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], Yash Chhabra[8], 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

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

ABSTRACT

We define a scalable and genome-wide metric based on the gene-repressive tri-methylation of histone 3 lysine 27 that strongly enriches for genes that drive cell diversification and determine their fates. Transcriptional Regulatory Inference Analysis from Gene Expression (TRIAGE) incorporates tendencies of H3K27me3 as deposited across over 100 representative cell states. As a consequence, inference of cell identity genes from expressed transcripts of any somatic cell type is made possible without requiring its correlative epigenetic data. We combine more than 1 million genome-wide data sets from different omics platforms to identify and experimentally validate cell type-specific regulatory mechanisms for organ systems in health and disease. The success with which driver genes are recovered attests to a repression-based regulatory logic conserved in species across the animal kingdom, and suggests a simple but effective computational approach to determine causal factors from gene output alone.

INTRODUCTION

Genetic regulation of cellular identity during differentiation and homeostasis is governed by highly conserved epigenetic regulation of chromatin states. Recent technological advances are enabling the capture of genome-wide information at the level of a single cells, providing an unprecedented level of detail into mechanisms of cellular identity. However, using genome-wide analyses of cellular states to identify regulatory cell identity programs is challenging due to the high abundance of structural and house-keeping genes. For this reason, predicting regulatory determinants of cell identity from genomic data involves cross-referencing between multiple data sets from complementary technologies[1-3] and remains an ongoing and essential strategy for gleaning novel insights into developmental biology, disease mechanisms and cellular reprogramming[1,4,5].

The breadth of histone modifications (HMs) has been shown to be structurally and functionally linked to cell-specific genome architecture and gene regulation[1,2,6-8]. For example, broad domains of H3K4me3 preferentially mark actively transcribed drivers of cell identity[1] and broad domains of repressive H3K27me3 are deposited over transcriptionally inactive drivers of cell identity[2,9]. Indeed, H3K27me3 functionally represses transcription[10,11] and therefore plays an essential juxtaposing role to H3K4me3[1] and H3K27ac[12] in regulating gene expression. Computational approaches incorporating genome-wide chromatin and transcriptional data have utilized these genomic features to predict the regulatory and cell-specific elements of cells[1-3,13]. We utilized histone methylation and transcriptional data from a diverse spectrum of cell states to evaluate genome-wide transcriptional features of genes safe-guarding cell identity to establish a computational logic that can be customized to infer regulatory drivers of *any* cell.

Here, we show the *a priori* probability of broad H3K27me3 domains signifies a repressive tendency of a gene in diverse cell states. This provides a quantifiable metric strongly predictive of regulatory genes governing mechanisms of cell differentiation and organ morphogenesis in health and disease. It defines a metric of gene regulatory potential that can be applied at scale to any quantitative readout of gene expression from theoretically any cell-state in an analysis termed TRIAGE (**T**ranscriptional **R**egulatory

Inference **A**nalysis of **G**ene **E**xpression). We demonstrate that TRIAGE can be used to analyze individual transcriptomes of millions of heterogeneous cells simultaneously to infer the cell type-specific regulatory genes driving somatic cell states across diverse species in the animal kingdom. With new capabilities in studying the genetic state of individual cells, these insights will potentially transform our capacity to understand the mechanistic basis of cellular heterogeneity in health and disease.

RESULTS

Broad histone domains demarcate genes with distinct regulatory roles

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. To predict the regulatory determinants of one state, information about the genome from diverse cell states is required to infer how variations in genome activity deliver biological complexity. We used NIH Epigenome Roadmap data[14], which contains ChIP-seq data for H3K4me3, H3K36me3, H3K27me3, H3K4me1, H3K27ac and H3K9me3 for 111 tissue or cell types **(Supplementary Table 1)**. To associate HM domains with proximal regulatory functions governing gene expression, we linked HM domains within 2.5 kb to known transcriptional start sites of RefSeq genes. For each of the six HMs, genes were annotated based on the broadest HM domain linked to the gene. For each HM, we found that the top 100 genes with the broadest domain were remarkably consistent between cell types **(Fig. 1a)**, however, broad domains of different HMs marked distinct sets of genes **(Supplementary Fig. 1a)**. We further noted that genes marked with broad repressive HMs (i.e. H3K9me3 or H3K27me3) were more consistently shared between cell types than genes marked by other HMs **(Fig. 1a, inset)** suggesting that broad repressive chromatin domains comprise a common strategy for epigenetic control of cell diversification.

We aimed to understand how the breadth of histone domains correlate with genes governing cell identity. To this end, we established a broadly applicable positive gene set for cell type-specific regulatory genes comprised of 634 variably expressed transcription factors (VETFs) having a coefficient of variation

greater than 1 (**Supplementary Table 2**) across 46 NIH Epigenome RNA-seq data sets[14,15]. We used Shannon entropy to quantify cell type specificity[16] and demonstrate that VETFs are significantly more cell type-specific, compared to non-VETFs or protein coding genes. Analysis of RNA-seq data sets from diverse cell and tissue types show that VETFs in each sample reflect appropriate tissue or cell type-specific regulatory functions **(Figs. 1b,c inset)**. Henceforth, VETFs provide a positive gene set where their enrichment is a performance metric for identifying cell type-specific regulatory genes.

We utilized VETFs to determine the relationship between cell type-specific regulatory genes and histone broad domains. To this end, all NIH Epigenome histone ChIP-seq data were ranked by domain 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 **(Fig. 1c)**, while repression of VETFs was significantly correlated with breadth of H3K27me3 domains **(Supplementary Fig. 1b)**. We also showed that genes frequently associated with broad H3K27me3 domains involve various morphogenic and developmental processes **(Fig. 1d)**. Taken together, we demonstrate that quantification of H3K27me3 broad domains from diverse cell and tissue types provides a powerful metric to reproducibly enrich for cell type-specific 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[17], we extracted expression and chromatin data from cardiomyocytes **(Figs. 1e,f)**. 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 **(Fig. 1e).** Furthermore, focusing on H3K27me3 of only the cardiomyocyte samples is uninformative in distinguishing structural from regulatory genes because these genes all lack repressive chromatin. In contrast, in all cell types *except* the heart, H3K27me3 domains broader than 30kb consistently identify cardiac regulatory genes from structural genes **(Fig. 1e)**.

No other HM analyzed demarcates cell type-specific regulatory genes from structural genes in this manner **(Fig. 1f, Supplementary Fig. 1c),** establishing the rationale that the frequency of H3K27me3 across heterogeneous cell types provides a novel strategy to infer the likelihood of a gene having cell type-specific regulatory function.

Cell type-specific regulatory genes tend to be marked by broad H3K27me3 domains

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[10,18]. Conversely, genes with housekeeping or non-regulatory roles rarely host broad H3K27me3 domains. We 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 **(Fig. 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) **(Supplementary Table 3).** Using the NIH Epigenome Roadmap data, the RTS is calculated for 99.3% (or 26,833 genes) of all RefSeq genes. To demonstrate that our 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 **(Supplementary Figs. 2a,b,e,f)**, with over 85% of assigned H3K27me3 domains overlapping only a single protein-coding gene **(Supplementary Fig. 2d)**. 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 that show a significant enrichment for genes underlying cellular

diversification including organismal development, pattern specification and multicellular organismal processes **(Fig. 2b),** and show they are cell type-specific (**Fig. 2c**) and lowly expressed (**Fig. 2d**). Among the 1,359 priority genes, we identified enrichment of regulatory gene sets, including 253 VETFs (**Fig. 2e,** p=3.49e-162, Fisher's exact test, one-tailed) in addition to 157 homeobox proteins[19] (p=6.43e-123). The gene set also includes 260 genes involved in at least one KEGG pathway[20] and 291 non-coding RNAs genes including FENDRR and HOTAIR, regulators of heart development and chromatin dynamics respectively[21,22]. We also demonstrate that genes with a high RTS are enriched in key regulators of processes underlying gastrulation and organ morphogenesis, comprise members of many of the major signaling pathways, as well as genes implicated in pathologies including cardiovascular disease, diabetes, neurological disorders and cancer **(Fig. 2f and Supplementary Table 4).** Taken together, these data indicate that ranking based on a gene's repressive tendency generates a simple and effective strategy to enrich for fundamental genetic determinants of biological complexity of cell states underlying health and disease.

Predicting cell type-specific regulatory genes based on H3K27me3

The transcriptome of a cell derives from a fraction of the genome comprising expression of structural, housekeeping and regulatory genes collectively representing the cell's state. The regulatory genes controlling the identity, fate and function of a cell, such as transcription factors (TFs), are a minor fraction of the transcriptome, and are non-trivial to identify due to their typical low abundance. Furthermore, regulatory genes are difficult to identify without drawing on prior knowledge, utilizing external reference points, or implementing other supervised methods. To address this, we established a mechanism for integrating genome-wide RTS values with cell type-specific transcriptomic data. Since every gene is assigned a fixed RTS value that hierarchically orders the genome based on regulatory likelihood, we devised a computational approach to integrate the signature of any cell's transcriptomic data with the RTS, a method we call TRIAGE (Transcriptional Regulatory Inference Analysis from Gene Expression).

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 introduces a non-linear, gene-specific weight (i.e. RTS) that prioritizes cell type-specific regulatory genes from expression data derived from any cellular state **(Fig. 3a, Supplementary Figs. 2c, 3a, 3b)**. Importantly, this strategy does not refer to any external data set, uses no arbitrary statistical cutoffs, does not require additional cell type-specific epigenetic data, does not focus on a specific gene type such as TFs, nor does it utilize external databases or prior knowledge to derive its prediction.

To evaluate TRIAGE, we identified known regulatory and structural genes from 5 tissue groups, analyzing H3K27me3 of cell-specific regulatory versus structural genes **(Fig. 3b)**. When applied to cell-specific transcriptional data, TRIAGE effectively reduces the relative abundance of structural and housekeeping genes, while enriching for regulatory genes in a cell type-specific manner **(Fig. 3c).** 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 exhibiting functional specificity unique to the input sample **(Fig. 3d and Supplementary Fig. 3c)**. Compared to the expression-based ranking, TRIAGE reduces the relative abundance of housekeeping genes **(Fig. 3d).** A tanglegram based on the Pearson distances between Roadmap tissue types[23] shows that relative to the total height of the dendrograms, TRIAGE increased the similarity between samples from the same tissue by ~29% when compared to distances calculated using absolute expression levels **(Supplementary Fig. 3d)**.

We subsequently evaluated TRIAGE against methods utilizing epigenetic data such as H3K4me3[1], H3K27me3[2] and markers for super-enhancers[8,24] to predict cell regulatory genes. These previous approaches require input cell type-specific data for HMs therefore significantly limiting their scalability.

Even when allowing for this, TRIAGE surpassed all other methods in terms of both sensitivity and precision across diverse tissue types **(Fig. 3e, f, Supplementary Figs. 4a-d)**. TRIAGE also outperforms routine approaches such as differentially expressed gene (DEG) analysis which is susceptible to variation based on the comparison point **(Fig. 4f, Supplementary Figs. 4g,h)**. Lastly, we show that TRIAGE predictions are not explained merely by prioritising TFs **(Supplementary Figs. 4g)** and does not simply recover lineage-specific genes based on the loss of H3K27me3 during development **(Supplementary Figs. 4e,f)**. Compared to alternative approaches, TRIAGE provides a powerful new strategy to recover unique regulators of cell identity that is largely non-overlapping with previous tested strategies (**Supplementary Fig. 4i**).

Identifying cell type-specific regulatory genes from any chordate somatic cell type

Regulatory genes underlying cell identity during development are 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*) **(Fig. 3g)**. 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 **(Fig. 3h).** 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**.**

Dissecting 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[25]. 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

requires no external reference points and therefore provides a distinctive advantage for identifying regulatory control mechanisms one cell transcriptome at a time.

To illustrate this, we analyzed 43,168 cells captured across a 30 day time-course of *in vitro* cardiac-directed differentiation from human pluripotent stem cells (hPSCs)[26]. Analysis of day-30 cardiomyocytes using standard expression data show that high abundance genes are dominated by housekeeping and sarcomere genes, whereas TRIAGE efficiently identifies regulatory genes governing cardiomyocyte identity including *NKX2-5*, *HAND1*, *GATA4*, *IRX4* within the top 10 most highly ranked genes **(Figs. 4a,b).** Importantly, TRIAGE retains highly expressed cell-specific structural genes providing an integrated readout of genes involved in cell regulation and function **(Fig. 4c)**. We used TRIAGE to convert the genes-by-cells matrix comprising ten different subpopulations spanning developmental stages including gastrulation, progenitor and definitive cell types **(Fig. 4d)**. 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 **(Fig. 4e, Supplementary Fig. 5)**. 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 **(Fig. 4f).**

Predicting regulatory drivers of cell identity using any genome-wide analysis of gene expression

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[27], TRIAGE consistently enriches for cell type-specific regulatory genes compared to original expression with no difference between droplet and smartseq2 data sets **(Fig. 4g and Supplementary Table 5).** Using the mouse organogenesis cell atlas (MOCA), which is among one of the largest single cell data sets generated to date[25], we demonstrated that TRIAGE outperformed the expression value 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, **Fig. 4h).** Lastly, we used benchmarking data to evaluate clustering accuracy based on ground truth[28] 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 **(Fig. 4i).**

We hypothesized that TRIAGE could be used to study any genome-wide quantitative measurement of gene expression. To test this, TRIAGE was applied using diverse quantitative readouts of gene expression across hundreds of different cell types. TRIAGE vastly outperforms original abundance metrics when measuring chromatin methylation for H3K36me3, a surrogate of RNA polymerase II activity deposited across gene bodies[6] collected from the 111 Roadmap samples **(Fig. 4j)**. 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 **(Fig. 4j and Supplementary Table 1)**[29]. 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[30] **(Fig. 4j)**. Taken together, these data illustrate the power of utilizing TRIAGE to predict regulatory drivers of cell states using diverse genome-wide multi-omic endpoints.

Determining the regulatory control points of disease

Strategies for identifying genetic determinants of disease have the potential to guide strategies for predicting or altering the natural course of disease pathogenesis. We analyzed genetic data from melanoma and heart failure (HF) pathogenesis to determine the utility of TRIAGE in identifying regulatory determinants of disease.

Treatment for melanoma has improved with the advent of drugs targeting proliferative cells, but highly metastatic and drug resistance subpopulations remain problematic. To assess the potential for TRIAGE

for informing disease mechanisms, we analyzed single cell RNA-seq data from 1,252 cells capturing a transition from proliferative to invasive melanoma[31]. Among the top ranked genes, TRIAGE consistently outperforms expression in prioritizing genes with known involvement in melanoma proliferation and invasion **(Fig. 5a)**. Using independently derived positive gene sets for proliferative versus invasive melanoma[31,32], TRIAGE recovers with high sensitivity the genetic signatures of these two cancer states **(Fig. 5b)**. Gene set enrichment analysis using TRIAGE identified *ETV5* and *TFAP2A* associated with proliferative melanoma versus *TFAP2C* and *TBX3* as regulators of invasive melanoma **(Fig. 5c and Supplementary Table 6).** *TFAP2A* and *TBX3* have been implicated in proliferative and invasive melanoma respectively[33,34], whereas *ETV5* and *TFAP2C* were novel predicted regulators. To validate this, we used *in vitro* nutrient restriction of melanoma cells to trigger a transition into an invasive phenotype[35]. In contrast to expression dynamics of *MITF*, a master regulator of melanocytic differentiation, and *TFAP2C* is upregulated together with *AXL*, a receptor tyrosine kinase associated with therapeutic resistance and transition to invasive melanoma **(Figs. 5d,e)**. Further genetic studies are required to determine the functional role of *TFAP2C* in governing invasive melanoma.

We next aimed to assess whether TRIAGE could identify transcriptional signatures of therapeutic interventions in heart failure (HF). Previous studies have shown that the epigenetic reader protein 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*[36,37] to prevent and treat HF by targeting gene programs linked to cardiac hypertrophy and fibrosis[37]. We analyzed RNA-seq data from adult mouse hearts where pre-established HF (transverse aortic constriction, TAC) was treated with JQ1. TRIAGE prioritized TFs and regulatory genes with known roles in HF pathogenesis **(Fig. 5f)**, outperforming expression ranked genes based on stress-associated gene sets **(Fig. 5g)**. Importantly, comparison between 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 **(Fig. 5g)**. Collectively, these data demonstrate the use of TRIAGE as a scalable strategy for studying the mechanistic basis of disease aetiology and therapy.

Identification of novel regulatory drivers of development

Lastly, we set out to demonstrate that TRIAGE can facilitate discovery of novel regulatory genes governing development *in vitro* and *in vivo*. Using data from single cell analysis of cardiac differentiation[26] we analyzed mesendoderm sub-populations at day 2. TRIAGE identified known regulatory genes governing sub-population identity among the top 10 highly ranked genes **(Fig. 6a)**. Among the TRIAGE identified genes was *SIX3*, a member of the sine oculis homeobox transcription factor family (RTS=0.54) **(Figs. 6a,b)**. Importantly, all pairwise differential expression analyses failed to enrich for *SIX3* **(Supplementary Fig. 6a).** Though the role of *SIX3* in neuroectoderm specification has been studied extensively, little is known about its role in other germ layer derivatives[38-40]. 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 **(Fig. 6c)** with enrichment of *SIX3*$^+$ cells associated with definitive endoderm **(Supplementary Figs. 6b,c**). Using previously published laser microdissection approaches, we captured the spatiotemporal transcriptional data from germ layer cells of mid-gastrula stage (E7.0) embryos[41], with an expanded analysis to include pre- (E5.5-E.6.0), early- (E6.5) and late-gastrulation (E7.5) mouse embryos (**Supplementary Fig. 6f**). Spatio-temporal expression of *SIX3* and other family members is observed in the epiblast and neuroectoderm, (**Fig. 6d, Supplementary Fig. 6g)** consistent with its known role in these lineages[38-40], as well as early endoderm lineages (**Fig. 6d**). Supporting this finding, *SIX3* has been identified as a gene distinguishing definitive from visceral endoderm[42] but no functional studies have validated this finding.

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 **(Figs. 6e,f)**. Cells were differentiated using monolayer cardiac differentiation and analyzed at day 2 **(Fig. 6g).** *SIX3* loss-of-function depleted endoderm and mesendoderm genes **(Fig. 6h**) consistent with FACs analysis showing depletion of CXCR4$^+$/EPCAM$^+$ endoderm cells **(Figs. 6i-k, Supplementary Fig. 6d, Supplementary Fig. 7)**. 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 **(Figs. 6l-n, Supplementary Fig. 6e).** 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.* 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 **(Fig. 6o)**[43]. The top ranked genes based on TRIAGE were analyzed **(Fig. 6p)**. *RNF220* (RTS=0.30, **Fig. 6q**), an E3 ubiquitin ligase governing Wnt signaling pathway activity through β-catenin degradation[44], 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. Collectively, these data illustrate that TRIAGE efficiently identifies novel functional regulatory determinants as a demonstration for discovering novel biology underlying mechanisms of development.

DISCUSSION

Understanding the genetic determinants of cell diversity is essential for establishing mechanisms of development, disease etiology and organ regeneration, as well as synthetic control of cell states including cell reprogramming. Recent advances in deriving genome-wide data at single cell resolution[25,27] as well as computational analysis and prediction algorithms[1,4,5,13,45] have revolutionized our capacity to study complex biological systems and understand the genomic basis of mechanisms governing cell identity[1,2,7,8].

Here, we take advantage of the logic that the selective absence of broad H3K27me3 domains at promoters across diverse cell states establishes a fixed metric that predicts cell identity genes. By drawing on the diverse cell types and epigenetic states recorded in consortium data sets, such as that hosted by the NIH Roadmap Epigenomics project[14], we calculate the direct relationship between core histone methylation

signatures and cell-specific regulatory genes. TRIAGE was designed to provide a *scalable* prediction metric and therefore focused on the main (and most confident) association between H3K27me3 domains and their protein-coding genes. Furthermore, focusing on protein coding genes provided a means for expanding its utility to cross-genome generalisations of the score by mapping to orthologous genes in other species. From this we define a, surprisingly simple, repressive tendency score based on H3K27me3 broad domains to identify cell type-specific regulatory genes underlying cell diversification in development and disease. We demarcate a priority gene-set of 1,359 genes consistently regulated by broad H3K27me3 domains across major cell lineages; we find that they represent key cell type specific drivers of germ layer diversification in development and control points of disease pathogenesis across diverse species.

We show that the genome-wide repressive tendency score of a gene can interface with any customized input data quantitatively measuring transcriptional activity of a cell including RNA-seq, proteomics, CAGE-seq and H3K36me3 tag density to effectively capture cell type-specific regulatory genes. In contrast to previous inference strategies[4,5], TRIAGE does not utilize comparative reference points nor any external metadata for its predictions and can be applied to data from theoretically any somatic cell type and animal phyla. In fact, TRIAGE performs convincingly against alternative formulations based on H3K4me3 domain breadth[1], super-enhancers[8,12,24], differential gene expression and functional heterogeneity (FH) score[2] across different biological contexts. We build on the prevailing view that epigenetic states guide regulatory processes; we note that while TRIAGE incorporates a single repressive mark mapped to transcription start sites of protein coding genes, it effectively identifies regulatory drivers with high sensitivity and precision.

While sufficiently diverse data sets on epigenetic control of cell states are currently available only for human and mouse, we show that the evolutionary conservation of gene regulation enables this quantitative strategy to predict regulatory genes across diverse species in the animal kingdom. We hypothesize that this approach can be applied across diverse cell and tissue types in species where gene

expression is governed by the polycomb group complex-2 (PRC2). While not perfectly conserved through evolution, PRC2 and its regulation of histone methylation are known to govern genes in protists, animals, plants, as well as fungi[46]. Indeed, zebrafish[47] and medaka[9] 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[10,48]. We note that cross-species applications are limited by gene mapping tools and the composition of cellular diversity that distinguishes animal phyla.

The application of TRIAGE to single cell resolution genomic data yields novel insights into the regulatory basis of heterogenous cell populations. With each transcriptome evaluated independently, TRIAGE effectively prioritizes cell-specific regulators of identity including TFs and signaling molecules, despite their low abundance in transcriptomics analyses. Importantly, TRIAGE prioritization of regulatory mechanisms governing cell states enables effective prediction of drivers of cell identity in a population-specific manner, providing a basis for cell type identification and an effective strategy for novel biological discovery.

The conservation of epigenetic regulatory logic of the genome provides an effective strategy for utilizing large, diverse genome-wide data to establish quantitative principles of cell states to infer cell type-specific mechanisms that explain the complexity of biological systems. We anticipate that this analytic approach can render customized inference predictions, based on chromatin transition, between diverse healthy and diseased tissues to reveal stress-sensitive loci and novel disease drivers. This conceptual and experimental framework can infer regulatory genes governing theoretically any cell state and has broad utility for studies in genome regulation of cell identity in health and disease.

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[14]. We used identifiers from Animal TFDB and DBD to identify 1,605 human TFs[49,50]. We also identified 3,804 housekeeping genes from the published literature[17].

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)[13]. We also used GTEx expression profiles of samples that belongs to the 5 tissue-groups[51]. 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[52]. 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 (**Supplementary Table 2**). 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) (**Supplementary Table 2**). This classification is based on previous observations that expression of actively regulated genes is highly variable both temporally and spatially[15]. 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 **(Fig. 1b)** as well as the expressional specificity **(Fig. 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[14,16]. The *relative* expression is calculated as

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

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, $D_{i,j} = \{d_1, d_2, \ldots, d_l, \ldots\}$ 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[53]. We removed any intergenic domains that did not overlap any proximal regions of genes to reduce potential bias of including putative enhancers[54].

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 (**Supplementary Fig. 2d**). 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 (**Fig. 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[1], (ii) breadth of H3K27me3 peaks at the gene locus multiplied by the corresponding expression level in a spatially heterogenous cell population[2] and (iii) super-enhancer (SE)[8,24] 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[1]. 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[55]. 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[14].

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[2]. 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)[2]. 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[56]. 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) **(Supplementary Fig. 4d)**.

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[8,12,57,58]. SEs have been studied as a genomic hot-spot to understand causal determinant of cell type-specific transcriptions in development and disease[57,58]. 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)[59]. We downloaded lists of SEs for 5 selected tissue-types from a published SE database SEdb (http://www.licpathway.net/sedb/index.php)[24]; (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[8]. 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 Astrocyte (*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[26] as well as selected bulk RNA-seq data for 3 distinct tissue groups (i.e. Blood, Brain and Heart) from the Roadmap project[14].

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[26]. 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[10,11]. 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)[3].

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 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.

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[60-62]. 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[62]. 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[60]. 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[60]. 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 (θ) 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 θ. 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 θ 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 (**Supplementary Figs. 2a and 2b**). 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[6] 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 **(Fig. 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 **(Fig. 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[28]. 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)[63-65]. 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)[66], 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)[65]

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[25-27], FANTOM5 cap analysis of gene expression (CAGE) peaks[29], human proteomes[30] and Roadmap H3K36me3 ChIP-seq data sets[14].

We applied TRIAGE to tabula muris mouse single-cell RNA-seq data sets for nearly 100,000 cells from 20 different tissue types[27]. 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[25]. 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 **(Supplementary Table 1).** 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[6]. 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[46], 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 (**Supplementary Table 1**). We then performed inter-species gene mapping using online Ensembl bioMart (http://asia.ensembl.org/biomart/martview/) by identifying human orthologues[67]. 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[31].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[31]. False discovery rate was used to correct for multiple hypothesis testing.

Melanoma regulator prediction. The ratio of the mean expression of the proliferative to invasive genes was used to order the melanoma gene expression profiles. For each of the melanoma single cell gene expression profiles, TRIAGE was applied to rank the genes according to their DS. Genes were identified which appeared in the top 50 of the ranked gene expression profiles at least 30 times. For each melanoma state, genes were ranked by frequency of appearing in the top 50 **(Supplementary Table 6)**. A modified gene set enrichment analysis was performed to determine if a given gene appearing within the top 50 of the expression profiles was enriched toward either end of the proliferative to invasive ordering. The enrichment score was calculated going from proliferative to invasive cells. The score was incremented when a given gene appeared in the top 50 of the ranked genes, and decremented otherwise, as per the weighting scheme described previously[68]. The largest absolute deviation from 0 was the final enrichment score. To determine the significance of the score 1,000 permutations of randomly ordering the expression profiles and re-calculating the enrichment score was used as the background. Bonferroni correction was used to correct for multiple hypothesis testing. For comparison, equivalent analyses were run for genes ranked according to absolute expression.

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)[37]. 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[41]. 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 **Supplementary Fig. 5f**). 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[69] 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[70]. 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** **5' – Forward Primer – 3'** **5' – Reverse Primer – 3'**

SIX3 gRNA1	F 5' TTGGGCTGAATCTTGACTCGGCGG 3'
	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[71]. 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).

Human melanoma cell lines A2058, MM96L and HT144 were cultured in RPMI 1640 media (Invitrogen-Life Technologies) supplemented with 5% foetal bovine serum, 2mM L-glutamine, and 5 mg/ml penicillin/streptomycin. Cells under glutamine deprivation were grown in RPMI 1640 media without

supplementation of L-glutamine while cells under glucose deprivation were grown in RPMI 1640 glucose-free media supplemented as described above.

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

RNA was extracted from melanoma cells using Trizol extraction (Sigma Aldrich) and first-strand cDNA prepared by reverse transcription using using iSCRIPT RT Supermix (Bio-rad). Gene expression was quantified using SYBR Green or Taqman mastermixes (Applied Biosystems) on a ViiA7 real time cycler (Applied Biosystems). SYBR based qRT-PCR primers used in the study were β-2-microglobulin (Fwd 5'CATTCGG5'GCCGAGATGTCT, Rev 5'CTCCAGGCCAGAAAGAGAGAGTAG), TFAP2C (Fwd 5' CTGTTGCTGCACGATCAGACA; Rev 5' CTCAGTGGGGTTCATTACGGC) and AXL (Fwd 5' TTTCCTGAGTGAAGCGGTCT; Rev 5' TCGTTCAGAACCCTGGAAAC). Taqman primers were β-2-microglobulin (Hs00187842_m1) and MITF (Hs01117294_m1) (Applied Biosystems). Comparative CT analysis was used to determine relative mRNA expression relative to the β-2-microglobulin housekeeping control gene. Relative expression data for each gene between the 0hr and 36/48 hour time points was transformed to z-scores and used to generate heat maps.

Cloning and lentivirus production. The open reading frame of the human TFAP2C gene was cloned into the pLVX-Tight-puro construct (Clonetech) and used to generate Lentivirus using the Lenti-X HT packaging system (Clontech) according to the manufacturer's protocol. Melanoma cell lines were transduced and selected in puromycin as described previously[72].

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[73]. 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[74]. 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.

Data availability. The source data underlying Figs 1,2,3 and 6 and Supplementary Fig. 2 are provided as a source data file. All relevant data are available from the authors upon request.

Code availability. 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 APP1143163) 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

YC: 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.

FIGURES

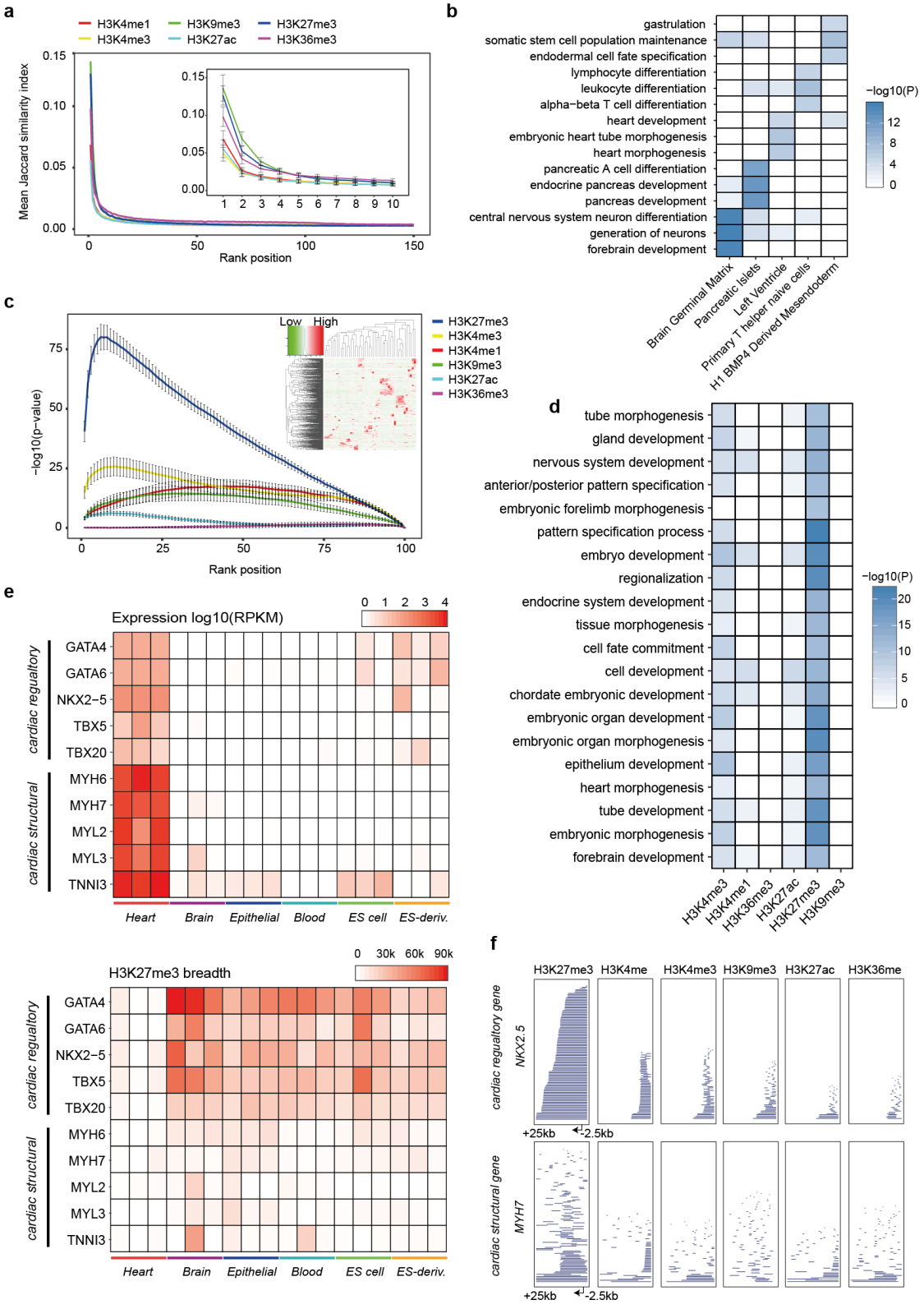

Fig. 1: Broad H3K27me3 domains are associated with cell type-specific regulatory genes.

(a) Broad HM domains identify a set of common genes. For each HM type, genes are ranked by the breadth of the associated HM domain across 111 Roadmap cell types and grouped into bins of 100 genes (*x*-axis). Mean similarity between cell types of the same rank bin is shown (*y*-axis). **(Inset)** Top 100 genes (i.e. first rank bin) are significantly more similar (p<2.2e-16 for all HMs, Wilcoxon rank-sum test, one-tailed) than genes with narrower domains. Scale bars show the 95% confidence interval. Source data are provided as a Source Data file.

(b) Enrichment of tissue-type-specific gene ontology (GO) biological process (BP) terms associated with most highly expressed 50 VETFs in 5 different cell types (Fisher's exact test, one-tailed); Brain germinal matrix (E070), Pancreatic islets (E087), Left ventricle (E095), Primary T helper naïve cells (E038) and H1 BMP4-derived mesendoderm (E004).

(c) 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). Source data are provided as a Source Data file. **(Inset)** Expression levels of the 634 VETFs (row) across the 46 Roadmap samples (column).

(d) 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).

(e) Gene expression level (top) and the H3K27me3 breadth (bottom) 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).

(f) 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.

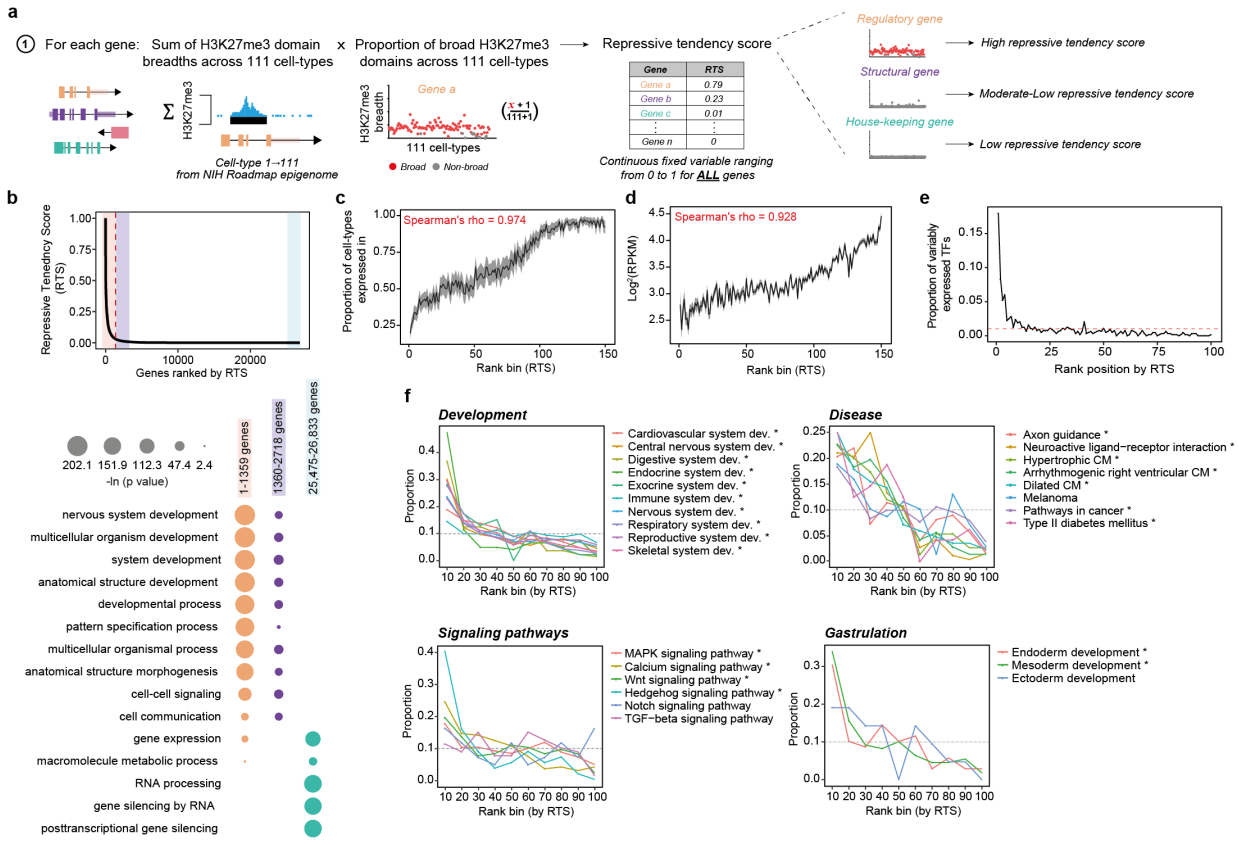

Fig. 2: Genes that are frequently associated with broad H3K27me3 domains regulate diverse developmental and disease processes.

(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) (Supplementary Table 4 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. Source data are provided as a Source Data file.

(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). Source data are provided as a Source Data file.

(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 a 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 (Supplementary Table 4 for the full list). Each rank bin includes 1% of all RefSeq genes with a 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).

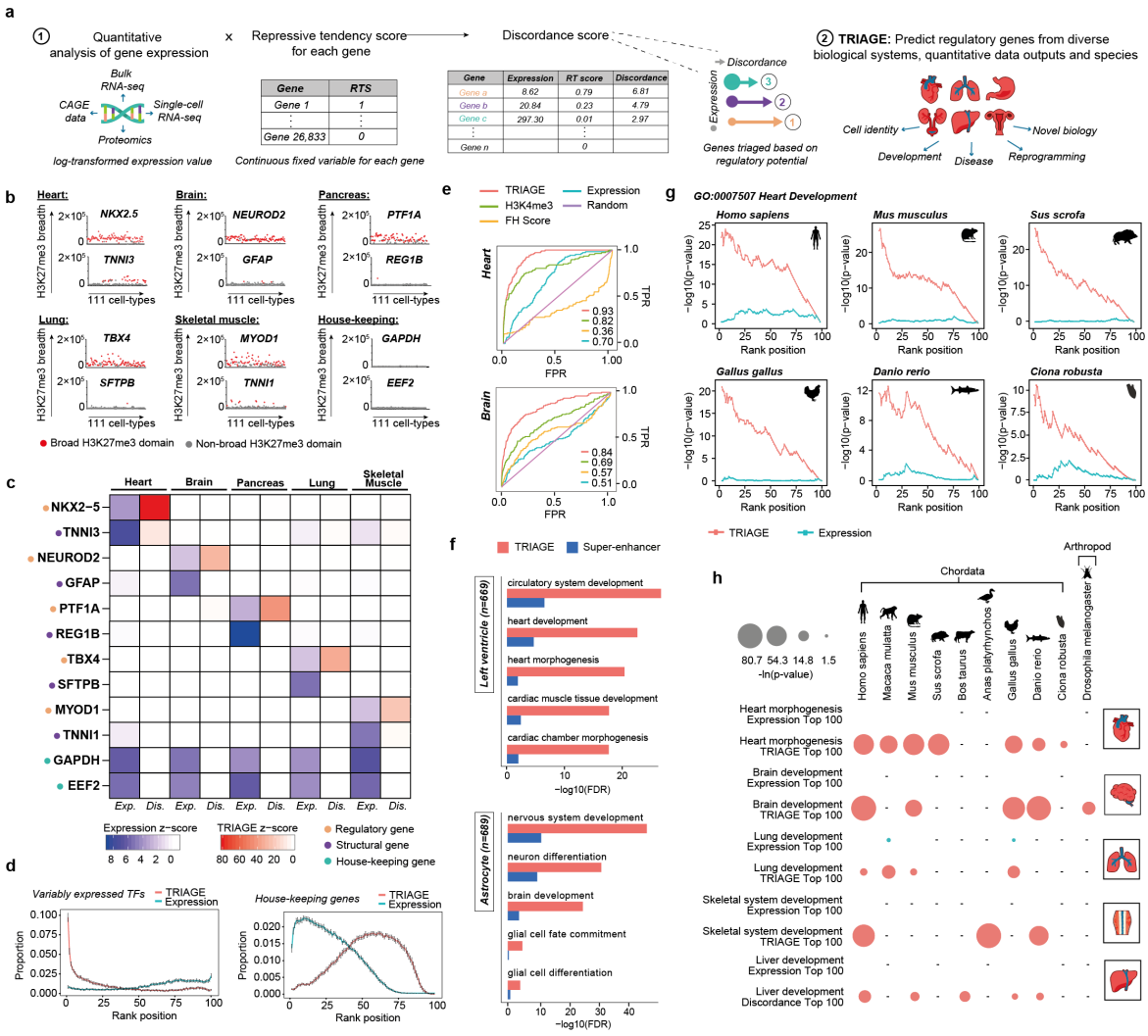

Fig. 3: TRIAGE effectively infers cell type-specific regulatory genes of somatic cell types across diverse species of the animal kingdom.

(a) TRIAGE transforms readouts of gene expression into discordance score (DS) to prioritize genes with important regulatory potential.

(b) Breadths of H3K27me3 domains (in base-pairs) associated with selected cell 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 given 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 after the transformation (p<2.2e-16 for top 1% genes, Wilcoxon rank-sum test, one-tailed). Scale bars show the 95% confidence interval. Source data are provided as a Source Data file.

(e) Receiver-operating characteristic (ROC) plots comparing performance of TRIAGE against existing methods on Roadmap heart (E095) and brain (E070) samples. Area under the curve (AUC) values are shown on the right bottom corner of the plot. For more extensive performance analysis (including other tissue types), see **Supplementary Fig. 4 and Supplementary Table 7** for comparison between TRIAGE and SE-based approach.

(f) TRIAGE effectively identifies genes with cell type-specific development functions on left ventricle (E095) and astrocyte (E125) Roadmap samples, with a greater sensitivity than super-enhancer (SE)-based approach. Enrichment of a given GO term is compared between (i) *n* number of genes as the nearest active gene to SEs and (ii) top *n* number of genes by the DS (**Supplementary Fig. 4b and Supplementary Table 7** for the full table).

(g) 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 given sample.

(h) 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.

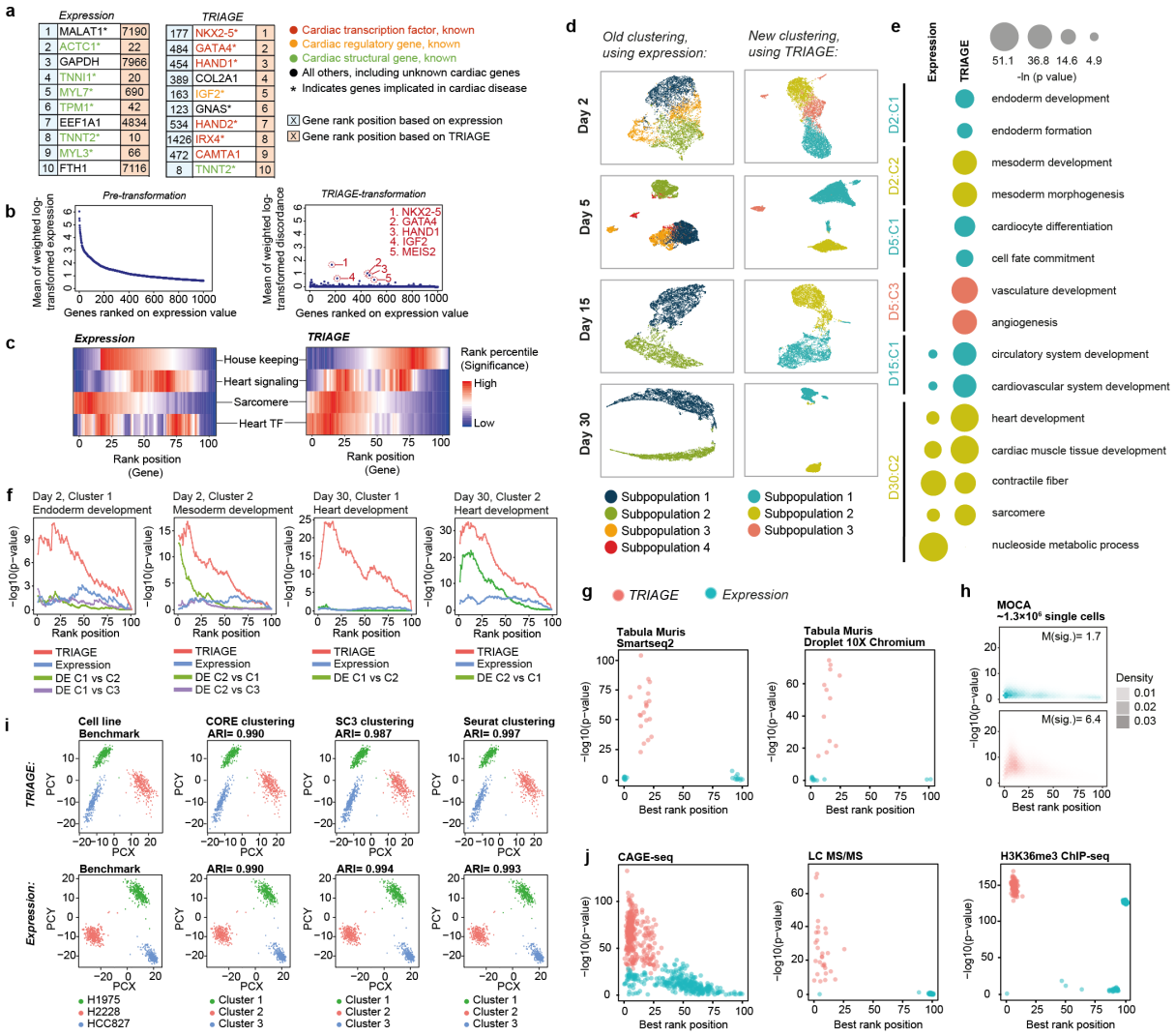

Fig. 4: TRIAGE is highly applicable and scalable for multi-omics datasets.

(a) Top 10 genes identified by expression or DS from hiPSC cardiomyocyte single-cell RNA-seq for *in vitro* cardiac-directed differentiation of day 30[26].

(b) TRIAGE transforms scRNA-seq for *in vitro* differentiation of cardiomyocytes at day 30 (left) to reveal known cardiac regulators (right).

(c) TRIAGE-transformation shifts distributions of different functional genes. Enrichment of a gene set with 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) in *in vitro* differentiated cardiomyocytes at day 30. Genes are sorted by either the expression value (left) or the DS (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 identified by TRIAGE but not original expression. 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,j) TRIAGE is applicable to **(g)** different scRNA-seq sequencing platforms (i.e. Smart-seq2 or Droplet 10X chromium, **(h)** the mouse organogenesis cell atlas (MOCA) data as well as **(j)** various readouts of the gene expression (i.e. CAGE-seq, proteome and H3K36me3 tag density). Genes are sorted by either expression value or DS and grouped into a percentile bin. Enrichment of VETFs for each sample is summarized by the most significant *p*-value (*y*-axis) the corresponding rank bin position (*x*-axis).

(i) Comparison of scRNA-seq cluster assignment efficiency between original expression and TRIAGE analyzed data using Mixology data sets[28].

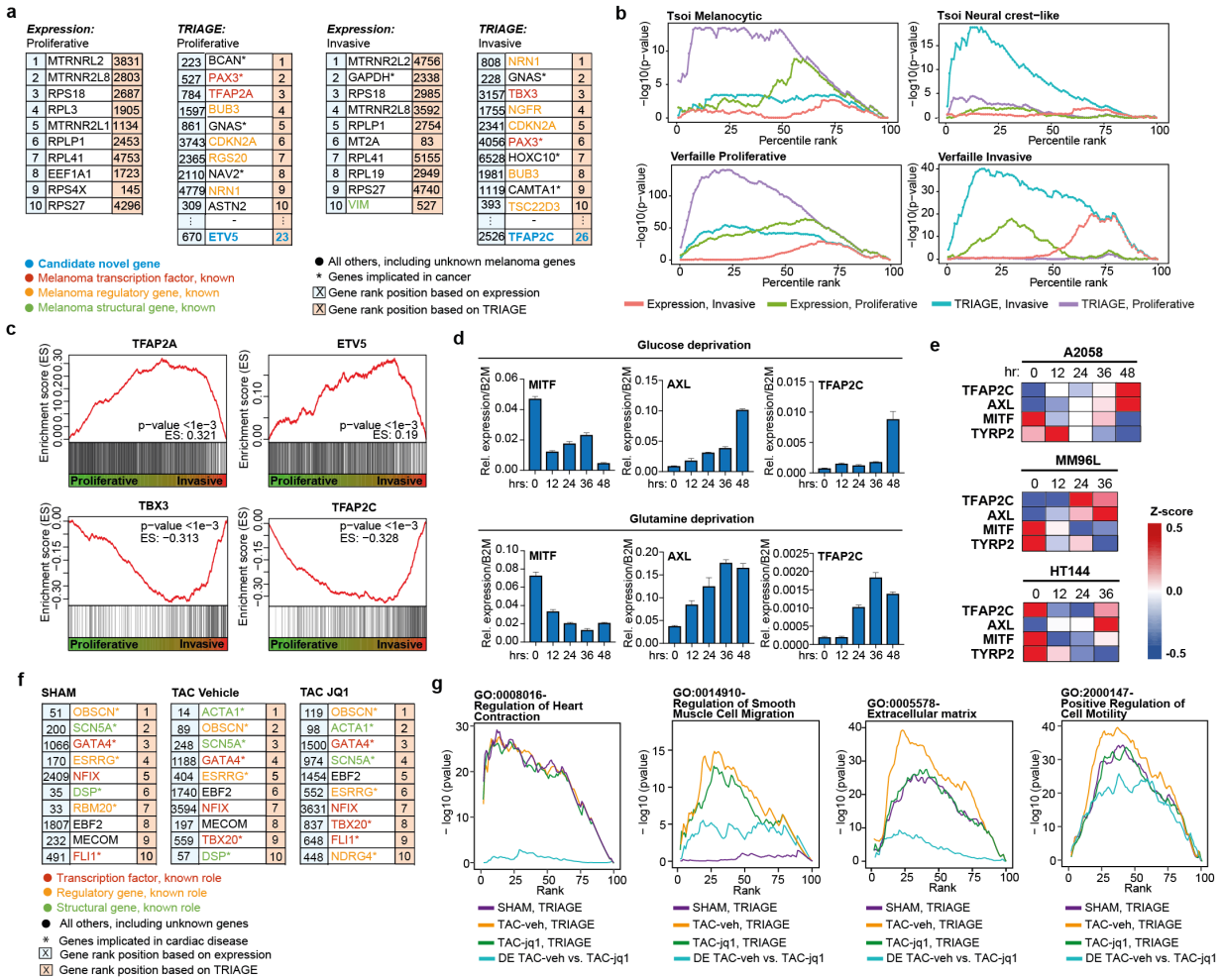

Fig. 5: Determining the regulatory basis of disease pathogenesis and therapy.

(a) Tables showing the top ranked genes from proliferative melanoma cells or invasive melanoma cells indicating rank position by original expression (left) or TRIAGE (right). Genes are identified based on their known roles as structural or regulatory genes in melanoma.

(b) Fisher's exact test enrichment of positive gene sets for proliferative and invasive melanoma states demonstrating high specificity of enrichment for cell type-specific gene signatures only with TRIAGE.

(c) Gene set enrichment analysis (GSEA) for *ETV5*, *TFAP2A*, *TBX3* and *TFAP2C*. The y-axis corresponds to the enrichment score with gene expression profiles ranked by TRIAGE. The x-axis shows cells ranked from proliferative to invasive. The vertical lines indicate when the respective gene was found in the top 50 of a ranked expression profile.

(d) qPCR analysis showing changes in expression of *MITF*, *AXL* and *TFAP2C* in A2058 melanoma cells over 48hours of glucose deprivation (top) and glutamine deprivation (bottom).

(e) *z*-score based heat maps showing changes in expression of melanoma genes in three individual BRAF mutant melanoma cell lines over 36-48 hours of glucose starvation.

(f) Top 10 genes ranked by expression (left) or TRIAGE (right) from SHAM, TAC-vehicle and TAC-JQ1 data from bulk RNA-seq of mice subjected to sham surgery (SHAM), transverse aortic constriction (TAC-vehicle) and TAC treated with JQ1 (TAC-JQ1).

(g) Enrichment of GO terms associated with cardiac biology and heart failure stress response mechanisms comparing each condition analyzed by TRIAGE or DE analysis (TAC-veh vs TACJQ1). Genes are ranked by either the expression value or TRIAGE and binned by rank (each bin includes 1% of all genes) and the enrichment is calculated at each rank position (y-axis, Fisher's exact test, one-tailed).

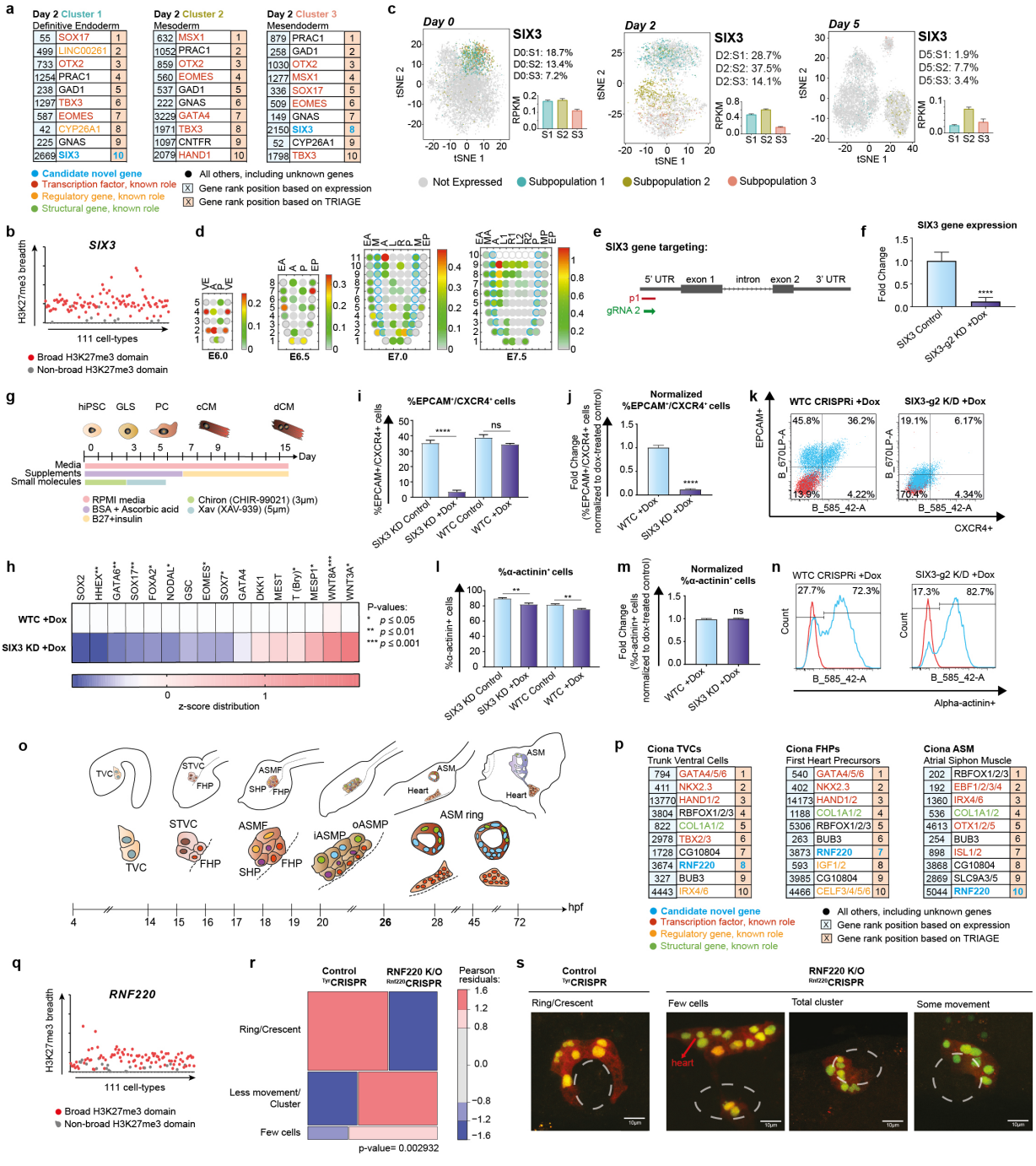

Fig. 6: Predicting novel regulators of heart development.

(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 Epigenomes 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 knockdown of *SIX3* in iPSCs.

(f) qPCR analysis of *SIX3* transcript abundance in control vs *SIX3* CRISPRi KD iPSCs. Source data are provided as a Source Data file.

(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) Cells were phenotyped on Day 2 of differentiation for endoderm markers by FACs analysis of 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. Source data are provided as a Source Data file.

(l-n) Analysis of cardiomyocytes by FACs for α-actinin. **(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 a-actinin analysis. Source data are provided as a Source Data file.

(o) Schematic overview of cardiac development in *Ciona* from 4 to 72 hours post fertilization (hpf) at 18ºC. Adapted from[75]. 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 DS (right) from populations found during *Ciona* heart development *in vivo*, highlighting *RNF220* as a candidate novel gene identified by TRIAGE.

(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

1 Benayoun, B. A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673-688, doi:10.1016/j.cell.2014.06.027 (2014).

2 Rehimi, R. *et al.* Epigenomics-Based Identification of Major Cell Identity Regulators within Heterogeneous Cell Populations. *Cell Rep* **17**, 3062-3076, doi:10.1016/j.celrep.2016.11.046 (2016).

3 Paige, S. L. *et al.* A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* **151**, 221-232, doi:S0092-8674(12)01058-6 [pii] 10.1016/j.cell.2012.08.027 (2012).

4 Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903-915 (2014).

5 Rackham, O. J. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nature genetics* **48**, 331 (2016).

6 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).

7 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279 (2011).

8 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).

9 Nakamura, R. *et al.* Large hypomethylated domains serve as strong repressive machinery for key developmental genes in vertebrates. *Development* **141**, 2568-2580, doi:10.1242/dev.108548 (2014).

10 Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353, doi:10.1038/nature04733 (2006).

11 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).

12 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

13 Palpant, N. J. *et al.* Chromatin and Transcriptional Analysis of Mesoderm Progenitor Cells Identifies HOPX as a Regulator of Primitive Hematopoiesis. *Cell Rep* **20**, 1597-1608, doi:10.1016/j.celrep.2017.07.067 (2017).

14 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

15 Perez-Lluch, S. *et al.* Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat Genet* **47**, 1158-1167, doi:10.1038/ng.3381 (2015).

16 Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* **6**, R33, doi:10.1186/gb-2005-6-4-r33 (2005).

17 Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569-574, doi:10.1016/j.tig.2013.05.010 (2013).

18 Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

19 Zhong, Y. F. & Holland, P. W. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev* **13**, 567-568, doi:10.1111/j.1525-142X.2011.00513.x (2011).

20 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30, doi:10.1093/nar/28.1.27 (2000).

21 Grote, P. & Herrmann, B. G. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol* **10**, 1579-1585, doi:10.4161/rna.26165 (2013).

22 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

23 Scornavacca, C., Zickmann, F. & Huson, D. H. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics* **27**, i248-256, doi:10.1093/bioinformatics/btr210 (2011).

24 Jiang, Y. *et al.* SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* **47**, D235-D243, doi:10.1093/nar/gky1025 (2019).

25 Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502, doi:10.1038/s41586-019-0969-x (2019).

26 Friedman, C. E. *et al.* Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. *Cell stem cell* **23**, 586-598. e588 (2018).

27 Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).

28 Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* **16**, 479-487, doi:10.1038/s41592-019-0425-8 (2019).

29 Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-470, doi:10.1038/nature13182 (2014).

30 Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581, doi:10.1038/nature13302 (2014).

31 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189, doi:10.1126/science.aad0501 (2016).

32 Verfaillie, A. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* **6**, 6683, doi:10.1038/ncomms7683 (2015).

33 Rambow, F. *et al.* New Functional Signatures for Understanding Melanoma Biology from Tumor Cell Lineage-Specific Analysis. *Cell Reports* **13**, 840-853, doi:10.1016/j.celrep.2015.09.037 (2015).

34 Peres, J. & Prince, S. The T-box transcription factor, TBX3, is sufficient to promote melanoma formation and invasion. *Molecular cancer* **12**, 117-117, doi:10.1186/1476-4598-12-117 (2013).

35 Falletta, P. *et al.* Translation reprogramming is an evolutionarily conserved driver of phenotypic plasticity and therapeutic resistance in melanoma. *Genes Dev* **31**, 18-33, doi:10.1101/gad.290940.116 (2017).

36 Anand, P. *et al.* BET bromodomains mediate transcriptional pause release in heart failure. *Cell* **154**, 569-582, doi:10.1016/j.cell.2013.07.013 (2013).

37 Duan, Q. *et al.* BET bromodomain inhibition suppresses innate inflammatory and profibrotic transcriptional networks in heart failure. *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aah5084 (2017).

38 Lagutin, O. V. *et al.* Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. *Genes & development* **17**, 368-379 (2003).

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

40 Steinmetz, P. R. *et al.* Six3 demarcates the anterior-most developing brain region in bilaterian animals. *EvoDevo* **1**, 14 (2010).

41 Peng, G. *et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev Cell* **36**, 681-697, doi:10.1016/j.devcel.2016.02.020 (2016).

42 Sherwood, R. I. *et al.* Prospective isolation and global gene expression analysis of definitive and visceral endoderm. *Developmental Biology* **304**, 541-555, (2007).

43 Wang, W. *et al.* A single cell transcriptional roadmap for cardiopharyngeal fate diversification. *BioRxiv* (2019).

44 Ma, P. *et al.* The ubiquitin ligase RNF220 enhances canonical Wnt signaling through USP7-mediated deubiquitination of beta-catenin. *Mol Cell Biol* **34**, 4355-4366, doi:10.1128/MCB.00731-14 (2014).

45 Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**, 364-376, doi:10.1038/nbt.3157 (2015).

46 Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343-349, doi:10.1038/nature09784 (2011).

47 Wu, S. F., Zhang, H. & Cairns, B. R. Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. *Genome Res* **21**, 578-589, doi:10.1101/gr.113167.110 (2011).

48 Fujikura, J. *et al.* Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev* **16**, 784-789, doi:10.1101/gad.968802 (2002).

49 Wilson, D., Charoensawan, V., Kummerfeld, S. K. & Teichmann, S. A. DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* **36**, D88-92, doi:10.1093/nar/gkm964 (2008).

50 Zhang, H. M. *et al.* AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**, D76-81, doi:10.1093/nar/gku887 (2015).

51 Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

52 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

53 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

54 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).

55 Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* **17**, 72, doi:10.1186/s13059-016-0944-x (2016).

56 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

57 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).

58 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

59 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

60 Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**, 1351-1359, doi:10.1038/nbt.1508 (2008).

61 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

62 Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-1740, doi:10.1038/nprot.2012.101 (2012).

63 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411, doi:10.1038/nbt.4096 (2018).

64 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14**, 483, doi:10.1038/nmeth.4236 (2017).

65 Senabouth, A. *et al.* ascend: R package for analysis of single-cell RNA-seq data. *Gigascience* **8**, doi:10.1093/gigascience/giz087 (2019).

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

67 Haider, S. *et al.* BioMart Central Portal--unified access to biological data. *Nucleic Acids Res* **37**, W23-27, doi:10.1093/nar/gkp265 (2009).

68 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

69 Mandegar, M. A. *et al.* CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell Stem Cell* **18**, 541-553, doi:10.1016/j.stem.2016.01.022 (2016).

70 Palpant, N. J. *et al.* Generating high-purity cardiac and endothelial derivatives from patterned mesoderm using human pluripotent stem cells. *Nat Protoc* **12**, 15-31, doi:10.1038/nprot.2016.153 (2017).

71 Burridge, P. W. *et al.* Chemically defined generation of human cardiomyocytes. *Nat Methods* **11**, 855-860, doi:10.1038/nmeth.2999 (2014).

72 Jagirdar, K. *et al.* The NR4A2 nuclear receptor is recruited to novel nuclear foci in response to UV irradiation and participates in nucleotide excision repair. *PLoS One* **8**, e78075, doi:10.1371/journal.pone.0078075 (2013).

73 Stolfi, A., Gandhi, S., Salek, F. & Christiaen, L. Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. *Development* **141**, 4115-4120, doi:10.1242/dev.114488 (2014).

74 Christiaen, L., Wagner, E., Shi, W. & Levine, M. Electroporation of transgenic DNAs in the sea squirt Ciona. *Cold Spring Harb Protoc* **2009**, pdb prot5345, doi:10.1101/pdb.prot5345 (2009).

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