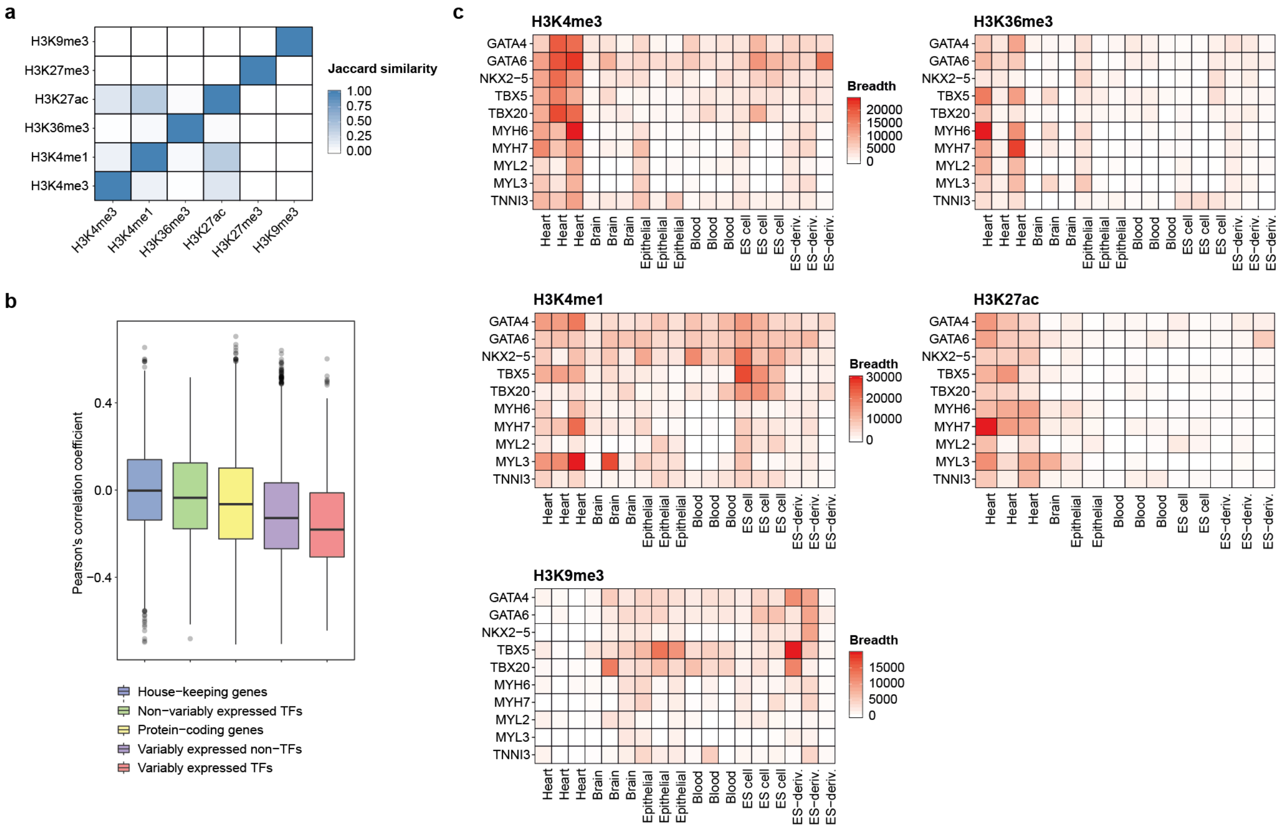
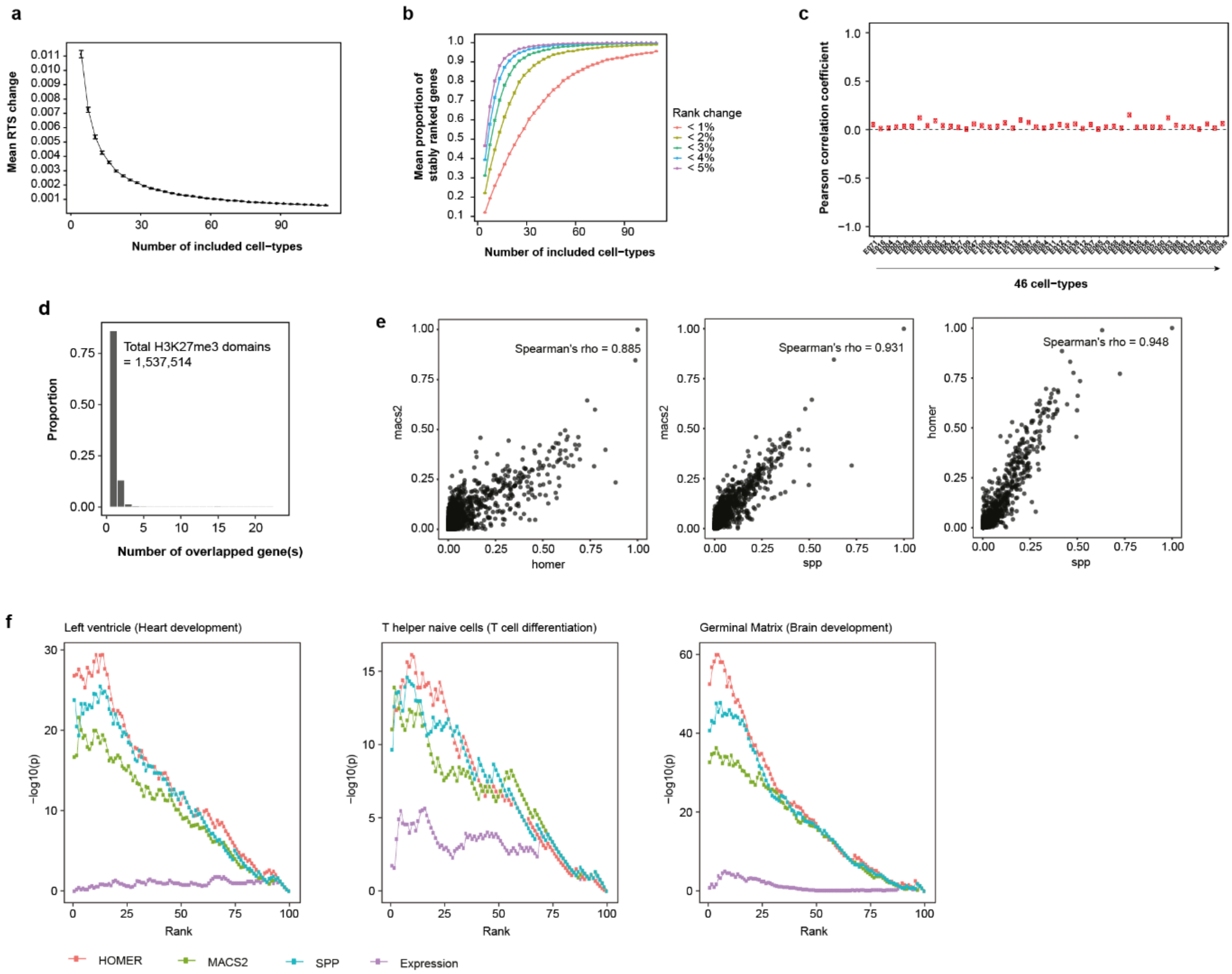
Conserved epigenetic regulatory logic infers genes governing cell identity

Supplementary Information



Supplementary Figure 1: H3K27me3 histone modification (HM) domains have distinct functional association with cell type specific regulatory genes (Supplementary to Figure 1).
(a) Distinct gene sets are identified by different types of broad HM domains. Similarity between top 200 genes that are most frequently associated with broad HM domains observed across 111 Roadmap tissue and cell types.
(b) Correlation between the H3K27me3 domain breadth and the corresponding expression value of genes observed across 46 Roadmap tissue and cell types. Stronger negative correlation is observed for variably expressed TFs (*n*=634, median Pearson's *r* = -0.181), compared to variably expressed non-TFs (*n*=7,406, median Pearson's *r* = -0.128, *p*=2.57e-07), all protein-coding genes (*n*=18,490, median Pearson's *r* = -0.064, *p*=2.55e-25), non-variably expressed TFs (*n*=793, median Pearson's *r* = -0.035, *p*=2.31e-23) and housekeeping genes (*n*=3,818, median Pearson's *r* = -0.002, *p*=7.7e-54) (Welch's t-test, one-tailed).
(c) Breadths of HM domains associated with selected cardiac-specific regulatory (*GATA4, GATA6, NKX2-5, TBX5, TBX20*) and structural (*MYH6, MYH7, MYL2, MYL3, TNNI3*) genes in 18 Roadmap sample; 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).



Supplementary Figure 2: Repressive tendency scores (RTS) estimated using the 111 Roadmap epigenomic datasets are stable and reproducible (Supplementary to Figures 2 and 3).

(a) Mean RTS changes (*y*-axis) with a cumulative addition of samples (*x*-axis). 111 Roadmap samples are randomly sorted and RTS values are calculated iteratively with an addition of 3 samples (without replacement) until all samples are included. This process is repeated 1,000 times. Error bars show 95% the confidence interval. Source data are provided as a Source Data file.

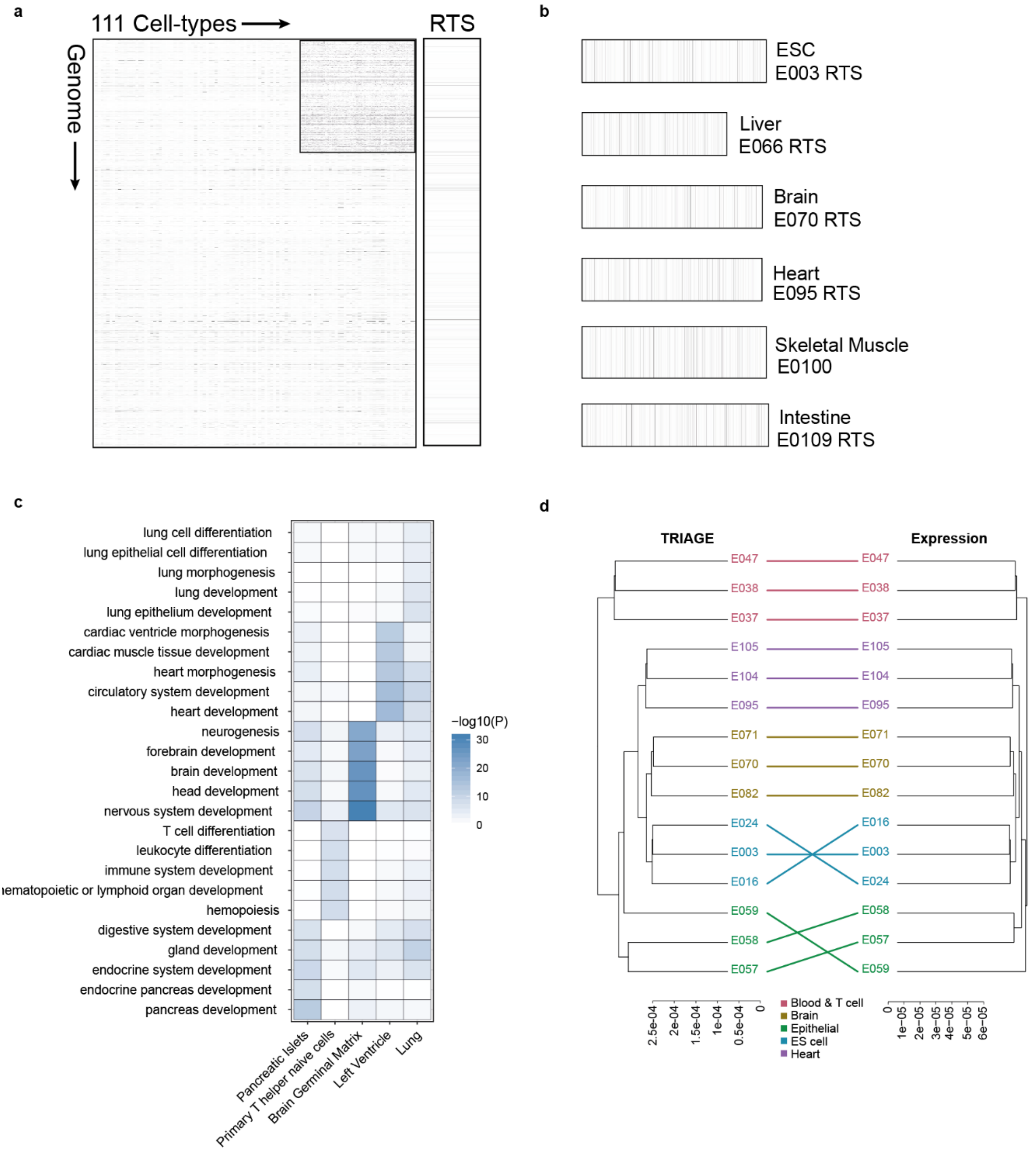
(b) Mean proportion of stably ranked genes (*y*-axis) with a cumulative addition of samples (*x*-axis) (see Methods). 111 Roadmap samples are randomly sorted and RTS values are calculated iteratively with an addition of 3 samples without replacement until all samples are included. This process is repeated 1,000 times. Error bars show 95% the confidence interval. Source data are provided as a Source Data file.

(c) Correlation between the expression value and the discordance score for protein coding genes (*n*=18,707) across the 46 Roadmap cell types.

(d) The number of genes overlapped by H3K27me3 domains assigned to protein coding genes (*n*=1,537,514). Approximately 85% of these domains overlap a single gene.

(e) Correlation between RTSs calculated from H3K27me3 domains identified by 3 different peak callers (i.e. MACS2, SPP2 and HOMER).

(f) Enrichment of cell type-specific regulatory genes (i.e. TFs with 'Heart development' GO:0007507, 'T cell differentiation' GO:0030217 and 'Brain development' GO:0007420 for left ventricle (E095), T helper naïve cell (E038) and brain germinal matrix (E070) samples respectively, from left to right). Genes are sorted by the DS (See Predicting cell-type-specific regulatory genes based on H3K27me3) calculated with 3 different versions of RTSs and binned into a percentile bin.



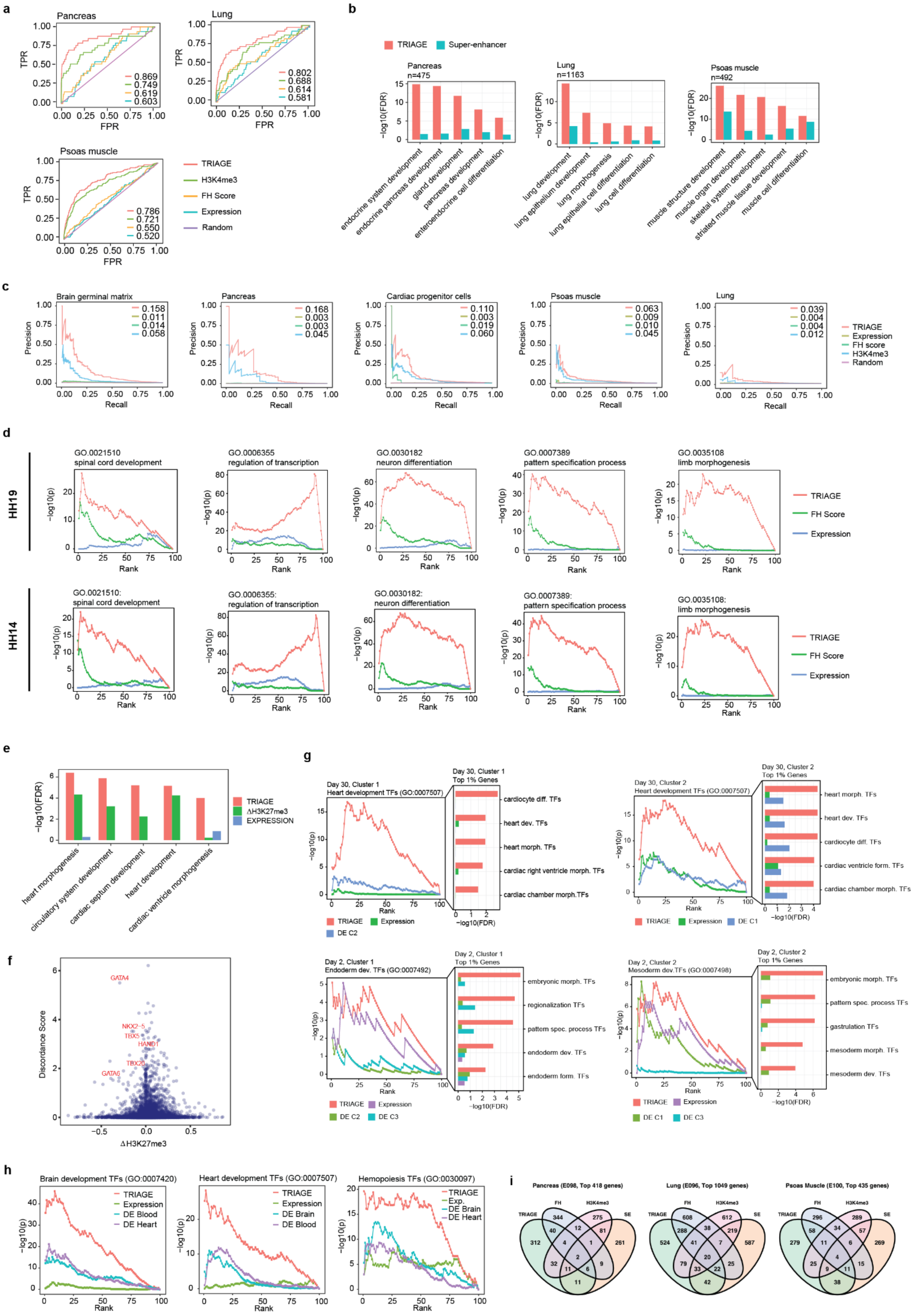
Supplementary Figure 3: TRIAGE is an application of gene-specific RTS to the input expression readout (Supplementary to Figures 2 and 3).

(a) (Left) Breadth of H3K27me3 domains assigned to 26,833 RefSeq genes (row) across the 111 Roadmap tissue and cell types (column). Darkness of the band shows the breadth of the assigned domain. **(Inset)** Top 5% broadest domains associated with the genes (row) across the Roadmap samples (column). **(Right)** RTS values corresponding the genes (row). Darkness of the band shows the RTS.

(b) TRIAGE collects a different set of RTS values specific for input transcriptome. Each collection represents a distinct Roadmap cell or tissue type. For each collection, all genes with an expression value (RPKM>0) are sorted from left (high) to right (low). Darkness of the band shows the RTS value.

(c) Enrichment of tissue or cell type-specific GO biological process terms among top 1% genes selected by using the DS in 5 distinct Roadmap tissue or cell types (Fisher's exact test, one-tailed); Pancreatic Islets (E087), Primary T helper naïve cells from peripheral blood (E038), Brain Germinal Matrix (E070), Left Ventricle (E095), Lung (E096).

(d) Similarity between 15 Roadmap samples based on (i) the original transcript abundance (Expression) and (ii) the DS (TRIAGE). Distance between samples is shown as 1-Pearson’s correlation coefficient. For description of samples, see Supplementary table1.



a
Pancreas
TPR
FPR
0.869
0.749
0.619
0.603
Lung
0.802
0.688
0.614
0.581
Psoas muscle
0.786
0.721
0.550
0.520
TRIAGE
H3K4me3
FH Score
Expression
Random
b
TRIAGE
Super-enhancer
Pancreas n=475
−log10(FDR)
endocrine system development
endocrine pancreas development
gland development
pancreas development
enteroendocrine cell differentiation
Lung n=1163
lung development
lung epithelium development
lung morphogenesis
lung epithelial cell differentiation
lung cell differentiation
Psoas muscle n=492
muscle structure development
muscle organ development
skeletal system development
striated muscle tissue development
muscle cell differentiation
c
Brain germinal matrix
0.158
0.011
0.014
0.058
Pancreas
0.168
0.003
0.003
0.045
Cardiac progenitor cells
0.110
0.003
0.019
0.060
Psoas muscle
0.063
0.009
0.010
0.045
Lung
0.039
0.004
0.004
0.012
Precision
Recall
TRIAGE
Expression
FH score
H3K4me3
Random
d
HH19
HH14
GO.0021510 spinal cord development
GO.0006355 regulation of transcription
GO.0030182 neuron differentiation
GO.0007389 pattern specification process
GO.0035108 limb morphogenesis
GO.0021510: spinal cord development
GO.0006355: regulation of transcription
GO.0030182: neuron differentiation
GO.0007389: pattern specification process
GO.0035108: limb morphogenesis
−log10(p)
Rank
TRIAGE
FH Score
Expression
e
−log10(FDR)
heart morphogenesis
circulatory system development
cardiac septum development
heart development
cardiac ventricle morphogenesis
TRIAGE
ΔH3K27me3
EXPRESSION
f
Discordance Score
ΔH3K27me3
GATA4
NKX2−5
TBX5
HAND1
TBX20
GATA6
g
Day 30, Cluster 1 Heart development TFs (GO:0007507)
Day 30, Cluster 1 Top 1% Genes
cardiocyte diff. TFs
heart dev. TFs
heart morph. TFs
cardiac right ventricle morph. TFs
cardiac chamber morph.TFs
TRIAGE
Expression
DE C2
Day 30, Cluster 2 Heart development TFs (GO:0007507)
Day 30, Cluster 2 Top 1% Genes
heart morph. TFs
heart dev. TFs
cardiocyte diff. TFs
cardiac ventricle form. TFs
cardiac chamber morph. TFs
TRIAGE
Expression
DE C1
Day 2, Cluster 1 Endoderm dev. TFs (GO:0007492)
Day 2, Cluster 1 Top 1% Genes
embryonic morph. TFs
regionalization TFs
pattern spec. process TFs
endoderm dev. TFs
endoderm form. TFs
TRIAGE
Expression
DE C2
DE C3
Day 2, Cluster 2 Mesoderm dev.TFs (GO:0007498)
Day 2, Cluster 2 Top 1% Genes
embryonic morph. TFs
pattern spec. process TFs
gastrulation TFs
mesoderm morph. TFs
mesoderm dev. TFs
TRIAGE
Expression
DE C1
DE C3
−log10(p)
Rank
−log10(FDR)
h
Brain development TFs (GO:0007420)
TRIAGE
Expression
DE Blood
DE Heart
Heart development TFs (GO:0007507)
TRIAGE
Expression
DE Brain
DE Blood
Hemopoiesis TFs (GO:0030097)
TRIAGE
Exp.
DE Brain
DE Heart
−log10(p)
Rank
i
Pancreas (E098, Top 418 genes)
TRIAGE
FH
H3K4me3
SE
344
275
40
12
81
312
4
1
261
2
32
11
6
9
11
Lung (E096, Top 1049 genes)
608
612
288
38
219
524
41
7
587
20
79
33
22
25
42
Psoas Muscle (E100, Top 435 genes)
296
289
58
34
57
279
11
6
269
4
25
9
11
15
38

Supplementary Figure 4: TRIAGE effectively prioritizes key regulators of cell identity across diverse cell states (Supplementary to Figures 3 and 4).

(a) Receiver operating characteristic (ROC) comparing performance of TRIAGE against other methods for pancreas (E087), lung (E096) and psoas muscle (E100). Area under curve (AUC) values are shown on the bottom right corner.

(b) Functional enrichment of top *n* genes identified by TRIAGE or super-enhancer (SE), where *n* is defined by the number of active genes nearest to SEs (Fisher's exact test, one-tailed, see Methods). For the full list, see Supplementary table 7.

(c) Precision-recall curves (PRC) for the 5 different tissue groups. Area under curve (AUC) values are shown on the top right corner.

(d) Enrichment of embryonic development GO terms for developing chicken embryo data[1] (Fisher's exact test, one-tailed). Genes are sorted by a given metric (i.e. DS, functional heterogeneity (FH) score or expression value) and binned into a percentile bin (*x*-axis). For the full list, see Supplementary table 8.

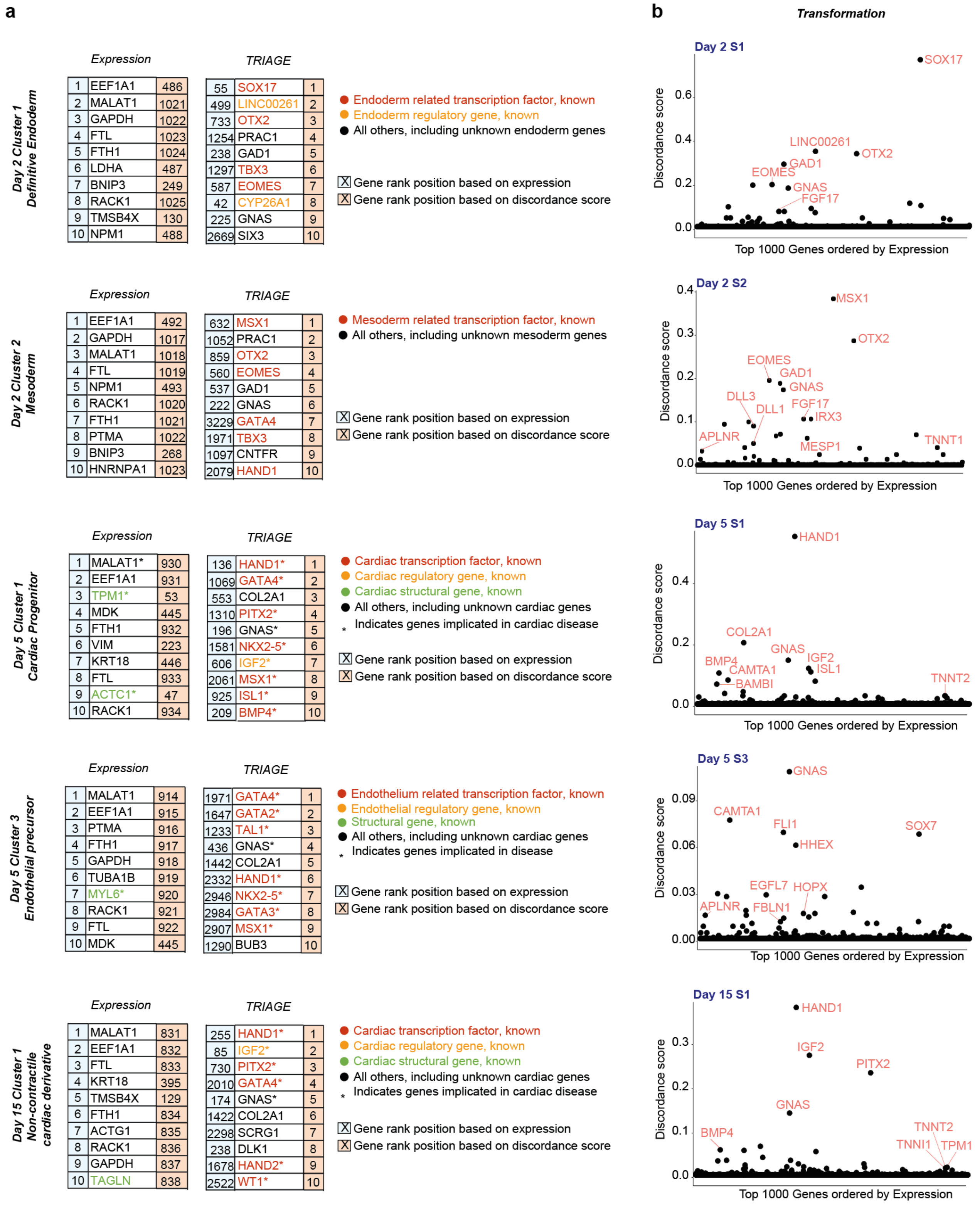
(e) Functional enrichment of top 100 genes ranked by TRIAGE, H3K27me3 loss (ΔH3K27me3) or expression value, for definitive cardiomyocyte (day 14) data[2]. For the full GO term list, see Supplementary table 9.

(f) Discordance score (*y*-axis) and H3K27me3 loss (ΔH3K27me3, *x*-axis) for selected cardiac specific regulatory genes (*GATA4, GATA6, HAND1, TBX5, NKX2-5 and TBX20*) for definitive cardiomyocyte (day 14) data.

(g) Enrichment of developmental GO terms among only expressed TFs for hiPSC data[3] (Fisher's exact test, one-tailed). TFs are sorted by a given metric (i.e. DS, differential expression (DE) fold-change or expression value) and binned into a percentile bin (*x*-axis). Also, top 1% TFs are extracted for detailed functional enrichment analysis. For the full list, see Supplementary table 10.

(h) Enrichment of TFs with a developmental GO term specific for 3 distinct Roadmap tissue types (Fisher's exact test, one-tailed). Genes are sorted by a given metric (i.e. DS, differential expression (DE) fold-change or expression value) and binned into a percentile bin (*x*-axis).

(i) Overlap of top ranked genes identified by different approaches. The number of genes compared is determined by the number of nearest active genes to SEs.



Day 2 Cluster 1 Definitive Endoderm

Expression

1	EEF1A1	486
2	MALAT1	1021
3	GAPDH	1022
4	FTL	1023
5	FTH1	1024
6	LDHA	487
7	BNIP3	249
8	RACK1	1025
9	TMSB4X	130
10	NPM1	488

TRIAGE

55	SOX17	1
499	LINC00261	2
733	OTX2	3
1254	PRAC1	4
238	GAD1	5
1297	TBX3	6
587	EOMES	7
42	CYP26A1	8
225	GNAS	9
2669	SIX3	10

● Endoderm related transcription factor, known
● Endoderm regulatory gene, known
● All others, including unknown endoderm genes

☒ Gene rank position based on expression
☒ Gene rank position based on discordance score

Day 2 Cluster 2 Mesoderm

Expression

1	EEF1A1	492
2	GAPDH	1017
3	MALAT1	1018
4	FTL	1019
5	NPM1	493
6	RACK1	1020
7	FTH1	1021
8	PTMA	1022
9	BNIP3	268
10	HNRNPA1	1023

TRIAGE

632	MSX1	1
1052	PRAC1	2
859	OTX2	3
560	EOMES	4
537	GAD1	5
222	GNAS	6
3229	GATA4	7
1971	TBX3	8
1097	CNTFR	9
2079	HAND1	10

● Mesoderm related transcription factor, known
● All others, including unknown mesoderm genes

☒ Gene rank position based on expression
☒ Gene rank position based on discordance score

Day 5 Cluster 1 Cardiac Progenitor

Expression

1	MALAT1*	930
2	EEF1A1	931
3	TPM1*	53
4	MDK	445
5	FTH1	932
6	VIM	223
7	KRT18	446
8	FTL	933
9	ACTC1*	47
10	RACK1	934

TRIAGE

136	HAND1*	1
1069	GATA4*	2
553	COL2A1	3
1310	PITX2*	4
196	GNAS*	5
1581	NKX2-5*	6
606	IGF2*	7
2061	MSX1*	8
925	ISL1*	9
209	BMP4*	10

● Cardiac transcription factor, known
● Cardiac regulatory gene, known
● Cardiac structural gene, known
● All others, including unknown cardiac genes
* Indicates genes implicated in cardiac disease

☒ Gene rank position based on expression
☒ Gene rank position based on discordance score

Day 5 Cluster 3 Endothelial precursor

Expression

1	MALAT1	914
2	EEF1A1	915
3	PTMA	916
4	FTH1	917
5	GAPDH	918
6	TUBA1B	919
7	MYL6*	920
8	RACK1	921
9	FTL	922
10	MDK	445

TRIAGE

1971	GATA4*	1
1647	GATA2*	2
1233	TAL1*	3
436	GNAS*	4
1442	COL2A1	5
2332	HAND1*	6
2946	NKX2-5*	7
2984	GATA3*	8
2907	MSX1*	9
1290	BUB3	10

● Endothelium related transcription factor, known
● Endothelial regulatory gene, known
● Structural gene, known
● All others, including unknown cardiac genes
* Indicates genes implicated in disease

☒ Gene rank position based on expression
☒ Gene rank position based on discordance score

Day 15 Cluster 1 Non-contractile cardiac derivative

Expression

1	MALAT1	831
2	EEF1A1	832
3	FTL	833
4	KRT18	395
5	TMSB4X	129
6	FTH1	834
7	ACTG1	835
8	RACK1	836
9	GAPDH	837
10	TAGLN	838

TRIAGE

255	HAND1*	1
85	IGF2*	2
730	PITX2*	3
2010	GATA4*	4
174	GNAS*	5
1422	COL2A1	6
2298	SCRG1	7
238	DLK1	8
1678	HAND2*	9
2522	WT1*	10

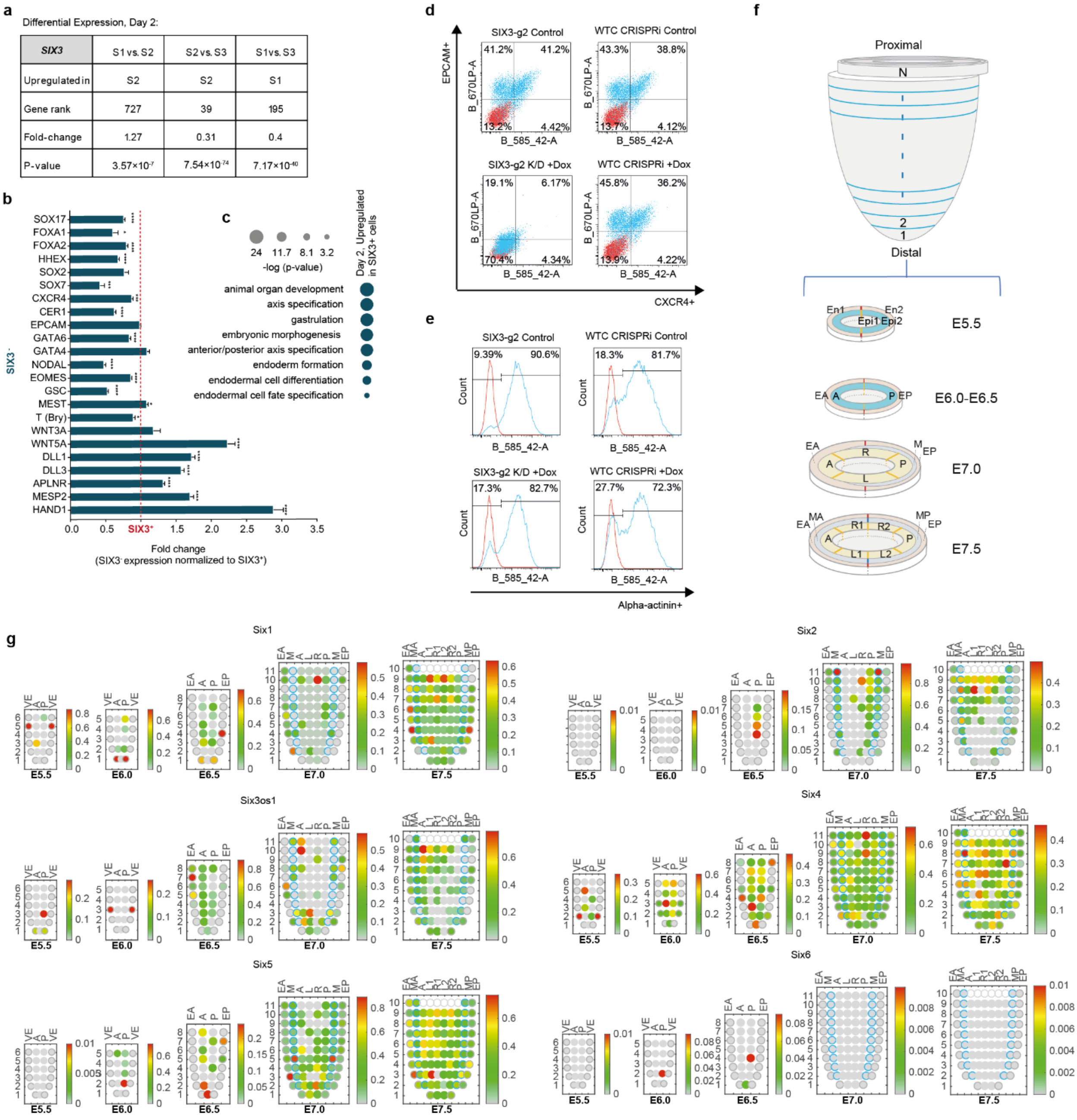
● Cardiac transcription factor, known
● Cardiac regulatory gene, known
● Cardiac structural gene, known
● All others, including unknown cardiac genes
* Indicates genes implicated in cardiac disease

☒ Gene rank position based on expression
☒ Gene rank position based on discordance score

Supplementary Figure 5: TRIAGE identifies known cell-type specific regulatory genes across diverse cell populations found during *in vitro* cardiac-directed differentiation (Supplementary to Figure 4).

(a) Top 10 genes ranked by the expression value or the discordance score from diverse cell populations from scRNA-seq of *in vitro* cardiac directed differentiation during germ layer specification (Day 2), cardiac progenitor specification (Day 5) and cardiomyocyte specification (Day 30).

(b) Transformation of the transcriptomic expression profile to the discordance score prioritizes known cell type specific TFs and regulatory genes across diverse cell populations.



Differential Expression, Day 2:

SIX3	S1 vs. S2	S2 vs. S3	S1vs. S3
Upregulated in	S2	S2	S1
Gene rank	727	39	195
Fold-change	1.27	0.31	0.4
P-value	3.57×10^{-7}	7.54×10^{-74}	7.17×10^{-40}

Supplementary Figure 6: Additional information on *SIX3*, a gene identified by TRIAGE to have a novel role in germ-layer specification during *in vitro* cardiac-directed differentiation (Supplementary to Figure 6).

(a) *SIX3* gene rank using differential expression analysis comparing clusters on Day 2 of scRNA-seq of *in vitro* cardiac directed differentiation shows that SIX3 was not prioritized in this expression-based analysis.

(b) Analysis of gene expression comparing *SIX*$^+$ vs *SIX3*$^-$ cells on Day 2 of scRNA-seq dataset assessing a panel of germ layer specification genes including markers of endodermal lineages (*SOX17, FOXA2, FOXA1, HHEX, SOX2, SOX7, CXCR4, CER1, GATA6*), mesendoderm (*EPCAM. NODAL, EOMES, GSC*) and mesodermal lineages (*GATA4, T-Bry, WNT3A, WNT5A, DLL1, DLL3, APLNR, MESP2, HAND1*).

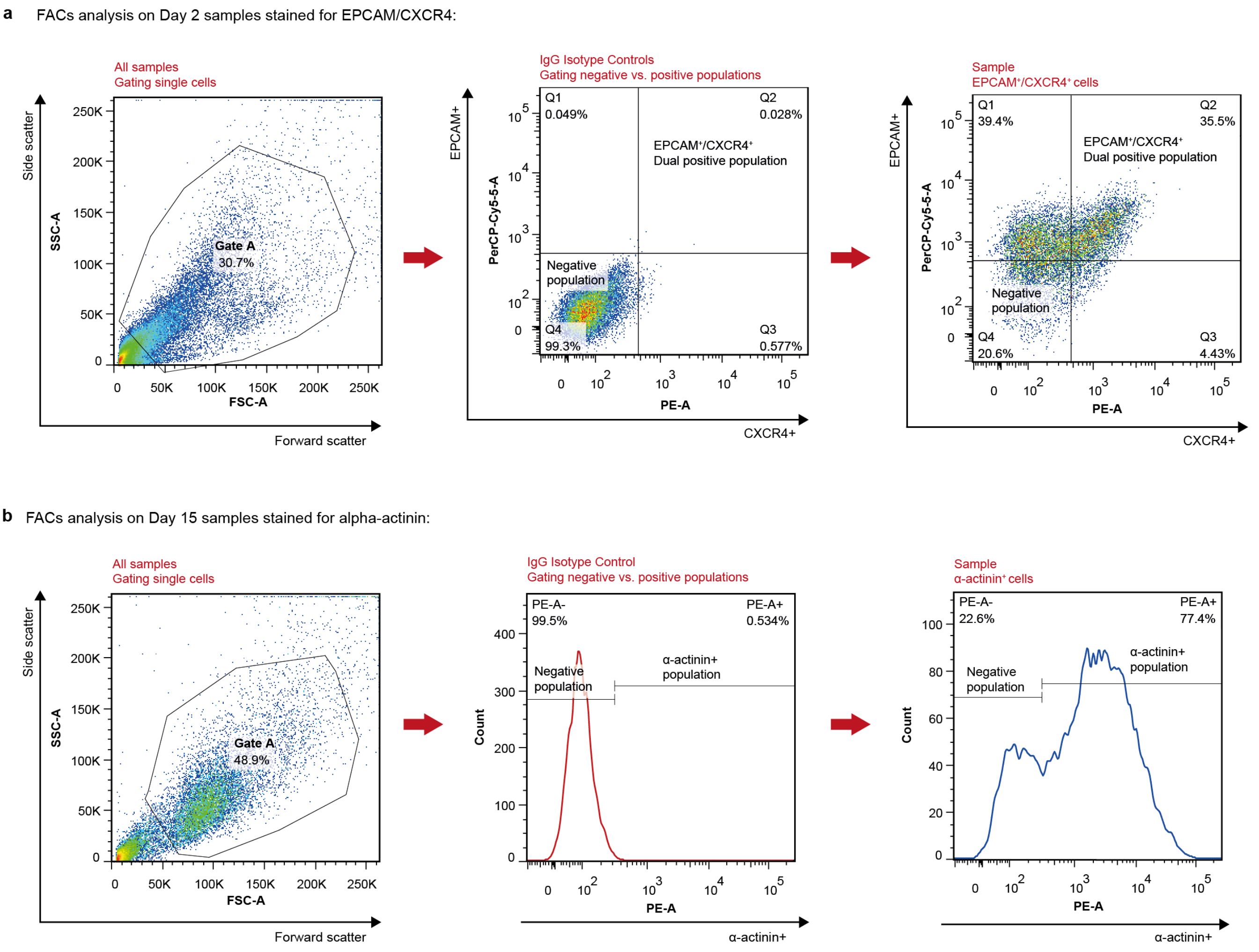
(c) Gene ontology (GO) analysis on day 2 of differentially expressed genes between *SIX3*$^+$/$^-$ populations displaying GO terms upregulated in *SIX3*+ cells.

(d) Raw FACS plots of EPCAM/CXCR4 analysis for all conditions tested (*n*=12-16 technical replicates per condition from 4-5 experiments).

(e) Raw FACS plots of alpha-actinin analysis for all conditions tested (*n*=6 technical replicates per condition from 3 experiments).

(f) Sample collection from E5.5-E7.5 embryos for analysis of spatiotemporal transcription. Positions of the cell populations in the embryo: the proximal-distal location in descending numerical order (1 = most distal site, N value of the most proximal section varied by the proximal-distal size of the embryo) 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).
(g) Corn plots showing spatial domains of *SIX* family genes expression in the germ layers of E5.5, E6.0, E6.5, E7.0 and E7.5 mouse embryos. The "kernels" in the plot represent the cell populations at different positions in the tissue layers of the embryo (panel **f**). Gradient scale shows the level of gene expression (by RNA-seq transcript reads) in each kernel.



a FACs analysis on Day 2 samples stained for EPCAM/CXCR4:

b FACs analysis on Day 15 samples stained for alpha-actinin:

Supplementary Figure 7. Flow cytometry gating strategy for analyses performed using SIX3-K/D iPSCs (Supplementary to Figure 6)

(a) For Day 2 samples dual-stained for EPCAM/CXCR4: first, all samples were gated to select singlets, excluding debris and doublets (Gate A= SSC-A vs. FSC-A). Singlets from fluorophore-specific IgG isotype controls (negative control) were used to distinguish between negative and positive populations in two parameter density plots (PerCP-Cy5.5-A vs. PE-A).

(b) For Day 15 samples stained for alpha-actinin: first, all samples were gated to select singlets (Gate A= SSC-A vs. FSC-A). Singlets from fluorophore-specific IgG isotype controls were used to distinguish between negative and positive populations in single parameter histograms (Count vs. PE-A).

Supplementary Tables

Supplementary Table 1. Descriptions of data sets used to estimate repressive tendency scores (RTS) and test performance of TRIAGE. These includes all 111 consolidated epigenomes from the NIH Roadmap project and a subset of 46 tissue or cell types with RNA-seq data, 329 CAGE-seq samples from the FANTOM5 project, 30 proteomes for distinct tissues and datasets from different species.

Supplementary Table 2. Lists of 634 variably expressed TFs (VETs) and tissue type specific regulatory genes used as the positive gene set for the performance analysis. We defined positive gene sets for the 5 distinct tissue types (i.e. heart, brain, lung, pancreas and skeletal muscle) by identifying TFs associated with a GO biological process term specific to a given tissue type.

Supplementary Table 3. Details of broad H3K27me3 domains from the 111 Roadmap cell types and the repressive tendency score (RTS) table.

Supplementary Table 4. Enrichment of KEGG pathway terms and selected GO biological process terms. Genes were ranked by the RTS and enrichment of a specific KEGG pathway or GO BP term was analyzed using Fisher's exact test (one-sided) across rank positions. Values are p-values or proportions.

Supplementary Table 5. Expression and discordance score (DS) across 12 different tissue types from Tabula Muris data sets. TRIAGE was applied to the expression data from two different single-cell RNA-seq platforms (i.e. 10x and SmartSeq2). Expression values of genes were averaged from all samples belonging to the tissue type.

Supplementary Table 6. Genes frequently ranked within top 50 by the expression value or the DS for proliferative and invasive melanoma states. Genes were sorted by the association frequency.

Supplementary Table 7. Functional enrichment for top ranked n genes (where n is the number of active genes nearest to super-enhancers, see Methods), comparing between TRIAGE and SE-based approach across 5 different Roadmap tissue types. Values shown are false discovery rate by Benjamini-Hochberg.

Supplementary Table 8. Enrichment of 19 GO terms relevant to embryonic neural differentiation extracted from Rehimi et al.[1], comparing between TRIAGE and functional heterogeneity (FH) score. Values shown are p-value at a given rank percentile bin (Fisher's exact test, one-tailed).

Supplementary Table 9. Functional enrichment of top 100 genes ranked by TRIAGE, H3K27me3 loss or expression value for definitive cardiomyocyte data (day 14)[2] (Fisher's exact test, one-sided).

Supplementary Table 10. Functional enrichment for top 1% TFs ranked by TRIAGE (Discordance), fold-change from DEG analysis and raw expression value.

References

1. Rehimi, R. *et al.* Epigenomics-Based Identification of Major Cell Identity Regulators within Heterogeneous Cell Populations. *Cell Rep* **17**, 3062-3076 (2016).
2. Paige, S.L. *et al.* A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* **151**, 221-32 (2012).
3. 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. e8 (2018).