Diverse Routes towards Early Somites in the Mouse Embryo

Carolina Guibentif[1,2,3,*], Jonathan A. Griffiths[4,†,*], Ivan Imaz-Rosshandler[1,2], Shila Ghazanfar[4], Jennifer Nichols[2,6], Valerie Wilson[5^], Berthold Göttgens[1,2,^], John C. Marioni[4,7,8,#,^]

1. Department of Haematology, University of Cambridge, CB2 0AW Cambridge, UK

2. Wellcome-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, CB2 0AW Cambridge, UK

3. Sahlgrenska Center for Cancer Research, Department of Microbiology and Immunology, University of Gothenburg, 413 90 Gothenburg, Sweden.

4. Cancer Research UK Cambridge Institute, University of Cambridge, CB2 0RE Cambridge, UK

5. MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, University of Edinburgh, EH16 4UU Edinburgh, UK

6. Department of Physiology, Development and Neuroscience, University of Cambridge, CB2 3DY Cambridge, UK

7. Wellcome Sanger Institute, Wellcome Genome Campus, CB10 1SA Cambridge, UK

8. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, CB10 1SD Cambridge, UK

* These authors contributed equally

† Current address: Genomics Plc, 50-60 Station Road, Cambridge, CB1 2JH, UK

Lead contact

^ Co-corresponding authors:

JCM (lead contact): john.marioni@cruk.cam.ac.uk

22 BG: bg200@cam.ac.uk

23 VW: V.Wilson@ed.ac.uk

24

Summary

The formation of somites lays down the segmental organization of vertebrates. Here, we describe three trajectories towards somite formation in the early mouse embryo. Precursors of the anterior-most somites ingress through the primitive streak before E7 and migrate anteriorly by E7.5, while a second wave of more posterior somites develops in the vicinity of the streak. Finally, neuromesodermal progenitors (NMPs) are set aside for subsequent trunk somitogenesis. Single-cell profiling of $T^{-/-}$ chimeric embryos shows that the anterior somites develop in the absence of T, and suggests a cell-autonomous function of T as a gatekeeper between paraxial mesoderm production and building of the NMP pool. Moreover, we identify putative regulators of early T-independent somites, and challenge the T-Sox2 cross-antagonism model in early NMPs. Our study highlights the concept of molecular flexibility during early cell type specification, with broad relevance for pluripotent stem cell differentiation and disease modelling.

Introduction

The recent emergence of high throughput single-cell RNA-sequencing assays has allowed researchers to survey entire transcriptional landscapes of development in numerous species (Cao et al., 2019; Packer et al., 2019; Pijuan-Sala et al., 2019; Wagner et al., 2018). Somites are transient segments of paraxial mesoderm that give rise to the axial skeleton and associated musculature. Following formation of the most anterior or occipital somites, subsequent axis elongation is fuelled by a pool of neuromesodermal progenitors (NMPs), which give rise to neural components of the spinal cord as well as the mesodermal tissue of the somites (Pourquie, 2001; Tzouanacou et al., 2009). NMPs are characterized by co-expression of transcription factors associated with gastrulation, mesodermal and neural development, including *Brachyury* (*T*), *Sox2* and *Nkx1-2* (Henrique et al., 2015; Steventon and Martinez Arias, 2017; Wilson et al., 2009).

Starting as uniform blocks of epithelium, somites compartmentalize into ventral sclerotome (which gives rise to major elements of the skeleton, such as the vertebrae and ribs) and dorsal dermomyotome (precursor of skeletal muscles and of the skin of the back; Keynes and Stern, 1988). Somitogenesis is often portrayed as a relatively uniform process, regulated by an interacting network of signalling pathways and transcription factors such as Fgf, Wnt, Notch, T and Tbx6 (Chapman and Papaioannou, 1998; Hubaud and Pourquie, 2014; Martin and Kimelman, 2008). However, multiple lines of evidence indicate that disruption of these canonical somite regulators has little effect on the formation of the first, most anterior, somites both in mouse and in fish (Nowotschin et al., 2012; Pourquie, 2001). Indeed, the molecular programs responsible for the formation of these occipital somites remain poorly defined. Occipital somites differentiate early in development, disintegrate quickly thereafter, and do not give rise to repetitive skeletal structures. In chick, gene expression analysis has demonstrated a specific molecular make-up of the anterior-most somites (Rodrigues et al., 2006), and in Amphioxus there are at least three distinct transcriptional networks regulating the emergence of specific anterior-posterior somite subsets (Aldea et al., 2019). Overall, these observations suggest that multiple, potentially independent, molecular pathways can generate somites.

Here we used trajectory inference in a transcriptional atlas of mouse gastrulation, as well as single-cell profiling of $T^{-/-}$ embryonic chimeras, to show that the somitic tissues present in the E8.5 mouse embryo emerge through different developmental pathways. A first wave arises from early mesoderm progenitors that ingress through the primitive streak before E7.0, and migrate anteriorly before E7.5, while a second wave of more posterior somitic progenitors remains in the streak region at the posterior end of the embryo. At E7.5, precursors of both waves are anatomically segregated, express different levels of *T* and *Tbx6*, and are exposed to distinct signalling environments. Nevertheless, both will activate the “core” somitic transcriptional program characterized by upregulation of *Tcf15* and *Meox1*. The presence of two distinct waves is

corroborated through analysis of $T^{-/-}$ chimeric embryos, where $T^{-/-}$ cells contribute normally to the first wave, but are highly depleted in the second wave. Depletion of $T^{-/-}$ cells from the second wave is accompanied by increased contribution to a third developmental trajectory, leading from epiblast to E8.5 NMPs, suggesting that T may function as a gatekeeper, regulating the allocation of streak cells to the NMP pool. Finally, we provide evidence that in E8.5 NMPs, T acts predominantly as a transcriptional activator, and may not be necessary for *Sox2* repression.

Results

The E8.5 mouse embryo contains somitic cells with distinct transcriptional signatures

A previously published reference atlas of mouse gastrulation reported 37 major cell types (Pijuan-Sala et al., 2019). To characterize the heterogeneity of E8.5 paraxial mesoderm, we sub-clustered cells belonging to the Somitic and Paraxial mesoderm clusters (Figure S1A-E). Pre-somitic mesoderm was identified by expression of *Fgf17*, *Tbx6*, *Cyp26a1*, *T*, *Hes7*, *Dll3*, *Lef1*, *Rspo3*, *Dkk1* (Bessho et al., 2001; Cao et al., 2004; Chal et al., 2015; Chapman et al., 1996; Galceran et al., 2004; Sakai et al., 2001; Takahashi et al., 2003; Wahl et al., 2007), and cranial mesoderm by elevated levels of *Tbx1*, *Foxl2* and *Pitx2* (Dastjerdi et al., 2007; Marongiu et al., 2015; Nandkishore et al., 2018; Sambasivan et al., 2011; Shih et al., 2007). Four somitic sub-clusters included two sets of uncompartmentalized somitic cells (co-expressing *Tcf15* and *Meox1*; Burgess et al., 1996; Mankoo et al., 2003) separated by clusters indicating commitment to sclerotome (*Pax1* and *Pax9*; Peters et al., 1999) and dermomyotome (*Dmrt2*, *Pax3* and *Meox2*; Kassar-Duchossoy et al., 2005; Sato et al., 2010; Figure S1E).

We next investigated the transcriptional similarity between these populations and other cell types related to axial elongation at E8.5 - NMPs and spinal cord. Diffusion processes revealed a one-dimensional ordering (Figure 1A-C) consistent with higher-dimensional representations (Figure

S1F), where cells are ordered from *Sox2* –expressing spinal cord, through NMPs co-expressing *Sox2*, *T* and *Nkx1-2*, to *Meox1*-expressing paraxial lineages (Figure 1D). Homeobox transcription factor expression supported an underlying spatial component to this ordering, with caudal Cdx genes peaking in the centre, at the position of NMPs (Figure 1E S1G). The cranial mesoderm signature is present at the rostral-most, paraxial end of the ordering (Figure 1E, S1E, G). Next in the ordering are somitic cells flanked by dermomyotome and sclerotome clusters. With this signature of ongoing compartmentalization, these represent the most developed, and therefore most anterior somites. They are followed by the uncompartmentalized, less mature, and more posterior somitic cells, followed by presomitic mesoderm and finally NMPs, present in the more posterior region of the E8.5 embryo (Figure 1B, C).

Inference of distinct developmental routes for E8.5 somitic tissues

Having characterized two transcriptionally distinct anterior and posterior sets of somitic cells as well as a caudal NMP pool at E8.5, we next investigated their putative developmental origins. We reconstructed developmental trajectories using an Optimal Transport approach (WOT; Schiebinger et al., 2019; Methods; Figure 2A, B), which assigns ‘mass’ to each cell at the clusters featuring the presumed trajectory endpoints, and then transfers that mass sequentially backwards between cells in adjacent time-points that are transcriptionally similar. For each cell, the ‘mass’ for each of the three end-points allowed us to allocate it to a given trajectory based on the highest mass contribution. As such, WOT enables incorporation of real-time information of the 9 sequential time points from E6.5 to E8.5 covered in the reference atlas; the classification of cells along a trajectory is thus not only based on their transcriptional similarity, but also on time-point progression. This analysis revealed that NMPs could be traced back to the epiblast at E6.5, while both somitic trajectories originate from E6.5 primitive streak cells. Separation between anterior and posterior somitic trajectories occurred within E7.0-E7.5 nascent mesoderm (Figure 2B and

S2A). This is consistent with a model whereby the diversification of these two populations occurs following ingression through the streak.

Consistent with reported features of gastrulation, both anterior and posterior somitic trajectories display a sharp early downregulation of *Nanog* coupled with a shift in cadherin expression (*Cdh1* to *Cdh2*) characteristic of epithelial-to-mesenchymal transition (EMT; Morgani et al., 2018). By contrast, for the NMP trajectory, these processes occur gradually after E7.0 (Figure 2C). The dynamic expression of NMP markers over time confirms the known NMP signature, with expression of *T*, *Sox2*, *Nkx1-2* and *Cdx2* at E7.5 being maintained up to E8.5. In the NMP trajectory, the persistence of *Cdh1* expression throughout upregulation of *Cdh2* between E7.0 and E8.25 is consistent with a delayed, or "incomplete" EMT in NMPs (Dias et al., 2020). Inspection of additional EMT genes, including *Epcam* (epithelial marker) and *Vim* (mesenchymal marker), reinforced this notion, with co-expression detected in half of the predicted NMP ancestors between E7.5 and E8.0 (Figure S2B-D).

Expression of gastrulation and early mesoderm markers *Eomes* and *Mixl1* in all three trajectories is followed by upregulation of somite markers *Meox1* and *Tcf15* specifically in the two somitic trajectories. Of note, these two trajectories showed clear molecular divergence at E7.5 (before up-regulation of *Meox1* and *Tcf15*), with upregulation of *Wnt3a*, *T* and *Tbx6* specific to the posterior trajectory (Figure 2C).

In addition to examining known regulators, we performed unbiased pair-wise comparisons of gene expression along the entire length of the three trajectories. Specifically, we examined for each gene whether its expression pattern significantly differed between each pair-wise combinations of trajectories, using as input data the mean expression level of each trajectory at each time-point (see Methods; Table S1). Gene Set Enrichment Analysis using the Molecular Signatures Database Hallmark gene set collection (Liberzon et al., 2015; Subramanian et al., 2005) revealed that genes displaying distinct behaviors between the three trajectories were enriched for the EMT

process (Figure S2E), consistent with our targeted analysis (Figure S2B-D). The process of myogenesis was enriched in the anterior vs posterior somitic trajectories comparison, likely due to the different maturation kinetics of these two sets of somites, reflected by the dynamics of the myogenesis regulator *Mef2c* (Figure S2E, F). The mTORC1 pathway was also enriched in the trajectory comparisons, with distinct expression of the upstream regulator *Pdk1* and of the downstream targets *Slc2a1* and *Slc2a3* (Figure S2E, F). Differences between anterior and posterior somitogenesis have been noted previously (Nowotschin et al., 2012; Rashbass et al., 1991). This newly inferred transcriptional trajectory leading from the primitive streak to anterior somitic tissues thus provides a first unbiased molecular description of this process.

Canonical regulators of somitogenesis are depleted in the anterior trajectory

The anterior and posterior somitic trajectories share early (E6.5-E7.0) transcriptional changes associated with gastrulation, as well as upregulation of somitic genes at E8.0-E8.5 (Figure 2C), but also show divergent expression at intermediate timepoints (E7.25-E7.75). Differential gene expression analysis at E7.5 showed earlier *Tcf15* expression in the anterior trajectory is consistent with more advanced somitic maturation compared to the posterior trajectory (Figure 2C, 3A, Table S2). Higher expression of *T* in the posterior trajectory was matched with higher expression of other canonical regulators of somitogenesis, including *Tbx6* and members of the Wnt, Notch, retinoic acid, Fgf and Nodal/Tgfb/BMP signalling pathways. Of note, formation of the earliest anterior somites has been observed in embryos that lack key somitic regulators such as *T*, *Tbx6*, *Wnt3a* and *Fgfr1a* (Takada et al., 1994; Xu et al., 1999). E7.5 cells on the anterior somitic trajectory instead showed strong up-regulation of the transcriptional regulator *Id3* as well as the homeobox transcription factor *Alx1*. Closer inspection of oscillating genes previously observed as part of the somitogenesis clock and wavefront model also revealed an overall reduced expression of these genes along the trajectory leading to anterior somitic tissues compared to that of posterior somitic tissues (Figure S3A).

We next interrogated the temporal dynamics of gene expression changes along the differentiation trajectory towards anterior paraxial mesoderm (Figure S3B). The transcription factor *Hand1* and adhesion molecule *Pmp22* showed an early peak of expression, the frizzled related protein *Sfrp1* and homeobox transcription factor *Alx1* peaked at a midpoint, and homeobox transcription factors of the *Irx* and *Prrx* family peaked towards the end of the trajectory. Many of the above candidate regulators have not previously been implicated in somite development, yet the anterior trajectory nevertheless culminates with induction of the somite master regulators *Tcf15* and *Meox1*.

Parallel spatial and transcriptional divergence of distinct somitic mesoderm programs

Complementary laser-capture microdissection experiments, measuring gene expression in contiguous segments of approximately 20 cells (Peng et al., 2019), have been performed at equivalent stages of mouse development, thus allowing us to interrogate spatial expression of genes differentially expressed between the posterior and anterior trajectories (Figure 3A). The E7.5 anterior somitic signature shows the strongest positional enrichment in anterior mesoderm, while the posterior signature is enriched in the posterior mesoderm, as well as in the posterior epiblast sections of the Peng et al. dataset (Figure 3B, C). We also performed a similar analysis, in the opposite direction, by extracting the genes enriched respectively in anterior and posterior mesoderm at E7.5 from the Peng et al. dataset (Table S3), and assessed their expression in our single-cell atlas, which highlighted the expected populations of anterior and posterior somitic trajectories (Figure S3C, Methods). This complementary analysis also highlighted additional expression sites (such as in lateral plate mesoderm lineages) for genes on the anterior trajectory. Taken together, this analysis supports the notion that at E7.5, posterior paraxial mesoderm precursor cells are still located close to the primitive streak, while the precursors of anterior paraxial mesoderm have already migrated to the anterior end of the embryo.

The clear separation of the two trajectories at E7.5 suggested they may be spatially segregated at earlier stages. We therefore employed a similar strategy to compare the two trajectories at E7.0 (Figure S3D, Table S2). Genes enriched in the E7.0 posterior paraxial mesoderm ancestors are most strongly associated with the primitive streak region, while genes specific to the E7.0 anterior paraxial mesoderm ancestors show highest enrichment in the mesoderm layer, suggesting that these cells have already ingressed through the primitive streak (Figure S3E). Interestingly, at this stage, genes enriched in the anterior somitic trajectory are expressed in more proximal regions of the egg cylinder compared to those of the posterior trajectory, highlighting an additional spatial segregation of the two sets of ancestors.

Next, we characterized the trajectory leading to NMPs. Comparison with the somitic trajectories suggested an early divergence, but also that ancestors of the posterior somitic tissues had a higher likelihood to contribute to the NMP trajectory than ancestors of anterior somitic tissues (Figure 2A, S3F). Differential gene expression analysis between NMP and somitic trajectory cells at E7.5 showed higher levels of NMP signature genes in NMP-fated cells (e.g., *Cdx1*, *Cdx2*, *Nkx1-2*, *Fst* and *Grsf1*; Gouti et al., 2017) but also a higher expression of epiblast markers *Dnmt3b*, *Epcam* and *Pou5f1* (Figure 3D, Table S4). Conversely, these NMP-fated cells had lower levels of mesoderm maturation genes *Mesp1*, *Aldh1a2*, *Cited1*, *Rspo3* relative to the E7.5 ancestors of somitic tissues. This suggests that at E7.5, the ancestors of NMPs have a more immature, epiblast-like signature compared with the early somite precursors. Consistent with this notion, spatial visualization of this NMP-enriched signature showed highest scores in epiblast sections of the E7.5 embryo (Figure 3E).

This spatiotemporal transcriptional analysis supports a model whereby rostrocaudal patterning of the first somites is concomitant with gastrulation. Mesoderm cells fated to an anterior paraxial fate ingress earlier through the primitive streak and likely acquire their somitic identity anteriorly (marked by an up-regulation of both *Tcf15* and *Meox1* after spatial segregation at E7.5), while

cells destined for a more posterior paraxial fate undergo gastrulation later and develop posteriorly in the embryo. Finally, NMP ancestors remain in the posterior epiblast, where they acquire an NMP signature, sustained up until at least E8.5 (last time point sampled in current atlas, Figure 3F).

T$^{-/-}$ chimaera scRNA-Seq analysis reveals alterations in common and rare cell types

Given the role of *Brachyury* (*T*) in axial elongation, we were intrigued to observe that *T* was the most differentially expressed gene between the two early somitic trajectories (Figure 3A, S3D). The homozygous *T* mutant mouse model is embryonic lethal, with a severe arrest of axis elongation, notochord and allantois defects, and a kinked neural tube (Beddington et al., 1992; Chesley, 1935; Rashbass et al., 1991). To study the cell-autonomous effects of T knockout, we performed single-cell RNA-sequencing (scRNA-Seq) on chimaeric mouse embryos. We generated *T*$^{-/-}$ cell lines from a mouse embryonic stem cell line constitutively expressing tdTomato (Pijuan-Sala et al., 2019; see Methods). We confirmed the disruption of the sequence encoding *T* by sequencing the Crispr/Cas9-targeted locus, which showed frameshift mutations and early stop codons precluding the generation of a functional protein, in two different clones (Figure S4A, B). Chimeric embryos generated with these two independent *T*$^{-/-}$ clones were harvested at E8.5, mutant and wildtype (WT) cells sorted based on tdTomato fluorescence, and scRNA-Seq performed on four independent pools of embryos, with a total of 14,048 *T*$^{-/-}$ and 13,724 WT single cell transcriptomes passing quality control (Figure S4C, Methods). Cell type identities were determined by mapping the chimeric embryo cells onto the reference atlas. As expected, the mutant cells still expressed the *T* transcript, although at reduced levels (Figure S4D, in agreement with self-regulation of this transcription factor; Beisaw et al., 2018). Importantly and in line with mutation analysis by sequencing, there is no detectable T protein in *T*$^{-/-}$ cells of embryo chimeras (Figure S4E), likely due to severe effects on the stability and/or conformation of any residual

peptide produced (as only the first 23% of the amino-acid sequence may be retained; Figure S4B).

We performed differential abundance testing of cell types with reference to matched wild-type chimaeras (Figure S4F, Methods). This demonstrated $T^{-/-}$ cell contribution to intermediate and somitic mesoderm was significantly reduced, while contribution to NMPs was increased (Figure 4A, S4F-G). Other T-expressing tissues including notochord and primordial germ cells (PGCs) also showed changes in differential abundance, but these changes did not reach statistical significance, likely due to the low overall abundance of these cell types in the embryo at this time-point. Interestingly, notochord cells showed perturbed gene expression patterns in $T^{-/-}$ cells (Figure S4H). Reduced contribution of $T^{-/-}$ cells to the PGC lineage has been reported, but no quantitative analysis was performed (Aramaki et al., 2013). Given a quantitative reduction rather than total absence seen by scRNA-Seq analysis of chimaeric embryos, we quantified the numbers of presumptive PGCs present from E7.5 (neural plate) to E8.5 (10 somite) stage in T-expressing embryos, and then compared presumptive PGC numbers at E7.75 (headfold stage) with $T^{-/-}$ embryos (Figure S4I, J). Counting presumptive PGCs in multiple embryos demonstrated that there is indeed a statistically significant reduction in the $T^{-/-}$ samples. Thus, even in rare cell types such as notochord and PGCs, combining mouse chimeras with scRNA-Seq reveals cell-autonomous roles for T.

$T^{-/-}$ chimaera analysis validates the two early somitic trajectories

To investigate the effect of T knockout on the distinct paths towards somite generation, we next focused on the sub-clusters of paraxial mesoderm defined in Figure 1. In agreement with previous findings (Beddington et al., 1992; Rashbass et al., 1991; Wilson et al., 1995), we observed a marked decrease in the contribution to posterior somitic tissue and presomitic mesoderm (Figure 4B). By contrast, cranial mesoderm and anterior somitic tissues showed only small changes in

abundance. This not only supports an essential cell-autonomous role for T specifically in the development of the E8.5 posterior somitic tissue, it also confirms that the two sets of somites present at E8.5 emerge in molecularly distinct developmental events as suggested by the two different trajectories defined above. To obtain a more fine-grained resolution, we assessed the distribution of chimeric cells mapped onto the transcriptional ordering from Figure 1C (Methods). In WT chimeras, tdTomato^{+} (tdTom^{+}) and tdTomato^{-} (tdTom^{-}) cells were similarly distributed ($p=0.14$, Kolmogorov–Smirnov (KS) test). By contrast, *T*$^{-/-}$ cells accumulated in the caudal-most portion of the ordering in *T*$^{-/-}$ chimaeras ($p \leq 10^{-15}$, KS test; Figure 4C).

Since the mapping above was based on transcriptomic features, we next used confocal imaging to visualise the distribution of tdTom^{+} cells in chimeric embryos, confirming a caudal accumulation as predicted from scRNA-Seq data (Figure 4D). Furthermore, examination of confocal Z-stacks of the primitive streak region suggested that caudal accumulation is primarily ectodermal and is therefore a consequence of failure to ingress through the primitive streak (Figure S4K), in agreement with prior observations (Wilson et al., 1995). Importantly, the confocal data also confirms that *T*$^{-/-}$ cells contribute normally to anterior somitic tissues and to other tissues, namely cranial regions, endoderm, cardiac cells, allantois and extra-embryonic mesoderm (Figure S4L, M). Of note, over-representation of *T*$^{-/-}$ cells in the caudal NMP subset supports a previously proposed model (Figure 4A and D), whereby higher levels of T favour ingression through the streak, while lower or absent T expression maintains cells in the streak region where they may ultimately contribute to the tail bud NMP pool (Wilson and Beddington, 1997).

Characterization of NMP over-production in the absence of T

To further investigate the relationships between the developmental trajectories for posterior somites and NMPs, we quantified the contribution of *T*$^{-/-}$ cells across all replicates of our E8.5 chimaeras to posterior somites and NMPs. To control for any potential differences in contribution

to lineages intrinsic to the chimaera assay, we considered the ratio of cell numbers in the injected population divided by the cell numbers in the host population for each lineage, in chimaeras generated by injection of $T^{-/-}$ and WT cells respectively. This confirmed the change in balance between the two lineages (Figure 5A). We next asked whether cells lacking T might already be differentially abundant between these two trajectories at E7.5 (Figure 2B). We thus generated a new set of chimaeras that were harvested at E7.5, and analysed by scRNA-Seq (Methods, Figure 5B, S5A-C). Quantitative analysis across replicate experiments confirmed the trend towards a reduced contribution to the posterior somitic trajectory, although at this stage mutant cells were only slightly overrepresented in the NMP trajectory (Figure 5B). Two other observations are noteworthy. For the E8.5 chimaeras, there is still a small contribution of $T^{-/-}$ cells to posterior somitic mesoderm, meaning that the *T* KO phenotype is not fully penetrant at this stage (Figure S4G). Secondly, when the E7.5 chimaera cells are mapped onto the landscape, a minority of cells is fairly far advanced, whereas the bulk still sits in a territory that overlaps with the NMP trajectory (Figure 5C, D). A model therefore emerges where the earlier cells contributing to the posterior somites may do so even in the absence of T, whereas the rest may be diverted along the NMP trajectory.

$T^{-/-}$ NMPs do not show molecular evidence of an early fate switch

The accumulation of $T^{-/-}$ cells in an NMP transcriptional state in E8.5 chimeras prompted us to characterize this over-represented mutant NMP subset by differential gene expression analysis (Figure 6A, S6A). The majority (75%) of genes differentially expressed in the absence of T were downregulated, suggesting that in these cells T functions mostly as a transcriptional activator (Figure 6A, Table S5). Moreover, 18 of the significantly downregulated genes have previously been identified by ChIP-Seq as direct targets of T in *in vitro* NMP models (Koch et al., 2017), which is significantly more than expected by chance (18 of 47 genes; $p < 10^{-11}$, Fisher's Exact test; Figure 6A, genes highlighted in yellow).

Genes downregulated in *T*$^{-/-}$ NMPs include major elements of the canonical signalling pathways of somitogenesis: Wnt (*Rspo3*), Fgf (*Fgf3*, *Fgf4*, *Fgf8* and *Fgf18*), Notch (*Dll1* and *Hes3*) and retinoic acid (*Cyp26a1*). This is consistent with the previously reported positive feedback loops between T and these pathways during axial extension (Diez del Corral et al., 2003; Hubaud and Pourquie, 2014; Kumar and Duester, 2014; Vermot and Pourquie, 2005). Less-well implicated but likely also important regulators include cell-cell adhesion and signal transduction genes *Sema6a*, *Epha1*, *Itgb8*, *Igfbp3*, *Penk*, *Nrxn1*, and *Fst*. The transcription factors *Mixl1*, *Ets2*, *Mycl* and *Dlx5* were also downregulated, and may therefore play previously unsuspected roles in NMP regulation and somitogenesis downstream of T.

It was proposed that the multipotent nature of NMPs relies on cross-antagonism between *T* and the neural determining factor *Sox2*, where each serves as a lineage-determining factor (Gouti et al., 2017; Koch et al., 2017). Furthermore, in our analysis of gene expression dynamics along the NMP trajectory, we observed a decline in *T* transcript concurrent with the increase in *Sox2* between E7.5 and E8.5 (Figure 2C), which would support this model. Accordingly, *T*$^{-/-}$ NMPs would be expected to express higher levels of *Sox2* than WT NMPs, which would in turn increase the production of spinal cord progenitors (Takemoto et al., 2011). However, neither *Sox2*, nor a broader neural signature, were upregulated in *T*$^{-/-}$ NMPs (Figure 6A, S6B, C). Moreover, spinal cord cells were not overproduced in the *T*$^{-/-}$ chimaeras (Figure 4A). Our analysis of primary cells therefore argues against a cell-autonomous mutually repressive model of *T* and *Sox2* as early NMP fate determinants.

To investigate earlier molecular consequences of *T* knockout, we next performed differential gene expression analysis within E7.5 chimaeras, focusing on the tdTom^{+} and tdTom^{-} cells mapping to each trajectory (Figure 6B; S6D-F, Table S6). There was little overlap between the sets of deregulated genes across the different trajectories, consistent with trajectory-specific effects at this early time-point. Among the genes upregulated in cells fated to anterior somitic tissues was

the T-box family transcription factor *Tbx3*. Cells fated to posterior somitic tissues showed downregulation of genes involved in cell migration including *Vim*, *Pdlim4* and *Htra1* (Fu et al., 2019; Singh et al., 2014; Ye and Weinberg, 2015). These cells also displayed downregulation of *Cited1*, previously shown to label specifically cells that have ingressed through the primitive streak (Garriock et al., 2015). Of note, genes related to an incomplete EMT state (Figure S2B-D) were not affected in $T^{-/-}$ NMP ancestors at any of the analysed time-points (Figure 6A and B).

Taken together, these results suggest that the precursors of anterior mesoderm are capable of undergoing gastrulation in the absence of T. Precursors of more posterior somites reach the streak later in development and require T to activate genes involved in EMT. In the absence of T, they remain in the streak region, where they may contribute to the developing pool of NMPs (Figure 6C).

Discussion

By integrating computational methods with scRNA-Seq of embryonic chimeras, we inferred the molecular maps for three distinct trajectories from pluripotent epiblast cells toward somite development. We revealed previously unknown dynamic gene expression during the emergence of the first, anterior-most somites, accompanied with a clear spatial separation at E7.5. Analysis of $T^{-/-}$ chimaeras validated the three trajectories, suggested reallocation of early posterior somite progenitors to the NMP pool in the absence of T, and supported a model whereby T does not inhibit expression of *Sox2* in NMPs.

To derive likely differentiation trajectories, we took advantage of the WOT approach (Schiebinger et al., 2019) that has the key advantage, compared to many trajectory inference methods, of incorporating real time information when analysing time-course datasets. Trajectory inference methods that do not take real time information into account can produce erroneous assignments when similar cell types are being produced over an extended period of time, or in “waves”. Here,

WOT allowed us to disentangle transcriptional trajectories with relatively similar signatures (in relation to the whole embryonic landscape), but with different time of developmental emergence. Importantly, additional independent analyses using spatial transcriptomic data (Peng et al., 2019), as well as the distinct effects of the *T* knockout in the chimera assays, were consistent with the trajectories inferred from the scRNA-Seq data.

Our results are consistent with a model whereby the first, anterior-most somites develop from mesodermal precursors that ingress early through the primitive streak and migrate anteriorly concurrently with the precursors of other anterior mesoderm tissues. This agrees with previous fate mapping experiments where precursors of the first pairs of somites are found in the same regions of the primitive streak as cardiac and cranial mesoderm, ingressing at around E7.0 (Kinder et al., 1999). The anterior somitic trajectory was characterized by higher levels of previously identified marker genes of lateral plate mesoderm (including *Hand1*, *Prrx1*, *Prrx2*), suggesting a shared ontogeny of the first somites with these progenitors. Different timing of ingression is further supported by the higher expression levels of caudal Cdx/Hox transcription factors in the E8.5 posterior paraxial tissues compared to anterior paraxial tissues, reflecting a later timing of ingression of precursors of posterior paraxial mesoderm (Forlani et al., 2003). One of the most noteworthy observations here is molecular convergence, where both the early anterior and posterior trajectories ultimately acquire a paraxial transcriptional identity, yet the journeys towards that shared identity are temporally, spatially, and molecularly distinct.

Quantitative and molecular analysis of $T^{-/-}$ embryos validated the distinct trajectories. The anterior somitic tissues identified here correspond to the first somite subsets, previously shown to form in the absence of T (Chesley, 1935). In E7.5 chimeric embryos, genes involved in cell migration were specifically downregulated in posterior somite-fated $T^{-/-}$ cells, providing a molecular explanation for previous reports where impaired cell migration was suggested to cause the observed accumulation of mutant cells in the remnants of the primitive streak of chimeric embryos

(Wilson and Beddington, 1997; Wilson et al., 1995). Our data further show that E8.5 caudal accumulation of *T*$^{-/-}$ cells is coupled with the acquisition of an aberrant NMP signature, consistent with the model proposed by Wilson and Beddington (1997), where primitive streak cells harbouring lower levels of T protein remain in the streak throughout gastrulation and contribute to the NMP pool of the developing tail bud to fuel subsequent axial elongation. Further studies will be required to functionally validate whether different levels of T protein regulate the allocation of individual streak cells to paraxial mesoderm or NMPs in the wild type setting.

The ability of anterior paraxial mesoderm precursors to ingress through the streak and migrate anteriorly in the absence of T suggests they rely on other factors. Our data indicate that other members of the T-box protein family may play this role: the anterior somite-fated cells ingress through the streak before E7.0, within the window of *Eomes* expression during gastrulation (Figure 2C), and with considerable overlap with T in gene targets (Tosic et al., 2019). Our molecular analysis revealed *Tbx3* as another possible candidate, due to its specific upregulation at the start of the developmental trajectory towards anterior somitic tissues, and in the E7.5 *T*$^{-/-}$ cells fated to the anterior somitic tissues (Figure 6B and S3B).

In line with prior mouse and zebrafish studies, we also observed a residual contribution of *T*$^{-/-}$ cells to the posterior somitic tissues (Martin and Kimelman, 2010; Wilson and Beddington, 1997). While expression of somitic markers had not been tested in these studies, our results suggest that some of these residual cells are indeed correctly transcriptionally patterned as somitic mesoderm.

Characterization of *T*$^{-/-}$ NMP-like cells suggested a model where T is required for NMPs to move down a somitic differentiation path, but where T has little bearing on NMPs moving along the neural lineage. Furthermore, the observation that many *T*$^{-/-}$ NMPs become trapped in the primitive streak, rather than produce excess neural tissue, suggests that at the single cell level in the intact embryo, many NMPs may not have both somitic and neural differentiation options available to

them, possibly due to spatial constraints. Indeed, although *in vivo* lineage tracing suggest widespread bipotency for larger NMP clones (Tzouanacou et al., 2009), heterotopic transplantation and live cell imaging studies suggest that many cells with NMP potential will only differentiate into one lineage in the embryo (Wood et al., 2019; Wymeersch et al., 2016).

In the present report, we show that single-cell transcriptional analysis of entire embryos provides a complementary approach towards a better understanding of long standing questions in developmental biology. Moving forward, the ability to couple such unbiased transcriptional profiling with information about a cell's location within the organism will further enable new biological discovery. Together with appropriate functional experiments, this promises to open an exciting new chapter in developmental biology, where hypotheses can be investigated *in vivo*, at single cell resolution, genome wide scale, and at the level of the whole organism.

Acknowledgements

We thank William Mansfield and the Gurdon Institute animal facility for blastocyst injections and support in embryo collection, the Flow Cytometry Core Facility at CIMR for cell sorting, Katarzyna Kania at the CRUK-CI genomics core for preparing the chimera 10X libraries, the Wellcome Sanger Institute DNA Pipelines Operations for sequencing, and Rebeca Hannah for re-analysis of the GSM2454138 dataset. Initial analysis of *T* knock-out embryos was inspired by Rosa Beddington, and performed in her laboratory. Research in the authors' laboratories is supported by the Wellcome Trust, MRC, CRUK, Blood Cancer UK, NIH-NIDDK, the Sanger-EBI Single Cell Centre; by core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust-MRC Cambridge Stem Cell Institute; and by core funding from Cancer Research UK and the European Molecular Biology Laboratory. J.A.G. was funded by Wellcome Trust award [109081/Z/15/A]. C.G. was funded by the Swedish Research Council (2017-06278). This work was funded as part of a Wellcome Strategic Award to study cell fate decisions during gastrulation (105031/D/14/Z) awarded to Wolf Reik, Berthold Göttgens, John

Marioni, Jennifer Nichols, Ludovic Vallier, Shankar Srinivas, Benjamin Simons, Sarah Teichmann, and Thierry Voet.

Author contributions

C.G. designed and performed the chimaera single cell analysis experiments, C.G., J.A.G., I.I-R., S.G analysed the data. J.C.M., B.G., J.N., V.W. supervised the study. C.G., J.A.G., J.C.M., B.G., V.W. wrote the manuscript. All authors read and approved the final manuscript.

Declaration of Interests

The authors declare no competing interests.

Main Figures

Figure 1: Two distinct transcriptional subsets of somites at E8.5

(A) UMAP representation of the axial elongation-related tissues present at E8.5.

(B) Schematic of the axial elongation-related tissues in the anatomy of the E8.5 mouse embryo. For colour code, refer to legend in (A).

(C) Distribution of E8.5 axial elongation-related tissues along one-dimensional transcriptional ordering. For colour code, refer to legend in (A).

(D) Marker expression along one-dimensional transcriptional ordering delimits neural and paraxial cell types, including bipotent NMPs. Expression levels are shown as the mean of the expression values in a sliding window of width 10% of the length of the ordering.

(E) Homeobox gene expression distribution provides rostrocaudal orientation of diffusion pseudotime ordering, with bipotent NMPs in the center of the ordering, corresponding to the caudal end of the embryo; and neural and paraxial cell types at the edges of the ordering expressing rostral Hox genes. Expression levels are shown as in (D).

See also Figure S1.

Figure 2: Identification of distinct developmental trajectories towards NMPs and Anterior and Posterior somitic cell subsets

(A) UMAP layout from Pijuan-Sala et al. (2019) highlighting cells belonging to the developmental trajectories for anterior somitic tissues, the newly formed posterior somitic tissues, and NMPs present at E8.5, predicted using WOT analysis. For visualization purposes, the rare populations of shared ancestors were plotted on top.

(B) UMAP layout from Pijuan-Sala et al. (2019) highlighting the same cells as in (A) coloured by sampling time-point.

(C) Gene expression dynamics along the three developmental trajectories reveals distinct transcriptional programs. y-axis: mean $\log_2$(normalised counts).

See also Methods, Figure S2 and Table S1.

Figure 3: Anterior-posterior patterning of paraxial mesoderm during gastrulation

(A) Differential expression analysis of E7.5 cells with predicted posterior somitic fate vs E7.5 cells with predicted anterior somitic fate. Genes queried individually in the eGastrulation tool (see (B and C)) are highlighted in bold.

(B) Overall “activity score” of the genes significantly enriched in the anterior trajectory (top) for E7.5 spatial data (Peng et al., 2019) and expression levels in $\log_{10}$(FPKM+1) for selected genes (bottom) highlighted in bold font in (A). Cornplots were generated using the eGastrulation tool, where the embryo is represented by anatomical sections featuring anterior-posterior and left-right axes for sections in distinct proximal-distal regions (10 being most proximal and 1 most distal; EA: Anterior Endoderm; MA: Anterior Mesoderm; A: Anterior epiblast; L1: Anterior Left lateral; R1:

485 Anterior Right lateral; L2: Posterior Left lateral; R2: Posterior Right lateral; P: Posterior epiblast;
486 MP: Posterior Mesoderm; EP: Posterior Endoderm).

487 (C) Overall “activity score” of the genes significantly enriched in the posterior trajectory (top),
488 and expression levels in $\log_{10}$(FPKM+1) for selected genes (bottom) highlighted in bold font in
489 (A). See also legend for panel (B).

490 (D) Differential gene expression of E7.5 cells with predicted NMP fate vs E7.5 cells with
491 predicted somitic fate. Adjusted p value is calculated for differential gene expression in the cells
492 with predicted NMP fate compared to either the anterior or the posterior somitic-fated cells. Log_2
493 fold-change is shown for posterior somitic mesoderm cells only. Genes queried individually in the
494 eGastrulation tool (see (E)) are highlighted in bold.

495 (E) Overall “activity score” of the genes significantly enriched in the NMP trajectory (top), and
496 expression levels in $\log_{10}$(FPKM+1) for selected genes (bottom) highlighted in bold font in (D).
497 See also legend for panel (B).

498 (F) Schematic of anterior-posterior patterning of paraxial mesoderm during gastrulation.
499 Tissues fated to the E8.5 anterior somites are coloured in red, those fated to the E8.5 posterior
500 somites are colored in yellow, and those fated to the E8.5 NMP pool are colored in green. A:
501 anterior; P: posterior.

502 See also Figure S2 and Tables S2 to S4.

503 **Figure 4: Development of *T*$^{-/-}$ cells in chimeric embryos reveals a differential**
504 **requirement of T in two developmental trajectories leading to somitic tissues**

505 (A) Differential abundance testing of cell types with most pronounced effects in *T*$^{-/-}$ chimeras
506 compared to WT controls, as well as other cell types relevant to axial elongation. *: BH-corrected
507 $p<0.1$, n=4 independent experiments.

(B) Differential abundance testing of the somitic subclusters identified in Figure 1 in $T^{-/-}$ chimeras compared to WT controls. * BH-corrected p<0.1, n=4 independent experiments.

(C) Density of mapped chimera cells along the one-dimensional diffusion pseudotime ordering characterized in Figure 1.

(D) Confocal image of a $T^{-/-}$ chimeric embryo stained with Phalloidin-Atto488 (green). Arrowhead points to accumulation of $tdTom^{+}$ cells in the caudal region of the embryo (red). *: somites; nt: neural tube; n: node; Scale bar: 100µm.

See also Figure S4.

Figure 5: Assessing allocation of $T^{-/-}$ cells to the NMP pool

(A) Relative contribution of injected cells to NMPs vs Posterior somites in E8.5 chimeras (p-values calculated by permutation). Each point is an independent experiment (pool of chimeric embryos), and calculated as: relative ratio = (number of $tdTom^{+}$ NMPs / number of $tdTom^{-}$ NMPs) / (number of $tdTom^{+}$ posterior somite cells / number of $tdTom^{-}$ posterior somite cells). Hollow circles: values for WT chimera assays; filled circles: values for $T^{-/-}$ chimeras.

(B) Relative contribution of injected cells to trajectories towards NMPs vs Posterior somites in E7.5 chimeras, showing significant bias towards the NMP fate in $T^{-/-}$ chimeras compared to WT (p-values estimated by permutation; values plotted as in (A)).

(C) UMAP layout from Pijuan-Sala et al. (2019), highlighting mapped nearest neighbours of injected ($tdTom^{+}$) and host cells ($tdTom^{-}$) in E7.5 and E8.5 chimeras.

(D) UMAP layout from Pijuan-Sala et al. (2019) with cells coloured by their relative mass from NMP vs posterior trajectories. Values are capped at -5 and 5 for better legibility. Arrowhead highlights the nascent mesoderm cell subset with balanced mass (i.e. equal likelihood) for both trajectories, according to WOT.

Figure 6: A two-step regulatory role of T in mammalian paraxial mesoderm – formation of the first posterior somites and establishment of the NMP pool for subsequent axis elongation

(A) Differential gene expression between E8.5 mutant cells accumulated in an NMP state and their WT counterparts within chimeric embryos (see inset and Figure 4C). Genes previously found to be bound by T (Koch et al., 2017) are highlighted in yellow.

(B) Differentially expressed genes in tdTom^{+} *T*$^{-/-}$ cells in E7.5 chimeric embryos compared to their tdTom^{-} WT counterparts (adjusted p<0.1), within the transcriptomes mapping to each of the developmental trajectories highlighted in Figure 2A and B. Genes also identified as differentially expressed in control chimeras (injected with WT tdTom^{+} cells) or significantly correlated with the tdTomato transcript were considered as results of a chimera assay-related technical bias, and excluded from the analysis (Figure S6D-F).

(C) Working model for cell-autonomous role of T in the formation of the first somites during gastrulation.

See also Figure S6 and Tables S5 and S6.

Star Methods

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, John C. Marioni (john.marioni@cruk.cam.ac.uk).

Materials Availability

Mouse embryonic stem cell lines generated in this study are available upon request.

Data and Code Availability

Raw sequencing data is available on Arrayexpress: T chimeras – E-MTAB-8811; WT chimeras – E-MTAB-7324 (as used in Pijuan-Sala et al., 2019) and E-MTAB-8812 (newly generated). Processed data is available from the Bioconductor package MouseGastrulationData (_____). This includes the single-cell RNA-seq data directly, as well as the NMP orderings, and somitogenesis trajectory labels used in this manuscript. An online visualisation tool is available at https://marionilab.cruk.cam.ac.uk/EarlySomites2020/.

EXPERIMENTAL MODELS AND SUBJECT DETAILS

Cell lines

All mouse embryonic stem cell lines were expanded under the 2i+LIF conditions (Ying et al., 2008), in a humidified incubator at 37°C and 7% CO_2, and routinely tested negative for mycoplasma infection. A male, karyotypically normal, tdTomato-expressing mouse embryonic stem cell line was derived from E3.5 blastocysts obtained by crossing a male ROSA26tdTomato (Jax Labs – 007905) with a wildtype C57BL/6 female. Competence for chimaera generation was assessed using morula aggregation assay. Targeting of the *T* locus was performed using the CRISPR/Cas9 system (see Method Details), mutant clones were assessed by next-generation sequencing (see Figure S4). Two mutant clones were used to generate $T^{-/-}$ embryonic chimeras.

Mouse models

All procedures were performed in strict accordance to the UK Home Office regulations for animal research. Chimeric mouse embryos were generated under the project licence number PPL 70/8406. Animals used in this study were 6-10 week-old females, maintained on a lighting regime of 14 hours light and 10 hours darkness with food and water supplied ad libitum. For chimera generation, E3.5 blastocysts were derived from wildtype C57BL/6 matings, and after injection of the mutant cells, the resulting chimeric embryos were transferred to C57BL/6 recipient females at 0.5 days of pseudopregnancy following mating with vasectomised males.

METHOD DETAILS

Somitic trajectory analysis from atlas data

Subclustering the atlas paraxial cell types. To dissect the Paraxial Mesoderm sub-populations present in the E8.5 embryo, cells from the reference Atlas (Pijuan-Sala et al., 2019) belonging to E8.5 time-point and to the cell types “Paraxial Mesoderm” and “Somitic Mesoderm” were extracted and re-clustered using igraph's Louvain algorithm. Clustering was performed on Mutual Nearest Neighbours (MNN) batch corrected principal components (top 50), and the resulting subclusters were annotated using differentially expressed genes. ***Transcriptional ordering of axial elongation cell types.*** The Atlas data were subset to E8.5 cells of spinal cord, NMP, caudal epiblast, caudal mesoderm, somitic mesoderm, and paraxial mesoderm cell types. A 50-dimensional principal component (PC) space was generated from these cells from log-transformed normalised gene counts (with an added pseudocount of 1), considering only highly-variable genes (HVGs, see Selection of HVGs in the “quantification and statistical analysis” section, below). Expression levels for each gene were centred, but not scaled, prior to PC computation. PCs were calculated using the irlba package. To ensure that the atlas manifold was continuous in the PC subspace, and so that batch-effects could not affect mapping of chimaera

data, it was batch-corrected as described below (see 'Batch correction'). As the manifold is largely a one-dimensional structure (see Figure 1A), it was summarised into a one-dimensional ordering using diffusion pseudotime (DPT; Haghverdi et al., 2016). DPT was computed from a diffusion map, itself computed from the atlas cells in the PC subspace, with DPT ordering from the spinal cord cell with most extreme value of the first diffusion component. ***Identifying somitic developmental trajectories.*** To reconstruct the lineages of cells in the reference atlas, we used the W-OT package 1.0.7 (Schiebinger et al., 2019) to estimate the sequence of ancestor distributions at earlier time points. Cells were allocated to the trajectory of their largest endpoint mass contribution, or to multiple trajectories if their mass contribution was at least 90% as large as their largest endpoint mass contribution (to capture apparently uncommitted cells). ***Spatial domains of trajectory-specific expression signatures.*** Genes that defined the posterior and anterior somitic trajectories at E7.0 and E7.5 (determined by differential expression, with adjusted P value < 0.1; differential expression testing was performed using the scran function findMarkers using default parameters) were introduced into the Gene Activity Score tool provided by the eGastrulation database (http://egastrulation.sibcb.ac.cn/; Peng et al., 2019) to generate 2-dimensional "corn plots". For the reverse analysis (Figure S3C), signature genes enriched in anterior and posterior mesoderm domains in the Peng et al. dataset were retrieved using the "Gene Search by Pattern" tool provided by the eGastrulation database. The following patterns were used as input: anterior – value of 80 for rows 3 to 7 of MA column (remaining slots were given 0); posterior – value 80 for rows 3 to 7 of MP column and value 60 for rows 3 to 7 of P column. Cutoff for correlation analysis: RCC > 0.4. We transformed the atlas expression levels onto a common scale (as a Z-score for each gene), and plotted the average Z-score of the Peng et al. signature genes on our transcriptional Atlas layout, which highlighted the expected populations of anterior and posterior somitic trajectories (Figure S3C). For details on gene expression comparisons along trajectories, see "quantification and statistical analysis" section below.

Chimera generation and sequencing

Embryo collection. All procedures were performed in strict accordance to the UK Home Office regulations for animal research under the project license number PPL 70/8406. ***Chimera generation.*** TdTomato-expressing mouse embryonic stem cells (ESC) were derived as previously described (Pijuan-Sala et al., 2019). Briefly, ESC lines were derived from E3.5 blastocysts obtained by crossing a male ROSA26tdTomato (Jax Labs – 007905) with a wildtype C57BL/6 female, expanded under the 2i+LIF conditions (Ying et al., 2008) and transiently transfected with a Cre-IRES-GFP plasmid (Wray et al., 2011) using Lipofectamine 3000 Transfection Reagent (ThermoFisher Scientific, #L3000008) according to manufacturer's instructions. A tdTomato-positive, male, karyotypically normal line, competent for chimaera generation as assessed using morula aggregation assay, was selected for targeting *T*. Two guides were designed using the http://crispr.mit.edu tool (guide 1: TGACGGCTGACAACCACCGC; guide 2: GCCCCAAAATTGGGCGAGTC) and were cloned into the pX458 plasmid (Addgene, #48138) as previously described (Ran et al., 2013). The obtained plasmids were then used to transfect the cells and single transfected clones were expanded and assessed for Cas9-induced mutations. Genomic DNA was isolated by incubating cell pellets in 0.1 mg/ml of Proteinase K (Sigma, #03115828001) in TE buffer at 50°C for 2 hours, followed by 5 min at 99°C. The sequence flanking the guide-targeted sites was amplified from the genomic DNA by polymerase chain reaction (PCR) in a Biometra T3000 Thermocycler (30 sec at 98°C ; 30 cycles of 10 sec at 98°C, 20 sec at 58°C, 20 sec at 72°C; and elongation for 7 min at 72°C) using the Phusion High-Fidelity DNA Polymerase (NEB, #M0530S) according to the manufacturer's instructions. Primers including Nextera overhangs were used (F-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCCGGTGCTGAAGGTAAAT; R-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTGCTTAACCCTCATCAGC), allowing library preparation with the Nextera XT Kit (Illumina, #15052163), and sequencing was performed

using the Illumina MiSeq system according to manufacturer's instructions. Two ESC clones showing frameshift mutations in exon 2 resulting in the functional inactivation of T were selected for injection into C57BL/6 E3.5 blastocysts. A total of 17 chimeric embryos were harvested at E8.5, dissected, and single-cell suspensions were generated from three independent pools of embryos by TrypLE Express dissociation reagent (Thermo Fisher Scientific) incubation for 7-10 minutes at 37°C under agitation. Single-cell suspensions were sorted into $tdTom^{+}$ and $tdTom^{-}$ samples using a BD Influx sorter with DAPI at 1µg/ml (Sigma) as a viability stain for subsequent 10X scRNA-seq library preparation (version 2 chemistry), and sequencing using the Illumina HiSeq 4000 platform, which resulted in 13,724 $tdTom^{-}$ and 14,048 $tdTom^{+}$ cells that passed quality control (see "Single-cell RNA sequencing analysis" below). To exclude transcriptional effects intrinsic to the chimera assay, chimeric embryos were generated by injecting the parental $tdTom^{+}$ $T^{+/+}$ (WT) line into C57BL/6 E3.5 blastocysts and processed as for the $T^{-/-}$ samples. Three independent embryo pools with a total of 13 embryos were used for scRNA-seq, and 1,077 $tdTom^{-}$ and 2,454 $tdTom^{+}$ cells passed quality control. ***Embryo staining and imaging.*** Following dissection, embryos were washed in PBS and fixed in 4% paraformaldehyde (PFA, Thermo Scientific) for 1 hour at room temperature. They were then washed three times for 15 minutes in wash buffer (0.1% fraction 5 bovine serum albumin, 0.1% Tween20, 5% DMSO, 0.1% Triton-X in PBS), permeabilized overnight at 4°C in permeabilization buffer (0.1% fraction 5 bovine serum albumin, 0.1% Tween20, 5% DMSO, 0.25% Triton-X in PBS) and washed three times for 15 minutes in wash buffer. Embryos were then incubated overnight in blocking solution (5% donkey serum and 1% BSA in wash buffer) at 4°C, washed three times for 15 minutes in wash buffer and incubated overnight at 4°C in blocking solution containing the goat anti mouse Brachyury primary antibody (1:200, R&D Systems, cat# AF2085). After three 15 minute washes, Phalloidin-AlexaFluor488 (Thermofisher Scientific) was added 1:1000 and 4′,6-Diamidino-2-phenylindole dihydrochloride (DAPI, Sigma) was added at 200ng/ml with the donkey anti-goat Alexa647 antibody (1:500, Invitrogen, cat# A21447) in blocking solution for another overnight incubation at

4°C. Embryos were then washed three times for 15 minutes in wash buffer and mounted in Vectashield mounting media (Vector laboratories, cat# H-1000) and imaged in a Confocal Leica TCS SP5 microscope. Images were captures with the Leica Application Suite software and processed for publication using Fiji.

Quantification of primordial germ cells

Following dissection, embryos were stained for Alkaline phosphatase activity as described previously (Ginsburg et al., 1990). Briefly, embryos were fixed in absolute ethanol with 12.5% glacial acetic acid at 4°C for 1 hour, followed by two 24h incubations in absolute ethanol at 4°C and two 1h washes in chloroform. They were mounted in wax, sectioned and incubated in freshly made staining solution (0.1mg/ml 1-Naphthyl phosphate, 0.5% borax solution, 0.5mg/ml Fast Red TR salt and 0.6% MgCl2, pH 9.2) for 15-30 minutes. For genotyping, extra-embryonic tissues of each embryo were digested with Proteinase K and tested by polymerase chain reaction for the presence of a 310bp region including the 3' coding region of the T gene, missing in $T^{-/-}$ embryos (primers: CCAGTTGACACCGGTTGTTACA and TATCCCAGTCTCTGGTCTGT). A 350bp fragment spanning the homeodomain of Hox 2.1 was used as a positive control (primers: GCGCCAGTGCAGGGAAGATTGGAA and GATATGACTGGGCCAGACGGAAA) (Rashbass et al., 1994).

Single-cell RNA sequencing analysis

10X data pre-processing. Raw files were processed with CellRanger 2.1.1 using default mapping arguments, with reads mapped to the mm10 genome and counted with GRCm38.92 annotation, including tdTomato sequence. This older annotation was used to ensure consistency with the reference atlas (Pijuan-Sala et al., 2019). Processed data and raw count matrices are available in the Bioconductor package MouseGastrulationData. ***Swapped molecule removal.*** Molecule counts that derived from barcode swapping were removed from all 10X samples by applying the DropletUtils function swappedDrops (default parameters) to groups of samples

(where a sample is a single lane of a 10X Chromium chip) that were multiplexed for sequencing. ***Cell calling.*** Cell barcodes that were associated with real cell transcriptomes were identified using emptyDrops (Lun et al., 2019), which assesses whether the RNA content associated with a cell barcode is statistically significantly distinct from the ambient background RNA present within each sample. A minimum UMI threshold was set at 5,000, and cells with an adjusted p-value < 0.01 (BH-corrected) were considered for further analysis. ***Quality control.*** Cells with mitochondrial gene expression fractions greater than 2.52% and 2.90%, for the $T^{-/-}$ chimaeras and WT chimaeras respectively, were excluded. These thresholds were determined by the data – we considered a median-centred MAD-variance normal distribution; cells with mitochondrial read fraction "outside" of the upper end of this distribution were excluded (adjusted p-value < 0.05; BH-corrected). ***Normalisation.*** Transcriptome size factors were calculated for each dataset separately ($T^{-/-}$ chimeras, WT chimeras), using computeSumFactors from the R scran package (Lun et al., 2019), using default parameters. Raw counts for each cell were divided by their size factors, and the resulting normalised counts were used for further processing.

Visualisation of Single-cell RNA sequencing data

Batch correction. Batch-effects were removed using the fastMNN function in scran on the first 50 PCs, computed from the HVG-subset logcount matrix. Default parameters were used. When correcting the reference atlas (Pijuan-Sala et al., 2019), correction was performed first between the samples within each time-point, merging sequentially from the samples containing the most cells to the samples containing the least. Time-points were then merged from oldest to youngest. When correcting the chimaeras, correction was performed on all samples within a genotype first, from largest sample to smallest, then across the two genotypes. ***UMAPs*** were calculated using the uwot R package with default parameters except for min_dist = 0.7. ***Diffusion maps*** were calculated using the R package destiny, with function DiffusionMap, using default settings. Batch-corrected principal components were used.

Chimaera cell type annotation

To annotate the cell types in the chimaeric embryos, we performed a transcriptional mapping to a large reference atlas of mouse embryonic development (Pijuan-Sala et al., 2019). Each stage of the atlas was sub-sampled at random to 10,000 cell libraries (i.e., including the technical artefacts of doublets and stripped nuclei) at each time-point. Cells from the mixed time-point were excluded. This subsampling reduces potential bias due to the different number of cells captured at each stage. Stages E6.5 and E6.75 contained fewer cells than other stages (3,697 and 2,169 respectively) and were not downsampled; however we do not expect cells from E8.5 or E7.5 chimeras to map to these time-points. A shared 50-dimensional PC subspace was constructed from the subsampled cells from the atlas, and all chimaera cells that were to be mapped. Batch-correction was then performed on the atlas cells in the PC space, as described above (Batch correction), to construct a single contiguous reference manifold. Samples to be mapped were then independently mapped onto the newly-corrected atlas data (scran function fastMNN), and the 10 nearest cells (by Euclidean distance) in the atlas to each chimera cell were recorded. Mapped time-point and cell type of chimera cells were defined as the most frequent of those of its 10 nearest-neighbours. Ties were broken by choosing the stage or cell type of the cell that had the lowest distance to the chimera cell. Cells that mapped to doublet- or stripped nucleus-labelled cells were excluded from downstream analyses. For cell type differential abundance testing in chimeric embryos, see “quantification and statistical analysis” section below.

Mapping chimaera cells onto the atlas backbone.

To map chimaera cells onto their appropriate positions on the atlas manifold, they were mapped onto it using a strategy similar to that used in Batch correction (above). Individual samples (i.e. one 10X channel) of the E8.5 chimaera datasets were mapped onto the corrected atlas using fastMNN, using coordinates from the PC subspace. This operation was repeated for each chimaera sample, retaining the mapped coordinate values for each cell. Performing this operation

in parallel across samples prevents any mapped chimaera cells affecting the future mapping of other samples. For the spinal cord to head mesoderm ordering, mapping was performed only using cells from the relevant cell types. DPT values (i.e., ordering positions) were inferred for chimaera cells by considering the mean DPT value for the 5 nearest atlas cells in the PC space, after performing the per-sample mapping. This value of DPT is, effectively, the position of a chimaera cell along the atlas backbone. For mapping chimaera cells to somite trajectories through the atlas, chimaera cells were mapped to the whole atlas (excluding cells from the “mixed gastrulation” atlas time-point, and with the subsampling described above), as above for cell type labelling. As for the previous approach, chimaera cells were considered a part of a trajectory if the most common trajectory state of their 10 nearest neighbours was one of the somite trajectories. For differential gene expression analyses, see “quantification and statistical analysis” section below.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of single-cell datasets

Selection of HVGs. HVGs were calculated using trendVar and decomposeVar from the scran R package, with loess span 0.05. Genes that had significantly higher variance than the fitted trend (BH-corrected $p < 0.05$) were retained. Genes with mean $\log_2$(normalised count) $< 10^{-3}$, genes on the Y chromosome, *Xist*, and *tdTomato* were excluded. *Gene expression comparisons along trajectories.* First, we selected genes that were variable along any of the three trajectories. We took the union of the genes calculated in each of the three trajectories, calculated according to the following procedure, considering only the cells from that trajectory: HVGs were first identified (see Selection of HVGs, below), and their mean expression level at each time-point was calculated; an order three polynomial linear model fit was compared to an intercept-only model by F-test (i.e., R function *anova*). We considered genes to be variable along a trajectory if the polynomial fit was significantly better than the intercept-only model (BH-corrected $p < 0.1$). In a pairwise manner across the three trajectories, we then tested these genes for differences in expression along them. As above, we calculated the mean expression level in each trajectory for the genes at each time-point. We then fitted a null model of an order three polynomial (i.e., the same model as for selecting genes above, except with the model using data from two, rather than one, trajectories at each time point). The alternative model allowed for trajectory-specific coefficients for each coefficient of the order three polynomial. We then compared the fit of the two models (by F-test) and considered genes to show different patterns of expression along the trajectories if they were fit better by the alternative model (BH-corrected FDR < 0.01). If the latter model fits better than the null, this suggests that the data are better described by different polynomials for each trajectory.

Overlap computation (GSEA). Following pair-wise comparisons of expression dynamics along the entire length of transcriptional trajectories (Table S1), resulting gene lists were used as input

for computing overlap with the Molecular Signatures Database Hallmark gene set collection using the Gene Set Enrichment Analysis tool (Liberzon et al., 2015; Subramanian et al., 2005). Results were plotted in Figure S2E using the calculated FDR q-values, analog of hypergeometric p-value after correction for multiple hypothesis testing according to Benjamini and Hochberg (Subramanian et al., 2005).

Analysis of embryonic chimeras

Differential abundance testing was performed using edgeR (McCarthy et al., 2012). Each 10x sample was considered as a replicate, and mapped cell type counts were used in place of gene counts. A separate linear model was fitted for E7.5 and E8.5 chimaeras. Each linear model contained an intercept value specific to each biological replicate (i.e., pools of chimaeric embryos – one sample $tdTom^+$ and the other $tdTom^-$). A factor term was included for the injected samples from the WT chimaeras, and another was included for the injected samples from the $T^{-/-}$ chimaeras. Differential abundance was tested using the contrast between these two factor terms, effectively asking whether the injected cell type frequency differed between the WT and $T^{-/-}$ chimaeras. This approach is preferable to testing entirely within the $T^{-/-}$ chimaeras, where the $tdTom^-$ fraction of cells may be influenced by aberrant behaviour of the $T^{-/-}$ cells. The intra-chimaera approach is also vulnerable to confounding injected status (which may subtly affect cell behaviours) with genotype; the inter-chimaera approach is not confounded. The use of wild-type chimaeras also allows incorporation of the intrinsic variability of a mutation-free chimaera system into the model. Finally, the use of edgeR allows sharing of uncertainty estimates across cell types with similar frequency in this sample-limited experiment. edgeR models were fitted and contrasts tested using the functions calcNormFactors, glmQLFit, and glmQLFTest.

Differential expression analyses. Differential expression testing was performed using the scran function findMarkers using default parameters. There was one exception. For the across-background NMP differential expression (Figure 6A), cells were selected with DPT values

between 1.25 and 1.6. However, different distributions of cells along this section could induce apparent differential expression due to positions along ordering, rather than due to differences in genetic background. Here, we used the more sophisticated edgeR model, where we also fit the centred DPT values as a model coefficient to control for different distributions along the cell ordering. For this model, we tested against an absolute $\log_2$ fold-change of 0.5 as the edgeR model proved extremely sensitive to very small differences in expression level.

Relative ratio comparisons. In Figure 5A and B, relative contribution of injected cells to NMPs vs Posterior somites trajectories are calculated in E8.5 and E7.5 embryonic chimeras, respectively. Each point corresponds to an independent experiment (pool of chimeric embryos), and calculated as: relative ratio = (number of $tdTom^+$ on NMPs trajectory / number of $tdTom^-$ on NMPs trajectory) / (number of $tdTom^+$ on posterior somites trajectory / number of $tdTom^-$ on posterior somites trajectory). This approach is robust to chimera-wide composition effects, as cell numbers are normalised using the host cells from each sample. To assess the difference in ratios between chimera types (i.e. WT into WT vs $T^{-/-}$ into WT chimeras). p-values were estimated from 1000 permutations of the cells' trajectory labels.

Quantification of primordial germ cells

Differences in Alkaline Phosphatase-positive PGC counts in *T*-expressing vs $T^{-/-}$ mouse embryos at the headfold stage were assessed using an unpaired two-sample t-test (Figure S4J).

ADDITIONAL RESOURCES

The code used to perform these analyses is available at https://github.com/MarioniLab/TChimeras2020. A singularity image that contains the exact

versions of software used can be downloaded from the Github repository. An online visualisation tool is available at https://marionilab.cruk.cam.ac.uk/EarlySomites2020/.

Supplemental Tables

Supplemental Table 1 (related to Figure 2):

Results of trajectory comparison for variable genes across trajectories. Lower p and FDR values correspond to more distinct expression patterns over time, genes with higher p and FDR values have more similar expression patterns. See also Methods.

Supplemental Table 2 (related to Figure 3 and S3):

Differential expression analysis results comparing cells fated towards Posterior Somitic tissues vs cells fated towards Anterior Somitic tissues at E7.5 and E7.0.

Supplemental Table 3 (related to Figure S3):

Genes enriched in the anterior and posterior mesoderm of the Peng et al. (2019) dataset; lists resulting from the “Gene Search by Pattern” tool provided by the eGastrulation database (see Methods), and used as input for Z-score calculations in Figure S3c.

Supplemental Table 4 (related to Figure 3):

Differential expression analysis results comparing cells fated towards NMP vs cells fated towards Somitic tissues at E7.5.

Supplemental Table 5 (related to Figure 6):

Differential expression analysis results comparing cells $tdTom^{+}$ vs $tdTom^{-}$ cells within the overrepresented NMP subset in E8.5 $T^{-/-}$ chimeras.

Supplemental Table 6 (related to Figure 6):

Differential expression analysis results comparing cells tdTom^{+} vs tdTom^{-} cells within the E7.5 cell subsets fated towards Anterior somitic tissues, Posterior somitic tissues and NMP *T*$^{-/-}$ chimeras and WT control chimeras.

References

Aldea, D., Subirana, L., Keime, C., Meister, L., Maeso, I., Marcellini, S., Gomez-Skarmeta, J.L., Bertrand, S., and Escriva, H. (2019). Genetic regulation of amphioxus somitogenesis informs the evolution of the vertebrate head mesoderm. Nat Ecol Evol *3*, 1233-1240.

Angerer, P., Haghverdi, L., Buttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 32, 1241-1243.

Aramaki, S., Hayashi, K., Kurimoto, K., Ohta, H., Yabuta, Y., Iwanari, H., Mochizuki, Y., Hamakubo, T., Kato, Y., Shirahige, K.*, et al.* (2013). A mesodermal factor, T, specifies mouse germ cell fate by directly activating germline determinants. Dev Cell *27*, 516-529.

Beddington, R.S., Rashbass, P., and Wilson, V. (1992). Brachyury--a gene affecting mouse gastrulation and early organogenesis. Dev Suppl, 157-165.

Beisaw, A., Tsaytler, P., Koch, F., Schmitz, S.U., Melissari, M.T., Senft, A.D., Wittler, L., Pennimpede, T., Macura, K., Herrmann, B.G.*, et al.* (2018). BRACHYURY directs histone acetylation to target loci during mesoderm development. EMBO Rep *19*, 118-134.

Bessho, Y., Miyoshi, G., Sakata, R., and Kageyama, R. (2001). Hes7: a bHLH-type repressor gene regulated by Notch and expressed in the presomitic mesoderm. Genes Cells *6*, 175-185.

Burgess, R., Rawls, A., Brown, D., Bradley, A., and Olson, E.N. (1996). Requirement of the paraxis gene for somite formation and musculoskeletal patterning. Nature *384*, 570-573.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J.*, et al.* (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496-502.

Cao, Y., Zhao, J., Sun, Z., Zhao, Z., Postlethwait, J., and Meng, A. (2004). fgf17b, a novel member of Fgf family, helps patterning zebrafish embryos. Dev Biol *271*, 130-143.

Chal, J., Oginuma, M., Al Tanoury, Z., Gobert, B., Sumara, O., Hick, A., Bousson, F., Zidouni, Y., Mursch, C., Moncuquet, P.*, et al.* (2015). Differentiation of pluripotent stem cells to muscle fiber to model Duchenne muscular dystrophy. Nat Biotechnol *33*, 962-969.

Chapman, D.L., Agulnik, I., Hancock, S., Silver, L.M., and Papaioannou, V.E. (1996). Tbx6, a mouse T-Box gene implicated in paraxial mesoderm formation at gastrulation. Dev Biol *180*, 534-542.

888 Chapman, D.L., and Papaioannou, V.E. (1998). Three neural tubes in mouse embryos with
889 mutations in the T-box gene Tbx6. Nature *391*, 695-697.

890 Chesley, P. (1935). Development of the short-tailed mutant in the house mouse. Journal of
891 Experimental Zoology *70*.

892 Dastjerdi, A., Robson, L., Walker, R., Hadley, J., Zhang, Z., Rodriguez-Niedenfuhr, M., Ataliotis,
893 P., Baldini, A., Scambler, P., and Francis-West, P. (2007). Tbx1 regulation of myogenic
894 differentiation in the limb and cranial mesoderm. Dev Dyn *236*, 353-363.

895 Dias, A., Lozovska, A., Wymeersch, F.J., Novoa, A., Binagui-Casas, A., Sobral, D., Martins, G.G.,
896 Wilson, V., and Mallo, M. (2020). A Tgfbr1/Snai1-dependent developmental module at the core
897 of vertebrate axial elongation. Elife *9*.

898 Diez del Corral, R., Olivera-Martinez, I., Goriely, A., Gale, E., Maden, M., and Storey, K. (2003).
899 Opposing FGF and retinoid pathways control ventral neural pattern, neuronal differentiation, and
900 segmentation during body axis extension. Neuron *40*, 65-79.

901 Forlani, S., Lawson, K.A., and Deschamps, J. (2003). Acquisition of Hox codes during gastrulation
902 and axial elongation in the mouse embryo. Development *130*, 3807-3819.

903 Fu, C., Li, Q., Zou, J., Xing, C., Luo, M., Yin, B., Chu, J., Yu, J., Liu, X., Wang, H.Y.*, et al.* (2019).
904 JMJD3 regulates CD4 T cell trafficking by targeting actin cytoskeleton regulatory gene Pdlim4. J
905 Clin Invest *129*, 4745-4757.

906 Galceran, J., Sustmann, C., Hsu, S.C., Folberth, S., and Grosschedl, R. (2004). LEF1-mediated
907 regulation of Delta-like1 links Wnt and Notch signaling in somitogenesis. Genes Dev *18*, 2718-
908 2723.

909 Garriock, R.J., Chalamalasetty, R.B., Kennedy, M.W., Canizales, L.C., Lewandoski, M., and
910 Yamaguchi, T.P. (2015). Lineage tracing of neuromesodermal progenitors reveals novel Wnt-
911 dependent roles in trunk progenitor cell maintenance and differentiation. Development *142*, 1628-
912 1638.

913 Ginsburg, M., Snow, M.H., and McLaren, A. (1990). Primordial germ cells in the mouse embryo
914 during gastrulation. Development *110*, 521-528.

915 Gouti, M., Delile, J., Stamataki, D., Wymeersch, F.J., Huang, Y., Kleinjung, J., Wilson, V., and
916 Briscoe, J. (2017). A Gene Regulatory Network Balances Neural and Mesoderm Specification
917 during Vertebrate Trunk Development. Dev Cell *41*, 243-261 e247.

918 Griffiths, J.A., Richard, A.C., Bach, K., Lun, A.T.L., and Marioni, J.C. (2018). Detection and
919 removal of barcode swapping in single-cell RNA-seq data. Nat Commun 9, 2667.

920 Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime
921 robustly reconstructs lineage branching. Nat Methods *13*, 845-848.

922 Henrique, D., Abranches, E., Verrier, L., and Storey, K.G. (2015). Neuromesodermal progenitors
923 and the making of the spinal cord. Development *142*, 2864-2875.

924 Hubaud, A., and Pourquie, O. (2014). Signalling dynamics in vertebrate segmentation. Nat Rev
925 Mol Cell Biol *15*, 709-721.

926 Kassar-Duchossoy, L., Giacone, E., Gayraud-Morel, B., Jory, A., Gomes, D., and Tajbakhsh, S.
927 (2005). Pax3/Pax7 mark a novel population of primitive myogenic cells during development.
928 Genes Dev *19*, 1426-1431.

929 Keynes, R.J., and Stern, C.D. (1988). Mechanisms of vertebrate segmentation. Development *103*,
930 413-429.

931 Kinder, S.J., Tsang, T.E., Quinlan, G.A., Hadjantonakis, A.K., Nagy, A., and Tam, P.P. (1999).
932 The orderly allocation of mesodermal cells to the extraembryonic structures and the
933 anteroposterior axis during gastrulation of the mouse embryo. Development *126*, 4691-4701.

934 Koch, F., Scholze, M., Wittler, L., Schifferl, D., Sudheer, S., Grote, P., Timmermann, B., Macura,
935 K., and Herrmann, B.G. (2017). Antagonistic Activities of Sox2 and Brachyury Control the Fate
936 Choice of Neuro-Mesodermal Progenitors. Dev Cell *42*, 514-526 e517.

937 Kumar, S., and Duester, G. (2014). Retinoic acid controls body axis extension by directly
938 repressing Fgf8 transcription. Development *141*, 2972-2977.

939 Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015).
940 The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417-425.

941 Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis
942 of single-cell RNA-seq data with Bioconductor. F1000Res 5, 2122.

943 Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., participants in the 1st Human Cell
944 Atlas, J., and Marioni, J.C. (2019). EmptyDrops: distinguishing cells from empty droplets in
945 droplet-based single-cell RNA sequencing data. Genome Biol *20*, 63.

946 Mankoo, B.S., Skuntz, S., Harrigan, I., Grigorieva, E., Candia, A., Wright, C.V., Arnheiter, H., and
947 Pachnis, V. (2003). The concerted action of Meox homeobox genes is required upstream of
948 genetic pathways essential for the formation, patterning and differentiation of somites.
949 Development *130*, 4655-4664.

950 Marongiu, M., Marcia, L., Pelosi, E., Lovicu, M., Deiana, M., Zhang, Y., Puddu, A., Loi, A., Uda,
951 M., Forabosco, A*., et al.* (2015). FOXL2 modulates cartilage, skeletal development and IGF1-
952 dependent growth in mice. BMC Dev Biol *15*, 27.

953 Martin, B.L., and Kimelman, D. (2008). Regulation of canonical Wnt signaling by Brachyury is
954 essential for posterior mesoderm formation. Dev Cell *15*, 121-133.

955 Martin, B.L., and Kimelman, D. (2010). Brachyury establishes the embryonic mesodermal
956 progenitor niche. Genes Dev *24*, 2778-2783.

957 McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor
958 RNA-Seq experiments with respect to biological variation. Nucleic Acids Res *40*, 4288-4297.

Morgani, S.M., Metzger, J.J., Nichols, J., Siggia, E.D., and Hadjantonakis, A.K. (2018). Micropattern differentiation of mouse pluripotent stem cells recapitulates embryo regionalized cell fate patterning. Elife *7*.

Nandkishore, N., Vyas, B., Javali, A., Ghosh, S., and Sambasivan, R. (2018). Divergent early mesoderm specification underlies distinct head and trunk muscle programmes in vertebrates. Development *145*.

Nowotschin, S., Ferrer-Vaquer, A., Concepcion, D., Papaioannou, V.E., and Hadjantonakis, A.K. (2012). Interaction of Wnt3a, Msgn1 and Tbx6 in neural versus paraxial mesoderm lineage commitment and paraxial mesoderm differentiation in the mouse embryo. Dev Biol *367*, 1-14.

Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J.*, et al.* (2019). A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. Science *365*.

Peng, G., Suo, S., Cui, G., Yu, F., Wang, R., Chen, J., Chen, S., Liu, Z., Chen, G., Qian, Y.*, et al.* (2019). Molecular architecture of lineage allocation and tissue organization in early mouse embryo. Nature *572*, 528-532.

Peters, H., Wilm, B., Sakai, N., Imai, K., Maas, R., and Balling, R. (1999). Pax1 and Pax9 synergistically regulate vertebral column development. Development *126*, 5399-5408.

Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L.*, et al.* (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature *566*, 490-495.

Pourquie, O. (2001). Vertebrate somitogenesis. Annu Rev Cell Dev Biol *17*, 311-350.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc *8*, 2281-2308.

Rashbass, P., Cooke, L.A., Herrmann, B.G., and Beddington, R.S. (1991). A cell autonomous function of Brachyury in T/T embryonic stem cell chimaeras. Nature *353*, 348-351.

Rashbass, P., Wilson, V., Rosen, B., and Beddington, R.S. (1994). Alterations in gene expression during mesoderm formation and axial patterning in Brachyury (T) embryos. Int J Dev Biol *38*, 35-44.

Rodrigues, S., Santos, J., and Palmeirim, I. (2006). Molecular characterization of the rostral-most somites in early somitic stages of the chick embryo. Gene Expr Patterns *6*, 673-677.

Sakai, Y., Meno, C., Fujii, H., Nishino, J., Shiratori, H., Saijoh, Y., Rossant, J., and Hamada, H. (2001). The retinoic acid-inactivating enzyme CYP26 is essential for establishing an uneven distribution of retinoic acid along the anterio-posterior axis within the mouse embryo. Genes Dev *15*, 213-225.

Sambasivan, R., Kuratani, S., and Tajbakhsh, S. (2011). An eye on the head: the development and evolution of craniofacial muscles. Development *138*, 2401-2415.

995 Sato, T., Rocancourt, D., Marques, L., Thorsteinsdottir, S., and Buckingham, M. (2010). A
996 Pax3/Dmrt2/Myf5 regulatory cascade functions at the onset of myogenesis. PLoS Genet *6*,
997 e1000897.

998 Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu,
999 S., Lin, S., Berube, P.*, et al.* (2019). Optimal-Transport Analysis of Single-Cell Gene Expression
1000 Identifies Developmental Trajectories in Reprogramming. Cell *176*, 1517.

1001 Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S.,
1002 Rueden, C., Saalfeld, S., Schmid, B., *et al.* (2012). Fiji: an open-source platform for biological-
1003 image analysis. Nat Methods *9*, 676-682.

1004 Shih, H.P., Gross, M.K., and Kioussi, C. (2007). Expression pattern of the homeodomain
1005 transcription factor Pitx2 during muscle development. Gene Expr Patterns *7*, 441-451.

1006 Singh, H., Nero, T.L., Wang, Y., Parker, M.W., and Nie, G. (2014). Activity-modulating monoclonal
1007 antibodies to the human serine protease HtrA3 provide novel insights into regulating HtrA
1008 proteolytic activities. PLoS One *9*, e108235.

1009 Steventon, B., and Martinez Arias, A. (2017). Evo-engineering and the cellular and molecular
1010 origins of the vertebrate spinal cord. Dev Biol *432*, 3-13.

1011 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich,
1012 A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a
1013 knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci
1014 U S A *102*, 15545-15550.

1015 Takada, S., Stark, K.L., Shea, M.J., Vassileva, G., McMahon, J.A., and McMahon, A.P. (1994).
1016 Wnt-3a regulates somite and tailbud formation in the mouse embryo. Genes Dev *8*, 174-189.

1017 Takahashi, Y., Inoue, T., Gossler, A., and Saga, Y. (2003). Feedback loops comprising Dll1, Dll3
1018 and Mesp2, and differential involvement of Psen1 are essential for rostrocaudal patterning of
1019 somites. Development *130*, 4259-4268.

1020 Takemoto, T., Uchikawa, M., Yoshida, M., Bell, D.M., Lovell-Badge, R., Papaioannou, V.E., and
1021 Kondoh, H. (2011). Tbx6-dependent Sox2 regulation determines neural or mesodermal fate in
1022 axial stem cells. Nature *470*, 394-398.

1023 Tosic, J., Kim, G.J., Pavlovic, M., Schroder, C.M., Mersiowsky, S.L., Barg, M., Hofherr, A., Probst,
1024 S., Kottgen, M., Hein, L.*, et al.* (2019). Eomes and Brachyury control pluripotency exit and germ-
1025 layer segregation by changing the chromatin state. Nat Cell Biol *21*, 1518-1531.

1026 Tzouanacou, E., Wegener, A., Wymeersch, F.J., Wilson, V., and Nicolas, J.F. (2009). Redefining
1027 the progression of lineage segregations during mammalian embryogenesis by clonal analysis.
1028 Dev Cell *17*, 365-376.

1029 Vermot, J., and Pourquie, O. (2005). Retinoic acid coordinates somitogenesis and left-right
1030 patterning in vertebrate embryos. Nature *435*, 215-220.

1031 Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018).
1032 Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science
1033 *360*, 981-987.

1034 Wahl, M.B., Deng, C., Lewandoski, M., and Pourquie, O. (2007). FGF signaling acts upstream of
1035 the NOTCH and WNT signaling pathways to control segmentation clock oscillations in mouse
1036 somitogenesis. Development *134*, 4033-4041.

1037 Wilson, V., and Beddington, R. (1997). Expression of T protein in the primitive streak is necessary
1038 and sufficient for posterior mesoderm movement and somite differentiation. Dev Biol *192*, 45-58.

1039 Wilson, V., Manson, L., Skarnes, W.C., and Beddington, R.S. (1995). The T gene is necessary
1040 for normal mesodermal morphogenetic cell movements during gastrulation. Development *121*,
1041 877-886.

1042 Wilson, V., Olivera-Martinez, I., and Storey, K.G. (2009). Stem cells, signals and vertebrate body
1043 axis extension. Development *136*, 1591-1604.

1044 Wood, T.R., Kyrsting, A., Stegmaier, J., Kucinski, I., Kaminski, C.F., Mikut, R., and Voiculescu,
1045 O. (posted 2019). Neuromesodermal progenitors separate the axial stem zones while producing
1046 few single- and dual-fated descendants. bioRXiv. doi: https://doi.org/10.1101/622571

1047 Wray, J., Kalkan, T., Gomez-Lopez, S., Eckardt, D., Cook, A., Kemler, R., and Smith, A. (2011).
1048 Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network
1049 and increases embryonic stem cell resistance to differentiation. Nat Cell Biol *13*, 838-845.

1050 Wymeersch, F.J., Huang, Y., Blin, G., Cambray, N., Wilkie, R., Wong, F.C., and Wilson, V. (2016).
1051 Position-dependent plasticity of distinct progenitor types in the primitive streak. Elife *5*, e10042.

1052 Xu, X., Li, C., Takahashi, K., Slavkin, H.C., Shum, L., and Deng, C.X. (1999). Murine fibroblast
1053 growth factor receptor 1alpha isoforms mediate node regression and are essential for posterior
1054 mesoderm development. Dev Biol *208*, 293-306.

1055 Ye, X., and Weinberg, R.A. (2015). Epithelial-Mesenchymal Plasticity: A Central Regulator of
1056 Cancer Progression. Trends Cell Biol *25*, 675-686.

1057 Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith,
1058 A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.

1059 Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B.,
1060 Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional
1061 profiling of single cells. Nat Commun *8*, 14049.