An integrated cell barcoding and computational analysis pipeline for scalable analysis of differentiation at single-cell resolution

Sophie Shen[1], Tessa Werner[1], Yuliangzi Sun[1], Woo Jun Shim[1], Samuel Lukowski[1,2], Stacey Andersen[1], Han Sheng Chiu[1], Di Xia[1,3], Xiaoli Chen[1], Duy Pham[1], Zezhuo Su[4,5], Joseph E. Powell[6,7], Patrick P. L. Tam[8,9], Mikael Bodén[10], Joshua W. K. Ho[4,5], Quan Nguyen[1], Nathan J. Palpant[1,*]

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

[2] Boehringer Ingelheim RCV, GmbH & Co KG., Cancer Immunology & Immune Modulation, Vienna, Austria

[3] Genome Innovation Hub, The University of Queensland, Brisbane, QLD, Australia

[4] School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

[5] Laboratory of Data Discovery for Health Limited ($D^2$4H), Hong Kong Science Park, Hong Kong SAR, China

[6] Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia

[7] University of New South Wales, Cellular Genomics Futures Institute, Sydney, NSW, Australia

[8] Embryology Research Unit, Children's Medical Research Institute, University of Sydney, Westmead, NSW, Australia

[9] Faculty of Medicine and Health, School of Medical Sciences, University of Sydney, Camperdown, NSW, Australia

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

[*] Corresponding Author

SUMMARY

This study develops a versatile cell multiplexing and data analysis platform to accelerate knowledge gain into mechanisms of cell differentiation. We engineer a cell barcoding system in human cells enabling multiplexed single-cell RNA sequencing for high throughput perturbation of customisable and versatile experimental conditions. This is coupled with a new computational analysis pipeline that overcomes the limitations of existing algorithms by using an unsupervised, genome-wide, orthogonal biological reference point to reveal the cell diversity and regulatory networks of any input scRNA-seq data set. We implement this pipeline by engineering transcribed barcodes into induced pluripotent stem cells and multiplex 62 independent experimental conditions comprising eight differentiation time points and nine developmental signalling perturbations with duplicates. We identify and deconstruct the temporal, signalling, and gene regulatory basis of multilineage iPSC differentiation into ectoderm, mesoderm, and endoderm cell types. This study provides a cellular and computational pipeline to study cell differentiation applicable to diverse fields of developmental biology, drug discovery, and disease modelling.

KEY WORDS

Stem cell, differentiation, scRNA-seq, multiplexing, barcoding, iPSC, epigenetic, regulation, cell identity, pipeline

Single-cell omics technologies have revolutionised capabilities to study dynamic and heterogeneous cell states involved in all biological processes and systems. These technologies are ideally suited to study population-scale transcriptional control of cell subtypes and their differentiation processes, enabling the capture of hundreds of thousands to millions of cells to study the cellular basis of development and disease[1-5].

Human pluripotent stem cell (hPSC) models are ideally suited to deconstruct early human developmental processes in a scalable and cell type-specific manner. Significant investment and sector growth in stem cell technologies highlight their importance in research, industry drug discovery pipelines, disease modelling, and regenerative medicine[6]. However, atlases of *in vitro* differentiation are uncommon and mechanisms governing cell differentiation and fate specification across cell lineages remain poorly understood[7-9] creating significant limitations in guiding hPSCs into cell types modelling adult tissues[10-12]. Scalable data generation to perturb and map sequential stages of multi-lineage cell differentiation decisions from pluripotency are critical to understanding and optimising methods for deriving cell types of the human body that closely model *in vivo* developmental and physiological niches.

While single cell technologies have delivered a new scale of knowledge gain into mechanisms of cell biology, current pipelines are limited by the following two major issues. First, sample multiplexing approaches rely either on intrinsic genetic differences between samples[13, 14], or use of post-experimental sample labelling[15-17] that limit versatility in experimental design and often require stressful sample handling to barcode cells. This study provides a novel cell barcoding method enabling integration of transcribed barcodes into the AAVS1 safe harbour locus of the human genome. The scale of barcode combinations coupled with the benefit of endogenously expressed cell barcoding enables scalable data generation in any human cell type including applications in stem cell biology for complex experimental designs in protocol development, analysis of cell-cell relationships in organoid models, and optimising therapeutics for organ regeneration. Using barcoded hPSCs, we perform a highly multiplexed perturbation study to characterise the diversification of cell types during multi-lineage differentiation from pluripotency *in vitro*.

Second, despite the ease of scaling up data generation, current single cell computational pipelines are designed using algorithmic methods with arbitrary threshold and parameter selection, rely heavily on prior knowledge and existing reference databases to delineate and annotate cell types, and routinely involve data subtraction by differential expression[6, 18]. These computational methods are fundamentally limited by the absence of a biological basis for defining distinct cell states in heterogeneous scRNA-seq data. We, therefore, implement a novel computational pipeline that uses an independent, unsupervised, genome-wide, quantitative reference point to efficiently identify biological diversity and underlying gene regulatory mechanisms of any orthologous single cell RNA-seq data set. The approach enables large-scale data interpretation through the identification of the most well-defined cell states. Collectively, this study provides a novel wet and dry lab pipeline for scalable knowledge gain into mechanisms of hPSC differentiation with demonstrated utility in revealing the temporal and signalling basis of differentiation from pluripotency (Fig. 1). A web accessible dashboard is available for user access which provides details on cell barcoding technology, data set availability, a navigable data analysis app, and download details for all computational methods (http://cellfateexplorer.d24h.hk/).

RESULTS

Derivation of novel barcoded iPSCs for multiplexed single-cell RNA-seq

Cell barcoding strategies to enable sample multiplexing reduce data generation costs and labour in scRNA-seq experiments. To enhance sample multiplexing capabilities in hPSCs, we designed a scalable sample barcoding method that enables diverse experimental design options to analyse isogenic human cell lines without additional cell handling. 10,000 15 base pair barcodes were randomly generated using an even 25% probability for the presence of each A, C, T and G nucleotides. Barcodes starting or ending with a stop codon or containing repetitive runs of 4 or more nucleotides were excluded and we required a minimum Hamming distance of 5 nucleotides between any pair of barcodes (Supplementary Table 1). We cloned barcodes to be transcribed downstream of the ubiquitously expressed chicken β-actin promoter (CAG) at the 3′ terminus of GFP in the AAVS1-2a-Puro gene targeting cassette[19, 20]. This enables gene targeting to the human AAVS1 safe harbour locus with selection of correctly targeted cells by puromycin and purity evaluated by fluorescence-activated cell sorting (FACS) for GFP (Supplementary Fig. 1a-b and Supplementary Fig. 2a). The design enables transcribed barcodes to be captured by any single-cell RNA-seq protocol and amplified in a separate sequencing library to allow each read from pooled experimental samples to be assigned back to its sample of origin without the need for exogenous labelling before sample pooling. With thousands of potential simultaneous barcode combinations, multiplexing is only limited by the scale of cell capture technologies. We engineered 18 isogenic barcoded WTC-11 hiPSC lines by this design, each of which underwent vigorous quality testing for karyotype, pluripotency, and purity (Supplementary Fig. 1b-d, Supplementary Fig. S2).

We performed a pilot scRNA-seq experiment pooling all 18 barcoding lines as undifferentiated iPSCs, labelling four lines (BC08, BC09, BC14, BC15) with exogenously applied cell hashing antibodies[16] to investigate the sensitivity and accuracy of identifying barcodes in post-processing computational protocols. We captured 4,829 high-quality cells and found that 1.3% of cells had no reads mapped to any genomic sample barcode, 42.0% mapped to one, and 56.7% to multiple barcodes. By comparison, 22% of cells barcoded using hashing antibodies lacked an assigned antibody barcode, demonstrating that our method has a lower false negative rate, but a lower singlet rate compared to the 71.9% of hashed cells. Dimensionality reduction analysis shows that transcriptomes are sufficiently homogenous to randomly distribute across UMAP space (Supplementary Fig. 1e)**,** and cells from each line reveal comparable reads per cell and expression of pluripotency markers *SOX2* and *OCT4* (Supplementary Fig. 1f,g), further confirming their homogeneity and suitability as isogenic barcoded cell lines. Notably, these quality control metrics do not differ in the four cell lines that underwent additional processing for cell hashing, and we observed 83.2% agreement between the genomic and cell hashing sample assignment, supporting the accuracy of our sample barcoding (Supplementary Fig. 1h). These data demonstrate the quality of barcoded iPS cell lines as a reliable method of multiplexing scRNA-seq experiments enabling a complex array of experimental designs.

Generation of a multiplexed atlas of iPSC differentiation *in vitro*

We set out to capture atlas-scale single-cell data revealing the temporal diversification of endoderm, mesoderm, and ectoderm cell lineages from pluripotency to link relationships between cell diversity and the signalling, molecular, and temporal determinants of differentiation for use in protocol development and analysis of genetic mechanisms of cell

differentiation. We used cell barcoding to multiplex scRNA-seq analysis of multi-lineage differentiation *in vitro* from hPSCs using a monolayer-based Wnt-induced differentiation protocol[21, 22] (Fig. 2a). First, we captured cells every 24 hours from day 2 to day 9 to generate a high-resolution time course of cell state transitions during differentiation (Fig. 2b). Second, we captured differentiating cells between gastrulation (day 3) and progenitor (day 5) differentiation stages, perturbed by small molecule inhibitors and activators of signalling pathways known to be essential regulators of development and differentiation: WNT[23], BMP[24], and VEGF[25] (Fig. 2c-e and Supplementary Table 1). This allowed us to enhance the lineage diversification of cell types captured at the progenitor and defined cell stages (days 5 and 9). Cells were captured before (day 2) and after signalling perturbations (day 5) and at a committed differentiated state (day 9) in biological duplicates (Fig. 2c,d and Supplementary Table 1). In total, this study multiplexed 62 independent barcoded cell samples into four scRNA-seq libraries. Standard quality control, pre-processing, and characterisation of individual datasets confirmed expression of marker genes indicating successful capture of mesoderm, endoderm, and ectoderm cell lineage differentiation (Supplementary Fig. 2,3). After data integration[26], marker gene expression shows cell types including early mesendoderm (*GSC*, *MIXL1*), cardiac mesoderm (*HAND1*), cardiomyocytes (*MYH6*), limb mesoderm (*PRRX1*), axial or notochord mesoderm (*T, NOG*), definitive endoderm (*FOXA2*), primitive gut (*AFP, TTR*), neural progenitor (*TTYH1*), and endothelium (*CDH5*) (Fig. 3a,b). Together, this dataset represents a high-resolution time-course and cell diversification atlas *in vitro* to facilitate discovering mechanisms underlying temporal and signalling control of cell heterogeneity during differentiation from pluripotency.

Using TRIAGE as an independent, unsupervised biological reference point to identify cell diversity

Sample barcoding and advances in single cell capture make data generation increasingly simple and cost-effective. However, downstream data analyses remain challenging because current methods primarily rely on mathematically defined dimensionality reduction and clustering approaches to partition cells based on user-defined thresholds rendering outputs that are subjective (Fig. 3c). We, therefore, set out to address this limitation by building a data analysis pipeline that anchors complex genomic data to a biological reference point. Detailed development and parameter testing of these methods are provided in our companion manuscript (Sun *et al.,* 2022).

In brief, we previously developed a method called *TRIAGE* (Transcriptional Regulatory Inference analysis from Gene Expression) that uses broad H3K27me3 domains from diverse cell types to identify cell type-specific regulatory genes[27]. The method results in an unsupervised, genome-wide, quantitative value applied to all protein-coding genes which we call a repressive tendency score (RTS) that quantifies the cell type-specific regulatory potential of any gene and offers a powerful approach to interpreting any complex genomic data type by reference to an independent biological reference point. We build on this observation by implementing *TRIAGE* for cell clustering using a topographical map-like approach where RTS-prioritised genes are used to identify cell populations in a low-dimensional space. Specifically, we used the rank order of RTS genes as a quantitative reference point to evaluate cell populations in the data set where every cell is anchored by the highest RTS-ranked gene expressed in the cell. This provides the computational basis for a weighted density function to identify cell populations in a 2-dimensional UMAP space. Cells falling within these "peak" regions are interpreted as defined cell types whereas non-peak cells are defined by lower RTS genes (less cell-type specificity) and therefore represent transitional or non-specific artifactual cells (Fig. 3d). We apply *TRIAGE*

clustering to our single cell differentiation data atlas to identify 48 cell peaks, each characterised by a *TRIAGE* anchor gene representing the RTS priority gene defining the greatest proportion of cells (Fig. 3e,f and Supplementary Table 2).

We show in our sister paper that the cell trimming resulting from *TRIAGE* clustering allows for an overall reduction in cell type-to-cell type similarity, compared to standard clustering methods that utilise all cells in the dataset (Sun *et al.,* 2022). Likewise, we observe that *TRIAGE* clustering on this novel dataset identifies more transcriptionally distinct cell populations among the 48 identified cell types when compared to equivalent whole-dataset clustering using Seurat (Fig 3g). To evaluate the performance of *TRIAGE* clustering independent of cell assignments represented in a 2D-UMAP space, we used VIA, which combines lazy-teleporting random walks with Monte-Carlo Markov Chain simulations in a graph directed analysis to evaluate relationships between cell states[28]. Compared to analysis using all cells in the data, restricting data input to only *TRIAGE* peaks results in the better peak-to-node assignment and greater correlation between pseudotime and actual time (Pearson correlation for all cells = -0.04; for cells in *TRIAGE* peaks = 0.51) (Supplementary Fig. 5). These data demonstrate that *TRIAGE* clustering is an improved approach for biologically determining cellular diversity in single cell data.

Annotating *in vitro* differentiated cells by reference to *in vivo* cell types

We next sought to broadly annotate cell peaks. Unfortunately, most annotation databases are developed using transcriptional characteristics of adult cell types and lack utility for evaluating stem cell differentiation which is enriched in progenitor, transitory, or cell types not well matched to any *in vivo* cell type. We took two approaches to evaluate *in vitro* derived cell peaks. We first used gene set scoring to analyse cells based on positive gene sets defining seven major cell type domains modelled from 921 RNA-seq datasets[29]. We found that among the high-scoring cells that scored 0 or greater, 5 general cell type domains are represented in our data: Mesoderm, Surface ectoderm, Neural crest, Germ, and Embryonic domains (Fig 3h and Supplementary Fig. 6a). We also used label transfer analysis to score *in vitro* derived cells based on their transcriptional similarity to single-cell atlases of mouse organogenesis[1] (Fig. 3i and Supplementary Fig. 6b) and human gastrulation[4] (Fig. 3j and Supplementary Fig. 6c). By setting a threshold to define cells scoring 0.5 above for transcriptional similarity, we find that using a combination of reference datasets provides a preliminary indication of the *in vitro*-derived cell type diversity. Though the general label transfer patterns are supported by marker gene expression, the results are highly dependent on accurate annotations in the reference data, and thus lack the capacity for comprehensive cell annotation.

Using TRIAGE to distinguish gene programs underpinning cell subtypes

To address the limitations of annotating *in vitro* hPSC-derived cell types, we evaluated each cell peak's gene expression profiles to dissect cell-type specific gene programs. While methods such as differential expression have historically been used to compare such gene programs, this approach is unstable and returns different gene sets depending on the comparison being made[27]. We, therefore, developed a novel approach called *TRIAGE-ParseR* that draws on the conceptual basis of *TRIAGE* to use shared H3K27me3 deposition patterns across diverse cell types to identify co-regulated, cell identity-defining biological processes[27]. Utilising consortium-level measurements of H3K27me3 across 111 cell types in the NIH Roadmap Epigenomics dataset[30], we extracted hidden orthogonal H3K27me3 deposition patterns using a Gaussian mixture model. These are used to group input gene lists into "co-modulated" gene sets with

shared patterns and functional roles in each cell type (Fig. 4a) (Sun *et al.,* 2022). As input gene lists, we used the top ranked genes for each peak following *TRIAGE*-transformation of the expression data. These genes represent the most cell type-defining regulatory and structural genes for each peak, which *TRIAGE-ParseR* separates into sub-groups, clarifying cell type annotation and enhancing enrichment readout from methods such as Gene Ontology (GO) or STRING network analyses.

Based on this strategy, we implemented a classification strategy using *TRIAGE* to transform input gene expression data from a given peak to prioritise cell type regulatory genes[27] and calculate pairwise Jaccard similarities between the identity-defining genes for each peak. The output dendrogram and heat map define the relationships between heterogeneous cell types in the dataset based on *TRIAGE* regulatory gene enrichment (Fig. 4b). We used *TRIAGE-ParseR* to evaluate the top *TRIAGE* gene lists from every cell type in the *in vitro* data set to characterise the underlying gene programs underpinning cell diversity. The utility of this method is demonstrated by analysing detailed data outputs for four cell peaks (32, 34, 37, 40) annotated as "cardiomyocytes" from the *in vivo* label transfer analysis with reference to mouse organogenesis data[1] (Fig. 4c-e and Fig. 3i). *TRIAGE* clustering identifies these cardiomyocyte peaks by the mesoderm regulatory gene *GATA4*, and all peaks are enriched in XAV treated conditions (Wnt inhibition) emerging during differentiation between days 7-9. To evaluate whether these peaks represent different cardiac lineage cell subtypes, we evaluated *TRIAGE* ranked genes and summary GO enrichment for *TRIAGE-ParseR* output gene modules (Fig. 4c-e and Supplementary Table 3). *TRIAGE* analysis of peak 37 shows enrichment in *TBX2*, a known regulator of non-chamber myocardium including septa, valves, and cardiac outflow tract[31, 32] and *TBX5*, which governs cardiac ventricle morphogenesis and ventricular septum formation[33] (Fig. 4c). *TRIAGE-ParseR* analysis of peak 37 reveals three main gene modules with GO enriched terms for cardiac developmental genes and enrichment for terms related to outflow tract and cardiac septum morphogenesis (Fig. 4d). We conclude that peak 37 is a cardiac cell subtype resembling the ventricular septum. Using the same deductive pathway, we define peaks 40, 32, and 34 to be closely related to a cardiac pacemaker, first heart field, and second heart field lineage cardiomyocytes, respectively (Fig. 4e). Taken together, this gene network analysis strategy provides a complementary unbiased approach to evaluate gene programs underpinning the identity of cell populations to facilitate the classification of cell subtypes in scRNA-seq data.

An integrated atlas for studying multi-lineage differentiation from pluripotency

We provide a detailed analysis of 48 *TRIAGE*-identified cell types in the dataset linking label transfer results, signalling treatment, timepoint, and gene regulatory metadata of cell subtypes derived from a time course of pluripotent stem cell differentiation (Fig. 5). We identify cells representing gastrulation stage cell types at day 2 including epiblast, ectoderm, and mesendoderm. Time course data between days 3-5 reveal the emergence of paraxial, cardiac, and forelimb mesoderm together with definitive endoderm and posterior foregut. A sampling of time course and signalling perturbation data on day 5 reveal the first appearance of endocardial endothelium, cells of the intestinal epithelium and anterior foregut, as well as notochord and neural tube. As cells progress between days 6 and 9 in the time course sampling data, we identify the emergence of first and second heart field cardiac lineage cells, pacemaker cells, and cells of gut lineages, including pancreas and liver. On day 9, integrated analysis of time course and signalling data show differentiation of additional definitive cell types including ventricular cardiomyocytes, intestinal epithelium, epithelial basal cells, tracheal/bronchial cell types, and neural plate derivatives. Collectively, this study provides wet and dry lab

methods to accelerate discovery into mechanisms of cell differentiation. All data, code, and information for access to resources described in this study are available at our online web dashboard, CellFateExplorer (http://cellfateexplorer.d24h.hk/).

DISCUSSION

This study introduces CRISPR barcoded hPSCs to facilitate efficient sample multiplexing in scRNA-seq experiments. We demonstrate the utility of this method as a powerful approach for accelerating knowledge gain by generating a single-cell transcriptomic atlas chronicling differentiation into cell types from mesoderm, endoderm, and ectoderm lineages. We then introduce a novel, two-part computational approach to cell clustering and classification that harnesses consortium-level epigenetic data to infer well-defined cell subtypes in scRNA-seq data. Together, we present a unique dataset of multi-lineage iPSC differentiation with broad applicability in studying mechanisms regulating cell type specification enabled by multiplexed perturbation platform for data generation and unbiased, biologically grounded data interpretation.

Unlike existing scRNA-seq sample multiplexing strategies[15-17], endogenously engineered cellular barcodes broaden the experimental design capabilities for *in vitro* and *in vivo* studies involving hPSCs. Barcoding isogenic lines via gene targeting to the AAVS1 locus allows for constitutively expressed barcodes in all cell types for direct comparison of perturbations, avoiding variables related to re-programming and genotype differences between hPSCs[34, 35] and removing cell processing stresses to enable higher quality transcriptomic data. As a pre-experimental sample multiplexing method, our barcoding strategy enables thousands of theoretical nucleotide combinations to allow scalable, complex experimental designs for high throughput parameter testing of protocol development, organoid biology, and in vivo cell transplantation studies.

We couple scalable data generation with a novel computational analysis pipeline that leverages the relationship between broad epigenetic repression and cell type-specific regulatory genes to anchor single-cell transcriptomic data by referencing biologically meaningful features to define cell types. By restricting the analysis to biologically grounded cell types, we eliminate noise arising from technical variability in scRNA-seq capture and hPSC culture that complicate data interpretation. We show that these cells facilitate more precise predictions of lineage, and through reference to *in vivo* benchmark datasets, can guide subsequent analysis into cell subtypes in the data.

Future work is needed to address several limitations of this approach. First, *TRIAGE-Clustering* filters the data depending on bandwidth selection, resulting in significant data trimming. This expensive loss of information can reduce statistical power for interpreting biological comparisons between cell types. Furthermore, while our method does not provide a means for solving cells between peaks, the continuous nature of the repressive tendency score across the UMAP space could be adapted to allow inference of transitions between cell states, akin to methods that calculate transition probability functions and energy landscapes over cell states[36, 37].

Cell annotation remains one of the most challenging issues in single cell studies because of the need to relate cell heterogeneity to known biology. While machine learning-based cell type annotation approaches are increasingly popular due to their automatic, unsupervised classification of cell types[38], they rely heavily on existing annotations. We show that use of *TRIAGE-ParseR* provides an unsupervised approach to interpreting gene programs underpinning each cell

type without using subjective cell-cell comparisons such as differential expression. Since existing annotations are often broad and do not inform on cell subtype identity, our method is complementary to automated cell classifiers. It can assist in cases where different clusters are assigned the same annotation.

In the effort to construct a comprehensive atlas of human development, the need for *in vitro* representations of early cell fate specification is critical[39]. To address this, our high-resolution time course from gastrulation-like cells to committed cell states provides unique insights into the transient progenitor cell states involved in this highly dynamic differentiation window. Positive and negative perturbation of developmentally relevant signalling pathways reveals dosage-dependent signalling cues that affect progenitor states during lineage specification. As with other *in vitro* datasets of differentiation[40, 41], our dataset should be benchmarked rigorously against *in vivo* development to assess how well each cell type transcriptomic profile matches between these models. Through understanding the sources of variation between *in vitro* and *in vivo* biology, improvements can be made to existing differentiation protocols to better represent *in vivo* cell differentiation and improve the accuracy of iPSC models for studies in developmental biology, drug discovery, disease modelling, and regenerative medicine.

LEAD CONTACT

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Nathan J. Palpant (n.palpant@uq.edu.au)

MATERIALS AVAILABIILITY

Details for all resources presented in this study are available at our web accessible dashboard (http://cellfateexplorer.d24h.hk/) and will be made available upon request.

DATA AND CODE AVAILABILITY

All three single-cell RNA sequencing raw datasets generated in this study have been deposited at [GEO]. Additionally, all figures and an interactive version of the combined, annotated dataset are available at http://cellfateexplorer.d24h.hk/ , where additional co-modulation analysis outputs, metadata, and gene expression for individual peaks can be queried. Custom scripts for analysis to generate depicted figures are available at http://github.com/palpant-comp/sc_barcoding_scripts. *TRIAGE* clustering and gene co-modulation scripts are available at https://github.com/palpant-comp/TRIAGE-Cluster and https://github.com/palpant-comp/TRIAGE-ParseR.

ONLINE METHODS

hiPSC culture and standard monolayer mesendoderm differentiation. All human pluripotent stem cell studies were carried out in accordance with consent from the University of Queensland's Institutional Human Research Ethics approval (HREC#: 2015001434). Undifferentiated hiPSCs were cultured on Vitronectin XF (STEMCELL Technologies #07180) coated plates in mTeSR1 media (STEMCELL Technologies #05850) with supplement at 37°C with 5% CO_2. On the day prior to differentiation (day -1), cells were dissociated using 0.5% EDTA solution and seeded onto separate coated plates in pluripotency media with ROCK Inhibitor (STEMCELL Technologies, #72308) and cultured overnight. Once forming an ~80% confluent monolayer, differentiation was induced (day 0) by changing the culture media to RPMI (ThermoFisher, #11845119) containing 3µM CHIR99021 (STEMCELL Technologies, #72054), 500mg/mL BSA (Sigma #A9418), and 213mg/mL ascorbic acid (Sigma #A8960). On days 3 and 5, the media was replaced with the same media cocktail excluding the CHIR99021 supplemental cytokine. On day 7 and every subsequent other day, the cultures were fed with RPMI containing 1xB27 (ThermoFisher #17504001) supplement plus insulin.

Barcode cassette design. The barcodes (Supplementary Table 1) were introduced into the cells as a part of a barcode cassette, which also incorporated the reverse complement of a partial Chromium Read2 adaptor sequence (truncated slightly to allow oligo length of <= 60bp). This Read2 sequence is used as a PCR handle to generate barcode-containing amplicons compatible with Chromium scRNA library preparation. Additionally, restriction enzyme recognition sequences were added or regenerated to enable easy transfer of the cassette between different vectors. The exact structure of the cassette is as follows: EcoRV site 3' 3bp – MluI site – partial 10x Read2 adaptor reverse complement – 15bp barcode – MluI complementary sequence. Barcoding cassettes were ordered as complementary single-stranded oligos, which could be annealed and ligated into a digested plasmid backbone.

Vector design. AAVS1-CAG-hrGFP (Addgene #52344) was used as the plasmid backbone. It contains hrGFP under the control of the CAG promoter, and AAVS1 homology arms to allow integration of the linearized plasmid into the genome when paired with the CRISPR system using well-described guide RNAs. The barcode cassette was introduced between the EcoRV and MluI sites of the plasmid between hrGFP and the poly-adenylation site, to enable expression as part of the hrGFP transcriptional unit.

Generation of barcoding plasmids. pAAVS1-CAG-hrGFP (Addgene #52344) was digested by incubation with EcoRV-HF (New England BioLabs; NEB #R3195S) followed by addition of MluI-HF (NEB #R3198S) and further incubation. Successful digestion was confirmed by running a small amount on an agarose gel, and remainder was purified using QIAQuick PCR Purification Kit (QIAGEN #28104). Top and bottom strands of barcode oligos were annealed by mixing 1uL each of 100 µM oligo in 1x T4 DNA ligase buffer in a volume of 10 µL, heating to 94°C for 2 min, then allowing to cool to 25°C at a rate of 1°C/s. Annealed oligos were further diluted 1 in 10 with nuclease-free water. Ligation of barcode cassettes to vector was performed by combining 100ng digested plasmid, 4 µL diluted annealed oligos, and 1 µL T4 DNA ligase (NEB #M0202S; 400U/µL) in 10µL total volume with 1x T4 DNA ligase buffer. Reaction was incubated at 16°C for 16 hours. Separate reactions were performed for each barcode oligo. 3µL of ligation reactions were added to 20 µL Stellar competent cells (Takara #636763) and heat shock transformation performed at 42°C for 1 min in 1.5mL tubes. 350 µL SOC media was added for recovery, with shaking at 37°C for 1 hour. 100 µL was spread onto selective ampicillin-containing agar plates, which were incubated overnight at 37°C. 3 colonies were picked from each plate for screening with colony PCR for expected insert in 10 µL reactions using MangoTaq (Bioline #BIO-21083). Colonies showing successful amplification of insert were grown overnight in 5 mL LB broth containing ampicillin, and plasmid purified using QIAGEN Plasmid Miniprep kit. 100% sequence identity was confirmed across the barcode insert by Sanger sequencing, performed by the Australian Genome Research Facility (AGRF). 50 mL cultures were grown from glycerol stocks of plasmids with confirmed barcode sequence insert, and plasmid was purified using Nucleobond Xtra Midi Kit (Macherey-Nagel #740410.50) to give endotoxin-free, high concentration stock for transfection.

Stable cell generation. For gene editing, cells were grown to about 50-80% confluency, dissociated using 1XTrypLE and 100-200K cells were used for each 10 µl reaction of the Neon Transfection System. The transfection mixture included 0.5 µg of Barcode plasmid DNA, 20 pmol AAVS1-taregting sgRNA (protospacer sequence: atcctgtccctagtggcccc, chemically synthsised by Agilent technology) and 20 pmol spCas9 protein (IDT). After electroporation with 1 pulse of 1300 V for 30 ms, cells were seeded in mTeSR1 with ROCK Inhibitor (Y-27632; STEMCELL Technologies #72308) and CloneR (STEMCELL Technologies #05888). Selection was performed with 1 µg/ml puromycin, and purified cell lines were frozen down in CryoStor CS10 Cell Freezing Medium (STEMCELL Technologies #07930) and stored in liquid nitrogen.

Quality control of cell lines. All cell lines underwent quality testing for correct genetic insertion, selection efficiency, pluripotency, chromosomal abnormalities, and mycoplasma contamination. Genomic DNA from all cell lines was extracted using QuickExtract DNA Extraction Solution (Epicentre #QE09050). Correct targeting of donor construct at the AAVS1 locus was confirmed by junction PCR using the following primer pair: AAVS1 F1: 5'-ggttcggcttctggcgtgtgacc-3', AAVS1 R1: 5'-tcaagagtcacccagagacagtgac-3'. The PCR product was then sent for Sanger

sequencing using a universal sequencing primer to validate correct barcode insertion in each cell lines. Flow cytometry was performed on live cells for endogenous GFP expression and after labelling for the pluripotency marker SSEA3 (BD Biosciences #562706) and a corresponding isotype control. Cells were analysed using a BD FACSCANTO II with FACSDiva software (BD Biosciences). Data analysis was performed using FlowJo (Tree Star). Karyotyping was carried out as a professional service by Sullivan Nicolaides Pathology. iPSCs were grown in a 25 cm^2 flask to about 70-80% confluency and sent for analysis. 15 cells were examined per culture and three exemplary karyotypes were provided as results.

Single cell RNA-sequencing of undifferentiated barcoded iPSCs. All barcoded iPS cell lines were cultured in parallel as described above, dissociated using 0.5% EDTA and 600K cells from each cell line were combined. Prior to this, four cell lines were additionally labelling with different TotalSeq-A cell hashing antibodies[16] according to the manufacturer's protocol. The combined sample was transferred into 2% BSA (Sigma #A9418) in PBS, stained with Propidium Iodide and 500K viable cells were sorted using a BD Influx™ Cell Sorter (BD Biosciences) with FACSDiva software (BD Biosciences).

Single cell RNA-seq libraries were generated using the 10X Genomics Chromium 3' Gene Expression (v2) protocol, with minor modifications to the workflow, outlined by Stoeckius and Smibert (https://citeseq.files.wordpress.com/2019/02/cell_hashing_protocol_190213.pdf) to capture the fraction of droplets containing the HTO-derived cDNA (<180bp). HTO additive primers and Illumina TruSeq DNA D7xx_s primer (containing i7 index) were ordered from IDT, and used according to the cell hashing protocol. Hashtag libraries were quantified using the Agilent Bioanalyzer. Sequencing was performed using the Illumina Nextseq instrument using the Nextseq High Output 150-cycle kit and the gene expression and HTO libraries were pooled on a single flowcell using a ratio of 90:10. The standard 10X Genomics v2 3' gene expression library was processed using the 10X Genomics cell ranger pipeline to derive gene expression count matrices.

HTO-tagged cells were identified and extracted from the fastq files using the CITE-seq-Count[42] program with default parameters, to generate a count matrix of cells and their respective HTO expression values. This allowed the pooled hashing cells to be independently deconvoluted. From the 3' gene expression data, barcoded cells were identified by the expression of a barcode from the whitelist, which was included in the transcriptome reference with unique identifiers (e.g., 'BC01').

Transcriptome-based quality control filtering was performed, keeping cells with library sizes between 7500 and 60,000 reads, between 2000 and 8000 distinct features, mitochondrial reads making up no greater than 20% of all reads, and ribosomal reads mapped to between 22 and 45% of total reads. The *scds* R package[43] was also used to remove doublets predicted based on transcriptomic features. For this pilot study, a simplistic demultiplexing method for both the genomic barcodes and the cell hashing multiplexing was used where the highest barcode was taken to be the sample identity for each cell, cells with no counts mapped to any barcode were assigned as 'Negative' and cells with two or more barcodes expressing the same maximum count were determined to be 'Doublets' or 'Multiplets'. The agreement between genomic and hashing barcode assignments used is the sum of the double positives and double negatives divided by the total

number of cells. Standard normalisation, scaling, and dimensionality reduction of this dataset was done using the *Seurat* R package to generate the plots in Supplementary Fig. 1.

Cell hashed high resolution time course of mesendoderm-directed differentiation. The cell line used to generate the time course scRNA-seq dataset were WTC CRISPRi GCaMP hiPSCs. 8 plates were cultured in temporally staggered setups of mesendoderm-directed differentiation to allow the collection of 8 different time points on the same day. 1E+06 cells from each sample were labelled with a different TotalSeq-A cell hashing antibody (BioLegend) (antibodies A2051-8) and pooled to generate in total 1 sequencing library.

Multiplexing of signalling perturbations during mesendoderm differentiation. 18 barcoded iPS cell lines (WTC BC01-BC18) were cultured in parallel and multiple temporally staggered set-ups of mesendoderm-directed differentiation. Cell lines were divided into 2 batches (BC01-BC09 and BC10-BC18) to allow for the capture of biological duplicates for all 9 conditions. Cells were collected for scRNA-seq on day 2 as a reference prior to any perturbations, as well as on days 5 and 9 of differentiation. In total, 3 sequencing libraries were generated, with each library consisting of all 18 cell lines. A combination of 2 different timepoints from the two batches was pooled in each library to allow for easy detection and removal of potential batch effects during downstream analysis (Supplementary Table 1).

Signalling pathway perturbation experiment single cell library preparation and sequencing. Sample pools were assessed for quality using a hemocytometer with Trypan Blue exclusion. Cell viability ranged from 82-88%; cell concentration was between 1.3E+06 and 2.1.9E+06 cells/mL. Chromium Single Cell 3' v3 (10x Genomics) reactions were performed for each sample according to the manufacturer's protocol, targeting 20,000 cells per reaction. 11 cycles of cDNA amplification were performed in a C1000 Touch thermocycler with Deep Well Reaction Module (Bio-Rad). After clean-up of full-length amplified cDNA, 10µL was used for the construction of the gene expression library according to the manufacturer's protocol, with 11 indexing PCR cycles. Additionally, 5µL of full-length amplified cDNA was used to generate a barcoding library for each sample pool. Briefly, a first round of PCR was performed to specifically amplify cDNA regions containing the barcode cassette and append partial P5 and P7 sequencing adaptors. Each reaction contained 1x KAPA HiFi HotStart ReadyMix and 300nM each barcode_amp_F (5'-ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T-3') and barcode_amp_R (5'-CTA CAC GAC GCT CTT CCG ATC T-3') primers in a final volume of 50µL. A 2-step PCR protocol was performed with annealing/extension at 71°C for 30s, for six cycles. After a 1.2X SPRI clean-up to remove primers, a second round of PCR was performed with the entire volume of purified product from PCR1. Each reaction contained 1x KAPA HiFi HotStart ReadyMix, 500nM SI-PCR primer (5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT C-3'), and 5µL of a unique i7 indexed R primer from Chromium i7 Multiplex Kit (10x Genomics), in a total volume of 50µL. PCR was performed as for the SI-PCR protocol in the Chromium gene expression library construction workflow. Eight indexing PCR cycles were performed, for a total of 14 cycles over two rounds of PCR. Final barcoding libraries were purified using 1X SPRI beads, and fragment size and library concentration verified along with gene expression libraries using a BioAnalyzer DNA High Sensitivity Kit (Agilent). Final gene expression libraries were 62-82nM, with average size 457-494bp. Barcoding libraries were 30-38nM, with average size 358-362bp.

A single pool was prepared from the three gene expression and three barcoding libraries for sequencing. The samples were pooled equimolar within each library type and combined so that the gene expression libraries together made up 90% of the pool, and the barcoding libraries 10%. Samples were sequenced on an Illumina NovaSeq 6000 (NovaSeq control software v1.6.0/Real Time Analysis v3.4.4) using a NovaSeq 6000 S4 Kit V1 (200 cycles) (Illumina, 20027466) in standalone mode as follows: 28 bp (Read 1), 8 bp (I7 Index), and 91 bp (Read 2). Gene expression count matrices were derived using the standard 10X cell ranger pipeline. For the time course dataset, the cell hashing antibody libraries were prepared and sequenced separately to the transcriptomes using the Illumina NexSeq500Dx.

Demultiplexing and quality control. Barcoding reads used to assign sample barcodes to each cell were from the separately amplified and sequenced barcode libraries, and cell barcodes without both transcriptome and sample barcoding reads were removed. For each barcode sequencing library, the 'HTODemux' function in the *Seurat* R package[44] was used to determine the dominant sample barcode for each cell and annotate negative and doublet cells based on their sample barcode reads alone. Three transcriptome-based doublet detection methods in the *scds* R package were used to further assign doublet annotations to each cell, and cells labelled as doublets by at least three methods were removed. Transcriptome-based cell filtering as part of the *Seurat* pipeline removed cells with fewer than 2000 and greater than 7500 detected genes; fewer than 5000 and greater than 50,000 total read counts; or mitochondrial reads accounting for greater than 25% of total reads. Following filtering, sample barcodes were assigned to the remaining cells based on the barcode with the highest expression in each cell.

Preliminary analysis of each single-cell dataset separately (Supplementary Fig. 4). The normalisation, UMAP dimensionality reduction, and clustering of the data were done following the standard pipeline in the *Seurat* pipeline. The clustering resolution used was 0.2, and we assigned cell type labels by interrogating marker gene expression in each cluster. To visualise the expression of marker gene expression in the UMAP plots, we used the R package *Nebulosa*[45] which represents gene expression using kernel density estimation to account for overplotting and noise from expression drop-out.

Integration of single-cell datasets. Integration of the two scRNA-seq datasets was done using the RCPI method as part of the *RISC* R package[26], following the suggested pipeline to pre-process raw counts. We used only genes expressed in both datasets to generate 15 gene eigenvectors and using the signalling pathway perturbation dataset as the reference dataset, performed integration to return 50 principal components for further analysis. UMAP dimensionality reduction of the integrated data was also done using the *RISC* package, using 15 components of the 'PLS' embeddings as recommended for integrated values. Gene expression UMAP plots were generated using the *Nebulosa* R package.

***TRIAGE* clustering.** To assign each cell a repressive tendency score (RTS), we identified the expressed gene with the highest *TRIAGE* RTS in each cell. The *RCPI*-generated UMAP coordinates were input into the python *scipy.stats gaussian_kde* function, using the RTS vector as the "weights" parameter to estimate a probability density function of the repressive tendency across the data in UMAP space. Using a bandwidth of 0.25, a contour representation of the data was generated, splitting the density function into 10 levels. For each level, the spatial clustering algorithm *DBSCAN* from the python module *sklearn* was used to identify all spatially separated groups of cells. *TRIAGE* "peak" regions

were found by identifying the *DBSCAN* clusters with no smaller clusters within them at higher contour levels. For more information, see Sun *et al.,* 2022, code is available at https://github.com/palpant-comp/TRIAGE-Cluster.

Benchmarking and general cell type annotation with reference to *in vivo* datasets. For gene set scoring of cell type domain genes from 921 mouse tissue RNA-seq samples[29], R package *biomaRt getLDS* function was used to map the mouse ensemble gene IDs to human Ensembl gene ids. R package *topGO* was used to determine enriched GO terms for each gene set. For each gene set, the *AddModuleScore* function from the *Seurat* package was used to assign a gene set score for each cell in each peak. After thresholding scores to be greater than 0, the gene set with the highest score was assigned as the cell type domain annotation for that cell. For label transfer, both our query peaks and the reference *in vivo* datasets[1, 4] were *TRIAGE*[27] transformed as it has been observed to result in higher prediction confidence in label transfer analysis. The label transfer prediction scores were thresholded only to consider cells with scores over 0.5, and the highest scoring cell type was used to annotate each cell. For both gene set scoring and label transfer annotations, each peak's overall annotation is defined as the most common label after thresholding and is assigned as NA if most cells do not have a score greater than the respective thresholds.

VIA. We used the integrated dataset as input for the python package *pyVIA*[28], following the recommended pipeline and parameters. The root cell used was the cell captured at day 2 with the highest *POU5F1* (pluripotency marker, OCT4) expression. This inference was run on just peak cells, as well as all cells in the dataset, and the results shown in Supplementary Fig. 6 are from running *VIA* with the "is_coarse" parameter set to False.

***TRIAGE-ParseR* analysis.** We calculated the average gene expression value for each peak and converted these values to the *TRIAGE* discordance score (DS) by multiplication of each gene to its repressive tendency score[27]. We used the top 100 genes ranked by DS from each peak to calculate the pairwise Jaccard similarity index between all peaks. Peaks were clustered based on Euclidean distance with complete-linkage using the *pheatmap* R package. The top 100 DS genes for each peak were then parsed into functionally related modules using gene co-modulation analysis as described in Sun *et al.,* 2022. In brief, principal component (PC) analysis was performed on H3K27me3 deposition breadth data across all genes and cell types in the NIH Roadmap Epigenomics dataset[30], where each PC represents a biologically meaningful H3K27me3 deposition pattern. The top 67 PCs, which capture over 96% of variance in the data, are used to inform probabilistic clustering of the input genes from each peak into co-modulated sets using a Gaussian Mixture Model. The optimal number of clusters is determined by Bayesian Information Criterion, and only clusters with significant ($p<0.001$, Fisher's Exact Test) protein-protein interactions as determined by STRING analysis[46] were kept for further analysis. For more information, see Sun *et al.,* 2022, and code is available at https://github.com/palpant-comp/TRIAGE-ParseR.

Circular plot of combined dataset annotations (Fig. 5a). To generate the curved dendrogram (ring 1), the Jaccard similarities for each pair of peaks was used to calculate a Euclidean distance matrix, from which a hierarchical clustering was calculated using *hclust* in R. The dendrogram representation of the clustering was curved using *circlize_dendrogram* from the *circlize* R package. Metadata rings 2-5, co-modulation gene rings 6-7, and the co-modulation gene ontology enrichment rings were generated using the *circos-heatmap* function, using the Jaccard dendrogram and similarities to

define their order. All rings were compiled for Figure 6 in Adobe Illustrator. Code is available at https://github.com/palpant-comp/sc_barcoding_scripts.

ACKNOWLEDGEMENTS

We thank Bruce Conklin (Gladstone Institutes) for the WTC iPSC line. Sanger sequencing was performed by the Australian Genome Research Facility. Karyotyping was carried out by Sullivan Nicolaides Pathology. Flow cell sorting for single-cell capture was performed by the Queensland Brain Institute Flow Cytometry Facility at the University of Queensland. Single-cell capture was performed by the Institute for Molecular Bioscience Sequencing Core at the University of Queensland. RNA Sequencing was performed by the Garvan Weizmann Centre for Cellular Genomics at the Garvan Institute of Medical Research. Nick Valmas provided all illustrations. Funding support was provided from the National Health and Medical Research Council of Australia (Grant 1143163, N.P.), the Australian Research Council (Grant SR1101002, N.P.), the National Heart Foundation of Australia (Grant 101889, N.P.), and AIR@InnoHK administered by Innovation and technology Commission (J.W.K.H).

AUTHOR CONTRIBUTIONS

To generate the barcoded iPS cell lines, S.L. designed the barcodes; S.A. cloned the plasmids; T.W. and H.C. carried out the cell culture; T.W. and D.X. performed the CRISPR gene editing. For the pilot barcoding scRNA-seq, T.W. and H.C. carried out the cell culture and collection; S.L. and S.A. processed and sequenced the samples; S.L. and S.S performed the computational sample demultiplexing and quality control analysis. For the time course scRNA-seq dataset, S.S. and X.C. carried out cell culture, collection, and sample preparation; S.S. performed quality control and preliminary computational analysis. For the signalling perturbation scRNA-seq dataset, T.W. and H.C. carried out cell culture, collection, and sample preparation; T.W. completed quality control and preliminary computational analysis. For computational analysis of the combined dataset, S.S. integrated the data and applied the whole analysis pipeline; Y.S. and S.S. conceptualised and implemented the *TRIAGE* clustering approach; Y.S. fine-tuned and benchmarked the clustering method; W.S. conceptualised and implemented the *TRIAGE-ParseR* analysis pipeline. J.P. assisted with sequencing of barcoded libraries and facilitated data interpretation. P.P.L.T. facilitated annotation of cell types. M.B. supervised development of the *TRIAGE-ParseR* software. For the online dataset visualisation platform, D.P. constructed the dashboard and website; Z.S. implemented the visualisation of gene expression to the platform; Q.N. supervised creation of the website; J.W.K.H supervised implementation of the gene expression visualisation. N.J.P. conceptualised the study, supervised the project, and raised funding. S.S. and N.J.P. wrote the manuscript.

DECLARATIONS OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495 (2019).
2. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
3. La Manno, G. et al. Molecular architecture of the developing mouse brain. *Nature* **596**, 92-96 (2021).
4. Tyser, R.C.V. et al. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, 285-289 (2021).
5. Tabula Sapiens Consortium et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
6. Shen, S. et al. Integrating single-cell genomics pipelines to discover mechanisms of stem cell differentiation. *Trends in Molecular Medicine* **27**, 1135-1158 (2021).
7. Volpato, V. et al. Reproducibility of Molecular Phenotypes after Long-Term Differentiation to Human iPSC-Derived Neurons: A Multi-Site Omics Study. *Stem Cell Reports* **11**, 897-911 (2018).
8. Jiang, C.L. et al. Cell type determination for cardiac differentiation occurs soon after seeding of human-induced pluripotent stem cells. *Genome Biology* **23**, 90 (2022).
9. Mellough, C.B. et al. Systematic Comparison of Retinal Organoid Differentiation from Human Pluripotent Stem Cells Reveals Stage Specific, Cell Line, and Methodological Differences. *Stem Cells Transl Med* **8**, 694-706 (2019).
10. Gomez-Garcia, M.J., Quesnel, E., Al-attar, R., Laskary, A.R. & Laflamme, M.A. Maturation of human pluripotent stem cell derived cardiomyocytes in vitro and in vivo. *Seminars in Cell & Developmental Biology* **118**, 163-171 (2021).
11. Subramanian, A. et al. Single cell census of human kidney organoids shows reproducibility and diminished off-target cells after transplantation. *Nature Communications* **10**, 5462 (2019).
12. Friedman, C.E. et al. Single-Cell Transcriptomic Analysis of Cardiac Differentiation from Human PSCs Reveals HOPX-Dependent Cardiomyocyte Maturation. *Cell Stem Cell* **23**, 586-598.e588 (2018).
13. Kang, H.M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* **36**, 89-94 (2018).
14. McFarland, J.M. et al. Multiplexed single-cell profiling of post-perturbation transcriptional responses to define cancer vulnerabilities and therapeutic mechanism of action. *bioRxiv*, 868752 (2019).
15. McGinnis, C.S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nature Methods* **16**, 619-626 (2019).
16. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* **19**, 224 (2018).
17. Rosenberg, A.B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176 (2018).
18. Kiselev, V.Y., Andrews, T.S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**, 273-282 (2019).
19. Hockemeyer, D. et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature Biotechnology* **27**, 851-857 (2009).
20. Oceguera-Yanez, F. et al. Engineering the AAVS1 locus for consistent and scalable transgene expression in human iPSCs and their differentiated derivatives. *Methods* **101**, 43-55 (2016).
21. Burridge, P.W., Keller, G., Gold, J.D. & Wu, J.C. Production of de novo cardiomyocytes: human pluripotent stem cell differentiation and direct reprogramming. *Cell Stem Cell* **10**, 16-28 (2012).
22. Lian, X. et al. Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *Proceedings of the National Academy of Sciences* **109**, E1848 (2012).
23. McMahon, A.P. & Moon, R.T. Ectopic expression of the proto-oncogene int-1 in Xenopus embryos leads to duplication of the embryonic axis. *Cell* **58**, 1075-1084 (1989).
24. Xu, P.-F., Houssin, N., Ferri-Lagneau, K.F., Thisse, B. & Thisse, C. Construction of a Vertebrate Embryo from Two Opposing Morphogen Gradients. *Science* **344**, 87-89 (2014).
25. Ferrara, N. et al. Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380**, 439-442 (1996).
26. Liu, Y., Wang, T., Zhou, B. & Zheng, D. Robust integration of multiple single-cell RNA sequencing datasets using a single reference space. *Nature Biotechnology* (2021).

27. Shim, W.J. et al. Conserved Epigenetic Regulatory Logic Infers Genes Governing Cell Identity. *Cell Systems* (2020).
28. Stassen, S.V., Yip, G.G.K., Wong, K.K.Y., Ho, J.W.K. & Tsia, K.K. Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nature Communications* **12**, 5528 (2021).
29. Hutchins, A.P. et al. Models of global gene expression define major domains of cell type and tissue identity. *Nucleic Acids Research* **45**, 2354-2367 (2017).
30. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
31. Aanhaanen, W.T. et al. The Tbx2+ primary myocardium of the atrioventricular canal forms the atrioventricular node and the base of the left ventricle. *Circ Res* **104**, 1267-1274 (2009).
32. Harrelson, Z. et al. Tbx2 is essential for patterning the atrioventricular canal and for morphogenesis of the outflow tract during heart development. *Development* **131**, 5041-5052 (2004).
33. Takeuchi, J.K. et al. Tbx5 specifies the left/right ventricles and ventricular septum position during cardiogenesis. *Development* **130**, 5953-5964 (2003).
34. Volpato, V. & Webber, C. Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility. *Dis Model Mech* **13**, dmm042317 (2020).
35. Brooks, I.R. et al. Functional genomics and the future of iPSCs in disease modeling. *Stem Cell Reports* (2022).
36. Chen, Z. et al. DensityPath: an algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell RNA sequencing data. *Bioinformatics* **35**, 2593-2601 (2019).
37. Fischer, D.S. et al. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nature Biotechnology* **37**, 461-468 (2019).
38. Huang, Y. & Zhang, P. Evaluation of machine learning approaches for cell-type identification from single-cell transcriptomics data. *Brief Bioinform* **22** (2021).
39. Haniffa, M. et al. A roadmap for the Human Developmental Cell Atlas. *Nature* **597**, 196-205 (2021).
40. Yu, Q. et al. Charting human development using a multi-endodermal organ atlas and organoid models. *Cell* **184**, 3281-3298.e3222 (2021).
41. Sozen, B. et al. Reconstructing aspects of human embryogenesis with pluripotent stem cells. *Nature Communications* **12**, 5550 (2021).
42. Roelli, P., bbimber, Flynn, B., santiagorevale & Gui, G. Hoohm/CITE-seq-Count: 1.4.2. *Zenodo* (2019).
43. Bais, A.S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150-1158 (2020).
44. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821 (2019).
45. Alquicira-Hernandez, J. & Powell, J.E. Nebulosa recovers single-cell gene expression signals by kernel density estimation. *Bioinformatics* **37**, 2485-2487 (2021).
46. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607-D613 (2019).

FIGURES AND FIGURE LEGENDS

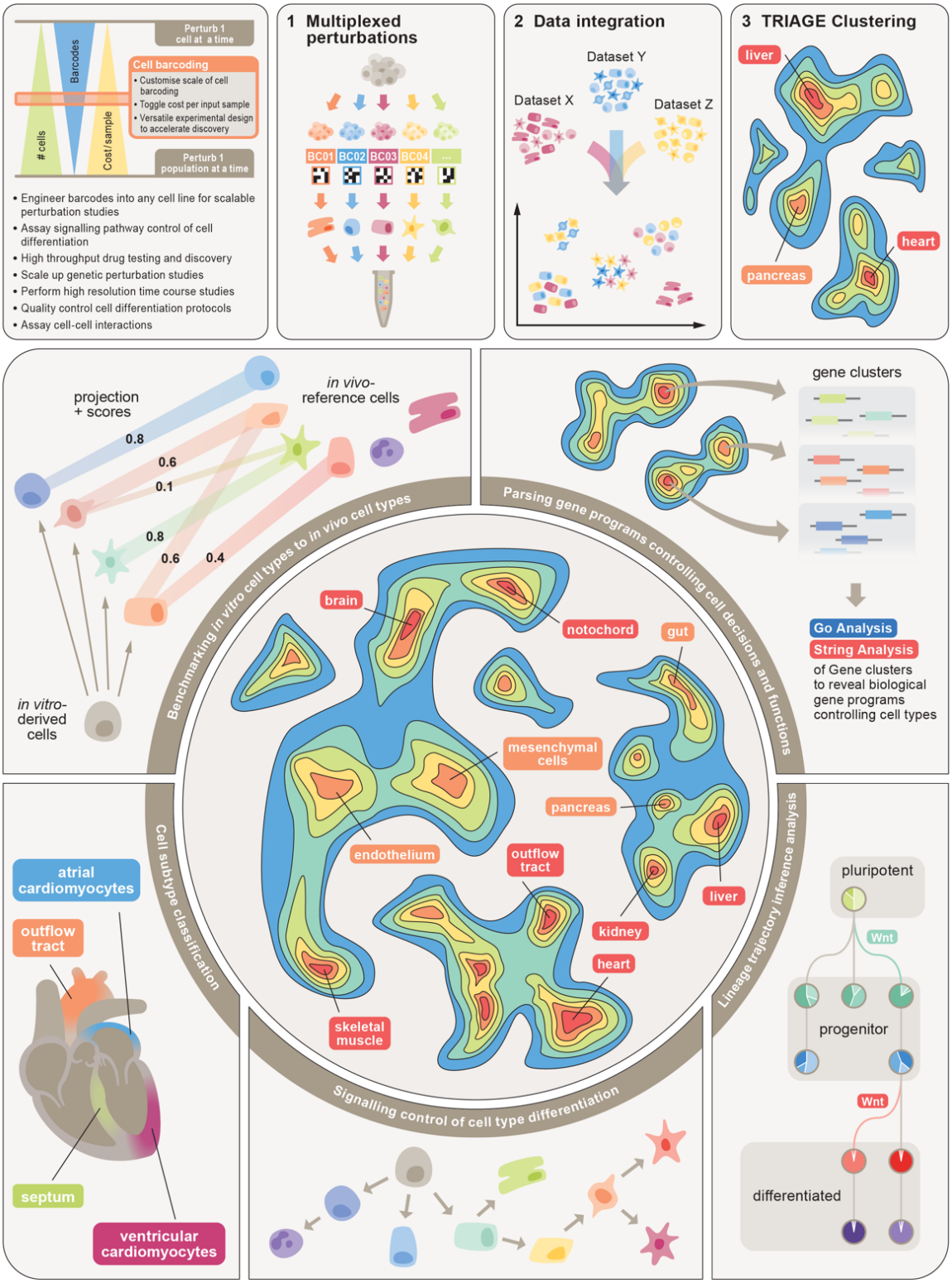
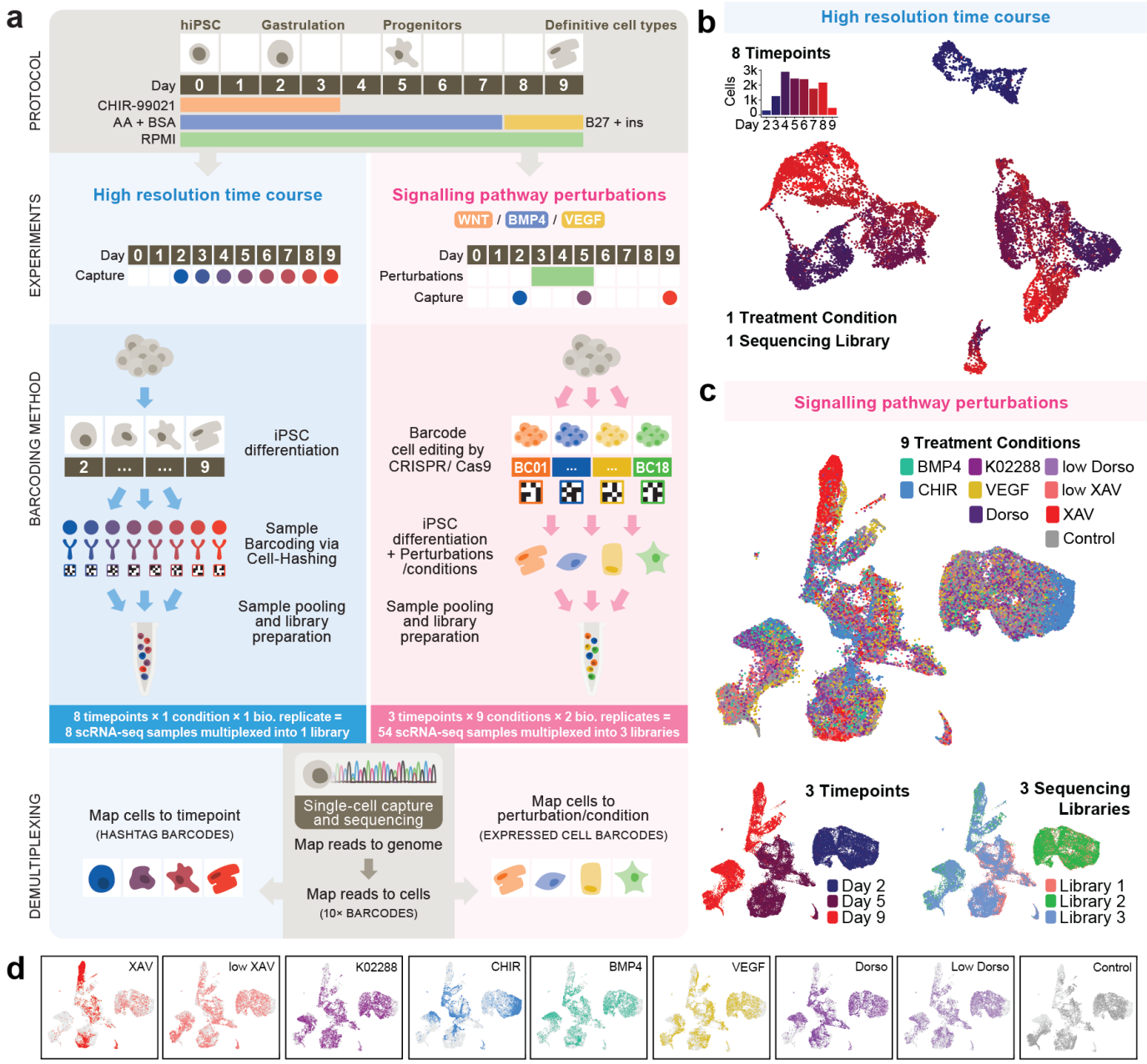


Fig. 1: Overview of pipeline for unsupervised, multiplexed analysis of lineage decisions. From top left: Multiplexed barcoding methods are implemented to enable high throughput analysis of cell differentiation. Downstream data integration and novel unsupervised analysis methods are implemented to study gene programs and signalling pathways controlling cell differentiation including methods for cell classification and benchmarking against *in vivo* cell types.



e

Condition	Molecule	Effect
XAV	5uM XAV939	Inhibition of Wnt/beta-catenin signalling via tankyrase TNKS1 & 2 inhibition
low XAV	3uM XAV939	Inhibition of Wnt/beta-catenin signalling via tankyrase TNKS1 & 2 inhibition
CHIR	3uM CHIR00921	Activation of Wnt/beta-catenin signalling via inhibition of glycogen synthase kinase 3 (GSK-3)
Dorso	4uM Dorsomorphin	Inhibition of BMP signalling, affecting BMP/Smad and TGFb pathways, also inhibits VEGFR2 and AMPK
low Dorso	0.4uM Dorsomorphin	Inhibition of BMP signalling, affecting BMP/Smad and TGFb pathways, also inhibits VEGFR2 and AMPK
K02288	0.5uM K02288	Inhibition of BMP/Smad signalling via BMP Type 1 receptors only
BMP4	10ng/mL BMP4 Ligand	Promotion of BMP signalling by introduction of recombinant BMP4 ligand
VEGF	100ng/mL VEGF Ligand	Promotion of VEGF signalling by introduction of recombinant VEGF ligand

Fig. 2: Multiplexed scRNA-seq datasets of mesendoderm differentiation.

a) Two datasets were generated. A high-resolution time course dataset (blue, left) was derived by capturing cells every 24 hours between days 2 and 9 of differentiation. A signalling pathway perturbation dataset (pink, right) was generated via perturbation of developmental signalling pathways between days 3 and 5 of differentiation with cells captured prior

to perturbation (day 2), immediately after (day 5), and at the committed cell stage (day 9). In total, 62 cell samples were captured and sequenced using 4 single cell libraries. Samples were demultiplexed by reference to RNA and sample barcodes.

b) UMAP plot showing all high quality cells (13,682 cells total) in the time course data after filtering. Cells are coloured by time point.

c) UMAP plot showing all high quality cells (48,526 cells total) in the signalling perturbation dataset after standard filtering and processing. Cells are coloured by treatment (upper), time point (lower left), and sequencing library (lower right). See Supplementary Table 1 for detailed breakdown of treatment, timepoint, and sequencing library combinations used.

d) Distribution of cells from each treatment in the signalling perturbation dataset are represented in separate UMAP plots.

e) Description of each small molecule, its concentration, and its effect used for each of the signalling perturbation samples.

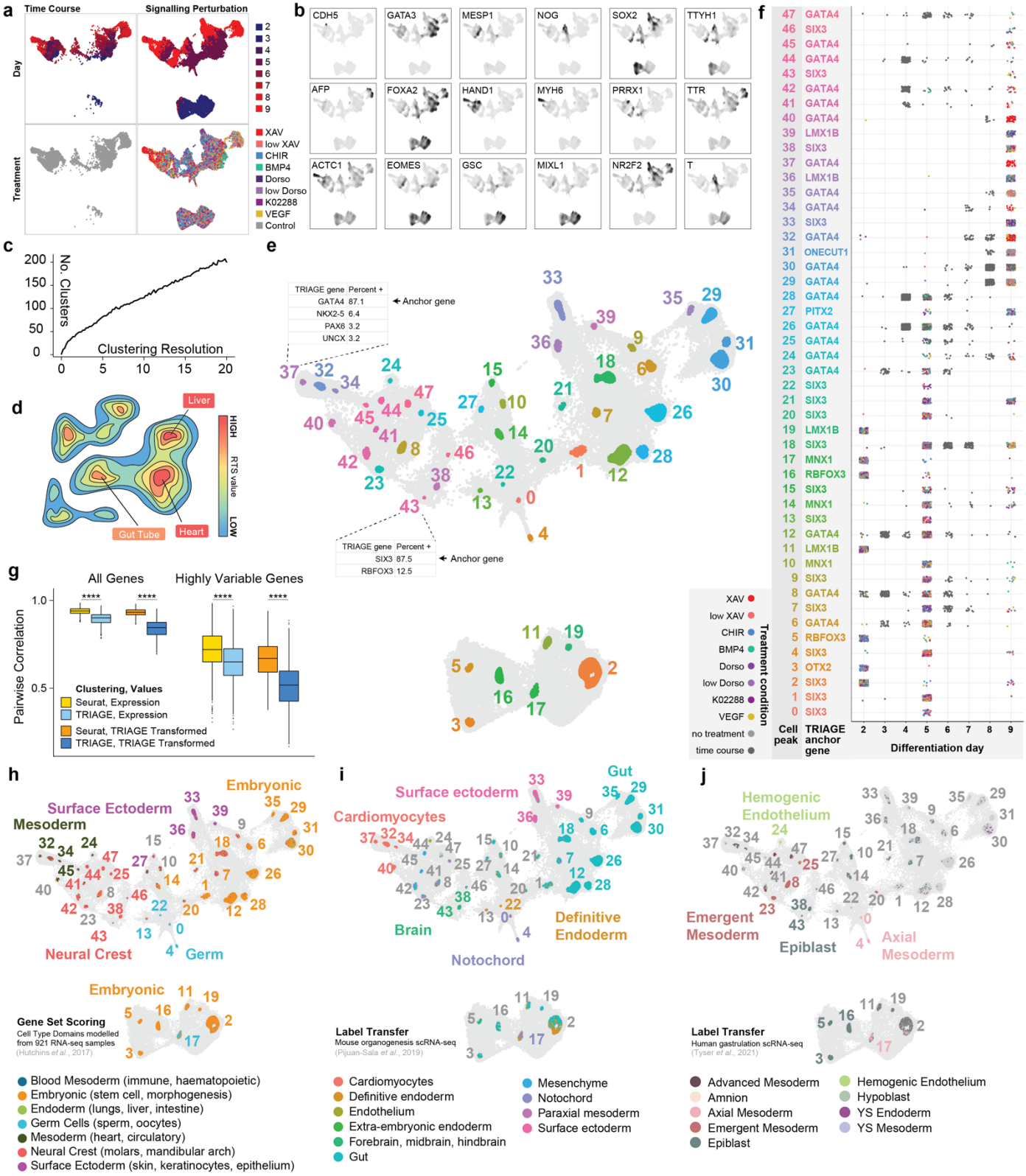


Fig. 3: Integrated atlas of *in vitro* iPSC differentiation.

a) UMAP plots of integrated datasets, split by dataset of origin (columns), coloured by time point (top row), and signalling perturbation treatment (bottom row).

b) Nebulosa[45] plots showing expression distribution of cell type marker genes.

c) Relationship between clustering resolution and number of clusters using Seurat.

d) Schematic of novel TRIAGE-based density clustering approach which anchors cell clusters by reference to regulatory genes governing cell diversity.

e) UMAP showing the 48 cell type peaks identified by *TRIAGE-Clustering*, where light grey cells do not fall within a peak. Highlighted are example proportions of *TRIAGE* genes defining two cell type peaks in the data. The *TRIAGE* gene defining the greatest percentage of cells within the peak is defined as the 'anchor gene', which provides insight into the shared regulatory basis of broad cell type groups.

f) Distribution of cells from each peak over time. Each point represents one cell and is coloured by treatment condition. TRIAGE anchor genes for each peak are listed.

g) Comparison of peak to peak transcriptomic similarity (Spearman rank correlation) using all genes (left) and just the top 1000 highly variable genes (right), compared between Seurat clustering (resolution 2.1 to achieve the same number of clusters identified in *TRIAGE*-Clustering) and *TRIAGE-clustering*, before and after TRIAGE transformation of the expression values. ****$p<2.22e-16$.

h-j) Annotation of TRIAGE cell type peaks by reference to (**h**) broad cell type domains from gene sets from 921 mouse tissue RNA-seq samples[29] using gene set scoring, (**i**) annotated cell types in single-cell datasets of mouse organogenesis[1], and (**j**) human gastrulation[4] using label transfer. Cells are coloured on a single-cell basis by their assigned label after thresholding (gene set scoring threshold = 0, label transfer threshold = 0.5), where dark grey cells do not have a label above the threshold. Peak number labels are coloured by the label that makes up the greatest proportion of cells in that peak, where grey labels indicate peaks where most cells do not score above the threshold. Light grey cells do not fall within a peak.

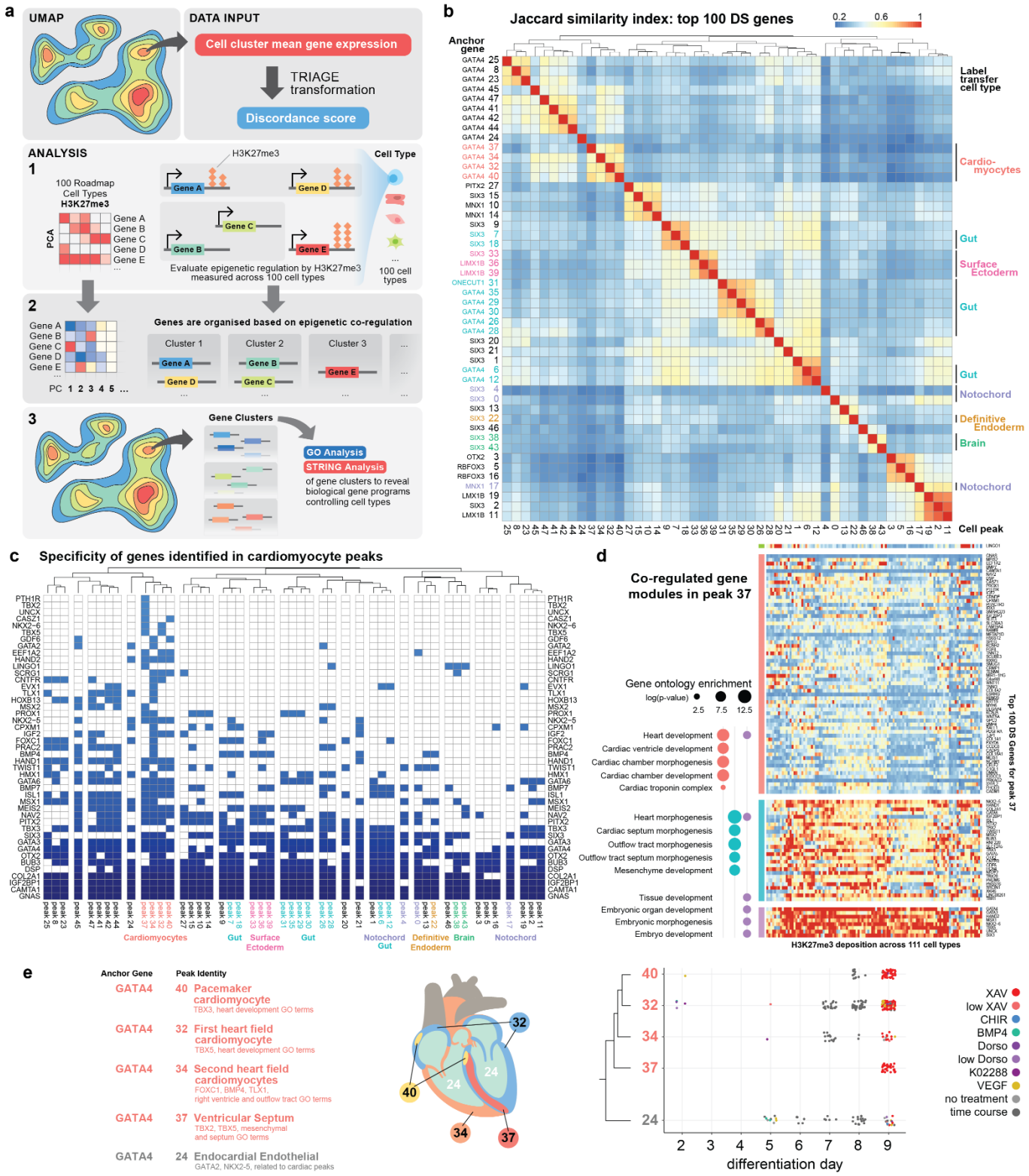


Fig. 4: Epigenetic gene co-modulation analysis identifies cell subtypes.

a) Schematic describing gene co-modulation analysis pipeline to cluster genes into related modules based on their regulation by H3K27me3 across 100 cell types from the NIH Roadmap Epigenome project. When organised into co-modulated gene sets, gene clusters from any given cell type can reveal functional and characteristic gene programs of cell subtypes.

b) Heatmap and dendrogram showing pairwise Jaccard similarity score between each peak using the top 100 genes by TRIAGE-derived discordance score (DS). *In vivo* annotated cell types from the label transfer analysis in Fig. 3i, and anchor genes from Fig. 3f for each peak are listed.

c) Specificity of the top 100 TRIAGE genes for the cardiomyocyte peaks. A coloured box indicates a gene (row) characterises that peak (column). Peaks sharing a characteristic gene indicate shared regulatory or structural features defining that cell type. Peak labels are coloured and annotated based on their label transfer annotation.

d) Heatmap showing H3K27me3 deposition across the 111 cell and tissue types in the NIH Epigenome data, over the 100 top DS genes identified for peak 37. The genes are grouped using the gene co-modulation analysis and GO term enrichment (-log(pvalue)) of select terms for each of the main gene clusters are shown (left).

e) Summarised interpretation of cell subtypes represented by each peak labelled as cardiomyocytes, their corresponding anatomical location, and distribution over time and treatment during *in vitro* differentiation from iPSCs.

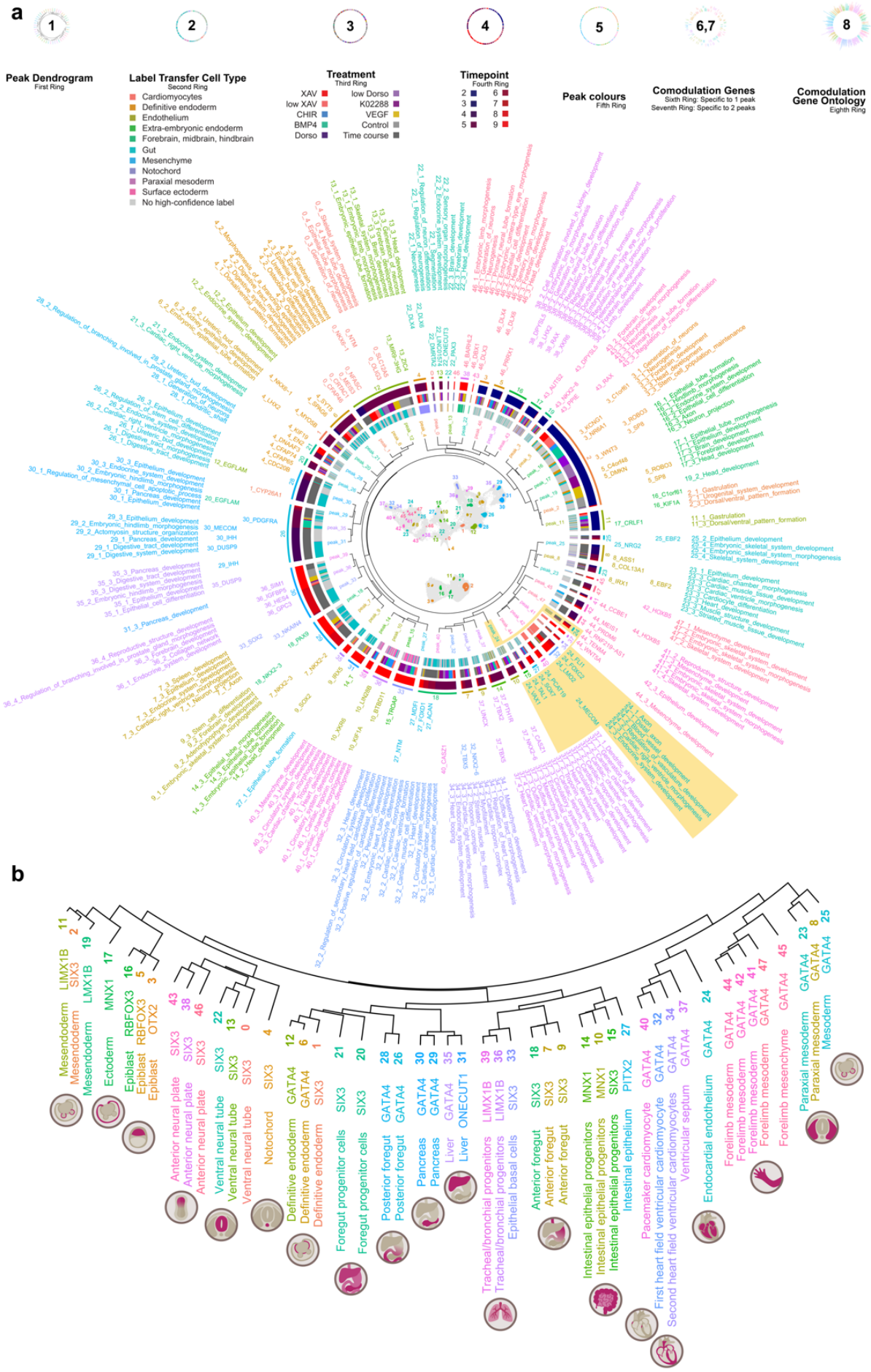

a
1
2
3
4
5
6,7
8
Peak Dendrogram
First Ring
Label Transfer Cell Type
Second Ring
Cardiomyocytes
Definitive endoderm
Endothelium
Extra-embryonic endoderm
Forebrain, midbrain, hindbrain
Gut
Mesenchyme
Notochord
Paraxial mesoderm
Surface ectoderm
No high-confidence label
Treatment
Third Ring
XAV
low XAV
CHIR
BMP4
Dorso
low Dorso
K02288
VEGF
Control
Time course
Timepoint
Fourth Ring
Peak colours
Fifth Ring
Comodulation Genes
Sixth Ring: Specific to 1 peak
Seventh Ring: Specific to 2 peaks
Comodulation Gene Ontology
Eighth Ring
b
Mesendoderm
Ectoderm
Epiblast
Anterior neural plate
Ventral neural tube
Notochord
Definitive endoderm
Foregut progenitor cells
Posterior foregut
Pancreas
Liver
Tracheal/bronchial progenitors
Epithelial basal cells
Anterior foregut
Intestinal epithelial progenitors
Intestinal epithelial progenitors
Pacemaker cardiomyocyte
First heart field ventricular cardiomyocyte
Second heart field ventricular cardiomyocytes
Ventricular septum
Endocardial endothelium
Forelimb mesoderm
Forelimb mesenchyme
Paraxial mesoderm
Mesoderm

Fig. 5: Combined dataset annotations and cell type classifications.

(a) Summarised view of all metadata for each peak – mouse gastrulation label transfer results, treatment condition, timepoint, specific genes identified in co-modulation analysis, and significantly enriched gene ontologies for each gene cluster in each peak. Endocardial endothelium peak is highlighted in yellow to provide reference for navigating across metadata for each cluster.

(b) Summarised interpretation of peak cell subtype identity using Jaccard similarity plot dendrogram (Fig. 4b) as organisation framework for cell cluster relationship.

LIST OF SUPPLEMENTARY CONTENT

Supplementary Fig. 1: Engineered barcoding cell line design and scRNA-seq pilot.

Supplementary Fig. 2: Quality control of engineered barcoding cell lines.

Supplementary Fig. 3: Quality control and filtering of scRNA-seq experiments.

Supplementary Fig. 4: Preliminary characterisation of individual scRNA-seq datasets.

Supplementary Fig. 5: Scoring and thresholding of cell type annotations with reference to *in vivo* datasets.

Supplementary Fig. 6: Performance testing of *TRIAGE* clustering analysis using VIA lineage inference.

Supplementary Table 1: Barcode sequences and multiplexed experimental design.

Supplementary Table 2: *TRIAGE* genes used for peak calling.

Supplementary Table 3: Specificity of cell identity-defining *TRIAGE-ParseR* genes across all peaks.