High-throughput total RNA sequencing in single cells using VASA-seq

Fredrik Salmen1,2,12, Joachim De Jonghe6,10,12, Tomasz S. Kaminski3,11, Anna Alemany3,1, Guillermo E. Parada4, Joe Verity-Legg1,2, Ayaka Yanagida5, Timo N. Kohler3,6, Nicholas Battich1,2, Floris van den Brekel1,2, Anna L. Ellermann3, Alfonso Martinez Arias7, Jennifer Nichols6,8, Martin Hemberg1,3,9, Florian Hollfelder3,13 and Alexander van Oudenaarden1,2,13

Most methods for single-cell transcriptome sequencing amplify the termini of polyadenylated transcripts, capturing only a small fraction of the total cellular transcriptome. This precludes the detection of many long non-coding, short non-coding and non-polyadenylated protein-coding transcripts and hinders alternative splicing analysis. We, therefore, developed VASA-seq to detect the total transcriptome in single cells, which is enabled by fragmenting and tailing all RNA molecules subsequent to cell lysis. The method is compatible with both plate-based formats and droplet microfluidics. We applied VASA-seq to more than 30,000 single cells in the developing mouse embryo during gastrulation and early organogenesis. Analyzing the dynamics of the total single-cell transcriptome, we discovered cell type markers, many based on non-coding RNA, and performed native splicing (AS) and alternative promoter (AP) usage analyses.

Full-length transcriptome sequencing methods10,11 have enabled AS profiling of polyadenylated RNA species at single-cell resolution12,13,14,15, but the exact quantification of splicing events is hampered by the lack of strand and unique molecular identifier (UMI) information along the whole gene body. Furthermore, neither full-length nor whole-transcriptome methods16,17 have been adapted to high-throughput droplet-based platforms, which offer at least one order-of-magnitude gain in throughput compared to plate-based methods18.

To overcome these challenges, we developed ‘vast transcriptome analysis of single cells by dA-tailing’ (VASA-seq), which captures both non-polyadenylated and polyadenylated transcripts across their length in both plate and droplet microfluidic formats. We first benchmarked VASA-seq against state-of-the-art methods using cultured cells. To our knowledge, VASA-seq is the only technology to combine excellent sensitivity, full-length coverage of total RNA and high throughput. Next, we used VASA-seq to sample more than 30,000 single cells from mouse post-implantation embryos at the following developmental stages: embryonic day (E) 6.5, E7.5, E8.5 and E9.5. Our resource provides a comprehensive analysis of mammalian post-implantation development by characterizing the total transcriptome at single-cell resolution. The analysis revealed layers of biological information that have been absent from recently published resources20–24. Indeed, VASA-seq’s increased sensitivity led to the discovery of several cell-type-specific marker genes and non-polyadenylated histone gene expression patterns, which were used to accurately determine cell cycle stage across tissues. Higher coverage of intronic regions in the full-length VASA-seq dataset led to more accurate RNA velocity measurements25,26 across differentiation trajectories. Finally, we used the full-length coverage to

1Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center, Utrecht, Netherlands. 2Onco&laquo; Institute, Utrecht, Netherlands. 3Department of Biochemistry, University of Cambridge, Cambridge, UK. 4Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. 5Division of Stem Cell Therapy, Center for Stem Cell Biology and Regenerative Medicine, Institute of Medical Science, University of Tokyo, Tokyo, Japan. 6Wellcome Trust - Medical Research Council Stem Cell Institute, University of Cambridge, Jeffrey Cheah Biomedical Centre, Cambridge, UK. 7Systems Bioengineering, DCEXS, Universidade Pompeu Fabra, Doctor Aiguader 88 ICREA (Institució Catalana de Recerca i Estudis Avançats), Barcelona, Spain. 8Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. 9Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women’s Hospital, Boston, MA, USA. 10Present address: Francis Crick Institute, London, UK. 11Present address: Department of Environmental Microbiology and Biotechnology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Warsaw, Poland. 12These authors contributed equally: Fredrik Salmen, Joachim De Jonghe. 13These authors jointly supervised this work: Florian Hollfelder, Alexander van Oudenaarden. 14E-mail: fh111@cam.ac.uk; a.vanoudenaarden@hubrecht.eu

Nature Biotechnology | www.nature.com/naturebiotechnology
Articles

NATURE BiOTECHNOLOGY

= n260, 10x Chromium: = n288). The number of detected annotated genes in HEK293T cells, for each method, is plotted against the number of reads (after quality end for the remainder of UMI-containing reads and toward the 3′ end. Smart-seq3 had a large bias toward the 5′ end for UMI-containing reads and toward the 3′ end for the remainder of the reads, which was also observed with Smart-seq-total (Fig. 1e).

Protein-coding genes were the most highly detected biotype across all methods. However, VASA-plate and VASA-drop both proportionally detected about twice as many long non-coding RNAs (lncRNAs) as 10x Chromium, Smart-seq3 and Smart-seq-total (Extended Data Fig. 2b). Only VASA-seq and Smart-seq-total detected short non-coding RNAs (snRNAs) (1.4% for VASA-plate, 2.5% for VASA-drop and 6.7% for Smart-seq-total), mainly miscellaneous RNA (miscRNA), small nucleolar RNA (snorRNA), ribozymes and small nuclear RNA (snRNA) for VASA-seq and miscRNA and pre-transfer RNA (irRNA) for Smart-seq-total (Extended Data Fig. 2c).

Next, the HEK293T datasets for each method were downsampled to determine the gene detection sensitivity and saturation rates of each method for all annotated genes. VASA-drop showed the highest sensitivity, followed by VASA-plate, with 9,825 ± 280 and 9,480 ± 1,252 (mean ± s.d.) detected genes per cell, respectively, at a sequencing depth of 75,000 trimmed reads per cell. Both exhibited a higher gene detection rate than Smart-seq3 (9,022 ± 1,455 genes per cell) and 10x Chromium (8,342 ± 1,430 genes per cell) and outperformed Smart-seq-total (4,243 ± 512 genes per cell) (Fig. 1f). Similarly, both VASA-seq workflows showed superior detection of protein-coding genes (Extended Data Fig. 2d). For the highest read coverage in our sequenced dataset (750,000 trimmed reads per cell (Extended Data Fig. 2f), only for VASA-plate, Smart-seq3 and Smart-seq-total), VASA-plate and Smart-seq3 showed similar sensitivities (15,248 ± 1,092 and 14,631 ± 988 genes per cell, respectively), whereas Smart-seq-total showed lower sensitivity (7,403 ± 938 genes per cell) (Extended Data Fig. 2e).

Because VASA-seq detects full-length transcripts and larger amounts of unspliced RNA due to the poly(A) tailing of RNA fragments across the transcript length, it can detect nascent transcripts at higher rates than other methodologies. To quantify this,

**Results**

**VASA-seq enables detection of both non-polyadenylated and polyadenylated transcripts in single cells using plates or droplets.** The first step in the VASA-seq protocol entails the fragmentation of RNA molecules from the single-cell lysate followed by end repair and poly(A) tailing, enabling cDNA synthesis from barcoded oligo-dT probes. In addition, a unique fragment identifier (UF1) allows for the accurate quantification of molecules with strand specificity. Barcoded cDNA is amplified using in vitro transcription, and the amplified ribosomal RNA (rRNA) is subsequently depleted. The reads, which was also observed with Smart-seq-total (Fig. 1e).

Protein-coding genes were the most highly detected biotype across all methods. However, VASA-plate and VASA-drop both proportionally detected about twice as many long non-coding RNAs (lncRNAs) as 10x Chromium, Smart-seq3 and Smart-seq-total (Extended Data Fig. 2b). Only VASA-seq and Smart-seq-total detected short non-coding RNAs (snRNAs) (1.4% for VASA-plate, 2.5% for VASA-drop and 6.7% for Smart-seq-total), mainly miscellaneous RNA (miscRNA), small nucleolar RNA (snorRNA), ribozymes and small nuclear RNA (snRNA) for VASA-seq and miscRNA and pre-transfer RNA (irRNA) for Smart-seq-total (Extended Data Fig. 2c).

Next, the HEK293T datasets for each method were downsampled to determine the gene detection sensitivity and saturation rates of each method for all annotated genes. VASA-drop showed the highest sensitivity, followed by VASA-plate, with 9,825 ± 280 and 9,480 ± 1,252 (mean ± s.d.) detected genes per cell, respectively, at a sequencing depth of 75,000 trimmed reads per cell. Both exhibited a higher gene detection rate than Smart-seq3 (9,022 ± 1,455 genes per cell) and 10x Chromium (8,342 ± 1,430 genes per cell) and outperformed Smart-seq-total (4,243 ± 512 genes per cell) (Fig. 1f). Similarly, both VASA-seq workflows showed superior detection of protein-coding genes (Extended Data Fig. 2d). For the highest read coverage in our sequenced dataset (750,000 trimmed reads per cell (Extended Data Fig. 2f), only for VASA-plate, Smart-seq3 and Smart-seq-total), VASA-plate and Smart-seq3 showed similar sensitivities (15,248 ± 1,092 and 14,631 ± 988 genes per cell, respectively), whereas Smart-seq-total showed lower sensitivity (7,403 ± 938 genes per cell) (Extended Data Fig. 2e).

Because VASA-seq detects full-length transcripts and larger amounts of unspliced RNA due to the poly(A) tailing of RNA fragments across the transcript length, it can detect nascent transcripts at higher rates than other methodologies. To quantify this,
RNA fragmentation

End repair & poly(A) tailing

Reverse transcription & second-strand synthesis

In vitro transcription & ribosomal aRNA depletion

Adapter ligation & reverse transcription

Indexing PCR

Number of UFIs assigned to human

Number of UFIs assigned to mouse

Average number of genes per cell

5' -> 3' gene body

Coverage

Reads per cell after trimming
we assigned reads that aligned either to introns or to exon–intron junctions as unspliced, whereas reads that exclusively aligned to exons were considered as spliced. VASA-seq showed the highest proportion of unspliced reads at 44.1 ± 10.1% (VASA-plate) and 56.5 ± 3.1% (VASA-drop) compared to Smart-seq3 (14.8 ± 2.5%), 10x Chromium (17.7 ± 12.8%) and Smart-seq-total (38.1 ± 4.6%) (Extended Data Fig. 2g).

Overall, VASA-seq combines the throughput offered by the 10x Chromium droplet microfluidic platform, the high sensitivity of the Smart-seq3 methodology and the broad-spectrum capture of non-coding RNAs offered by Smart-seq-total in a single experimental workflow. In addition, the method preserves even coverage across the gene body for the unbiased capture of unspliced regions and splicing junctions.

VASA-seq expands the list of cell-type-specific marker genes in the mouse embryo. Next, we used these advantages to extend and improve current atlases of mouse development. We used VASA-drop (referred to as VASA-seq in the remainder of the manuscript) to generate a single-cell total RNA-seq atlas of murine gastrulation and early organogenesis, with a total of 33,662 single cells sequenced from mouse embryonic post-implantation stages E6.5, E7.5, E8.5 and E9.5 (Fig. 2a and Extended Data Fig. 3a). The VASA-seq datasets from post-implantation E6.5, E7.5 and E8.5 were directly compared to a reference dataset generated using the 10x Chromium platform.

Proportionally, VASA-seq detected a slightly lower fraction of protein-coding transcripts, but IncRNAs and transcription factors (TFs) were detected at about 2–3-fold-higher levels, whereas snRNAs were captured only in the VASA-seq dataset (Fig. 2b).

Overall, most genes were identified in both methods across timepoints (70.8–76.2%) (Fig. 2c), but 18.7–25.3% of the genes were detected only in the VASA-seq dataset, whereas a much smaller fraction was observed uniquely in the 10x Chromium dataset (2.4–5.1%).

To explore whether our total scRNA-seq atlas provided more marker genes for different cell types, we identified groups of equivalent cell clusters present in both VASA-seq and 10x Chromium and compared them through differential gene expression analysis, using only the reads that map to the 3′ terminal 20% of the gene bodies in both technologies (Fig. 2d,e and Extended Data Fig. 3b–d). For E8.5 embryos, we identified 43 equivalent clusters shared between the 10x Chromium and the VASA-seq datasets, allowing for systematic differential expression analysis for spliced/unspliced protein-coding transcripts as well as IncRNAs. Overall, VASA-seq detected a higher number of differentially upregulated genes (log, fold change >2 and P < 0.01) for most equivalent comparisons with 10x Chromium (Fig. 2f and Extended Data Fig. 3e,f). Based on previous cell type annotations, examples include the detection of Foxl2os as a paraxial mesoderm progenitor marker, A115009 as a marker for mesenchyme and C130021120Rik as a specific marker for forebrain/midbrain/hindbrain and surface ectoderm (Fig. 2h).

These results demonstrated that VASA-seq could expand the list of known marker genes, especially for unspliced protein-coding and lncRNA genes.

Histone genes as in vivo markers for cycling cells. To further identify global gene signatures intrinsic to VASA-seq, we performed differential gene expression analysis by comparing the mean expression values for all genes across equivalent clusters and timepoints. This analysis identified a subset of genes that were significantly higher expressed in VASA-seq (22 genes; log, fold change >4 and P < 0.001), of which many were canonical histone genes (Fig. 3a and Supplementary Table 3). Consistently, most of the highly differentially expressed genes in the VASA-seq dataset are classified as non-polyadenylated (Fig. 3a).

We reasoned that histone gene expression could be further used to identify cell cycle state, because most canonical histone genes are strongly upregulated during the S-phase. A histogram of total histone gene expression per cell revealed a bimodal distribution for VASA-seq, in contrast to 10x Chromium (Fig. 3b). Detection of S-phase using canonical cell cycle gene expression did not overlap with histone content measurements, illustrating their benefit to confidently assign cell cycle phase in total RNA-seq datasets (Fig. 3c). We further embedded all cells from the different timepoints into a single UMAP and visualized the total expression of histone genes across the dataset (Extended Data Fig. 4a). Cells with high histone expression were clearly segregated in the uniform
**c**  Both VASA 10×

- E6.5 3.6%
- E7.5 2.4%
- E8.5 2.4%
- E9.5 2.4%

**d**  Target technology vs. Reference technology

- High assignment score to purple cluster
- Low assignment score to purple cluster

**e**  UMAP 1 vs. UMAP 2

- Blood
- Ectoderm
- Endoderm
- Mesoderm
- Epiblast

**f**  Differential genes VASA 10×

- Spliced prot vs. Unspliced prot vs. IncRNA

**g**  Merged clusters

- log₂ (VASA/10x)

**h**  Fox2os, A115009, C13002120Rik

- Surface ectoderm
- Paraxial mesoderm progenitor
identified seven cell types that had at least 30 cells in each of three specific cell types across the probed developmental timepoints. We also explored if the percentage of cells in S-phase changed for consistent with results obtained using cell cycle reporter cell lines33. cells, with 20% and 30% of the cells in S-phase, respectively (Fig. 3f), late primitive erythrocytes (84%), whereas node cells and cells from cell types displayed higher proportions of cells in S-phase, such as is specific to the ectoderm germ layer and epiblasts (cell types 12 and 30). Figure 3i, right panel).

In conclusion, VASA-seq detected a high number of histone genes that enabled robust cell cycle and cell-type-specific histone usage determination across the dataset.

**Increased intron coverage with VASA-seq allows for improved RNA velocity estimates.** The large proportion of unspliced transcripts detected with VASA-seq suggested that RNA velocity profiles5, calculated using the ratio of unspliced-to-spliced counts for each gene, could be enhanced using this method. We, therefore, computed the velocities and confidence values using the scVelo package2 in stochastic mode for all cells across all four timepoints (E6.5–E9.5). The velocity vector directions clearly followed the manifold approximation and projection (UMAP) from cells with low histone expression, a feature that was not detected using standard scRNA-seq cell cycle scoring methods (Extended Data Fig. 4a). The bimodal distribution of histone expression in the VASA-seq datasets enabled the classification of cells as being in S-phase (high total histone expression) or non-S-phase (low total histone expression) (Fig. 3d, left panel). Differential gene expression analysis between S-phase and non-S-phase cells was performed for either pooled or separate timepoints, which provided us with an extended list of cell cycle genes co-expressed with histones during mouse embryonic development (Supplementary Table 4).

We then regressed out cell cycle effects by removing the cell cycling genes from our dataset and produced an improved UMAP with reduced cell cycle patterning (Fig. 3d, right panel). We clustered the regressed data using the Leiden algorithm and assigned a cell type annotation to each cluster based on markers obtained through differential gene expression (Fig. 3e, Extended Data Fig. 4b and Supplementary Table 5). Next, we investigated if certain cell types were cycling more frequently. The proportion of cells in S-phase for each cell type in the mouse embryo was 65 ± 11%. However, some cell types displayed higher proportions of cells in S-phase, such as late primitive erythrocytes (84%), whereas node cells and cells from the primitive heart tube (PHT) showed lower proportions of cycling cells, with 20% and 30% of the cells in S-phase, respectively (Fig. 3f), consistent with results obtained using cell cycle reporter cell lines14. We also explored if the percentage of cells in S-phase changed for specific cell types across the probed developmental timepoints. We identified seven cell types that had at least 30 cells in each of three consecutive sampled timepoints: endoderm (cell type 8), allantois (cell type 16), lateral plate mesoderm (cell type 18), trophectoderm (cell type 20), endoderm (cell type 26), visceral endoderm (cell type 27) and outflow tract (cell type 34). In this subset, only the trophectoderm showed unaltered proportions of cells in S-phase from E6.5 to E8.5 (Fig. 3g, left panel). The other six cell types showed a reduction in the number of cells in S-phase across timepoints, with the allantois showing the most striking decrease from 79% to 38% between E7.5 and E9.5 (Fig. 3g, right panel).

Additionally, we performed differential histone gene expression analysis between cell types (Supplementary Table 6). Because histones from the same family (H1, H2a, H2b, H3 and H4) have extensive sequence similarity, not all reads could be uniquely assigned to a single histone gene. We found ten single-annotated (Fig. 3h) and 14 multi-annotated (Extended Data Fig. 4c) genes significantly upregulated in at least one cell type (log fold change > 2; P < 0.01). Some histone genes showed germ layer and/or cell-type-specific expression. For example, H2aw was upregulated in the ectoderm. H2bc15 was ubiquitously expressed in most cell types but absent in the node (cell type 40) and the visceral endoderm (cell type 27) (Fig. 3i, left panel). H2bc1 expression was detected only in epiblast at E6.5 (cell type 30) (Fig. 3i, middle panel). H2bua2 displayed specific gene expression in the ectoderm germ layer and epiblast (cell types 12 and 30) (Fig. 3i, right panel).
H3c11, H3c6, H3c7, H4c1, H4c8, Kmt2d, Notch2, H4c12, H2ac12, H4c11

% of cells in S-phase

Cell types

Node (40) Primitive heart tube (33) Late primitive erythrocytes (1)

Assigned phase using cell cycle scoring

Histone expression

Counts

Cell type/germ layers

H2bc1, H2bc15, H2bu2

Histone genes

Cell type group/germ layers

H2bc7, Hist1h2br, H190, H3c13, H2bc15, Hist1h2bp, H2aw, H2bu2, H116, H2bc1

% of cells in S-phase

Timepoint

E6.5, E7.5, E8.5, E9.5

log2(mean expression + 1)

Blood, Ectoderm, Epiblast, Endoderm, Mesoderm, Node, Trophoblast
consecutive timepoints and cell type progression in the UMAP, recapitulating previously characterized trajectories in the developing mouse embryo (Fig. 4a). To contrast with the equivalent 10x Chromium dataset, we repeated the analysis for both datasets using the E6.5, E7.5 and E8.5 timepoints. The RNA velocity vectors for VASA-seq had higher confidence metrics overall (0.84 ± 0.12) compared to the 10x dataset (0.65 ± 0.12). This indicates that increased intronic capture with VASA-seq improves RNA velocity measurements.
Fig. 5 | Landscape of AS events across cell types involved in gastrulation and early organogenesis. a, Schematic representation of the different types of AS nodes categorized using Whippet. Colors highlight the exonic parts that correspond to the indicated splicing nodes. Arrows indicate different possible splice junctions that can be connected to the nodes. Red arrows represent inclusion events; black arrows represent exclusion events. b, PAGA projecting cell types onto a force field revealing connectivity between cell clusters. Edge width represents the connectivity values between cell clusters. Selected cell types are highlighted by color. c, Pairwise comparisons that exhibited the highest number of DISNs are represented by a dot plot, where the position of each point on the x axis indicates the cell clusters that were involved in each comparison (P1–P15) across the types are highlighted by color.

each point on the x axis indicates the cell clusters that were involved in each comparison (P1–P15) across the types are highlighted by color. Cell clusters

Cluster class
Brain Cardiac Endoderm Epiblast Erythroid Extra-embryonic Mesoderm Neural crest Other Spinal cord

Cluster connectivity

PBI (ψ) = Inclusion + Exclusion

Inclusion size

Set intersection

Erythropoiesis Primitive erythroid progenitors Endothelium Allantois Mesendodermal unknown Secondary heart field Myoblasts Lateral plate mesoderm Primitive heart tube First heart field Pharyngeal arches I Pharyngeal arches II Primitive endoderm Intermediated mesoderm I Cranial mesoderm Early caudal endoderm Spinal cord (differentiated neurons) Spinal cord (cycling progenitors) Rhombomeres (hindbrain) Forebrain/hindbrain (lateral) Forebrain Neural crest Anterior neural crest Caudal epithelium (E7.5) Epiblast (E6.5) Prepharyngeal/pedal region Endoderm Dorsal surface ectoderm Outflow tract Trophoblast Epiblast and visceral endoderm Nodal Pronephric

Set size
0 50 100

Intersection size
0 25 50 75

Cell clusters
AA AD AF AL BS CE RI TE TS

PBI
1.00 0.75 0.50 0.25 0.00

Cell clusters

Set intersection

Erythropoiesis Primitive erythroid progenitors Epiblast and visceral endoderm Nodal Pronephric
compared to 10x Chromium (0.65 ± 0.12) (Fig. 4b), highlighting higher average correlation of the velocity vectors for a given cell and its neighbors. Next, we extracted the number of genes that contributed significantly to the RNA velocity vectors. We found that most significant genes were shared between the methods (1,492). However, VASA-seq detected a large number of additional genes (1,069) that contributed to the RNA velocity vector (Fig. 4c and Supplementary Table 7). For the genes that were shared between both methods, we quantified the goodness of fit ($r^2$) to the gene phase diagrams with the prediction made by scVelo (Fig. 4d). The stochastic model of the scVelo package fitted better to VASA-seq in terms of goodness of fit (0.74 ± 0.18) compared to the 10x Chromium data (0.38 ± 0.25). Examples of genes with an $r^2$ about 1.s.d. above average for both VASA-seq and 10x Chromium are shown in Extended Data Fig. 5a. To determine whether these measurements would enable a more accurate trajectory prediction across our atlas, the velocity vectors from the 10x Chromium dataset were projected on the UMAP spanning the developmental timepoints E6.5, E7.5 and E8.5. This analysis revealed a discrepant trajectory across blood maturation (Fig. 4e) that was not observed in our dataset (Fig. 4a). Latent time predictions using scVelo's dynamical modeling on the blood cell types across E7.5 and E8.5 further highlighted trajectory inconsistencies for the 10x Chromium dataset (Fig. 4f,g), which has previously been associated with confounding effects from multiple rate kinetics genes in the overlapping first and second blood waves. These observations were not replicated with VASA-seq, which accurately reported on blood maturation across physically sampled timepoints (Fig. 4h). These findings highlight the benefits of more sensitive RNA velocity measurements using VASA-seq to agnostically identify trajectories across cell types. Based on the capture of non-coding species across their gene body using VASA-seq, lncRNA kinetics across tissues can be determined. For example, the endothelium showed (1) the induction of Hoxa11or in the yolk sac at E7.5 and E8.5; (2) the induction of Gms50321 at E7.5 and split induction and repression at E8.5 and E9.5; and (3) the induction of D03007L05Rik at E7.5 and progressive repression across E8.5 and E9.5 (Extended Data Fig. 5c,d). These observations could not be replicated in the 10x Chromium dataset because unspliced molecules for these lncRNAs could not be detected. Therefore, VASA-seq showed better reconstruction of RNA velocity vectors guiding differentiation trajectories and identification of novel gene expression dynamics.

**Comprehensive profiling of AS across mouse gastrulation and early organogenesis.** The ability to profile full-length transcripts at scale using VASA-seq allows for the identification of AS patterns across cell types by quantifying the inclusion rates of non-overlapping exonic parts, herein referred to as ‘splicing nodes’. Every splicing node is associated with different types of AS, alternative transcriptional start sites or alternative polyadenylation events, and their inclusion rates are calculated as percent-spliced-in (ψ) values, which is quantified by taking the ratio of reads that support the inclusion of a given splicing node (Fig. 5a). To quantify AS patterns, we used Whippet, a computationally lightweight and accurate quantification method, previously integrated in computational workflows dedicated to process scRNA-seq data. Because splicing node coverage is limited at the single-cell level (Extended Data Fig. 5d), we implemented a pseudo-bulk pooling approach, developed as part of MicroExonator, where reads from the same cell type are pooled in silico before differential splicing node quantification. Pooling reads into pseudo-bulks from each cell type substantially increased our ability to quantify splicing nodes (Extended Data Fig. 5e–g). To detect differentially included splicing nodes (DISNs) across cell types, we implemented a method developed as part of MicroExonator to detect robust AS changes across pairwise comparisons of closely related cell types. For this purpose, we computed the k-nearest neighbor connectivity values across cell types to generate a coarse-grained graph with partition-based graph abstraction (PAGA) (Fig. 5b). This enabled us to compute 72 pairwise comparisons between related cell types, from which we identified a total of 979 DISNs (Supplementary Table 8). We found that 45.8% of DISNs were core exon (CE) nodes, which correspond to cassette exons involved in exon skipping, the most abundant type of AS event across vertebrates.

To further investigate the AS profile across cell types, we focused on our analyses on the 15 pairwise comparisons that detected the highest amount of DISNs, which accounted for 67.6% of the total. These comparisons were overrepresented with cell types involved in heart morphogenesis, early gastrulation, extra-embryonic tissues and blood development, showing widespread involvement of AS in these processes. Again, CE was the most abundant type across DISNs, except for comparison 7 (trophectoderm (cell type 20) versus parietal endoderm (cell type 37)), where 40.7% of the DISNs were classified as retained intron (RI) (Fig. 5c,d). Further intersection analyses between the set of DISNs revealed differential splicing nodes that were recurrently detected across different comparisons. The biggest set of common DISNs was found across comparisons that share cell clusters, such as P1/P3/P6 or P6/P13/P14, which all correspond to cell types involved in heart development (Fig. 5e). However, 68.7% of the identified DISNs were found exclusively across individual pairwise comparisons, suggesting a prevalence of AS events that are specific to certain differentiation transitions.

To gain further insights of global splicing patterns in relation to cell types, we identified splicing nodes with $\psi$ values that strongly

---

**Fig. 6 | AS patterns across heart morphogenesis and blood formation.** a, Schematic of mouse heart development. b, Volcano plot illustrating the DISNs detected between the PHT (positive ΔΨ values) and the FH (negative ΔΨ values). c, Sashimi plot of Tpm1 showing a coordinated AS switch from smooth muscle to striated muscle conformation after heart development from ECE to PHT. Box annotation on top illustrates exon order for Tpm1. Color-coding indicates the splicing node. Line in the bottom single-cell ψ plots across cell types delineates the mean $\psi$ value, and the shading indicates the 95% confidence interval (CD). d, Single-cell gene expression UMAP plot for Tpm1 (top left) and single-cell $\psi$ projection for Tpm1L14, Tpm1L22 and Tpm1L25 across the global splicing analysis. Each node is color-coded and highlighted in a single-cell line plot representing single-cell $\psi$ values across all three clusters. e, Schematic representation of murine blood development through the profiled timepoints. f, Volcano plot illustrating the DISNs detected in the pairwise comparison between erythrocytes at E7.5 and E9.5. Color-coding indicates proteins with calmodulin-binding and/or spectrin-binding domains or RNA splicing proteins as determined by GO analysis. NS annotation stands for non-significant (gray color). g, Sashimi, domain annotation and line plots representing the skipping of exon 16 (Epb41_30) between E7.5 and E9.5. Line in the bottom single-cell $\psi$ plots across timepoints delineates the mean $\psi$ value at each timepoint, and the shading indicates the 95% CI. h, Sashimi, domain annotation and line plots representing the inclusion of Add1_37 leading to a premature stop codon inclusion at E9.5 removing the C-terminus calmodulin-binding domain. Line in the bottom single-cell $\psi$ plots across timepoints delineates the mean $\psi$ value at each timepoint, and the shading indicates the 95% CI. i, Sashimi, domain annotation and line plots representing the gradual exclusion of the AnkL_43 microexon in a disordered domain. Line in the bottom single-cell $\psi$ plots across timepoints delineates the mean $\psi$ value at each timepoint, and the shading indicates the 95% CI.
deviated from the rest of the cell types and denoted these as splice node markers (SNMs) (Fig. 5f). In total, we identified 996 SNMs (Supplementary Table 9), 27.7% of which were also detected as DISNs (Extended Data Fig. 5g). In agreement with our previous analyses, we detected an elevated number of SNMs for cell types involved in heart development and early gastrulation. Among all the cell types, the PHT (cell type 33) had the most divergent splicing patterns and featured the highest number of SNMs (263), both included and excluded (Fig. 5f). Moreover, we found 132 SNMs for the second heart field (cell type 35), supporting the observation of extensive AS activity during heart morphogenesis. Extra-embryonic cell types that were sampled in the earlier timepoints mainly (E6.5 and E7.5), such as the trophectoderm (cell type 20), parietal endoderm (cell type 37) and visceral endoderm (cell type 27), also exhibited a higher-than-average proportion of SNMs (62, 56 and 132, respectively).
Taken together, we show that sequencing transcripts across their length with high cellular coverage using VASA-seq enabled the identification of extensive AS patterns during mouse development.

AS analysis of blood and heart-related cell types. Across all cell types, the PHT showed considerable AS signatures compared to the first heart field (FHF) (comparison 1; Fig. 5c–f). These changes occur while the heart undergoes extensive morphogenesis with the formation of the cardiac crescent, consisting of the FHF and second heart field (SHF) at E7.5, which subsequently re-arranges to form the PHT at E8.0 (ref. 22) (Fig. 6a).

The detected splicing events for comparison 1 (Supplementary Table 10) were coordinated with the differential expression of heart-specific RNA-binding proteins (RBPs), such as Ptbp1 (Extended Data Fig. 6a), that are likely orchestrating the observed AS events20,21. In addition to changes in gene expression for RBPs, a pair of mutually exclusive exons for Rbfox2 (Rbfox2_143/144, commonly referred to as B40 and M43 in the literature) were among the most significant DISNs identified in the FHF to PHT comparison (Fig. 6b and Extended Data Fig. 6bc). Our results showed that B40 and M43 were preferentially included in FHF and PHT cells, respectively, which is in line with previous findings42. In addition, tropomyosin 1 (Tpm1) stood out with three DISNs, a CE and a pair of mutually exclusive exons, which had some of the highest confidence levels detected (Tpm1_29, Tpm1_22 and Tpm1_25 corresponding to exons 9a, 6a and 6b, respectively). These splicing events are part of a coordinated transition between a smooth muscle and striated muscle program, orchestrated by PTBp1 and RBFOX2 (ref. 22). This transition was captured along a differentiation trajectory encompassing the early caudal epiblast (ECE), the FHF and PHT (Fig. 6c and Extended Data Fig. 5d), which also highlighted a switch for the N- (Tpm1_14, exon 1b) and C- (Tpm1_32, exon 9b) terminal that modulate the protein's interaction with actin and troponin.54−61. Because Tpm1 has many cell-type-specific isoforms,62 we further visualized single-cell UMI values for the aforementioned splicing nodes on the UMAP, which showed cell-type-specific patterns across the atlas (Fig. 6d and Extended Data Fig. 6ef).

At E7.25, primitive erythrocytes emerge from the blood islands in the yolk sac and enter the bloodstream at E9.0 (ref. 63) (Fig. 6e). The erythrocyte cytoskeleton then undergoes gradual rearrangements that increase their deformability when circulating in the narrow network of fetal vasculature, a change catalyzed by the adoption of the erythrocyte-specific transmembrane spectrin–actin backbone.50,64 To determine if we could identify DISNs that mediate such rearrangements, we performed pairwise differential splicing analysis between erythroid cells from E7.5 (early progenitors, primitive erythroblasts) and E9.5 (early differentiating proerythroblasts, ProE) (Extended Data Fig. 6g). The analysis uncovered 210 DISNs that showed an enrichment for Gene Ontology (GO) terms relating to spectrin (GO:00030507; false discovery rate (FDR) = 4.8 × 10−3) and calmodulin (GO:0005516; FDR = 7.05 × 10−4). For example, Mbnl1 (ref. 65), showed a skipped exon (Mbnl1_37) encoding for a nuclear localization signal (Fig. 6d and Extended Data Fig. 6hi). Nuclear localization signal skipping of this exon leads to its localization in the nucleus and cytoplasm rather than exclusively in the nucleus,66 likely affecting the spectrum of AS events depicted across early erythroid progenitor differentiation.

These results show that VASA-seq can inform on cell-type-specific gene function by measuring AS across cell types.

Discussion

VASA-seq is a novel technology that enables the sequencing of the total transcriptome from single cells. The protocol demonstrates best-in-class RNA capture efficiency and provides full-length coverage of coding sequences and enriches for non-coding RNA bio-types. In our datasets, the latter were detected at much higher levels compared to current state-of-the-art methods: 10x Chromium and Smart-seq3 (ref. 13). VASA-seq also outperformed Smart-seq-total67 in terms of scalability, sensitivity, balance in gene body coverage and lncRNA detection. However, Smart-seq-total may complement our approach for the study of specific sncRNAs. VASA-seq does not rely on random priming, which has been shown to induce sequence-specific biases in transcriptome composition.58 Fragment quantification is also ameliorated, as it employs UMI/IFI tagging across the whole gene body, and the reads retains strand specificity, which improves the quantification of overlapping transcripts69.

The excellent performance of the method was maintained for both plate-based (VASA-plate) and droplet-based (VASA-drop) formats in our benchmarking effort. However, a discrepancy arose for sncRNAs and unspliced molecules, which were detected at lower levels in VASA-plate compared to VASA-drop, possibly due to inefficient nuclei lysis in the plate experiment or different length exclusion during the DNA purification steps. On the other hand, rRNAs were not depleted as efficiently in the VASA-drop datasets, maybe because the increased barcode length for the method decreases the ability to exclude short ribosomal fragments that remain after depletion using DNA purification methods. Nevertheless, rRNA depletion in both VASA-seq methods outperformed Smart-seq-total. For integration with datasets generated with 3′ capture methods, we recommend using the 20% terminal fragments of gene bodies to generate a shared embedding.

The throughput of the method is an order of magnitude larger than previously described total RNA-seq methodologies70−73. This allowed us to generate a large-scale total-RNA-seq atlas to profile mouse gastrulation and early organogenesis. The high sensitivity and increased coverage of non-coding RNA molecules enabled us to expand the current list of cell-type-specific markers that will complement previous findings20−24. We further provide a detailed map of cell-type-specific AS events encompassing mouse development from E6.5 to E9.5, which underlined the predominance of alternative cassette exon usage throughout the timepoints investigated. Our resource provides a comprehensive analysis of AS during post-implantation mammalian development.

Furthermore, VASA-seq enables the accurate estimation of cell cycle stage from direct measurements of histone content. Because most histone genes are non-polyadenylated74 and because canonical histone expression is a marker for S-phase75, VASA-seq outperformed previous cell cycle scoring methods based solely on polyadenylated marker expression in our dataset. This is especially useful to determine cell cycling across developmental phases or between different
populations of cells. The workflow also enables effective removal of cell cycle effects on profiled transcriptomes, which is important for cell type classification and unbiased analysis.

Because VASA-seq captures RNA molecules across their entire length, RNA velocity predictions were ameliorated. This offers a resource for further explorations that go beyond transcriptional kinetics, such as the detection of splicing dynamics across developmental trajectories.

The modularity afforded by the microfluidic workflow will expand the number of single-cell assays that can be performed at high throughput. Indeed, consecutive injections of reaction mixes in droplets enables multi-step processes will benefit complex multi-omic workflows. Moreover, lower reagent costs due to smaller volumes, associated with droplet miniaturization, and the lack of reliance on commercial kits for the VASA-drop workflow will enable inexpensive, large-scale, in-depth transcriptomic profiling at a cost of approximately $0.11 USD per cell for sequencing-ready libraries compared to $0.5 USD per cell for the 10x Chromium v2 kit56. The VASA-plate method has a library preparation cost of $0.98 USD, which is similar to the estimated range between $0.57 USD and $1.14 USD per cell for Smart-seq3 (ref. 27) (Supplementary Table 13).

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01361-8.

Received: 20 November 2021; Accepted: 13 May 2022; Published online: 27 June 2022

References
1. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. Cell Rep. 2, 666–673 (2012).
2. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21, 1160–1167 (2011).
3. Ramsköld, D. et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 30, 777–782 (2012).
4. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382 (2009).
5. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015).
6. Macosko, E. Z. et al. High-parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).
7. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 8, 14049 (2017).
8. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251–255 (2015).
9. Jainin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776–779 (2014).
10. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240 (2013).
11. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015).
12. Hagemann-Jensen, M. et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq2. Nat. Biotechnol. 38, 708–714 (2020).
13. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9, 171–181 (2014).
14. Feng, H. et al. Complexity and graded regulation of neuronal cell-type-specific alternative splicing revealed by single-cell RNA sequencing. Proc. Natl Acad. Sci. USA 118, e203056118 (2021).
15. Lukaczovych, D. et al. Single-cell RNA-seq reveals developmental origins and ontogenetic stability of neurexin alternative splicing profiles. Cell Rep. 27, 3752–3759 (2019).
16. Hayashi, T. et al. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat. Commun. 9, 619 (2018).
17. Verboom, K. et al. SMaRTer single cell total RNA sequencing. Nucleic Acids Res. 47, e93–e93 (2019).
50. Zhou, J., Zhao, S. & Dunker, A. K. Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. *J. Mol. Biol.* **430**, 2342–2359 (2018).

51. Cheng, A. W. et al. Muscleblind-like 1 (*Mbnl1*) regulates pre-mRNA alternative splicing during terminal erythropoiesis. *Blood* **124**, 598–610 (2014).

52. Gates, D. P., Coonrod, L. A. & Berglund, J. A. Autoregulated splicing of muscleblind-like 1 (*MBNL1*) pre-mRNA. *J. Biol. Chem.* **286**, 34224–34233 (2011).

53. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).

54. Zhao, S. et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics* **16**, 675 (2015).

55. El Kennani, S. et al. MS_HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones. *Epigenetics Chromatin* **10**, 2 (2017).

56. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

57. Salomon, R. et al. Droplet-based single cell RNAseq tools: a practical guide. *Lab Chip* **19**, 1706–1727 (2019).

58. Zhang, X. et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *Mol. Cell* **73**, 130–142 (2019).

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/). © The Author(s) 2022
Cell lines. HEK293T cells were passaged every second day and cultured in T75 flasks. The culture media was DMEM-F12 (Thermo Fisher Scientific) supplemented with 10% heat-inactivated FBS (Thermo Fisher Scientific) and 100 μM l-phenylalanine–streptomycin (Thermo Fisher Scientific). For passaging, the cells were washed with 10 ml of ice-cold 1× PBS (Lonza) twice. Then, 9 ml of PBS was added to the flask, and cells were detached by adding 1 ml of 10× trypsin-EDTA (Sigma-Aldrich) and incubated at 37 °C for 5 minutes. Trypsin-EDTA was then inactivated with 1.5 ml of DMEM-F12 + 10% FBS and incubated at 37 °C for 5 minutes. The cells were then pelleted at 300 g for 3 minutes, and the supernatant was aspirated. After aspiration of the supernatant, the cells were washed twice in PBS and viability-assessed and counted before encapsulation.

mESCs were passaged every other day and cultured in 2i+LIF medium. In brief, DMEM/F-12 nutrient mixture without L-glutamine (Thermo Fisher Scientific) and neurobasal medium without L-glutamine (Thermo Fisher Scientific) in a 1:1 ratio, 0.1% sodium bicarbonate (Thermo Fisher Scientific), 0.11% bovine albumin fraction V solution (Thermo Fisher Scientific), 0.5× B-27 supplement (Invitrogen), 1× N-2 supplement (Cambridge Stem Cell Institute, made in-house), 50 μM 2-mercaptoethanol (Thermo Fisher Scientific), 2 mM L-glutamine (Thermo Fisher Scientific), 100 μM l-phenylalanine–streptomycin (Thermo Fisher Scientific), 12.5 μg/ml of insulin zinc (Thermo Fisher Scientific), 0.2 μg/ml of mLIF (Cambridge Stem Cell Institute), 3 μM CHIR99021 (Cambridge Stem Cell Institute) and 1 μM PD0325901 (Cambridge Stem Cell Institute). Culture dishes were coated with 0.1% gelatine in PBS for at least 30 minutes. Cells were detached with 500 μl per six-well of Accutase (Merck) for 3 minutes at 37 °C. The detached cells were transferred into 9.5 ml of washing medium (DMEM/F-12 with 1% bovine albumin fraction V solution) and centrifuged at 300 g for 3 minutes. The supernatant was aspirated, and the cell pellet was resuspended in 2i+LIF medium and re-plated at 80,000 cells per six-well. For the encapsulation process, the cells were washed twice in PBS, viability-assessed and counted before dilution to the correct concentration.

Murine embryo collection and dissociation. Pregnancy C57BL/6 female mice were purchased from Charles River Laboratories or obtained from natural mating of C57BL/6 mice in-house. Mice were maintained on a lighting regimen of 12-hour light/dark cycle with food and water supplied ad libitum. Detection of a copulation plug after natural mating indicated E9.5. After euthanasia of the females using cervical dislocation, the uteri were collected into PBS (Lonza) with 0.1% pronase (Sigma-Aldrich) and 0.2% IGEPAL CA-630 (Sigma-Aldrich) were added to each well. Plates were sealed and spun down at 2,000 r.c.f. for 2 minutes. Reverse and tailing was carried out at 37 °C for 1 hour. Plates were snap-chilled, spun down at 2,000 r.c.f. for 1 minute and stored on ice before next dispensation.

VASA: Selection of the droplet generation device. The droplet generation device for compressible barcoded bead and single-cell co-encapsulation (Extended Data Fig. 1c) was modified from previous designs16. The flow-focusing junction (80 μm) was narrowed to generate smaller droplets (0.5-5 μl) at high throughput (115 Hz).

VASA: Design of droplet picoinjection devices. The design of both droplet picoinjector devices is based on the findings of a previous study17. Several key features were added to the architecture of previous designs to ameliorate the robustness of the injections in large droplets containing compressible barcoded beads:

1. Emulsion-diluting oil inlet, number 2 (Extended Data Fig. 1d), which reduces the packing of the emulsion to eliminate fragmentation of densely packed droplets before being re-injected in the picoinjection channel. This design feature allows for packed droplets to arrange into an evenly spaced monolayer that reduces fluctuations in volume of droplets after picoinjection.
2. Smooth narrowing of the re-injection chamber facilitating the ordering of droplets before spacing, which reduced droplet break-up.
3. Deepening of the outlet junction between the outlet, number 5 (Extended Data Fig. 1d,e, deep blue color), which stabilizes droplets and reduces droplet merging, which was observed during the rapid transition from the shallow microfluidic channel to a wide tubing or collection tip.

VASA: photolithography of microfluidic molds. The channel layout for the microfluidic chips was designed using AutoCAD (Autodesk) and printed out on a high-resolution film photomask (Micro Lithography Services). The designs in Extended Data Fig. 1 are deposited on https://openwetware.org/wiki/DropBase:Devices and can be found in the supplementary file ‘VASAdrop CAD designs.smaks.dsX’. The microfluidic devices were fabricated following standard hard and soft lithography protocols that can be performed in local cleanrooms or outsourced to contract manufacturing companies. First, microfluidic molds were patterned on a 3-inch silicon wafer (MicroChemicals) using high-resolution film masks (Micro Lithography Services) and SU-8 2075 photoresists (Kayaku Advanced Materials). An MJB4 mask aligner (SUSS MicroTeC) was used to UV expose all the SU-8 spin-coated wafers. The thickness of the structures (corresponding to the depth of channels in the final microfluidic devices) was measured using a DeektAxT stylus profiler (Bruker).

We used the following settings for photolithography:
To collect the droplets after the initial encapsulation or at the end of the first pico-injection, the tip can be connected to the outlet. The bead device, inlet number 5 (Extended Data Fig. 1c,d), and the syringe is disconnected to allow the evacuation of mineral oil as the tip is loaded. For each of the two droplet pico-injection steps, the mineral oil can be pushed using a Nemesys syringe pump (Cetonti) to re-inject droplets into the pico-injectors, inlet number 1 (Extended Data Fig. 1d,e). For each of the re-injection and collection steps, the PDMS-punched holes on the microfluidic device need to be primed with 5% w/w 008-FluoroSurfactant (RAN Biotechnologies) in HFE-7500 (3M) to avoid a trapped air bubble to perturbate the stability of re-injection or the integrity of the emulsions. After droplet collection during the encapsulation and first pico-injection, the tip can be closed by inserting the narrower end of the tip into a glass-bonded PDMS plug, which closes the system and allows for incubation of the tip in the water bath (Extended Data Fig. II). The glass-bonded PDMS plug was fabricated before the experiment by punching a 8-mm-thick piece of PDMS with a 1.5-mm biopsy puncher that was then bonded to microscopy glass using an oxygen plasma.

### VASA-drop: microfluidic device operation.

**Polyacrylamide beads manufacturing.** Barcoded polyacrylamide beads were manufactured following a previously described protocol. In brief, a polyacrylamide mix was used to generate 60 µm of water-in-oil emulsions using a single-inlet flow-focusing device and collected in a 1.5-ml LoBind tube (Eppendorf) containing 200 µl of mineral oil (Sigma-Aldrich). The droplets were solidified overnight at 65 °C, de-emulsified using a 20% (w/v) 1H,1H,2H,2H-perfluoro-1-octanol (Alfa Aesar) in HFE-7500 (3M) solution and stored at 4 °C for up to 6 months.

**Co-encapsulation of cells and barcoded beads.** A detailed protocol was used as a reference for droplet generation. First, the microfluidic droplet generation chip was installed on the stage of an inverted microscope (Olympus X73). Next, two pieces of polyethylene tubing (Portex, Smiths Medical) were connected to two 1-ml gas-tight syringes (SGE) and filled with PBS (Lonza). The tubing was manually filled with PBS, and a small, 1-cm-long air bubble was left at the end tip of each tubing. The bead suspension and lysis mix were manually aspirated to the tubing, and the small air bubble provided a separation between the reagents and the PBS buffer. Then, 150 µl of cell suspension was manually aspirated into the cell loading tip pre-filled with mineral oil (Sigma-Aldrich). A fourth 2.5-ml glass syringe (SGE) was filled with 5% w/v 008-FluoroSurfactant (RAN Biotechnologies) in HFE-7500 (3M). Next, all three tubings and the cell chamber with cell suspension were inserted to the corresponding inlets of the droplet generation chip (Extended Data Fig. II). From Nemesys syringes pumps (Cetonti) were used to drive the droplet generation, and the droplet formation was monitored using x4 or x10 objectives (Olympus) and a fast camera (Phantom Miro eX4) connected to the inverted microscope. After the device was primed and droplet generation was stabilized, the collection chamber was connected to the outlet.

### Microfluidic device operation—pico-injection.

Before starting the pico-injection of droplets containing single-cell lysates, the electrode sections, numbers 6 and 7 (Extended Data Fig. 1d,e) of the devices, were pre-filled with filtered 5 M NaCl as previously described. The pico-injection chip was filled with 5% (w/v) 008-FluoroSurfactant (RAN Biotechnologies) in HFE-7500 (3M) using a pre-filled 2.5-ml glass syringe (SGE) connected to a piece of tubing (Portex, Smiths Medical). The reaction mix was primed, and the tip containing the emulsions (with fluorinated oil evaporated by pushing the glass syringe until the emulsions reached the exit of the tip) was primed and connected to the device. Next, flows of droplet emulsions, the reaction mix, the emulsion–diluting oil, number 2 (Extended Data Fig. 1d,e), and the droplet-spacing oil, number 3 (Extended Data Fig. 1d,e) were applied using the Nemesys syringe pumps (Cetonti). The droplets were diluted in a first instance in the re-injection chamber and then spaced with the second stream of oil in a flow-focusing re-injection junction. 5% (w/v) 008-FluoroSurfactant (RAN Biotechnologies) in HFE-7500 (3M) was used for both diluting and spacing of droplets. The function generator (AIM & Thurlby Thandar Instruments) was set to generate square waves of 2.5 V amplitude and 10 kHz frequency, which were further amplified 100 times to 250 V by a Trek 601C-1 amplifier, which enabled coalescence-activated reaction of the reagent into the droplets. The droplets were collected in a 1-ml collection tip connected at the outlet, number 5 (Extended Data Fig. 1d,e).

### VASA-drop: polycrylamide bead barcoding.

The bead barcoding procedure was performed as previously described with the indrop v3 barcoding scheme. In brief, the solidified barcoded beads were filtered and dispersed in four 96-well plates containing the first barcode from the indrop v3 design, and the bead-bound ascorbic acid-derivatized 5′-modified oligonucleotides (introduced by Clontech) after the barcoded oligonucleotides. The reaction was then stopped, and the second strand was removed using a sodium hydroxide treatment. The second barcode was added in a similar fashion, and the beads were stored for up to 6 months at 4 °C.

### VASA-drop: cell encapsulation in water-in-oil emulsions.

For the cultured cells and the embryos, we used a loading concentration of 450 cells per µl in 1× PBS (Lonza) with 15% OptiPrep (Sigma-Aldrich). The lysis mix was made fresh before each encapsulation, as follows: 0.5 mM DNTPs each (Thermo Fisher Scientific),

---

**VASA-drop: soft lithography.** To manufacture PDMS microfluidic devices, 20–30 g of silicone elastomer base and curing agent (Sylgard 184, Dow Corning) were mixed at a 10:1 (w/w) ratio in a plastic cup and de-gassed in a vacuum chamber for 30 minutes. PDMS was then poured on a master wafer with SU-8 structures and cured in the oven at 65 °C for at least 4 hours. Next, the inlet holes were punched using two types of biopsy punchers with plungers (Kai Medical Laboratory): a 1.5-mm-diameter punch was used to make the inlet for the cell delivery tip, number 2 (Extended Data Fig. 1c); outlet for droplet collection tip, inlet number 5 (Extended Data Fig. 1c,d); and the inlets for droplet re-jection, number 1 (Extended Data Fig. 1d,e); whereas other inlets were made using a 1-mm-wide biopsy puncher. The patterned PDMS chip was then placed on a 52 mm × 76 mm × 1 mm (length × width × thickness) glass slide (VWR) in a low-pressure oxygen plasma generator (Femto, Diener Electronis). Next, the hydrophobic modification of microfluidic channels was performed by flushing the device with 1% (v/v) trichloro(1H,1H,2H,2H-perfluorooctyl)silane (Sigma-Aldrich) in HFE-7500 (3M) and baked on a hot plate at 75 °C for at least 30 minutes to evaporate the fluorocarbon oil and silane mix.

Although we have not used these commercial suppliers, we propose the following list of contract manufacturers for users who may not have access to photolithography/soft lithography: Five photon Biochemicals, Darwin Microfluidics, uFluidics, Flowjem and Microfactory.

| Fabrication step (no. of layers) | 1st layer | 2nd layer (used for picoinjectors only) |
|--------------------------------|----------|----------------------------------------|
| Exposure (at -10 mW cm²)        | 2x 10 s  | 2x 10 s                                |
| Post-baking                     | 2 min at 65 °C | 2 min at 65 °C                        |
| Hard baking (optional)          | 10 min at 200 °C (only for a 1-layer chip) | 10 min at 200 °C                      |
| Measured range of thicknesses   | 80-84 µm | 168-178 µm (second layer is usually -20% thicker than nominal) |

---

**VASA-drop: cell loading and droplet collection/re-injection chamber manufacturing.** Cell injection chamber. The cells were loaded in a cell loading tip pre-filled with mineral oil (Sigma-Aldrich). To manufacture the cell loading tip, a low-retention pipette tip with 200-µl volume capacity (Aswyn) was cut at the top, in parallel to the rim and under the filter. A solidified 3-mm-thick piece of PDMS (Dow Corning) was punched from a slab of PDMS with a 5.0-mm sampling tool (EMS-Core). The circular piece of PDMS was then biopsy-punched with a 1-mm-wide biopsy puncher (Kai Medical Laboratory) in the middle. The circular piece of PDMS was pushed inside the tip while remaining parallel to the upper rim of the tip. A 1-ml glass syringe (SGE) was then pre-filled with 1 ml of mineral oil and connected to a 30-cm-long tubing (Portex, Smiths Medical) that can be inserted to a hole in the middle of the circular PDMS piece in the tip. Next, the tip was pre-filled with mineral oil by manually pushing the syringe, and the cell-containing solution was further aspirated with care as to not introduce any air bubble in the system. The tip can then be connected to the cell-encapsulation PDMS device, inlet number 2 (Extended Data Fig. 1c,g), and injection rates are modulated by a Nemesys syringe pump (Cetonti).

**Droplet collection and re-injection chamber.** A second type of tip chamber was designed to collect, incubate and re-inject droplets for each microfluidic step. To this end, a 5-mm-thick PDMS piece was punched from a slab of PDMS with an 8.0-mm sampling tool (EMS-Core) and re-punched in its center using a 1-mm-wide biopsy puncher (Kai Medical Laboratory), and a 30-cm-long tubing (Portex, Smiths Medical) was connected to the latter punched hole. The resulting piece of PDMS was then inserted into a 1-ml filterless pipette tip (Sigma-Aldrich) with a parallel orientation to the rim. Unsolidified PDMS (Dow Corning,1:10 w/v) was then deposited to the space between the circular PDMS piece at the top. The tip was then incubated at 65 °C for at least 4 hours and connected to a 1-ml glass syringe (SGE) pre-filled with mineral oil. The tip was then pre-filled with mineral oil by manually pushing the connected syringe.
For V ASA-plate: after IVT, 2 V ASA-plate and V ASA-drop: downstream library preparation and sequencing. 1H,1H,2H,2H-perfluoro-1-octanol. The tubes were centrifuged for 5 seconds on of filtered HFE-7500 was added to the emulsions, followed by 200 µl of 100% 0.52% IGEPAL-CA630 (Sigma-Aldrich), 40 mM UltraPure Tris-HCl pH 8 (Life Technologies). The droplets were centrifuged for 5 seconds on of loaded droplets) in 1-ml LoBind tubes (Eppendorf) and bead suspensions were loaded in the tubing of two individual 1-ml SGE glass-bonded PDMS plug. The tip container was then incubated for 25 minutes at the proper length on a Bioanalyzer (Agilent), and concentration was measured with a Qubit (Invitrogen). A detailed catalog of reagents and instrumental materials is provided in Supplementary Table 14. The V ASA-drop samples were sequenced on a NovaSeq 6000 S2, 300 cycles flow cell (Illumina), with the following parameters: Read1 247 cycles, Index1 31 cycles, Index2 46 cycles, Read2 14 cycles. 10x Chromium) were extracted. To determine the number of cells in each sample, first the total number of raw reads was determined for each possible barcode. Next, we plotted the histogram of (read number) for each possible barcode, which we fitted to a polynomial function that shows two or three minima. The position of the minimum with the highest value of (reads) as the threshold: only barcodes with reads above this threshold were used for downstream analysis. We merged sequenced barcodes that can be uniquely assigned to an accepted barcode with a Hamming distance of 2 nt or less. FASTQ file pre-processing in V ASA-drop and 10x Chromium. Raw reads for VASA-drop were pre-processed with a Python script to have a favorable format for the pipeline (four reads were demultiplexed and re-arranged into two reads). We merged sequenced barcodes that can be uniquely assigned to an accepted barcode with a Hamming distance of 2 nt or less.
scRNA seq analysis for mouse 10x Chromium libraries and individual timepoints.
The alignments for each cluster were assembled and then merged using StringTie2 (ref. 12). The resulting GTF file was then compared to the input transcriptome annotation using gtfcompare35, which assigns a classification code to each assembled transcript, which is subsequently used to filter transcripts with codes that indicate additional portions of annotated transcripts or novel genes. Novel transcripts spanning three or more exons that were classified under code k, m, n, j, x, i or y were appended to the input transcriptome annotation, expanding the original set of annotated transcripts. Finally, to further improve the quality of potentially novel transcripts, additional custom filtering steps were implemented to avoid novel transcripts due to false-positive novel exons. This filter is particularly important for transcripts assembled from reads that are mapped to repetitive sequences or exons that are ≤30 nt, which can arise from HISAT2 misalignments. To annotate potentially novel microexons, we used MicroExonator, a specialized computational workflow for discovering and quantifying microexons. After running MicroExonator's discovery module, we obtained a transcriptome annotation, which was later processed with custom scripts to limit the number of alternative transcription start and end sites.

Quantification of AS events across cell types. The final GTF from the expanded transcriptome annotation was used to quantify isoforms and AS events using Whippet. We ran Whippet through MicroExonator's downstream module to profile AS events using scRNA-seq data, which enabled randomized aggregations of cells into pseudo-bulk and pairwise comparisons of AS profiles across cell types. To determine relevant pairs for comparison of AS profiles across cell types, we used PAGA to calculate connectivities between cell clusters based on gene expression. We then compared the 72 pairs of clusters that have a connectivity

Comparison between 10x Chromium and VASA-seq embryo data. For the comparison, only reads mapping at the 80% 3' end of gene bodies were used to generate count tables for both VASA-seq and 10x Chromium. Only genes expressed in both technologies were used for the comparison. The technology and the number of counts were regressed out from the combined VASA–10x Chromium dataset, and dimensionality reduction was performed by PCA. Manhattan-based distances between cells were calculated in the combined PCA space. Equivalent clusters were defined by fist clustering each dataset for each technology independently. Second, for a given cluster and a reference technology (for example, VASA-seq), a background histogram of the distances between cells in that cluster and their corresponding first nearest neighbor in the target technology (for example, 10x Chromium) was obtained. Finally, each cell in the target technology was assigned to the cluster of its nearest neighbor in the reference technology. Cells with low transfer scores were excluded, and equivalent clusters technology was assigned to the cluster of its nearest neighbor in the reference technology. Equivalent clusters between VASA-seq and 10x Chromium were defined as groups of cells with identical 10x Chromium and VASA cluster assignments. To assign a gern name to each equivalent cluster, published annotations for the 10x Chromium datasets were used (epiblast: epiblast, primitive streak, anterior primitive streak, caudal epiblast and NMP; ectoderm: Ex ectoderm, caudal neural crest, rostral neuroectoderm, surface ectoderm, forebrain/midbrain/hindbrain, neural crest and spinal cord; mesoderm: nesent mesoderm, caudal mesoderm, ExE mesoderm, intermediate mesoderm, mesenchyme, mixed mesoderm, paraxial mesoderm, pharyngeal mesoderm, somitic mesoderm and cardiomyocytes; endoderm: endoderm, epiblast endoderm and visceral endoderm; blood: blood progenitors 1, blood progenitors 2, erythro1, hematopoietic progenitors, endothelium, erythroid2 and erythroid3; and PCG: PGCG). The prevalent annotation for each equivalent cluster was used.

Master UMAP for VASA-drop mouse embryo data. The master UMAP, where all cells for all timepoints are integrated together, was obtained as previously described32. In brief, we first built a directed graph. For each cell in each timepoint, we found the top 30 nearest neighbors in the subset of cells from the same timepoint and the previous timepoint (cells from E6.5 are only connected to cells from E6.5). To do so, all the cells in the subset are projected to the PCA space of the current timepoint, and distances are calculated using the Manhattan metric. Next, the undirected graph was extracted and used to project the data to the two-dimensional UMAP.

Expanding the transcriptome annotation. A total of 33,662 demultiplexed and ribo-depleted FASTQ files for each cell were used to reconstruct the transcriptome and quantify AS events. To this end, we implemented a custom computational workflow using Snakemake36 based on Hisat2/StringTie2 (ref. 12) and additional custom scripts. First, PCR duplicates were removed through a custom Python script that calculates pairwise identity across UMIIs for each sequenced read within single cells. Then, reads were grouped by previously obtained Leiden clusters and mapped to genome assembly, from E9.5) reads and fewer than 106 transcripts were kept. Next, only cells in which 85–95% reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used reads (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used. For Smart-seq3, both reads containing a UMI (5′ reads and non-UMI-containing reads were used together. For the gene expression analysis between cells in S-phase and not S-phase was performed using the Foldchange function in Scanpy), for separate timepoints and all data together (Supplementary Table 4). Next, highly variable genes with mean log expression between 0.0125 and 1.5 were selected, and cell cycle graphs were regressed out (Scapiny function scapiny:pp:regress out), and data were z-transformed (scapiny:pp:scale). For all timepoints, we selected the top 50 principal components (except for E6.5, for which we selected the first 20). For each timepoint, and computed a directed graph connecting nearest neighbor cells in the reduced principal components analysis (PCA) space, using the Manhattan metric as previously described32. Initially, for each cell, we identified its ten nearest neighbors. An outgoing edge from cell i to cell j was kept if the distance d was less than the mean ± 1.5x s.d. among all the distances connecting ten nearest neighbors. Cells that were connected to any other cell were filtered out. The directed graph was converted to an undirected graph, and a two-dimensional UMAP was obtained as previously described32. We clustered the data using the Leiden algorithm (scapiny:tl:leiden, resolution set to 1) and performed differential gene expression between Leiden clusters using the t-test (default scapiny:tl:rank_genes_groups).

scRNA seq analysis for mouse 10x Chromium libraries and individual timepoints. 10x data were analyzed similarly to the VASA-seq data. Here, we kept cells with more than 105 and fewer than 106 uniquely detected transcripts and with 85–97% protein-coding transcripts. Cell cycle genes were not removed from the set of highly variable genes, and cell cycle regression was not performed. The effect of the libraries was regressed out before Z-score scaling.

Benchmarking against other methods. To determine the number of potential doublets, barcodes with more than 75% of the genes assigned to only one of either mouse or human were considered singlets. Cells with fewer than 7,500 UFIs were filtered out and not assigned to any organism. For gene body coverage, the BAM files for all single cells were used as a bulk. QoRTs was used to calculate coverage, and only protein-coding genes were kept. For Smart-seq3, both reads containing a UMI (5′ reads) and non-UMI-containing reads were used together. Average coverages were used for the plotting. To determine percentages of different biotypes, all single cells were used as a bulk. UMI/UMI filtering was carried out for reads where this was possible. For Smart-seq3, both reads containing a UMI (5′ reads) and non-UMI-containing reads were used together. For the gene detection assay, only cells that had been sequenced to the highest number of reads (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used (reads with proper barcode and quality/homopolymers trimming) were used. For Smart-seq3, both reads containing a UMI (5′ reads) and non-UMI-containing reads were used together. Mean and standard deviation were calculated and plotted.

Filtered reads were mapped to the mouse GRCm38 genome (Ensembl 99) or to the human GRCh38 genome (Ensembl 99) using STAR with default parameters. Assignment of reads to gene biotypes was performed according to the following hierarchical:

- All mappings falling in TEC transcripts were discarded.
- Reads fully falling inside a region annotated as miscRNA, miRNA, mRNA, TrlGene, miRNA, RNA, ribozymes, ssaRNA, snRNA or snoRNA (for example, biotypes that do not have annotated introns) were assigned to unique regions.
- When a read maps to multiple genes simultaneously (because of annotation overlap in the reference GTF file), exonic annotations were given preference to introns. In case all references are exonic or intronic, the read is assigned to a gene whose name is the sequence of the target gene names.
- Reads falling into exon-intron junctions or inside introns are assigned to splipced transcripts. Reads falling inside exonic regions are assigned to spliced transcripts.
- If at least one UFI of the same cell from the same transcript has been assigned to an unspliced transcript (because it is mapped in an intron or an exon–exon junction), the all other reads with the same UFI of the same cell for the same transcript are automatically assigned to unspliced transcripts even if they mapped to exons exclusively.

85–97% protein-coding transcripts. Cell cycle genes were not removed from the set of highly variable genes, and cell cycle regression was not performed. The effect of the libraries was regressed out before Z-score scaling.
≥0.05. For each comparison, cells from each cluster were randomly pooled to form at least three different pseudo-bulks of 200 or fewer cells. To detect reproducible changes of splicing node inclusion across cell types, random pseudo-bulk pooling and differential inclusion steps were repeated 50 times for each pairwise comparison, avoiding the detection of spurious splicing events. As part of MicroExonator’s workflow, the obtained probabilities of each splicing node to be differentially included were used to fit a beta distribution model and calculate CDF-beta values for each event. DSSNs were defined as events with CDF-beta values equal to or lower than 0.05. To identify SNMs, we calculated the average ψ values for each splicing node across three randomly defined pseudo-bulk samples for each cell cluster. For splicing nodes where ψ values could be quantified based on at least ten reads across at least 50 pseudo-bulks, we calculated the Z-score by comparing to all other pseudo-bulks. We considered a splicing node as an SNM for a given cell type if at least two pseudo-bulks had significant Z-scores (P ≤ 0.05) and an absolute difference of at least 0.3 from the mean across all pseudo-bulks. To show some functional consequences of detected AS events for protein function, we used the drawProteins package75 to draw scaled diagrams of protein domains and other features annotated in UniProt76.

**References**

59. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
60. Sciambi, A. & Abate, A. R. Generating electric fields in PDMS microfluidic devices with salt water electrodes. *Lab Chip* **14**, 2605–2609 (2014).
61. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
62. Adiconis, X. et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
63. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
64. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
65. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
66. Hartley, S. W. & Mullikin, J. C. QoTic: a comprehensive toolset for quality control and data processing of RNA-seq experiments. *BMC Bioinformatics* **16**, 224 (2015).
67. Lum, A. et al. DropletUtils. https://bioc conductor.org/packages/release/bioc/html/DropletUtils.html
68. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 (2019).
69. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
70. McMinn, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
71. Koster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
72. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
73. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
74. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
75. Brennan, P. DrawProteins: a Bioconductor/R package for reproducible and programmatic generation of protein schematics. *F1000Res.* **7**, 1105 (2018).
76. Bairach, A. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2004).

**Acknowledgements**

We thank R. van der Linden for assistance during experiments. We thank B. Blencowe and F. Stahl, for providing valuable feedback on the manuscript. We also thank all members of the van Oudenaarden, Hollfelder and Hemberg laboratories and A. Hita for scientific discussions. This work was supported by a European Research Council (ERC) Advanced Grant (ERC-AdG 742225-IntScoMics), a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) TOP award (NWO-CW 714.016.001) and the Wellcome Trust (WT108438/C/15/Z). This work is part of the Oncode Institute, which is partly financed by the Dutch Cancer Society. J.D.J received scholarship support from the Biotechnology and Biological Sciences Research Council (BBSRC), J.N. from AstraZeneca, A.L.E. from the Cambridge Trusts and the EU H2020 Marie Curie ITN MMBio and T.S.K. from an EU H2020 Marie Sklodowska-Curie Actions Individual fellowship (MSCA-IF 750772). E.H. is an H2020 ERC Advanced Investigator (69566). M.H. was supported by a core grant from the Wellcome Trust and by funding from the Evergrande Center for Immunologic Diseases. J.N. was funded by the Wellcome Trust (03151Z/16/2Z). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. A.Y. was funded by the BBSRC (BB/RS88855), and the mice used in the study are associated with the Wellcome Trust Strategic Grant (105031). Parts of the illustrations were designed using BioRender.

**Author contributions**

F.S., J.D.J., E.H. and A.v.O. designed the project. F.S. developed the initial VASA-seq molecular workflow in plates. F.S., J.V.L., N.B. and F.v.d.B. optimized the VASA plate workflow. J.D.J. and T.S.K. performed and optimized the VASA-drop workflow. A.Y., T.N.K. and J.N. retrieved and dissociated the mouse embryos. A.I.E. and N.B. retrieved and dissociated the cultured cells. J.D.J. and T.S.K. processed the cells for the mouse atlas. J.D.J. and E.H. performed library preparation and sequencing. A.A. and E.H. developed a versatile VASA-seq specific data analysis pipeline. A.A., F.S. and J.D.J. performed downstream analysis, including benchmarking against other datasets, histone content and velocity analyses. G.E.P. and J.D.J. performed alternative splicing analysis and developed the associated pipelines. A.A., A.M.A. and J.D.J. performed cell type annotation. T.S.K., A.A. and G.E.P. contributed equally. M.H., E.H. and A.v.O. supervised the work. F.S., J.D.J., A.A., G.E.P., T.S.K. and E.H. co-wrote the manuscript, with input from all authors.

**Competing interests**

F.S., A.v.O., J.D.J., T.S.K. and E.H. are inventors on patent applications submitted by the Stichting Oncode Institute on behalf of Koninklijke Nederlandse Akademie Van Wetenschappen and the University of Cambridge (via its technology transfer office, Cambridge Enterprise). A.v.O. is a member of the advisory board of Single-Cell Discoveries.

**Additional information**

Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01361-8. Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01361-8. Correspondence and requests for materials should be addressed to Florian Hollfelder or Alexander van Oudenaarden.

**Peer review information** Nature Biotechnology thanks Kun Zhang, Bart Deplancke and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Overview of the sequencing and droplet microfluidic process and benchmarking analysis. a, The two platforms for VASA-seq, using a microfluidic device (left, VASA-drop) and a plate dispenser (right, VASA-plate). The microfluidic device allows the generation of single-cell libraries from thousands of cells, while the plate-based approach is better for rare cell types where a prior sorting is required. Each library contains the transcriptome from a large mix of cells, which are demultiplexed based on their barcode and index sequences. b, Library barcoding design for the VASA-drop workflow. Mouse embryo libraries were sequenced with the Illumina NovaSeq platform. To avoid index hopping, a custom dual indexing strategy was used. For the index i7 read, which usually only contains barcode 1 (inDrop v3), we inserted a 8-bp second index directly after a 15 bp common sequence. Only reads that had the correct combination of i5 and i7 index were further used for downstream processing. c, Design of the device used for barcoded bead and single-cell co-encapsulation. 1) input channel for the lysis and fragmentation mix, 2) input channel for the cell loading, 3) input channel for the barcoded compressible bead loading, 4) input channel for the fluorinated oil with admixed surfactant, 5) droplet exit channel. d, Design of the first picoinjector device, to inject the repair and poly(A) polymerase. 1) input channel for droplet reinjection, 2) emulsion diluting oil inlet, 3) droplet spacing oil inlet, 4) inlet for the repair and poly(A) polymerase to be picoinjected, 5) droplet exit channel, 6) positive electrode (red), 7) negative (moat) electrode (black). e, Design of the second picoinjector device, to inject the RT enzyme mix. 1) input channel for droplet re-injection, 2) emulsion diluting oil inlet, 3) droplet spacing oil inlet, 4) inlet for the RT mix to be picoinjected, 5) droplet exit channel, 6) positive electrode (red), 7) negative (moat) electrode (black). f, Photography of the container tip used for droplet collection and reinjection in the picoinjector devices. The tip is connected to a glass syringe which enables aspiration and delivery of emulsions. The tip can be connected to a PDMS plug to close the system for incubation in the water bath. g, Photography of the encapsulation process. The container tip collecting the emulsions is plugged into the outlet of the encapsulation device, while a tip containing cells is plugged in one of the inputs to deliver single-cells in the droplets. h, Photography of the bead barcode photocleavage set-up. The container tip is surrounded by aluminium foil to reflect the UV light and is kept on a container filled with ice.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Comparison to other methods. a, Species mixing histogram plotted as a percentage of UFIIs quantified from mapping events to a human reference genome, divided by the sum of UFIIs quantified from mapping events to mouse and human reference genomes. b, Proportion of mapped reads to all annotated genes for each biotype using HEK293T cells across all methods. VASA-seq detected proportionally larger amounts of lncRNA genes (light blue) compared to the other technologies. The proportion of detected sncRNAs in VASA-seq methods was higher than 10x Chromium and Smart-seq3, but lower than with Smart-seq-total (grey). c, Proportion of sncRNA biotypes captured for HEK293T cells across methods for reads mapped to all annotated genes. Only VASA-seq and Smart-seq-total detected a significant proportion of sncRNAs biotypes, with Smart-seq-total providing the best performance in terms of relative distribution of biotypes, followed by VASA-drop and VASA-plate. MiscRNA (brown), snoRNA (pink), Ribozyme (grey-green) and snRNA (red) took up the largest proportion of measured biotypes. d, The number of detected protein coding genes in HEK293T, for each method, is plotted against the number of reads (after quality filtering, adapter removal and homopolymer trimming), per cell across different downsampling thresholds. The saturation curves showed that VASA-seq was the most sensitive of the methods. Curvature of gene detection indicated that full complexity was not reached for the method when 75,000 reads were allocated to each cell. Only cells that were sequenced to at least 75,000 reads were used. e, Number of detected genes per cell for Smart-seq3 (red), Smart-seq-total (black) and VASA-plate (blue) when sequenced at a depth of approximately 750,000 reads per cell. Data in boxplot represent the 25%, median (centre) and 75% percentiles with minimum and maximum values. The number of cells sampled were n = 113 (Smart-seq3), 260 (Smart-seq-total) and 68 (VASA-plate). f, Percentage of sequenced reads with proper barcodes that survived trimming, rRNA filtering and mapping for each method using HEK2993T cells (VASA-plate, VASA-drop, 10x Chromium, Smart-seq3 and Smart-seq-total). g, Percentage of unspliced reads for each method for HEK293T cells. VASA-seq detected more unspliced reads (44.1–56.5%) than the alternative technologies (12.8–38.1%). Data in boxplot represent the 25%, median (centre) and 75% percentiles with minimum and maximum values. The number of cells sampled were n = 976 (10x Chromium), 117 (Smart-seq3), 260 (Smart-seq-total), 726 (VASA-drop) and 192 (VASA-plate).
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Mouse gastrulation and organogenesis atlas. a, Brightfield microscope images of the embryos collected before dissociation. Two collections were performed for E6.5 (39 embryos total), whereas single collections were performed for E7.5 (8 embryos total), E8.5 (7 embryos total) and E9.5 (6 embryos total). Scale indicates a 1 mm scale for the background gridlines. b, Average gene expression correlation values ($r^2$) per biotype across equivalent clusters between 10x Chromium and VASA-seq at stage E8.5. n number of cells were 8,365 (VASA-seq) and 9,939 (10x). The equivalent clusters are annotated by using the percentage of cells assigned to a cell type in 10x Chromium. Only cell types present in more than 30% of the equivalent cluster are indicated. The points are the mean and standard error of the mean obtained by bootstrapping genes for each equivalent cluster and biotype 1000 times are represented. c, UMAP of E6.5 mouse embryo cells from 10x (n = 640) and VASA-seq (n = 298) that were part of equivalent clusters. Clusters that are detected in both technologies are marked with numbers 1-16 and each cluster is colored according to the cell type category: blue = ectoderm and grey = epiblast. Grey fill in cluster label indicates extra-embryonic contribution, black fill indicates embryonic contribution. d, UMAP of E7.5 mouse embryo cells from 10x (n = 3,319) and VASA-seq (n = 1,892) that were part of equivalent clusters. Clusters that are detected in both technologies are marked with numbers 1-38 and each cluster is colored according to the cell type category: green = blood, blue = ectoderm, purple = endoderm, orange = mesoderm and grey = epiblast. Grey fill in cluster label indicates extra-embryonic contribution, black fill indicates embryonic contribution. e, Scatter plot showing the number of differentially expressed genes per cluster at E6.5 in VASA-seq (x axis) vs. 10x Chromium (y axis) for spliced protein coding (left panel), unspliced protein coding (middle panel) and lncRNA (right panel) counts. Numbers indicate clusters where a higher number of marker genes were detected in 10x. Clusters are colored according to the cell type category: blue = ectoderm and grey = epiblast. f, Scatter plot showing the number of differentially expressed genes per cluster at E7.5 in VASA-seq (x axis) vs. 10x Chromium (y axis) for spliced protein coding (left panel), unspliced protein coding (middle panel) and lncRNA (right panel). Numbers indicate clusters where a higher number of marker genes were detected in 10x. Clusters are colored according to the cell type category: green = blood, blue = ectoderm, purple = endoderm, orange = mesoderm and grey = epiblast.
Extended Data Fig. 4 | Histone expression and regression of cell-cycle. a, UMAPs showing the log10 total counts for histone genes (left panel), S-phase genes (middle panel) and GM genes (right panel). Only cell cycle scoring using solely histone genes shows a clear cell cycle segregation in VASA-seq. b, Core expression of cell-type specification markers during gastrulation and early organogenesis projected on the 10x Chromium and regressed VASA-seq UMAP. c, Heatmap showing differentially expressed multi annotated histone genes. Rows display genes, and columns display cell types. Cell type categories/germ layers are colored above the heatmap.
Extended Data Fig. 5 | RNA velocity and global splicing marker analysis. 

a, Velocity of Adgrf5 shown on diagrams of spliced vs. unspliced counts, along with UMAPs highlighting velocity and expression for the gene for VASA-seq (top) and 10x Chromium (bottom) showing both induction and repression in the endothelium for VASA-seq with high goodness of the fit. Goodness of the fit are values approximately one SD above average for each method to show genes that are good in both datasets. Black arrow indicates the endothelium in the VASA-seq dataset.

b, Velocity of Cacna2d2 shown on diagrams of spliced vs. unspliced counts, along with UMAPs highlighting velocity and expression for the gene for VASA-seq (top) and 10x Chromium (bottom) showing induction of the gene in the Primitive heart tube and first heart field for VASA-seq. Goodness of the fit are values approximately one SD above average for each method to show genes that are good in both datasets.

c, Velocity of lncRNAs with unspliced molecules uniquely detected in the VASA-seq dataset for the endothelium. Phase diagrams of spliced vs. unspliced counts, along with UMAPs highlighting velocity and expression for VASA-seq show early induction of Hoxa11os in the yolk sac, followed by induction of Gm50321 across the endothelium (yolk sac and embryonic) and selective repression of D03007L05Rik at E9.5. Dots in the diagram are labelled according to developmental time points.

d, Violin plot of the velocities across timepoints E6.5, E7.5 and E8.5 in the endothelium for the VASA-seq dataset showing differential induction and repression for lncRNAs. Dashed line indicates null velocity.

e, Violin plot showing the distribution of coverage values obtained for splice nodes when computed at single-cell or pseudo-bulk level.

f, Violin plot showing the number quantified spliced nodes (read coverage>5) obtained when quantified at the single-cell or pseudo-bulk level.

g, Euler diagram showing the splicing node intersection between the DISN and SNM sets.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Heart and blood development reveal tissue-specific AS patterns across developmental trajectories.

a, Differential gene expression analysis using a two-sided Wilcoxon rank sum test between the FHF (negative average log2 fold-change values) and the PHT (positive average log2 fold-change values) with differentially expressed RBPs highlighted. Significance levels are indicated by color (grey non-significant and black significant), and determined by the following threshold: |average log2 fold-change| > 0.5 and Bonferroni adjusted p-value < 1E-05. 
b, Rbfox2_143 and 144 mutually exclusive exon usage in the FHF and PHT respectively.
c, Rbfox2 gene expression across the atlas, log2 normalized values.
d, Tpm1 sashimi plot between the ECE, FHF and PHT, dashed square highlights the region of interest plotted in Fig. 6c.
e, Tpm1_29 single-cell PSI UMAP plot across the atlas highlighting a PHT specific core exon usage at the C-terminus.
f, Tpm1_32 single-cell PSI UMAP plot across the atlas highlighting a PHT specific core exon usage at the C-terminus.
g, UMAP plot across timepoints depicting erythropoietic cell types.
h, Single-cell Ψ UMAPs of Epb41_30, Add1_37, Ank1_43 and Mbnl1_37 depicting alternative exonic usage across blood maturation trajectories.
i, Single-cell gene expression UMAP plot depicting differences in gene expression for Epb41, Add1, Ank1 and Mbnl1 illustrating differences in gene expression that differ from the AS patterns observed across blood maturation.
Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

| Data collection | The VASA-drop samples were sequenced on a Novaseq 6000 S2, 300 cycles flow cell (Illumina), with the following parameters: Read1 247 cycles, Index1 31 cycles, Index 2 8 cycles, Read2 14 cycles. VASA-plate samples were sequenced on a Nextseq 500 High output 150 cycles flow cell (Illumina), with the following parameters: Read1 26 cycles, Index 8 cycles, Read2 135 cycles. |
| Data analysis | Data was analyzed using the following packages for the analysis from Figure 1-4: Scrapy v1.5.1, Trim Galore v0.4.3, Cutadapt v2.10, STAR aligner v2.7.3a, bwa v0.7.10, scVelo v0.2.2, Samtools v1.3.1, bedtools v2.29.2, Scrublet v0.2.1, QoI TS v1.2.42 and dropletUtils v1.6.1. For the analysis in Figures 5 and 6, we used the following software: Whippet v1.5.1, MicroExonator v1.0.0, bwa v0.7.15, Hisat2 v2.1.4, Snakemake 5.32.0, bedtools v2.29.2, ggashami v0.5.0, GFF utilities v0.9, UpsetR v1.4.0, drawProteins v1.0.2, StringTie v2.1.4. Autocad 2018 was used for generating the microfluidic designs. Link to code: https://github.com/hemberg-lab/VASAseq_2022 |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data is available at GEO with accession number: GSE176588. For benchmarking, we used the following accession numbers: E-MTAB-8735 [Smart-seq3], GSE151334
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](http://nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size   | >30,000 cells sequenced from C57Bl/6 embryos at stage E6.5, E7.5, E8.5 and E9.5. Sample size for each timepoint up to E8.5 was determined to match the equivalent dataset from Pijuan-Sala et al. (Nature, 2019). E9.5 sample size was estimated relatively to E8.5 sample size. |
|---------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions| Cellular doublets and low complexity cells were excluded from the analysis.                                                                                                                         |
| Replication   | The datasets generated were benchmarked against equivalent technologies: Smart-seq3, Smart-seq total and 10x Genomics Chromium. Samples for VASA-drop were collected in batches of ~1,000 cells to underline the absence of batch effect. |
| Randomization | N/A                                                                                                                                                                                                                                                             |
| Blinding      | N/A                                                                                                                                                                                                                                                             |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a                              | n/a     |
| - Antibodies                     | - Involved in the study |
| - Eukaryotic cell lines          | - ChiP-seq |
| - Palaeontology and archaeology  | - Flow cytometry |
| - Animals and other organisms    | - MRI-based neuroimaging |
| - Human research participants    |         |
| - Clinical data                  |         |
| - Dual use research of concern   |         |

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)  HEK293T cells were a generous gift from Prof. Marc de la Roche (Department of Biochemistry, University of Cambridge). Mouse embryonic stem cells E14Tg2a wild-type were a generous donation from Prof. Austin Smith. The C57Bl/6 mouse embryos were obtained from Charles River or the Cambridge Stem Cell Institute.

Authentication          None of the cells were authenticated

Mycoplasma contamination Cells were not screened for Mycoplasma contamination.

Commonly misidentified lines (See [ICLAC register](#))  N/A

Animals and other organisms

Policy information about [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals  *Mus musculus C57Bl/6* males and females at stages E6.5, E7.5, E8.5 and E9.5.
| Wild animals      | N/A                      |
|-------------------|--------------------------|
| Field-collected samples | N/A                      |
| Ethics oversight  | Use of animals in this project was approved by the Animal Welfare and Ethical Review Body for the University of Cambridge and relevant Home Office licence PPL [76777883] is in place. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.