LETTER

The *cis*-regulatory dynamics of embryonic development at single-cell resolution

Darren A. Cusanovich[1]*, James P. Reddington[2]*, David A. Garfield[2]†*, Riza M. Daza[1], Delasa Aghamirzaie[1], Raquel Marco-Ferreres[2], Hannah A. Pliner[1], Lena Christiansen[3], Xiaojie Qiu[1], Frank J. Steemers[3], Cole Trapnell[1], Jay Shendure[1,4]§ & Eileen E. M. Furlong[2]§

Understanding how gene regulatory networks control the progressive restriction of cell fates is a long-standing challenge. Recent advances in measuring gene expression in single cells are providing new insights into lineage commitment. However, the regulatory events underlying these changes remain unclear. Here we investigate the dynamics of chromatin regulatory landscapes during embryogenesis at single-cell resolution. Using single-cell combinatorial indexing assay for transposase accessible chromatin with sequencing (sci-ATAC-seq)[1], we profiled chromatin accessibility in over 20,000 single nuclei from fixed *Drosophila melanogaster* embryos spanning three landmark embryonic stages: 2–4 h after egg laying (predominantly stage 5 blastoderm nuclei), when each embryo comprises around 6,000 multipotent cells; 6–8 h after egg laying (predominantly stage 10–11), to capture a midpoint in embryonic development when major lineages in the mesoderm and ectoderm are specified; and 10–12 h after egg laying (predominantly stage 13), when each of the embryo's more than 20,000 cells are undergoing terminal differentiation. Our results show that there is spatial heterogeneity in the accessibility of the regulatory genome before gastrulation, a feature that aligns with future cell fate, and that nuclei can be temporally ordered along developmental trajectories. During mid-embryogenesis, tissue granularity emerges such that individual cell types can be inferred by their chromatin accessibility while maintaining a signature of their germ layer of origin. Analysis of the data reveals overlapping usage of regulatory elements between cells of the endoderm and non-myogenic mesoderm, suggesting a common developmental program that is reminiscent of the mesendoderm lineage in other species[2–4]. We identify 30,075 distal regulatory elements that exhibit tissue-specific accessibility. We validated the germ-layer specificity of a subset of these predicted enhancers in transgenic embryos, achieving an accuracy of 90%. Overall, our results demonstrate the power of shotgun single-cell profiling of embryos to resolve dynamic changes in the chromatin landscape during development, and to uncover the *cis*-regulatory programs of metazoan germ layers and cell types.

We adapted our sci-ATAC-seq protocol[1] to work with nuclei from formaldehyde-fixed *Drosophila* embryos and concurrently implemented optimizations to increase the sensitivity by roughly an order of magnitude. The nuclei processed from each developmental time point were derived from hundreds of embryos of both sexes, and were therefore likely to include intermediate developmental states. Of 431 million sequenced read pairs, 70% mapped to the nuclear reference genome and were assigned a cell barcode (Extended Data Fig. 1a, b). Altogether, we obtained chromatin accessibility profiles for 23,085 cells across the three time points ($12,904 \pm 10,979$ (mean $\pm$ s.d.) reads per cell after

de-duplication; minimum of 500 unique reads per cell (Extended Data Fig. 1c)). Sequenced fragments exhibited nucleosomal banding and were strongly enriched in DNase-hypersensitive sites (DHS) that have been defined in bulk *Drosophila* embryos[5] (Extended Data Fig. 1d).

We partitioned the genome into 2-kb windows and scored each cell by whether any reads were observed in each window. For each time point, we performed latent semantic indexing[1] (LSI) using the 20,000 most frequently accessible windows and discarding the 10% of cells with the fewest reads. Of the 20,000 windows, 14,295 were common across all three time points (Extended Data Fig. 1e). Although measurements of accessibility in individual cells are naturally sparse (as there are only 2–4 genome equivalents per nucleus), the data are sufficiently structured to reveal subsets of cells with similar chromatin accessibility (Fig. 1a–c). To map the underlying regulatory elements, we aggregated data from cells within each of the largest 4–5 clades per time point to call peaks and summits of accessibility for each 'in silico-sorted' clade (Fig. 1d). Merging summits across all time points and clades identified 53,133 potential *cis*-regulatory elements, 40,967 of which have clade-specific accessibility in at least one time point (Supplementary Table 1); including 12,605 at 2–4 h, 25,615 at 6–8 h and 28,253 at 10–12 h after egg laying (Extended Data Fig. 1f). These results reveal the highly dynamic and heterogeneous nature of chromatin accessibility during embryogenesis, with roughly twice as many differentially accessible sites identified at the later time points compared with the earlier one.

To determine the identity of each cell clade, we compared accessible regions to 3,841 developmental enhancers[6–8] and 9,356 gene promoters[9,10] with characterized tissue activity across embryogenesis. The enrichments of clade-specific promoter-distal (putative enhancers) and promoter-proximal (putative promoters) elements gave consistent results (Supplementary Table 2). The four major clades at 6–8 h and 10–12 h correspond to the three major germ layers, with two subdivisions: ectoderm, which is split into neurogenic (clade 1) and non-neurogenic (clade 2) lineages, and mesoderm, which is split into myogenic mesoderm (clade 3) and non-myogenic mesoderm (such as fat body and haemocytes) combined with endoderm (clade 4) (Extended Data Fig. 2, Supplementary Table 2). The latter indicates that non-myogenic mesoderm and endoderm exhibit similar chromatin accessibility, suggesting a shared developmental program. Although, to our knowledge, *Drosophila* mesoderm and endoderm have not been shown to share a common origin, this is highly reminiscent of the mesendoderm lineage in *Caenorhabditis elegans*[2], sea urchins[3] and vertebrates[4]. Of the 53,133 potential *cis*-regulatory elements, 35,963 are distal (putative enhancers); 12% overlap characterized developmental enhancers and 48% overlap putative enhancers identified from bulk DHS data[5] (based on 1-bp overlap). Conversely, of the

---

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. [3]Illumina, San Diego, California, USA. [4]Howard Hughes Medical Institute, Seattle, Washington, USA. †Present address: IRI Life Sciences, Humboldt Universität zu Berlin, Berlin, Germany.
*These authors contributed equally to this work.
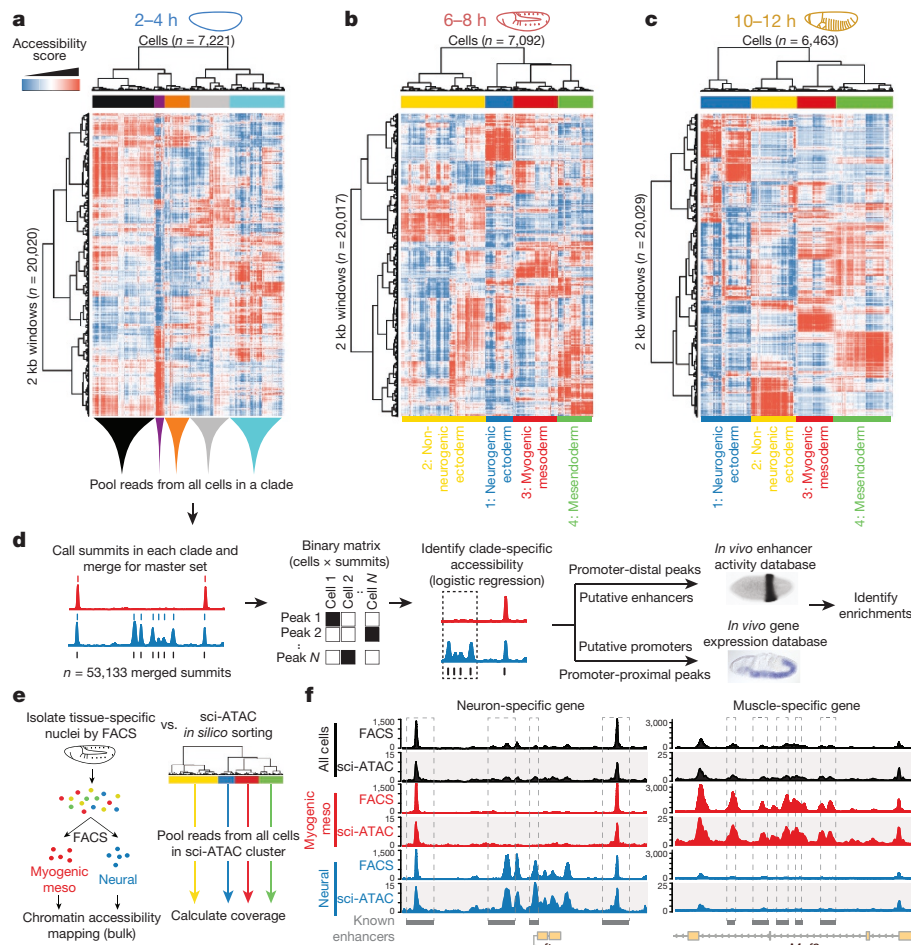§These authors jointly supervised this work.

Figure 1 | Single-cell profiling of chromatin accessibility across *Drosophila* embryogenesis. a–c, Heat maps of binarized, LSI-transformed, clustered read counts for single cells (columns) in 2-kb windows across the genome (rows) at 2–4 h (**a**), 6–8 h (**b**) and 10–12 h (**c**) after egg laying. Major clades are assignable to germ layers at post-gastrulation time points (**b**, **c**). **d**, Approach to annotation of clades by intersecting clade-specific peaks of chromatin accessibility with enhancer activity and gene expression. *In situ* image of enhancer activity (black stain) from ref. 7; RNA *in situ* (blue stain) from the Berkeley *Drosophila* Genome Project[10,31,32]. **e**, Comparing fluorescence-activated cell sorting combined with DNase I sequencing (FACS–DNase-seq) and *in silico*

sorting with sci-ATAC-seq. Nuclei from myogenic mesoderm and neurons were isolated from 6–8-h embryos using antibodies against tissue-specific regulatory proteins Mef2 (myogenic mesoderm) and Elav (neurons), sorted by FACS and analysed by DNase-seq. *In silico* sorts from sci-ATAC-seq were built by pooling reads from all cells within each LSI-defined clade. **f**, Library-size-normalized coverage tracks from FACS–DNase-seq (top graph for each clade) and sci-ATAC-seq *in silico* sorts (bottom graph for each clade) for whole embryo (black), mesodermal (red), and neuronal (blue) at 6–8 h. Shown are *ftz* (neuronal; left) and *Mef2* (mesodermal; right) loci. Known enhancers for each tissue are indicated.

3,841 characterized developmental enhancers, 2,533 (66%) overlapped regions of accessible chromatin identified in this study.

To validate *in silico* sorting and clade assignments, we used fluorescence-activated cell sorting (FACS) to isolate myogenic mesoderm and neuronal nuclei from 6–8-h embryos[11] to approximately 98% purity. Sorted nuclei were subjected to DNase I-hypersensitive-site sequencing (DNase-seq) in bulk, and the resulting accessibility maps were compared to our *in silico*-sorted (that is, clade-defined) sci-ATAC-seq data from 6–8-h embryos (Fig. 1e). The comparison shows notable similarity both globally (Spearman's $\rho > 0.85$ for matched versus 0.53 for non-matched comparisons) and at individual loci. For example, both methods show that previously characterized neuronal enhancers near the *ftz* gene are accessible in neurogenic ectoderm but not in myogenic mesoderm (Fig. 1f, left) and, conversely, that muscle enhancers of *Mef2* are accessible in myogenic mesoderm but not in neurogenic ectoderm (Fig. 1f, right).

The clade assignments are further supported by motif enrichments for transcription factor binding sites and transcription factor occupancy at putative enhancers. For example, at mid and late embryogenesis, motifs for the lineage-specifying factors Krüppel (Kr), tramtrack (Ttk) and runt (Run) were among the most enriched in neurogenic

ectoderm[12] (clade 1), Mef2 and Cf2 motifs were enriched in myogenic mesoderm[13] (clade 3) and GATA motifs were enriched in mesendoderm (clade 4) (Extended Data Fig. 3a–c, Supplementary Table 3). The presence of GATA motifs may reflect the conserved role of GATA factors in the specification of both non-myogenic mesoderm[14] and endoderm[15]. Similarly, regions occupied by transcription factors with more constitutive roles, such as CTCF, exhibit similar accessibility across all clades (Extended Data Fig. 3d–g), whereas regions bound by myogenic transcription factors are more accessible in the myogenic mesodermal clade[16] (Extended Data Fig. 3h–l).

Cells examined at 2–4 h after egg laying fall into five major clades (Fig. 1a) in which regulatory identities are clearly distinct from later stages in embryogenesis (Extended Data Fig. 4, Supplementary Table 2). The 2–4-h nuclei span embryos from the syncytial blastoderm, cellularization, gastrulation and early germ-band extension (stages 5–8), with the majority of embryos being pre-gastrulation (stage 5). Developmental transitions during these stages are very rapid, with cellularization (stage 5) lasting 40 min and onset of gastrulation (stage 6) lasting only 10 min. To capture finer granularity across these dynamic transitions, we applied *t*-distributed stochastic neighbour embedding (*t*-SNE)[17] to the binary sci-ATAC-seq matrix of cells versus summits of
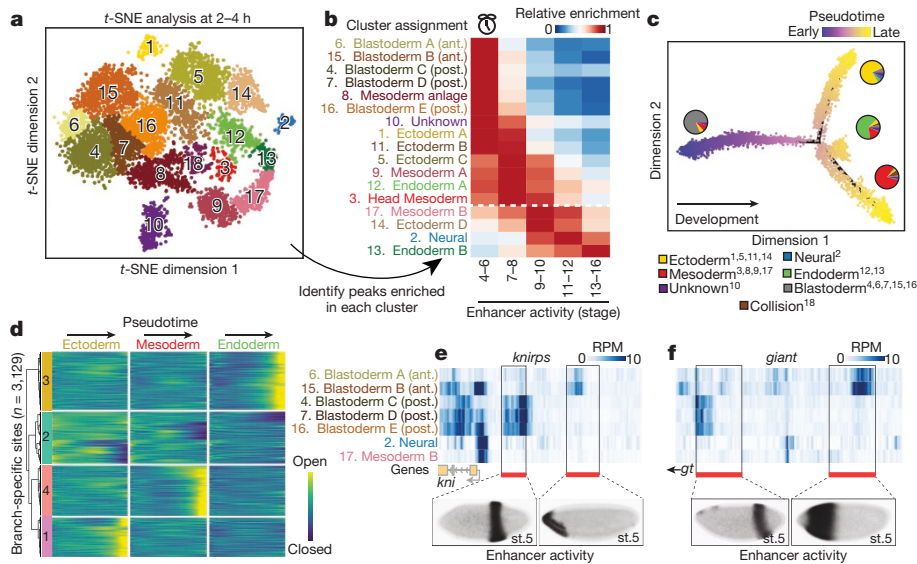
Figure 2 | Temporal dynamics and spatial heterogeneity in chromatin accessibility in the early embryo. a, *t*-SNE analysis of cells at 2–4 h. Clusters were defined by a density peak clustering algorithm (see Methods) and annotated on the basis of overlaps between cluster-enriched peaks and known tissue-specific enhancers or genes. **b**, Relative enrichment of enhancers that are active at different developmental stages in each cluster. Clusters below the white dashed line are likely to be derived from embryos outside the 2–4-h window, owing to female holding of older embryos. Ant., anterior; post., posterior. **c**, Pseudotime ordering of cells along a developmental trajectory. Cells were ordered in three dimensions (only two are shown) with DDRTree. Point colours correspond to cells'

progression along the trajectory. Pie charts indicate relative frequencies of germ-layer assignments for cells in each branch. Superscript numbers in the key indicate which clusters from **a** were included in each category. **d**, Heat map of smoothed accessibility curves fit to sites (rows) for 100 bins of cells progressing through pseudotime (columns). Sites were clustered into four groups on the basis of their temporal dynamics. Only sites classified as branch-specific are shown. **e, f**, Heat maps of library-size-normalized read counts in the vicinity of the gap genes *knirps* (**e**) and *giant* (**f**). In each case, one characterized enhancer is known to drive anterior expression and the other drives posterior expression in blastoderm embryos (stage 5). *In situ* images of enhancer activity obtained from ref. 7.

accessibility. Because of confounding differences in sex chromosome copy number between male and female nuclei (Extended Data Fig. 5), we restricted the matrix to autosomal elements.

Density-peak clustering[18] of cells after *t*-SNE enabled identification of 18 cell clusters at 2–4 h (Fig. 2a). Analysis of the relative enrichment of these clusters for active enhancers and transcription-factor occupancy (Supplementary Tables 4, 5) revealed marked differences in their developmental stages (Fig. 2b), highlighting developmental time as a major axis of variation within this time point. Notably, two of the developmentally early clusters were sex-biased (cluster 10: 85% male; cluster 1: 69% female). Whereas the identity of the male-biased cluster remains unclear, the female-biased cluster is enriched for enhancers that are active in brain anlage.

To evaluate this temporal ordering more formally, we used a graph-based method to arrange single cells into a developmental trajectory[19]. This 'pseudotemporal' ordering agreed well with the observed enrichments in cell clusters for active enhancers (Extended Data Fig. 6a–c). Notably, the trajectory split cells into three major branches that were consistent with our annotations of the major germ layers (neuronal cells are rare at this time point, as expected) (Fig. 2c). Pseudotemporal ordering also enabled us to explore the dynamics of sites that open or close within the 2–4-h window. We identified 12,165 sites with significant pseudotime-dependent temporal changes (1% false discovery rate (FDR)). Using a simple heuristic, we classified 5,219 (43%) of these sites as closing as pseudotime progressed; 5,133 (42%) as opening; and the remaining 1,813 (15%) as having more complex dynamics (Extended Data Fig. 6d–i, Supplementary Table 6). Many of the most pronounced changes match expectations, falling within gene loci that have dynamic roles during early embryogenesis. For example, the most significant closing site (*P* value $= 5 \times 10^{-224}$) is within the *slam* locus, a gene that is essential for blastoderm cellularization during a very brief temporal window[20] (Extended Data Fig. 6g).

To identify sites that open or close specifically within individual germ-layer trajectories, we tested for pseudotime-dependent changes along each of the three paths (Fig. 2c) independently (with the potential

caveat that these branches may be contaminated to some degree by cells from older embryos, owing to female 'holding'). This test identified 3,129 sites that were significantly pseudotime-dependent in only one branch, with 992, 1,071, and 1,066 restricted to the ectoderm, mesoderm and endoderm, respectively (Fig. 2d, Supplementary Tables 7–10). As with the global pseudotime ordering, sites associated with lineage-specific pseudotime exhibited dynamics consistent with biological expectation (for example, sites in the *heartless* (*htl*)[21], *GATAe*[22], and *dachsous* (*ds*)[23] loci are accessible specifically in mesoderm, endoderm and ectoderm, respectively; Extended Data Fig. 6j–l).

Therefore, germ layers appear late in pseudotime at 2–4 h (Fig. 2c), yet developmentally early nuclei in this same window (as defined in Fig. 2b; clusters 6, 15, 4, 7, 8, 16) exhibit heterogeneous chromatin accessibility that reflects enhancer activity in refined spatial domains along the embryo's antero-posterior (A–P) and dorso-ventral (D–V) axes (Supplementary Table 5). For example, chromatin accessibility surrounding two gap genes, *knirps* (*kni*) and *giant* (*gt*), varies among developmentally early clusters (Fig. 2e, f). The expression of *knirps* and *giant* is spatially patterned in two broad stripes along the A–P axis of the embryo, each controlled by two enhancers driving either the posterior or the anterior expression[7]. The anterior enhancers of both genes have greater accessibility in cells of the presumptive anterior blastoderm clusters (clusters 6 and 15), while the posterior enhancers exhibit greater accessibility in the presumptive posterior blastoderm clusters (clusters 4, 7, and 16) (Fig. 2e, f). This example illustrates how despite being untargeted, sci-ATAC-seq can identify regulatory regions that are specifically accessible in spatially refined subsets of cells without the need for FACS sorting. Classic lineage-tracing and transplantation experiments showed that the broad fate and developmental potential of cells are largely determined at the cellular blastoderm stage, leading to the concept of a blastoderm fate map[24]. Our data support the view that these early pre-gastrulation cell specification events are underpinned by spatial heterogeneity in chromatin accessibility.

Applying *t*-SNE to the later time points, during lineage commitment (6–8 h) and differentiation (10–12 h), revealed a fine-grained map of
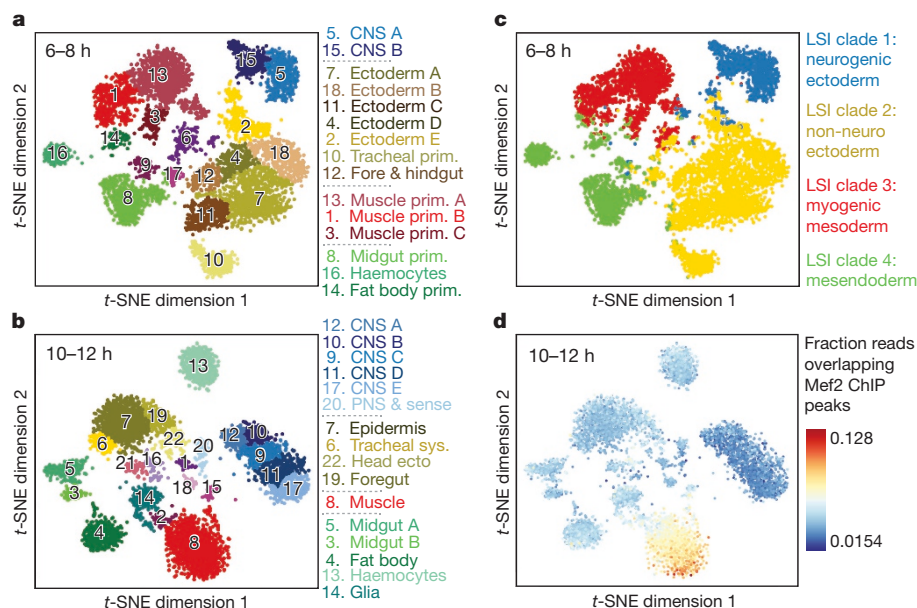
Figure 3 | Single cells are readily assigned to tissues and cell types based on chromatin accessibility. a, b, Clustering of sci-ATAC-seq data from the 6–8-h (**a**) and 10–12-h (**b**) time points after *t*-SNE dimensionality reduction. Clusters were annotated based on overlaps between cluster-enriched peaks and enhancers or genes with known tissue-specific activity. Three 6–8-h (6, 9, 17) and six 10–12-h (1, 2, 15, 16, 18, 21) clusters are likely to comprise multi-cell collisions based on library complexity and the distribution of reads mapping to the X chromosome (Extended Data Fig. 7). **c,** The 6–8-h *t*-SNE shown in **a**, coloured according to the original germ-layer assignment. **d,** The 10–12-h *t*-SNE shown in **b**, coloured according to the fraction of reads falling in Mef2 ChIP–seq peaks.

cell clusters that could be readily assigned to specific tissues or cell types (Fig. 3a, b; Supplementary Table 4). A few small clusters were identified as likely 'collisions' resulting from the combinatorial indexing, and were therefore discarded (purple clusters in Fig. 3a, b, Extended Data Fig. 7). For all remaining clusters, the cell-type assignments are broadly consistent with the germ-layer clade assignments (Fig. 3c, Extended Data Fig. 8), but with much finer granularity, whether we use information from either enhancer or gene-activity databases (Extended Data Fig. 9). For example, mesendoderm (clade 4 in Figs 1, 3c) is resolved into three separate clusters at 6–8 h, comprising the fat body (cluster 14) and haemocytes (cluster 16) from the non-myogenic mesoderm, and midgut (cluster 8) from the endoderm (Fig. 3a). Although we are clearly undersampling the number of cells present at these stages, the data are not obviously biased towards any particular tissue or cell type. The clusters' tissue identities also match transcription factor occupancy by tissue-specific factors (Supplementary Table 4). For example, cells in cluster 8 (muscle) at 10–12 h are enriched for reads that overlap chromatin immunoprecipitation (ChIP) peaks for the key myogenic factor Mef2 at 10–12 h (Fig. 3d).

A major advantage of profiling chromatin accessibility is its potential to identify distal regulatory elements that shape gene expression. To determine whether elements that exhibit tissue-specific chromatin accessibility corresponded to bona fide tissue-specific enhancers, we tested 31 elements in transgenic embryos. We selected promoter-distal elements exhibiting clade-specific accessibility at 6–8 h and/or 10–12 h that did not overlap with previously characterized enhancers (Supplementary Table 11). No other criteria were used to bias the selection towards different classes of distal regulation (for example, enhancers versus insulators). Each putative regulatory element was cloned upstream of a minimal promoter driving a *lacZ* reporter and stably integrated into a common location in the *Drosophila* genome to minimize positional effects. Enhancer activity was then assessed across all stages of embryogenesis by *in situ* hybridization.

Notably, given the simple selection strategy, 94% (29 of 31) of tested regions functioned as developmental enhancers *in vivo* (Fig. 4, Extended Data Fig. 10, Supplementary Table 11). Furthermore, 90% (26 of 29) of active enhancers showed activity in the predicted tissue, with 23 being exclusive to that tissue (Extended Data Fig. 10, Supplementary Table 11). For example, elements specifically accessible in the neuronal, ectodermal or muscle clades show enhancer activity in the developing central nervous system (with some amnioserosa) (Fig. 4a), epidermis (Fig. 4b) and muscle (Fig. 4c), respectively. Elements that are specifically accessible in the mesendoderm clade

often act as enhancers in either the gut endoderm or haemocytes (mesoderm). Enhancer 4, for example, is accessible in cells of the developing midgut (endoderm) at both 6–8 h and 10–12 h, matching its activity in the anterior–posterior midgut during these stages (Fig. 4d). The only exceptions to our predictions were three of the seven elements that are specifically accessible in clade 4, which when tested were active in yolk nuclei (Extended Data Fig. 10). As the yolk is extra-embryonic, this was unexpected, and suggests a potential regulatory link between the yolk and mesendodermal tissues, which is supported by the role of the GATA transcription factor serpent in both yolk[25] and non-myogenic mesoderm[14].

In summary, our results demonstrate the power of sci-ATAC-seq to not only elucidate the developmental dynamics of chromatin accessibility, but also for the large-scale prediction of *in vivo* enhancer activity. Altogether, we identified 30,075 putative distal regulatory elements exhibiting clade-specific accessibility (Supplementary Table 1). By combining reads from cells within each *t*-SNE cluster, we generated cell-type-specific tracks of chromatin accessibility, which reveal a wealth of differences between cell types, and a powerful resource for future investigations (http://shiny.furlonglab.embl.de/scATACseqBrowser/). We also provide site-by-cell matrices and vignettes to facilitate further exploration of the data (http://atlas.gs.washington.edu).

The sparsity of data from single-cell molecular profiling technologies, including sci-ATAC-seq, remains a challenge. Although insights can be derived by aggregating observations across subsets of cells, as done here, increasing the number of reads per cell will increase the granularity at which chromatin accessibility can be explored. Combinatorial indexing is subject to collisions; with our current strategy, around 12% of cell barcodes are expected to represent aggregates of two or more cells. Analogous to doublets in emulsion-based single-cell RNA sequencing, collisions primarily add noise to the aggregate profiles of clades, but can sometimes lead to artefactual clusters. We present a strategy for identifying such clusters here; however, collisions are likely to be more effectively overcome by additional rounds of combinatorial indexing[26], which would also increase throughput.

Looking forward, an expanded dataset that includes many more cells per time point and covers the entirety of *Drosophila* development has the potential to identify rarer cell types and reveal a fully continuous view of the landscape of chromatin accessibility as it unfolds. Our ability to understand how changes in the regulatory landscape underlie lineage commitment would be greatly aided by the concurrent measurement of chromatin accessibility and transcription. In the long term, the integration of chromatin state, transcriptional output[26], lineage history[27,28]
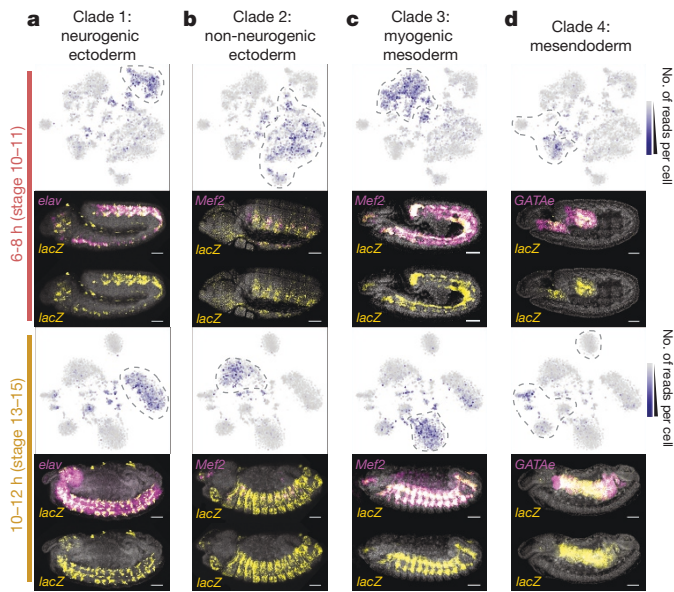
Figure 4 | Prediction of tissue-specific enhancer activity using sci-ATAC-seq. a–d, Examples of candidate LSI clade-specific enhancers tested with transgenic reporters. For each time point, upper panels show the *t*-SNE map with blue intensity representing the number of sci-ATAC-seq reads obtained from each tested element. Cell clusters bounded by dashed lines correspond to the predicted clade of activity. Lower panels show transgenic embryos with DAPI-stained nuclei (grey), *in situ* hybridization of the *lacZ* reporter gene driven by the enhancer (yellow), and a tissue marker (magenta). All embryo images are lateral views, with anterior left and dorsal up, and are representative of observations across hundreds of embryos. Scale bar, 50 μm. The activity and an overview of all tested enhancers are shown in Extended Data Fig. 10.

and spatial information[29,30] at single-cell resolution has the potential to unlock how an organism's genome encodes its development.

1. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348,** 910–914 (2015).
2. Maduro, M. F., Meneghini, M. D., Bowerman, B., Broitman-Maduro, G. & Rothman, J. H. Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3β homolog is mediated by MED-1 and -2 in *C. elegans. Mol. Cell* **7,** 475–485 (2001).
3. Sethi, A. J., Wikramanayake, R. M., Angerer, R. C., Range, R. C. & Angerer, L. M. Sequential signaling crosstalk regulates endomesoderm segregation in sea urchin embryos. *Science* **335,** 590–593 (2012).
4. Rodaway, A. & Patient, R. Mesendoderm. An ancient germ layer? *Cell* **105,** 169–172 (2001).
5. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biol.* **12,** R43 (2011).
6. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44,** 148–156 (2012).
7. Kvon, E. Z. *et al.* Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo. Nature* **512,** 91–95 (2014).
8. Gallo, S. M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila. Nucleic Acids Res.* **39,** D118–D123 (2011).
9. Frise, E., Hammonds, A. S. & Celniker, S. E. Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. *Mol. Syst. Biol.* **6,** 345 (2010).
10. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* **3,** research0088.1 (2002).
11. Bonn, S. *et al.* Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. *Nat. Protoc.* **7,** 978–994 (2012).
12. Doe, C. Q. Temporal patterning in the *Drosophila* CNS. *Annu. Rev. Cell Dev. Biol.* **33,** 219–240 (2017).
13. Ciglar, L. & Furlong, E. E. Conservation and divergence in developmental networks: a view from *Drosophila* myogenesis. *Curr. Opin. Cell Biol.* **21,** 754–760 (2009).
14. Spahn, P. *et al.* Multiple regulatory safeguards confine the expression of the GATA factor serpent to the hemocyte primordium within the *Drosophila* mesoderm. *Dev. Biol.* **386,** 272–279 (2014).
15. Reuter, R. The gene *serpent* has homeotic properties and specifies endoderm versus ectoderm within the *Drosophila* gut. *Development* **120,** 1123–1135 (1994).
16. Cannavò, E. *et al.* Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541,** 402–406 (2017).
17. Van Der Maaten, L. & Hinton, G. H. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
18. Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **344,** 1492–1496 (2014).
19. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell developmental trajectories. *Nat. Methods* **14,** 979–982 (2017).
20. Lecuit, T., Samanta, R. & Wieschaus, E. *slam* encodes a developmental regulator of polarized membrane growth during cleavage of the *Drosophila* embryo. *Dev. Cell* **2,** 425–436 (2002).
21. Beiman, M., Shilo, B. Z. & Volk, T. Heartless, a *Drosophila* FGF receptor homolog, is essential for cell migration and establishment of several mesodermal lineages. *Genes Dev.* **10,** 2993–3002 (1996).
22. Okumura, T., Matsumoto, A., Tanimura, T. & Murakami, R. An endoderm-specific GATA factor gene, dGATAe, is required for the terminal differentiation of the *Drosophila* endoderm. *Dev. Biol.* **278,** 576–586 (2005).
23. Clark, H. F. *et al. Dachsous* encodes a member of the cadherin superfamily that controls imaginal disc morphogenesis in *Drosophila. Genes Dev.* **9,** 1530–1542 (1995).
24. Simcox, A. A. & Sang, J. H. When does determination occur in *Drosophila* embryos? *Dev. Biol.* **97,** 212–221 (1983).
25. Tingvall, T. O., Roos, E. & Engström, Y. The GATA factor serpent is required for the onset of the humoral immune response in *Drosophila* embryos. *Proc. Natl Acad. Sci. USA* **98,** 3884–3888 (2001).
26. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357,** 661–667 (2017).
27. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353,** aaf7907 (2016).
28. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain by scGESTALT. Preprint at https://doi.org/10.1101/205534 (2017).
29. Karaiskos, N. *et al.* The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358,** 194–199 (2017).
30. Frieda, K. L. *et al.* Synthetic recording and *in situ* readout of lineage information in single cells. *Nature* **541,** 107–111 (2017).
31. Tomancak, P. *et al.* Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8,** R145 (2007).
32. Hammonds, A. S. *et al.* Spatial expression of transcription factors in *Drosophila* embryonic organ development. *Genome Biol.* **14,** R140 (2013).

METHODS

Fixation of embryos and nuclear isolation. Wild-type *D. melanogaster* embryos were collected and fixed as previously described[33]. In brief, embryos were collected on apple-agar plates in two-hour windows following three one-hour pre-collections to synchronize the collections. After ageing (at 25 °C) to the appropriate time window, embryos were washed from the plates, cleaned and dechorionated in 50% bleach for 2 min, followed by 15-min fixation with shaking at room temperature in cross-linking solution (50 mM Hepes, 1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, pH 8, 1.8% formaldehyde v/v) with a heptane layer. Fixation was stopped by washing with 125 mM glycine in PBS. The embryos were washed, dried and frozen at −80 °C in ~1-g aliquots. Embryo dissociation and nuclear isolation were performed as described previously (steps 1–10)[11] using a dounce homogenizer and a 22G needle. The resulting nuclei were pelleted at 2,000*g* at 4 °C, resuspended in nuclear freezing buffer (50 mM Tris at pH 8.0, 25% glycerol, 5 mM Mg(OAc)$_2$, 0.1 mM EDTA, 5 mM DTT, 1× protease inhibitor cocktail (Roche), 1:2,500 super-asin (Ambion)) and flash frozen in liquid nitrogen.

Collection of sci-ATAC-seq data. Our protocol for generating sci-ATAC-seq data was largely as previously described[1], but with a few important improvements. Frozen nuclei were thawed quickly in a 37 °C water bath and then pelleted at 500*g* for 5 min at 4 °C, aspirated and resuspended in cold lysis buffer (supplemented with protease inhibitors). Nuclei were stained with 3 μM DAPI and 2,500 DAPI$^+$ nuclei were sorted into each well of a 96-well plate containing 9 μl lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl$_2$ and 0.1% IGEPAL CA-630[34], supplemented with protease inhibitors (Sigma)) and 10 μl TD buffer (Illumina, part of FC-121-1031) in each well. One microlitre of each of the 96 custom and uniquely indexed Tn5 Transposomes (Illumina, 2.5 μM)[35] was then added to each well and nuclei were incubated at 55 °C for 30 min. Following tagmentation, 20 μl 40 mM EDTA (supplemented with 1 mM spermidine) was added to stop the reaction and the plate was incubated at 37 °C for 15 min. All wells of the plate were then pooled, nuclei were stained again with 3 μM DAPI and 25 DAPI$^+$ nuclei were sorted into each well of a second set of 96-well plates that contained 12 μl reverse crosslinking buffer (11 μl EB buffer (Qiagen) supplemented with 0.5 μl 20 mg/ml Proteinase K (Qiagen) and 0.5 μl 1% SDS). For each time point, we collected four plates of nuclei at this stage. We expect that sorting 25 nuclei into each well at this stage will result in approximately 12% of barcodes representing more than one nucleus (collisions)[1]. Nuclei were then incubated overnight at 65 °C. Proceeding from reverse-crosslinking, we added primers (0.5 μM final concentration, Supplementary Table 12), 7.5 μl NPM polymerase master mix (Illumina, FC-121-1012) and BSA (2× final concentration; NEB) to each well. Tagmented DNA was then PCR amplified. To determine the number of cycles required, we first amplified several test wells of nuclei that had been sorted onto an additional plate and monitored the reactions with SYBR green on a qPCR machine to establish when the libraries reached saturation. The cycling conditions were as follows: 72 °C 3 min, 98 °C 30 s; 98 °C 10 s, 63 C 30 s, 15–25 cycles; 72 °C 1 min, hold at 10 °C.

We have found that the optimal number of cycles can vary from one experiment to the next, but is usually in the range of 15–25 cycles. After PCR amplification, all wells were pooled and split across four DNA Clean & Concentrator-5 columns (Zymo) and all four products were then pooled and cleaned again using Ampure beads (Agencourt). Finally, the concentration and quality of the libraries was determined using the BioAnalyzer 7500 DNA kit (Agilent). For sequencing, equimolar libraries from the three time points were pooled and loaded at 1.5 pM on a NextSeq High output 300 cycle kit and sequenced using custom primers and a custom sequencing recipe[35]. Fifty base pairs were sequenced from each end, in addition to the barcodes introduced during tagmentation and PCR amplification. This improved protocol resulted in roughly an order of magnitude more reads per cell than previously reported.

Read alignment, cell assignment and duplicate removal. To process the data, BCL files were converted to fastq files using bcl2fastq v.2.16 (Illumina). Each read was assigned a barcode which was actually made up of four individual components: a tagmentation barcode and a PCR barcode added to the P5 end of the molecule, and a distinct tagmentation and PCR barcode added to the P7 end of the molecule. To correct for sequencing and/or PCR amplification errors, we broke the barcode into its constituent parts and matched each piece against all possible barcodes. If the component was within three edits of an expected barcode and the next best matching barcode was at least two edits further away, we fixed the barcode to its presumptive match. Otherwise, we classified the barcode as ambiguous or unknown. We next mapped each read to the dm3 reference genome using bowtie2[36] with '-X 2000 -3 1' as options and filtered out read pairs that did not map uniquely to autosomes or sex chromosomes with a mapping quality of at least 10, as well as reads that were associated with ambiguous or unknown barcodes. Of 430,658,635 sequenced read pairs, 301,314,040 (70%) mapped to the nuclear reference genome with an assigned cell barcode. By contrast, only 366,468 read pairs (0.09%) mapped to the mitochondrial genome, with an assigned cell

barcode. We subsequently removed PCR duplicates for all reads that mapped to the nuclear genome using a custom Python script that only considered reads assigned to the same barcode. Finally, to determine which barcodes represented genuine cells (as opposed to background reads assigned to improper barcodes), we counted the number of reads assigned to each barcode and log-transformed those counts and then used the mclust package in R[37,38], which fits the data using a mixture model and determines the maximum likelihood parameters for a given number of distributions, to define two distributions of barcodes—setting the read depth cut-off for a cell at the point at which we were 95% confident that the barcode belonged to the higher read-depth distribution. Considering the distribution of barcodes for all three experiments at the same time, we determined this read-depth cut-off to be 500 reads (that is, we required a barcode to be associated with at least 500 reads to be considered a true cell; Extended Data Fig. 1). See http://atlas.gs.washington.edu for more details on data processing.

Latent semantic indexing. To further process the raw data, we first broke the genome into 2-kb windows and then scored each cell according to whether it had any insertions in each window, creating a large binary matrix of windows by cells for each time point. Based on this binary matrix, we retained only the top 20,000 most commonly used sites (this number could extend a little above 20,000 because we retained all sites that were tied at the threshold for cell counts) and then filtered out the 10% of cells with the smallest number of accessible sites. We then normalized and re-scaled these large binary matrices by using the term frequency–inverse document frequency (TF–IDF) transformation. We first weighted each site that was accessible in an individual cell by the total number of sites accessible in that cell. We then multiplied these weighted values by log(1 + the inverse frequency of each site across all cells). Subsequently, we performed singular value decomposition on the TF–IDF matrix and then generated a lower-dimensional representation of the data by only considering the second to sixth dimensions (because we have found that the first dimension is always highly correlated with read depth). These LSI scores were then used to cluster cells and windows on the basis of cosine distances using the ward algorithm in R. Scores of accessibility were standardized by row and capped at ± 1.5 for visualization. Visual examination of the resulting bi-clustered heat map identified 4–5 major clades for each time point.

Peak calling. To identify specific regulatory elements within each of the major clades at each time point, we aggregated the data across cells from each clade using a process we call '*in silico* cell sorting'. To do so we collected all the unique mapped reads associated with cells that were assigned to a given clade and saved them as a distinct bam file. Then for each bam file representing a clade, we used MACS2[39] to identify peaks of increased insertion frequency, as well as summits of accessibility within each of those peaks. For MACS, we used the macs2 callpeak command with the following parameters: "--nomodel --keep-dup all --extsize 200 --shift -100 --format BAM --gsize mm --call-summits". For downstream analyses we generated a master list of potential regulatory elements by taking 150-bp windows centred on all summits called in each clade in each time point and merged them with the BEDTools program[40]. For Extended Data Fig. 1d, we also compared our sci-ATAC-seq data to previously collected DNase-seq bulk data[5] on whole embryos at similar time points. To be consistent in our comparisons (and provide a comprehensive list of peaks), we downloaded the raw DNase-seq reads (36 bp, single-end), remapped them with our pipeline and called peaks with MACS2 as described above. Specifically, we downloaded two replicates for each of three time points: stage 5, stage 11 and stage 14. Peaks called on each replicate independently were intersected to create a master list of peaks for each time point, which were then intersected with our sci-ATAC-seq data.

Identification of differentially accessible sites. To identify regulatory elements that were more specifically accessible in individual clades, we generated a new binary matrix of insertion scores for individual cells using the master list of summits of accessibility described above. We then used a logistic regression framework to test whether cells of a given clade were more likely to have insertions at a given site relative to all other cells. To identify sites that were specifically more accessible in a single clade, we first found summits that were significantly more open in a given clade at a 1% FDR, including log$_{10}$(total unique reads) for each individual cell as a covariate. To ensure that these sites were specific to any one clade, we also filtered out sites that were significantly accessible in any other clade at a relaxed 20% FDR. All testing of differential accessibility was implemented with the Monocle 2 package[19,41] using the binomialff test. For this analysis, only sites observed in at least 50 cells in a given time point were tested.

***k*-mer discovery.** We used SeqGL[42] to identify motifs that were enriched in clade-specific elements. To do so, we started with all clade-specific sites, based on our logistic regression testing described above. Because our master list of sites included sites of variable length (after merging all sites from all clusters), we only considered 150-bp windows centred on summit midpoints. We also removed sites within 500 bp of a transcription start site (TSS), to focus on tissue-specific distal elements. As a background set of regions we randomly selected an equal number of

sites from the master summit list that matched the GC and repeat element content of the test set (this was controlled using a script provided in the gkm-SVM software package)[43]. Finally, instead of default parameters, we used 200 groups and 30,000 features, similar to the parameters used to analyse DNase-seq data in the original SeqGL publication[42].

Enrichments for tissue or cell-type activity and transcription factor binding data. To perform categorical enrichments, we annotated regions, windows and peaks of the non-coding genome using two types of experimental information: (1) tissue-specific expression of the nearest gene comprising *in situ* hybridization data from the Berkeley *Drosophila* Genome Project (http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl) and a download of the FlyBase gene-expression annotations (May 2016); (2) a custom enhancer database of ~8,000 transgenic reporter assays covering 15% of the non-coding genome, containing spatio-temporal information of ~4,000 active developmental enhancers (CAD4; Supplementary Table 13). We compiled the enhancer database (CAD4) from three primary resources: our previous CRM Activity Database (CAD)[6], entries from the RedFly enhancer database (Release 5)[8], and data from the Vienna Tiling Project[7]. We compiled this dataset in two steps. First, all expression terms (and timing terms, where available) were mapped to a common standard (FlyBase anatomy terms v.1.47) and, when timing information was available, a common set of stage windows (stages 1–3, stages 4–6, stages 7–8, stages 9–10, stages 11–12, stages 13–16). In most cases, the mapping was automatic and unambiguous. In some cases, manual term matching was required (generally unambiguous). In the second step, we merged overlapping entries from CAD3 and the RedFly database and manually removed redundant information. Given the different methodologies used in the compilation of the data sources, no attempt was made to merge entries from CAD3/RedFly with the Vienna Tiles.

Almost all expression terms for both the gene and enhancer annotations could be mapped to a common set of hierarchically organized anatomical terms (FlyBase anatomy OBO file v.1.47). In the few cases where an exact match could not be found, a choice was made manually or using the map provided by FlyBase (FBrf0219073). The stage or timing information from both datasets was shifted as needed to match a common set of grouped stages (stages 1–3, stages 4–6, stages 7–8, stages 9–10, stages 11–12, stages 13–16). The compiled data are shown in Supplementary Table 13. In addition to BDGP/FlyBase gene expression data, we made use of *Drosophila*-specific gene-level functional information (biological process, molecular function and cellular compartment) downloaded from the Gene Ontology Consortium (v.1.2) and additional, higher-level functional annotations downloaded from the PANTHER classification system (v.8) corresponding roughly to the higher-level categories of the GO-SLIM ontology.

To further explore the functions of specific regions of noncoding DNA, we also made use of a custom compilation of high-quality transcription factor binding data from ChIP studies during embryogenesis (taken from ref. 16) that allowed us to assign transcription factor binding events to each sciATAC window or peak. Transcription factor binding motifs were taken from this same dataset. To infer likely transcription factor binding events, we scanned under published ChIP peaks for instances of the motif using FIMO[44]. Enrichments for these data are listed under the category name 'custom' in the enrichment data tables.

Categorical enrichments. To identify enriched categories within the LSI clades, we first assigned categorical labels by looking for overlaps between our summit regions and our enhancer activity database, with summits inheriting the timing and expression labels of all overlapping enhancers. Gene-based annotations (expression, GO and PANTHER terms) were assigned by association to the nearest gene.

To identify differentially accessible summit regions, we used a logistic-regression framework (see above) as applied to all summit regions containing reads in at least 50 cells. Enriched summit regions constituted the foreground set for any clade, with the remaining tested summit regions constituting the background set. For each of our category sets (for example, enhancer expression, gene expression or GO), we used a Fisher's exact test to look for over-representation of each category among our foreground set relative to the background set. Because many of our categories are strongly overlapping, we have applied no formal correction for multiple comparison, choosing instead to focus on large, consistent enrichments with highly significant P values. Overlaps among significant categories were visualized by plotting distances between categories using the pyEnrichment package (https://github.com/ofedrigo/pyEnrichment) to avoid overcalling a category.

Categorical enrichment within our t-SNE clusters was assessed similarly. Foreground sets per cluster (within each time point) were assessed using the results of our binomial enrichment test (q value ≤ 0.01 and a beta > 0). The background set consisted of all other tested summits at that time point (see above).

***t*-SNE and cluster identification.** To identify clusters of cells with finer resolution than the LSI-based clades, we used t-SNE[17] for dimensionality reduction. We started with the same binary matrix of insertions in summits that we used to identify clade-specific differentially accessible sites. We again filtered out the lowest

10% of cells (in terms of site coverage) and in this case we retained only sites that were observed in at least 5% of cells. We then transformed this matrix with the TF–IDF algorithm described above. Finally, we generated a lower-dimensional representation of the data by including the first 50 dimensions of the singular value decomposition of this TF–IDF-transformed matrix. This representation was then used as input for the Rtsne package in R[17,45,46]. To identify clusters of cells in this 2D representation of the data, we used the density peak clustering algorithm[18] as implemented in Monocle 2[19,41]. Rho and delta parameters were chosen to be very inclusive of outlier peak centres (based on the decision plot), while making sure that the clusters were sensible based on visual inspection of the cluster assignments on the t-SNE plot.

***t*-SNE differential accessibility.** To identify summits that were significantly more accessible in t-SNE-defined cell clusters, we used a similar framework to the one described for LSI-based clades above. There were, however, a few differences. In this case, we consider sites that were seen in at least 10 cells in any time point (instead of 50). In addition, we did not use a second cut-off to determine specificity within a time point.

Sexing individual nuclei. Another biological axis of the data that came to light through the use of t-SNE plots was that we were able to clearly distinguish nuclei from male and female embryos. In an initial analysis, we included data from the sex chromosomes while clustering cells (as was done for the germ-layer analysis). This resulted in many individual cell clusters appearing 'bi-lobed' (Extended Data Fig. 5a), which prompted us to explore whether there was sex bias in the lobes of individual cell clusters. We found that the distribution of reads mapping to the X chromosome in individual cells was distinctly bimodal (Extended Data Fig. 5b), allowing us to assign a sex to each cell. When we coloured the t-SNE plots according to these sex assignments we found that the lobes of individual cell clusters almost perfectly segregated the sexes (Extended Data Fig. 5c). Although this may be very useful for future studies, we alleviated this bi-lobed problem here by excluding sex chromosome reads from our analysis and re-clustered cells with t-SNE. This resolved the bi-lobed structures and removed the sex bias from almost every individual cluster (Extended Data Fig. 5d).

Arranging single cells from 2–4-h embryos along developmental trajectories. Because we noted that cells from 2–4-h embryos were distributed across the t-SNE map in a manner consistent with their developmental stage, we sought to more formally evaluate the arrangement of individual cells along a temporal trajectory. We used Monocle 2[19,41] v.2.5.3, which uses a reverse graph embedding algorithm to learn trajectories in single-cell data and was recently extended to single-cell ATAC-seq data[47]. To define sites to use for ordering cells, we combined the t-SNE clusters into major groups on the basis of our annotations—blastoderm, mesoderm, endoderm, ectoderm, neural ectoderm, unknown and collisions—and identified sites that were differentially accessible (1% FDR) between each cluster and all other cells within that time point (with the exception of the collision and unknown clusters). We then took the union of sites that were among the 100 most differentially accessible for each cluster and used this set of sites to order cells in Monocle. In order to reduce the sparsity of the data, we aggregated all sites that were within 1 kb of each other and summed their reads to obtain a regional score accessibility. Using these aggregated sites as features, cells were ordered by the DDRTree algorithm in three dimensions ('max_components = 3'), with the ncenter parameter set to 200 and the maxIter parameter set to 1,000 during the dimensionality-reduction step. Only the first two dimensions are visualized and the coordinates of the first dimension were multiplied by −1 so that pseudotime would run from left to right (Fig. 2c). This resulted in a tree with four differentiated branches representing the major germ layers (one is a possibly spurious, short branch along the ectodermal lineage). On the basis of this ordering, we aimed to identify sites that were significantly associated with progression in pseudotime using the likelihood-ratio testing framework in Monocle 2 (Supplementary Table 6). As with ordering the cells, we adopted a strategy to reduce the sparsity of our data. Specifically, we binned the pseudotime into 100 bins and counted how many cells had accessible chromatin in each pseudotime bin for each site. All sites that were accessible in more than ten cells were tested. To identify sites that were associated with pseudotime in a lineage-specific fashion we used a similar framework. First, we separated out cells along each unbranched path through the trajectory to test separately for pseudotime dependence. We took the cells at the tip of each lineage state and traversed the graph to the root state (that is, beginning of the pseudotime), collecting the cells that were arranged along this path. As mentioned above, there was a small branch off of the ectodermal lineage that was ignored for this analysis. Then we binned the cells along this single pseudotime branch and performed likelihood ratio testing for each lineage as we did for the global pseudotime measure (Supplementary Tables 7–9). After testing all three lineages, we defined a site as specific to a lineage if it was significantly associated with pseudotime in that lineage (1% FDR) but was not significantly associated with pseudotime in the other two lineages at a relaxed threshold (20% FDR).

Identifying clusters of cells that are likely artefacts of barcode collisions. Several small clusters (for example, cluster 6 at 6–8 h) appear to be mixtures of cells from different germ layers and/or tissues, based on our enrichment analysis. To determine whether these were technical (due to barcode collisions, where one cell barcode represents the nuclear contents of two cells) or biological, we used two metrics to identify collisions (instances wherein two or more cells coincidentally pass through the same combination of wells during sci-ATAC-seq). First, we looked at the estimated complexity of individual cells that make up these small clusters, as collisions are expected to be twice as complex on average as barcodes that truly represent an individual cell. To calculate the estimated library complexity (that is, the estimated total number of unique reads per cell in the library), we used the same algorithm as implemented in Picard (http://broadinstitute.github.io/picard) on a cell-by-cell basis. Second, we considered whether the proportion of reads mapping to the X chromosome for cells in these clusters was distinctly bimodal, as collisions would be just as likely to combine data from cells of the opposite sex as from two cells of the same sex (Extended Data Fig. 7). While the vast majority of clusters exhibited distributions of complexity and X chromosome coverage consistent with single nuclei, a small subset of clusters in each time point showed either higher complexity than expected, more unimodality of reads mapping to the X chromosome, or both—consistent with our suspicion that these are cell collision clusters (Extended Data Fig. 7). At 2–4 h, we identified one (2.3% of cells), at 6–8 h we identified three (5.8% of cells) and at 10–12 h we identified six (7.3% of cells) potential collision clusters (Figs 2a, 4a, b, purple clusters).

Transgenic enhancer assays. Candidate clade-specific enhancers were selected from sci-ATAC-seq summits using the following criteria only: (1) summit shows enriched accessibility specifically in the target cell clade at 6–8 h and/or 10–12 h (q value < 0.01 and beta > 0 in target clade, q value > 0.2 in all other clades); (2) summit does not fall within 500 bp of an annotated transcription start site; (3) summit does not overlap a region already in our database of characterized developmental enhancers. Summits showing a range of effect sizes (beta) were selected (minimum beta approximately 1.9; see Supplementary Table 11). The selected regions, plus 100–200 bp of flanking sequence, were PCR amplified from genomic DNA (primers are listed in Supplementary Table 11) and cloned upstream of a minimal *hsp70* promoter driving a *LacZ* reporter gene in an attB-containing plasmid. All constructs were injected into embryos according to standard methods[48] and inserted into the attP landing site line M{3×P3-RFP.attP′}ZH-51C via PhiC31 integrase insertion[49], yielding integration at chromosomal position 51C1. Transgenic lines were generated by BestGene. Ten elements from each of the four germ-layer clades were initially selected—some failed at the cloning or transgenesis phase. We obtained 31 transgenic lines, representing six candidate regions with specific accessibility in neurogenic ectoderm, ten in non-neurogenic ectoderm, eight in myogenic mesoderm and seven in non-myogenic mesoderm plus endoderm.

Overnight collections of homozygous embryos spanning all stages of embryogenesis were formaldehyde-fixed, stained by double fluorescent *in situ* hybridization[50], and mounted in ProLong Gold with DAPI (Invitrogen; cat. #P36931). Antisense *in situ* probes against *LacZ* and a tissue marker gene were used: *Mef2* marking myogenic mesoderm was used for predicted myogenic mesoderm and non-neurogenic ectoderm enhancers; *GATAe* was used for predicted non-myogenic mesoderm and endoderm enhancers. For the predicted neurogenic ectoderm enhancers, neurons were marked by immunostaining with antibodies against the Elav protein (Elav-9F8A9; Developmental Studies Hybridoma Bank). The annotation of enhancer activity is based on observations across hundreds of embryos. Representative images were acquired with a Zeiss LSM780 laser-scanning confocal microscope using a PlanApo 20×/NA 0.8 objective at an effective pixel size of 461 nm in the *x–y* plane. Images were processed using Fiji[51]. Annotated *t*-SNE plots for each candidate enhancer were produced by plotting the sum of sci-ATAC-seq reads per cell that overlapped each tested genomic region.

FACS isolation of tissue-specific nuclei and DNase-seq. Target populations of cell nuclei from staged fixed embryos were obtained by FACS as previously described[11] with the following modifications. Prior to incubation with primary antibodies, nuclei from 6–8-h embryos were incubated in PBS supplemented with 5% BSA, 0.1% TritonX-100 and 0.2% Igepal-630 on a rotator at 4 °C for 30 min. Primary antibody staining was performed overnight at 4 °C in 3 ml PBS supplemented with 5% BSA and 0.1% TritonX-100 per 1g frozen embryos. Primary antibodies used were monoclonal anti-Elav (Developmental Studies Hybridoma Bank 9F8A9 at 1:100 dilution) to mark postmitotic neurons and anti-Mef2 (produced and pre-cleared in the Furlong laboratory and used at 1:200 dilution) to mark myogenic mesoderm. Secondary antibody staining was performed for 1 h at 4 °C in the same buffer. Following each antibody staining, nuclei were washed twice by pelleting and resuspending in 10 ml PBS supplemented with 5% BSA. An aliquot of stained, unsorted nuclei was put aside to represent the whole embryo. For DNase digestion, nuclei were resuspended in R buffer (7.5mM Tris pH8, 45mM NaCl, 30mM

KCl, 6mM MgCl₂, 1mM CaCl₂) and 10–20 million nuclei were digested using 5–20 U DNaseI at 37 °C for 3 min, and the reaction was stopped by adding 500 μl stop buffer (50mM Tris pH8, 100 mM NaCl, 0.1% SDS, 100 mM EDTA pH8). A small control digest without DNaseI was performed to assess DNA integrity. Following addition of RNaseA, samples were incubated at 55 °C for 10 min, then 25 μl proteinase K (25 mg/ml) was added and the samples were incubated overnight at 65 °C to reverse cross-links. A small aliquot was run on a 1% agarose gel to assess digestion levels, and optimal digests were size-fractionated using 10–40% sucrose gradients. DNA fragments ~100–500 bp in length were isolated from fractions using a Qiagen PCR clean up kit and checked for enrichment in known hypersensitive sites by qPCR. The digests with the highest qPCR enrichment were selected for library preparation using the NextFlex qRNA-seq Kit v.2 (Bioscientific #NOVA-5130-12). In brief, ~10–30 ng DNA consisting of ~100–500 bp fragments that result from DNase digestion was end-repaired and terminal adenosine residues were added. Adapters containing in-line molecular barcodes were ligated, after which the material was size selected using AMPure beads (negative selection with 0.6× beads, then positive selection with 0.98× beads). PCR amplification was performed using barcoded primers to introduce sample barcodes for 12–16 cycles, depending on input amount. The PCR-amplified library was purified using AMPure beads, quantified using a Qubit High-sensitivity DNA kit (Invitrogen), and sized on a Bioanalyzer High-Sensitivity DNA chip (Agilent). Libraries were pooled and sequenced in paired-end mode on a HiSeq2000 (Ilumina). Reads were mapped to the Dm3 reference genome using BWA aln[52], keeping only reads with a mapping quality score greater than 20. Duplicate reads originating from PCR were removed using the Je suite[53] making use of the molecular indices.

Ethics statement. Anti-Mef2 antibodies were generated from rabbits at EMBL in accordance with European Law and EMBL ethical guidelines. *Drosophila melanogaster* were reared and collected at EMBL in accordance with standard practice and the ethical standards of the European research community.

Code availability. Most of the code used in processing and analysis of the data in this article is available at http://atlas.gs.washington.edu. Any code not provided there will be made available upon request.

Data availability. All raw ATAC-seq and DNase-seq data are available through GEO (accession GSE101581) and ArrayExpress (E-MTAB-5999). BigWig files for coverage within each clade, regions of accessibility (peak calls) and a master list of all potential regulatory elements (Supplementary Table 1) will be made available on the Furlong laboratory web page (http://furlonglab.embl.de/data). To make the data easily accessible we have generated a searchable html page where users can select a *t*-SNE cluster or genomic locus of interest and visualize the data throughout the genome (http://shiny.furlonglab.embl.de/scATACseqBrowser/) and site-by-cell matrices and vignettes to facilitate further exploration of the data (http://atlas.gs.washington.edu).

33. Sandmann, T., Jakobsen, J. S. & Furlong, E. E. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nat. Protoc.* **1,** 2839–2855 (2006).
34. 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. *Nat. Methods* **10,** 1213–1218 (2013).
35. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46,** 1343–1349 (2014).
36. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
37. Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.* **97,** 611–631 (2002).
38. Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. *Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597* (Department of Statistics, Univ. of Washington, 2012).
39. Zhang, Y. *et al.* Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9,** R137 (2008).
40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
41. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
42. Setty, M. & Leslie, C. S. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. *PLOS Comput. Biol.* **11,** e1004271 (2015).
43. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLOS Comput. Biol.* **10,** e1003711 (2014).
44. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).
45. Van Der Maaten, L. Accelerating *t*-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15,** 3221–3245 (2014).

46. Krijthe, J. H. Rtsne: t-distributed stochastic neighbor embedding using a Barnes–Hut implementation. https://github.com/jkrijthe/Rtsne (2015).
47. Pliner, H. *et al.* Chromatin accessibility dynamics of myogenesis at single cell resolution. Preprint at https://doi.org/10.1101/155473 (2017).
48. Rubin, G. M. & Spradling, A. C. Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218,** 348–353 (1982).
49. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proc. Natl Acad. Sci. USA* **104,** 3312–3317 (2007).
50. Furlong, E. E., Andersen, E. C., Null, B., White, K. P. & Scott, M. P. Patterns of gene expression during *Drosophila* mesoderm development. *Science* **293,** 1629–1633 (2001).
51. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (2012).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
53. Girardot, C., Scholtalbers, J., Sauer, S., Su, S. Y. & Furlong, E. E. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17,** 419 (2016).
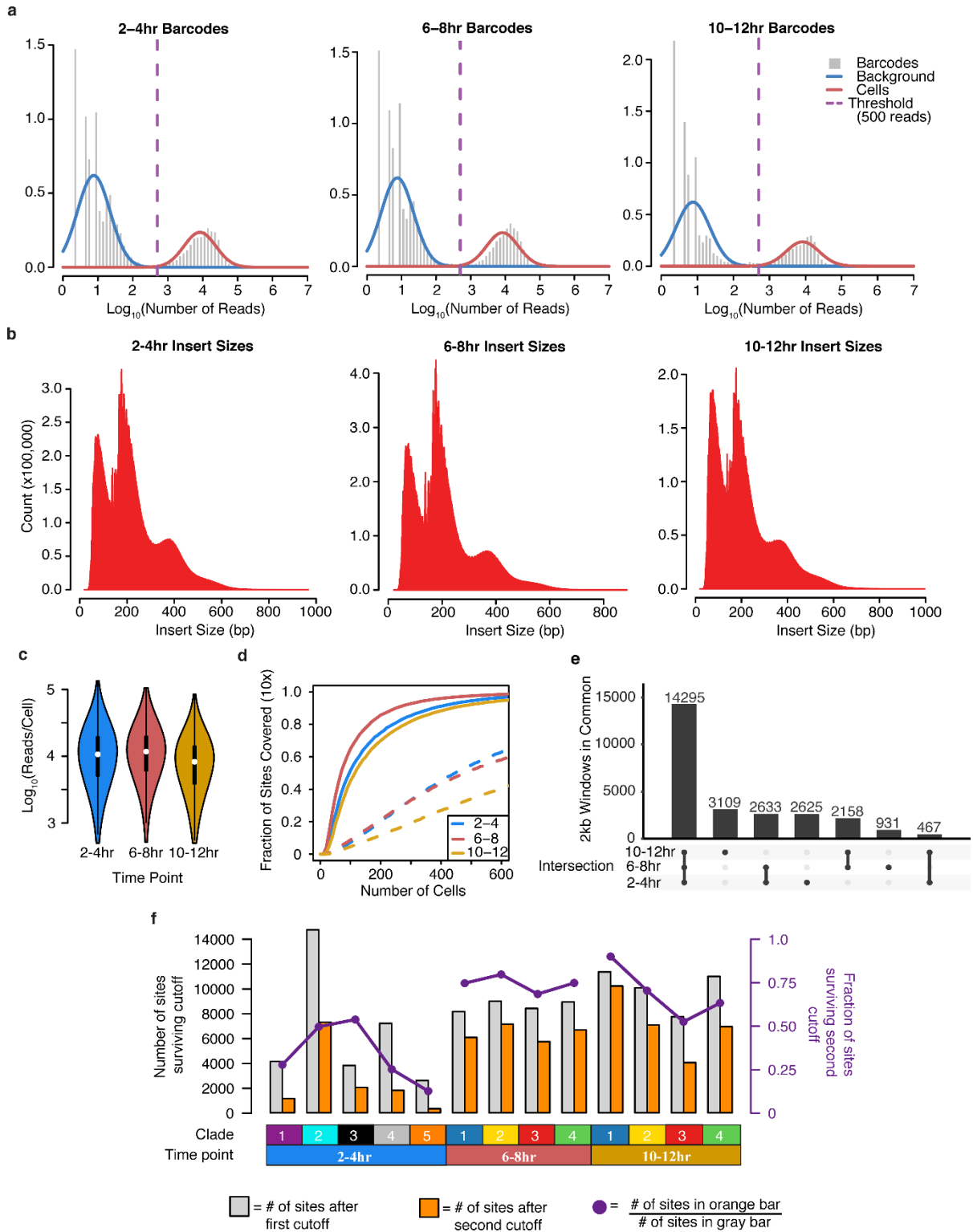
Extended Data Figure 1 | Summary of read distributions across the three sampled time points. a, $\log_{10}$ counts of sci-ATAC-seq reads per barcode at each time point are bimodally distributed. A threshold of 500 reads was used to identify barcodes corresponding to valid cells versus background. **b**, Fragment size distribution at each time point is consistent with expected nucleosomal banding pattern of standard (bulk) ATAC-seq experiments. **c**, Violin plot for distribution of unique, mappable reads per cell at each time point (2–4 h, $n = 8,024$; 6–8 h, $n = 7,880$; 10–12 h, $n = 7,181$) plotted on a logarithmic scale. White point indicates median value, thick black line extends to 25th and 75th percentile, and thin black lines extend to most extreme values within 1.5 times the interquartile range of the median. The filled colour width represents a density estimate

of the distribution of cells along the y axis. **d**, Fraction of previously characterized DHS covered in at least 10 cells upon sampling a given number of cells (solid lines) as compared to random genomic windows (dashed lines). **e**, An UpSet plot shows the degree to which the top 20,000 windows overlap between the three time points. Each bar shows the number of sites included in a specific intersection and the 'peg board' below shows which particular comparison is included in that bar. **f**, Bar plot of the number of sites identified as significantly open in each clade (1% FDR; grey bar, first cutoff) and the number of sites specific to that clade (orange bar, second cutoff). Overlaid on the barplot (purple points) is the fraction of sites passing the first cut-off that also pass the second cut-off (count of orange bar/count of grey bar).
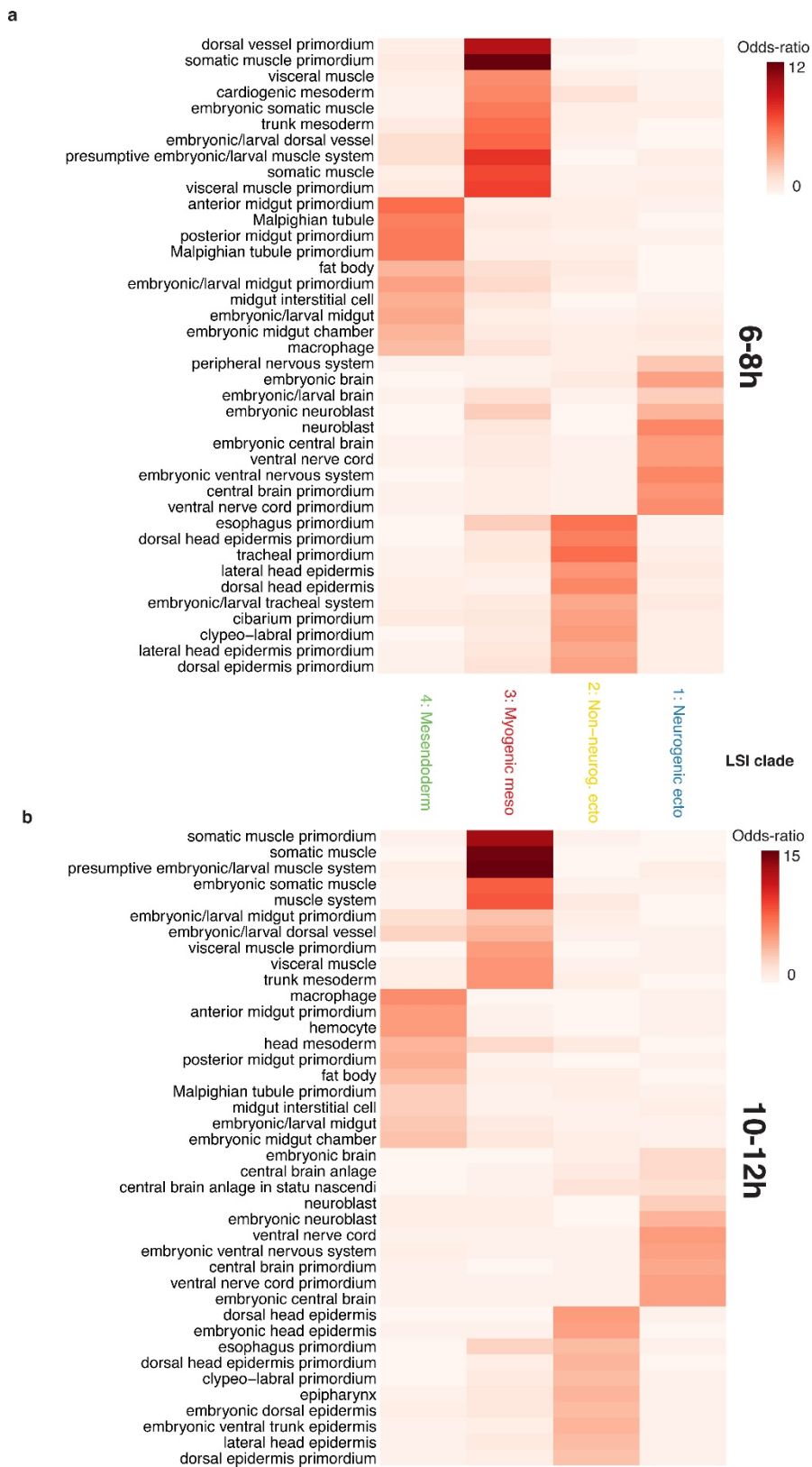
Extended Data Figure 2 | Enhancer enrichments for LSI clades at 6–8 h and 10–12 h. Enrichment for tissue-of-expression information for characterized distal enhancers overlapping clade-specific peaks at 6–8 h (**a**) and 10–12 h (**b**). Each column represents a different clade and each row represents an annotation term assigned to tested enhancer elements. Shading indicates the odds ratio for the intersection of enhancers sharing a given annotation with clade-specific accessible sites. Shown are all categories in the top ten enrichments of any clade (enrichment scores capped at 15 for display) containing at least 35 known enhancer overlaps.
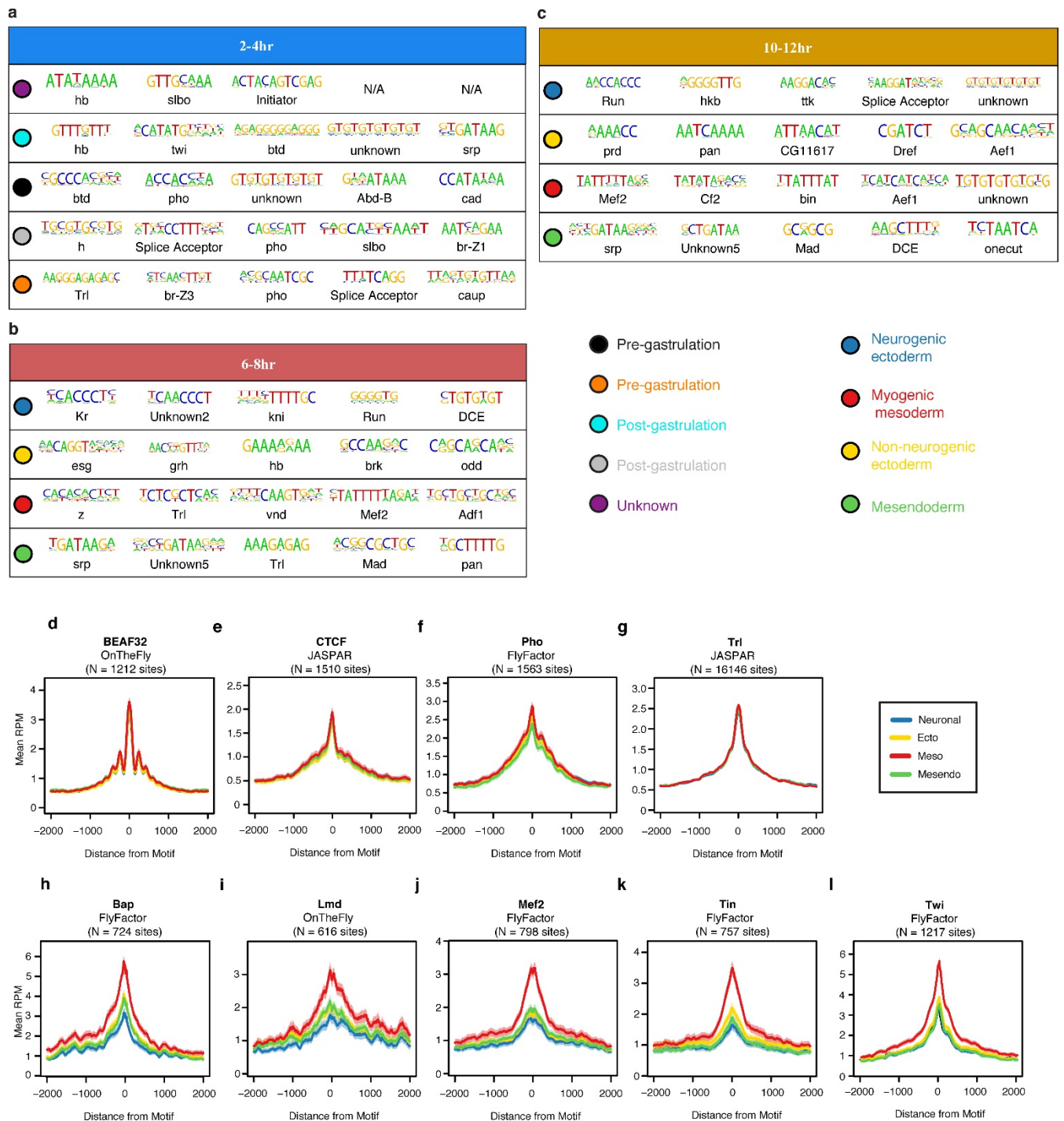
Extended Data Figure 3 | Relationship between transcription-factor binding motifs and occupancy, and LSI clade-specific accessibility. **a–c**, SeqGL was run on LSI clade-specific distal peaks at each time point to identify enriched sequence motifs. The top five most-enriched unique motifs for each clade are displayed. Coloured circles indicate the clade represented by each line. For the later time points (6–8 h and 10–12 h), blue is neurogenic ectoderm, yellow is non-neurogenic ectoderm, red is myogenic mesoderm and green is mesendoderm. The results show an enrichment of motifs for factors associated with early development at 2–4 h with more tissue-specific factor motifs (for example, mesodermal factor Mef2 or neural regulator Tramtrack) within germ-layer annotated clades at later stages of development. **d–l**, Using ChIP occupancy data (peaks) and transcription factor binding motifs compiled previously[16], we scanned for all transcription factor motif instances under ChIP peaks from datasets spanning 6–8 h of development using FIMO. Aggregate read counts in 4-kb windows centred on each identified motif instance are shown for each of the four LSI clades at 6–8 h. Green, endoderm; red, myogenic mesoderm; yellow, non-neurogenic ectoderm; blue, neurogenic ectoderm. Light shading in the same colours indicates 95% confidence intervals. **d–g**, Aggregate plots for four ubiquitous transcription factors (BEAF32, CTCF, Pho, and Trl) at 6–8 h. **h–l**, Aggregate plots for mesodermal transcription factors (Bap, Lmd, Mef2, Tin, Twi) at 6–8 h.
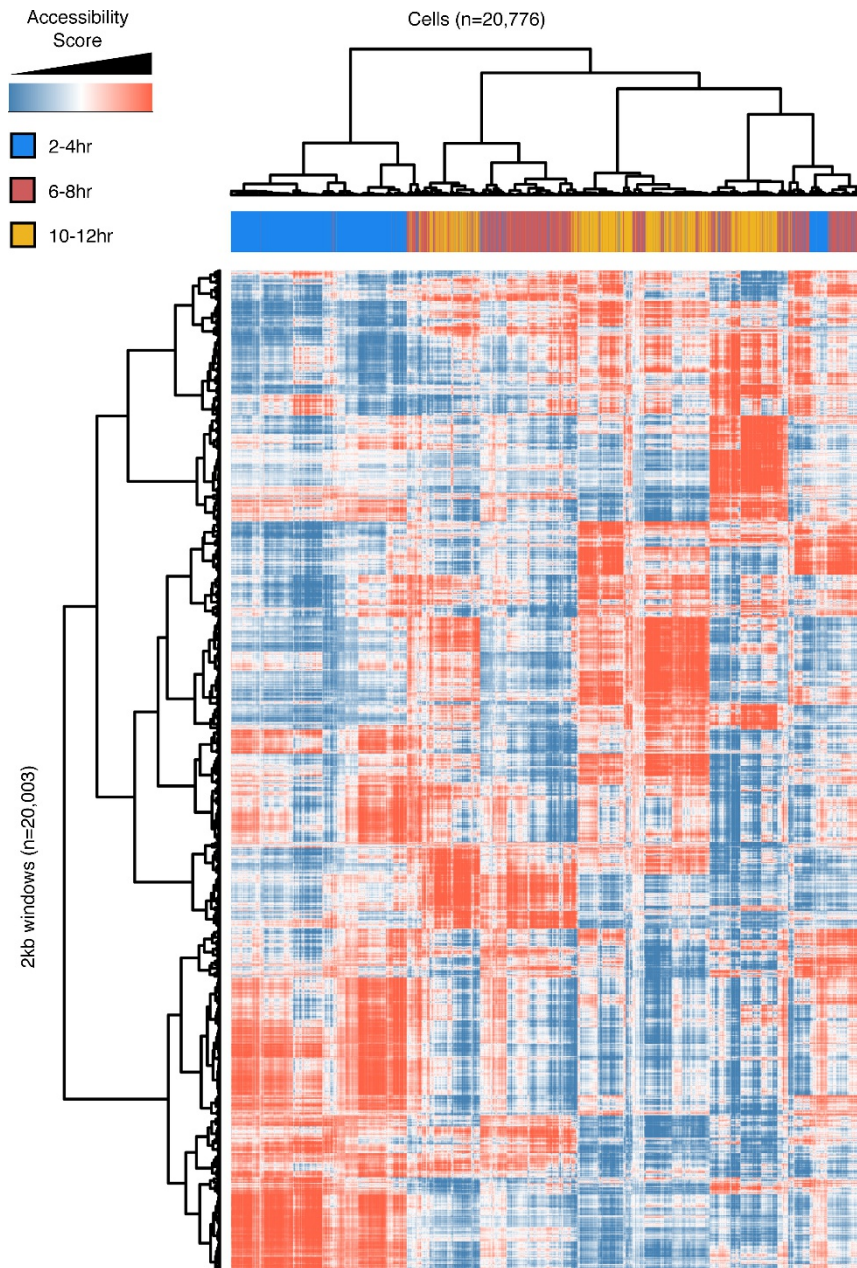
Extended Data Figure 4 | Similarities and differences in accessibility across all three time points. In addition to processing data from each time point independently, data from all cells can be analysed together (with the caveat that time point and batch are confounded). Here, we show binarized, LSI-transformed and clustered count data for 2-kb windows across the genome for cells from all three time points (blue, 2–4 h; red, 6–8 h; orange, 10–12 h) processed together. The predominant pattern is one in which 2–4-h cells cluster separately from 6–8-h and 10–12-h cells. Cells from 6–8 h and 10–12 h are intermingled, clustering first (roughly) by germ layer of origin.
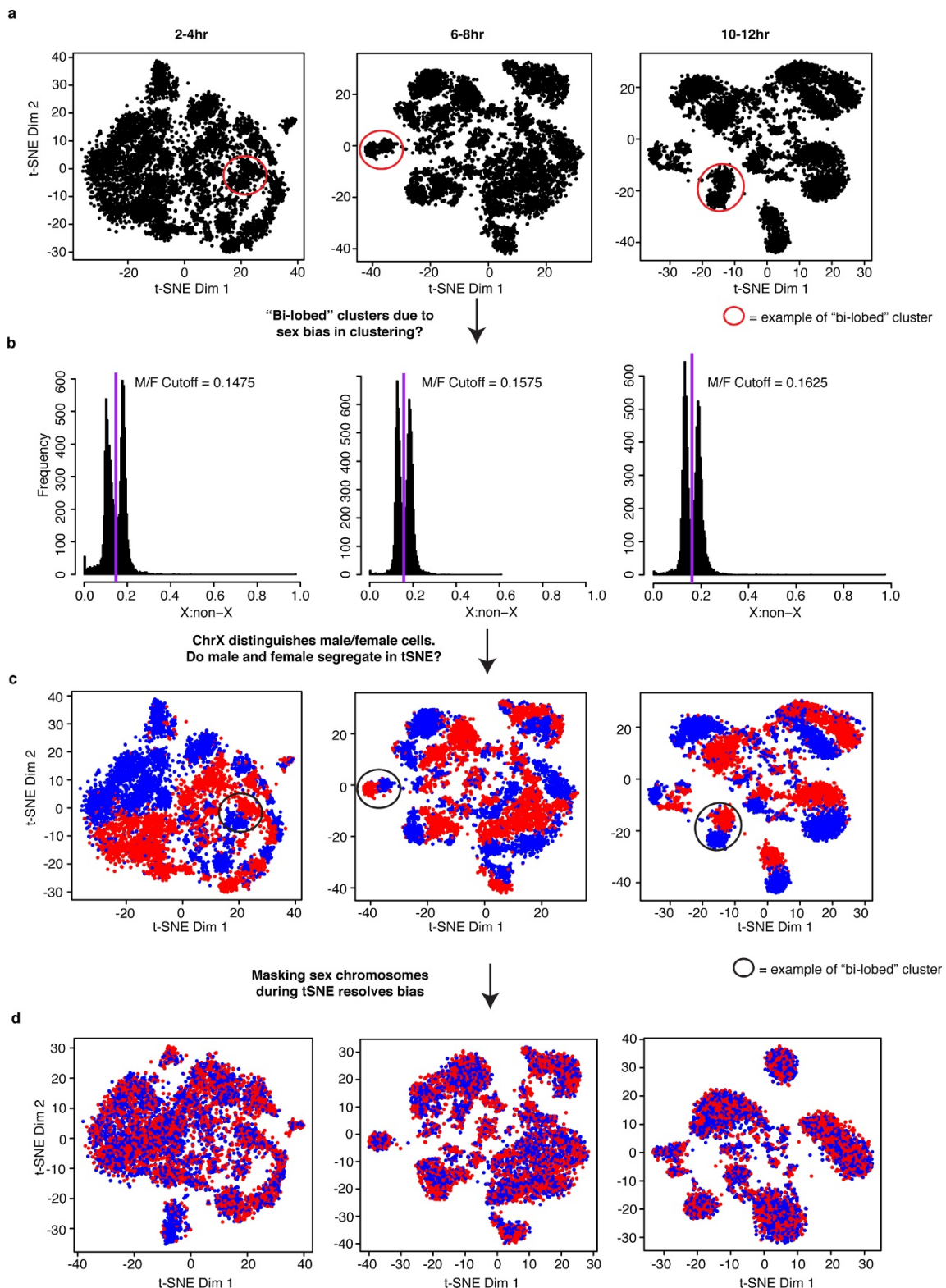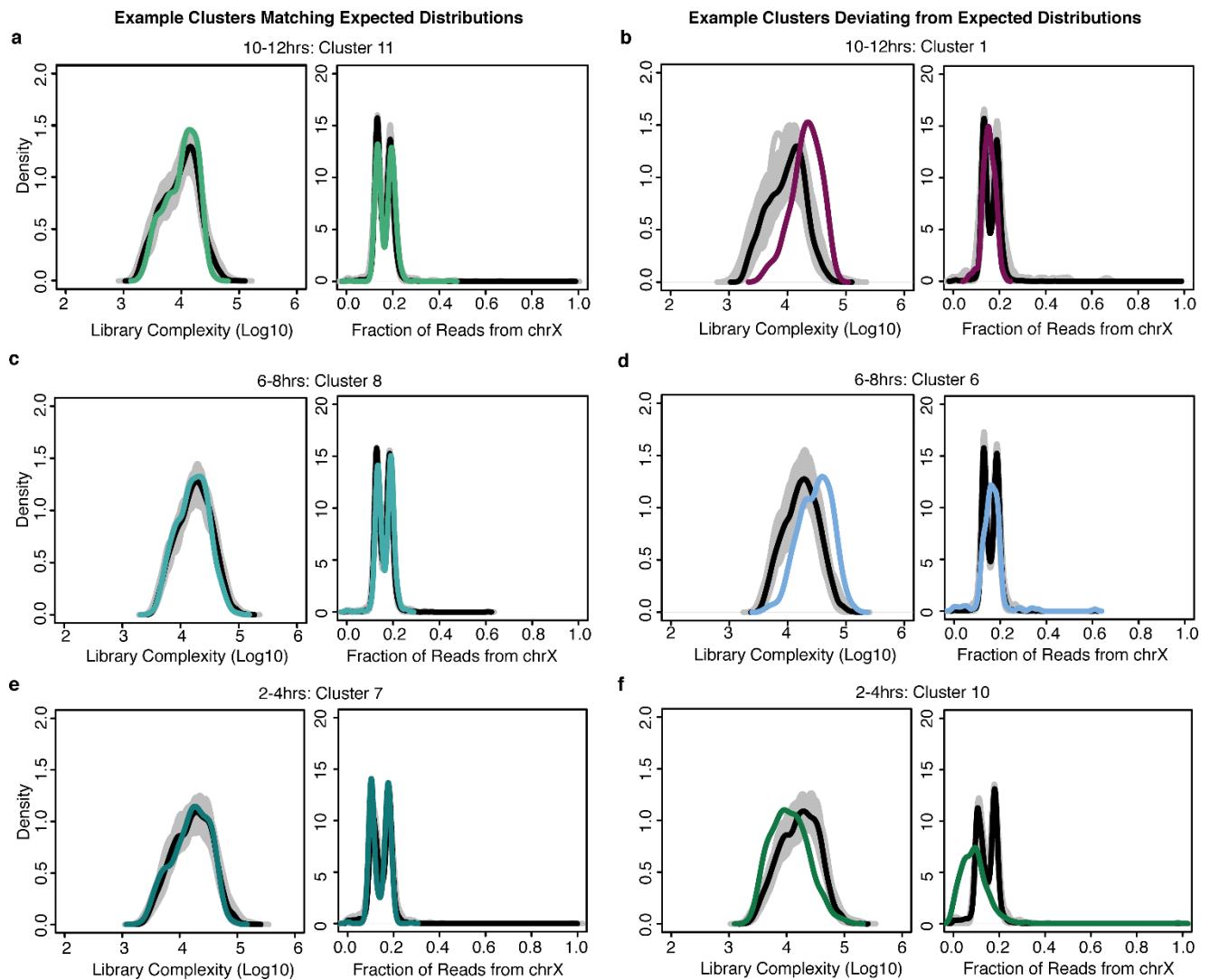
Extended Data Figure 5 | Sex of individual cells identified by ratio of X-chromosome to autosomal reads. Embryos at all stages consist of a mixture of male and female embryos (males, XY; females, XX). **a**, *t*-SNE plots of three time points from analysis in which sex chromosome sites were not excluded. Many clusters exhibited a bi-lobed structure, where each individual cluster was made up of two mirrored lobes (red circles identify one example of a bi-lobed cluster from each time point). This was most apparent at the 10–12-h time point. **b**, Histogram of the ratio of X-chromosome to autosomal reads in individual cells. To explore whether this bi-lobed structure was a function of sex biases in clustering, we attempted to sex individual cells. The ratio of X-chromosome to autosomal reads shows a bimodal distribution, as expected in a system with heterogametic (XY) males and no evidence of imprinting. The purple line marks the local minimum between the two peaks of the histograms. **c**, Initial *t*-SNE clusters coloured according to sex assignment. Red indicates female cells and blue indicates male cells. Colouring individual cells by their sex reveals that the bi-lobed architecture is largely driven by sex biases in clustering. **d**, After removing X-chromosomal reads, data were re-clustered and individual cells were recoloured according to the ratio of X-chromosome to autosomal reads (red, female; blue, male). The resulting clusters showed an approximately equal number of male and female cells except for clusters 1 and 10 at the 2–4-h time point.
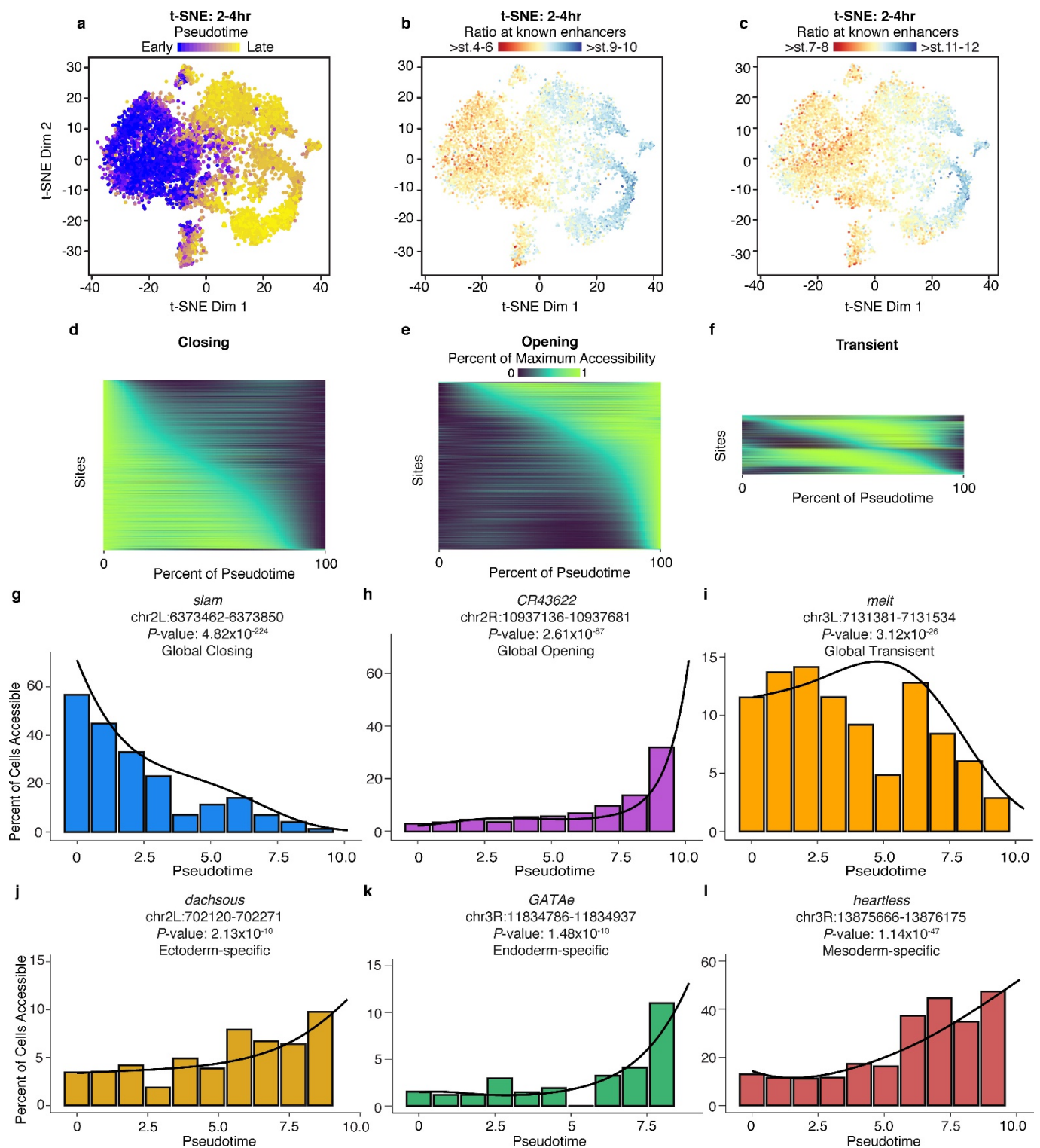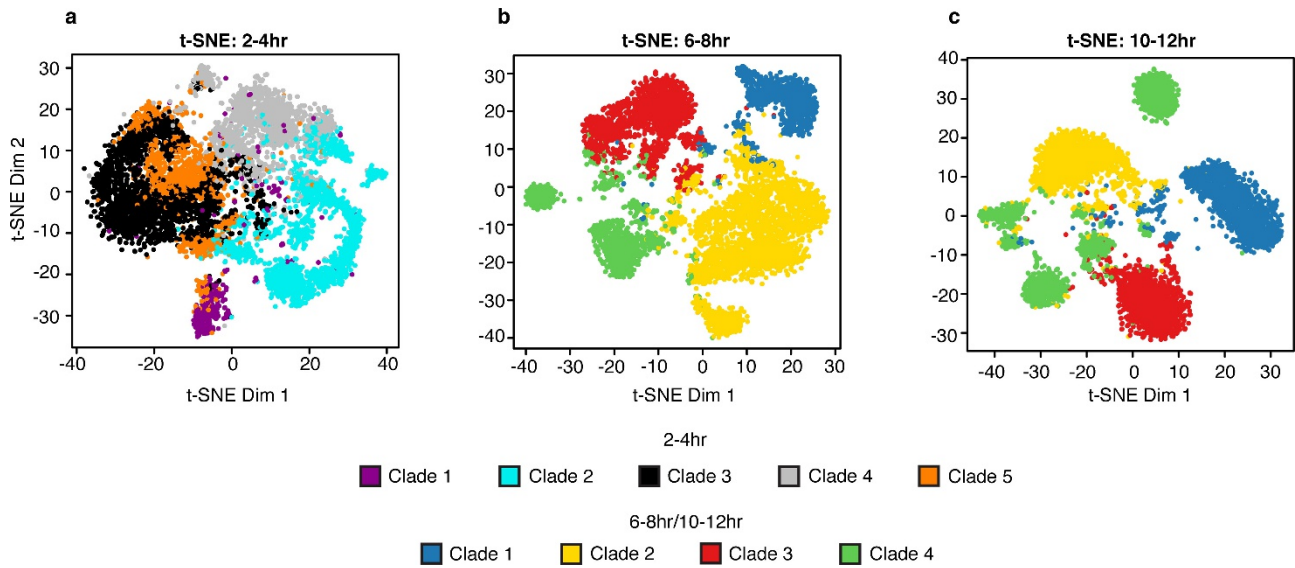
Extended Data Figure 6 | Temporal ordering of cells at 2–4 h using Monocle. a–c, t-SNE maps of cells at 2–4 h with colour representing either the Monocle-inferred pseudotime of each cell (**a**) or the ratio of reads per cell at enhancers active at different stages of development (**b, c**). Read counts within temporally characterized enhancers provide insight into the specific stage of development from which a cell is derived. Plotted here are ratios of counts in earlier versus later active enhancers showing a rough temporal progression from left to right that is also inferred by Monocle. **d–f**, Heat maps of sites that are significantly associated with pseudotime (based on a likelihood ratio test). For each site, a spline was fitted to the data across pseudotime. Sites (rows) were ordered for the heat maps based on the pseudotime at which they first reached half the maximum predicted accessibility from the fit curve. The colours indicate the spline-predicted accessibility across pseudotime, scaled as the fraction of the maximum accessibility for that site. **g–i**, Single-locus plots of the most significant closing, opening and transient sites. Histogram of percentage of cells in which the specified site is accessible in 10 bins across pseudotime, within the 2–4-h time point. The curve is from spline fit for accessibility in cells through pseudotime. **j–l**, As in **g–i**, examples of sites with lineage-specific association with pseudotime. One example of a branch-specific opening site for each germ layer: ectoderm (**j**), endoderm (**k**) and mesoderm (**l**). In **g–i**, P values were calculated for likelihood ratio tests evaluating the effect of progress through pseudotime on accessibility ($n = 100$ bins of cells; see Methods for details). Note that the branch point in pseudotime occurs at approximately 5.6 on the x axis.

Example Clusters Matching Expected Distributions

a 10-12hrs: Cluster 11

c 6-8hrs: Cluster 8

e 2-4hrs: Cluster 7

Example Clusters Deviating from Expected Distributions

b 10-12hrs: Cluster 1

d 6-8hrs: Cluster 6

f 2-4hrs: Cluster 10

Extended Data Figure 7 | Library complexity and fraction of X chromosome reads highlights clusters of collisions between cells from different tissues. Density plots of the estimated library complexity (using the same equation implemented in Picard; left) and the representation of X-chromosome reads (right) in individual clusters. While most of the clusters defined by *t*-SNE are readily biologically interpretable, a small number of clusters (containing relatively few cells) were not easily characterized and are marked by an increase in both estimated library complexity and an unusual distribution of X chromosome to autosomal reads. These clusters are likely to be clusters of collisions; that is, cases in which two or more distinct cells share the same barcode as a consequence of the combinatorial indexing protocol. The black line is the global distribution for all cells in that time point. The grey lines show the results of randomly sampling an equal number of cells to the cluster in question. The coloured line marks the distribution for the cluster being interrogated. **a, c, e,** Most clusters show relatively similar distributions of library complexity (left) and a characteristic, bimodal distribution among cells in the ratio of X-chromosome to autosomal reads (reflecting our use of a pool of male (XY) and female (XX) embryos, right). **b, d,** Putative collision clusters show a clear increase in the average library complexity (left) and a unimodal rather than bimodal distribution of X-chromosome to autosomal reads (right). **f,** These features are not universally diagnostic (for example, cluster 10 at 2–4 h seems to show a strong, bona fide sex bias), but the combination of features is strongly predictive of clusters containing few cells and conflicting biological annotations based on gene or enhancer overlaps.

a t-SNE: 2-4hr

b t-SNE: 6-8hr

c t-SNE: 10-12hr

2-4hr

■ Clade 1 ■ Clade 2 ■ Clade 3 ■ Clade 4 ■ Clade 5

6-8hr/10-12hr

■ Clade 1 ■ Clade 2 ■ Clade 3 ■ Clade 4

Extended Data Figure 8 | LSI-defined clades and *t*-SNE clusters show strong correspondence. *t*-SNE maps of cells from each of the three time points coloured according to the LSI clade to which they were previously assigned (Fig. 1a–c). For the post-gastrulation time points, green is endoderm, red is myogenic mesoderm, yellow is non-neurogenic ectoderm and blue is neurogenic ectoderm. There is strong correspondence between the germ-layer-level clade annotations from the LSI analysis and tissue-specific *t*-SNE clusters, particularly at the post-gastrulation time points (6–8 h and 10–12 h).
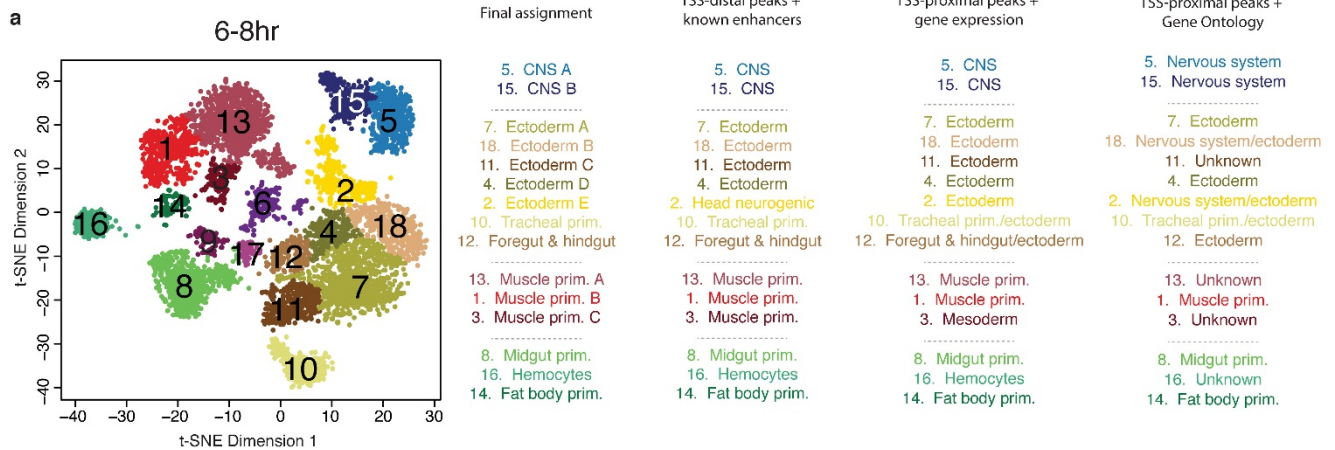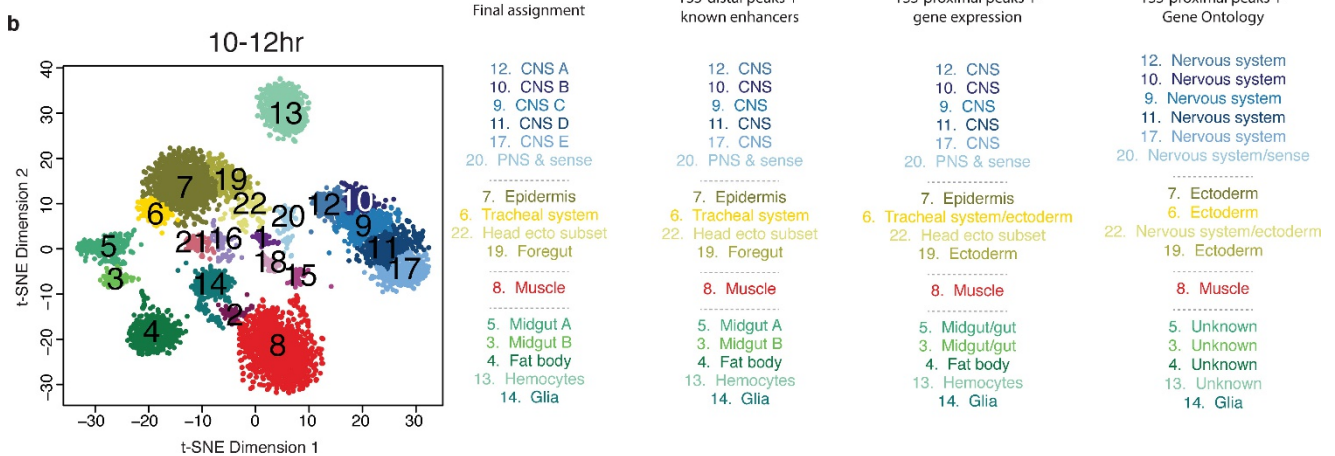
**Extended Data Figure 9 | Cell cluster assignment is similar using either enhancer or gene tissue activity.** For each time point, cell clusters were annotated by first dividing peaks into TSS-distal (putative enhancers) and TSS-proximal (gene promoters) peaks. Each cell cluster was then annotated separately by overlaps between cluster-enriched peaks and: (1) enhancers, comparing the TSS-distal elements to the tissue or cell-type activity of characterized enhancers; (2) genes, comparing TSS-proximal elements to the tissue expression of genes; and (3) Gene Ontology (GO) information (see Methods). Shown are the cluster assignments based on enhancer, gene expression, or Gene Ontology annotation alone. The final assignment, used in the main figures, combines all enrichment information to produce more robust assignments.
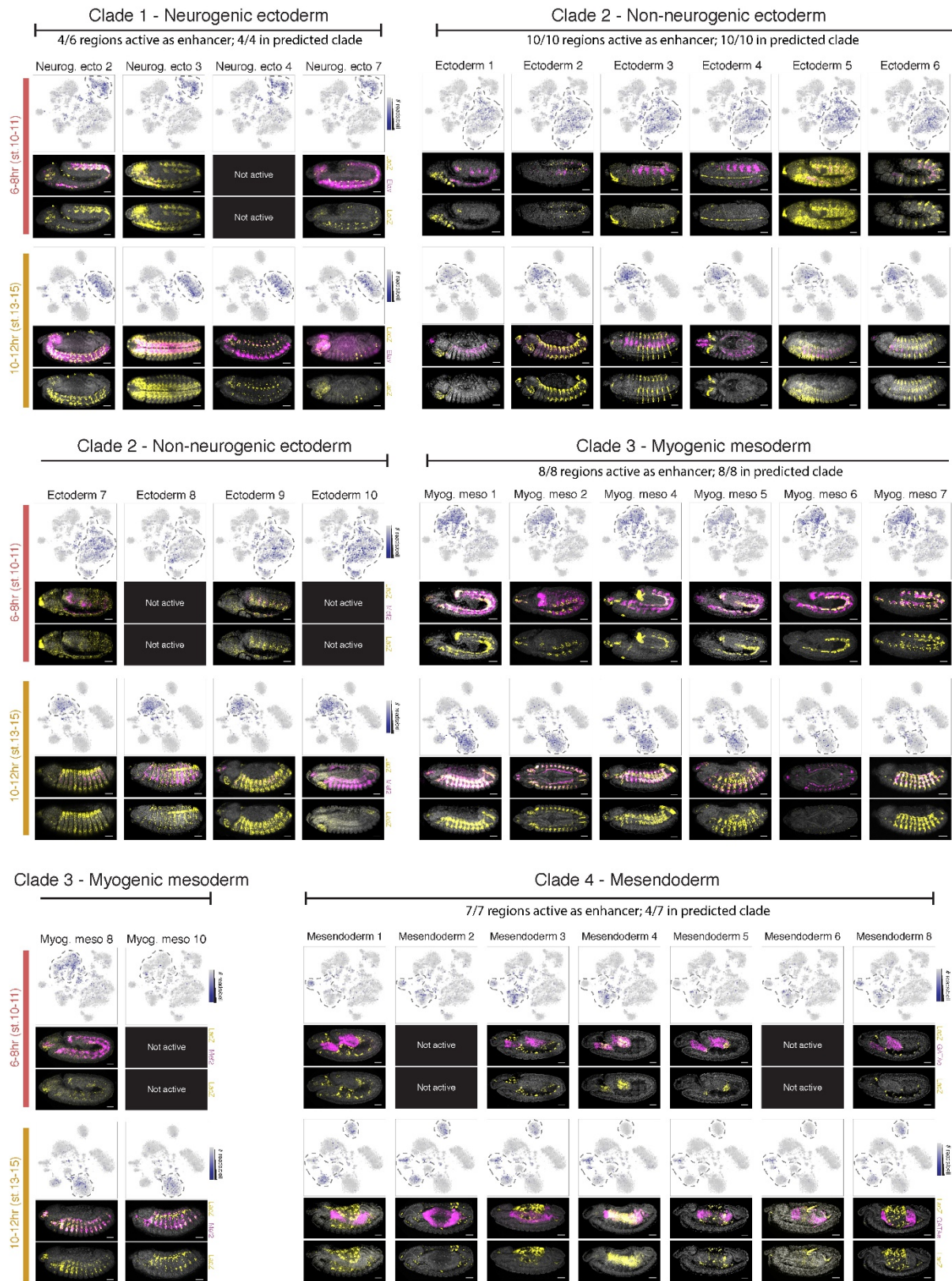
Extended Data Figure 10 | sci-ATAC-seq can predict tissue-specific enhancer usage during development. All candidate clade-specific enhancers tested in transgenic reporters. For each time point, upper panels show single cells visualized by *t*-SNE with the blue intensity representing the number of sci-ATAC-seq reads obtained from each tested element in each individual cell. Cell clusters bounded by dashed lines correspond to the predicted clade of activity. Lower panels show representative embryos for each time point with nuclei stained with DAPI (grey), *in situ* hybridization of the reporter gene driven by the enhancer (yellow) and a tissue marker (magenta). Annotation of each element's activity involved observations across hundreds of embryos. Scale bar, 50 μm.

nature research

Corresponding author(s):   Eileen Furlong & Jay Shendure

☐ Initial submission   ☐ Revised version   ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▶ Experimental design

1. **Sample size**

   Describe how sample size was determined.

   Using our sci-ATAC-seq protocol for collecting chromatin accessibility data on individual cells at high throughput, we collected data on 384 wells of cells for each of three time points. This resulted in data from ~7,000 cells for each time point. As this was the first time this protocol was applied to Drosophila, it was not possible to a priori estimate the sufficient sample size beforehand. However, we note that this sample size met the criteria of both being experimentally manageable and yet still generating one of the larger single cell data sets to date.

2. **Data exclusions**

   Describe any data exclusions.

   There were two reasons for excluding data. (1) We excluded information from cell barcodes that had very few reads associated with them. Looking at the distribution of reads assigned to individual barcodes (Extended Data Figure 1), we noted a bimodal distribution that likely arose from the combination of a normally distributed population of barcodes that represent true cells and a second population near zero that represents background noise barcodes that likely do not represent individual cells. (2) We additionally excluded from analyses the bottom 10% of cells (in terms of feature coverage) to exclude any noisy, low-coverage cells.

3. **Replication**

   Describe whether the experimental findings were reliably reproduced.

   Because we collected data on many thousands of cells we have replicate measurements for cells from each major cell group. We do not present the results of replicate experiments.

4. **Randomization**

   Describe how samples/organisms/participants were allocated into experimental groups.

   Samples from each time point were processed separately and sequenced as a single pool. However, within a time point all cells were processed in parallel in a highly randomized fashion.

5. **Blinding**

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Researchers were not blind to the identity of the time points. However (as noted above), all cells within a time point were processed in parallel.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The <u>exact sample size</u> (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▶ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	bcl2fastq v2.16, trimmomatic v0.32, bowtie2 v2.2.3, samtools v1.1, MACS v2.1.0, BEDTools v2.21.0, deepTools v2.6.0. R and python were used for additional data processing and plotting. The scripts necessary for the primary processing of the raw data will be made available on GitHub, upon request.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▶ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No restrictions.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Mef2 positive cells were stained with a rabbit polyclonal antibody generated at EMBL which was initially described and validated in PMID:16740481 . Elav positive cells were stained with a monoclonal mouse antibody against Drosophila Elav (Elav-9F8A9 from the Developmental Studies Hybridoma Bank) which was initially described in PMID:1716300.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eurkaryotic cell lines were used
b. Describe the method of cell line authentication used.	No eurkaryotic cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.	No eurkaryotic cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eurkaryotic cell lines were used

▸ Animals and human research participants

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

> Mef2 antibodies were generated from rabbits at EMBL in accordance with European Law and EMBL ethical guidelines. Wild-type (CantonS strain) Drosophila melanogaster were reared and collected at EMBL in accordance with standard practice and the ethical standards of the European Research community. Embryo collections (mixed male and female) were performed at 2-4, 6-8, and 10-12 hours after egg laying.

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

> This study did not involve human participants.