.A., Da Rocha, E.L., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903-915.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., *et al.* (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496-502.

Carl, M., Loosli, F., and Wittbrodt, J. (2002). Six3 inactivation reveals its essential role for the formation and patterning of the vertebrate eye. Development *129*, 4057-4063.

Christiaen, L., Wagner, E., Shi, W., and Levine, M. (2009). Electroporation of transgenic DNAs in the sea squirt Ciona. Cold Spring Harb Protoc *2009*, pdb prot5345.

Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol *17*, 72.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., *et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M., *et al.* (2015). A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. Stem Cell Reports *5*, 763-775.

Davis, F.P., and Eddy, S.R. (2013). Transcription factors that convert adult cell identity are differentially polycomb repressed. PLoS One *8*, e63407.

Duan, Q., McMahon, S., Anand, P., Shah, H., Thomas, S., Salunga, H.T., Huang, Y., Zhang, R., Sahadevan, A., Lemieux, M.E., *et al.* (2017). BET bromodomain inhibition suppresses innate inflammatory and profibrotic transcriptional networks in heart failure. Sci Transl Med *9*.

Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet *29*, 569-574.

Evans Anderson, H., and Christiaen, L. (2016). Ciona as a simple chordate model for heart development and regeneration. Journal of cardiovascular development and disease *3*, 25.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728-1740.

Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., *et al.* (2014). A promoter-level mammalian expression atlas. Nature *507*, 462-470.

Friedman, C.E., Nguyen, Q., Lukowski, S.W., Helfer, A., Chiu, H.S., Miklas, J., Levy, S., Suo, S., Han, J.-D.J., and Osteil, P. (2018). Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. Cell stem cell *23*, 586-598. e588.

Fujikura, J., Yamato, E., Yonemura, S., Hosoda, K., Masui, S., Nakao, K., Miyazaki Ji, J., and Niwa, H. (2002). Differentiation of embryonic stem cells is induced by GATA factors. Genes Dev *16*, 784-789.

Grote, P., and Herrmann, B.G. (2013). The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. RNA Biol *10*, 1579-1585.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). BioMart Central Portal--unified access to biological data. Nucleic Acids Res *37*, W23-27.

Heinaniemi, M., Nykter, M., Kramer, R., Wienecke-Baldacchino, A., Sinkkonen, L., Zhou, J.X., Kreisberg, R., Kauffman, S.A., Huang, S., and Shmulevich, I. (2013). Gene-pair expression signatures reveal lineage control. Nat Methods *10*, 577-583.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.

Jiang, Y., Qian, F., Bai, X., Liu, Y., Wang, Q., Ai, B., Han, X., Shi, S., Zhang, J., Li, X., *et al.* (2019). SEdb: a comprehensive human super-enhancer database. Nucleic Acids Res *47*, D235-D243.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol *26*, 1351-1359.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., *et al.* (2014). A draft map of the human proteome. Nature *509*, 575-581.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., *et al.* (2017). SC3: consensus clustering of single-cell RNA-seq data. Nature Methods *14*, 483.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Lagutin, O.V., Zhu, C.C., Kobayashi, D., Topczewski, J., Shimamura, K., Puelles, L., Russell, H.R., McKinnon, P.J., Solnica-Krezel, L., and Oliver, G. (2003). Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. Genes & development *17*, 368-379.

Lang, A.H., Li, H., Collins, J.J., and Mehta, P. (2014). Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput Biol *10*, e1003734.

Lattanzi, L., Salvatori, G., Coletta, M., Sonnino, C., Cusella De Angelis, M.G., Gioglio, L., Murry, C.E., Kelly, R., Ferrari, G., Molinaro, M., *et al.* (1998). High efficiency myogenic conversion of human fibroblasts by adenoviral vector-mediated MyoD gene transfer. An alternative strategy for ex vivo gene therapy of primary myopathies. J Clin Invest *101*, 2119-2128.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., *et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.

Ma, P., Yang, X., Kong, Q., Li, C., Yang, S., Li, Y., and Mao, B. (2014). The ubiquitin ligase RNF220 enhances canonical Wnt signaling through USP7-mediated deubiquitination of beta-catenin. Mol Cell Biol *34*, 4355-4366.

Mandegar, M.A., Huebsch, N., Frolov, E.B., Shin, E., Truong, A., Olvera, M.P., Chan, A.H., Miyaoka, Y., Holmes, K., Spencer, C.I., *et al.* (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell *18*, 541-553.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343-349.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Morris, S.A., and Daley, G.Q. (2013). A blueprint for engineering cell fate: current technologies to reprogram cell identity. Cell Res *23*, 33-48.

Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., *et al.* (2018). Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol *36*, 428-431.

Nakamura, R., Tsukahara, T., Qu, W., Ichikawa, K., Otsuka, T., Ogoshi, K., Saito, T.L., Matsushima, K., Sugano, S., Hashimoto, S., *et al.* (2014). Large hypomethylated domains serve as strong repressive machinery for key developmental genes in vertebrates. Development *141*, 2568-2580.

Paige, S.L., Thomas, S., Stoick-Cooper, C.L., Wang, H., Maves, L., Sandstrom, R., Pabon, L., Reinecke, H., Pratt, G., Keller, G., *et al.* (2012). A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. Cell *151*, 221-232.

Palpant, N.J., Pabon, L., Friedman, C.E., Roberts, M., Hadland, B., Zaunbrecher, R.J., Bernstein, I., Zheng, Y., and Murry, C.E. (2017a). Generating high-purity cardiac and endothelial derivatives from patterned mesoderm using human pluripotent stem cells. Nat Protoc *12*, 15-31.

Palpant, N.J., Wang, Y., Hadland, B., Zaunbrecher, R.J., Redd, M., Jones, D., Pabon, L., Jain, R., Epstein, J., Ruzzo, W.L., *et al.* (2017b). Chromatin and Transcriptional Analysis of Mesoderm Progenitor Cells Identifies HOPX as a Regulator of Primitive Hematopoiesis. Cell Rep *20*, 1597-1608.

Peng, G., Suo, S., Chen, J., Chen, W., Liu, C., Yu, F., Wang, R., Chen, S., Sun, N., Cui, G., *et al.* (2016). Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Dev Cell *36*, 681-697.

Perez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., and Guigo, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. Nat Genet *47*, 1158-1167.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rackham, O.J., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., and Shin, J.W. (2016). A predictive computational framework for direct reprogramming between human cell types. Nature genetics *48*, 331.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279.

Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nat Methods *14*, 263-266.

Rehimi, R., Nikolic, M., Cruz-Molina, S., Tebartz, C., Frommolt, P., Mahabir, E., Clement-Ziza, M., and Rada-Iglesias, A. (2016). Epigenomics-Based Identification of Major Cell Identity Regulators within Heterogeneous Cell Populations. Cell Rep *17*, 3062-3076.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Roost, M.S., van Iperen, L., Ariyurek, Y., Buermans, H.P., Arindrarto, W., Devalla, H.D., Passier, R., Mummery, C.L., Carlotti, F., de Koning, E.J., *et al.* (2015). KeyGenes, a Tool to Probe Tissue Differentiation Using a Human Fetal Transcriptional Atlas. Stem Cell Reports *4*, 1112-1124.

Russ, A.P., Wattler, S., Colledge, W.H., Aparicio, S.A., Carlton, M.B., Pearce, J.J., Barton, S.C., Surani, M.A., Ryan, K., Nehls, M.C., *et al.* (2000). Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature *404*, 95-99.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat Biotechnol *33*, 495-502.

Schaum, N., Karkanias, J., Neff, N., May, A.P., S.R., Q., Wyss-Coray, T., Batson, J., Botvinnik, O., Chen, M.B., Chen, S., *et al.* (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367-372.

Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol *6*, R33.

Scornavacca, C., Zickmann, F., and Huson, D.H. (2011). Tanglegrams for rooted phylogenetic trees and networks. Bioinformatics *27*, i248-256.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. The R journal *8*, 289-317.

Senabouth, A., Lukowski, S.W., Hernandez, J.A., Andersen, S.B., Mei, X., Nguyen, Q.H., and Powell, J.E. (2019). ascend: R package for analysis of single-cell RNA-seq data. Gigascience *8*.

Steinmetz, P.R., Urbach, R., Posnien, N., Eriksson, J., Kostyuchenko, R.P., Brena, C., Guy, K., Akam, M., Bucher, G., and Arendt, D. (2010). Six3 demarcates the anterior-most developing brain region in bilaterian animals. EvoDevo *1*, 14.

Stolfi, A., Gandhi, S., Salek, F., and Christiaen, L. (2014). Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. Development *141*, 4115-4120.

Takeuchi, J.K., and Bruneau, B.G. (2009). Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors. Nature *459*, 708-711.

Tian, L., Dong, X., Freytag, S., Le Cao, K.A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T.S., Seidi, A., Jabbari, J.S., *et al.* (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods *16*, 479-487.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., *et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189.

Van Handel, B., Montel-Hagen, A., Sasidharan, R., Nakano, H., Ferrari, R., Boogerd, C.J., Schredelseker, J., Wang, Y., Hunter, S., Org, T., *et al.* (2012). Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. Cell *150*, 590-605.

Verfaillie, A., Imrichova, H., Atak, Z.K., Dewaele, M., Rambow, F., Hulselmans, G., Christiaens, V., Svetlichnyy, D., Luciani, F., Van den Mooter, L., *et al.* (2015). Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. Nat Commun *6*, 6683.

Waardenberg, A.J., Ramialison, M., Bouveret, R., and Harvey, R.P. (2014). Genetic networks governing heart development. Cold Spring Harb Perspect Med *4*, a013839.

Wang, W., Niu, X., Jullian, E., Kelly, R., Satija, R., and Christiaen, L. (2019). A single cell transcriptional roadmap for cardiopharyngeal fate diversification. BioRxiv.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. (2008). DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res *36*, D88-92.

Wu, S.F., Zhang, H., and Cairns, B.R. (2011). Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. Genome Res *21*, 578-589.

Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res *43*, D76-81.

Zhong, Y.F., and Holland, P.W. (2011). HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evol Dev *13*, 567-568.

Figure 1

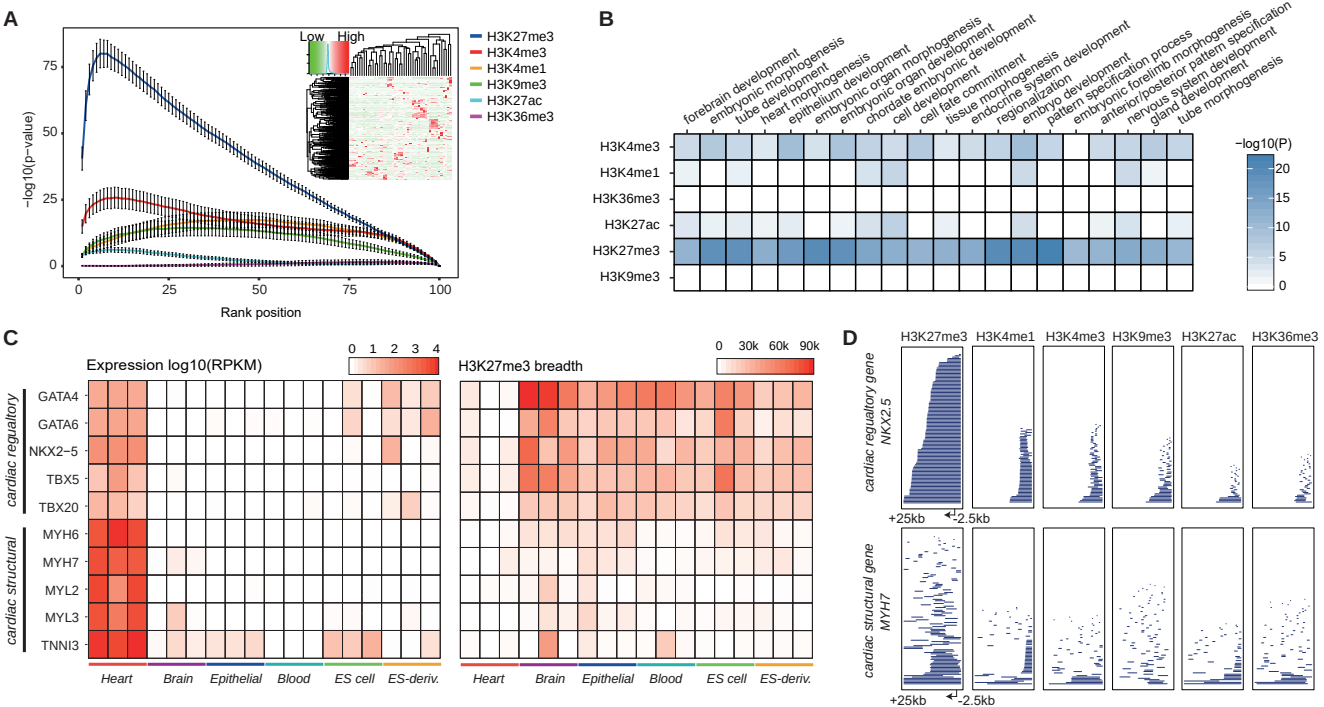

A
−log10(p−value)
Rank position
Low
High
H3K27me3
H3K4me3
H3K4me1
H3K9me3
H3K27ac
H3K36me3
B
forebrain development
embryonic morphogenesis
tube development
heart morphogenesis
epithelium development
embryonic organ morphogenesis
embryonic organ development
chordate embryonic development
cell development
cell fate commitment
tissue morphogenesis
endocrine system development
regionalization
embryo development
pattern specification process
embryonic forelimb morphogenesis
anterior/posterior pattern specification
nervous system development
gland development
tube morphogenesis
H3K4me3
H3K4me1
H3K36me3
H3K27ac
H3K27me3
H3K9me3
−log10(P)
C
Expression log10(RPKM)
H3K27me3 breadth
cardiac regulatory
cardiac structural
GATA4
GATA6
NKX2−5
TBX5
TBX20
MYH6
MYH7
MYL2
MYL3
TNNI3
Heart
Brain
Epithelial
Blood
ES cell
ES-deriv.
D
H3K27me3
H3K4me1
H3K4me3
H3K9me3
H3K27ac
H3K36me3
cardiac regulatory gene
NKX2.5
cardiac structural gene
MYH7
+25kb
-2.5kb

Figure 2

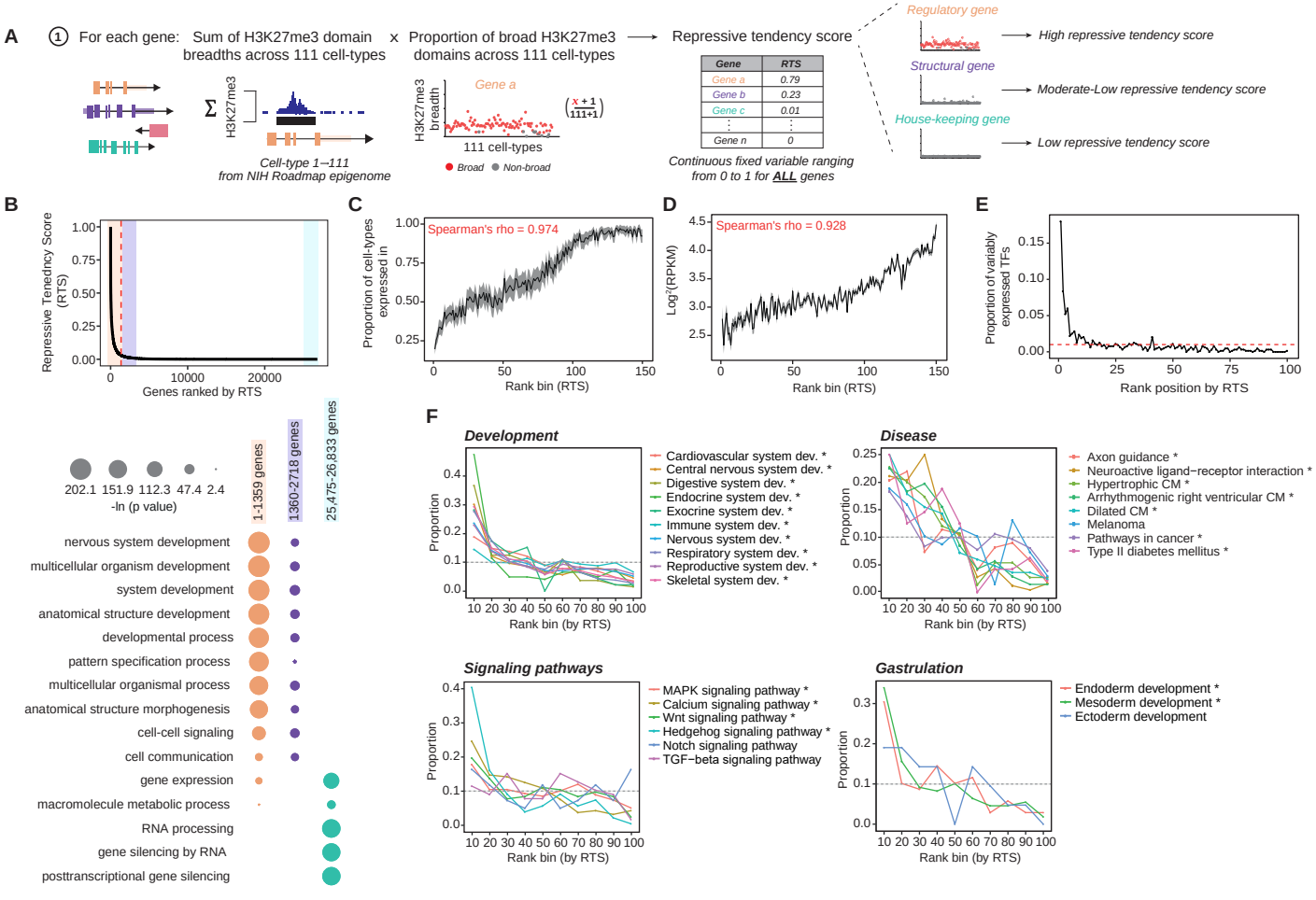

A
① For each gene: Sum of H3K27me3 domain breadths across 111 cell-types × Proportion of broad H3K27me3 domains across 111 cell-types → Repressive tendency score
H3K27me3
Cell-type 1–111 from NIH Roadmap epigenome
H3K27me3 breadth
Gene a
111 cell-types
Broad
Non-broad
Gene | RTS
Gene a | 0.79
Gene b | 0.23
Gene c | 0.01
Gene n | 0
Continuous fixed variable ranging from 0 to 1 for ALL genes
Regulatory gene
High repressive tendency score
Structural gene
Moderate-Low repressive tendency score
House-keeping gene
Low repressive tendency score
B
Repressive Tenedncy Score (RTS)
Genes ranked by RTS
C
Spearman's rho = 0.974
Proportion of cell-types expressed in
Rank bin (RTS)
D
Spearman's rho = 0.928
Log2(RPKM)
Rank bin (RTS)
E
Proportion of variably expressed TFs
Rank position by RTS
202.1 151.9 112.3 47.4 2.4
-ln (p value)
1-1359 genes
1360-2718 genes
25,475-26,833 genes
nervous system development
multicellular organism development
system development
anatomical structure development
developmental process
pattern specification process
multicellular organismal process
anatomical structure morphogenesis
cell-cell signaling
cell communication
gene expression
macromolecule metabolic process
RNA processing
gene silencing by RNA
posttranscriptional gene silencing
F
Development
Cardiovascular system dev. *
Central nervous system dev. *
Digestive system dev. *
Endocrine system dev. *
Exocrine system dev. *
Immune system dev. *
Nervous system dev. *
Respiratory system dev. *
Reproductive system dev. *
Skeletal system dev. *
Disease
Axon guidance *
Neuroactive ligand−receptor interaction *
Hypertrophic CM *
Arrhythmogenic right ventricular CM *
Dilated CM *
Melanoma
Pathways in cancer *
Type II diabetes mellitus *
Signaling pathways
MAPK signaling pathway *
Calcium signaling pathway *
Wnt signaling pathway *
Hedgehog signaling pathway *
Notch signaling pathway
TGF−beta signaling pathway
Gastrulation
Endoderm development *
Mesoderm development *
Ectoderm development
Proportion
Rank bin (by RTS)

Figure 3

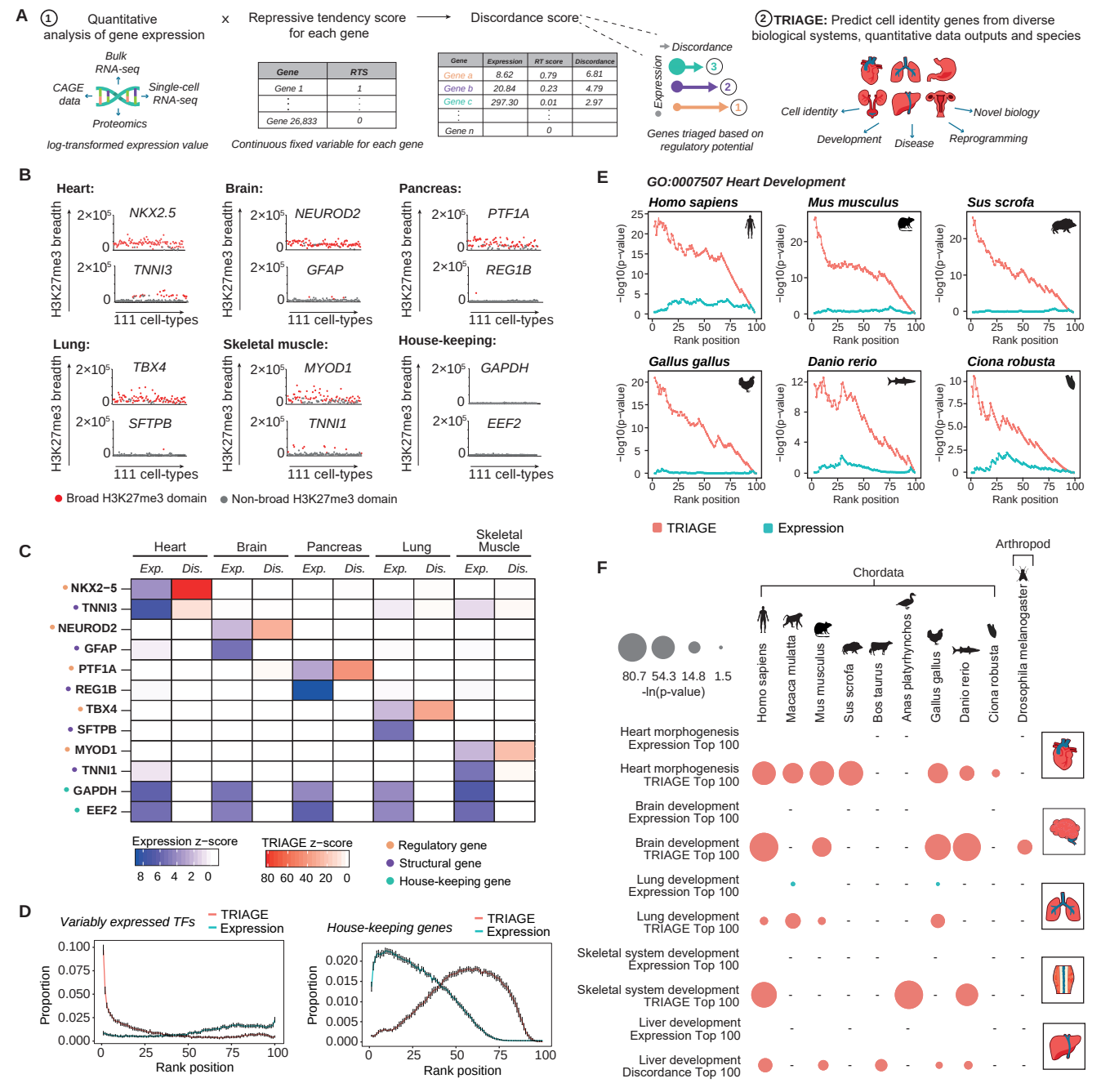

A
① Quantitative analysis of gene expression
Bulk RNA-seq
CAGE data
Single-cell RNA-seq
Proteomics
log-transformed expression value
x
Repressive tendency score for each gene
Gene | RTS
Gene 1 | 1
Gene 26,833 | 0
Continuous fixed variable for each gene
Discordance score
Gene | Expression | RT score | Discordance
Gene a | 8.62 | 0.79 | 6.81
Gene b | 20.84 | 0.23 | 4.79
Gene c | 297.30 | 0.01 | 2.97
Gene n | | 0 |
Discordance
Expression
Genes triaged based on regulatory potential
② TRIAGE: Predict cell identity genes from diverse biological systems, quantitative data outputs and species
Cell identity
Novel biology
Development
Disease
Reprogramming
B
Heart:
Brain:
Pancreas:
Lung:
Skeletal muscle:
House-keeping:
H3K27me3 breadth
NKX2.5
TNNI3
NEUROD2
GFAP
PTF1A
REG1B
TBX4
SFTPB
MYOD1
TNNI1
GAPDH
EEF2
111 cell-types
Broad H3K27me3 domain
Non-broad H3K27me3 domain
C
Heart
Brain
Pancreas
Lung
Skeletal Muscle
Exp.
Dis.
NKX2−5
TNNI3
NEUROD2
GFAP
PTF1A
REG1B
TBX4
SFTPB
MYOD1
TNNI1
GAPDH
EEF2
Expression z−score
TRIAGE z−score
Regulatory gene
Structural gene
House-keeping gene
D
Variably expressed TFs
House-keeping genes
TRIAGE
Expression
Proportion
Rank position
E
GO:0007507 Heart Development
Homo sapiens
Mus musculus
Sus scrofa
Gallus gallus
Danio rerio
Ciona robusta
−log10(p−value)
Rank position
TRIAGE
Expression
F
Arthropod
Chordata
Homo sapiens
Macaca mulatta
Mus musculus
Sus scrofa
Bos taurus
Anas platyrhynchos
Gallus gallus
Danio rerio
Ciona robusta
Drosophila melanogaster
80.7 54.3 14.8 1.5
-ln(p-value)
Heart morphogenesis Expression Top 100
Heart morphogenesis TRIAGE Top 100
Brain development Expression Top 100
Brain development TRIAGE Top 100
Lung development Expression Top 100
Lung development TRIAGE Top 100
Skeletal system development Expression Top 100
Skeletal system development TRIAGE Top 100
Liver development Expression Top 100
Liver development Discordance Top 100

Figure 4

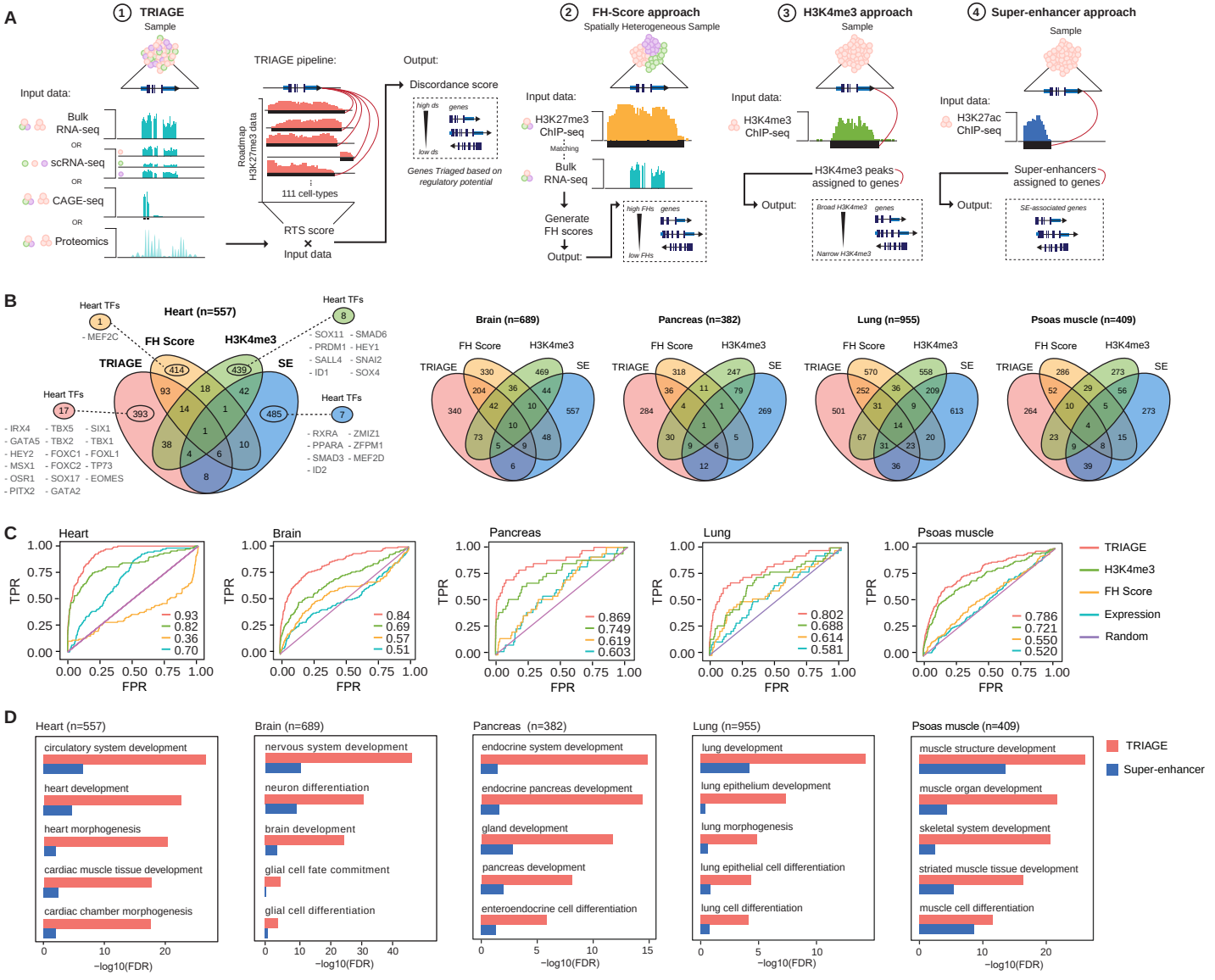

A
1 TRIAGE
Sample
Input data:
Bulk RNA-seq
OR
scRNA-seq
OR
CAGE-seq
OR
Proteomics
TRIAGE pipeline:
Roadmap H3K27me3 data
111 cell-types
RTS score × Input data
Output:
Discordance score
high ds
genes
low ds
Genes Triaged based on regulatory potential
2 FH-Score approach
Spatially Heterogeneous Sample
Input data:
H3K27me3 ChIP-seq
Matching
Bulk RNA-seq
Generate FH scores
Output:
high FHs
genes
low FHs
3 H3K4me3 approach
Sample
Input data:
H3K4me3 ChIP-seq
H3K4me3 peaks assigned to genes
Output:
Broad H3K4me3
genes
Narrow H3K4me3
4 Super-enhancer approach
Sample
Input data:
H3K27ac ChIP-seq
Super-enhancers assigned to genes
Output:
SE-associated genes
B
Heart (n=557)
Heart TFs
1
- MEF2C
FH Score
H3K4me3
TRIAGE
SE
414
439
93
18
42
14
1
1
38
10
4
6
8
393
485
Heart TFs
17
- IRX4 - TBX5 - SIX1
- GATA5 - TBX2 - TBX1
- HEY2 - FOXC1 - FOXL1
- MSX1 - FOXC2 - TP73
- OSR1 - SOX17 - EOMES
- PITX2 - GATA2
Heart TFs
8
- SOX11 - SMAD6
- PRDM1 - HEY1
- SALL4 - SNAI2
- ID1 - SOX4
Heart TFs
7
- RXRA - ZMIZ1
- PPARA - ZFPM1
- SMAD3 - MEF2D
- ID2
Brain (n=689)
330
469
204
36
44
42
10
340
557
10
73
5
9
48
6
Pancreas (n=382)
318
247
36
11
79
4
1
284
269
30
9
6
5
12
Lung (n=955)
570
558
252
36
209
31
9
501
613
14
67
20
31
23
36
Psoas muscle (n=409)
286
273
52
29
56
10
5
264
273
4
23
9
8
15
39
C
Heart
Brain
Pancreas
Lung
Psoas muscle
TPR
FPR
0.00
0.25
0.50
0.75
1.00
0.93
0.82
0.36
0.70
0.84
0.69
0.57
0.51
0.869
0.749
0.619
0.603
0.802
0.688
0.614
0.581
0.786
0.721
0.550
0.520
TRIAGE
H3K4me3
FH Score
Expression
Random
D
Heart (n=557)
circulatory system development
heart development
heart morphogenesis
cardiac muscle tissue development
cardiac chamber morphogenesis
Brain (n=689)
nervous system development
neuron differentiation
brain development
glial cell fate commitment
glial cell differentiation
Pancreas (n=382)
endocrine system development
endocrine pancreas development
gland development
pancreas development
enteroendocrine cell differentiation
Lung (n=955)
lung development
lung epithelium development
lung morphogenesis
lung epithelial cell differentiation
lung cell differentiation
Psoas muscle (n=409)
muscle structure development
muscle organ development
skeletal system development
striated muscle tissue development
muscle cell differentiation
−log10(FDR)
TRIAGE
Super-enhancer

Figure 5

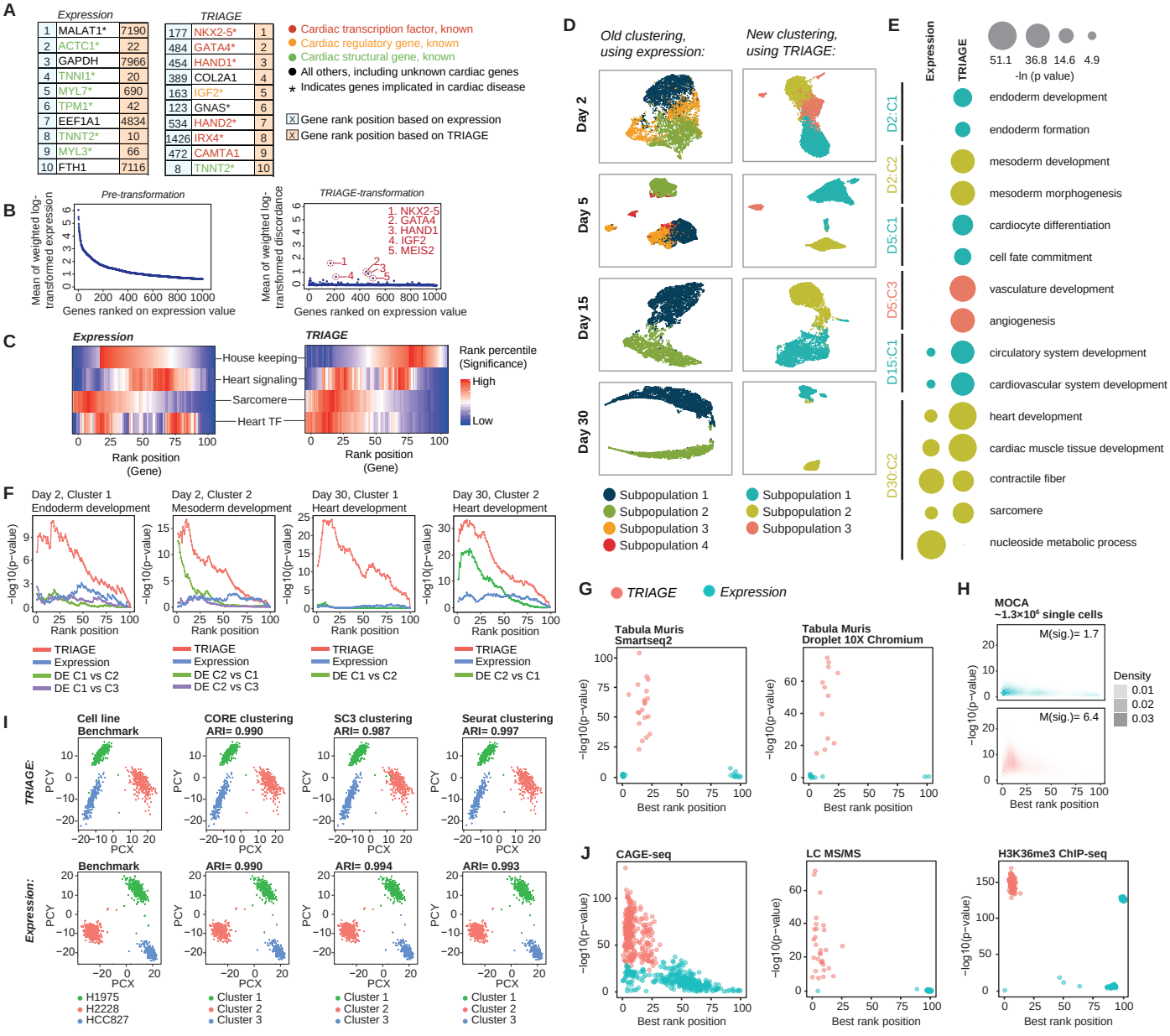
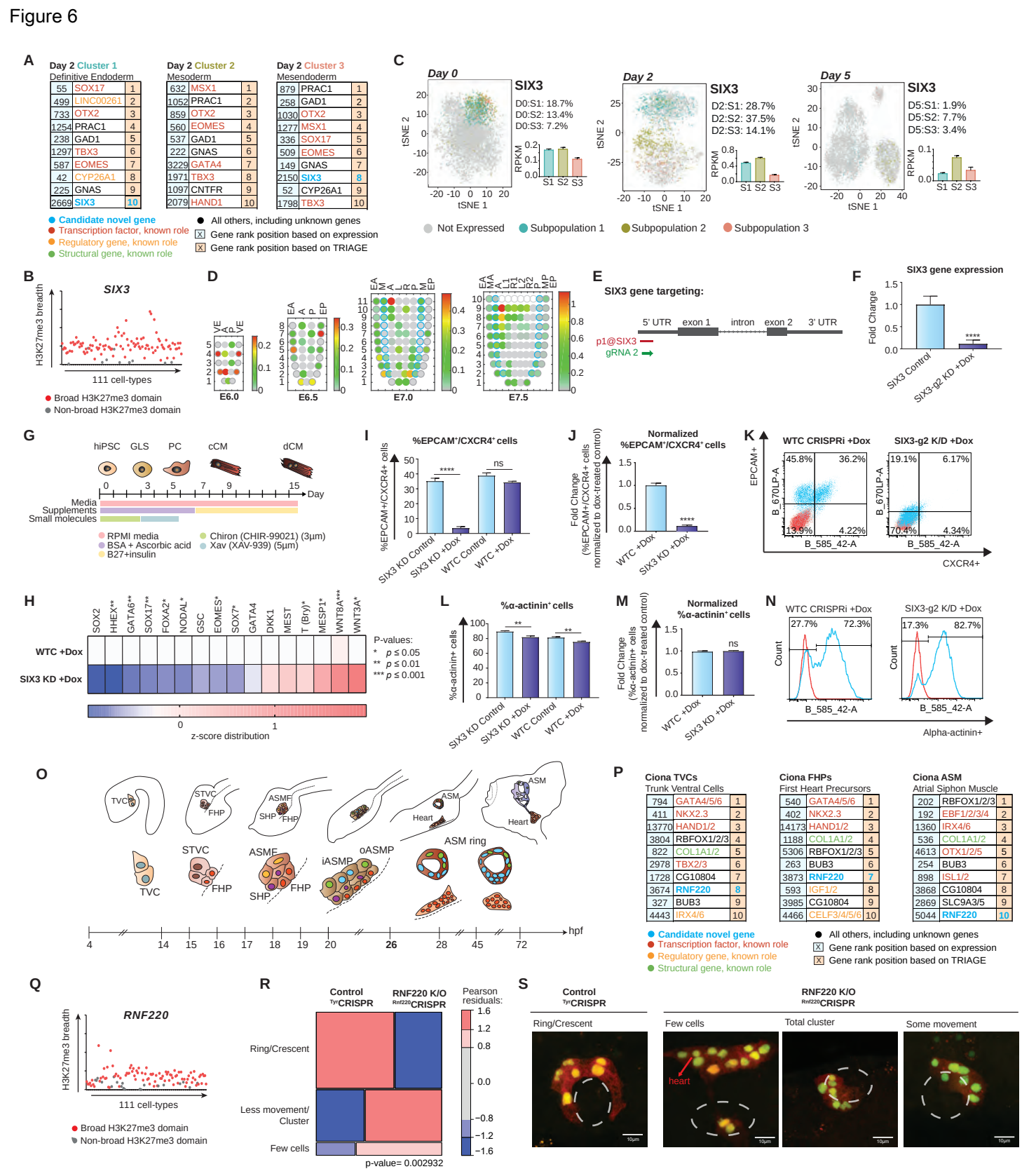

A
Expression
TRIAGE
Cardiac transcription factor, known
Cardiac regulatory gene, known
Cardiac structural gene, known
All others, including unknown cardiac genes
Indicates genes implicated in cardiac disease
Gene rank position based on expression
Gene rank position based on TRIAGE
B
Pre-transformation
TRIAGE-transformation
Mean of weighted log-transformed expression
Mean of weighted log-transformed discordance
Genes ranked on expression value
1. NKX2-5
2. GATA4
3. HAND1
4. IGF2
5. MEIS2
C
Expression
TRIAGE
House keeping
Heart signaling
Sarcomere
Heart TF
Rank percentile (Significance)
High
Low
Rank position (Gene)
D
Old clustering, using expression:
New clustering, using TRIAGE:
Day 2
Day 5
Day 15
Day 30
Subpopulation 1
Subpopulation 2
Subpopulation 3
Subpopulation 4
E
Expression
TRIAGE
51.1
36.8
14.6
4.9
-ln (p value)
D2:C1
D2:C2
D5:C1
D5:C3
D15:C1
D30:C2
endoderm development
endoderm formation
mesoderm development
mesoderm morphogenesis
cardiocyte differentiation
cell fate commitment
vasculature development
angiogenesis
circulatory system development
cardiovascular system development
heart development
cardiac muscle tissue development
contractile fiber
sarcomere
nucleoside metabolic process
F
Day 2, Cluster 1 Endoderm development
Day 2, Cluster 2 Mesoderm development
Day 30, Cluster 1 Heart development
Day 30, Cluster 2 Heart development
−log10(p−value)
Rank position
TRIAGE
Expression
DE C1 vs C2
DE C1 vs C3
DE C2 vs C1
DE C2 vs C3
G
TRIAGE
Expression
Tabula Muris Smartseq2
Tabula Muris Droplet 10X Chromium
Best rank position
H
MOCA ~1.3×10^6 single cells
M(sig.)= 1.7
M(sig.)= 6.4
Density
0.01
0.02
0.03
I
Cell line Benchmark
CORE clustering ARI= 0.990
SC3 clustering ARI= 0.987
Seurat clustering ARI= 0.997
TRIAGE:
Benchmark
ARI= 0.990
ARI= 0.994
ARI= 0.993
Expression:
PCX
PCY
H1975
H2228
HCC827
Cluster 1
Cluster 2
Cluster 3
J
CAGE-seq
LC MS/MS
H3K36me3 ChIP-seq

Figure 6


A
Day 2 Cluster 1
Definitive Endoderm
55 SOX17 1
499 LINC00261 2
733 OTX2 3
1254 PRAC1 4
238 GAD1 5
1297 TBX3 6
587 EOMES 7
42 CYP26A1 8
225 GNAS 9
2669 SIX3 10
Day 2 Cluster 2
Mesoderm
632 MSX1 1
1052 PRAC1 2
859 OTX2 3
560 EOMES 4
537 GAD1 5
222 GNAS 6
3229 GATA4 7
1971 TBX3 8
1097 CNTFR 9
2079 HAND1 10
Day 2 Cluster 3
Mesendoderm
879 PRAC1 1
258 GAD1 2
1030 OTX2 3
1277 MSX1 4
336 SOX17 5
509 EOMES 6
149 GNAS 7
2150 SIX3 8
52 CYP26A1 9
1798 TBX3 10
Candidate novel gene
Transcription factor, known role
Regulatory gene, known role
Structural gene, known role
All others, including unknown genes
Gene rank position based on expression
Gene rank position based on TRIAGE
C
Day 0
SIX3
D0:S1: 18.7%
D0:S2: 13.4%
D0:S3: 7.2%
Day 2
SIX3
D2:S1: 28.7%
D2:S2: 37.5%
D2:S3: 14.1%
Day 5
SIX3
D5:S1: 1.9%
D5:S2: 7.7%
D5:S3: 3.4%
tSNE 1
tSNE 2
RPKM
S1 S2 S3
Not Expressed
Subpopulation 1
Subpopulation 2
Subpopulation 3
B
SIX3
H3K27me3 breadth
111 cell-types
Broad H3K27me3 domain
Non-broad H3K27me3 domain
D
E6.0
E6.5
E7.0
E7.5
E
SIX3 gene targeting:
5' UTR
exon 1
intron
exon 2
3' UTR
p1@SIX3
gRNA 2
F
SIX3 gene expression
Fold Change
SIX3 Control
SIX3-g2 KD +Dox
G
hiPSC
GLS
PC
cCM
dCM
Day
Media
Supplements
Small molecules
RPMI media
BSA + Ascorbic acid
B27+insulin
Chiron (CHIR-99021) (3μm)
Xav (XAV-939) (5μm)
I
%EPCAM+/CXCR4+ cells
SIX3 KD Control
SIX3 KD +Dox
WTC Control
WTC +Dox
J
Normalized %EPCAM+/CXCR4+ cells
Fold Change (%EPCAM+/CXCR4+ cells normalized to dox-treated control)
K
WTC CRISPRi +Dox
SIX3-g2 K/D +Dox
EPCAM+
CXCR4+
B_670LP-A
B_585_42-A
45.8%
36.2%
13.9%
4.22%
19.1%
6.17%
70.4%
4.34%
H
SOX2
HHEX**
GATA6**
SOX17**
FOXA2*
NODAL*
GSC
EOMES*
SOX7*
GATA4
DKK1
MEST
T (Bry)*
MESP1*
WNT8A***
WNT3A*
WTC +Dox
SIX3 KD +Dox
P-values:
* p ≤ 0.05
** p ≤ 0.01
*** p ≤ 0.001
z-score distribution
L
%α-actinin+ cells
SIX3 KD Control
SIX3 KD +Dox
WTC Control
WTC +Dox
M
Normalized %α-actinin+ cells
Fold Change (%α-actinin+ cells normalized to dox-treated control)
N
WTC CRISPRi +Dox
SIX3-g2 K/D +Dox
27.7%
72.3%
17.3%
82.7%
Count
B_585_42-A
Alpha-actinin+
O
TVC
STVC
FHP
ASMF
SHP
ASM
Heart
ASM ring
oASMP
iASMP
hpf
4
14
15
16
17
18
19
20
26
28
45
72
P
Ciona TVCs
Trunk Ventral Cells
794 GATA4/5/6 1
411 NKX2.3 2
13770 HAND1/2 3
3804 RBFOX1/2/3 4
822 COL1A1/2 5
2978 TBX2/3 6
1728 CG10804 7
3674 RNF220 8
327 BUB3 9
4443 IRX4/6 10
Ciona FHPs
First Heart Precursors
540 GATA4/5/6 1
402 NKX2.3 2
14173 HAND1/2 3
1188 COL1A1/2 4
5306 RBFOX1/2/3 5
263 BUB3 6
3873 RNF220 7
593 IGF1/2 8
3985 CG10804 9
4466 CELF3/4/5/6 10
Ciona ASM
Atrial Siphon Muscle
202 RBFOX1/2/3 1
192 EBF1/2/3/4 2
1360 IRX4/6 3
536 COL1A1/2 4
4613 OTX1/2/5 5
254 BUB3 6
898 ISL1/2 7
3868 CG10804 8
2869 SLC9A3/5 9
5044 RNF220 10
Candidate novel gene
Transcription factor, known role
Regulatory gene, known role
Structural gene, known role
All others, including unknown genes
Gene rank position based on expression
Gene rank position based on TRIAGE
Q
RNF220
H3K27me3 breadth
111 cell-types
Broad H3K27me3 domain
Non-broad H3K27me3 domain
R
Control
Tyr CRISPR
RNF220 K/O
Rnf220 CRISPR
Pearson residuals:
Ring/Crescent
Less movement/ Cluster
Few cells
p-value= 0.002932
S
Control
Tyr CRISPR
RNF220 K/O
Rnf220 CRISPR
Ring/Crescent
Few cells
Total cluster
Some movement
heart
10μm