

SCIENTIFIC DATA

OPEN Data Descriptor: An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development

Received: 02 August 2016
 Accepted: 19 October 2016
 Published: 20 December 2016

Pang Wei Koh^{1,*}, Rahul Sinha^{2,*}, Amira A. Barkal², Rachel M. Morganti², Angela Chen², Irving L. Weissman^{2,**}, Lay Teng Ang^{3,**}, Anshul Kundaje^{1,**} & Kyle M. Loh^{2,**}

Mesoderm is the developmental precursor to myriad human tissues including bone, heart, and skeletal muscle. Unravelling the molecular events through which these lineages become diversified from one another is integral to developmental biology and understanding changes in cellular fate. To this end, we developed an *in vitro* system to differentiate human pluripotent stem cells through primitive streak intermediates into paraxial mesoderm and its derivatives (somites, sclerotome, dermomyotome) and separately, into lateral mesoderm and its derivatives (cardiac mesoderm). Whole-population and single-cell analyses of these purified populations of human mesoderm lineages through RNA-seq, ATAC-seq, and high-throughput surface marker screens illustrated how transcriptional changes co-occur with changes in open chromatin and surface marker landscapes throughout human mesoderm development. This molecular atlas will facilitate study of human mesoderm development (which cannot be interrogated *in vivo* due to restrictions on human embryo studies) and provides a broad resource for the study of gene regulation in development at the single-cell level, knowledge that might one day be exploited for regenerative medicine.

Design Type(s)	cell type comparison design • cell differentiation design • organism development design
Measurement Type(s)	transcription profiling assay • single-cell gene expression analysis • assay for transposase-accessible chromatin using sequencing • cell phenotyping
Technology Type(s)	RNA-seq assay • RNA sequencing • next generation DNA sequencing • flow cytometry assay
Factor Type(s)	tissue • cell line
Sample Characteristic(s)	Homo sapiens • embryonic stem cell • primitive streak • paraxial mesoderm • lateral plate mesoderm • somite • cardiogenic mesoderm • dermomyotome • sclerotome

¹Department of Genetics and Department of Computer Science, Stanford University, Stanford, California 94305, USA. ²Department of Developmental Biology, Institute for Stem Cell Biology and Regenerative Medicine, Ludwig Center for Cancer Stem Cell Biology and Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. ³Stem Cell & Regenerative Biology Group, Genome Institute of Singapore, A*STAR, Singapore 138672, Singapore. *These authors contributed equally to this work. **These authors jointly supervised this work. Correspondence and requests for materials should be addressed to R.S. (email: sinhar@stanford.edu) or to A.K. (email: akundaje@stanford.edu).

Background & Summary

A longstanding goal of regenerative medicine has been to efficiently differentiate stem cells into pure, functional populations of desired cell types. This has been challenging to achieve in practice: many extant differentiation methods take weeks or months to complete and result in heterogeneous mixtures of the target lineage and other contaminating lineages. Difficulties in differentiating stem cells into desired cell-types *in vitro* might stem from incomplete knowledge of how stem cells naturally develop into these lineages during the course of embryonic development.

We focus here on human mesoderm development, which starts with the differentiation of pluripotent stem cells into the primitive streak (PS) and then into paraxial and lateral mesoderm^{1–3}. Paraxial mesoderm subsequently buds off into tissue segments known as somites⁴, with dorsal somites (dermomyotome) giving rise to brown fat, skeletal muscle, and dorsal dermis, and ventral somites (sclerotome) yielding the bone and cartilage of the spine and ribs⁵. Separately, lateral mesoderm goes on to form limb bud mesoderm⁶ and cardiac mesoderm⁷, the latter of which generates cardiomyocytes and other heart constituents.

Our related publication⁸ delineated a comprehensive roadmap for human mesoderm development that outlined key intermediate stages and defined the minimal combinations of extrinsic signals sufficient to induce differentiation at each stage. To elicit differentiation at defined stages, in addition to identifying the necessary inductive cues at each stage (as is typical), we also identified pathways leading to ‘unwanted’ cell fates and systematically repressed them at each lineage branchpoint. We used this strategy to efficiently differentiate pluripotent stem cells, through anterior and mid primitive streak, into paraxial and lateral mesoderm, and subsequently into somites, sclerotome, dermomyotome, and cardiac mesoderm (Fig. 1). The identity and purity of these cell types was respectively assessed by transplantation into mouse models or single-cell gene expression profiling⁸.

Here we describe in detail the materials and methods used to generate and profile these distinct cell types, with an eye towards promoting reproducibility and reuse of our data. We focus on the biological methods used to generate the data; the computational pre- and post-processing of the data; and the technical validation of the quality of our data. In contrast, our related publication⁸ focused on experimentally validating the biological function and purity of the differentiated cell types and on extracting developmental insights from the data.

Our dataset comprises three main types of data -- gene expression, chromatin accessibility, and surface marker expression -- across 10 different cell types (pluripotent stem cells, anterior PS, mid PS, paraxial mesoderm, somitomers, somites, sclerotome, dermomyotome, lateral mesoderm and cardiac mesoderm). For expression, we performed bulk-population RNA-seq as well as single-cell RNA-seq (using the Fluidigm C1 system) on a total of 651 cells spanning all lineages. Chromatin accessibility across the genome was measured by ATAC-seq⁹. For each lineage, two to six biological replicates were assayed for bulk-population RNA-seq and ATAC-seq. Finally, the expression of 332 cell-surface markers was ascertained on most lineages by means of high-throughput antibody screening.

Taken together, this dataset will constitute a useful resource for the study of human mesoderm development. For example, this dataset enabled us to identify novel marker genes in somitogenesis (a transient process which cannot be observed *in vivo* due to restrictions on the use of human embryos); identify the putative cell-of-origin for different subtypes of congenital scoliosis; and infer the activity of transcription factors at each stage of mesodermal development⁸. The data from the high-throughput surface marker screen will also be helpful in purifying desired cell types for transplantation or further study.

Moreover, we believe that this dataset will be useful as a broader resource for the analysis of a timecourse data, e.g., as a testing ground for algorithms that aim to reconstruct developmental paths from single-cell RNA-seq data^{10,11}, or for the study of how changes in chromatin accessibility are correlated with, and are ultimately causative of, changes in gene expression across developmental time and space.

Methods

We reproduce here the experimental protocols included in our related publication⁸, with added detail on our computational processing steps, RNA library construction, and surface marker screening. A list of all experiments reported here, together with accession codes of the corresponding data, can be found in Table 1 (available online only).

Bulk-population RNA-seq

RNA extraction, library preparation and sequencing. For bulk-population RNA-seq, RNA was extracted from either whole cell populations or alternatively, cell subsets purified by fluorescence activated cell sorting (FACS). In brief, RNA was obtained from undifferentiated H7 hESCs (day 0 of *in vitro* differentiation), H7-derived anterior primitive streak populations (day 1), H7-derived mid primitive streak populations (day 1), H7-derived lateral mesoderm (day 2), H7-derived FACS-purified GARP+ cardiac mesoderm (day 3), H7-derived FACS-purified DLL1+ paraxial mesoderm populations (day 2), H7-derived day 3 early somite progenitor populations (day 3), H7-derived dermomyotome populations (day 5, treated with BMP4+CHIR99021+Vismodegib on days 4–5), and H7-derived FACS-purified PDGFR α + sclerotome populations (day 6).

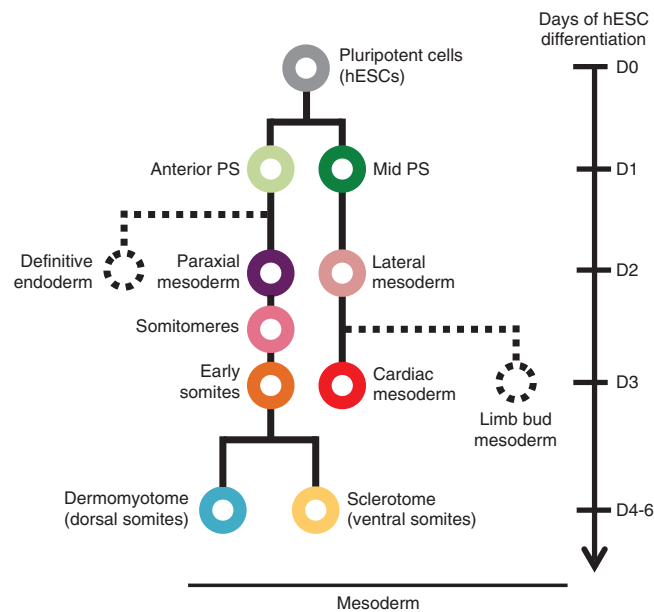


Figure 1. A schematic of human mesoderm development. We differentiate and profile each of the 10 cell types shown in color here, starting with pluripotent stem cells and ending in dermomyotome, sclerotome, and cardiac mesoderm.

Total RNA from the above cell populations was isolated using Trizol (Thermo Fisher) as per the manufacturer's recommendations, with the additional use of linear polyacrylamide (Sigma) as a carrier to facilitate RNA precipitation. Purified total RNA was treated with 4 units of RQ1 RNase-free DNase (Promega) at 37 degrees Celsius for 1 h to remove trace amounts of genomic DNA. The DNase-treated total RNA was cleaned-up using the RNeasy Micro Kit (Qiagen). Subsequently, the integrity of extracted RNA was assayed by on-chip electrophoresis (Agilent Bioanalyzer) and only samples with a high RNA integrity (RIN) value were used for subsequent cDNA library preparation.

Purified total RNA (10–50 ng) was reverse-transcribed into cDNA and amplified using the Ovation RNA-seq System V2 (NuGEN). Amplified cDNA was sheared using the Covaris S2 (Covaris) with the following settings: total volume 120 μ l, duty cycle 10%, intensity 5, cycle/burst 100 and total time 2 min. The sheared cDNA was cleaned up using Agencourt Ampure XP beads (Beckman Coulter) to obtain cDNA fragments \geq 400 base pairs (bp). 500 ng of sheared and size-selected cDNA was used as input for library preparation using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs) as per the manufacturer's recommendations. Resulting libraries (fragment distribution: 300–700 bp; peak 500–550 bp) were pooled (multiplexed) and sequenced using either a HiSeq 4000 or NextSeq 500 (Illumina) at the Stanford Functional Genomics Facility to obtain 2×150 bp paired-end reads. For each RNA-seq library, the effectiveness of adapter ligation and the effective library concentration was determined by qPCR and Bioanalyzer (Agilent) prior to pooling and loading them onto the sequencers.

Each sample in our data constitutes a separate biological replicate. Bulk population RNA-seq libraries were prepared in three batches (Table 2).

Quantification and processing. Obtained RNA-seq reads were trimmed for base call quality (PHRED score \geq 21) and for adapter sequences (using Skewer¹²), and then were subsequently processed using a slightly-modified version of the ENCODE long RNA-seq pipeline for quantification of mRNA expression (<https://www.encodeproject.org/rna-seq/long-rnas/>)¹³. Specifically, reads were aligned to hg38 using STAR 2.4 (ref. 14); gene-level expression was then quantified using RSEM 1.2.21 (ref. 15). We only kept samples with at least 10,000,000 uniquely mapping reads and with at least 50% of reads uniquely mapping, which meant rejecting one sample (from sclerotome) out of 34. The numbers and percentages of uniquely mapping reads for each sample are listed in Table 2. The full parameter settings used can be found in our versions of `STAR_RSEM.sh` and `STAR_RSEM_prep.py` (see Code Availability below).

To facilitate global comparisons of gene expression levels across cell types, we first took the log₂TPM (transcripts per million) values for each gene, before filtering out all genes where there was a difference of less than 2 (in log₂TPM units, i.e., a 4-fold difference in expression) between the cell types with the highest and lowest expression. Next, we used ComBat with non-parametric priors¹⁶ (as implemented through the *sva* R package¹⁷) to correct for batch effects. This sometimes left small negative values for the expression of some genes, which we set to 0. The R Markdown script implementing this batch correction is `bulkDataViz.Rmd`.

Sample ID	Celltype	Batch	Number of uniquely mapped reads	Percentage of uniquely mapping reads
H7hESC_1	D0 H7 hESC	1	29077933	77.64
H7hESC_2	D0 H7 hESC	1	32593810	72.39
H7hESC_3	D0 H7 hESC	1	30800393	74.51
H7Trzl	D0 H7 hESC	3	60399642	76.01
APS_1	D1 Anterior Primitive Streak	1	29444532	71.43
APS_2	D1 Anterior Primitive Streak	1	30585671	72.73
APS_3	D1 Anterior Primitive Streak	3	62559303	74.76
MPS_1	D1 Mid Primitive Streak	1	32195105	74.5
MPS_2	D1 Mid Primitive Streak	1	29668483	76.07
MPS_3	D1 Mid Primitive Streak	3	64236835	73.91
MPS_4	D1 Mid Primitive Streak	3	76385886	72.9
DLL1nD2nonPXM_1	D2 DLL1 – Paraxial Mesoderm	2	28555523	64.46
DLL1nD2nonPXM_2	D2 DLL1 – Paraxial Mesoderm	2	24999466	66.51
DLL1pPXM_3	D2 DLL1+ Paraxial Mesoderm	2	11948788	50.27
DLLpPXM_1	D2 DLL1+ Paraxial Mesoderm	2	29595403	73.74
DLL1pPXM_2	D2 DLL1+ Paraxial Mesoderm	2	25504394	71.61
D2LtM_1	D2 Lateral Mesoderm	3	79623847	72.3
D2LtM_2	D2 Lateral Mesoderm	3	85281213	70.89
D3GARPpCrdcM_1	D3 GARP+ Cardiac Mesoderm	3	87530457	69.6
D3GARPpCrdcM_2	D3 GARP+ Cardiac Mesoderm	3	94587346	74.36
Smt_1	D3 Somite	1	15272975	57.37
Smt_2	D3 Somite	1	24702016	60.63
D3EarlySmt_1	D3 Somite	3	60723730	65.19
D3EarlySmt_2	D3 Somite	3	64467176	64.97
Smt_4	D3 Somite	2	14235841	62.14
Smt_3	D3 Somite	2	20701016	62.06
Drmmtm_1	D5 Dermomyotome	1	32318123	73.43
Drmmtm_2	D5 Dermomyotome	1	18890789	63.49
D5CentralDrmmtm	D5 Dermomyotome	3	126799726	69.06
Drmmtm_3	D5 Dermomyotome	2	10583923	65.92
Scrltm_1	D6 PDGFRA+ Sclerotome	1	20751549	58.61
Scrltm_2	D6 PDGFRA+ Sclerotome	1	23117071	66.89
D6PDGFRApScrltm_1	D6 PDGFRA+ Sclerotome	3	102352286	74.86

Table 2. Bulk-population RNA-seq metadata and mapping statistics.

For ease of use, we also prepared a spreadsheet with TPM values for each gene, augmented with the following information on each gene: 1) whether the gene product is present on the cell surface (GO code GO:0009986); 2) for each pair of adjacent conditions, whether the gene was differentially expressed between those conditions; and 3) the shrunken log-fold-change for that gene between those conditions. We provide (1) as a convenience to help in finding potential surface markers that were not included in our high-throughput screen (e.g., because an antibody was not available). (2) and (3) were calculated by DESeq2 (ref. 18) using batch information; genes were called as differentially expressed at a false discovery rate (FDR) of 0.1.

The raw data from the bulk-population RNA-seq can be found in [Data Citation 1]. A spreadsheet of TPM values can be found in [Data Citation 2]. The annotated spreadsheet, as described in the previous paragraph, is in [Data Citation 3].

Single-cell RNA-seq

Library preparation and sequencing. Cells were briefly washed (DMEM/F12), dissociated (TrypLE Express), strained (100 μ m filter), pelleted and re-suspended in DMEM/F12 for counting. Before single-cell capture, two quality control steps were implemented. First, cell size was estimated in order to determine whether cells should be loaded onto C1 capture arrays of either 10–17 μ m or 17–25 μ m size. Arrays were chosen for each lineage by estimating the median cell size of each given population on a flow cytometer on the basis of the FSC-W signal¹⁹ and choosing an array with an appropriate pore size to

accommodate such cells. Second, to ensure the high viability of *in vitro*-differentiated cells prior to commencing single-cell RNA-seq, for each population a separate aliquot of cells was stained with 1.1 μ M DAPI and analyzed by flow cytometry; for all cell populations that were used for single-cell RNA-seq, >98% of cells were viable (i.e., DAPI negative).

For single-cell capture, cells were diluted to a concentration of 1000 cells per μ l, diluted in a 3:2 mixture of C1 Cell Suspension Reagent and DMEM/F12, and then loaded onto a Fluidigm C1 single-cell capture array chip for automated capture on a Fluidigm C1 Machine (Stanford Stem Cell Institute Genomics Core). 10–17 μ m array chips were used for hESCs, day 1 anterior PS, day 2 sorted DLL1+ paraxial mesoderm, day 2.25 somitomeres, day 3 early somites, day 2 lateral mesoderm, day 3 sorted GARP+ cardiac mesoderm, day 5 central dermomyotome, and day 6 sorted PDGFRA+ sclerotome while a 17–25 μ m array chip was used for day 1 mid PS.

After loading, the efficiency of single-cell capture was verified using an automated microscope that imaged each captured cell on the chip. Subsequent cell lysis, cDNA synthesis, and amplification was executed within each microfluidic chamber in the array chip in an automated fashion with the Fluidigm C1 machine using the reagents from SMARTer Ultra Low RNA Kit (Clontech, 634833), as per the manufacturers' instructions (Fluidigm, PN 100–7168 Rev. A2). The amplified cDNA from individual cells was harvested into a nuclease-free 96-well plate and diluted using the C1 harvesting reagent (Fluidigm). The concentration and integrity of amplified cDNA were assessed using a Fragment Analyzer (Advanced Analytical) in 96-well plate format. Amplified cDNAs from only those wells that (1) were not degraded and (2) originated from wells that were microscopically verified manually to contain a single cell, were carried forward for subsequent library construction. It is important to note that because of manual verification, we were able to effectively rule out doublets if captured in the medium (10–17 μ m) or the large (17–25 μ m) array chips.

A single-channel liquid handling robot, Mosquito X1 (TTP Labtech), was used to simultaneously, 1) dilute amplified cDNAs from single cells from all lineages to a concentration range of 0.05–0.16 ng per μ l with C1 Harvest Reagent (Fluidigm) as a diluent and 2) consolidate the diluted cDNA into 384 well plates. The diluted single-cell cDNAs were tagged and converted to sequencing libraries in the 384 well plates using the Nextera XT DNA Sample Prep Kit (Illumina, FC-131–1096) in an automated fashion using another 16-channel pipetting robot, Mosquito HTS (TTP Labtech), and 384 distinct Illumina-compatible molecular barcodes. The resulting sequencing libraries from a single such 384 well plate were then pooled and cleaned up using Agencourt AMPure XP beads (Beckman Coulter). The pooled libraries were then analyzed for quality and concentration using Bioanalyzer (Agilent) and qPCR and loaded on a single lane of NextSeq 500 or two lanes of HiSeq 4000 to obtain 1–2 million 2×150 bp reads per cell. The reads obtained were trimmed for base call quality (PHRED score ≥ 21) and the presence of adapter sequences using Skewer¹².

Quantification and processing. We quantified single-cell gene expression using the ENCODE long RNA-seq pipeline (with the same parameter settings as employed for analysis of bulk-population RNA-seq). We only kept samples with at least 1 million uniquely mapped reads and at least 70% of reads uniquely mapping, which meant keeping data from 498 single cells out of 651. The numbers and percentages of uniquely mapping reads for each cell are listed in Table 3 (available online only).

We next filtered out genes with low or undetectable expression by only considering genes with least 20 cells (across all 498 retained cells) showing a \log_2 (TPM+1) value of at least 10 for that gene. As with the data from the bulk-population RNA-seq, when performing analyses comparing cell types to one another, we additionally filtered out genes whose \log_2 (TPM+1) values did not vary by a difference of at least 2 (i.e., a 4-fold difference in expression) between the cell types with the highest and lowest expressions.

The raw data from the single-cell RNA-seq can be found in [Data Citation 1]. A spreadsheet of TPM values can be found in [Data Citation 2].

ATAC-seq

Library preparation and sequencing. ATAC-seq was performed as described previously⁹, with minor modifications. In brief, for each replicate, 50,000 cells were lysed in lysis buffer containing 0.01% IGEPAL CA-630 (Sigma, I8896) to obtain nuclei, which were directly used in the Tn5 transposition reaction (reagents from Nextera DNA Sample Preparation Kit; Illumina, FC-121–1030). Immediately following transposition, DNA fragments were purified (MinElute Kit, Qiagen) and PCR amplified for a total of 12–13 cycles using previously-designed primers that included Illumina compatible adapters and barcodes⁹. The resulting ATAC-seq libraries were purified (MinElute Kit, Qiagen) and pooled, and final library-pool concentrations were assessed (Bioanalyzer) prior to next-generation sequencing. The quality of ATAC-seq libraries was confirmed by a shallow sequencing run using a MiSeq v3 (Stanford Functional Genomics Facility, 2×75 bp reads) before deep sequencing was performed on a NextSeq 500 (2×75 bp reads). Two replicates were analyzed per cell-type.

Quantification and processing. We used the ATAc pipeline²⁰ to process the ATAC-seq reads, starting with adapter trimming and then alignment to hg19 (Bowtie2 (ref. 21)). While we used hg38 for RNA-seq alignment, we opted for hg19 for ATAC-seq because of the availability of a curated blacklist of

artifactual regions in hg19 (ref. 13). We then filtered out reads based on a variety of criteria (excluding unmapped reads, mate-unmapped reads, secondary alignments, duplicates (using Picard's MarkDuplicates²²), multi-mapping reads (MAPQ < 30), and mitochondrial reads), retaining only high-read-quality, properly-paired reads.

Two biological replicates were assayed by ATAC-seq for each cell-type. As the post-filtering sequencing depth varied between replicates and cell types, we subsampled each replicate to a maximum of 35 M uniquely-mapping reads (post-filtering) to improve comparability between samples. We next used MACS2 (ref. 23) to call peaks for each replicate, with a relaxed false discovery rate (FDR) threshold of 0.01, and then created a unified peak list for each cell type by selecting only peaks that were reproducible between both replicates. This was done through an irreproducible discovery rate (IDR) analysis²⁴, similar to what was previously described by the ENCODE Consortium²⁵. In brief, the IDR method takes in peak calls from a pair of replicates, filters out all peaks that only appear in one replicate, and then uses a copula mixture model to model the remaining peaks as belonging to either a reproducible 'signal' population or an irreproducible 'noise' population'. We used an IDR threshold of 0.1, i.e., we only retained peaks that were deemed to have come from the 'signal' population with a probability of more than 0.9 after a multiple testing correction. Finally, we filtered out all peaks that appeared in the aforementioned blacklist of artifactual regions in hg19 (<https://www.encodeproject.org/annotations/ENCSR636HFF/>).

We note that this ATAC-seq analysis pipeline is an improved version of the one used for analysis in our related publication⁸. In particular, here we adjusted the IDR threshold, the shift size parameter for MACS2, and a multi-mapping parameter, resulting in increased sensitivity for peak detection.

To obtain a universal list of peaks across all cell-types, we used BEDtools²⁶ to merge the lists of filtered, reproducible peaks for each cell-type, resulting in a total of 166,256 peaks. For each cell-type, we then pooled its two biological replicates together and called peaks (MACS2) on the pooled reads. To obtain a single measure of confidence at each peak P in the universal list for each cell-type C, we took the highest $-\log_{10}$ P-value out of all peaks in the pooled replicates for C that intersected with P.

The raw ATAC-seq data can be found in [Data Citation 1]. The peak calls can be found in [Data Citation 2]. ATAC-seq metadata is tabulated in Table 4 (available online only).

High-throughput surface marker screening

High-throughput, antibody-based screening of surface markers expressed on various mesodermal progenitors was performed as described in our related publication⁸ and explained in further detail here. The following lineages, derived from the indicated embryonic stem cell lines, were screened using this approach: undifferentiated H7 hESCs ('undifferentiated hESCs'), H7-derived day 2 paraxial mesoderm ('paraxial mesoderm'), H7-derived day 3 early somite progenitors ('early somite'), H7-derived day 5 dermomyotome ('dermomyotome'), H7-derived day 6 sclerotome ('sclerotome'), *MIXL1-GFP* reporter HES3 hESC-derived day 1 anterior primitive streak ('primitive streak') and finally, *NKX2.5-GFP* reporter HES3 hESC-derived day 3 cardiac mesoderm ('cardiac mesoderm'). 10–70 million cells of each lineage were used in each surface-marker screen. Due to limited resources, we did not include mid primitive streak and lateral mesoderm in this screen.

Prior to antibody staining, hESCs or their differentiated mesodermal progeny were dissociated by brief 37 C incubation in TrypLE Express (Gibco). TrypLE Express was chosen as a dissociation reagent, as it has been previously shown to minimally cleave cell-surface epitopes²⁷, which would otherwise confound surface marker screening data. After cell detachment, they were washed off plates in a large excess of DMEM/F12 to neutralize the dissociation reagent, filtered to remove large cell clumps, pelleted by centrifugation, and re-suspended in approximately 30 ml of Cell Suspension Buffer (Biological).

To conduct antibody screening, a multichannel pipette was used to plate the cell suspension into individual wells of four 96-well plates, each well containing a distinct PE-conjugated antibody against a human cell-surface antigen, altogether totaling 332 unique cell-surface markers across multiple 96-well plates (LEGENDScreen PE-Conjugated Human Antibody Plates; Biologend, 700001). Cells were stained with respective antibodies for 30 min at 4 C, washed twice with Cell Staining Buffer and then finally re-suspended in Cell Staining Buffer containing 1.1 μ M DAPI (Biologend) as a viability dye before analysis on an LSR Fortessa (Stanford Stem Cell Institute FACS Core). Stained cells were not fixed prior to FACS analysis.

The percentage of viable (DAPI-negative cells) for each lineage that expressed each given surface marker was determined by rigorously gating the PE fluorescent signal such that no more than several percent of negative control cells (unstained cells or cells that were stained with an isotype control antibody directed against no known cellular antigen) were regarded positive. For analysis of surface-marker expression on *MIXL1-GFP* reporter HES3 hESC-derived primitive streak or *NKX2.5-GFP* reporter HES3 hESC-derived day 3 cardiac mesoderm, cells were respectively pre-gated on the *MIXL1-GFP+* and *NKX2.5-GFP+* fractions before analysis of PE signal intensity. Multicolor compensation was conducted to control for fluorescent bleedthrough between the PE and GFP channels.

A table with the percentage of viable cells in each lineage that expressed each given surface marker can be found in [Data Citation 4]. Metadata for the surface marker screen is tabulated in Table 5 (available online only).

Code availability

All custom code used in this work is available at <https://github.com/kundajelab/mesoderm>. This includes R Markdown files that reproduce the figures in this paper.

For RNA-seq processing and quantification, we used STAR 2.4 (ref. 14), RSEM 1.2.21 (ref. 15), and Skewer 0.1.127 (ref. 12). The full parameter settings for STAR and RSEM can be found in STAR_RSEM.sh and STAR_RSEM_prep.py in the Github repository above. For bulk-population RNA-seq read processing, we used the following parameters for Skewer:

```
-x AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTG
CTTG
-y AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
-t 16 -q 21 -l 21 -n -u -f sanger
```

For single-cell RNA-seq read processing, we used the following parameters for Skewer:

```
-x CTGTCTCTTATACACATCTCCGAGCCCACGAGACNNNNNNNNNATCTCGTATGCCGTCTTCTG
CTTG
-y CTGTCTCTTATACACATCTGACGCTGCCGACGANNNNNNNNNGTGTAGATCTCGGTGGTCG
CCGTATCATT
-t 16 -q 21 -l 21 -n -u -f sanger
```

For ATAC-seq processing, we used commit 9077b9... of the ATAC pipeline²⁰. In turn, this used MACS2 2.1.0 (ref. 23) and Bowtie2 2.2.6 (ref. 21).

Differentiation

Human pluripotent stem cell culture. H7, *MIXL1-GFP* HES3, *NKX2.5-GFP* HES3, *SOX17-mCherry* H9, *pCAG-GFP* H7, *EF1A-BCL2-2A-GFP* H9 and *UBC-Luciferase-2A-tdTomato*; *EF1A-BCL2-2A-GFP* H9 hESCs and BJC1 hiPSCs were routinely propagated feeder-free in mTeSR1 medium (StemCell Technologies)+1% penicillin/streptomycin (Gibco) on cell culture plastics coated with Geltrex basement membrane matrix (Gibco). Undifferentiated human pluripotent stem cells (hPSCs) were maintained at high quality with particular care to avoid any spontaneous differentiation, which would confound downstream differentiation. Unless otherwise indicated, the majority of experiments performed in this study were conducted using H7 hESCs, including all bulk-population RNA-seq, single-cell RNA-seq, and ATAC-seq experiments.

Directed differentiation in defined medium. Partially-confluent wells of undifferentiated hPSCs were dissociated into very fine clumps using Accutase (Gibco) and sparsely passaged 1:12-1:20 onto new Geltrex-coated cell culture plates in mTeSR1 supplemented with 1 μ M thiazovivin (Tocris; a ROCK inhibitor to prevent cell death after dissociation) overnight. Seeding hPSCs sparsely prior to differentiation was critical to prevent cellular overgrowth during differentiation, especially during long-duration differentiation. hPSCs were allowed to plate overnight. The following morning, they were briefly washed (in DMEM/F12) before the addition of differentiation medium. All differentiation was conducted in serum-free, feeder-free and monolayer conditions in chemically-defined CDM2 basal medium.

The composition of CDM2 basal medium²⁸ was as follows: 50% IMDM (+GlutaMAX, +HEPES, +Sodium Bicarbonate; Gibco, 31980-097)+50% F12 (+GlutaMAX; Gibco, 31765-092)+1 mg ml⁻¹ polyvinyl alcohol (Sigma, P8136-250G)+1% v/v concentrated lipids (Gibco, 11905-031)+450 μ M monothioglycerol (Sigma, M6145)+0.7 μ g ml⁻¹ insulin (Roche, 1376497)+15 μ g ml⁻¹ transferrin (Roche, 652202)+1% v/v penicillin/streptomycin (Gibco). Polyvinyl alcohol was brought into solution by gentle warming and magnetic stirring in IMDM/F12 media before addition of additional culture supplements.

Primitive streak induction. As previously described⁸, after overnight plating, hPSCs were briefly washed (with DMEM/F12) and then differentiated into either anterior primitive streak (30 ng ml⁻¹ Activin A+4 μ M CHIR99021+20 ng ml⁻¹ FGF2+100 nM PIK90; for subsequent paraxial mesoderm induction) or mid primitive streak (30 ng ml⁻¹ Activin A+40 ng ml⁻¹ BMP4+6 μ M CHIR99021+20 ng ml⁻¹ FGF2+100 nM PIK90; for subsequent cardiac mesoderm induction) for 24 h. Though both types of primitive streak broadly expressed pan-primitive streak markers (e.g., *MIXL1* and *BRACHYURY*), anterior and mid primitive streak lineages were distinguished by expression of distinct region-specific markers and differing developmental competence to develop into downstream lineages⁸.

Subsequently, day 1 anterior primitive streak was briefly washed (DMEM/F12) and differentiated towards day 2 paraxial mesoderm for 24 h (1 μ M A-83-01+3 μ M CHIR99021+250 nM LDN-193189 [DM3189]+20 ng ml⁻¹ FGF2). Separately, day 1 mid primitive streak was differentiated towards day 2 lateral mesoderm for 24 h (1 μ M A-83-01+30 ng ml⁻¹ BMP4+1 μ M C59; with 2 μ M SB-505124 sometimes used instead of A-83-01)⁸.

Paraxial mesoderm downstream differentiation. Day 2 paraxial mesoderm was briefly washed (DMEM/F12) and further differentiated into day 3 early somite precursors for 24 h (1 μ M A-83-01+250 nM LDN-193189+1 μ M C59+500 nM PD0325901). Subsequently, day 3 early somites were dorsoventrally patterned into either ventral somites/sclerotome (5 nM 21 K+1 μ M C59) or dorsal somites/

dermomyotome (3 μM CHIR99021+150 nM Vismodegib). Sclerotome induction was conducted for 48–72 h (leading to day 5–6 ventral somite progenitors). For dermomyotome induction, sometimes dermomyotome was induced in the presence of 50 ng ml^{-1} BMP4 to upregulate *PAX7* after 48 h of BMP4+CHIR99021+Vismodegib differentiation (leading to day 5 dermomyotome progenitors)⁸. Media was changed every 24 h for all steps. The small-molecule Hedgehog agonist 21 K²⁹ was commercially synthesized.

Lateral/cardiac downstream differentiation. Day 2 lateral mesoderm was differentiated into day 4 cardiac mesoderm by treating them with 1 μM A8301+30 ng ml^{-1} BMP4+1 μM C59+20 ng ml^{-1} FGF2 for 48 h, or alternatively, with 1 μM A8301+30 ng ml^{-1} BMP4+20 ng ml^{-1} FGF2 for 24 h followed by 25 ng ml^{-1} Activin+30 ng ml^{-1} BMP4+1 μM C59 for the next 24 h. Subsequently, day 4 cardiac mesoderm was briefly washed (DMEM/F12) and treated with 30 ng ml^{-1} BMP4+1 μM XAV939+200 $\mu\text{g/ml}$ 2-phospho-ascorbic acid (Sigma) for 48–96 h to yield day 6–8 cardiomyocyte-containing populations. Spontaneously contracting cardiomyocyte foci were evident from day 8 onwards⁸.

Data Records

The raw RNA-seq data (bulk-population and single-cell) and ATAC-seq data can be found at SRA under BioProject PRJNA319573 (accession number SRP073808) [Data Citation 1].

Reproducible peak calls on our ATAC-seq data, as well as transcript per million (TPM) values for each gene and sample in our bulk-population and single-cell RNA-seq data, can be found at GEO under accession number GSE85066 [Data Citation 2]. Bulk-population RNA-seq metadata and mapping statistics can be found in Table 2, while single-cell RNA-seq mapping statistics are in Table 3 (available online only). ATAC-seq metadata can be found in Table 4 (available online only).

For ease of usage, the collated bulk-population RNA-seq data can be viewed at http://cs.stanford.edu/~zhenghao/mesoderm_gene_atlas. As described above, an augmented spreadsheet with TPM values for each gene (for bulk-population RNA-seq data) and additional annotations about whether each gene corresponds to a potential cell surface marker and whether the gene was differentially expressed between conditions can be found on Figshare with DOI 10.6084/m9.figshare.3842835 [Data Citation 3].

Processed surface marker data (a table with the percentage of cells expressing each marker in each cell type) can be found on Figshare with DOI 10.6084/m9.figshare.3505817 [Data Citation 4]. Surface marker screening metadata is in Table 5 (available online only).

The full set of ATAC-seq quality control graphs for all of our samples can be found on Figshare with DOI 10.6084/m9.figshare.3507167 [Data Citation 5].

Technical Validation

Bulk-population RNA-seq

As mentioned above (see Methods), we only analyzed samples with at least 10,000,000 uniquely mapping reads and with at least 50% of reads uniquely mapping. On average, each sample had 45 M uniquely mapping reads with 69% of reads uniquely mapping; full numbers and percentages are in Table 2.

We used FastQC³⁰ to measure the per-base sequence quality for each of our bulk-population RNA-seq experiments. All of the samples passed this quality check (i.e., for each base, the distribution of quality scores had a lower quartile of more than 10 and a median of more than 25). We show a representative FastQC plot (of the first sample we assayed) in Fig. 2a.

We also used principal component analysis (PCA) to visually inspect how the samples were distributed in $\log_2(\text{TPM})$ space. Applying PCA to the 500 genes with highest variance across all samples revealed the presence of batch effects. After correcting for batch effects (see Methods), the PCA plot showed tight clustering (Fig. 2b) among samples and implicitly suggested the developmental trajectory of the cells, starting from human embryonic stem cells in the bottom left and moving upwards towards cardiac mesoderm and somites and their derivatives. An R Markdown script to reproduce Fig. 2b is provided in `bulkDataViz.Rmd` in our Github repository.

Lastly, in our related publication⁸, we independently validated our RNA-seq results by qPCR. Specifically, we conducted qPCR to measure the mRNA expression levels of key genes known to be lineage markers for the various cell types in our study (e.g., *TBX6* and *MSGN1* for paraxial mesoderm; *PARAXIS*, *MEOX1*, and *FOXC2* in the somites). These qPCR expression patterns corroborated our RNA-seq results⁸.

Single-cell RNA-seq

Before sequencing, we used an automated microscope to image each of the cell-capture wells on our Fluidigm C1 chips and manually inspected each image; for subsequent single-cell RNA-seq library construction we only used libraries from wells that contained exactly one cell. After sequencing, we filtered out cells with fewer than 1 million uniquely mapping reads or with fewer than 70% of reads uniquely mapping. Unfortunately, under these stringent selection criteria, all cardiac mesoderm cell RNA-seq libraries were discarded; we ultimately retained 498 single cells out of 651. Full statistics of the cells are provided in Table 3 (available online only).

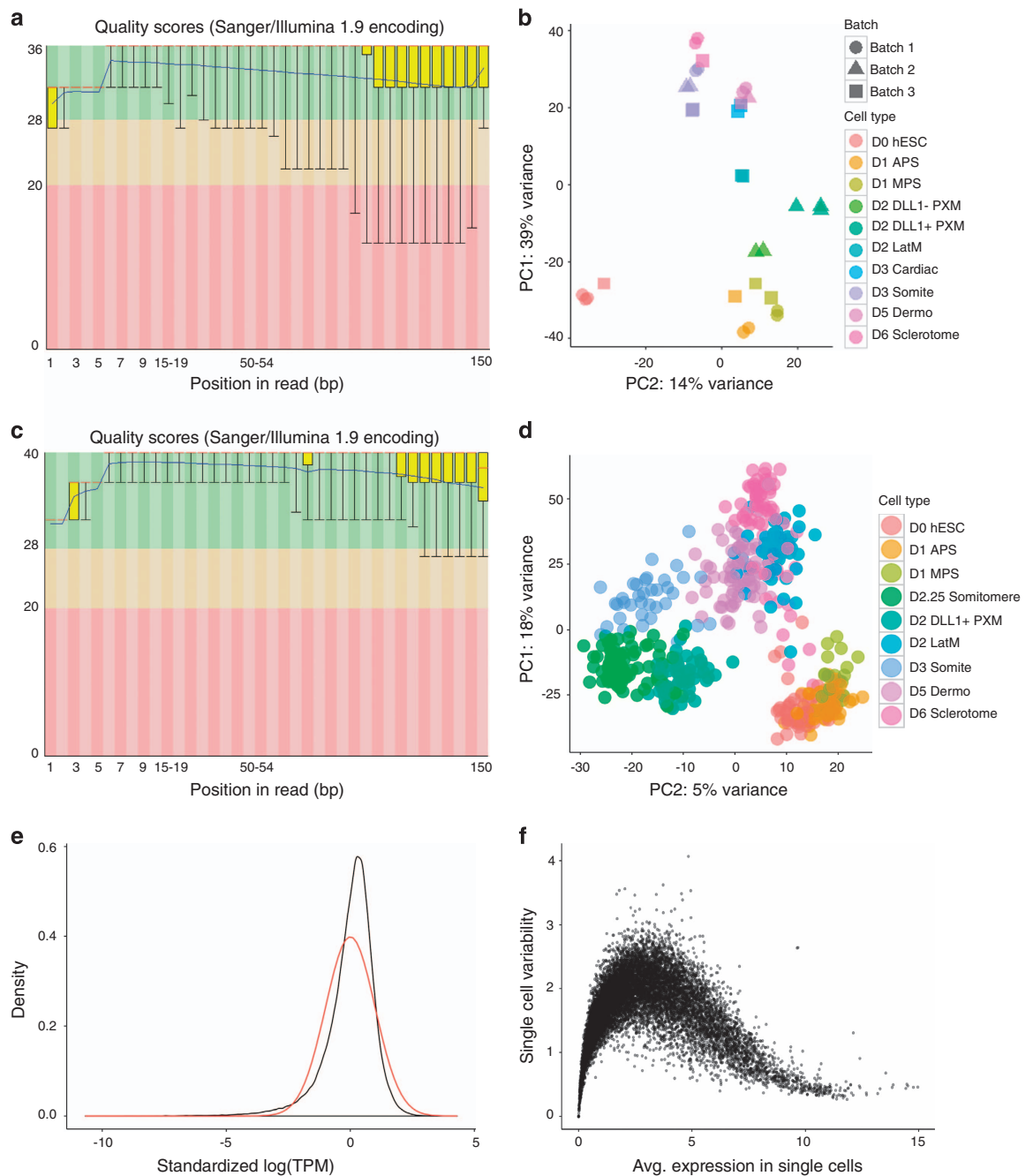


Figure 2. RNA-seq data quality and visualization. (a). Bulk-population RNA-seq FastQC quality scores across read position, shown for a representative sample (D0 hESC). (b). PCA plot of bulk-population RNA-seq data, based on the top 500 genes by variance across all samples, and using log₂ TPM values. (c). Single-cell RNA-seq FastQC quality scores across read position, shown for a representative sample (D2.25 somitomeres). (d). PCA plot of single-cell RNA-seq data, based on the top 500 genes by variance across all cells, and using log₂ TPM values. (e). Black: Plot of density against standardized log₂ TPM for single-cell RNA-seq data across all genes in all cells, after removing zeroes. Red: Fitted normal distribution. (f). Plot of single-cell variability (s.d.) against mean expression value for each gene, shown for a representative cell type (paraxial mesoderm).

As with the bulk-population RNA-seq data, we used FastQC³⁰ to check the per-base sequence quality of each experiment. All of the cells passed this quality check (i.e., for each base, the distribution of quality scores had a lower quartile of more than 10 and a median of more than 25), with a representative FastQC plot in Fig. 2c.

To visualize the distribution of single cells, we once again used PCA on the 500 genes with highest variance (Fig. 2d) in $\log_2(\text{TPM})$ space. As expected, the single-cell RNA-seq libraries separated by cell type, with cell types that are closer to each other biologically (and temporally) tending to cluster together. We note that each cell type was loaded onto a different Fluidigm C1 chip, and due to resource constraints we were only able to use one chip per cell type. This means that cell type is perfectly confounded with chip in our single-cell RNA-seq experiments, and in particular, we cannot tell from the PCA the degree to which batch/chip effects are responsible for the observed separation between cell types.

To tackle this problem, for each cell type, we measured the overall Pearson correlation between the average expression in the single cells and the corresponding average expression in the bulk-population RNA-seq experiments, all in \log_2 TPM units. On average, correlation was 0.82, varying from 0.76 to 0.87 depending on cell type. To ensure that this behavior was not driven solely by housekeeping genes, we looked at key marker genes expressed across our cell types (e.g., *MIXL1* and *BRACHYURY* in primitive streak; *MSGN1* and *DLL3* in paraxial mesoderm; *HAND1* and *FOXF1* in lateral mesoderm; *HOPX* in somitomeres; *FOXC2* and *PAX9* in sclerotome). Single-cell RNA-seq expression patterns of these archetypic marker genes were consistent with independent measures from bulk-population RNA-seq, qPCR, flow cytometry, and immunostaining (data in (ref. 8)).

As technical checks, we also examined the distribution of TPM values across all genes and cells. This followed a roughly log-normal distribution (Fig. 2e) after removing zeros, as expected. Finally, for each cell type, we plotted the standard deviation of each gene against its mean expression value (shown for paraxial mesoderm in Fig. 2f), obtaining for each cell type an expected curve where standard deviation is lowest when average expression is very low (because the expression of the gene in each cell is close to zero) or very high (because high expression translates into a large number of reads, allowing us to reduce technical variation from sampling error).

The script to reproduce Fig. 2d–f is provided `scDataViz.Rmd`. The correlation between average expression in single cells and the bulk population can be analyzed by running `scAverageCorrelation.r`.

ATAC-seq

Through the ATAC-seq pipeline²⁰, we calculated a variety of quality metrics to validate our ATAC-seq data. First, we looked at how many reads remained in each replicate after removing reads that did not successfully align, multi-mapping reads, duplicate reads, and mitochondrial reads. We had two replicates per cell type, and on average, each replicate had 46 M reads remaining, enough to robustly call peaks.

We then looked at the fragment length distribution of the remaining reads; we show a representative plot from lateral mesoderm in Fig. 3a. A ‘good’ ATAC-seq experiment will have a majority of reads falling in the nucleosome-free region (NFR), with a mono-nucleosomal peak representing reads that cut on both sides of a nucleosome (≈ 200 bp in length). All of our samples displayed a mono-nucleosome peak, with 60–70% of reads falling in the NFR.

We also studied the enrichment of reads falling into transcription start sites (TSS), as TSS are known to be open chromatin sites (Fig. 3b; lateral mesoderm). On average, the enrichment of reads at TSS was 10.4x, with a range from 4.6x to 22.3x.

Next, we looked at the number of peaks called across each replicate. Because of the variability in quality across the experiments (e.g., some experiments had a higher TSS enrichment and/or more reads), we first subsampled each replicate to have a maximum of 35 M reads (post-filtering). We then used MACS2 (ref. 23) to call peaks on each replicate independently, before using an IDR analysis²⁴ to identify peaks that were reproducible between the two replicates for each cell type (Fig. 3c). Using an IDR threshold of 0.1, we found an average of 91 K reproducible peaks per cell type.

Full statistics and metadata for each replicate is provided in Table 4 (available online only), including additional quality metrics such as library complexity metrics, the fraction of NFR to mono-nucleosome reads, and the number of reads falling in universal DNase-I hypersensitive regions, promoter regions, enhancer regions, and called peak regions. To compute these, we used putative promoter, enhancer, and DHS annotations from 127 cell types and tissues from the Roadmap Epigenomics Project. These annotations are provided in the flagship Roadmap Epigenomics Project publication³¹ and are available from the supplementary website <http://compbio.mit.edu/roadmap> in the ‘DNase-I accessible regulatory regions’ section. In brief, DNase-seq based chromatin accessible regions were labeled as promoter or enhancer based on chromatin state maps learned using 5 core histone modifications across the 127 cell types and tissues.

The graphs in Fig. 3 were taken from the output of the ATAC-seq pipeline for one representative sample (lateral mesoderm). The full set of graphs for all of our samples can be found in [Data Citation 5].

High-throughput surface marker screening

For validation, we focused on surface markers with lineage-specific expression (Fig. 4), and chose two surface markers for in-depth *in vivo* and *in vitro* validation: *DLL1* (a marker of paraxial mesoderm) and *GARP* (a marker of cardiac mesoderm). *In situ* hybridization of zebrafish homologs of these genes (*deltaC*, the homolog of human *DLL1* and *lrrc32*, the homolog of human *GARP*) was conducted in

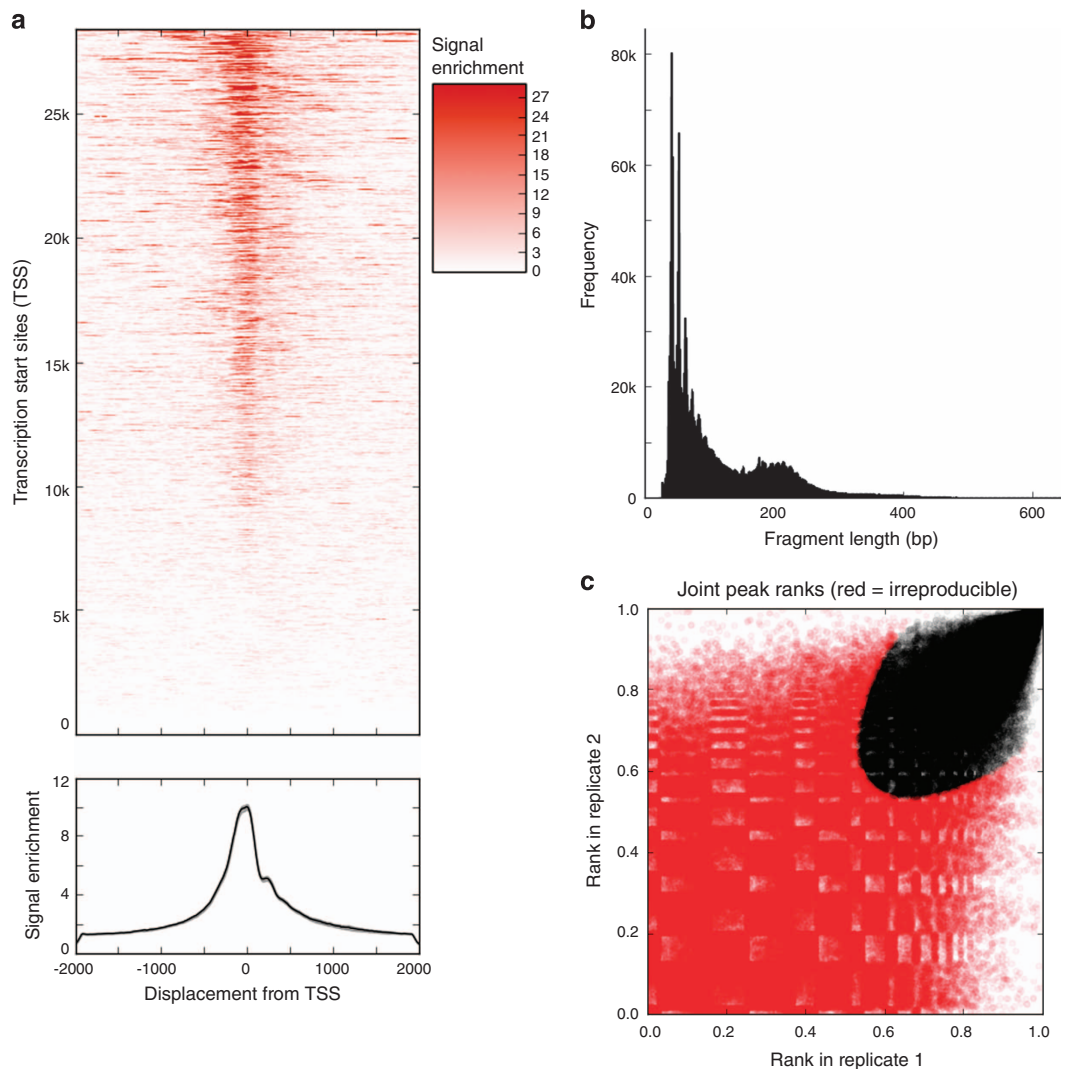


Figure 3. ATAC-seq data quality metrics. (a). Enrichment of ATAC-seq signal around transcription start sites (TSS), shown for a representative sample (lateral mesoderm). Top: enrichment around individual TSS. Bottom: aggregated enrichment around all TSS's. (b). Fragment length distribution of ATAC-seq reads from a representative sample (lateral mesoderm). Most of the reads fall into the nucleosome-free region (< 150 bp) and a clear mono-nucleosome peak can be seen. (c). Irreproducible rate (IDR) analysis of ATAC-seq peaks from lateral mesoderm. The scatter plot shows one point for every peak, with its location representing in rank in each replicate. For downstream analysis, we only consider peaks shown in black (reproducible at an IDR rate of 0.1), which have ranks that are consistent between replicates.

zebrafish embryos, which revealed fairly specific expression of *deltaC* in paraxial mesoderm and that of *lrrc32* in the developing heart tube *in vivo*⁸. Additionally, fluorescence-activated cell sorting (FACS) of DLL1+ cells from hESC-derived day 2 paraxial mesoderm cultures followed by bulk-population and single-cell RNA-seq revealed that all DLL1+ cells essentially expressed paraxial mesoderm transcription factors at the single-cell level⁸. Collectively, this reaffirmed that DLL1 and GARP respectively mark human paraxial and cardiac mesoderm.

Differentiation

Our related publication⁸ focused on establishing the identity and function of the derived cell types, and we refer readers interested in those details to that manuscript. In brief, we verified cellular function through *in vivo* transplantation experiments and we assessed cellular identity and purity through molecular analyses of marker expression (RNA-seq, ATAC-seq, immunostaining, and flow cytometry).

On the molecular side, for each cell type we identified archetypic genes and surface markers based on biological knowledge and prior literature. We confirmed that the key genes were expressed at a population level through bulk RNA-seq and qPCR; then, through single-cell RNA-seq, immunostaining,

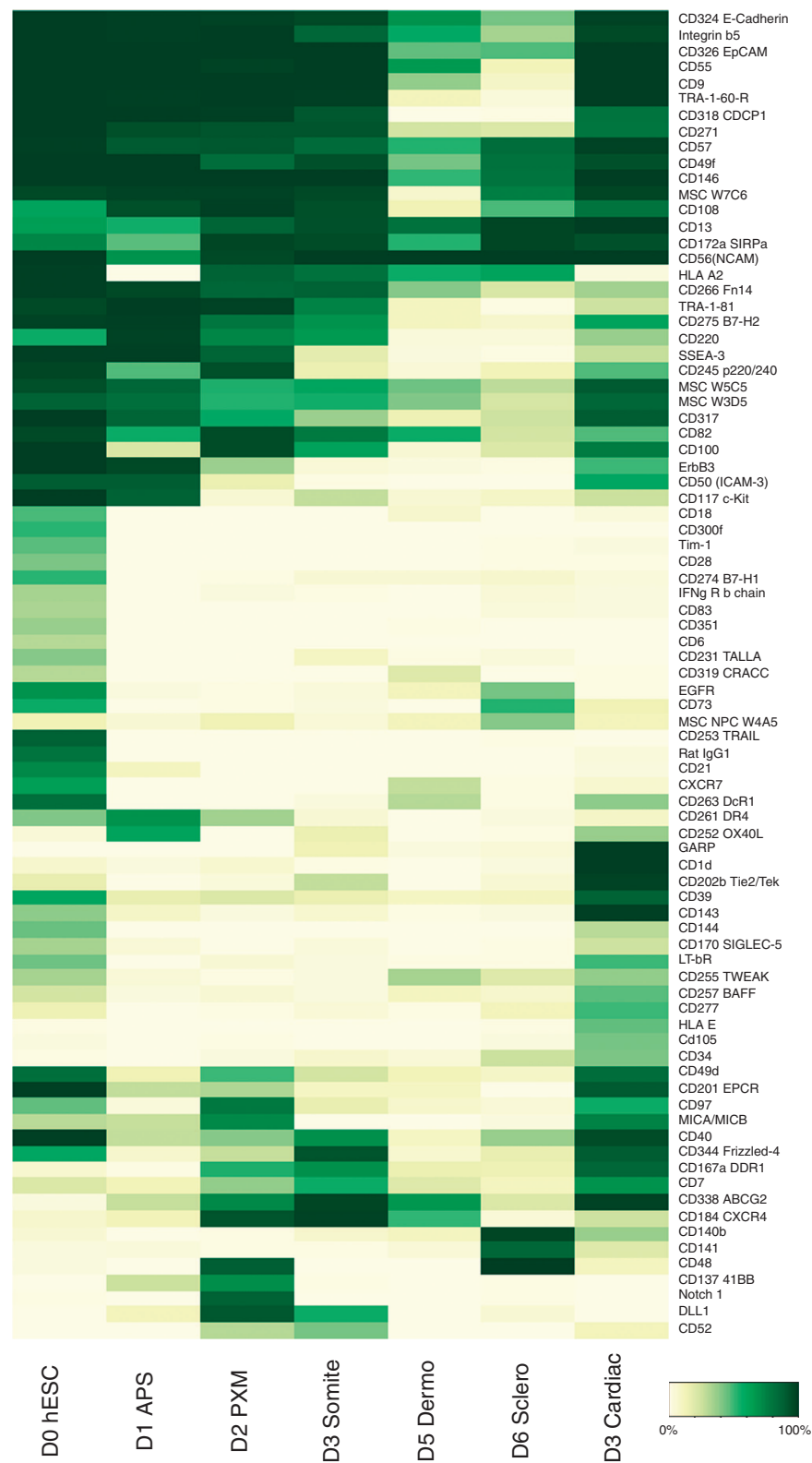


Figure 4. High-throughput surface marker screening. We show here a heatmap of all surface markers whose expression varied considerably across cell types, filtering out markers where less than 30% or more than 70% of cells across all cell types expressed the marker. The % refers to the percentage of cells of a given type that expressed a marker.

or flow cytometry, we verified that the population was suitably homogeneous for those genes and surface markers⁸. On the basis of those metrics, the cell populations we derived were generally between 80 and 99% pure. Motif enrichment analysis of the open chromatin regions in each cell type (as measured by ATAC-seq) also yielded results consistent with their cellular identity, e.g., GATA motifs were significantly enriched in lateral and cardiac mesoderm.

We conducted transplantation experiments in immunodeficient mice to further verify the function of two human mesodermal cell-types derived from our differentiation process, namely sclerotome and cardiac mesoderm⁸. Sclerotome cells subcutaneously injected into immunodeficient mice self-organized to form an ectopic human bone, undergoing ossification, displaying spatial structure expected of human bone, and even attracting and becoming vascularized by mouse blood vessels. For the cardiac mesoderm, we first further differentiated them into cardiomyocytes through WNT blockade and BMP inhibition for four days and engineered them to express a constitutively-expressed *luciferase* reporter gene. To further test the functionality of these ESC-derived human cardiomyocytes, we developed an experimental system wherein ventricular fragments from week 15–17 human fetal heart³² were subcutaneously implanted into the mouse ear. We then transplanted ESC-derived cardiomyocytes directly into the human fetal heart graft and found that they engrafted the human heart tissue for at least 10 weeks, as measured by bioluminescence imaging of *luciferase*-expressing cardiomyocytes *in vivo*.

Usage Notes

Researchers studying the single-cell RNA-seq data reported herein should note that inferences made from global comparisons (e.g., PCA or clustering) may be limited by experimental design, as each individual cell-type was processed on a separate Fluidigm C1 chip. Hence when comparing single-cell RNA-seq data from different cell-types it is difficult to account for batch effects arising from different chips. Our analysis, including comparisons of key lineage marker genes known to vary between distinct cell lineages, shows that the data are still valid. However, care should be taken in global comparisons that involve aggregating large numbers of genes, as the noise from batch effects could be substantial in that context; the bulk-population RNA-seq data could be used to verify results from such comparisons.

In our related publication⁸, we applied principal component analysis to the single-cell RNA-seq data to reconstruct the differentiation trajectory of paraxial mesoderm, somitomeres, and early somites in ‘pseudotime’, a concept first introduced in the context of single-cell RNA-seq by other groups (ref. 10 and ref. 33). Interested readers might want to apply these, and other more sophisticated trajectory reconstruction methods, to our single-cell data.

The variance in data quality across the ATAC-seq experiments, due to technical reasons (e.g., different numbers of starting reads or varying cell lysis) and biological reasons (e.g., distinct cell types may have different amounts of open chromatin), mean that care must also be taken when conducting global comparisons of ATAC-seq data. We found that rank-normalization (or, at one extreme, binarization) makes it easier to compare ATAC-seq data across cell-types, as opposed to using P-values, local IDR values, or measures of signal intensity to score each peak. Sub-sampling reads before peak-calling, as we did in our analysis, should only be done for global comparisons; researchers who are doing an in-depth study of one cell type should use all available reads that pass the filtering criteria for maximal information.

We are currently using this dataset to study the temporal changes in alternative splicing and long non-coding RNA expression as differentiation progresses. In addition, we are actively exploring the expression of repeat elements, including dormant retrotransposons, interspersed nuclear elements, Alu elements, and human endogenous retrovirus elements that may have a role in early human embryonic development. Readers who are interested in similar questions are welcome to contact us to discuss methods and collaborations.

References

1. Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113**, 891–911 (1991).
2. Rosenquist, G. C. Location and movements of cardiogenic cells in the chick embryo: the heart-forming portion of the primitive streak. *Developmental Biology* **22**, 461–475 (1970).
3. Tam, P. P. & Beddington, R. S. The formation of mesodermal tissues in the mouse embryo during gastrulation and early organogenesis. *Development* **99**, 109–126 (1987).
4. Pourquié, O. Vertebrate segmentation: from cyclic gene networks to scoliosis. *Cell* **145**, 650–663 (2011).
5. Christ, B. & Scaal, M. *Formation and differentiation of avian somite derivatives*, vol. 638 of *Adv Exp Med Biol* 1–41 (Landes Bioscience, 2008).
6. Tanaka, M. Molecular and evolutionary basis of limb field specification and limb initiation. *Dev. Growth Differ* **55**, 149–163 (2013).
7. Später, D., Hansson, E. M., Zangi, L. & Chien, K. R. How to make a cardiomyocyte. *Development* **141**, 4418–4431 (2014).
8. Loh, K. M. *et al.* Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* **166**, 451–467 (2016).
9. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).

10. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotech.* **32**, 381–386 (2014).
11. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
12. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 1–12 (2014).
13. ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**, 57–74 (2012).
14. Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
15. Li, B. & Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
16. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics (Oxford, England)* **8**, 118–127 (2007).
17. Leek, J. *et al.* sva: Surrogate variable analysis. *r* package version 3.18.0 (2015).
18. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
19. Tzur, A., Moore, J. K., Jorgensen, P., Shapiro, H. M. & Kirschner, M. W. Optimizing optical flow cytometry for cell volume-based sorting and analysis. *PLoS ONE* **6**, e16053 (2011).
20. Kundaje Lab. Atac-seq processing pipeline. https://github.com/kundajelab/atac_dnase_pipelines (2016).
21. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nature Methods* **9**, 357–359 (2012).
22. Broad Institute. Picard Tools. <https://broadinstitute.github.io/picard/> (2016).
23. Zhang, Y. *et al.* Model-based analysis of chip-seq (macs). *Genome Biol.* **9**, R137 (2008).
24. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**, 1752–1779 (2011).
25. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from encode data. *Nature* **488**, 91–100 (2012).
26. Quinlan, A. R. & Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
27. Panchision, D. M. *et al.* Optimized flow cytometric analysis of central nervous system tissue reveals novel functional relationships among cells expressing cd133, cd15, and cd24. *Stem Cells* **25**, 1560–1570 (2007).
28. Loh, K. M. *et al.* Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell* **14**, 237–252 (2014).
29. Brunton, S. A. *et al.* Potent agonists of the hedgehog signaling pathway. *Bioorganic & Medicinal Chemistry Letters* **19**, 4308–4311 (2009).
30. Andrews, S. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).
31. Roadmap Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
32. Ardehali, R. *et al.* Prospective isolation of human embryonic stem cell-derived cardiovascular progenitors that integrate into human fetal heart tissue. *Proc Natl Acad Sci USA* **110**, 3405–3410 (2013).
33. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).

Data Citations

1. NCBI Sequence Read Archive, SRP073808 (2016).
2. Gene Expression Omnibus, GSE85066 (2016).
3. Koh, P. W. *et al.* Figshare <https://dx.doi.org/10.6084/m9.figshare.3842835> (2016).
4. Loh, K. *et al.* Figshare <https://dx.doi.org/10.6084/m9.figshare.3505817.v3> (2016).
5. Koh, P. W. *et al.* Figshare <https://dx.doi.org/10.6084/m9.figshare.3507167.v2> (2016).

Acknowledgements

We thank Jin-Wook Lee, Nathan Boley, and Daniel Kim for assistance with ATAC-seq analyses, Norma Neff and Olive Curreri for assistance with single-cell RNA-seq, Cole Trapnell and Dana Pe'er for advice on the computational pipeline for single-cell RNA-seq, and the Stanford Functional Genomics and Stem Cell Genomics Core Facilities for infrastructure support. K.M.L. was supported as a Siebel Investigator of the Siebel Stem Cell Institute with additional support from the Fannie and John Hertz Foundation, the U.S. National Science Foundation, and the Davidson Institute for Talent Development. A.K. was supported by the Sloan Foundation Research Fellowship. This study was also supported by the California Institute for Regenerative Medicine (RT2-02060, RT3-07683, TB1-01195 and a GC1R-06673-A Collaborative Research Program) and the U.S. National Institutes of Health (T32GM007365).

Author Contributions

P.W.K. developed the computational pipeline, analyzed the data, and wrote the first draft of the manuscript. R.S. developed the RNA-seq pipeline and constructed single-cell and bulk-population RNA-seq libraries and performed the critical reading of the manuscript. A.A.B. prepared ATAC-seq libraries. R.M.M. constructed single-cell and bulk-population RNA-seq libraries under the supervision of R.S. A.C. developed the mesoderm differentiation system, purified cells for transcriptional and chromatin analyses, and conducted surface-marker screens. I.L.W. supervised mesoderm differentiation studies. L.T.A. developed the mesoderm differentiation system and supervised cell differentiation studies. A.K. supervised the development of the computational pipeline and the data analysis. K.M.L. developed the mesoderm differentiation system, purified cells for transcriptional and chromatin analyses, conducted surface-marker screens, and wrote the first draft of the manuscript.

Additional Information

Tables 1,3,4 and 5 are only available in the online version of this paper.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Koh, P. W. *et al.* An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci. Data* 3:160109 doi: 10.1038/sdata.2016.109 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

Metadata associated with this Data Descriptor is available at <http://www.nature.com/sdata/> and is released under the CC0 waiver to maximize reuse.

© The Author(s) 2016