Comprehensive perturbation of transcription factors in human cardiomyocytes reveals the regulatory architecture of congenital heart disease

Chikara Takeuchi[1,*], Sushama Sivakumar[2,*], Anjana Sundarrajan[1,‡], Yihan Wang[1,‡], Sean C Goetsch[2], Huan Zhao[1], Lei Wang[1], Mpathi Nzima[1], Minnie Deng[2], Kartik N Kulkarni[2], Lin Xu[7,8], Jun Wu[6], Bruce A Posner[5], Maria H Chahrour[3,4,9,10,11], W. Lee Kraus[1,12], Nikhil V Munshi[1,2,4,6,13,#], Gary C Hon[1,12,14,#]

[1]Cecil H. and Ida Green Center for Reproductive Biology Sciences,
[2]Department of Internal Medicine, Division of Cardiology,
[3]Department of Neuroscience,
[4]Eugene McDermott Center for Human Growth and Development,
[5]Department of Biochemistry,
[6]Department of Molecular Biology,
[7]Quantitative Biomedical Research Center, Peter O'Donnell Jr School of Public Health,
[8]Department of Pediatrics, Division of Hematology/Oncology,
[9]Department of Psychiatry,
[10]Center for the Genetics of Host Defense,
[11]Peter O'Donnell Jr Brain Institute,
[12]Department of Obstetrics and Gynecology
[13]Hamon Center for Regenerative Science and Medicine,
[14]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA

*These authors contributed equally (co-first author).
‡These authors contributed equally (co-second author).
#Correspondence to: Gary.Hon@utsouthwestern.edu or Nikhil.Munshi@utsouthwestern.edu

Abstract

Over 100 genes have been implicated in congenital heart disease (CHD), yet the genetic basis for >50% of CHD remains unexplained. A key challenge is to define the regulatory architecture of CHD genes. Here, we systematically perturb 1983 transcription factors (TFs) during cardiomyocyte differentiation of human stem cells. Our analysis links TFs to gene expression phenotypes across scales and nominates TFs in cardiac cell fate commitment, dosage sensitivity, and transposable element regulation. By deriving TF-gene regulatory networks from experimental perturbation, we gain insights into how developmental networks are structured, use this map to interpret the regulatory architecture of CHD, and deorphanize TFs with roles in CHD that have been under-sampled by genetic studies. To extend these networks to include enhancer-TF-gene linkages, we also perturb 981 putative enhancers of TFs. Finally, we develop a deep learning transformer model to accurately predict perturbed TFs from altered transcriptomes, which we apply to interpret CHD patient transcriptomes. This reference map represents a foundational platform to model the functions of TFs and aids in interpreting CHD variants and mechanisms.

Highlights

1. Systematic perturbation of 1983 human TFs in cardiac differentiation.
2. Defining the regulatory architecture of CHD genes and cardiac cell fate commitment.
3. Nominating TFs in dosage sensitivity and regulation of transposable elements.
4. Extended enhancer-TF-gene networks improve fine mapping of patient-derived variants.
5. Perturbation models nominate TFs driving altered regulatory networks in patients.

Keywords: congenital heart disease, development, transcription factor, functional genomics, gene regulatory network, machine learning, enhancers

Introduction

The human genome encodes ~2000 transcription factors (**TF**s) that regulate diverse gene expression programs, notably for lineage specification[1]. Since cell fate commitment is essential for normal development, genetic mutations of TFs have been associated with developmental defects such as congenital heart disease (**CHD**). For example, GATA4, NKX2-5, and TBX5 have well-established roles in cardiac development in both mice and humans[2–4], and mutations in these TFs cause CHD[5–8]. More broadly, human genetic studies have implicated dozens of TFs and epigenetic regulators in CHD[9], suggesting that altered gene regulation is an important component of this disease. Furthermore, mutations in several chromatin regulatory proteins have been linked to CHD by comprehensive exome sequencing of parent-offspring trios[10]. However, the genetic basis for 54% of familial and 80% of sporadic CHD cases remains unknown[11], indicating that disease mechanisms remain incompletely understood.

Traditional strategies to define TF function have relied on computational inference of TF-gene regulatory networks. These correlative approaches often use gene co-expression, chromatin accessibility, TF binding, and motif enrichment to model the relationship between TFs and the genes they regulate. However, accurate inference of gene regulatory networks (**GRN**s) remains a well-recognized computational challenge[12,13]. An alternative approach is to experimentally perturb TFs and directly measure regulated genes, an approach accomplished in yeast[14]. While single-cell screens have recently been developed to perform genome-scale perturbations in human cells (Perturb-Seq)[15], they have been more challenging to implement in stem cell-based models of human development[16–19].

Here, we systematically perturb TFs and map their transcriptional activities to gain insights into cardiac development and disease. Using an optimized Perturb-Seq platform in human iPS cell-derived cardiomyocytes (**CM**s)[20], we characterize the impact of 1983 TFs during cardiac differentiation. We perform multiscale analysis linking TFs to transcriptome-wide phenotypes, altered activity of gene programs, and changes in expression of individual genes, which we use to derive TF-gene regulatory networks from experimental perturbation. Our analysis gives insights on how developmental networks are structured and the regulatory architecture of CHD. We use these maps of TF function to nominate understudied TFs in cardiac cell fate commitment and deorphanize TFs relevant to CHD that have been undersampled in human genetic studies. We also demonstrate the utility of this unique resource to experimentally identify dosage-sensitive TFs and to define regulatory relationships between TFs and transposable elements. Next, we extend this cardiac regulatory network to include enhancers that regulate TFs and impact cardiac cell fate commitment. Finally, we apply these perturbation maps to develop Percoder, a machine learning model that predicts underlying TF perturbations given transcriptomes from patient-derived cardiomyocyte disease models.

Results

Systematic perturbation of 1983 human TFs in cardiac differentiation

To comprehensively survey the impact of transcription factors on cardiac phenotypes, we perform Perturb-Seq on 1983 TFs (1747 DNA binding proteins and 236 epigenetic regulators) during iPSC differentiation to cardiomyocytes (**Fig. 1a**, **Supplementary Tables 1 and 2**). After filtering, this high quality dataset consists of 765,351 cells (ave 3588 transcriptome unique molecular identifiers (UMIs), ave 1004 sgRNA UMIs, ave 1.9 significant sgRNAs called per cell)(**Fig. 1b**). We confirm

on-target repression efficiency (**Extended Data Fig. 1a**) and replicate reproducibility among libraries (**Extended Data Fig. 1b-c**). Our differentiation protocol yields cell states across cardiomyocyte differentiation, from early CMs (FN1+) to late CMs (TNNT2+) (**Extended Data Fig. 1d**). Expectedly, we observe reduced cell recovery for well-known regulators of cardiac development including GATA4, NKX2-5, HAND2, ISL1, and TBX5 (**Extended Data Fig. 2a-b**).

Transcriptome-wide phenotypes prioritize 209 TFs in cardiomyocyte differentiation

Perturb-Seq enables the analysis of TF functions at multiple transcriptional resolutions. At the level of transcriptome-wide phenotypes, we adopt a principal components-based energy distance metric to identify TFs that impact global cell state (**Fig. 1c**)[21]. This analysis identifies 209 regulators with consistent sgRNAs that are significantly different from non-targeting negative controls (**Extended Data Fig. 3 and 4a-c**, **Supplementary Tables 3-7**). While most significant regulators are relatively highly expressed, a handful of lowly expressed TFs also exhibit significant impact on the transcriptome (**Fig. 1d**). Notably, significant regulators are highly enriched in known CHD genes (p=4.4E-14), and are enriched in TFs with de novo variants (p=1.7E-4) (**Fig. 1e**)[10,22,23].

Using energy distance to visually embed and cluster TFs, we identify several notable functional modules (**Fig. 1f**). Expectedly, we identify clusters representing the hypoxia pathway (Cluster 1: HIF1A, ARNT)[24], DNA damage (Cluster 24: TP53, RB1)[25,26], transcriptional coactivators (Cluster 27: CREBBP, EP300)[27], and the WNT pathway (Cluster 20: TCF7, PIN1)[28,29] (**Fig. 1f**). Interestingly, Cluster 2 contains regulators highly enriched in known CHD TFs, including TBX5 [2,3], SOX4 [30], CHD7[31], and SMAD2-4[10,32,33] (**Fig. 1g**). This observation indicates that CHD TFs share similar transcriptional impacts, and suggests potential shared mechanisms. In addition, Cluster 7 represents regulators with known functions in heart development (**Fig. 1h**), including MEF2A[34], MEF2C[35], and ISL1[36].

Through guilt-by-association, we can implicate lesser studied TFs from Clusters 2 and 7 to CHD and heart development. For example, CHAMP1 is a member of Cluster 7 (development) and human genetic studies have identified one CHD patient with a 1-MB deletion including CHAMP1 and one CHD patient with an early termination de novo variant in CHAMP1 (**Extended Data Fig. 4d**)[22,37,38]. Our Perturb-Seq analysis thus further prioritizes CHAMP1 as a regulator of cardiac differentiation and CHD. To test this hypothesis, we knocked down CHAMP1 and observed delayed cardiac differentiation by TNNT2 staining (**Fig. 1i**). A similar analysis implicates the High-Mobility Group TF HMG20A (a member of Cluster 2) in cardiac cell fate commitment[39], as well as several zinc-finger containing genes (ZBTB12, ZC3H15, ZNF384)

Deorphanization of CHD TFs regulating cardiac differentiation

To increase the resolution of transcriptional phenotypes, we apply consensus, non-negative matrix factorization (cNMF)[40] **(Fig. 2a)** to define 250 gene programs (GPs) and 241 associated TF perturbations across perturbed transcriptomes (mean 7.2 TFs per GP, mean 6.9 GPs per TF, p_{adj} = 0.001) (**Extended Data Fig. 5a**, **Supplementary Tables 8 and 9**, **Supplementary Note 2**). Gene Program 1 (GP1) is highly enriched in CHD genes[9] (p_{adj} = 1.6e-06), including MYH7, MYH6, and MEIS2, as well as cardiac development and differentiation ontologies **(Fig. 2b, Supplementary Tables 10 and 11)**. We identify 24 TF-perturbations that down-regulate GP1 (**Fig. 2c, Supplementary Table 12**), many of which have been genetically associated with CHD and have well-known roles in cardiac development. These TFs have distinct differentiation phenotypes. For example, knockdown of TBX20 or NKX2-5 causes severe stalling of

differentiation, while GATA4 knockdown cells exhibit an attenuated phenotype (**Fig. 2d**). Independent qPCR and immuno-staining experiments confirm the roles of these TFs in regulating markers of cardiac differentiation (**Fig. 2e, Extended Data Fig. 5b-c**). Consistently, the findings from Perturb-Seq analyses in our *in vitro* model of human development are also recapitulated by *in vivo* knockout studies from the International Mouse Phenotyping Consortium[41] (**Extended Data Fig. 5d**). These results indicate that altered regulation of cardiac differentiation is a common mechanism for CHD pathogenesis.

Four TFs (ZSCAN10, GATA4, BPTF, CHAMP1) regulating GP1 have been implicated in CHD cohort sequencing studies (**Supplementary Table 13**)[10,22,23], but below the threshold of statistical significance as these studies are still underpowered. For example, BPTF, ZSCAN10 and CHAMP1 are not commonly mutated in CHD patients, but there have been isolated instances of de novo mutations identified in CHD patients (missense for ZSCAN10 and BPTF, and nonsense for CHAMP1) (**Extended Data Fig. 4d**). Importantly, these results are corroborated by independent TNNT2 immuno-staining experiments (**Fig. 1i**, **Extended Data Fig. 5e**). Since mis-regulation of cardiac differentiation is a common feature for CHD pathogenesis, these results can be used to deorphanize several de novo TF mutations associated with CHD.

Beyond GP1, GP28 is also significantly enriched for CHD genes including HAND1, FOXC2, and PRDM6 (p_{adj} = 1.4e-03) (**Fig. 2f, Supplementary Table 14**), with functional enrichment for genes of the extracellular matrix and collagen-containing structures **(Fig. 2b, Supplementary Table 10)**. Our analysis indicates GP28 is regulated by TFs that are associated with CHD (AFF4, ZFPM2, and GATA4) as well as TFs whose roles in CHD remain uncharacterized (SOX11, GTF2I, and MTF2). These results further demonstrate the potential of this approach to potentially deorphanize novel TF regulators of CHD.

Co-regulation analysis reveals novel interactions between TFs and chromatin regulators

We noticed groups of TFs and chromatin regulators that often co-regulate multiple gene programs more than would be expected by chance (**Fig. 2g**). For example, we found that HIF1A and ARNT, both well-known regulators of the hypoxia response, co-regulate 29 gene programs (**Fig. 2h**). This analysis also identifies frequent co-regulation of the MEF2 family of TFs (MEF2A, MEF2C, and MEF2D) with the members of the Polycomb Repressive 1 Complex (PRC1) (PCGF6, PCGF2, and RNF2) (**Supplementary Table 15**). These TFs regulate the expression of multiple gene programs associated with CHD (**Supplementary Table 9**). In general, chromatin regulators do not directly bind DNA and thus require recruitment by context-dependent TFs, which makes target gene identification particularly challenging. Interestingly, we notice a partitioning of gene programs regulated by the distinct complexes of PRC1: PRC1.2 (containing PCGF2 and RNF2) generally regulates a distinct set of gene programs from PRC1.6 (PCGF6). In contrast, TF usage is more varied: MEF2D only co-regulates PRC1.2 programs while MEF2A and MEF2C co-regulates with both PRC1 complexes (**Fig. 2i**). To test if MEF2A coregulates genes with RNF2 (member of PRC1.2) we performed qPCR in individual MEF2A or RNF2 CRISPRi CMs. We confirm that HEY1 is modestly upregulated and NKX2-5 is downregulated when MEF2A or RNF2 is repressed during CM differentiation (**Fig. 2j, Extended Data Fig. 5f**). To understand the molecular mechanism behind co-regulation of the downstream genes, we performed co-immunoprecipitation of Myc-RNF2, and confirmed co-localization and physical interaction between RNF2 and MEF2A, at least in a heterologous overexpression system (**Fig. 2k,l**). Overall, this network analysis suggests possible molecular mechanisms of coordinated gene regulation by TFs and chromatin regulators.

The regulatory architecture of CHD TFs and genes

We next construct a gene regulatory network (GRN) with 25,615 TF-gene edges connecting 1316 TFs to 7122 unique genes that are differentially expressed upon TF knockdown (ave 19.5 edges per TF; ave 3.6 edges per gene.) (**Fig. 3a-b**, **Supplementary Table 16**). This network is consistent with a scale-free network with power-law exponent of 2.04 (**Extended Data Fig. 6a**), which has implications for the hub-like structure of GRN in cardiomyocytes.

Next, we assess the properties of Perturb-Seq-derived GRNs to understand how CHD genes are regulated. Applying reverse pagerank on CHD genes (**Fig. 3c**), we identify ISL1 as a highly influential regulator of CHD genes[36,42], followed by known cardiac regulators (TBX20, NKX2-5), and other CHD genes (SOX11[43], ZFPM2[44]). These TFs regulate a significant number of CHD genes, both positively and negatively (**Fig. 3d**). Many well-known CHD genes are connected by this network. Consistent with a scale-free network, while some CHD genes like MYH7 are regulated by as many as 71 TFs, others like GJA1 are only regulated by 6 TFs (**Fig. 3e**). Each of 50 CHD genes is regulated by on average 7.4 TFs, thus implicating TFs that could influence CHD by modifying the expression of CHD genes. For example, TBX20 and NKX2-5 regulate 18 and 12 CHD genes, respectively. Indeed, the top TFs genetically associated with CHD convergently regulate multiple CHD genes (**Fig. 3f**). For example, MYH7 is regulated by ISL1, TBX20, and NKX2-5. This regulation of multiple CHD genes underscores the impact of TFs in CHD pathogenesis by regulating cardiomyocyte developmental programs.

50 TFs have been associated with CHD[9]. To understand how the TF-TF regulatory network is structured to control the expression of developmental and disease genes, we examine the GRN subnetwork connecting these 50 CHD TFs (**Extended Data Fig. 6b**)[9]. The TF-TF network is consistent with a small world network (**Extended Data Fig. 6c**), highlighting their pivotal roles of central regulatory hub genes(HAND1, NKX2-5, PITX2, and TBX20), in cardiac development and CHD[45]. Network motifs further reveal TF crosstalk and compensatory mechanisms (**Extended Data Fig. 6d**). For example, NKX2-5 and TBX20 co-regulate multiple downstream TFs including MYOCD, which we independently confirm (**Extended Data Fig. 6e**). Consistent with studies in mice[46,47], we find that three regulators (HAND1, NKX2-5, and PITX2) form a feedback loop, representing a self-sustaining regulatory module that maintains stable expression patterns in heart development.

In summary, experimentally derived GRNs reveal a scale-free network with hub structure regulating CHD genes (**Fig. 3j**).This regulatory architecture where many TFs can converge to regulate CHD genes highlights the multiple paths to disease pathogenesis.

Regulation of transposable elements by TFs during cardiac differentiation

Our unbiased TF Perturb-seq dataset can also capture non-canonical transcripts, including transposable elements (TEs)[48]. Given recent evidence suggesting that TEs modulate gene expression through enhancer-like activity[49] and cause developmental disease by retrotransposition[50,51], we sought to identify TF-TE regulatory interactions. Thus, we examined how knockdown of TFs, especially Zinc-finger TFs, impacts the expression of TEs during cardiomyocyte differentiation[52]. Highlighting the sensitivity of Perturb-Seq to capture TF-TE regulatory interactions, we recover the published finding that knockdown of the KRAB zinc finger TF ZNF91 results in derepression of the SVA family of TEs[53] (**Extended Data Fig. 7a**). Consistently, these TF-TE interactions are supported by published TF binding at TEs[54] (**Extended Data Fig. 7b**). These results suggest that Perturb-Seq can efficiently capture TE regulation. Next,

we systematically identify 144 TF-TE regulatory relationships as those TEs that are differentially expressed upon TF knockdown (**Fig. 3g**, **Supplementary Table 17**) (FDR = 10%). Consistent with the role of Zinc finger TFs in repressing TEs during differentiation, knockdown of TF hits generally results in TE de-repression (**Fig. 3g, red**). Examples with potential relevance to cardiac development include CHAMP1 regulating MER61A, and RCOR2 regulating LTR2B, which we validate by qPCR (**Fig. 3h**). Consistently, these regulated TEs are down-regulated during normal in vivo human development in the PCW4 embryo[55], with higher expression in early cardiomyocytes (MYH7-) compared to late cardiomyocytes (MYH7+) (**Fig. 3i**). In addition, regulated TEs are enriched in evolutionarily young (Haplorhini and Simiiformes) stages (**Extended Data Fig. 7c**). Finally, we find that almost all Zinc finger TF hits regulate a single TE (**Extended Data Fig. 7d**) and that most TEs are regulated by a single TF (**Extended Data Fig. 7e**), consistent with their specialized role for targeting TEs[54]. One notable exception is the WNT pathway transcription factor TCF7, whose knockdown leads to the up-regulation of 54 TEs (**Extended Data Fig. 7d**), including multiple LINEs (**Extended Data Fig. 7f**), and HERV-H which functions as an insulator in cardiomyocytes[56].

Identifying dosage-sensitive TFs of cardiac cell fate commitment

CHD TFs can be dosage sensitive, which is exemplified by Holt-Oram Syndrome caused by TBX5 haploinsufficiency. Indeed, using the LOEUF metric of haploinsufficiency from gnomAD[57], we find that TF hits from Perturb-Seq are significantly enriched in dosage sensitive TFs (**Fig. 4a**). However, while LOEUF scores haploinsufficiency, it lacks resolution of cell type. Thus, we adopt a framework to distinguish dosage sensitive from dosage buffered TFs using Perturb-Seq data. We define dosage sensitive TFs as those where small changes in TF expression cause large changes in transcriptional phenotypes, compared to dosage buffered TFs that require large changes in TF expression (**Fig. 4b**). We distinguish these two possibilities by whether TF-phenotype relationships fit linearly (sensitive) or exponentially (buffered)[58]. (**Fig. 4c**, **Supplementary Table 18**). This analysis is made possible because we targeted each TF with multiple sgRNAs, each of which may have differential CRISPRi repression efficiency. This analysis confirms that TBX5 is dosage sensitive while RCOR2 is dosage buffered, which is validated by independent qPCR experiments (**Fig. 4d**). Expanding this analysis across all testable TFs where sgRNAs span a large dynamic range of repression, we identify 5 TFs that are dosage sensitive in CMs (ARID4B, RB1CC1, SOX4, TBX5, and ZNF608) (**Fig. 4e**). Notably, SOX4 mutations cause Coffin-Siris syndrome with heart phenotypes and SOX4 haploinsufficiency has been implicated in developmental disorders[30]. Consistently, these TFs also have low LOEUF scores, while dosage buffered TFs tend to have higher scores.

Enhancer regulatory networks connect non-coding CHD variants to TF phenotypes

Since enhancers regulate TFs and can contain CHD variants, a more comprehensive enhancer-TF-gene regulatory network can give further insights into cardiac differentiation and CHD. Thus, we prioritized Perturb-Seq experiments of 1267 enhancers, of which 981 are in the regulatory neighborhoods of 107 TFs (**Fig. 5a, Extended Data Fig. 8**, **Supplementary Tables 19 and 20**). Energy distance analysis identifies 116 genes and 194 enhancers as regulators that are significantly different from non-targeting negative controls (**Fig. 5b**, **Supplementary Tables 21-23**). Clustering and embedding analysis shows that enhancer perturbations phenocopy promoter perturbations on a transcriptome-wide level (**Fig. 5c**, **Extended Data Fig. 9**), including two enhancers regulating HOXB4 (**Fig. 5d, Extended Data Fig. 10a**) and three enhancers regulating TBX5 (**Fig. 5e**).

Accurate identification of enhancers is challenging because of the low detection sensitivity of scRNA-Seq. However, for enhancers of TFs, perturbing strong enhancers phenocopies promoter perturbation because both induce robust transcriptome-wide phenotypes. This is also true for weaker enhancers (**Fig. 5f**): CRISPRi of Element293 moderately represses TBX5 and displays modest transcriptome-wide similarity with promoter perturbation, as compared with stronger enhancers (Element 290/291/294) (**Fig. 5g**). These observations motivate us to develop a new approach to increase enhancer target gene sensitivity by examining the similarity of transcriptome-wide phenotypes (**Extended Data Fig. 10b-c**). This analysis identifies 207 enhancers that regulate the expression of 64 TFs (**Fig. 5h-i, Extended Data Fig. 10d**, **Supplementary Table 24**), compared to only 76 enhancers (77.6% overlap) identified from traditional differential expression analysis (**Fig. 4h, Extended Data Fig. 10e**). Enhancer perturbations also facilitate better characterization of TF dosage sensitivity by achieving a broad spectrum of repression levels. We observed that the dosage-sensitive TF TBX5 also exhibits a continuous, linear relationship between its expression levels and transcriptome-wide phenotypes (**Extended Data Fig. 11a–b, Supplementary Table 25**). In contrast, TBX20 displays a non-linear, dosage-buffered pattern (**Extended Data Fig. 11c–d**). These findings underscore the critical role of enhancers regulating dosage-sensitive TFs in development and disease, consistent with previous reports on the distal TBX5 enhancer [19,59].

We illustrate one example of how this enhancer-TF-gene network can help interpret CHD variants. We identify an enhancer (Enhancer781) associated with a non-coding CHD variant[60] about 4-kb from the gene for the transcription factor MEIS1. Perturb-Seq of Enhancer781 and the MEIS1 promoter share modest transcriptome-wide similarity (**Fig. 5j**), with upregulation of FBN2 and KDR and downregulation of TBX18 and GAP43, which we independently validate (**Fig. 5k**). Importantly, a previous study characterized a non-coding variant at Element781 that significantly modulates activity in an enhancer reporter assay[61] (**Fig. 5l**). Together, these data confirm that this non-coding variant contributes to CHD pathogenesis by modifying the activity of Element781 and the expression of the transcription factor MEIS1, resulting in altered expression of its downstream genes and programs (**Fig. 5m**).

Perturbation modeling nominates TFs driving altered regulatory networks in patients

One application of experimentally-derived TF-gene networks is to interpret transcriptomes from CHD patients. Thus, we developed **Percoder** (a portmanteau of Perturb-Seq + Decoder) to predict underlying TF perturbations given perturbed (but unlabeled) transcriptomes (**Fig. 6a, Extended Data Fig. 12a**). Percoder learns the transcriptional phenotypes of TF knockdown as a distribution over a set of cells using a Set transformer architecture[62], which we use to predict perturbation labels (See Supplementary Note) (**Extended Data Fig. 12b**). To test how accurately Percoder can learn perturbation labels from the transcriptome, we trained on 209 significant TF perturbations. We found that 48.7% of cell sets are correctly classified by cross validation (102x better than random). Notably, 193 out of 210 perturbations achieve AUC > 0.75 (**Fig. 6b; left**), with strong perturbations like sgNKX2-5 or sgGATA4 yielding near-perfect accuracy (**Fig. 6b; right**). Instances of misclassification are biologically meaningful examples where TFs have shared functions: sgARNT and sgHIF1a (components of the hypoxia pathway), MEF2A and MEF2C (cardiac MEF2 family), EP300 and CREBBP (histone acetyl-transferases), and TBX5 and GATA4 (cardiac TFs and known interactors) (**Fig. 6c**). Finally, we also test Percoder's generalizability on out-of-sample data using the enhancer perturbation dataset (**Fig. 6d and 5a**), where 86 shared TFs were perturbed (**Fig. 6e**). Using a conditional variational autoencoder to calibrate for batch effects in a shared embedding space (**Fig. 6f, Extended Data Fig. 12c and**

13)[63] (see **Supplementary Note 1**), we achieve prediction accuracy far exceeding random chance (**Fig. 6g**). Overall, these results indicate that Percoder can effectively learn diverse perturbation responses with out-of-sample data.

Next, we tested Percoder to predict dysregulated TFs from patient-derived iPSC-cardiomyocytes (iPSC-CMs) disease models. First, we analyze single-cell RNA-Seq datasets of iPSC-CMs derived from patients with Hypoplastic Left Heart Syndrome (HLHS)[64]. Percoder identified YAP1 as a top candidate dysregulated TF, consistent with the original published study identifying a defective YAP-regulated antioxidant response in these patients (**Fig. 6h**, **Supplementary Table 26**). Importantly, the original study also observed significant dysregulation of genes in the hypoxia pathway in patients, which is also consistent with Percoder's prediction of ARNT and HIF1A as likely dysregulated TFs in patient transcriptomes (**Fig. 6i**). We also apply Percoder to a single-cell RNA-Seq dataset of iPSC-CMs modeling a GLYR1 P496L point mutation found in CHD patients[65] (**Fig. 6j**). While the original study identified differentially expressed genes in heart development and mitotic cell cycle in the iPSC-CMs, Percoder goes further by predicting that these dysregulated networks could be driven by TBX20, RB1, and ATMIN (**Fig. 6k**, **Supplementary Table 27**). Analysis of scRNA-Seq data from iPSC-derived cardiomyocytes of a Hypoplastic Right Heart Syndrome patient[66], also identifies the myogenesis gene MTF1 as a potential driver (**Extended Data Fig. 12d**, **Supplementary Table 28**)[67]. Overall, these results highlight the utility of Percoder to interpret patient-derived transcriptomes to predict underlying perturbed transcriptional regulators, which could be targets for intervention.

Discussion

In this study, we experimentally perturbed and measured transcriptome-wide phenotypes for ~2000 human transcription factors during cardiomyocyte differentiation to illuminate their roles in cardiac gene regulation, development, and disease. Our analysis systematically catalogs the impact of each TF on the transcriptome, gene programs, and individual genes. This reference map of experimentally-defined regulatory interactions in cardiomyocytes represents a foundational platform to model the functions of TFs, as well as to interpret the impact of genetic variants and cellular perturbations on cardiac differentiation and CHD-relevant transcriptional phenotypes.

Human genetic studies have implicated transcriptional regulators in CHD. Early genome linkage studies implicated GATA4, NKX2-5, and TBX5 in familial CHD[2,7,68]. Consistent with their well-studied roles in cardiac development, these TFs exhibit significant transcriptional phenotypes in our knockdown studies. More recently, exome studies have identified de novo mutations for dozens of genes in sporadic CHD cases[10,22,23]. Many of the genes reaching statistical significance are transcription factors and epigenetic factors including CHD7, KDM5B, KMT2A, PITX2, SMAD2, and SMAD6. Notably, our CRISPRi studies show that these genes are necessary for proper cardiac cell fate commitment. However, as CHD cohort sequencing studies are undersampled, many genes with de novo mutations in a small number of patients have not reached statistical significance. By attributing experimental evidence to the functions of TFs, our functional datasets nominate additional TFs with roles in CHD, including: AHDC1, CHAMP1, CRAMP1L, EP300, KDM5B, KMT2A, LHX2, TADA2B, and TBX18. For example, exome sequencing identified a de novo nonsense mutation of CHAMP1 in a CHD proband[22] and our analysis shows that CHAMP1 knockdown significantly delays cardiomyocyte differentiation and phenocopies MEF2A and MEF2C knockdown (both of which are CHD TFs). This example highlights the utility of our

resource to deorphanize TFs with roles in CHD and to interpret the significance and mechanisms of CHD patient variants.

The lack of experimentally-derived gene regulatory networks, especially in biological systems relevant to cardiac development and CHD, has been a bottleneck in the field. Our comprehensive experimental mapping of TF-gene interactions has several implications. First, regarding the regulatory architecture of CHD, we find that some CHD genes like MYH7 act as “hubs” in the regulatory network and are regulated by many TFs, suggesting convergent molecular routes to CHD. In addition, these TFs can have varying impact on the expression of a CHD gene. For example, small changes in NKX2-5 have a large impact on MYH7 expression, whereas even large changes of TBX20 have a minor impact on MYH7. Second, regarding potential applications of experimentally defined regulatory networks, we find that machine learning models trained on these networks can be used to accurately predict perturbed TFs from unseen transcriptomes, even those from patient-derived cells. We anticipate that future maps of networks in other developmental states will expand our ability to predict the underlying TFs that are dysregulated in diseases across cellular contexts.

Figure and Table Legends

FIGURES

Fig. 1: Perturb-Seq identifies 209 transcription factors regulating cardiac differentiation.

a. Breakdown of the 1983 transcriptional regulators perturbed in this study.
b. Schematic of TF Perturb-Seq experiment design, filtering strategy and key statistics.
c. Energy distance analysis prioritizes 209 TFs significantly different from negative control sgRNAs (p=0.001). We calculated a TF-TF distance matrix, followed by clustering and embedding to visualize regulatory relationships.
d. Distribution of gene expression level for TFs perturbed in this study. Orange bars indicate significant TFs by energy distance.
e. Enrichment of TFs for known congenital heart diseases genes (middle) and TFs with de novo variants (right).
f. t-SNE embedding where each dot represents a TF perturbation. Colors indicate clusters. Labels of clusters and representative genes are annotated manually.
g. Enrichment of known CHD genes in each cluster. The names of TFs in Cluster 2 are listed. Magenta TF names indicate CHD genes within the cluster. Bold gene was tested in panel i.
h. Enrichment of heart developmental genes in each cluster. The names of TFs in Cluster 7 are listed. Orange TF names indicate heart developmental genes within the cluster. Bold gene was tested in panel i.
i. Immunofluorescent staining of TNNT2 in iPSC-CM of Control KD (Left), HMG20A KD (middle) and CHAMP1 KD (right).

Fig. 2: Perturb-Seq de-orphanizes CHD TFs and identifies novel TF-TF interactions.

a. Overview of gene program analysis, which simultaneously maps gene programs to: the group of genes representing the program, the cells expressing it, and the perturbations that regulate it.
b. Heatmap showing Gene Ontology terms associated with five gene programs.
c. Summary of Gene Program 1, associated with cardiac muscle development and CHD. (Left) Feature plots showing: expression of an early CM marker (FN1), expression of a late CM marker (TNNT2), and GP1 usage. (Center) List showing the top 10 genes associated with GP1. (Right) Volcano plot of TFs regulating GP1. Known CHD TFs are highlighted in red.
d. Feature plots illustrating the cells with sgRNAs targeting NKX2-5, TBX20, RCOR2, and GATA4.
e. Immunofluorescence staining of TNNT2 in iPSC-CM of Control KD, NKX2-5, TBX20, ROCR2, GATA4(Left to right). Note that the Control IF image is identical to Fig. 1I.
f. Summary of Gene Program 28, associated with CHD.
g. Schematic of co-regulation analysis of Gene Programs. We identified TF pairs that co-regulate multiple gene programs more often than expected by chance.
h. Volcano plot showing occurrence of co-regulation between TF pairs. Significant pairs (FDR<0.05) are indicated as red. (right) Example of co-regulation of 29 gene programs between HIF1A and ARNT.
i. An example of co-regulation between the MEF2 family and PRC1 complex genes. Note that while MEF1A and MEF2C are associated with both PRC1.2 (PCGF2) and PRC1.6 (PCGF6), MEF2D is only associated with PRC1.2.
j. qPCR results showing HEY1(left), and NKX2-5 (right) after MEF2A and RNF2 KD in iPSC-CM.

k. Immunofluorescence staining of Myc-RNF2 and fluorescence of MEF2A-GFP in the heterologous overexpression expression system.
l. Co-immunoprecipitation of MYC-RNF2 and MEF2A-GFP, followed by Western blotting, to confirm physical interaction between RNF2 and MEF2A.

Fig. 3: Regulatory architecture of CHD TFs, genes, and transposable elements.

a. Schematic to construct a gene regulatory network from Perturb-Seq data. An edge between a TF and a gene exists if 1) knockdown of the TF alters the expression of the gene and 2) the TF's motif exists in the gene's promoter.
b. A sub-network of 92 CHD genes regulated by TFs.
c. PageRank scores for each TF in the experimentally derived GRN.
d. Bar plot showing the number of CHD genes regulated by TFs. The number of total DEG (Green), up-regulation (Red), down-regulation (Blue) are shown. Bold indicates known CHD-related TFs from CHDgene[9].
e. Bar plot showing the number of TFs regulating CHD genes.
f. Venn diagram showing shared target of 4 cardiac TFs (TBX20, ISL1, SOX11, NKX2-5).
g. This heatmap globally summarizes TF-TE relationships. Rows indicate regulated TEs, and columns indicate targeted TFs in Perturb-Seq. Colors indicate significance scores of statistical enrichment between TF-TE interactions.
h. qPCR result validating the observed regulation in Perturb-Seq experiment.
i. Feature plots showing genes and TEs expressed in human fetal CMs in vivo.
j. Summary of the GRN analysis of CHD regulatory architecture.

Fig. 4: Analytical framework to identify TF dosage-sensitivity in hiPSC-CM

a. LOEUF distributions of significant TFs and the non-significant TFs.
b. Illustration of the concept of transcription factor dosage sensitivity. The schematic illustrates scenarios for dosage sensitive versus dosage buffered effects. Dots represent each gRNA targeting a given TF. An orange line shows linear fit (representing dosage sensitivity) and a blue line shows curved fit (representing dosage buffered).
c. Scatterplot showing TF dosage versus E-distance effect for (top) TBX5 and (bottom) RCOR2 promoter perturbations. Dots represent each sgRNA for the target TF. Lines for linear and curve fits are indicated.
d. Scatterplot showing LOEUF from gnomAD data and goodness-of-fit to curve function compared to linear fit. (top) TFs with better curve fit (dosage buffered by Perturb-Seq). (bottom) TFs with better linear fit (dosage sensitive by Perturb-Seq).

Fig. 5: Enhancer regulatory networks connect non-coding CHD variants to TF phenotypes.

a. Overview of enhancer Perturb-Seq experiment design, filtering strategy, and key statistics.
b. Prioritization of target genes and enhancers based on energy distance statistics. Each dot represents a perturbation of promoters or enhancers. Perturbations with $p<0.001$ in energy statistics are shown in orange. 116 promoters and 194 enhancers are prioritized.
c. t-SNE embedding where each dot represents a promoter or enhancer perturbation. Colors indicate clusters.
d. (top) Zoomed-in view of HOXB4 promoters and enhancers cluster in the embedding and genomic positions of these elements. ATAC-Seq is from Day12 WTC11-derived cardiomyocytes and H3K27ac is from D15 H9-derived cardiomyocytes[56]. (bottom) qPCR validation of downstream gene expression changes after promoter/enhancer CRISPRi.

e. (top) Zoomed-in view of TBX5 promoters and enhancers cluster in the embedding and genomic positions of these elements. (bottom) qPCR validation of downstream gene expression changes after promoter/enhancer CRISPRi.
f. Strategy to map enhancers to target genes with transcriptome-wide phenotypes.
g. Impact of enhancer perturbations on TBX5 expression and transcriptome-wide energy distance. Shown are all enhancers near TBX5. The x-axis indicates mean change in TBX5 expression (CPM) and the y-axis indicates energy-distance between promoter perturbations and enhancer perturbations. The right vertical dashed line indicates TBX5 expression with non-targeting perturbations and the left dashed line indicates TBX5 expression with TBX5 promoter KD.
h. Venn diagram of enhancers identified by differential expression analysis (DEG, green), compared to our transcriptome-wide approach (blue).
i. Impact of all perturbed enhancers on normalized target gene expression and transcriptome-wide energy distance. Color indicates the hits of enhancers found by different metrics.
j. Genomic track showing the MEIS1 gene locus with enhancer Element781. Non-coding CHD de novo variants at Element 781[60] are also shown.
k. scRNA-Seq and qPCR results of the downstream impact by MEIS1 KD or MEIS1 enhancer (Element781) KD.
l. MPRA result of reference and altered sequence at chr2:66439493-66439664 (overlapped with Element781). Y-axis indicates MPRA activity. Data From Xiao F et al., 2024[61].
m. Schematic summarizing a MEIS1 variant-enhancer-target-GeneProgram-Phenotype network.

Fig. 6: Machine-learning approach can predict key dysregulated TFs.

a. Schematic of Percoder framework, which uses a Set Transformer to pool and embed an input set of cells for the classification task of predicting perturbed genes.
b. Histogram showing distribution of AUC per perturbation label. Red dashed bar indicates AUC=0.75 (threefold). (Right) Example ROC curves of key cardiac TFs.
c. Heatmap of confusion matrix for TFs with related functions.
d. Schematic of out-of-sample prediction. External datasets (other Perturb-Seq experiments or patient-derived samples) are integrated into the shared embeddings with the TF Perturb-Seq dataset. We then use the Percoder classifier to predict dysregulated TFs in the set cells from the external single-cell RNA-Seq dataset.
e. Venn diagram showing shared perturbations from two batches of Perturb-Seq experiments: TF Perturb-Seq (Fig. 1) and enhancer Perturb-Seq (Fig. 5).
f. Feature plot of TNNT2 expression in UMAPs from (left) PCA space and (right) scVI[63].
g. Top k prediction accuracy for shared perturbations between two Perturb-Seq batches. Blue bar indicates accuracy in the classifier and orange bar is the result from random classifier.
h. We applied Percoder on scRNA-Seq data derived from HLHS patient-derived iPSCs differentiated to cardiomyocytes[64]. The volcano plot shows Percoder predictions of dysregulated regulators. Color indicates the Percoder score in the patient sample (not normalized relative to control).
i. Gene Ontology analysis of differentially expressed genes between control and patient samples shown in Xu X et al.
j. We applied Percoder on scRNA-Seq data derived from GLYR1 (P496L) patient-derived iPSCs differentiated to cardiomyocytes[65]. The volcano plot shows Percoder predictions of dysregulated regulators. Color indicates the Percoder score in the patient sample (not normalized relative to control).

k. Gene Ontology analysis of differentially expressed genes between control and patient samples shown in Gonzalez-Teran B et al.

EXTENDED DATA FIGURES

Extended Data Fig. 1

a. Knockdown efficiency of target genes in Perturb-Seq. Lowly expressed genes (Log10(CPMnon-targeing)<0.5) are excluded in the plot. The density plot is shown on the right.
b. Feature plot illustrating the cell distribution for each of the 28 TF Perturb-Seq libraries. Library 5 recovered fewer cells because a lower volume was recovered during droplet generation.
c. Feature plot of the transcriptome UMIs per cell.
d. Feature plot of the late-stage CM marker TNNT2 and the early stage CM marker FN1.

Extended Data Fig. 2

a. Heatmap showing the Pearson correlation of gRNA abundancy at Day 0, 7 (by bulk amplicon sequencing), 12 (by scRNA-Seq).
b. Comparison of sgRNA abundance: (left) Day 0 vs Day 7 (bulk vs bulk); (right) Day 0 vs Day 12 (bulk vs single-cell).

Extended Data Fig. 3

a. An example of using energy distance analysis to filter out poorly performing sgRNAs. One sgRNA targeting HIF1A (top, 62162262.23-P1P2-1) does not correlate with other sgRNAs targeting HIF1A and is an outlier. One sgRNA targeting ARNT (150848808.23-P1P2-2) has weaker correlation with other sgRNAs targeting ARNT.
b. Schematic of our strategy to filter out poorly performing sgRNAs. (left) We are given sgRNAs (g1, g2, g3, g4) targeting a given gene. Each dot represents a cell where the sgRNA is detected. (right) Our strategy compares each sgRNA to all possible combinations of other sgRNAs targeting the same gene.
c. Schematic of our strategy to filter out poorly performing sgRNAs. Continuing from (b), we then calculate the energy distance between cells with a given sgRNA versus cells from all possible combinations of sgRNAs targeting the same gene. Next, we rank order these comparisons by energy distance. If a given sgRNA exhibits significant skewing in this rank ordering, then the sgRNA is an outlier.
d. Example outlier analysis for sgRNAs targeting HIF1A, where we successfully identify 62162262.23-P1P2-1 as an outlier.

Extended Data Fig. 4

a. Bar plot showing significant genes called at different p-value cut-offs. Orange bars indicate the number of hits called for Olfactory Receptor genes (OR genes, which represent targeting negative controls).
b. Bar plot showing enrichment of OR genes in the significant hits at different p-values.
c. Heat map matrix of the energy distance between 209 significant perturbations.
d. Schematic of known CHAMP1 mutation in neuro developmental diseases and de novo mutations found in a CHD patient. Clinical information is from Jin et al[22].

Extended Data Fig. 5

a. Summary statistics for gene program analysis. (top) Shown is the distribution of the number of programs significantly associated with each TF perturbation. (bottom) Shown is the distribution of the number of TFs significantly associated with each gene program.
b. Summary of Gene Program 1 analysis. This heatmap shows the impact of significant TFs in GP1 (rows) on the expression of 300 genes in GP1 (columns). Red indicates up-regulation and blue indicates down-regulation upon CRISPRi.

c. We generated independent stem cells with knockdown of NKX2-5, TBX20, RCOR2, and GATA4. After differentiation to cardiomyocytes, we performed qPCR validation experiments on markers of differentiation (MYH7, FBG2, TNNT2).
d. For each significant TF associated with GP1, we cross-referenced early (E9.5) mouse cardiac phenotypes from the International Mouse Phenotyping Consortium[41]. TFs were sorted by impact on GP1. The top plot in blue represents the density of mouse phenotypes.
e. Immunofluorescence staining of TNNT2 in iPSC-CM of Control KD and ZSCAN10 (left to right). Note that the Control IF image is identical to Fig. 1i.
f. Manhattan plot showing differentially expressed genes in RNF2 KD. The x-axis indicates the genomic position, the y-axis indicates the "significance score" (equivalent to −log10(FDR)), and the size of the dots indicates the fold change of expression. HEY1 and NKX2-5 (validated genes in Figure 2j) are highlighted.

Extended Data Fig. 6

a. Scatterplot showing the degree distribution of nodes in the TF-gene regulatory network (power-law exponent: 2.04).
b. An experimentally-derived TF-TF regulatory network derived from Perturb-Seq data. Bold gene names indicate CHD-associated TFs.
c. Barplot showing clustering coefficient and average shortest path length in TF-TF regulatory network compared with random network with same number of nodes and edges.
d. Examples of network motifs of TF-TF interactions in the network.
e. qPCR result validating MYOCD expression change after TBX20 and NKX2-5 CRISPRi KD.

Extended Data Fig. 7

a. Recovery of a known regulatory interaction between ZNF91 and the expression of the SVA_D family of TEs.
b. (left) Comparison of TF-TE regulatory hits identified by Perturb-Seq (x-axis) and chromatin interactions identified by ChIP-Seq enrichment of Zinc Finger TFs at the TEs. (right) An example of ZNF561 ChIP-Seq enrichment at the LTR10C TE, which corroborates the TF-TE interaction identified by Perturb-Seq.
c. Bar plot showing distributions of regulated and all TEs by the evolutional age (from Dfam database).
d. Histogram showing the number of TEs regulated by each TF.
e. Histogram showing the number of TFs regulating each TE.
f. Number of the TEs regulated by TCF7 by the TE family.

Extended Data Fig. 8

(top) Feature plots for TNNT2 (late cardiac marker) and FN1 (early cardiac marker).
(bottom) Feature plot illustrating the cell distribution for each of the 26 Enhancer Perturb-Seq libraries.

Extended Data Fig. 9

t-SNE embedding where each dot represents a perturbation in the Enhancer Perturb-Seq experiment. Colors indicate clusters. Enhancers are labeled as “Element” plus a numerical identifier (see **Supplementary Table 19**).

Extended Data Fig. 10

a. Relative expression(Upper) and CPM (bottom) of HOXB cluster genes after CRISPRi of two enhancers near HOXB4.
b. Schematic of normalized energy distance. In this example, if the energy distance between background and promoter perturbation is 20 and the energy distance between promoter and enhancer perturbation is 8, then the normalized energy distance for the enhancer perturbation is 0.4.
c. Distribution of normalized e-dist between TF-OR genes and TF-neighbor regulatory elements. (OR: olfactory receptor genes, which serve as negative controls)
d. Distributions of normalized target expression level and normalized energy distance. Color indicates the genomic promoter-enhancer distance.
e. The proportion of identified enhancers in different metrics by target-enhancer distance. A number of enhancer hits are shown at the top of each bar.

Extended Data Fig. 11

a. Scatter plot showing responsiveness of the TF dosages. The x-axis indicates relative expression of TBX5, and the y-axis indicates the energy-distance between top 3 strongest promoter perturbations and each gRNA perturbation. Color indicates target enhancers/promoters near target genes.
b. (Upper) Line plot showing LOWESS regression fits between TBX5 relative expression and downstream targets (TECRL, NPPA, TRH, CORIN, PLXNA4). (Lower) targeted enhancers near TBX5 promoter.
c. Scatter plot showing responsiveness of the TF dosages. The x-axis indicates relative expression of TBX20, and the y-axis indicates the energy-distance between top 3 strongest promoter perturbations and each gRNA perturbation. Color indicates target enhancers/promoters near target genes.
d. (Upper) Line plot showing LOWESS regression fits between TBX20 relative expression and downstream targets (SLC8A1, HSPB7, TRH, TPM1, MLY3). (Lower) targeted enhancers near TBX20 promoter.

Extended Data Fig. 12

a. Schematic architecture of Percoder.
b. Bar plot showing Percoder's prediction accuracy depends on the number of cells used in the input cell sets.
c. Percoder prediction accuracy using the out of sample Enhancer Perturb-Seq dataset, when using PCA (blue) or scVI (orange) embeddings.
d. We applied Percoder on scRNA-Seq data derived from HLHS patient-derived iPSCs differentiated to cardiomyocytes[66]. The volcano plot shows Percoder predictions of dysregulated regulators. Color indicates the Percoder score in the patient sample (not normalized relative to control).

Extended Data Fig. 13

Feature plots illustrating the quality of single-cell data integration with (left) PCA and (right) scVI-based UMAP. Shown are the cells from two Perturb-Seq experiments (Figure 1 data; Figure 5 data), patient-derived cardiomyocytes from three published studies (Xu, Gonzalez, Yang), and two marker genes (TNNT2, FN1).

SUPPLEMENTARY TABLES

Supplementary Table 1: List of sgRNAs used in TF Perturb-Seq screen.
Supplementary Table 2: Key statistics of TF Perturb-Seq libraries.
Supplementary Table 3: Outlier analysis of sgRNAs in TF Perturb-Seq screen.
Supplementary Table 4: Permutation analysis for sgRNAs in TF Perturb-Seq screen.
Supplementary Table 5: Filtered list of non-targeting gRNAs used for background.
Supplementary Table 6: Bulk sgRNA sequencing at D0 and D7 of differentiation.
Supplementary Table 7: Summary of energy distance TF hits.
Supplementary Table 8: Top 300 genes in each gene program identified by cNMF.
Supplementary Table 9: Gene programs associated with at least 1 TF perturbation.
Supplementary Table 10: Gene Ontology terms associated with each gene program.
Supplementary Table 11: Overlap analysis between gene programs and known CHD genes.
Supplementary Table 12: Perturbations associated with Gene Program 1.
Supplementary Table 13: Overlap of energy distance hits with de novo CHD variants.
Supplementary Table 14: Perturbations associated with Gene Program 28.
Supplementary Table 15: Summary of TF-TF interaction analysis.
Supplementary Table 16: Differentially expressed genes associated with each perturbed TF.
Supplementary Table 17: Differentially expressed TEs associated with each perturbed TF.
Supplementary Table 18: Summary statistics for dosage analysis.
Supplementary Table 19: List of sgRNAs used in Enhancer Perturb-Seq screen.
Supplementary Table 20: Key statistics of Enhancer Perturb-Seq libraries.
Supplementary Table 21: Outlier analysis of sgRNAs in Enhancer Perturb-Seq screen.
Supplementary Table 22: Permutation analysis for sgRNAs in Enhancer Perturb-Seq screen.
Supplementary Table 23: Filtered list of non-targeting gRNAs used for background.
Supplementary Table 24: Summary of enhancer hits.
Supplementary Table 25: Summary of gRNA-level transcriptome/target gene effects in TBX5/TBX20 enhancer perturbations.
Supplementary Table 26: Summary of Percoder predictions in Xu et al dataset.[64]
Supplementary Table 27: Summary of Percoder predictions in Gonzalez-Teran et al dataset.[65]
Supplementary Table 28: Summary of Percoder predictions in Yu et al dataset.[67]
Supplementary Table 29: Primers used in this study.

Methods

Experimental details

Cell culture

Human pluripotent stem cells (WTC11 and H9) are cultured as previously described[20]. In short, H9/WTC11 cells were maintained under feeder-free conditions in mTeSR plus medium (STEMCELL Technologies) and incubated in 37 ºC 5% CO2. For Perturb-Seq, we used the WTC11-CLYBL-dCas9-KRAB cell line[69].

Cardiac differentiation

WTC11 and H9 are differentiated to cardiomyocytes as previously described[19,20]. In short, 3 or 4 µM CHIR99021 supplemented CDM3 (RPMI 1640; 0.5 mg/ml human albumin, ScienCell OsrHSA; 211 µg/ml L-ascorbic acid 2-phosphate) for 48 hr followed by 2 µM Wnt-C59 supplemented CDM3 for 48 hr with subsequent media changes every 2 days with CDM3 alone. On day 12 post-differentiation, samples were washed with DPBS and dissociated to single cells with Accutase.

Design of sgRNA library

sgRNA sequences are from Replogle et al[70]. 6 gRNAs per region are used for each transcript. We included 600 non-targeting gRNAs and 600 targeting gRNAs as a control in the final library (**Supplementary Table 1**). For enhancer Perturb-Seq, we designed gRNA using FlashFry[71]. 10 gRNAs per target are used (**Supplementary Table 19**)[72].

Perturb-Seq

Library construction, virus packaging, transduction of virus, single cell RNA-Seq is performed as previously reported[20]. Briefly, DNA oligos (Twist Bioscience) are cloned into LentiGuide(10X)-BFP-Puro (LW203), and electroporated in Endura Duo competent cells (Lucigen). For virus packaging, 8µg of these plasmids are transfected in HEK293T with the 6µg of PMD2.G and 2µg of psPAX2 (Addgene ID: 12259 and 12260). The media is changed the next day and viruses are collected after 48hr of media change. These viruses are transfected to the WTC11-CLYBL-dCas9-KRAB cells at 80% confluency using the standard protocol, followed by BFP-positive sorting and 2 weeks culture. These cells proceed to the cardiac differentiation. For scRNA-Seq, cells were loaded onto a Next GEM Chip N with hash-tag oligo superloading[20,73]. Raw data were processed according to the manufacturer (10x Genomics, User Guide, CG000513).

RNA extraction, cDNA synthesis and RT-qPCR.

RNA extraction was done using Trizol according to previously published protocols[20]. cDNA was synthesized using lunascript RT (NEB) according to manufacturers instructions. qPCR was performed using SYBR-green master mix and specific primers to detect transcript levels. Biorad's CFX 384 real time qPCR machine was used to quantify transcript levels.

Immunostaining, Immuno-precipitation and western blotting

For immunostaining, differentiated cardiomyocytes were grown on matrigel coated coverslips. Immunostaining was performed as previously described[74]. For immunoprecipitation and immunostaining in Figure 2, HeLa cells were transiently transfected with PCS2-Mef2A-GFP, PCS2-Myc-RNF2 constructs using Lipofectamine 3000 according to manufacturers instructions. MYC co-immunoprecipitation and immunoblotting was done as reported previously[75]. Input indicates 10% of lysates on western blot.

Analysis details

scRNA-seq data pre-processing, filtering low-quality cells and doublets

We used our in-house preprocessing pipeline Perturb-Seq-Processing-Pipeline available on GitHub (https://github.com/Hon-lab/Perturb-Seq-Processing-Pipeline/). In short, we mapped the transcriptome to hg38 with Cell Ranger (version 7.0.0, 10x Genomics). Cells with low UMI count or higher than 20% mitochondrial content were removed, as well as genes with fewer than 1 counts. Also, gRNAs and HTO are assigned to cells using FBA[76]. Singlet cells (only 1 HTO) with at least 1 gRNAa are used in the following analysis. For visualization, we applied UMAP based on 50 principal components from principal component analysis (PCA).

Energy distance overview

To compare the effect of perturbations and associations between perturbations, we use energy distance analysis[15,21,77] . To balance computational cost and precision, we applied principle component analysis(PCA) (described above), and kept the top 50 principle components for the downstream analysis.

For samples $x_1, x_1, \ldots, x_n$ from X and $y_1, y_1, \ldots, y_n$ from Y, energy distance E(X,Y) is calculated in the following formula:

$$E(X,Y) = 2A - B - C$$

where A, B, and C are simply averages of pairwise euclidean distances of population (1) X and Y, (2) X and X and (3) Y and Y:

$$A = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} ||x_i - y_j||, B = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||x_i - x_j||, C = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} ||y_i - y_j||$$

|| • || denotes the Euclidean norm of its argument. We implemented this calculation using pytorch to boost speed with the Tesla V100 32GB (see code availability).

Removing outlier gRNAs

The first step of analysis is removing outlier gRNAs within each target. Although CRISPRi has less off-target activity, we observed some examples of off-target activity gRNAs and weak gRNA, that affect associations between perturbations (**Extended Data Fig. 4a**). Some methods filter out cells based on target gene expression level[78] or differential expressed genes[79]. However, these methods are not appropriate for our setting because (1) gene expression is highly changing over time in differentiating dataset, and (2) cell states are not uniform even with non-targeting gRNAs and (3) quantification of transcription factor expression is sometimes difficult especially for transient or lowly expressed but important genes. Therefore, we propose a new algorithm based on energy-distance. Assumptions of this algorithm are (1) cells with the same gRNA have the same off-target and on-target (repression efficiency) effects (2) cells with outlier gRNAs have a different distribution in the embedding space. These assumptions motivate us to filter gRNAs based on metrics which do not depend on target gene expression level. This algorithm consists of three steps.

(1) *Filtering gRNA based on # of gRNAs and cells*
If less than 20 cells are assigned to a gRNA, this gRNA is not considered in the following analysis because it would be noisy to use a small number of cells in the analysis. After this filtering based on cell number, if only 1 or 2 gRNAs are left after drop-off or filtering out gRNAs with a small number of cells, all gRNAs are kept.If not (if the number of gRNAs

are more than 2), proceed to step 2.

(2) *Test equal distributions among gRNAs using DISCO*

First step is to distinguish whether outlier gRNAs are included in a given gRNA set. The purpose of this process is to judge outlier gRNAs are included in the gRNAs or not. Then, we tested the equal distribution of all gRNAs using DISCO (distance components) , which is a nonparametric extension of ANOVA using energy distance. Let $A_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$ is a set of cells with the gRNA i. To calculate F statistics of DISCO:

$$W = W(A_1, A_2, \ldots, A_K) = \sum_{j=1}^{K} \frac{n_j}{2} g(A_j, A_j)$$

$$\text{where } g(A_1, A_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} \| a_{1i} - a_{2m} \|$$

And total dispersion T is $T = \frac{N}{2} g(A, A)$ where A is the pooled sample of $(A_1, A_2, \ldots, A_K)$. Then, The between-sample energy statistic S is calculated as $S = T - W$.

$$D = \frac{S / (K-1)}{W / (N-K)}$$

This D is an analogue of F statistics in ANOVA and calculated empirically using a permutation test. Practically, we shuffled the label of gRNA assignment within tested gRNAs (we use 1000 permutations for each target). p-value is calculated by the following formula.

$p = \#(D_{permute} > D_{observe}) / \#(Permutation)$ where $D_{permute}$ represents for D statistics after permutation of gRNA labels and $D_{observe}$ represents observed D statistics with right labels. If p-value of DISCO is more than 0.05, gRNAs per target is judged as all equally distributed. In other words, each gRNA behaves similarly. If p-value is less than 0.05, these gRNAs are likely to contain outliers within these, therefore proceed to the "Finding outlier gRNAs" step.

(3) *Finding outlier gRNAs*

The assumption here is that "clean" gRNAs, which show no off-target activity and strong repression of targeted regions, are distributed similarly. Thus, the energy distance between cells with distinct clean gRNAs will be smaller than between outlier gRNAs. This idea is similar to the Local outlier factor analysis of k-nearest neighborhood method, with some modification: (1) to calculate p-value in an unbiased way and (2) to detect outlier gRNA in small numbers of samples (max 6 gRNAs per target).

1. Let g_1, g_2, g_3, g_4 be gRNAs targeting a given gene, and $X_m = \{x_{m,1}, x_{m,2}, \ldots, x_{m,n}\}$ are cells with g_m detected. Among these gRNAs, there may exist outlier gRNAs with distributions that significantly differ from other gRNAs. We propose the algorithm below to detect these outlier gRNAs.

2. Generate all possible pair combinations of gRNAs (targeting the same gene). For instance, given five gRNAs [A, B, C, D, E], we consider pairs such as [(A,B), (E)] or [(A,C,D), (B,E)] where both the total number and the combination of labels within each pair are arbitrary. We denote the first group as G1 and the second group as G2. For example, If we test the group of [(A,C,D), (B,E)], G1 is (A,C,D) and G2 is (B,E) (**Extended Data Fig. 4b**).

3. Calculate the Energy distance between G1 and G2 for all possible combinations.

4. Rank the calculated Energy distances in ascending order. If a gRNA is an outlier, the distances between groups containing this label and other groups should be larger, resulting in higher ranks for combinations including the outlier label (**Extended Data Fig. 4c**). If there is no outlier gRNA, combinations containing any given gRNA will be equally distributed or skewed to lower rank. Therefore, the frequency of combinations where a specific gRNA is skewed to higher rank can be used as a metric of how much this gRNA is outlier. We test this hypothesis using a hypergeometric test for each label, specifically examining the probability of combinations containing that label falling into the lower 50% of ranks. An adjusted p-value after Bonferroni correction is less than 0.05 is detected as outlier gRNAs and filtered out in the following analysis unless mentioned.

Detecting significant hit regions

After removing outlier gRNAs we performed a permutation test using energy distance as described[15,21]. In short, we calculate energy distance between cells with non-targeting control gRNAs and cells with targeting gRNAs. For 1 batch, we randomly subsampled 2000 non-targeting cells to faster computation, and 1000 permutation was performed per batch. We repeated this step for 20 batches to avoid a bias effect on choice of background non-targeting cells. In total 20000 permutations were tested per target. To note, p-values are consistent among different sets of 2000 non-targeting cells (**Supplementary Table 4**). In this study, perturbations whose permutation-based p-value is less than 0.05 are referred to as significant hits.

Association analysis, clustering, and embedding of perturbations

Within significant TF hits, we calculated energy distance between perturbations (**Extended Data Fig. 3**). Clustering of these significant hits are performed in the following method. First, distance matrices from energy distance analysis are embedded in high dimension space with Isomap (implemented by sklearn.manifold.Isomap). This step is important because biologically associated TF pairs such as EP300 and CREBBP (both are histone acetyltransferase), HIF1A and ARNT (both are hypoxia pathway genes) significantly show smaller distances than other TFs, whereas larger distances between perturbations generally have little information, and Isomap is able to retain information of local structure in distance space. After embedding in high dimension space with Isomap, we clustered perturbations using affinity propagation (implemented by sklearn.cluster.AffinityPropagation) in Isomap embedded space to find clusters in an unbiased way without defining the number of clusters.

Comparison between CHD genes and de novo variants

The list of congenital heart disease genes are obtained from CHDgene[9]. Among CHD genes, we extracted overlapped genes with our tested TFs, and calculated p-value of enrichment using a hypergeometric test. For de novo variants, we obtained the list of de novo variants from multiple studies[10,22,23].

Transposable element (TE) analysis

scRNA-Seq transcriptome reads were mapped with scTE[48]. Differentially expressed TEs were detected by pySpade[80]. The phylogenetic distribution and family of each TE is annotated from Dfam 3.3[81]. ZNF561 ChIP-Seq data shown in **Fig. 4E** is downloaded from the ENCODE consortium (https://www.encodeproject.org/experiments/ENCSR350GZC/).

TF dosage analysis

We performed energy distance analysis at the gRNA level compared to 2000 randomly selected cells with non-targeting gRNAs. We excluded gRNAs with less than 30 cells assigned. Next, we identified testable transcription factors for dosage analysis using the following criteria: (1) at least

3 gRNAs are analyzable after the previous filtering step (2) dynamic range of expression level (maximum normalized expression level - minimum normalized expression level) is more than 0.2 (3) Pearson correlation coefficient between normalized expression and energy distance is less than -0.5. (4) Constraint information is available in the gnomAD database. After these filtering steps, 24 TFs remained. To note, the bottleneck of this filtering step is dynamic range for each TFs because we observed that most gRNAs work efficiently. We fit this expression-energy distance plot with two different two-parameter functions that satisfy (0,0) potentially (non-targeting control). For linear fitting, $y = a(x-1)+b$ is used. For curve fitting, we used $y = a(exp(-b(x-1)) - 1)$ is used. In this formula, y is energy distance and x is relative expression. Data are fit by least squares method using curve_fit function in scipy.optimize library. Details can be found in the code, deposited on pubmed (see code availability). For the comparison between dosage sensitivity and the gnomAD database, we used transcripts marked as "canonical" in the gnomAD[57] v4.1.0 database.

Gene programs analysis

Gene program analysis was performed with consensus non-negative matrix factorization (cNMF)[40]. cNMF was run for 100 iterations for top 2000 highly variable genes using 0.1 as clustering metric. The optimal k (k=250) was identified based on unique gene ontology terms[82], significant perturbations identified, and recovery of a rare perturbation (REST) as a positive control.

Significant perturbation calculation. cNMF assigns a score for each cell for every program. The mean score was calculated for each target for each program. Similarly the mean score was calculated for non-targeting control. Log2 fold change was calculated between targeting and non-targeting for every program. Mann-Whitney U test was performed to determine significance and Benjamin-Hochberg for adjusted p-value. A cutoff P-value of 0.001 was used based on the olfactory receptor genes as negative control[83].

Public mouse datasets

Mouse heart phenotype data was obtained from International Mouse Phenotype Consortium[41]. We queried the website (https://www.mousephenotype.org/) for individual genes to identify the earliest heart phenotype in knockout mice.

Identification of TF-TF interactions from gene program analysis

TF-gene interactions were quantified across 250 gene programs. To assess the significance of observed interactions, we applied a bipartite shuffle algorithm, which preserves the number of nodes and edges while randomizing connections. To determine the best-fit distribution for modeling the null distribution of TF interactions, we used the Fitter package to compare multiple candidate distributions (normal, gamma, exponential, beta, lognormal, Poisson, and binomial). The most frequent best-fit distribution across all tested interactions was identified and used to estimate p-values. Adjusted p-values were calculated using the Benjamini-Hochberg method, with a significance threshold of 0.05. To ensure biologically meaningful interactions, we applied two additional filters: (1) Interactions with an **observed count <2** were removed; (2) Interactions between **two promoters of the same gene** were excluded.

Differential expression (DE) analysis

pySpade (version 0.1.6) is used to calculate differential expressed genes[80]. pySpade 'process' function takes the output files from our in-house pre-processing pipelines. The randomization method of the DErand function is set to be 'equal' for all the datasets. Complementary cells are used in the background sets except time-course datasets. Non-target cells are the background in

time-course datasets. All hits reported are FDR < 0.1, with fold change >20% as cutoff, and genes expressed in at least 5% of cells.

Motif scanning

To search for transcription factor motifs in the promoters of DE genes, the JASPAR database (2024) is used to scan the TF motifs within the promoter region of DE genes[84]. The promoter regions are defined as +/-250 bp of transcription start site.

Network Construction and visualization

To construct the CHD-centric gene regulatory network (GRN), CHD-associated genes were identified as previously described and subsequently classified as either transcription factors (TFs) or non-TFs.

For the sub-network regulating non-TF CHD genes, we utilized the global hit table generated by pySpade. Non-TF CHD genes were first identified from the gene_names column. These genes were then mapped to their corresponding regulatory regions (from the region column), which were associated with specific TF perturbations. This set of putative interactions was filtered to retain only those where a binding motif of the perturbing TF was present within the promoter region of the target non-TF CHD gene.

The resulting TF-to-gene interactions were used to construct a directed graph using the NetworkX library[85] (version 2.6.3). The network was then visualized using the Kamada-Kawai layout algorithm.

Data availability

The complete FASTQ data, sgRNA sequences, Differential Expression tables and all relevant metadata is available on the IGVF Consortium Data Portal (https://data.igvf.org/). The data used throughout this paper can be accessed through the following accession numbers and links:

- https://data.igvf.org/measurement-sets/IGVFDS5398VEFY/
- https://data.igvf.org/measurement-sets/IGVFDS3275TXAD/
- https://data.igvf.org/measurement-sets/IGVFDS9367JFNB/
- https://data.igvf.org/measurement-sets/IGVFDS6847GSEQ/
- https://data.igvf.org/measurement-sets/IGVFDS0860HOHE/
- https://data.igvf.org/measurement-sets/IGVFDS3599AKSM/
- https://data.igvf.org/measurement-sets/IGVFDS1042ITOF/
- https://data.igvf.org/measurement-sets/IGVFDS9568EYXJ/
- https://data.igvf.org/measurement-sets/IGVFDS8548HJBX/
- https://data.igvf.org/measurement-sets/IGVFDS0119GHBM/
- https://data.igvf.org/measurement-sets/IGVFDS2933CBPV/
- https://data.igvf.org/measurement-sets/IGVFDS1867JBBD/
- https://data.igvf.org/measurement-sets/IGVFDS6990PWDJ/
- https://data.igvf.org/measurement-sets/IGVFDS6980CXOR/
- https://data.igvf.org/measurement-sets/IGVFDS9699UXJW/
- https://data.igvf.org/measurement-sets/IGVFDS7696ONVJ/
- https://data.igvf.org/measurement-sets/IGVFDS4080NNVH/
- https://data.igvf.org/measurement-sets/IGVFDS6472JBHY/
- https://data.igvf.org/measurement-sets/IGVFDS0814DYDR/
- https://data.igvf.org/measurement-sets/IGVFDS7181KBXZ/
- https://data.igvf.org/measurement-sets/IGVFDS5151YCLI/
- https://data.igvf.org/measurement-sets/IGVFDS5304ZOXD/
- https://data.igvf.org/measurement-sets/IGVFDS6688BORQ/

- https://data.igvf.org/measurement-sets/IGVFDS1584ZJUM/
- https://data.igvf.org/measurement-sets/IGVFDS8493OWMF/
- https://data.igvf.org/measurement-sets/IGVFDS2653YSTF/

Code availability

Software related to this study are on GitHub:

- https://github.com/Hon-lab/TFperturb_CM
- https://github.com/Chikara-Takeuchi/Percoder
- https://github.com/Hon-lab/Perturb-Seq-Processing-Pipeline

Supplementary Note 1: Architecture of Percoder

1. Architecture of Percoder

The architecture of the Percoder is based on the Set Transformer architecture. A set of N cells is denoted as $C = \{x^{(i)}\}_{i=1}^{N}$ where $x^{(i)} \in R^{G}$ represents G-dimentional embedding of the i-th cell (e.g., raw or normalized counts, PCA components, etc.). First, each cell embedding $x^{(i)}$is passed through a Multi-layer perception (MLP), f_{Cell} , with ReLU activation, followed by a Dropout layer. This transforms the input from a G-dimensional vector to a G'-dimensional vector, so $x'^{(i)} \in R^{G'}$.

Next, the transformed cell set $C' = \{x'^{(i)}, p\}_{i=1}^{N}$ is processed by a **Set Attention Block (SAB)**. The structure of this SAB reflects the Set Transformer++ architecture, incorporating residual connections:

$$SAB(X) = X + rFF(H)), H = X + Multihead(X_{norm}, X_{norm}, X_{norm}; \omega))$$

where $X_{norm} = LayerNorm(X)$, and ω is the parameter for the Multihead attention layer. The output is denoted as $Z = SAB(C')$.

After the cell embeddings are processed by the attention mechanism, they are aggregated into a single vector using a Pooling by Multihead Attention (PMA) layer:

$$PMA(X) = S + rFF(X'), X' = S + Multihead(S, X_{norm}, X_{norm}; \omega'))$$

where $X_{norm} = LayerNorm(X)$ and $S \in R^{G'}$is a learnable seed vector. S is seed vector with G' dimension, and ω' is the parameter for the Multihead attention layer.

The resulting output vector, $V = PMA(Z) \in R^{G'}$ is then passed through a final MLP, f_{class} , with ReLU activation, and a Softmax layer to predict the probability for each class (i.e., perturbation).

$$P = Softmax(f_{class}(V)),$$

The final output $P \in R^{210}$is a probability distribution over the 210 classes.

2. Optimization of cell set-based classification problem

The reverse-perturbation problem is discussed in the previous foundation model studies[86,87]. The innovations in Percoder to predict perturbations is using sets of the cells for training, instead of single cells for training. We observe significant improvement of prediction accuracy if we use more cells for training/validation (**Extended Data Fig. 12b**). This indicates that heterogeneity in Perturb-Seq data is meaningful in a biological way to distinguish perturbations. A recently published perturbation prediction model called STATE also uses a similar approach in terms of using sets of the cells as input[88]. In the STATE model, the prediction accuracy from a set of cells dramatically outperforms one from using the mean of cells. This also suggests that heterogeneity of the cell transcriptome encodes biological information.

To achieve best prediction accuracy, we noticed that Online Label Smoothing[89] significantly improved final model prediction (**Fig. 5c**). Online label smoothing can learn conceptually similar labels. It is true that some of the TFs are known to form protein complexes (e.g. HIF1A and ARNT), or work in the same pathway (CREBBP and EP300).

We speculate that the loss function of OLS helps to distinguish these similar concepts, and leads to better classification accuracy.

3. Optimization of generalization
Finally, to achieve the generalizability of Percoder, we embedded scRNA-Seq using the conditional variational auto encoder (cVAE) using scVI[63]. During hyperparameter tuning, we noticed that some of the parameters are critical to generalizable embeddings. First ZINB modeling of transcriptome outperforms NB modeling. Second, not highly-variable genes but using all of the genes improves the final prediction accuracy in the out-of-sample experiment. These results indicate that ZINB models can express single-cell transcriptomes well and technological/biological effects are expressed even in the lowly expressed genes.

Supplementary Note 2: Discussion of gene programs identified by cNMF

- Differentiation programs: Program 1 was identified as a differentiation program based on mature cardiac genes and GO terms. Please see main text for more in-depth discussion.

- Metabolic programs: Among the significant programs 2 of them particularly had metabolism related GO terms. Program 13 had glycolysis process related terms and program 250 with cholesterol related terms. The metabolism related programs had depletion of MEF2C, ISL1 and CHAMP1 perturbation.

- CHD program: CHD program was chosen based on overlapping 142 genes from the CHDgene database[9] with the 300 genes defining each program. Program 1 and 28 were the top programs with significant overlap with CHD genes.

- Neuronal program: REST gene is known to repress neuronal genes[83] in non-neuronal lineages. Thus, repression of REST promoters is known to upregulate these neuronal genes. Gene Program 67 is enriched in neuronal genes, and is significantly associated with both targeted promoters of REST.

Acknowledgements

We acknowledge the BioHPC computational infrastructure at UT Southwestern for providing HPC and storage resources that have contributed to the research results reported within this paper. This work is supported by the National Human Genome Research Institute (NHGRI): UM1HG011996.

Competing Interests

The authors declare no competing interests.

Author contributions

Chikara Takeuchi: Methodology, Software, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization **Sushama Sivakumar:** Methodology, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization **Anjana Sundarrajan:** Methodology, Software, Formal analysis, Writing - Original Draft, Visualization **Yihan Wang:** Methodology, Formal analysis, Writing - Original Draft, Visualization **Sean C Goetsch:** Methodology, Investigation **Huan Zhao:** Investigation, Visualization **Lei Wang:** Investigation **Mpathi Nzima:** Data Curation **Minnie Deng:** Validation **Kartik N Kulkarni:** Validation **Lin Xu:** Funding acquisition **Jun Wu:** Funding acquisition **Bruce A Posner:** Funding acquisition **Maria H Chahrour:** Funding acquisition **W. Lee Kraus:** Funding acquisition **Nikhil V Munshi:** Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition **Gary C Hon:** Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

References

1. Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132**, 661–680 (2008).
2. Basson, C. T. *et al.* Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat. Genet.* **15**, 30–35 (1997).
3. Li, Q. Y. *et al.* Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat. Genet.* **15**, 21–29 (1997).
4. Williams, K., Carson, J. & Lo, C. Genetics of Congenital Heart Disease. *Biomolecules* **9**, (2019).
5. Kathiriya, I. S. *et al.* Modeling Human TBX5 Haploinsufficiency Predicts Regulatory Networks for Congenital Heart Disease. *Dev. Cell* (2020) doi:10.1016/j.devcel.2020.11.020.
6. McElhinney, D. B., Geiger, E., Blinder, J., Benson, D. W. & Goldmuntz, E. NKX2.5 mutations in patients with congenital heart disease. *J. Am. Coll. Cardiol.* **42**, 1650–1655 (2003).
7. Garg, V. *et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
8. Bruneau, B. G. *et al.* A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106**, 709–721 (2001).
9. Yang, A. *et al.* CHDgene: A curated database for congenital heart disease genes. *Circ. Genom. Precis. Med.* **15**, e003539 (2022).
10. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).
11. Zaidi, S. & Brueckner, M. Genetics and Genomics of Congenital Heart Disease. *Circ. Res.* **120**, 923–940 (2017).
12. Badia-I-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-

omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).

13. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
14. Hu, Z., Killion, P. J. & Iyer, V. R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.* **39**, 683–687 (2007).
15. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e28 (2022).
16. Nourreddine, S. *et al.* A perturbation cell atlas of human induced pluripotent stem cells. *bioRxivorg* (2024) doi:10.1101/2024.11.03.621734.
17. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365–372 (2023).
18. Leng, K. *et al.* CRISPRi screens in human iPSC-derived astrocytes elucidate regulators of distinct inflammatory reactive states. *Nat. Neurosci.* **25**, 1528–1542 (2022).
19. Armendariz, D. A. *et al.* CHD-associated enhancers shape human cardiomyocyte lineage commitment. *Elife* **12**, (2023).
20. Sivakumar, S. *et al.* Benchmarking and optimizing Perturb-seq in differentiating human pluripotent stem cells. *bioRxiv* (2025).
21. Peidli, S. *et al.* scPerturb: harmonized single-cell perturbation data. *Nat. Methods* **21**, 531–540 (2024).
22. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
23. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
24. Ke, Q. & Costa, M. Hypoxia-inducible factor-1 (HIF-1). *Mol. Pharmacol.* **70**, 1469–1480 (2006).

25. Horowitz, J. M. *et al.* Frequent inactivation of the retinoblastoma anti-oncogene is restricted to a subset of human tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2775–2779 (1990).
26. Di Leonardo, A., Linke, S. P., Clarkin, K. & Wahl, G. M. DNA damage triggers a prolonged p53-dependent G1 arrest and long-term induction of Cip1 in normal human fibroblasts. *Genes Dev.* **8**, 2540–2551 (1994).
27. Gerritsen, M. E. *et al.* CREB-binding protein/p300 are transcriptional coactivators of p65. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 2927–2932 (1997).
28. Behrens, J. *et al.* Functional interaction of beta-catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
29. Ryo, A., Nakamura, M., Wulf, G., Liou, Y. C. & Lu, K. P. Pin1 regulates turnover and subcellular localization of beta-catenin by inhibiting its interaction with APC. *Nat. Cell Biol.* **3**, 793–801 (2001).
30. Zawerton, A. *et al.* De Novo SOX4 variants cause a neurodevelopmental disease associated with mild dysmorphism. *Am. J. Hum. Genet.* **104**, 246–259 (2019).
31. Corsten-Janssen, N. *et al.* The cardiac phenotype in patients with a CHD7 mutation. *Circ. Cardiovasc. Genet.* **6**, 248–254 (2013).
32. van de Laar, I. M. B. H. *et al.* Mutations in SMAD3 cause a syndromic form of aortic aneurysms and dissections with early-onset osteoarthritis. *Nat. Genet.* **43**, 121–126 (2011).
33. Le Goff, C. *et al.* Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome. *Nat. Genet.* **44**, 85–88 (2011).
34. Naya, F. J. *et al.* Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor. *Nat. Med.* **8**, 1303–1309 (2002).
35. Lin, Q., Schwarz, J., Bucana, C. & Olson, E. N. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. *Science* **276**, 1404–1407 (1997).
36. Cai, C.-L. *et al.* Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev. Cell* **5**, 877–889 (2003).

37. Huang, C. *et al.* Congenital heart defect and mental retardation in a patient with a 13q33.1-34 deletion. *Gene* **498**, 308–310 (2012).
38. Yang, Y.-F. *et al.* A 1.1Mb deletion in distal 13q deletion syndrome region with congenital heart defect and postaxial polydactyly: additional support for a CHD locus at distal 13q34 region. *Gene* **528**, 51–54 (2013).
39. Herchenröther, A. *et al.* The H2A.Z and NuRD associated protein HMG20A controls early head and heart developmental transcription programs. *Nat. Commun.* **14**, 472 (2023).
40. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, (2019).
41. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).
42. McCulley, D. J. & Black, B. L. Transcription factor pathways and congenital heart disease. *Curr. Top. Dev. Biol.* **100**, 253–277 (2012).
43. Okamoto, N., Ehara, E., Tsurusaki, Y., Miyake, N. & Matsumoto, N. Coffin-Siris syndrome and cardiac anomaly with a novel SOX11 mutation. *Congenit Anom (Kyoto)* **58**, 105–107 (2018).
44. Pizzuti, A. *et al.* Mutations of ZFPM2/FOG2 gene in sporadic cases of tetralogy of Fallot. *Hum. Mutat.* **22**, 372–377 (2003).
45. Franco, D. & Campione, M. The role of Pitx2 during cardiac development. Linking left-right signaling and congenital heart diseases. *Trends Cardiovasc. Med.* **13**, 157–163 (2003).
46. Srivastava, D. Making or breaking the heart: from lineage determination to morphogenesis. *Cell* **126**, 1037–1048 (2006).
47. Olson, E. N. Gene regulatory networks in the evolution and development of the heart. *Science* **313**, 1922–1927 (2006).
48. He, J. *et al.* Identifying transposable element expression dynamics and heterogeneity

during development at the single-cell level with a processing pipeline scTE. *Nat. Commun.* **12**, 1456 (2021).

49. Cao, Y. *et al.* Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res* **29**, 40–52 (2019).
50. Gardner, E. J. *et al.* Contribution of retrotransposition to developmental disorders. *Nat Commun* **10**, 4630 (2019).
51. Chu, C., Ljungström, V., Tran, A., Jin, H. & Park, P. J. Contribution of de novo retroelements to birth defects and childhood cancers. *medRxiv* 2024.04.15.24305733 (2024).
52. de Tribolet-Hardy, J. *et al.* Genetic features and genomic targets of human KRAB-zinc finger proteins. *Genome Res.* **33**, 1409–1423 (2023).
53. Jacobs, F. M. J. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).
54. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
55. Zeng, B. *et al.* The single-cell and spatial transcriptional landscape of human gastrulation and early brain development. *Cell Stem Cell* **30**, 851–866.e7 (2023).
56. Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* **51**, 1380–1388 (2019).
57. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
58. Naqvi, S. *et al.* Precise modulation of transcription factor levels identifies features underlying dosage sensitivity. *Nat. Genet.* **55**, 841–851 (2023).
59. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263 (2012).
60. Richter, F. *et al.* Genomic analyses implicate noncoding de novo variants in congenital

heart disease. *Nat. Genet.* **52**, 769–777 (2020).

61. Xiao, F. *et al.* Functional dissection of human cardiac enhancers and noncoding de novo variants in congenital heart disease. *Nat. Genet.* **56**, 420–430 (2024).
62. Lee, J. *et al.* Set Transformer: A framework for attention-based permutation-invariant neural networks. *arXiv [cs.LG]* (2018).
63. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
64. Xu, X. *et al.* Uncompensated mitochondrial oxidative stress underlies heart failure in an iPSC-derived model of congenital heart disease. *Cell Stem Cell* **29**, 840–855.e7 (2022).
65. Gonzalez-Teran, B. *et al.* Transcription factor protein interactomes reveal genetic determinants in heart disease. *Cell* **185**, 794–814.e30 (2022).
66. Yu, Y. *et al.* Abnormal Progenitor Cell Differentiation and Cardiomyocyte Proliferation in Hypoplastic Right Heart Syndrome. *Circulation* **149**, 888–891 (2024).
67. Tavera-Montañez, C. *et al.* The classic metal-sensing transcription factor MTF1 promotes myogenesis in response to copper. *FASEB J* **33**, 14556–14574 (2019).
68. Schott, J. J. *et al.* Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* **281**, 108–111 (1998).
69. Yang, X. *et al.* Functional characterization of Alzheimer's disease genetic variants in microglia. *Nat Genet* **55**, 1735–1744 (2023).
70. Replogle, J. M. *et al.* Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors. *Elife* **11**, (2022).
71. McKenna, A. & Shendure, J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.* **16**, 74 (2018).
72. Xie, S., Armendariz, D., Zhou, P., Duan, J. & Hon, G. C. Global analysis of enhancer targets reveals convergent enhancer-driven regulatory modules. *Cell Rep.* **29**, 2570–2578.e5 (2019).

73. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
74. Sivakumar, S. *et al.* TP53 promotes lineage commitment of human embryonic stem cells through ciliogenesis and sonic hedgehog signaling. *Cell Rep.* **38**, 110395 (2022).
75. Sivakumar, S. *et al.* The human SKA complex drives the metaphase-anaphase cell cycle transition by recruiting protein phosphatase 1 to kinetochores. *Elife* **5**, (2016).
76. Duan, J. & Hon, G. C. FBA: feature barcoding analysis for single cell RNA-Seq. *Bioinformatics* **37**, 4266–4268 (2021).
77. Rizzo, M. L. & Székely, G. J. Energy distance. *Wiley Interdiscip. Rev. Comput. Stat.* **8**, 27–38 (2016).
78. Tian, R. *et al.* CRISPR interference-based platform for multimodal genetic screens in human iPSC-derived neurons. *Neuron* **104**, 239–255.e12 (2019).
79. Duan, B. *et al.* Model-based understanding of single-cell CRISPR screening. *Nat. Commun.* **10**, 2233 (2019).
80. Wang, Y. *et al.* Enhancer regulatory networks globally connect non-coding breast cancer loci to cancer genes. *Genome Biol.* **26**, 10 (2025).
81. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
82. Schnitzler, G. R. *et al.* Convergence of coronary artery disease genes onto endothelial cell programs. *Nature* **626**, 799–807 (2024).
83. Aoki, H., Hara, A., Era, T., Kunisada, T. & Yamada, Y. Genetic ablation of Rest leads to in vitro-specific derepression of neuronal genes during neurogenesis. *Development* **139**, 667–677 (2012).
84. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **52**, D174–D182 (2024).

85. Hagberg, A., Swart, P. & S Chult, D. *Exploring Network Structure, Dynamics, and Function Using NetworkX*. https://www.osti.gov/biblio/960616 (2008).

86. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

87. Zeng, Y. *et al.* CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nat Commun* **16**, 4679 (2025).

88. Adduri, A. K. *et al.* Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv* (2025) doi:10.1101/2025.06.26.661135.

89. Zhang, C.-B. *et al.* Delving deep into label smoothing. *arXiv [cs.CV]* (2020).

Figure 1

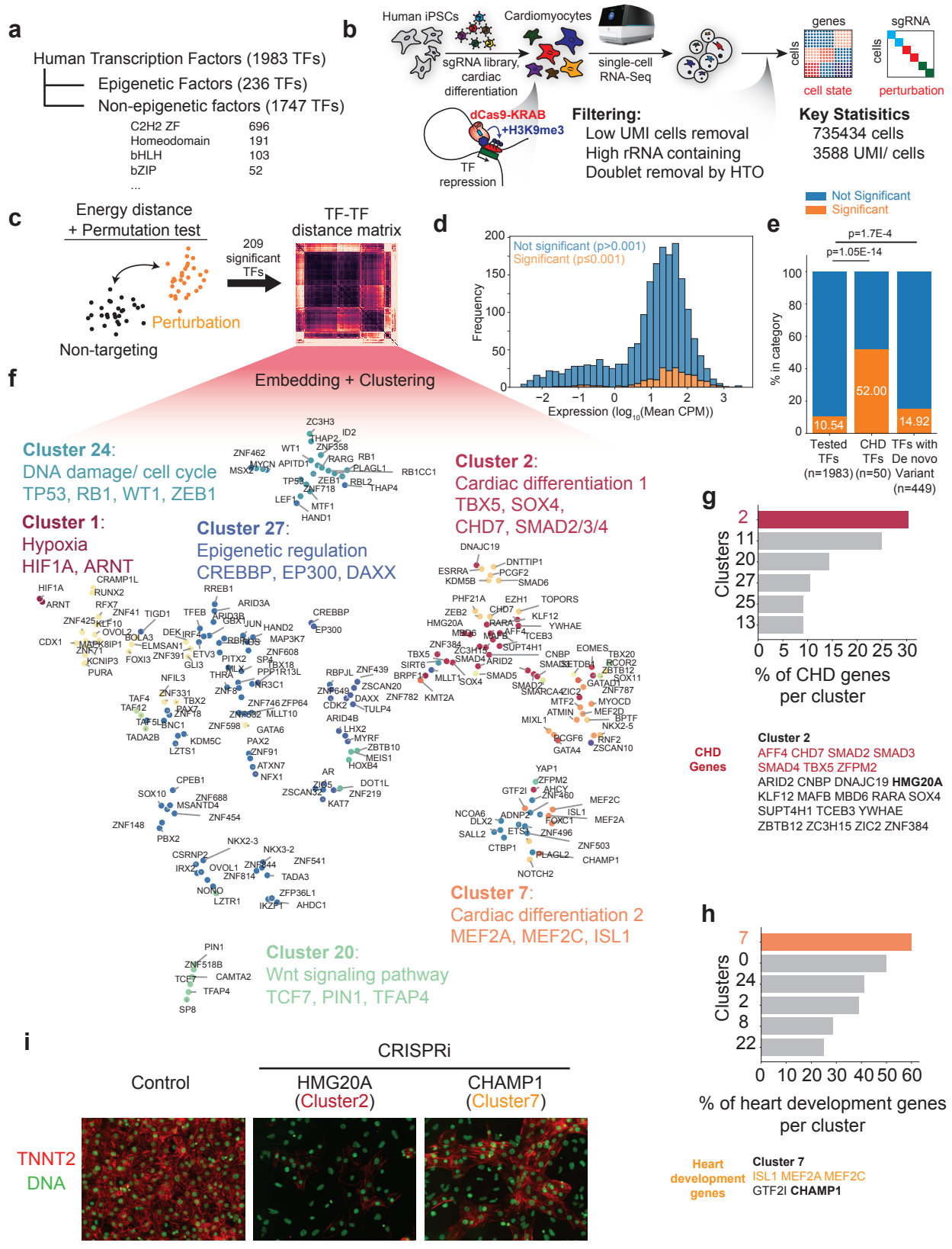

a
Human Transcription Factors (1983 TFs)
Epigenetic Factors (236 TFs)
Non-epigenetic factors (1747 TFs)
C2H2 ZF 696
Homeodomain 191
bHLH 103
bZIP 52
...
b
Human iPSCs
Cardiomyocytes
sgRNA library, cardiac differentiation
single-cell RNA-Seq
genes
cells
sgRNA
cell state
perturbation
dCas9-KRAB
+H3K9me3
TF repression
Filtering:
Low UMI cells removal
High rRNA containing
Doublet removal by HTO
Key Statisitics
735434 cells
3588 UMI/ cells
c
Energy distance + Permutation test
Perturbation
Non-targeting
209 significant TFs
TF-TF distance matrix
d
Not significant (p>0.001)
Significant (p≤0.001)
Frequency
Expression (log10(Mean CPM))
e
Not Significant
Significant
p=1.7E-4
p=1.05E-14
% in category
10.54
52.00
14.92
Tested TFs (n=1983)
CHD TFs (n=50)
TFs with De novo Variant (n=449)
f
Embedding + Clustering
Cluster 24:
DNA damage/ cell cycle
TP53, RB1, WT1, ZEB1
Cluster 1:
Hypoxia
HIF1A, ARNT
Cluster 27:
Epigenetic regulation
CREBBP, EP300, DAXX
Cluster 2:
Cardiac differentiation 1
TBX5, SOX4,
CHD7, SMAD2/3/4
Cluster 7:
Cardiac differentiation 2
MEF2A, MEF2C, ISL1
Cluster 20:
Wnt signaling pathway
TCF7, PIN1, TFAP4
g
Clusters
% of CHD genes per cluster
CHD Genes
Cluster 2
AFF4 CHD7 SMAD2 SMAD3
SMAD4 TBX5 ZFPM2
ARID2 CNBP DNAJC19 HMG20A
KLF12 MAFB MBD6 RARA SOX4
SUPT4H1 TCEB3 YWHAE
ZBTB12 ZC3H15 ZIC2 ZNF384
h
Clusters
% of heart development genes per cluster
Heart development genes
Cluster 7
ISL1 MEF2A MEF2C
GTF2I CHAMP1
i
CRISPRi
Control
HMG20A (Cluster2)
CHAMP1 (Cluster7)
TNNT2
DNA

Figure 2

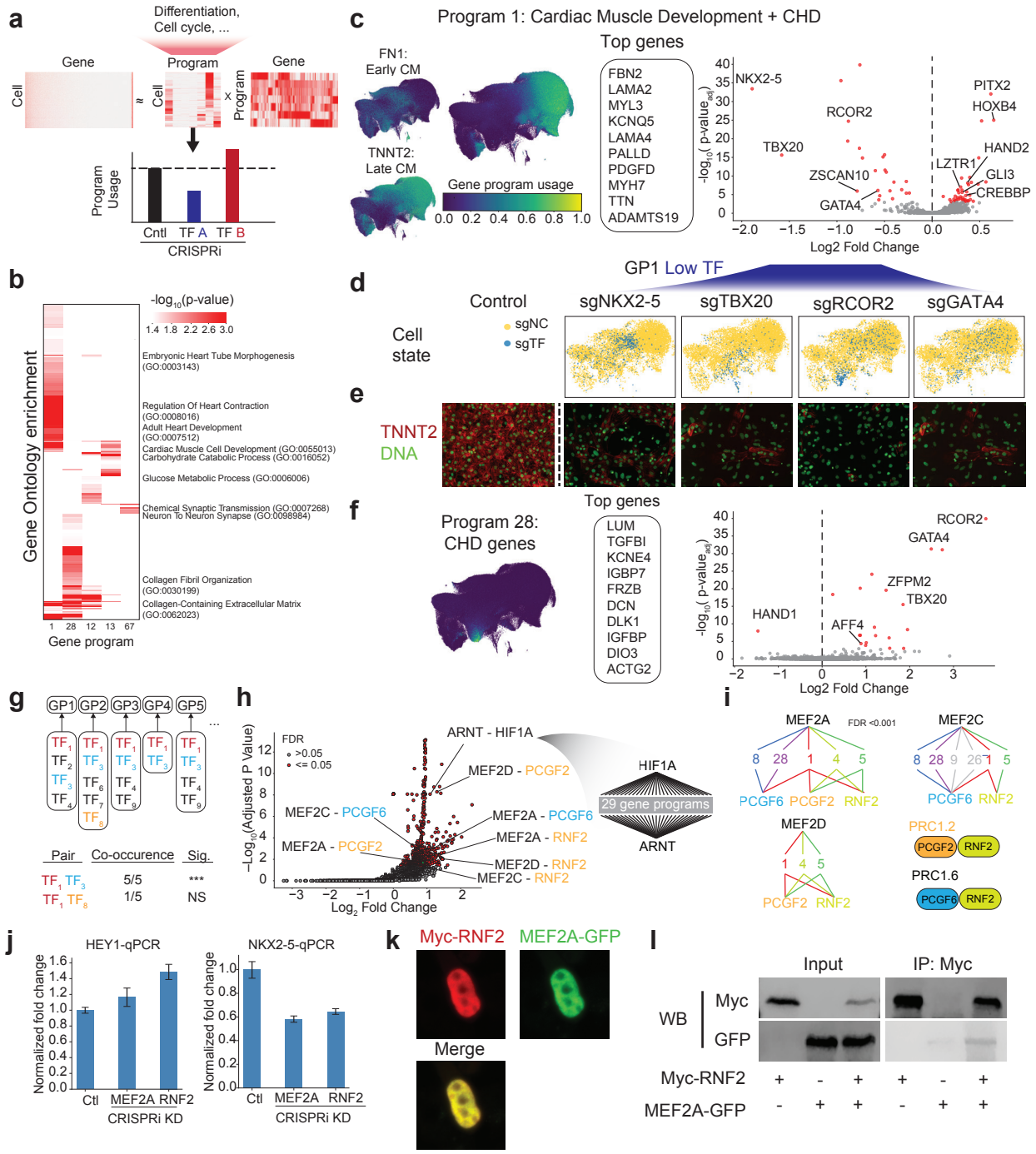

a
Differentiation, Cell cycle, ...
Gene
Program
Gene
Cell
Cell
Program
Program Usage
Cntl
TF A
TF B
CRISPRi
b
-log10(p-value)
1.4 1.8 2.2 2.6 3.0
Gene Ontology enrichment
Embryonic Heart Tube Morphogenesis (GO:0003143)
Regulation Of Heart Contraction (GO:0008016)
Adult Heart Development (GO:0007512)
Cardiac Muscle Cell Development (GO:0055013)
Carbohydrate Catabolic Process (GO:0016052)
Glucose Metabolic Process (GO:0006006)
Chemical Synaptic Transmission (GO:0007268)
Neuron To Neuron Synapse (GO:0098984)
Collagen Fibril Organization (GO:0030199)
Collagen-Containing Extracellular Matrix (GO:0062023)
1 28 12 13 67
Gene program
c
Program 1: Cardiac Muscle Development + CHD
FN1: Early CM
TNNT2: Late CM
Gene program usage
0.0 0.2 0.4 0.6 0.8 1.0
Top genes
FBN2
LAMA2
MYL3
KCNQ5
LAMA4
PALLD
PDGFD
MYH7
TTN
ADAMTS19
NKX2-5
RCOR2
TBX20
ZSCAN10
GATA4
PITX2
HOXB4
HAND2
LZTR1
GLI3
CREBBP
-log10(p-value)
Log2 Fold Change
d
GP1 Low TF
Control
sgNKX2-5
sgTBX20
sgRCOR2
sgGATA4
Cell state
sgNC
sgTF
e
TNNT2
DNA
f
Program 28: CHD genes
Top genes
LUM
TGFBI
KCNE4
IGBP7
FRZB
DCN
DLK1
IGFBP
DIO3
ACTG2
RCOR2
GATA4
ZFPM2
TBX20
HAND1
AFF4
-log10(p-value)
Log2 Fold Change
g
GP1 GP2 GP3 GP4 GP5
Pair
Co-occurence
Sig.
TF1 TF3
5/5
***
TF1 TF8
1/5
NS
h
FDR
>0.05
<= 0.05
ARNT - HIF1A
MEF2D - PCGF2
MEF2C - PCGF6
MEF2A - PCGF6
MEF2A - PCGF2
MEF2A - RNF2
MEF2D - RNF2
MEF2C - RNF2
−Log10(Adjusted P Value)
Log2 Fold Change
HIF1A
29 gene programs
ARNT
i
MEF2A
FDR <0.001
MEF2C
PCGF6 PCGF2 RNF2
PCGF6 RNF2
MEF2D
PCGF2 RNF2
PRC1.2
PCGF2 RNF2
PRC1.6
PCGF6 RNF2
j
HEY1-qPCR
NKX2-5-qPCR
Normalized fold change
Ctl
MEF2A
RNF2
CRISPRi KD
k
Myc-RNF2
MEF2A-GFP
Merge
l
Input
IP: Myc
WB
Myc
GFP
Myc-RNF2
MEF2A-GFP

Figure 3

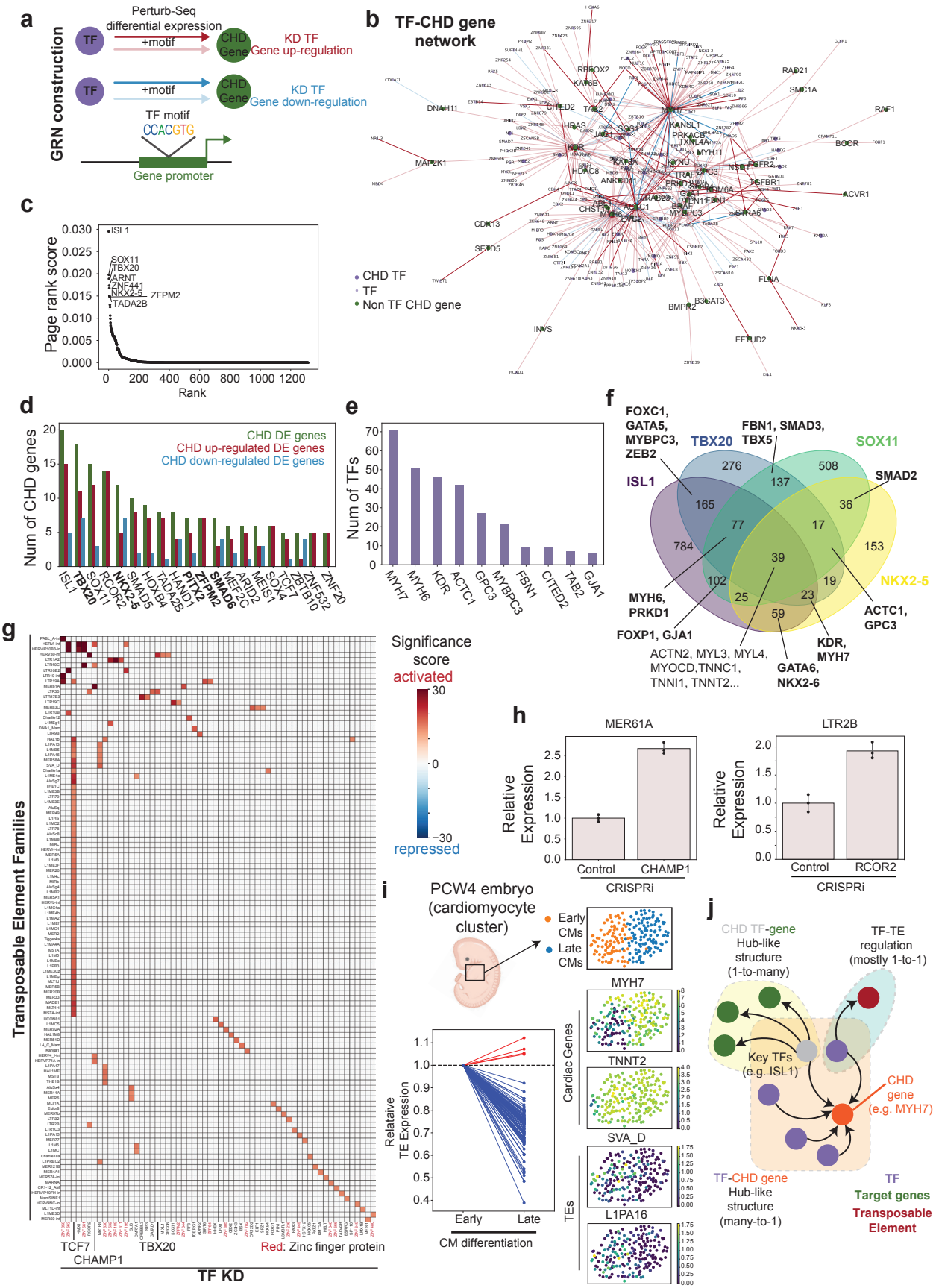

a
GRN construction
Perturb-Seq differential expression
TF
+motif
CHD Gene
KD TF Gene up-regulation
KD TF Gene down-regulation
TF motif
CCACGTG
Gene promoter
b
TF-CHD gene network
CHD TF
TF
Non TF CHD gene
c
Page rank score
Rank
ISL1
SOX11
TBX20
ARNT
ZNF441
NKX2-5
ZFPM2
TADA2B
d
Num of CHD genes
CHD DE genes
CHD up-regulated DE genes
CHD down-regulated DE genes
ISL1
TBX20
SOX11
RCOR2
SMAD5
NKX2-5
HOXB4
TADA2B
HAND1
ZFPM2
PITX2
SMAD6
ARID2
MEF2C
MEIS1
SOX7
ZBTB10
ZNF532
ZNF20
e
Num of TFs
MYH7
MYH6
KDR
ACTC1
GPC3
MYBPC3
FBN1
CITED2
TAB2
GJA1
f
FOXC1, GATA5, MYBPC3, ZEB2
TBX20
FBN1, SMAD3, TBX5
SOX11
ISL1
SMAD2
NKX2-5
MYH6, PRKD1
FOXP1, GJA1
ACTN2, MYL3, MYL4, MYOCD,TNNC1, TNNI1, TNNT2...
GATA6, NKX2-6
KDR, MYH7
ACTC1, GPC3
g
Transposable Element Families
Significance score
activated
repressed
TCF7
CHAMP1
TBX20
Red: Zinc finger protein
TF KD
h
MER61A
LTR2B
Relative Expression
Control
CHAMP1
RCOR2
CRISPRi
i
PCW4 embryo (cardiomyocyte cluster)
Early CMs
Late CMs
MYH7
TNNT2
SVA_D
L1PA16
Cardiac Genes
TEs
Relataive TE Expression
Early
Late
CM differentiation
j
CHD TF-gene Hub-like structure (1-to-many)
TF-TE regulation (mostly 1-to-1)
Key TFs (e.g. ISL1)
CHD gene (e.g. MYH7)
TF-CHD gene Hub-like structure (many-to-1)
TF
Target genes
Transposable Element

Figure 4

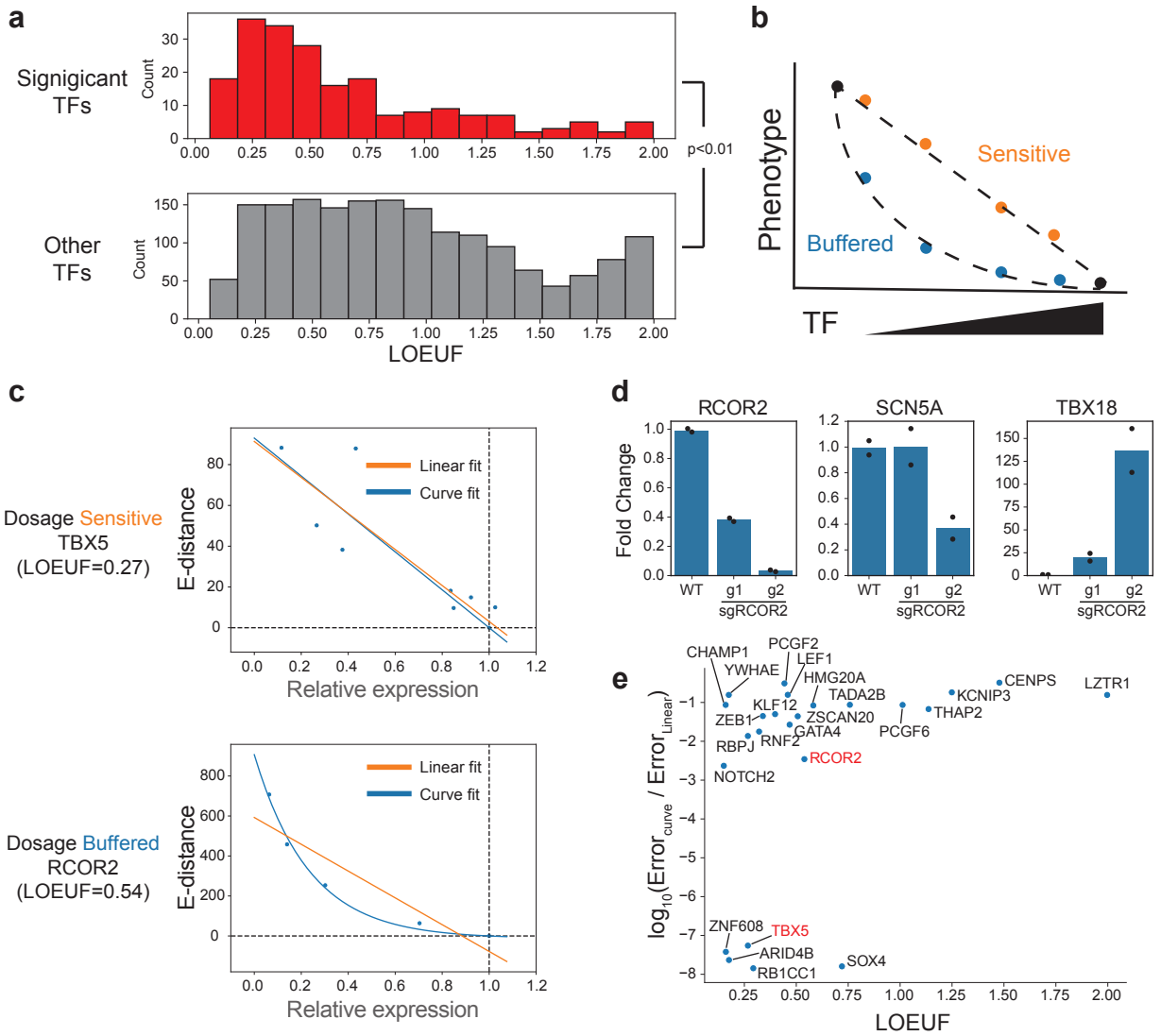

a
Signigicant TFs
Other TFs
Count
LOEUF
p<0.01
b
Phenotype
Sensitive
Buffered
TF
c
Dosage Sensitive
TBX5
(LOEUF=0.27)
Dosage Buffered
RCOR2
(LOEUF=0.54)
E-distance
Relative expression
Linear fit
Curve fit
d
RCOR2
SCN5A
TBX18
Fold Change
WT
g1
g2
sgRCOR2
e
CHAMP1
PCGF2
LEF1
YWHAE
HMG20A
TADA2B
CENPS
LZTR1
KCNIP3
KLF12
ZEB1
ZSCAN20
THAP2
GATA4
PCGF6
RBPJ
RNF2
RCOR2
NOTCH2
ZNF608
TBX5
ARID4B
SOX4
RB1CC1
log10(Error_curve / Error_Linear)
LOEUF

Figure 5

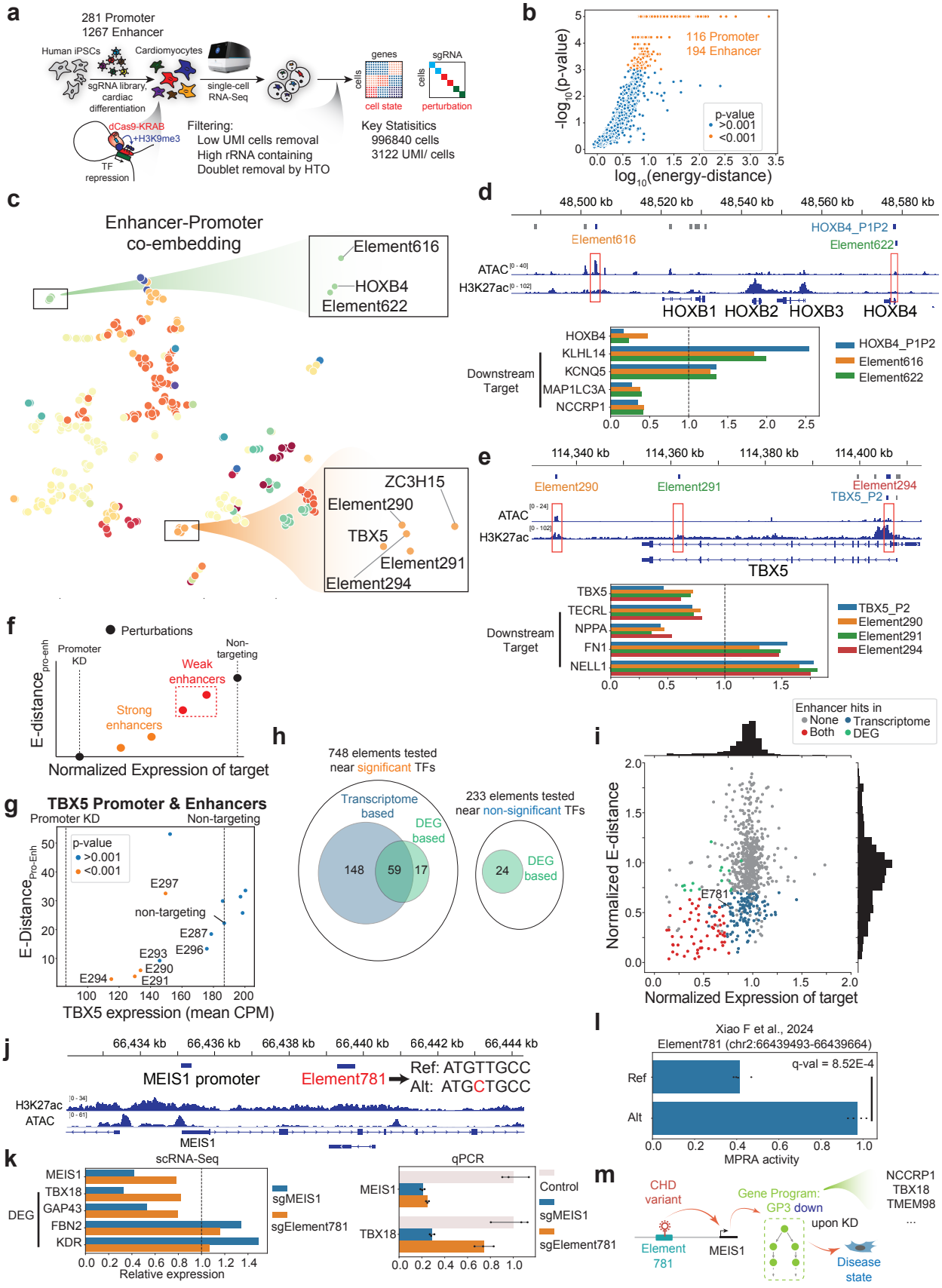
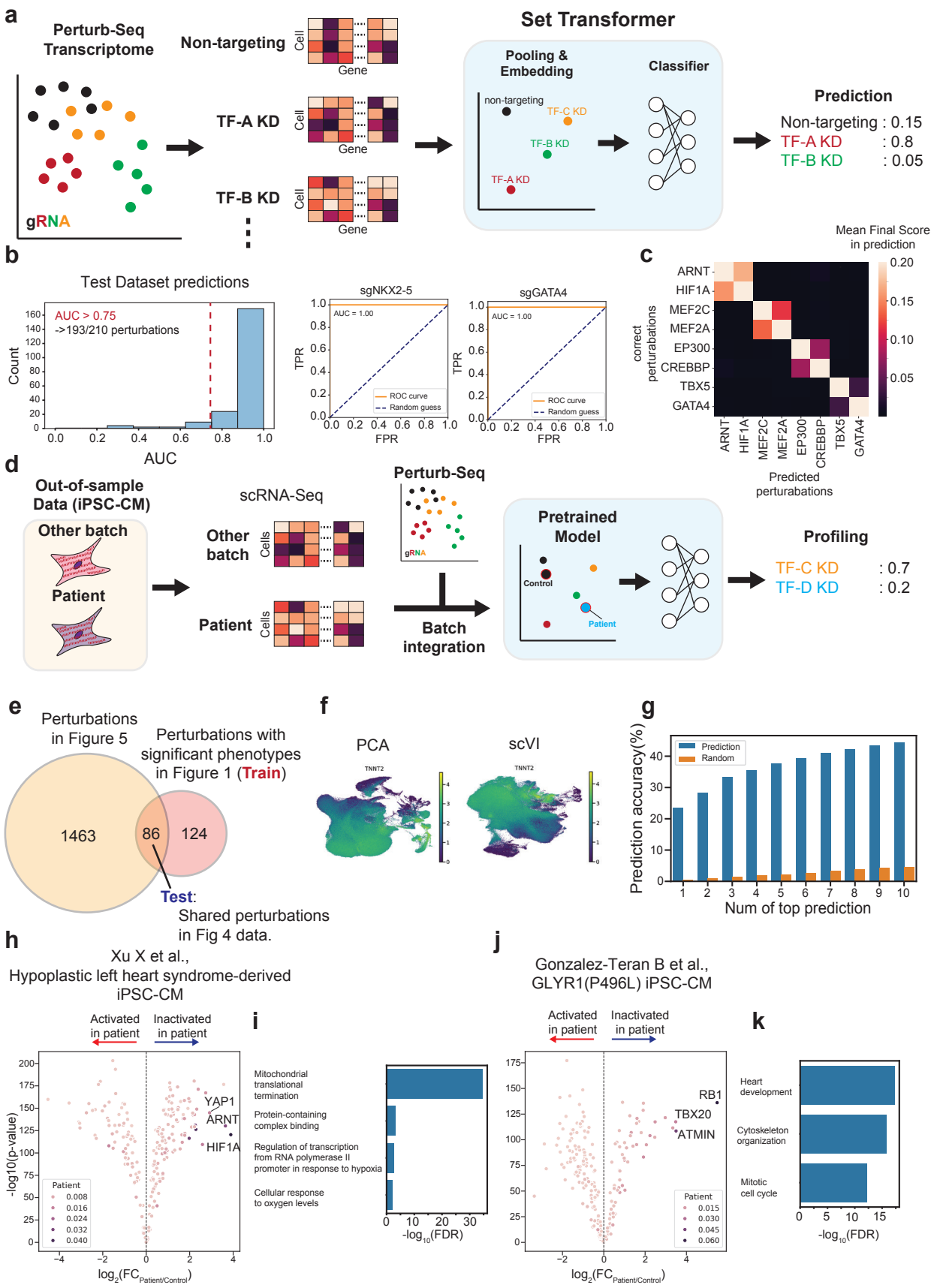

a
281 Promoter
1267 Enhancer
Human iPSCs
Cardiomyocytes
sgRNA library, cardiac differentiation
single-cell RNA-Seq
genes
cells
cell state
sgRNA
perturbation
dCas9-KRAB
+H3K9me3
TF repression
Filtering:
Low UMI cells removal
High rRNA containing
Doublet removal by HTO
Key Statisitics
996840 cells
3122 UMI/ cells
b
116 Promoter
194 Enhancer
p-value
>0.001
<0.001
-log10(p-value)
log10(energy-distance)
c
Enhancer-Promoter co-embedding
Element616
HOXB4
Element622
ZC3H15
Element290
TBX5
Element291
Element294
d
48,500 kb
48,520 kb
48,540 kb
48,560 kb
48,580 kb
HOXB4_P1P2
Element616
Element622
ATAC
H3K27ac
HOXB1 HOXB2 HOXB3 HOXB4
Downstream Target
HOXB4
KLHL14
KCNQ5
MAP1LC3A
NCCRP1
HOXB4_P1P2
Element616
Element622
e
114,340 kb
114,360 kb
114,380 kb
114,400 kb
Element290
Element291
Element294
TBX5_P2
ATAC
H3K27ac
TBX5
Downstream Target
TBX5
TECRL
NPPA
FN1
NELL1
TBX5_P2
Element290
Element291
Element294
f
Perturbations
Promoter KD
Non-targeting
Weak enhancers
Strong enhancers
E-distance pro-enh
Normalized Expression of target
g
TBX5 Promoter & Enhancers
Promoter KD
Non-targeting
p-value
>0.001
<0.001
E297
non-targeting
E287
E296
E293
E290
E291
E294
E-Distance Pro-Enh
TBX5 expression (mean CPM)
h
748 elements tested near significant TFs
Transcriptome based
DEG based
148
59
17
233 elements tested near non-significant TFs
24
DEG based
i
Enhancer hits in
None
Transcriptome
Both
DEG
E781
Normalized E-distance
Normalized Expression of target
j
66,434 kb
66,436 kb
66,438 kb
66,440 kb
66,442 kb
66,444 kb
MEIS1 promoter
Element781
Ref: ATGTTGCC
Alt: ATGCTGCC
H3K27ac
ATAC
MEIS1
k
scRNA-Seq
DEG
MEIS1
TBX18
GAP43
FBN2
KDR
sgMEIS1
sgElement781
Relative expression
qPCR
MEIS1
TBX18
Control
sgMEIS1
sgElement781
l
Xiao F et al., 2024
Element781 (chr2:66439493-66439664)
q-val = 8.52E-4
Ref
Alt
MPRA activity
m
CHD variant
Element 781
MEIS1
Gene Program: GP3 down
upon KD
Disease state
NCCRP1
TBX18
TMEM98
...

Figure 6


a
Perturb-Seq Transcriptome
gRNA
Non-targeting
TF-A KD
TF-B KD
Cell
Gene
Set Transformer
Pooling & Embedding
Classifier
non-targeting
TF-C KD
TF-B KD
TF-A KD
Prediction
Non-targeting : 0.15
TF-A KD : 0.8
TF-B KD : 0.05
b
Test Dataset predictions
AUC > 0.75
->193/210 perturbations
Count
AUC
sgNKX2-5
sgGATA4
AUC = 1.00
TPR
FPR
ROC curve
Random guess
c
Mean Final Score in prediction
correct perturbations
Predicted perturbations
ARNT
HIF1A
MEF2C
MEF2A
EP300
CREBBP
TBX5
GATA4
d
Out-of-sample Data (iPSC-CM)
Other batch
Patient
scRNA-Seq
Other batch
Patient
Cells
Perturb-Seq
gRNA
Batch integration
Pretrained Model
Control
Patient
Profiling
TF-C KD : 0.7
TF-D KD : 0.2
e
Perturbations in Figure 5
Perturbations with significant phenotypes in Figure 1 (Train)
1463
86
124
Test:
Shared perturbations in Fig 4 data.
f
PCA
scVI
TNNT2
g
Prediction
Random
Prediction accuracy(%)
Num of top prediction
h
Xu X et al.,
Hypoplastic left heart syndrome-derived iPSC-CM
Activated in patient
Inactivated in patient
-log10(p-value)
YAP1
ARNT
HIF1A
Patient
0.008
0.016
0.024
0.032
0.040
log2(FC Patient/Control)
i
Mitochondrial translational termination
Protein-containing complex binding
Regulation of transcription from RNA polymerase II promoter in response to hypoxia
Cellular response to oxygen levels
-log10(FDR)
j
Gonzalez-Teran B et al.,
GLYR1(P496L) iPSC-CM
Activated in patient
Inactivated in patient
RB1
TBX20
ATMIN
Patient
0.015
0.030
0.045
0.060
log2(FC Patient/Control)
k
Heart development
Cytoskeleton organization
Mitotic cell cycle
-log10(FDR)