Comparative analysis of diverse cell states establishes an epigenetic basis for inferring regulatory 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], Deepak Srivastava[4], Lionel Christiaen[3], Quan Nguyen[2], Mikael Bodén[1,14], Nathan J. Palpant[2,11,14]

[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

[14] 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

Understanding genetic control of cell diversification is essential for establishing mechanisms controlling biological complexity. This study demonstrates that the a priori deposition of H3K27me3 associated with gene repression across diverse cell states provides a genome-wide metric that enriches for genes governing fundamental mechanisms underlying biological complexity in differentiation, morphogenesis, and disease. We use this metric in combination with more than 1 million genome-wide data sets from diverse omics platforms to identify cell type specific regulatory mechanisms underlying diverse organ systems from species across the animal kingdom. From this analysis, we identify and genetically validate multiple novel genes controlling development in diverse chordates including humans and the tunicate, Ciona robusta. This study demonstrates that the conservation of epigenetic regulatory logic provides an effective strategy for utilizing large, diverse genome-wide data to establish quantitative basic principles of cell states to infer cell-type specific mechanisms that underpin the complexity of biological systems.

Introduction

Capturing the information basis of a cell through genome-wide sequencing is a powerful mechanism for understanding the complexities of development and disease. However, the information collated is often limited, reflecting only a snapshot of the steady state of the genome. Enhancing the strategies for predicting regulatory determinants of cell identity has proven to be essential for gleaning novel insights into developmental biology, disease mechanisms and cell reprogramming (Benayoun et al., 2014; Cahan et al., 2014; Rackham et al., 2016). Here, we demonstrate an approach to infer regulatory drivers of any cell state, without the requirement of external reference data or prior knowledge, by analyzing the landscape of diverse chromatin states for distinguishing features of cell specificity. We demonstrate that the *a priori* probability that the presence of a broad repressive H3K27me3 histone modification mark, which signifies the repressive tendency across a gene locus in diverse cell states, provides a quantifiable metric that strongly predicts regulatory genes governing mechanisms of cell differentiation and organ morphogenesis in health and disease. We show that the repressive tendency can be used to analyze individual transcriptomes of millions of heterogeneous cells simultaneously to infer the cell type-specific regulatory genes controlling 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 focused on the breadth of histone modifications (HMs) which has been shown to be structurally and functionally linked to cell-specific genome architecture and gene regulation (Barski et al., 2007). We used NIH Epigenome Roadmap data (Kundaje et al., 2015), which contains ChIP-seq data for H3K4me3, H3K36me3, H3K27me3, H3K4me1, H3K27ac and H3K9me3 for 111 tissue or cell types (**Table S1)**. 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 **(Figure 1A)**, however, broad domains of different HMs marked distinct sets of genes **(Figure S1A).** 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 **(Figure 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; this set is comprised of 634 variably expressed transcription factors (TFs) having a coefficient of variation greater than 1 (**Table S2**) and detected in 46 NIH Epigenome RNA-seq data sets (Perez-Lluch et al., 2015). We used Shannon entropy to quantify cell type-specificity (Schug et al., 2005) and demonstrate that variably expressed TFs are significantly more cell type-specific, compared to non-variably expressed TFs or protein coding genes **(Supplementary Methods).** Analysis of RNA-seq data sets from diverse cell and tissue types show that variably expressed TFs in each sample reflect appropriate tissue or cell type-specific regulatory functions **(Figures 1B and 1C inset)**. Henceforth, variably expressed TFs provide a

positive gene set where their enrichment is a performance metric for identifying cell type-specific regulatory genes.

We utilized variably expressed TFs 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 variably expressed TFs. These data show that H3K27me3 uniquely and significantly enriches for variably expressed TFs within the top 5% of broad domains **(Figures 1C, 1D and S1B**). This demonstrates 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 (Eisenberg and Levanon, 2013), we extracted expression and chromatin data from cardiomyocytes **(Figures 1E and 1F)**. We show that the transcript abundance of cardiac regulatory genes (i.e. *GATA4*, *GATA6*, *NKX2-5*, *TBX5* and *TBX20*) and structural sarcomere genes (i.e. *MYH6*, *MYH7*, *MYL2*, *MYL3* and *TNNI3*) are all significantly elevated in cardiac cells compared to other cell types, but cannot be distinguished as regulatory or structural genes except by differential expression **(Figure 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 **(Figures 1E and 1F)**. No other HM analyzed demarcates cell type-specific regulatory genes from structural genes in this manner **(Figures 1F and S1C),** 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 (Boyer et al., 2006; Lee et al., 2006). 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 **(Figure 2A).** This approach quantifies a single value for every gene that defines its association with broad H3K27me3 domains which we call its repressive tendency score (RTS) **(Table S3).** Using the NIH Epigenome Roadmap data, the RTS is calculated for 99.3% (or 26,833 genes) of all RefSeq genes. To demonstrate that 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. Furthermore, we note that the 111 cell types provided sufficient sample size to calculate a stable RTS **(Figures S2A and S2B)**, with a majority of assigned H3K27me3 domains (over 85%) overlapping a single gene **(Figures S2D, S2E, S3A and S3B)**. Importantly, 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 **(Figure 2B),** and show they are cell type-specific (**Figure 2C**) and lowly expressed (**Figure 2D**). Among the 1,359 priority genes, we identified 318 TFs, including variably expressed TFs which had a significantly higher RTS overall (mean=0.083) compared to the background (mean=0.006, **Figure 2E**) in addition to 155 homeobox proteins, 291 non-coding RNAs genes (e.g. FENDRR and HOTAIR (Grote and Herrmann, 2013; Rinn et al., 2007)), and 260 genes involved in cell signaling. 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 **(Figure 2F and Table S4).** Taken together, these data indicated 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 comprises a small fraction of the genome and represents the signature of structural, housekeeping and regulatory genes underlying a cell state. Identifying the regulatory genes controlling the identity, fate and function of a particular cell state is difficult to determine from thousands of expressed genes. 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 distinctive signature of any cell's transcriptomic data with the RTS, a method we call TRIAGE (Transcriptional Regulatory Inference Analysis from Gene Expression). TRIAGE theoretically provides a means to identify cell type-specific regulatory genes for any cell type **(Figure 3A).** 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 reflects the juxtaposition of a gene's association with being epigenetically repressed and the observed transcriptional abundance of that gene in the input data. Collectively, TRIAGE introduces a non-linear, gene-specific weight that prioritizes cell type-specific regulatory genes based on the input expression signature of any cellular state. Of importance, this strategy does not require reference 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 demonstrate TRIAGE, we identified known regulatory and structural genes from 5 tissue groups, analyzing H3K27me3 of cell-specific regulatory versus structural genes **(Figures 3B)**. When applied to cell-specific transcriptional data, TRIAGE reduces the relative abundance of structural and housekeeping genes, while enriching for regulatory genes in a cell type-specific manner **(Figure 3C).** Taken to scale, TRIAGE transformation of all 46 Roadmap cell types results in enrichment of cell type-specific TFs among the top 1% in every cell type. Compared to the expression-based ranking, TRIAGE reduces the relative abundance of housekeeping genes **(Figures 3D and S2C).** Constructing a tanglegram based on the Pearson distances between Roadmap tissue types (Scornavacca et al., 2011)**,** 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 **(Figure S4A)**.

Previous work by Benayoun et al. ranked genes based on broad H3K4me3 domains to enrich for cell type-specific regulatory genes (Benayoun et al., 2014). Using diverse cell and tissue types in which

expression and H3K4me3 data are available, we demonstrate that TRIAGE outperforms original expression and H3K4me3 broad domains in both sensitivity and precision of identifying cell type-specific regulatory genes **(Figures 3E, S4B and S4C)**.

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*) **(Figure 3F)**. In contrast to expression alone, TRIAGE recovered cardiac regulatory genes with high efficiency across all species. More broadly, we used TRIAGE to enrich for relevant tissue morphogenesis biological processes from diverse cell types and species including arthropods **(Figure 3G).** 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 (Cao et al., 2019). Determining mechanistic information from diverse cell states captured using single-cell analytics remains a challenge. TRIAGE is scalable for studies of cell heterogeneity because it 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) (Friedman et al., 2018). 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 **(Figures 4A and 4B).** Importantly, TRIAGE retains highly expressed cell-specific structural genes providing an integrated readout of genes involved in cell regulation and function **(Figure 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 **(Figure 4D)**. In contrast to expression data, which significantly enriches for structural and housekeeping genes, TRIAGE consistently identifies gene sets associated with development of every subpopulation through differentiation **(Figures 4E and Figure S5)**. Lastly, standard -omics analysis pipelines implement differential expression (DE) followed by gene ontology, pathway or network analysis. We show that DE results in variable outcomes depending on the comparison and consistently under-performs against TRIAGE, which identifies population-specific regulatory genes across diverse cell states without any external reference comparisons **(Figure 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. Variably expressed TFs **(Figure 1B)** were used as a positive gene set to test enrichment of regulatory genes across diverse tissue types. For each tissue type we plotted the rank position of the peak significance ($-\log_{10}p$) value in a Fisher's exact test. Using tabula muris data of nearly 100,000 cells from 20 different mouse tissues at single-cell resolution (Schaum et al., 2018), TRIAGE consistently enriches for cell type-specific regulatory genes compared to original expression with no difference between droplet and smartseq2 data sets **(Figure 4G and Table S5).** Using the mouse organogenesis cell atlas (MOCA), which is among one of the largest

single cell data sets generated to date (Cao et al., 2019), we demonstrated that TRIAGE outperformed the expression value alone in prioritizing cell type-specific regulatory genes across more than 1.3 million mouse single-cell transcriptomes **(Figure 4H).** Lastly, we used benchmarking data for assessing clustering accuracy (Tian, 2018) to assess the performance of TRIAGE using three independent algorithms (i.e. CORE, sc3, and Seurat) and show no difference in accurately assigning cells to the reference (ARI > 0.98) using original expression or TRIAGE transformed expression **(Figure 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 (Barski et al., 2007) collected from the 111 Roadmap samples **(Figure 4J)**. Similarly, cap analysis of gene expression (CAGE), which measures genome-wide 5' transcription activity, showed significant enrichment of variably expressed TFs using TRIAGE from 329 selected FANTOM5 CAGE samples **(Figures 4J and Table S1)** (Forrest et al., 2014). Lastly, analysis of a draft map of the human proteome shows that TRIAGE enriches for regulatory drivers of 30 different tissue types from high resolution Fourier transform mass spectrometry data (Kim et al., 2014) **(Figure 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 (Tirosh et al., 2016). Among the top ranked genes, TRIAGE consistently outperforms expression in prioritizing genes with known involvement in melanoma proliferation and invasion **(Figures 5A and Table S6)**. Using independently derived positive gene sets for proliferative versus invasive melanoma (Tirosh et al., 2016; Verfaillie et al., 2015), TRIAGE recovers with high sensitivity the genetic signatures of these two cancer states **(Figure 5B)**. Gene set enrichment analysis using TRIAGE identified *ETV5* and *TFAP2A* associated with proliferative melanoma versus *TFAP2C* and *TBX3* as regulators of invasive melanoma **(Figure 5C).** *TFAP2A* and *TBX3* have been implicated in proliferative and invasive melanoma respectively (Peres and Prince, 2013; Rambow et al., 2015), 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 (Falletta et al., 2017; Ferguson et al., 2017). 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 **(Figures 5D and 5E)**. These data demonstrate the ability for TRIAGE to effectively identify genetic signatures of functionally distinct cancer cell states without external reference points.

We 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* (Anand et al., 2013; Duan et al., 2017; Spiltoir et al., 2013) to prevent and treat HF by

targeting gene programs linked to cardiac hypertrophy and fibrosis (Duan et al., 2017). We analyzed RNA-seq data from adult mouse hearts where pre-established HF (transverse aortic constriction, TAC) was treated with JQ1. TRIAGE prioritized TFs and regulatory genes with known roles in HF pathogenesis **(Figure 5F)**, outperforming expression ranked genes based on stress-associated gene sets **(Figure 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 **(Figure 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 (Friedman et al., 2018) we analyzed sub-populations at day 2. TRIAGE identified known regulatory genes governing sub-population identity among the top 10 highly ranked genes **(Figure 6A)**. Among the TRIAGE identified genes was *SIX3*, a member of the sine oculis homeobox transcription factor family (RTS=0.54) **(Figures 6A and 6B)**. Importantly, all pairwise differential expression analyses failed to enrich for *SIX3* **(Figure S6A).** Though the role of *SIX3* in neuroectoderm specification has been studied extensively, little is known about its role in other germ layer derivatives (Carl et al., 2002; Lagutin et al., 2003; Steinmetz et al., 2010). Analysis of *SIX3* in hPSC *in vitro* cardiac differentiation shows robust expression in day 2 definitive endoderm (DE) (28.7%) and mesoderm (37.5%) cell populations **(Figure 6C)** with enrichment of *SIX3*$^+$ cells associated with definitive endoderm **(Figures S6B and S6C**). Using previously published laser microdissection approaches, we captured the spatiotemporal transcriptional data from germ layer cells of mid-gastrula stage (E7.0) embryos (Peng et al., 2016), with an expanded analysis to include pre- (E5.5-E.6.0), early- (E6.5) and late-gastrulation (E7.5) mouse embryos (**Figure S6F**). Spatio-temporal expression of *SIX3* and other family members is observed in the epiblast and neuroectoderm, (**Figures 6D and S6G)** consistent with its known role in these lineages (Carl et al., 2002; Lagutin et al., 2003; Steinmetz et al., 2010), as well as early endoderm lineages (**Figure 6D**). Supporting this finding, *SIX3* has been identified as a gene distinguishing definitive from visceral endoderm (Sherwood et al., 2007) 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 **(Figures 6E and 6F)**. Cells were differentiated using monolayer cardiac differentiation and analyzed at day 2 **(Figure 6G).** *SIX3* loss-of-function depleted endoderm and mesendoderm genes **(Figure 6H**) consistent with FACs analysis showing depletion of CXCR4$^+$/EPCAM$^+$ endoderm cell **(Figures 6I-K and S6D)**. 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 **(Figures 6L-N and S6E).** Taken together, these data demonstrate a novel role of *SIX3* in 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 time-course of cardiac development were analyzed with TRIAGE using a customized gene mapping tool to link human to *Ciona* genes **(Figure 6O)** (Wang et al., 2019). The top ranked genes based on TRIAGE were analyzed **(Figure 6P)**. *RNF220* (RTS=0.30, **Figure 6Q**), an E3 ubiquitin ligase governing Wnt signaling pathway activity through β-catenin degradation (Ma et al., 2014; Tsoi et al., 2018), 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 (Cao et al., 2019; Schaum et al., 2018) as well as computational analysis and prediction algorithms (Benayoun et al., 2014; Cahan et al., 2014; Palpant et al., 2017; Rackham et al., 2016) have revolutionized our capacity to study complex biological systems. This study demonstrates the power of analyzing cell heterogeneity to understand genome regulation at scale and revealing a repressive tendency metric that provides a strong, quantitative prediction value for cell type-specific regulatory genes controlling cell diversification in development and disease. 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 H3K27me3 data from diverse cell and tissue types in species where gene expression is governed by the polycomb group complex. 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. The conservation of this regulatory logic provides an effective strategy of utilizing large, diverse genome-wide data to establish quantitative basic principles of cell states to infer cell-type specific mechanisms that underpin the complexity of biological systems. We anticipate furthermore that this analytic approach can be applied to 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.

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: Assisted in 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
DS: Supervised work on analysis of HF pathogenesis data
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

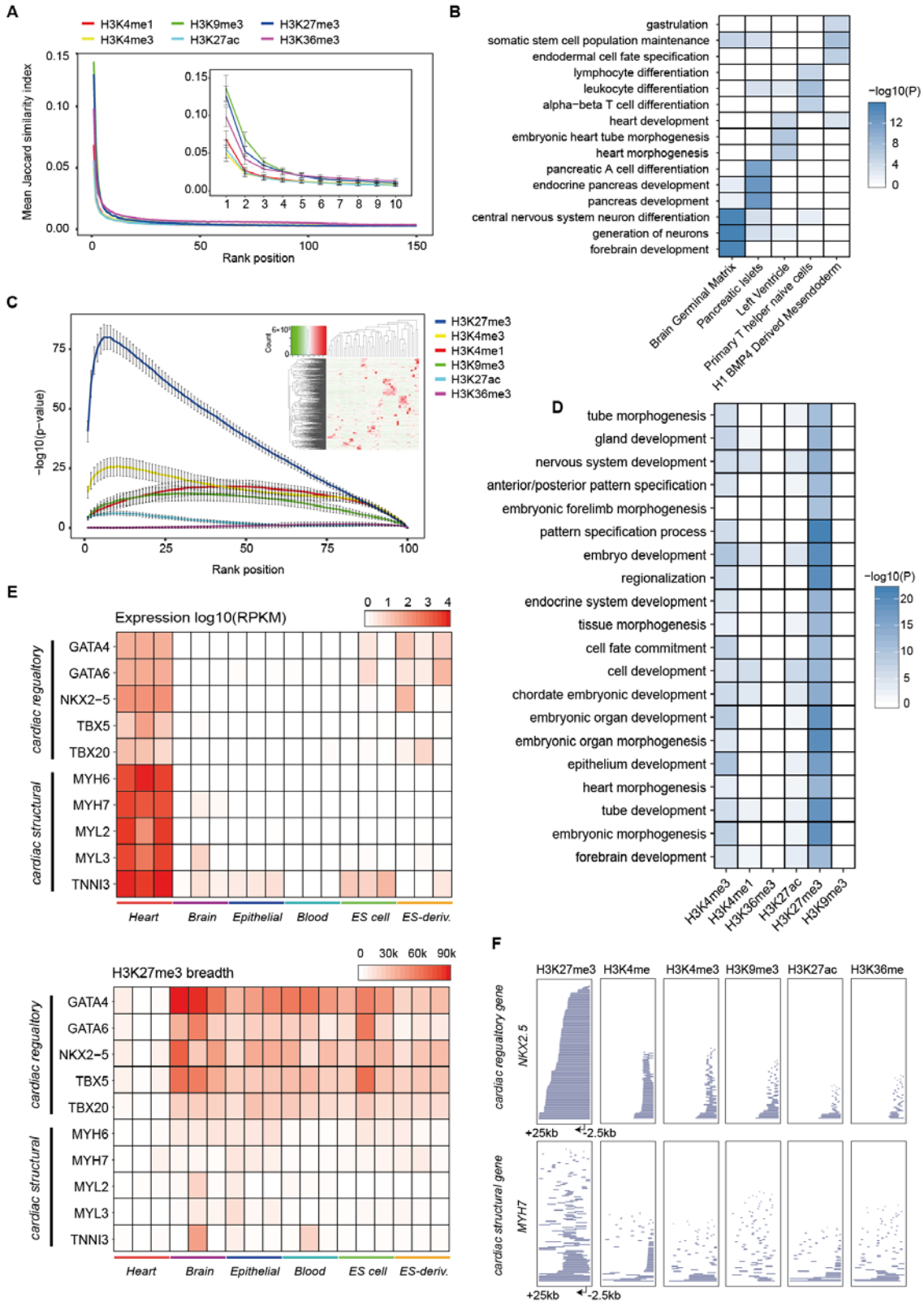
A
H3K4me1
H3K4me3
H3K9me3
H3K27ac
H3K27me3
H3K36me3
Mean Jaccard similarity index
Rank position
B
gastrulation
somatic stem cell population maintenance
endodermal cell fate specification
lymphocyte differentiation
leukocyte differentiation
alpha−beta T cell differentiation
heart development
embryonic heart tube morphogenesis
heart morphogenesis
pancreatic A cell differentiation
endocrine pancreas development
pancreas development
central nervous system neuron differentiation
generation of neurons
forebrain development
Brain Germinal Matrix
Pancreatic Islets
Left Ventricle
Primary T helper naive cells
H1 BMP4 Derived Mesendoderm
−log10(P)
C
−log10(p−value)
Rank position
D
tube morphogenesis
gland development
nervous system development
anterior/posterior pattern specification
embryonic forelimb morphogenesis
pattern specification process
embryo development
regionalization
endocrine system development
tissue morphogenesis
cell fate commitment
cell development
chordate embryonic development
embryonic organ development
embryonic organ morphogenesis
epithelium development
heart morphogenesis
tube development
embryonic morphogenesis
forebrain development
H3K4me3
H3K4me1
H3K36me3
H3K27ac
H3K27me3
H3K9me3
−log10(P)
E
Expression log10(RPKM)
cardiac regulatory
GATA4
GATA6
NKX2−5
TBX5
TBX20
cardiac structural
MYH6
MYH7
MYL2
MYL3
TNNI3
Heart
Brain
Epithelial
Blood
ES cell
ES-deriv.
H3K27me3 breadth
F
H3K27me3
H3K4me
H3K4me3
H3K9me3
H3K27ac
H3K36me
cardiac regulatory gene
NKX2.5
cardiac structural gene
MYH7
+25kb
-2.5kb

Figure 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 within each cell type and grouped into bins of 100 genes. Mean Jaccard similarity index is calculated by comparing gene sets of equivalent bins between all pair-wise cell types. Regardless of HM type, top 100 genes were significantly more shared between cell types compared to genes with narrower domains ($p < 2.2e-16$ for all HMs, Wilcoxon rank-sum test). Scale bars shows the 95% confidence interval.

(B) Enrichment of tissue type specific GO biological process terms associated with the top 50 variably expressed TFs by their expression level in 5 selected Roadmap samples (brain germinal matrix (E070), pancreatic islets (E087), left ventricle (E095), primary T helper naïve cells (E038) and H1 BMP4-derived mesendoderm (E004)). Fisher's exact test (one-tailed) was used for enrichment analysis.

(C) Variably expressed TFs are strongly associated with broad H3K27me3 domains. For each cell type, genes are ranked by the breadth of the associated HM peak and grouped into percentile bins (e.g. genes with top 1% broadest peaks are grouped into the rank bin position 1 in the x-axis). H3K27me3 significantly enriches for variably expressed TFs within the top 5% of broad domains ($p=6.66e-16$, Fisher's exact test, one-tailed). Mean enrichment of variably expressed genes across the cell types at each rank position is shown on the y-axis, with scale bars showing the 95% confidence interval. Inset heatmap shows row-normalized expression levels of the 634 variably expressed TFs across the 46 Roadmap samples.

(D) Top 200 genes that are most frequently associated with broad HMs across the 111 Roadmap cell types. H3K27me3 is uniquely associated with regulation of development and morphogenesis. Enrichment of selected GO biological process terms is calculated using Fisher's exact test (one-tailed).

(E) Analysis of gene expression (top) vs H3K27me3 breadth (bottom) for cardiac-specific regulatory genes vs structural genes. H3K27me3 uniquely identifies cardiac regulatory from cardiac structural genes when analyzed on all samples except heart. 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) Distribution of HM domains in the proximal region of selected cardiac RefSeq TSSs *NKX2-5* and *MYH7*. Each line represents an associated HM domain aligned to the proximity of the gene (+25kb upstream of the TSS to -2.5kb downstream) in 111 cell types.

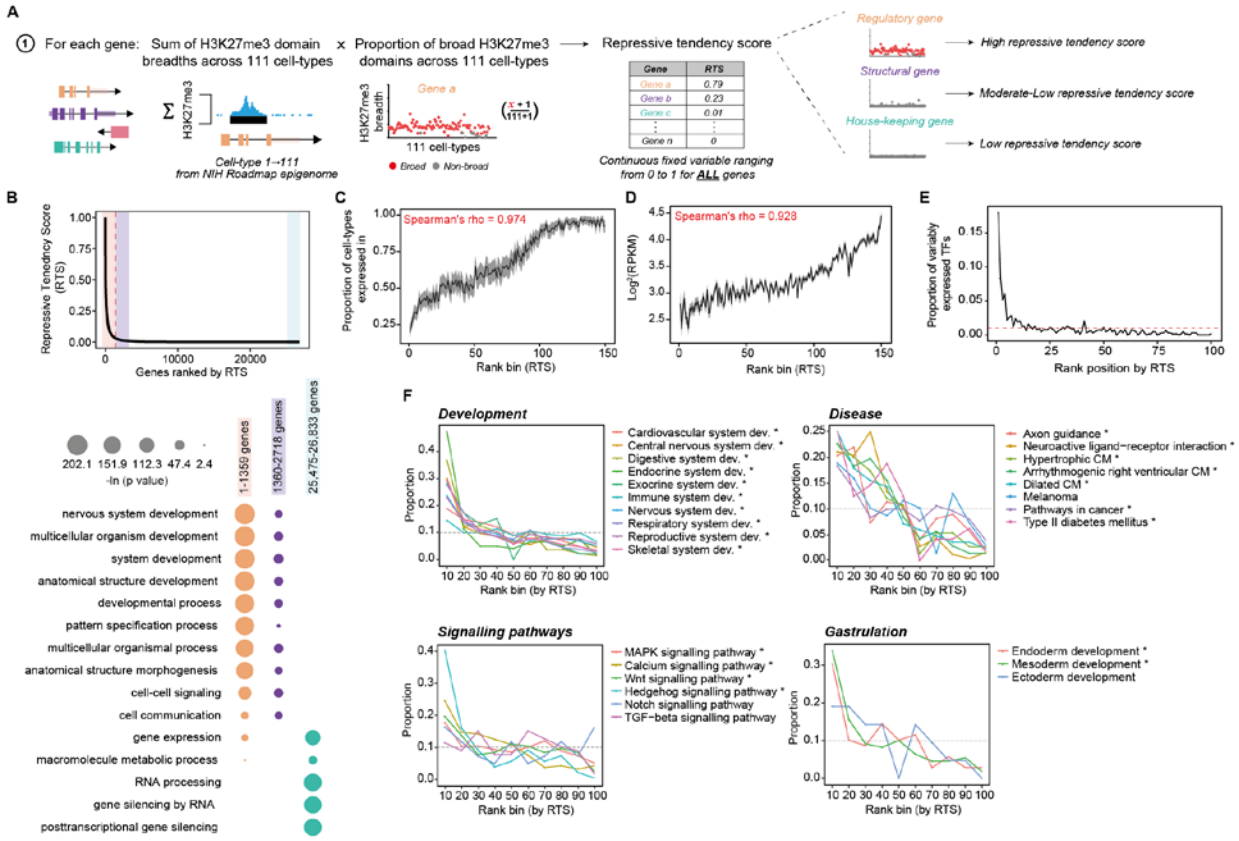

Figure 2: Genes with frequent broad H3K27me3 domains are cell type-specific regulatory genes.

(A) Schematic diagram showing the calculation basis for the repressive tendency score (RTS) for any gene based on breadth information of assigned H3K27me3 domains observed across the 111 NIH Epigenome data sets.

(B) Genes ranked by the RTS (top). Red dashed line indicates the inflection point on the interpolated curve (RTS=0.03022) above which genes exhibit substantially higher RTS than the rest (*n*=1,359). (bottom) Enrichment of selected GO biological process terms associated with the RTS priority genes, in comparison to genes with a lower RTS (Fisher's exact test, one-tailed) **(Supplementary Table 4).**

(C) Expressional specificity of protein-coding genes sorted by the RTS (x-axis). Each rank bin includes 100 genes sorted in a descending order of the RTS (e.g. top 100 genes in rank position 1 etc.). For each gene, the proportion of cell types (out of the 46 Roadmap samples) where the gene is detected (RPKM > 1) is calculated. Average proportions are subsequently calculated for each bin. Shaded regions mark the 95% confidence interval.

(D) Relationship between the expression level and the RTS. Each rank bin includes 100 genes sorted in a descending order (e.g. top 100 genes in rank position 1 etc.). For each bin, an average expression value of genes is calculated. Shaded regions mark the 95% confidence interval.

(E) Distribution of variably expressed TFs sorted by the RTS in a descending order (x-axis). Each rank bin includes 1% of the total genes included. Red dashed line represents a uniform distribution.

(F) Distribution of genes with selected GO biological process (Development and Gastrulation) or KEGG pathway (Disease and Signaling pathways) terms **(Supplementary Table 4).** Asterisk marks indicate significant enrichment of a given term at rank bin position 10 (i.e. top 10% genes) (Benjamini-Hochberg FDR < 0.05, Fisher's exact test).

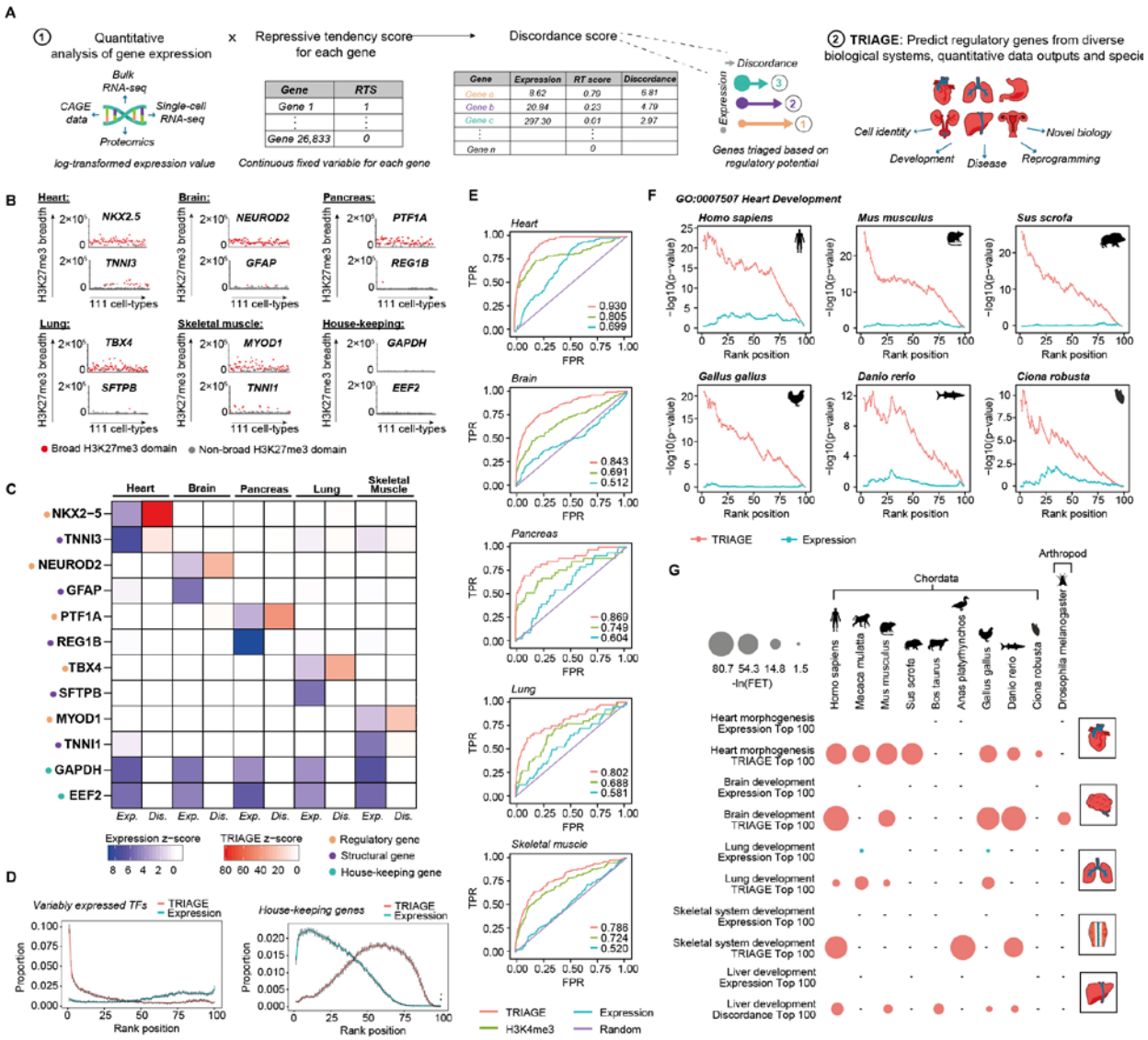

Figure 3: Inferring cell type-specific regulatory genes from somatic cell types.

(A) Schematic outline showing integration of the product of the RTS with any genome wide readout of gene expression establishes a discordance score as the basis for TRIAGE – a computational analysis strategy for inferring the regulatory basis of cell identity underlying development, disease, and cell reprogramming.

(B) Breadths of H3K27me3 domains (in base-pairs) associated with selected cell type-specific regulatory and structural genes, observed across the 111 NIH Epigenomes data sets.

(C) TRIAGE non-linearly transforms the expression value (Exp.) to the discordance score (Dis.), consistently enriching for regulatory genes in all tissue types. Expression profiles were collected from GTEx samples and averaged for the tissue type (Lonsdale et al., 2013).

(D) Distributions of variably expressed TFs (left) and housekeeping genes (right) when genes are ordered by the expression value (blue) or the discordance score (red). Each rank bin (x-axis) includes 1% of all genes and sorted in descending order (e.g. rank position 1 represents top 1% genes etc.). TRIAGE identifies variably expressed TFs among the top 1% of genes in every cell type analyzed **(**p<2.2e-16 for all cell types,

Fisher's exact test, one-tailed) while reducing the relative abundance of housekeeping genes compared to the expression-based ranking **(**$p<2.2e-16$, Wilcoxon rank-sum test, one-tailed).

(E) Receiver-operating characteristic (ROC) plots comparing sensitivity of identifying tissue type specific regulatory genes (**Supplementary Table 2**) of the 5 distinct tissue types. Area under the curve (AUC) values are shown on the right bottom corner of the plot. Performance comparison is between TRIAGE (red), H3K4me3 broad domains (green) (Benayoun et al., 2014), expression value (blue), and random (purple).

(F) Inter-species enrichment analysis of TFs annotated with 'heart development' GO term (GO:0007507) in cardiac samples. For each species, genes are ranked by either the expression value (blue) or the discordance score (red) and binned into a rank bin (each bin includes 1% of all genes) and the enrichment is calculated at each rank position (y-axis, Fisher's exact test, one-tailed).

(G) Enrichment of tissue type specific TFs within selected GO BP term in the top 100 genes ranked by the expression value (blue) or the discordance score (red), across different species. Enrichment of a given gene set is calculated using Fisher's exact test (one-tailed). Hyphen (-) indicates no data set available.

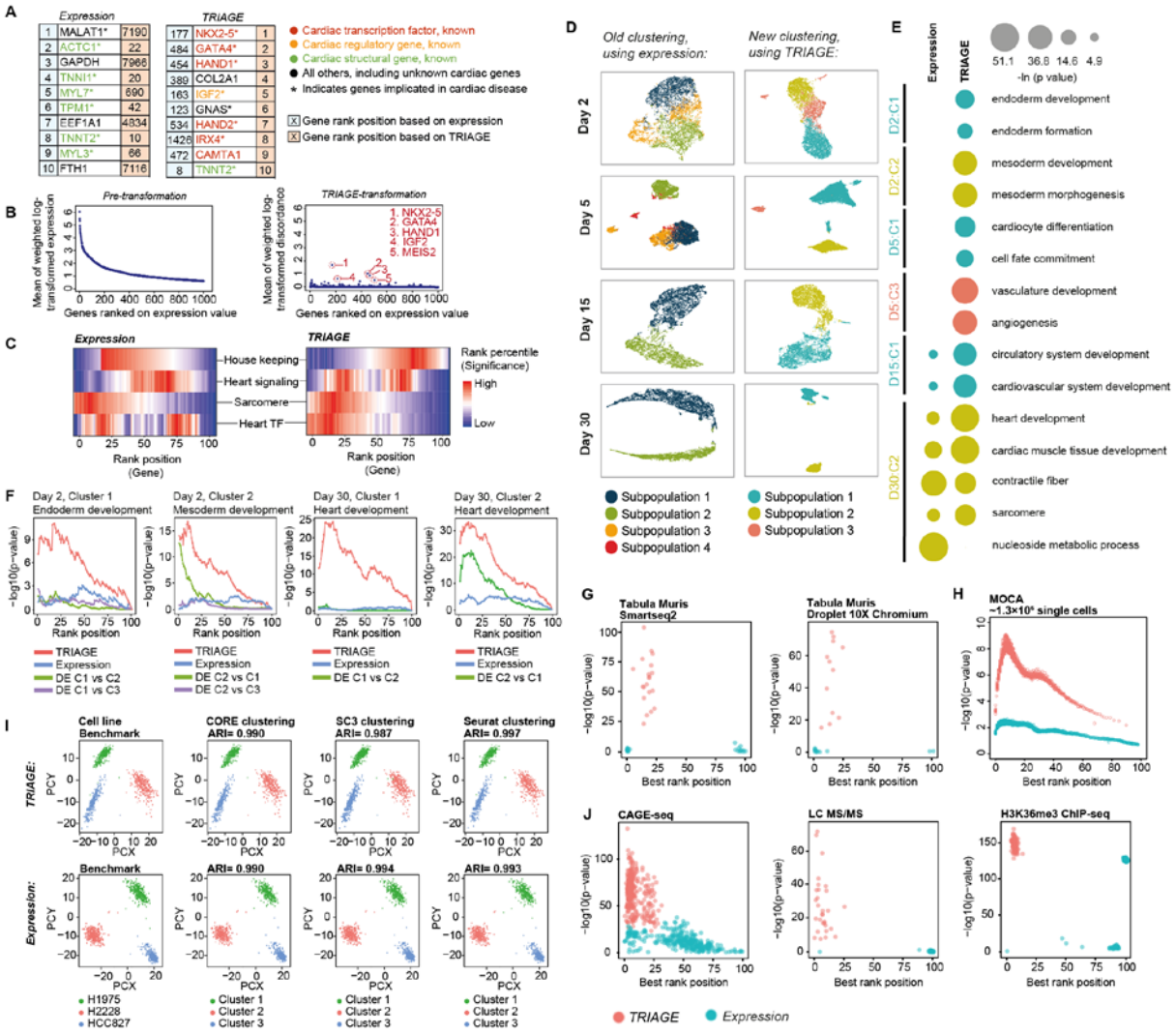

Figure 4: Identifying regulatory genes using diverse multi-omics data sets.
(A) Top 10 genes ranked by expression (left) or discordance score (right) from hPSC-derived cardiomyocyte using scRNA-seq from *in vitro* cardiac-directed differentiation (Friedman et al., 2018).
(B) Cardiomyocyte expression data ranked by original expression (left). Using the same rank position, genes are re-calculated using TRIAGE (right) showing the dramatic quantitative change in values for all genes resulting in quantitative prioritization of cell-type specific cardiac regulatory genes.
(C) Enrichment of four different functional gene sets (i.e. housekeeping (defined in (Eisenberg and Levanon, 2013), heart signaling (genes with 'heart development (GO:0007507)' term and any KEGG signaling pathway term(s)), sarcomere (genes with 'sarcomere (GO:0030017)' term) or heart TF (TFs with 'heart development (GO:0007507) term') genes in cardiomyocytes. Genes are ranked by either the expression value (left) or TRIAGE (right) and enrichment of a given gene set is calculated at each rank position using 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) that

are consistently identified by TRIAGE but not original expression. In contrast, expression data strongly detect structural and housekeeping genes.

(G) Enrichment of developmental GO BP terms during *in vitro* cardiac-directed differentiation. Genes are ranked by TRIAGE (red), expression (blue) or fold change of gene expression between cell clusters using differentially expressed gene (DEG) analysis (green or purple). Each rank bin includes 1% of all genes and is sorted in a descending order (i.e. top 1% gene in the rank position 1 etc.).

(G, H) Application of TRIAGE to scRNA-seq data generated from Smart-seq2 or Droplet 10X chromium **(G)** as well as the mouse organogenesis cell atlas (MOCA) **(H)**. For each sample, the rank position (x-axis) of the highest enrichment (defined as the lowest p-value using Fisher's exact test, y-axis) of variably expressed TFs is plotted. For MOCA data set, each data point represents an average of 1,000 samples**.**

(I) Comparison of single cell RNA-seq cluster assignment efficiency between original expression and TRIAGE analyzed data using Mixology data sets.

(J) Application of TRIAGE to various quantitative readouts of gene expression including CAGE-seq, protein abundance, and tag density of H3K36me3. For each sample, the rank position (x-axis) of the highest enrichment (defined as the lowest p-value using Fisher's exact test, y-axis) of variably expressed TFs is plotted.

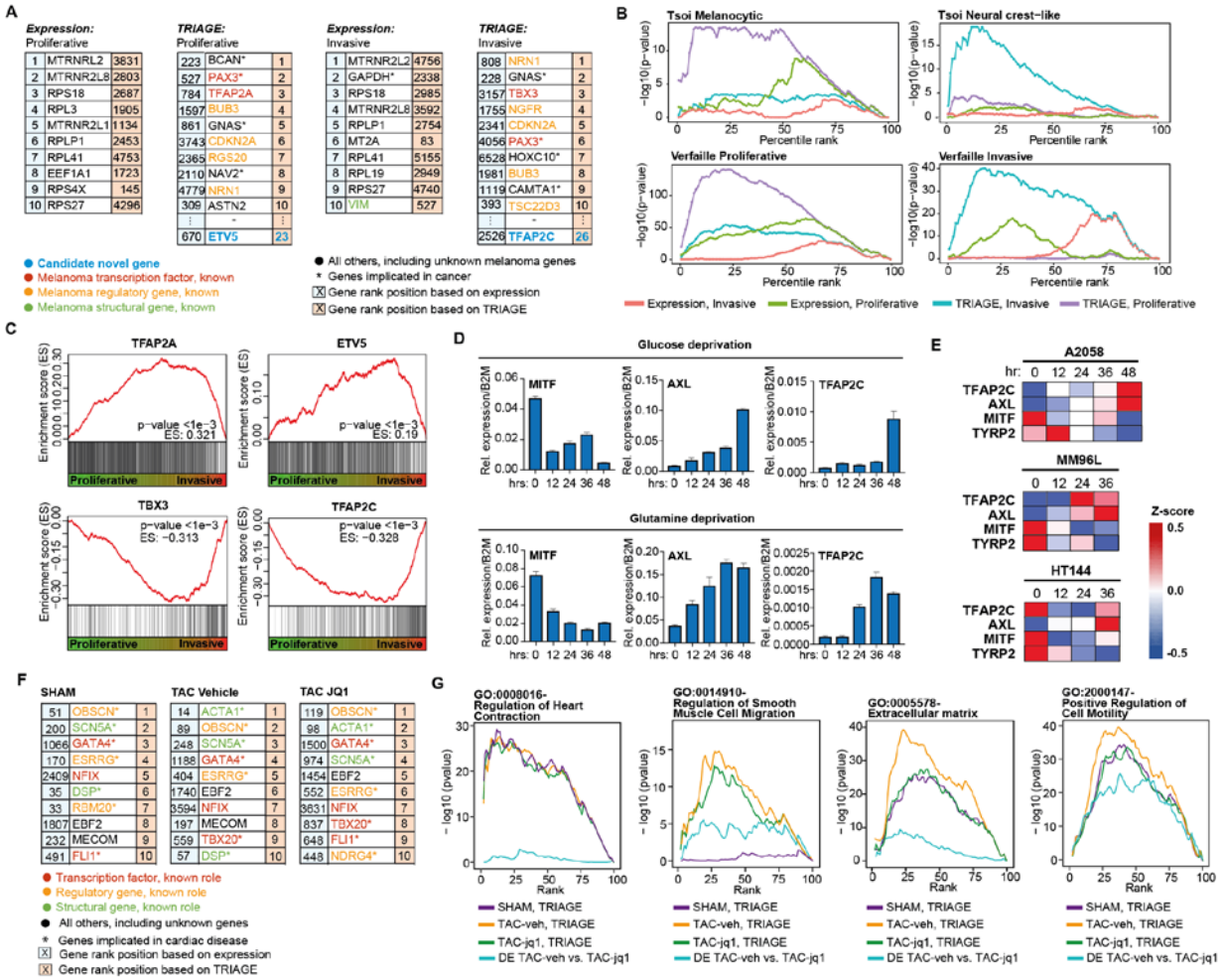

Figure 5: Determining the regulatory basis of disease pathogenesis and therapy.

(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 analysis of genes annotated with GO terms associated with cardiac biology and heart failure stress response mechanisms comparing each sample individually analyzed by TRIAGE and

compared against outcomes resulting from DE analysis of 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).

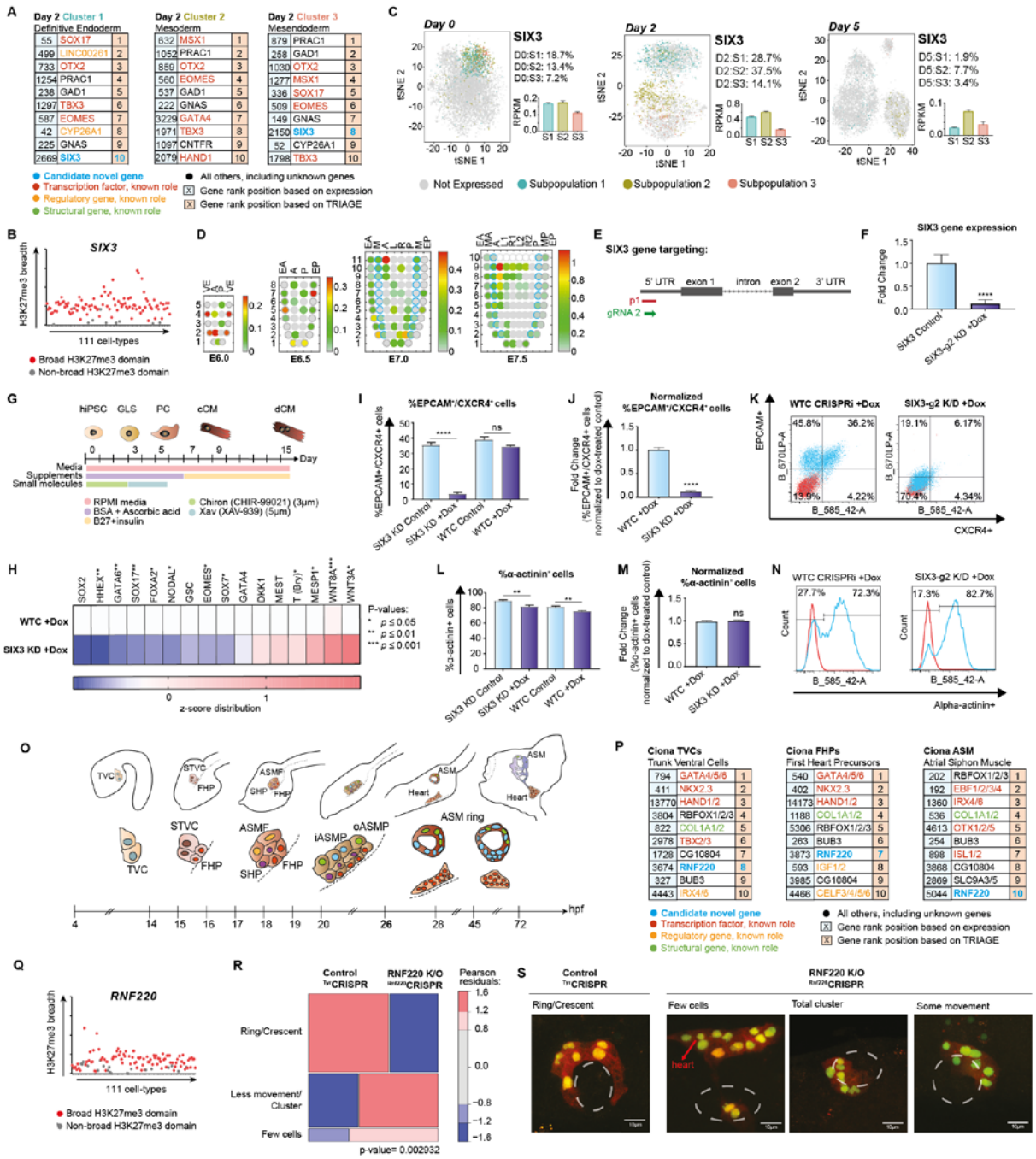

Figure 6: Predicting novel regulators of heart development.

(A) Top 10 genes ranked by expression value (left) or TRIAGE (right) from populations found on day 2 (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.
(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.
(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.
(O) Schematic overview of cardiac development in *Ciona* from 4 to 72 hours post fertilization (hpf) at 18°C. Adapted from (Evans Anderson and Christiaen, 2016). TVC: trunk ventral cells; STVC: second TVC; FHP: first heart precursor; SHP: second heart precursor; ASMF: atrial siphon muscle founder cells; iASMP: inner atrial siphon muscle precursor; oASMP: outer atrial siphon muscle precursor.
(P) Top 10 genes ranked by expression value (left) or discordance score (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

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

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

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

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

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

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

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

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

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

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

Falletta, P., Sanchez-Del-Campo, L., Chauhan, J., Effern, M., Kenyon, A., Kershaw, C.J., Siddaway, R., Lisle, R., Freter, R., Daniels, M.J., *et al.* (2017). Translation reprogramming is an evolutionarily conserved driver of phenotypic plasticity and therapeutic resistance in melanoma. Genes Dev *31*, 18-33.

Ferguson, J., Smith, M., Zudaire, I., Wellbrock, C., and Arozarena, I. (2017). Glucose availability controls ATF4-mediated MITF suppression to drive melanoma cell growth. Oncotarget *8*, 32946-32959.

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

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

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

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., *et al.* (2014). A draft map of the human proteome. Nature *509*, 575-581.
Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.
Lagutin, O.V., Zhu, C.C., Kobayashi, D., Topczewski, J., Shimamura, K., Puelles, L., Russell, H.R., McKinnon, P.J., Solnica-Krezel, L., and Oliver, G. (2003). Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. Genes & development *17*, 368-379.
Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., *et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.
Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.
Ma, P., Yang, X., Kong, Q., Li, C., Yang, S., Li, Y., and Mao, B. (2014). The ubiquitin ligase RNF220 enhances canonical Wnt signaling through USP7-mediated deubiquitination of β-catenin. Molecular and cellular biology *34*, 4355-4366.
Palpant, N.J., Wang, Y., Hadland, B., Zaunbrecher, R.J., Redd, M., Jones, D., Pabon, L., Jain, R., Epstein, J., Ruzzo, W.L., *et al.* (2017). Chromatin and Transcriptional Analysis of Mesoderm Progenitor Cells Identifies HOPX as a Regulator of Primitive Hematopoiesis. Cell Rep *20*, 1597-1608.
Peng, G., Suo, S., Chen, J., Chen, W., Liu, C., Yu, F., Wang, R., Chen, S., Sun, N., Cui, G., *et al.* (2016). Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Dev Cell *36*, 681-697.
Peres, J., and Prince, S. (2013). The T-box transcription factor, TBX3, is sufficient to promote melanoma formation and invasion. Molecular cancer *12*, 117-117.
Perez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., and Guigo, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. Nat Genet *47*, 1158-1167.
Rackham, O.J., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., and Shin, J.W. (2016). A predictive computational framework for direct reprogramming between human cell types. Nature genetics *48*, 331.
Rambow, F., Job, B., Petit, V., Gesbert, F., Delmas, V., Seberg, H., Meurice, G., Van otterloo, E., Dessen, P., Robert, C., *et al.* (2015). New Functional Signatures for Understanding Melanoma Biology from Tumor Cell Lineage-Specific Analysis. Cell Reports *13*, 840-853.
Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.
Schaum, N., Karkanias, J., Neff, N., May, A.P., S.R., Q., Wyss-Coray, T., Batson, J., Botvinnik, O., Chen, M.B., Chen, S., *et al.* (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367-372.
Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol *6*, R33.
Scornavacca, C., Zickmann, F., and Huson, D.H. (2011). Tanglegrams for rooted phylogenetic trees and networks. Bioinformatics *27*, i248-256.
Sherwood, R.I., Jitianu, C., Cleaver, O., Shaywitz, D.A., Lamenzo, J.O., Chen, A.E., Golub, T.R., and Melton, D.A. (2007). Prospective isolation and global gene expression analysis of definitive and visceral endoderm. Developmental Biology *304*, 541-555.
Spiltoir, J.I., Stratton, M.S., Cavasin, M.A., Demos-Davies, K., Reid, B.G., Qi, J., Bradner, J.E., and McKinsey, T.A. (2013). BET acetyl-lysine binding proteins control pathological cardiac hypertrophy. J Mol Cell Cardiol *63*, 175-179.

Steinmetz, P.R., Urbach, R., Posnien, N., Eriksson, J., Kostyuchenko, R.P., Brena, C., Guy, K., Akam, M., Bucher, G., and Arendt, D. (2010). Six3 demarcates the anterior-most developing brain region in bilaterian animals. EvoDevo *1*, 14.
Tian, L., Dong, X., Freytag, S., LeCao, K., et al. (2018). scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. BioRxiv.
Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., *et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189.
Tsoi, J., Robert, L., Paraiso, K., Galvan, C., Sheu, K.M., Lay, J., Wong, D.J.L., Atefi, M., Shirazi, R., Wang, X., *et al.* (2018). Multi-stage Differentiation Defines Melanoma Subtypes with Differential Vulnerability to Drug-Induced Iron-Dependent Oxidative Stress. Cancer Cell *33*, 890-904.e895.
Verfaillie, A., Imrichova, H., Atak, Z.K., Dewaele, M., Rambow, F., Hulselmans, G., Christiaens, V., Svetlichnyy, D., Luciani, F., Van den Mooter, L., *et al.* (2015). Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. Nat Commun *6*, 6683.
Wang, W., Niu, X., Jullian, E., Kelly, R., Satija, R., and Christiaen, L. (2019). A single cell transcriptional roadmap for cardiopharyngeal fate diversification. BioRxiv.