

A transcriptional and regulatory map of mouse somitogenesis

Ximena Ibarra-Soria^{1,6,*#}, Elodie Thierion^{1,*}, Gi Fay Mok², Andrea E. Münsterberg², Duncan T. Odom^{1,3,#}, John C. Marioni^{1,4,5,#}

¹ Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK.

² School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK.

³ DKFZ, Division of Regulatory Genomics and Cancer Evolution B270, Im Neunheimer Feld 280, Heidelberg, 69120, Germany.

⁴ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK.

⁵ Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK.

⁶ Present address: Genomic Sciences, GSK R&D, Stevenage, SG1 2NY, UK.

* Contributed equally

Correspondence: marioni@ebi.ac.uk, d.odom@dkfz-heidelberg.de, ximena.x.ibarra-soria@gsk.com

ABSTRACT

The mammalian body plan is shaped by rhythmic segmentation of mesoderm into somites, which are transient embryonic structures consisting of hundreds of cells that form down each side of the neural tube. We have systematically analysed the genome-wide transcriptional and chromatin dynamics occurring within nascent somites, from early inception of somitogenesis to the latest stages of body plan establishment. We created matched gene expression and open chromatin maps for the three leading pairs of somites at six time points during embryonic development. Here we show that the rate of somite differentiation accelerates as development progresses. We identified a conserved maturation programme followed by all somites after segmentation, but somites from more developed embryos concomitantly switch on differentiation programmes from derivative cell lineages soon after segmentation. Integrated analysis of the somitic transcriptional and chromatin activities revealed opposing regulatory modules controlling the onset of differentiation. We identified transcription factors expressed during early development that inhibit the activity of proteins required for commitment and differentiation of skeletal cell populations. Our results provide a powerful, high-resolution view of the molecular genetics underlying somitic development in mammals.

INTRODUCTION

The segmentation of the body plan during early embryogenesis is a fundamental and conserved feature of all vertebrate species. It results in the metamerized organisation of the vertebrae and the associated skeletal muscles, nerves, and blood vessels. This segmentation

5 is established via formation of somites, which are transient embryonic structures consisting of hundreds of cells that bud off from the anterior tip of the presomitic mesoderm (PSM) on each side of the neural tube. Each pair of somites is symmetrically and rhythmically formed along the anterior-posterior axis according to the *clock and waveform model* (Cooke and Zeeman 1976).

10

This model integrates spatiotemporal information from waves of transcriptionally oscillating genes in the PSM (the molecular clock) and antagonistic signalling gradients along the embryo axis (the waveform). The molecular oscillator is known as the segmentation clock, which drives cyclic and synchronised gene expression along the PSM (Palmeirim et al. 1997). The so-called

15 *clock genes* belong to the Notch, Wnt and FGF signalling pathways (Dequéant and Pourquié 2008). The waveform involves posterior gradients of Wnt and FGF signalling that are counteracted by an opposing gradient of retinoic acid secreted from the somites (Bénazéraf and Pourquié 2013). When the segmentation clock reaches cells that have passed the waveform, segmentation genes, including *Mesp2*, are activated, leading to the specification of

20 the somite boundary (Saga 2012). As well as specifying the somite boundaries, retinoic acid signalling suppresses signals that break left-right symmetry, ensuring that somite production is bilaterally symmetric (J. Vermot and Pourquié 2005; Julien Vermot et al. 2005). This periodic addition of somites underlies body plan generation in all vertebrates, and the oscillating signals from Notch, Wnt and FGF pathways are conserved in the PSM of model organisms as diverse

25 as mouse, chicken, and zebrafish (Krol et al. 2011).

The specification of somites along the anterior-posterior axis is determined before somitogenesis by Hox gene expression (Krumlauf 1994) and the specific combination of Hox genes expressed along the axis establishes the identity of the resulting vertebrae (Wellik

30 2007). Somites are further patterned along the dorso-ventral and medio-lateral axes, giving rise to two somitic derivatives found in all vertebrates: the sclerotome (precursor of vertebral and rib cartilage, tendons, and blood vessels) and the dermomyotome (precursor of skeletal muscles and back dermis). Fate specification to either derivative is controlled by signals from adjacent tissues. Ventral cells of the somite differentiate into sclerotome under the influence of Shh signals from the notochord and the floor-plate of the neural tube. Dorsal cells instead receive Wnt signals from the neural tube and the ectoderm and BMP4 from the lateral mesoderm, to give rise to the dermomyotome (Weldon and Münsterberg 2022).

Master transcriptional regulators driving somite differentiation have been identified through classical genetic approaches (Yusuf and Brand-Saberi 2006; Christ, Huang, and Scaal 2007). However, how these master regulators orchestrate somitogenesis through embryonic space
5 and time, and indeed what genes they directly regulate, remains less clear. While several studies have used gene expression microarrays to characterise gene expression patterns during somitogenesis, these studies have all been performed in the presomitic mesoderm (Krol et al. 2011; Dequéant et al. 2006; Ozbudak, Tassy, and Pourquié 2010).

10 Here, we map the transcriptional and chromatin changes that occur across somitogenesis by performing high-resolution RNA and ATAC-sequencing of individual, manually microdissected somites at six developmental stages. By comparing the three most recently segmented somites, we characterised the molecular basis of the earliest stages of somite maturation. Additionally, we identified patterns of dynamic regulatory activity across development, with
15 pronounced differences between somites that give rise to differing types of vertebrae. By characterising the biological processes dominating each stage, we found that somite differentiation accelerates with developmental progression. Finally, we used the combined information from the transcriptional and chromatin maps to define regulatory modules with differing activity during early and late development. These molecular programmes control the
20 onset of differentiation, thus regulating the timing of skeletal system development.

RESULTS

A high resolution transcriptional and regulatory map of somitogenesis

To characterise the transcriptional changes that orchestrate mouse somitogenesis, we generated coupled transcriptional and chromatin accessibility profiles of individual somite pairs, across embryonic development. Each somite typically contains 500-1000 cells, which is sufficient to generate high-resolution small bulk data. We first compared the transcriptomes of matched left and right somites dissected from 20-25 somite embryos, and observed no significant differences in expression (Figure S1), indicating that from a molecular genetics perspective, the two somites were indistinguishable. Therefore, for each somite pair, we used one somite to map the transcriptome (RNA-seq) and one to map matched open chromatin (ATAC-seq) (Figure 1A).

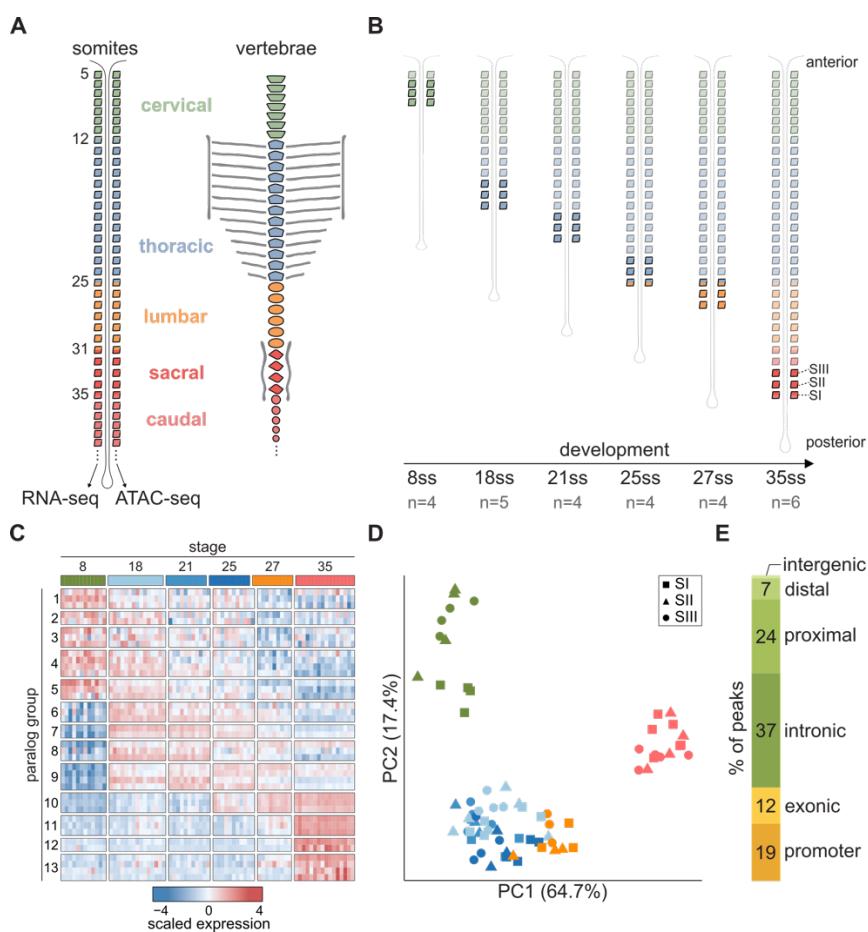


Figure 1 | Expression and chromatin profiling of mouse somitogenesis. **A)** Schematic of somite pairs on each side of the neural tube, and the corresponding vertebrae structures they will form. One somite from each pair was used for RNA-seq, and the other for ATAC-seq. Somites and vertebrae are coloured based on their vertebral identity (cervical, thoracic, lumbar, sacral or caudal). Somites at boundaries between two different types of vertebrae are numbered. The first four somite pairs (occipital) are not shown. **B)** Somites collected in this study. From each embryo, the three most posterior somites (SI-SIII) were collected. Embryos profiled were from six different developmental stages, determined by

the number of somites. ss = somite stage. n = number of embryos collected. **C)** Heatmap of the expression of all Hox genes. Samples are ordered in columns according to their observed somite stage. Each row is a different Hox gene, ordered by paralogous groups from 1 to 13. Expression is represented as z-scores. **D)** Principal component analysis of the expression of Hox genes orders somites consistently with their observed somite stage. **E)** Proportion of open chromatin regions classified based on their genomic context.

After segmentation, somites maintain a round shape for several hours before undergoing an epithelial-to-mesenchymal transition (EMT) when cells commit to somite-derived lineages and initiate migration (Yusuf and Brand-Saberi 2006). To study the molecular changes associated

5 with fate commitment, we collected the three most posterior pairs of somites, which correspond to those most recently segmented, and that have not yet begun EMT (Christ, Huang, and Scaal 2007; M. Jacob, Christ, and Jacob 1975) (Figure 1B). To understand how somitogenesis progresses across embryonic development, we sampled these somite trios from embryos at 10 six different developmental stages. We defined the embryonic stage by counting the total number of somite pairs, and profiled at least four different embryos containing 8, 18, 21, 25, 27 and 35 pairs of somites (Figure 1B and Table S1). These stages span four of the five different types of vertebrae (cervical, thoracic, lumbar and sacral; Figure 1A), providing profiles of somites that will contribute to all four structures.

15 We generated matched transcriptome (RNA-seq) and open chromatin (ATAC-seq) maps from the vast majority (71/81) of the samples (Tables S1). From the 77 RNA-seq libraries, all but one produced good-quality transcriptomes (Tables S2). We normalised for sequencing depth and corrected for batch effects associated with the date of somite collection (Methods; Figure S2).

20 Somites from specific axial levels express particular combinations of Hox genes (Wellik 2007), which are directly associated with segment identity (Kmita and Duboule 2003; Mallo, Wellik, and Deschamps 2010). Somites from different embryonic stages consistently showed clear 25 differences in the class and expression level of Hox genes (Figure 1C), and the expression of Hox genes alone accurately ordered samples according to our observed somite stage (Figure 1D).

30 ATAC-seq libraries were successfully produced from 75 samples, but 25 of these were removed after applying stringent quality control criteria (Figure S3 and Table S3). The remaining 50 open chromatin maps showed efficiency biases, which were correlated with mean fragment abundance (Figure S4A-B). To compensate for this trend, we used a loess-

based normalisation strategy (Figure S4C). Additionally, we applied the same batch correction approach that was used on the RNA-seq data to remove technical variation (Figure S4D-E).

We classified the possible functional role of open chromatin regions based on their genomic location. Peaks that were within 200 bp of an annotated transcription start site were deemed promoter-like elements and represent 19% of total peaks; an additional 12% of peaks overlapped gene exons. The remainder of the peaks were annotated as enhancer-like elements, and subdivided into *proximal* (24%) or *distal* (7%) if they were within 25 and 100 kb of an annotated gene, respectively; or *intergenic* (1%) (Figure 1E).

10

Epithelial somites deploy a shared maturation programme across embryonic development

We systematically profiled the three most recently segmented somites, which are at the beginning of the differentiation process that will give rise to all somitic derivatives, including muscle, bone, cartilage and dermis (H. J. Jacob, Christ, and Jacob 1974; Christ et al. 1992; Ordahl and Le Douarin 1992; Aoyama 1993). Following the nomenclature proposed by Christ and Ordahl (Christ and Ordahl 1995), we refer to each somite in these trios as somites I, II and III, from the most posterior to the most anterior, respectively (Figure 2A).

20

To characterise the molecular changes underlying somite maturation, we compared all pairwise combinations of somites I, II and III at each developmental stage. We identified a median of 453 genes that significantly differ per stage (FDR < 5% and |fold-change|>1.5). Most differentially expressed genes had subtle changes in expression, with half showing less than a two-fold difference between any two somites. To increase statistical power, we repeated the 25 analysis using samples from different stages as replicates, and detected genes that showed consistent changes regardless of developmental stage. Altogether, we identified 2,977 significantly differentially expressed genes. Similar numbers of genes were up and down-regulated, with the strongest differences manifested between somites I and III (Figure 2B). The vast majority of differentially expressed genes (75.8%) showed consistent expression 30 dynamics across different stages. However, most genes (86.9%) also showed differences in expression levels across developmental time, illustrating the complex regulatory dynamics prevalent during embryonic development (Figure 2C).

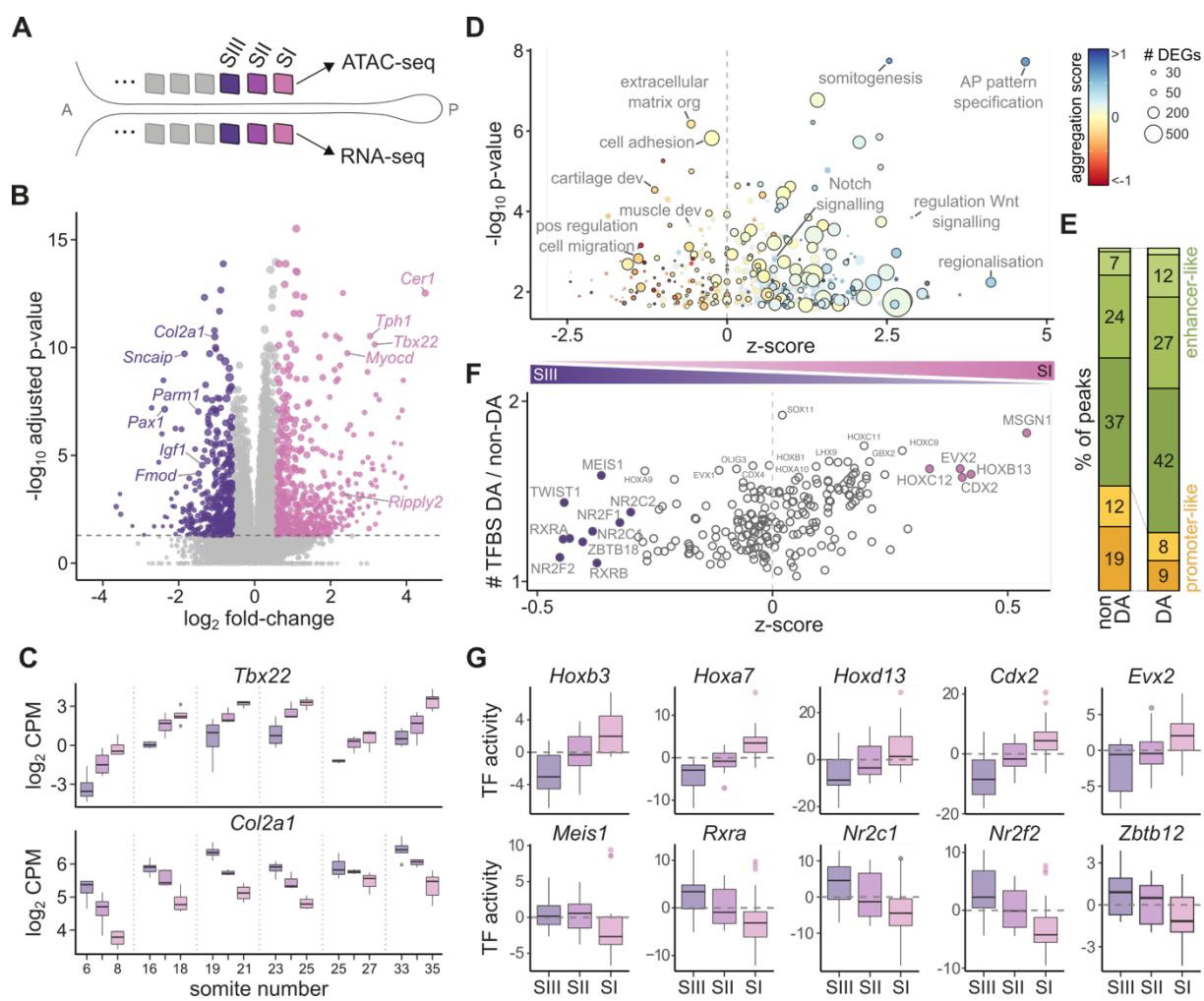


Figure 2 | Somites follow a conserved maturation programme across development. **A)** Schematic indicating the somite trios profiled from each embryo. A: anterior; P: posterior. Colour and shading scheme is preserved throughout all figures to indicate the different somites. **B)** Volcano plot of expression changes between somites I and III. Genes significantly differentially expressed are coloured. **C)** Differences in gene expression for two representative genes (*Tbx22* and *Col2a1*) among somites I,II,III are consistently maintained across developmental stages. **D)** Significantly enriched Gene Ontology functional categories in the set of differentially expressed genes. Enrichment significance is shown on the y axis. The x-axis indicates whether a term contains a majority of genes that are downregulated (positive) or upregulated (negative). Points are coloured based on an 'aggregation score', which corresponds to the average fold-change of all differentially expressed genes in the GO term. The size of the points indicates the number of differentially expressed genes in each term. Outlined points correspond to terms that are also significantly enriched in the set of differentially accessible chromatin regions. **E)** Barplot of the proportion of peaks falling in different genomic contexts. Differentially accessible (DA) regions between somites I,II,III are enriched for enhancers. Colours indicate the same classes as in Figure 1E. **F)** Similar to D but showing the enrichment of transcription factor binding sites (TFBS) in differentially accessible peaks. **G)** Representative examples of motif activity dynamics for TFs that are enriched in differentially accessible peaks. Positive (negative) activity scores indicate the regions are more (less) accessible compared to background chromatin. Hox and

other hoemeodomain TF binding sites close in mature somites, while C4 zinc finger class of receptors (*Rxra*, *Nr2c1*, *Nr2f2*, *Zbtb12*) sites become more accessible in SIII.

The genes downregulated along somite maturation were enriched for biological processes related to regionalisation and pattern specification, which are active in the PSM and lead to somite segmentation. Consistently, both the Wnt and Notch signalling pathways were

5 preferentially downregulated (Dequéant and Pourquié 2008) (Figure 2D). In contrast, a steady progression towards EMT during somitogenesis was reflected by the upregulation of cell adhesion and migration programs, together with a switch to positive regulation of Rho and ERK signalling (Figure 2D).

10 The open chromatin landscape was similarly dynamic across the somite trios, with 2,701 genomic regions showing significantly different accessibility levels (FDR < 5% and $|fold-change| > 1.5$). Open chromatin regions that actively changed between somites were enriched for enhancer-like regions, with fewer promoter elements (Figure 2E). Indeed, only 506 (18.7%) differentially accessible regions were located within 5 kb of a differentially expressed gene,
15 indicating that the regulatory mechanisms driving expression changes operate through distal regulatory elements, rather than by directly modulating chromatin accessibility at promoters. The coordination of dynamic chromatin accessibility and gene expression changes was also reflected in their shared over-representation of the same biological functions (Figure 2D).

20 Next, we annotated transcription factor (TF) binding motifs within ATAC-seq peaks and identified 201 regulators whose binding motifs were significantly enriched in the dynamic regions, when compared to static open chromatin (Figure 2F). These included Hox factors, as well as multiple members of the homeodomain, Tal, Sox and NK families. For example, binding motifs for MSGN1 were present in 22% of all dynamic regions (compared to 12% in non
25 differentially accessible chromatin), and most of these peaks showed reduced accessibility in more mature somites, consistent with the role of this protein as a master regulator of PSM differentiation (Chalamalasetty et al. 2014) (Figure 2F). In contrast, the dynamic peaks with binding motifs for TWIST1, a critical factor mediating EMT, were more accessible in the most mature somites (Lamouille, Xu, and Derynck 2014) (Figure 2F).

30 Finally, to understand how the overrepresented transcription factors regulate somite maturation, we analysed the accessibility dynamics of the genomic loci with binding sites for each of these TFs. Binding sites for all Hox proteins, regardless of their paralogous group or stage activity pattern, showed decreased accessibility upon somite maturation (Figure 2G).

35 This behaviour also extended to most of the other TFs enriched in differentially accessible

peaks (Figure 2G). One notable exception gained accessibility as somites matured: the C4 zinc finger class of receptors, which includes the retinoid X receptor-related factors, critical in mediating the biological effects of retinoid signalling and its differentiation-inducing activity (Draut, Liebenstein, and Begemann 2019) (Figure 2G).

5

Molecular remodelling across development regulates somite responses to the signalling environment

Our data revealed profound changes in transcriptional and regulatory activity in somites I-III across development. We compared the RNA-seq profiles of somites among all different

10 developmental stages (Figure 3A) and identified 10,691 genes with significant changes in expression (FDR < 5% and $|fold-change| > 1.5$; see Methods for details; Figure 3B), including most known transcription factors (838 from a total of 1,310 expressed). The chromatin landscape was also remodelled extensively, with 33,013 open chromatin regions showing significant differences in accessibility (Figure S5A). In contrast to the changes observed across

15 somite maturation, a much higher proportion of differentially accessible loci were in promoters or close to differentially expressed genes (Figure S5B), indicating that across development widespread chromatin remodelling plays a crucial role in controlling the genes available for expression.

20 When ordered by developmental stage, somites from different vertebral fates showed waves of temporally restricted transcriptional and chromatin remodelling (Figure 3B, S5A). We analysed these patterns of coordinated gene expression by performing enrichment analysis of Gene Ontology functional terms. Differentially expressed genes with highest expression in cervical somites (clusters 1-4) were related to epithelial cell development and response to 25 retinoic acid signalling, including several retinoic acid receptors (Figure 3B-C). Additional genes involved in somitogenesis and embryonic patterning were prevalent in both cervical and thoracic somites (clusters 1-6, Figure 3C), and were generally expressed at the highest level in the most-recently segmented somite. This suggests somites at these early developmental stages closely resemble their PSM lineage and are only beginning to activate a somite-specific 30 transcriptional profile.

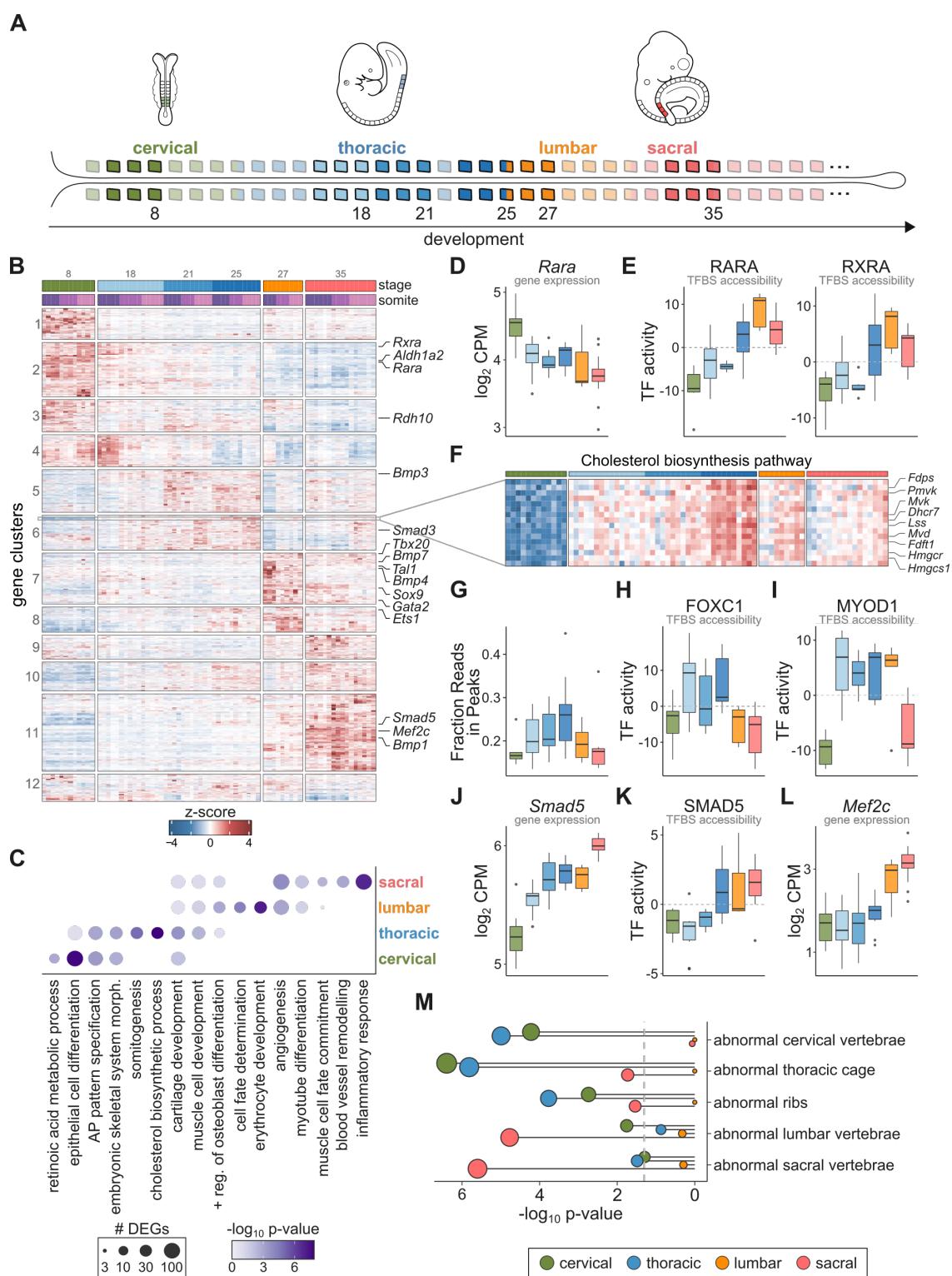


Figure 3 | Epithelial somites at late development activate differentiation programmes of derivative lineages absent in early stages. A) Schematic indicating the somites profiled across development, and their vertebral fate. Colour scheme is preserved throughout all figures to indicate the different stages. **B)** Heatmap of expression of genes differentially expressed across development. Samples (columns) are ordered based on their somite number, and their stage and somite level are indicated at the top. **C)** Gene ontology term enrichment analysis results for sets of genes with highest activity at particular vertebral fates. The size of the circles indicates the number of differentially

expressed genes in each term-fate combination, and the intensity of the colour corresponds to the significance of the enrichment. **D)** Expression of the retinoic acid receptor *Rara* across development. Related factors such as *Rxra* show a similar pattern. **E)** Chromatin activity scores from chromVAR for the genome-wide binding sites (TFBS) of RARA and RXRA. Positive/negative scores indicate higher/lower accessibility than background chromatin. **F)** Zoomed-in region of the heatmap in B, showing the expression of genes in the cholesterol biosynthesis pathway. **G)** Same as E but for FOXC1. **H-I)** MYOD1 chromatin activity at its TFBSs (H) and gene expression (I) across stages. **J-K)** *Smad5* gene expression (J) and chromatin activity at its TFBSs (K) across stages. **L)** Gene expression levels for *Mef2c* across development. **M)** Significance scores for enrichment of chromatin regions associated with genes that show skeletal abnormalities in KO mice. The significance of each term is shown separately for the sets of regions with highest activity at each vertebral fate, indicated by the colour of the circle. Circle size is proportional to the significance level.

We used our data to dissect the complex interplay between metabolite production, TF activity and chromatin dynamics involved in retinoic acid (RA) signalling, which requires precise spatiotemporal regulation for adequate differentiation of progenitor cells (Draut, Liebenstein, 5 and Begemann 2019). RA signalling effects are mediated by the retinoic acid receptor (RAR) and retinoid X receptor (RXR) families. These ligand-dependent transcription factors can recruit either corepressors or coactivators to induce changes in chromatin condensation and regulate transcription (Draut, Liebenstein, and Begemann 2019). Expression of the enzymes involved in RA production (*Aldh1a2* and *Rdh10*) as well as of several RAR/RXR TFs peaked 10 early in development (Figure 3B,D). However, the accessibility of loci with binding sites for RAR/RXR factors was lowest at this stage (Figure 3E), suggesting their association with corepressors to induce chromatin condensation. As development proceeded these chromatin loci progressively increased in accessibility, maintaining an open-chromatin configuration from stage 25 onwards (Figure 3E). These data indicate that the epigenetic profile of somites is 15 reshaped across development from a repressive to a permissive state for RA signalling activity.

Thoracic somites showed strong enrichment for genes involved in the development of the skeletal system, including both the muscle and cartilage lineages (clusters 5-6; Figure 3C). We observed prominent expression of many components of the TGFbeta, BMP and Smad 20 signalling pathways, which are fundamental in orchestrating skeletal system development (Wu, Chen, and Li 2016) (Figure 3B). We also observed coordinated expression of cholesterol biosynthesis, with maximal expression of 17 metabolically central genes at stage 25 before being downregulated at later stages (Figure 3F). Cholesterol plays important roles in the transduction of hedgehog signalling (S. Xu and Tang 2022; Stottmann et al. 2011) and is 25 required for the correct development of muscle and bone (Campos et al. 2015; Anderson et al. 2020). Sonic hedgehog (SHH) is secreted by the notochord and controls the specification

of the sclerotome during patterning of epithelial somites (Murtaugh, Chyung, and Lassar 1999). Defective cholesterol biosynthesis leads to impaired response to Shh signalling and skeletal defects (Stottmann et al. 2011; S. Xu and Tang 2022). As expected, we did not detect significant *Shh* expression in the somites. However, the tightly controlled expression of the 5 cholesterol pathway components suggests a mechanism to control when somites are most responsive to extrinsic hedgehog signalling.

Somite differentiation accelerates across development

Our data revealed that chromatin remodelling is concentrated in thoracic somites. In addition 10 to activating the gene programmes controlling skeletal system development, somites at stages 18 to 25 generally showed higher levels of open chromatin, compared to other stages. We observed a sharp increase in the fraction of reads in peaks in somites from stage 18 embryos, with further increases at stages 21 and 25, before dropping in the stage 27 somites (Figure 3G). Consistently, over half of all differentially accessible chromatin loci showed highest 15 accessibility in thoracic somites (Figure S5A). These chromatin loci were enriched for many transcription factor motifs, including several with well described roles in skeletal system development such as forkhead TFs, implicated in both skeletal muscle and cartilage development (Sanchez, Candau, and Bernardi 2014; J. Xu et al. 2021) (Figure 3H). We also observed increased accessibility at loci harbouring binding sites for MYOD1 (Figure 3I) and 20 MYF5 (Figure S5C), which are essential for cell commitment to the myogenic lineage (Chal and Pourquié 2017).

Previous work (Borman and Yorde 1994; Berti et al. 2015; Gi Fay Mok, Mohammed, and Sweetman 2015; Maschner et al. 2016) have characterised the expression dynamics of 25 sclerotome and myotome markers, including *Myod1*, at several embryonic stages. These studies showed that somites from younger embryos take longer to activate marker gene expression compared to somites from more advanced embryos. We hypothesised that the shorter times required for marker expression onset in late development stem from a change in the permissiveness of the chromatin landscape, which allows lineage-defining TFs to activate 30 their downstream pathways sooner. Analysis of the active biological processes prevalent at later developmental stages indeed revealed a switch from cell development and morphogenesis programmes to lineage commitment and differentiation (Figure 3C). These transitions were often accompanied by shifts in the active components of canonical signalling pathways, both by altering the expression of key TFs and by changing the permissiveness of 35 the chromatin at their effector sites throughout the genome. For example, while expression of *Smad3* and *Bmp3* was at its highest levels in thoracic somites, increasing expression of *Smad5* alongside *Bmp1*, *Bmp4* and *Bmp7* was observed later in development (Figure 3B). *Smad5*

was expressed at all stages, albeit at lower levels early on (Figure 3J); however, its binding sites only became accessible from stage 25 onwards, when expression was highest (Figure 3K). Signalling through SMAD2/3 and SMAD1/5 have opposing effects on differentiation. For example, while BMP3-SMAD3 block osteogenesis, BMP1 and BMP7, downstream of 5 TGFbeta, promote osteoblast production (Wu, Chen, and Li 2016). Thus, the switch in usage of the opposing arms of the SMAD-BMP or SMAD-TGFbeta signalling pathways suggests cells at later developmental stages have progressed further in their differentiation trajectory. Consistent with this, genes crucial for fate determination and commitment also increased in expression across time: *Mef2c* (Figure 3L), which is fundamental in myogenic differentiation, 10 and *Sox9* (Figure S5D), which specifies chondrocytes, peaked in sacral somites (Green et al. 2015; Molkentin et al. 1995).

Additionally, we observed significant upregulation of *Bmp4* specifically in lumbar somites (cluster7, Figure 3B). Besides regulating the development of the skeletal system, BMP4 has 15 been shown to induce the expression of *Flk1 (Kdr)* in epithelial somites (Nimmagadda et al. 2005), a factor essential for vasculogenesis and angiogenesis. The dermomyotome derivative lineages include vascular endothelial cells. Concomitant with *Bmp4* upregulation, we also observed increased expression of many other genes involved in angiogenesis, such as *Gata2*, *Tal1*, *Ets1* and *Tbx20* (cluster7, Figure 3B). Later in development, sacral somites continued to 20 express high levels of angiogenic factors and further activated more mature programmes involved in blood vessel remodelling (Figure 3C).

Finally, to assess whether our molecular atlas captures regulators important in determining 25 vertebral fate identity, we tested for enrichment of genes with specific mouse knock-out phenotypes. Genomic loci with highest activity in cervical or thoracic somites were strongly enriched for phenotypes affecting cervical vertebrae and thoracic and rib morphology. In contrast, regions active in sacral somites were associated with abnormal lumbar and sacral vertebrae (Figure 3M). Thus, our catalogue of differential activity along the axial skeleton can be utilised to identify genes and regulatory elements important for the specification of the 30 different vertebral structures.

Skeletogenesis is shaped by opposing regulatory modules

Next, we leveraged the paired design of our dataset to map chromatin-transcription regulatory 35 interactions driving somite maturation and differentiation, by applying the Functional Inference of Gene Regulation (FigR) method (Kartha et al. 2022) (with minor modifications, see Methods for details). First, we computed the correlation between the activity levels of all differentially expressed genes and the open chromatin peaks within 100 kb to identify regulatory elements

likely to direct nearby gene expression changes. Peak-gene pairs were considered significantly associated if they had stronger correlation values compared to randomised interactions. After restricting results to pairs with moderate to high correlation scores (>0.3), we identified 12,803 putative regulatory interactions, involving 47% of all differentially expressed genes.

5 Although a small fraction of the interactions included promoter-like peaks in the immediate vicinity of the genes, most resembled enhancers and were dozens of kilobases away, with a median distance of 38 kb (interquartile range: 13.1-67.5 kb). Linked peaks overlapped more often with FANTOM5 and ENCODE enhancers (H3K27ac-high H3K4me3-low signature) compared to all peaks, lending support for their regulatory activity (Figure 4A).

10 This proportion sharply increased when links were restricted to those with the strongest correlations (Figure 4A). Thus, this strategy serves to enrich the set of chromatin loci for enhancer elements, and to associate their activity to dynamically regulated genes.

Most (93.7%) differentially expressed genes significantly linked with chromatin changes were

15 associated with one to five putative enhancer regions but a few hundred genes were linked to many more enhancers (Figure 4B). The set of 349 strongly-connected genes contained factors key in controlling somite development and differentiation, suggesting that these processes are under complex regulatory control. Some of the most highly connected genes were Hox factors from late paralogous groups (Figure 4B), consistent with chromatin remodelling playing a 20 crucial role in controlling their timely expression (Soshnikova and Duboule 2009). Next, we scanned the peaks associated with these highly-regulated genes and identified enriched TF motifs. TF-gene pairs were assigned a *regulation score* that favours TFs showing correlated expression to the accessibility dynamics of linked peaks (Figure 4C).

25 We identified opposing transcriptional programmes active in early and late development by clustering transcription factors with large regulation scores (Figure 4D-E). Four different modules of regulatory activity were evident. Module 1 acted on genes that show differences in expression between the somite trios, including genes involved in the establishment of anterior-posterior patterning and somitogenesis, such as *Cdx1/2*, *Gbx2*, *Lef1*, and Hox genes from 30 early paralogous groups (Figure 4D,E). CDX1 and GBX2 themselves, together with SALL4, showed some of the strongest regulation scores on these genes. Consistent with their role in the specification and patterning of somitic mesoderm, their expression was highest in the most immature somite I (Figure 4F).

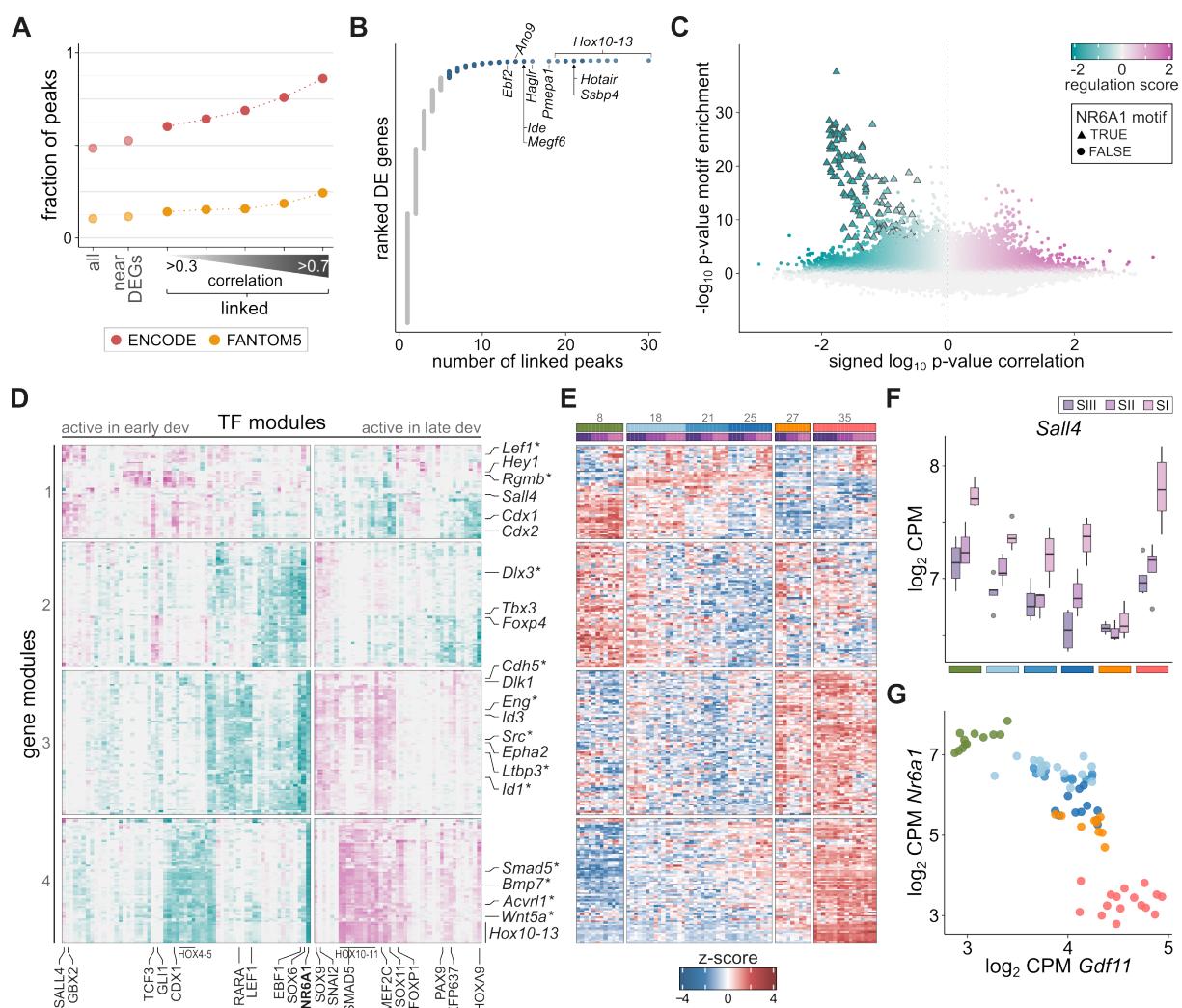


Figure 4 | Regulatory modules with opposing activity along embryonic development ensure timely activation of skeletogenesis pathways. **A)** Fraction of peaks that overlap enhancer elements from the ENCODE and FANTOM5 catalogues. Peaks identified as putative regulators of differentially expressed genes (linked peaks) are more likely to be annotated enhancers. This fraction increases as the set of peaks is restricted to stronger interactions, as shown by limiting to linked peaks with correlation scores higher than 0.3-0.7. **B)** For each differentially expressed (DE) gene, the number of significantly associated peaks within 100kb. Several hundred genes are linked to a large number of peaks and these include many late Hox genes. **C)** Regulation scores predicted by FigR between transcription factors (TFs) and genes with many linked peaks (blue set from B). The x-axis indicates the strength of the correlation between TF expression and peak accessibility; the y-axis corresponds to the significance of the enrichment of the TF binding sites in the linked peaks. Interactions involving NR6A1 are highlighted with triangles. **D)** Heatmap depicting patterns of regulatory activity between TFs (columns) and genes (rows). Genes are split into four modules by hierarchical clustering. Genes from the TGFbeta and BMP signalling pathways are highlighted with asterisks. Colour scale is the same as in C. **E)** Heatmap showing the expression levels of the same genes as in D across all somites profiled in this study. Samples (columns) are ordered based on their observed somite stage (indicated at the top). **F)** Expression levels of *Sall4*, one of the TFs with large regulation scores on module 1 genes. **G)** Expression levels of *Nr6a1* and *Gdf11* in all somites show a strong antagonistic relationship.

The other three modules were instead related to genes that are downregulated (module 2) or upregulated (modules 3 and 4) with developmental progression (Figure 4E). Module 2 activity was influenced by Shh signalling (GLI1; Figure 4D), while genes expressed late in 5 development were under the control of several transcription factors. Among these, NR6A1 showed a prominent role, particularly in module 4, with its binding sites highly enriched in the peaks linked to these genes (Figure 4C). Expression of *Nr6a1* was negatively correlated with peak accessibility, indicating a repressive regulatory effect. A number of other transcription factors, including late Hox TFs, showed large positive regulation scores on the same genes 10 (Figure 4D), suggesting antagonistic regulatory activities to NR6A1. Among this set of activating TFs were factors key in the specification of the muscle and cartilage lineages (MEF2C and SOX9), as well as proteins involved in balancing proliferation and differentiation of the progenitor cells, with many required to avoid premature differentiation (SNAI2, ZFP637, HOXA9 and FOXP1; Figure 4D).

15

Recently, NR6A1 was shown to be a key regulator of the trunk-to-tail transition in the tailbud, where its expression early in development prevents premature activation of late-expressing genes, including late Hox genes. *Nr6a1* expression is then terminated by *Gdf11* to allow the trunk-to-tail transition (Chang et al. 2022). Although these regulatory interactions were 20 dissected in undifferentiated tailbud mesoderm cells, we observed the antagonistic expression between *Nr6a1* and *Gdf11* is retained in segmented somites (Figure 4G), suggesting this regulatory programme remains at play as cells commit to the somitic lineage. Consistently, FigR predicted the strongest effects exerted by NR6A1 to affect all Hox genes in paralogous 25 groups 10 to 13 (Figure 4D). Additional predicted regulatory interactions included several components of the TGFbeta and BMP signalling pathways from modules 3 and 4 (highlighted with asterisks in Figure 4D), with the peaks associated with these genes also showing significant enrichment for NR6A1 binding motifs. Among these genes was *Smad5*, with a regulation score only slightly lower than those observed for late Hox genes. Further, SMAD5 30 itself was identified as a positive regulator of module 4 genes (Figure 4D). Thus, these data suggest that the regulatory network controlled by NR6A1 not only regulates the trunk-to-tail transition in paraxial mesoderm, but it may participate in the timely activation of differentiation pathways required for skeletogenesis. As *Nr6a1* expression diminishes in later development, so does its repression of TGFbeta and BMP signalling. In turn, other transcription factors 35 increase in activity to enhance these pathways and drive the commitment and differentiation of cells down the various somitic derivative lineages.

DISCUSSION

The establishment of the vertebrate body plan through somitogenesis is deeply conserved. Although the molecular mechanisms driving the segmentation process are shared, alterations to the number and class of segments between different species allows facile generation of 5 vastly different body structures among vertebrates (Gomez et al. 2008). Previous studies have characterised the transcriptional changes accompanying the transition from unsegmented mesoderm to nascent and differentiated somites in the chick (G. F. Mok et al. 2021), mouse (Chal et al. 2015) and human embryos (Xi et al. 2017), at a single developmental stage. These studies have provided insights into the molecular pathways controlling segmentation and the 10 subsequent differentiation of somitic mesoderm derivatives.

Here, we have experimentally analysed how the three most recently segmented somites of mouse embryos remodel their transcriptional and chromatin landscapes at six different developmental stages, capturing the earliest molecular mechanisms that give rise to cervical, 15 thoracic, lumbar and sacral structures. We identified three thousand genes transcriptionally remodelled during somite maturation. The expression of these genes after segmentation follows the same pattern at independent stages, indicating that somites from different axial levels adhere to a conserved differentiation trajectory. However, we also identified genes expressed in specific developmental stages of the embryo, reflecting changes in the 20 microenvironment in which somites develop.

Although the pace with which new somites are formed is roughly constant across development (Palmeirim et al. 1997), our data reveals that somite differentiation accelerates as embryos grow. We observed that somites from stage 8 embryos maintain a naive transcriptional profile 25 for the entirety of the 6 hours captured in our data, but progression along development results in a shortening of the time spent in such an undifferentiated state. At later stages, somitic gene regulation is dominated by transcription factors controlling cell fate commitment. By the time embryos have formed 35 pairs of somites, differentiation programmes of derivative lineages are upregulated within a few hours post-segmentation and can already be observed in somite 30 III. This is consistent with previous work in chick embryos (Borman and Yorde 1994; Berti et al. 2015; Gi Fay Mok, Mohammed, and Sweetman 2015; Maschner et al. 2016) that showed pronounced differences in the onset of expression of key factors for the commitment of cells to the myogenic lineage, depending on embryonic age. Our data characterises the regulatory mechanisms controlling cell fate specification and commitment well before the onset of 35 definitive lineage markers, and revealed that the acceleration of somite differentiation in later development initiates soon after somite segmentation. This phenomenon is not restricted to the myogenic lineage, but extends to other cell type populations, including chondrogenic and

endothelial cells. In sum, our high-resolution view of the molecular mechanisms underlying the specification and development of somitic lineages has revealed novel features of somitogenesis and is a powerful resource for the developmental biology community to study its progression in mammals.

METHODS

Mouse embryo collection and dissection

All experiments followed the Animals (Scientific Procedures) Act 1986 (United Kingdom) and
5 with the approval of the Cancer Research UK Cambridge Institute Animal Welfare and Ethical
Review Body (form number: NRWF-DO-01- v3). Animal experiments conformed to the Animal
Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines developed by the National
Centre for the Replacement, Refinement and Reduction of Animals in Research
(NC3Rs). C57BL/6J strain mice were obtained from Charles River Laboratories and
10 maintained under standard husbandry practices.

Mouse embryos from the appropriate somite stages (8 to 35 somites) were dissected in
RNase-free conditions in cold PBS, on silicone plates. Utmost care was taken to accurately
count the number of somite pairs of each embryo; however, for the 35-somite stage embryos
15 this task becomes very difficult and it is possible that there is a one or two somite error range
in the number estimated. Photos of all the embryos profiled are provided in Supplementary
File 1. To dissect out the somites, embryos were treated with dispase II (1mg/mL in DMEM)
for 30-45 seconds at 37°C. The three most posterior pairs of somites were then collected using
20 tungsten needles, dissecting out every somite separately. We labelled each somite pair as
somite I, II or III from the most posterior to the most anterior, respectively. Thus, somite I
corresponds to the most recently segmented somite, while somites II and III were segmented
~2 and ~4 hours before (Dequéant and Pourquié 2008); we refer to this as the somite's age.
Each individual somite was placed in 10µL of lysis buffer (Takara) containing RNase inhibitor.
25 One somite from each pair was flash frozen in liquid nitrogen and stored at -80°C for later
processing for RNA-seq. The matching somites were directly processed to generate ATAC-
seq libraries.

Experimental design

We collected at least four different embryos from each developmental stage (Table S1).
30 Dissections were performed on ten different days with every stage represented on at least two
separate collection dates. However, samples from three different stages were collected on five
days, without overlap with the samples from the remaining three stages, resulting in a partially
confounded design.

35 RNA-seq experiments and library preparation

Reverse transcription was performed directly on frozen lysed somites and cDNA was amplified
with 8 cycles of PCR, using the SMART-seq v4 Ultra Low Input RNA Kit for Sequencing

(TaKaRa, 634891). RNA-seq libraries were generated from 100pg of amplified cDNA using the NEXTERA XT DNA Library Preparation kit (Illumina, FC-131-1096), according to the manufacturer's instructions, except only a quarter of the recommended reagents' amount was used. The resulting libraries were quantified using a Qubit instrument and their size 5 distributions were assessed with a TapeStation machine. Pooled libraries were sequenced on an Illumina HiSeq 4000 according to manufacturer's instructions to produce paired-end 150bp reads.

ATAC-seq experiments and library preparation

10 ATAC-seq experiments were performed following the protocol from Corces and colleagues (Corces et al. 2017). Briefly, individual somites were lysed and transposed with 1µL of transposome (Nextera DNA Sample Preparation kit FC-121-1030) at 37°C for 30 minutes. Samples were then purified with the Zymo Clean & Concentrator kit and eluted in 21µL of elution buffer. Transposed DNA was quantified by qPCR using 5 µl of PCR products. The 15 number of additional cycles was determined by plotting linear Rn versus cycle and corresponded to one third of the maximum fluorescence intensity. Transposed DNA was then amplified with 13 cycles of PCR. The final products were double size-selected with AMPure beads (0.55X - 1.5X) to obtain fragments between 100bp and 700bp. Libraries were quantified and the sizes were assessed with a TapeStation machine. Pooled libraries were sequenced 20 on an Illumina HiSeq 4000 according to manufacturer's instructions to produce paired-end 150bp reads. Samples were sequenced to a median depth of 70.5 million fragments.

RNA-seq data processing and quality control

RNA-seq paired-end fragments were aligned to the mouse reference genome (GRCm38) with 25 STAR 2.6.0c (Dobin et al. 2013) with options --outFilterMismatchNmax 6 --outFilterMatchNminOverLread 0.5 --outFilterScoreMinOverLread 0.5 --outSAMtype BAM SortedByCoordinate --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --alignIntronMin 20 --alignIntronMax 30 1000000 --alignMatesGapMax 1000000 --outSAMstrandField intronMotif. On average, 84% of the sequencing fragments mapped uniquely. We also set the option --quantMode GeneCounts to quantify the number of fragments overlapping annotated transcripts, using Ensembl's genome annotation version 96 (<http://apr2019.archive.ensembl.org/index.html>).

35 Samples were sequenced to a median depth of 17.4 million paired-end fragments. One sample had a library size of only 88 thousand fragments and was discarded. All other samples showed a uniform number of fragments mapped uniquely (median 85.6%, standard deviation (SD)

5.6%) and most of these were within annotated exons (median 84.9%, SD 2.8%). On average, we detected around 22 thousand expressed genes per sample (Table S2).

To validate the staging of samples we exploited the Hox code that serves as a molecular 5 indicative of developmental stage. As shown in Figure 1C-D, our embryo stages defined by the observed number of somites agreed with the expected expression levels of Hox genes. However, samples from one stage 27 embryo were more similar to the stage 35 somites, showing expression of several late Hox genes from paralogous groups 12 and 13 that are only observed in the stage 35 samples. These data suggest this particular embryo was likely of a 10 more advanced stage than 27 somites and was removed from downstream analyses (Table S1).

RNA-seq data normalisation

Downstream analyses were restricted to genes with at least 10 counts in three or more 15 samples, as implemented in the `filterByExpr` function from the `edgeR` package (Mark D. Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012); this represents 20,062 genes. To normalise for differences in sequencing depth we used the `calcNormFactors` function that implements the weighted trimmed mean of M-values method (M. D. Robinson and Oshlack 2010), and generated counts-per-million normalised expression estimates.

20 A PCA of the thousand most variable genes (determined from variance-stabilised data, computed with the `vst` function from the `DESeq2` package) revealed good separation of samples from different developmental stages (Figure S2A). However, we also observed subgrouping by the date of collection, indicating substantial batch effects (Figure S2A). Since 25 the experimental design is partially confounded with the date of collection of the samples, we were unable to include this as a covariate in downstream analyses. Instead, to control for technical variation unrelated to the biological variables of interest, we used the function `ImFit` from the `limma` package (Ritchie et al. 2015) to fit a linear model of the combination of developmental stage and somite age for each sample. We then performed PCA on the 30 residuals from the fit (function `residuals`) to capture systematic variation unrelated to the biological design of interest. To determine how many principal components (PCs) captured significant variation we used the `parallelPCA` function from the `scran` package (Lun, McCarthy, and Marioni 2016) on the normalised counts; this function estimates, via permutation analysis, the number of PCs that explain more variation than expected by chance, which in our case 35 was 14. Thus, the 14 first PCs were used as covariates in downstream analyses to control for unwanted variation (Figure S2B). We note that this procedure captures both technical and biological variation not modelled in our design of interest (i.e. the sex of the embryos).

RNA-seq differential expression analysis

To identify genes significantly differentially expressed across conditions we used edgeR (Mark D. Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012), with a design

5 matrix of the interaction of each sample's age and developmental stage, plus the 14 PCs representing technical variation as covariates. Dispersion was estimated with the `estimateDisp` function (setting `robust = TRUE`) and fitting the model with `glmQLFit`. Specific contrasts were then tested with the `glmQLFTest` function.

10 To identify the regions that change as somites differentiate we compared samples from somites I, II and III. To identify conserved differences across development we defined contrasts for all three pairwise comparisons, using the average of same-age samples from all six stages:

$$\begin{aligned} \text{somite}_{i.\text{vs.}j} &= \sum_{k \in \{8, 18, 21, 25, 27, 35\}} \text{stage}_k.\text{somite}_i / 6 \\ &\quad - \sum_{k \in \{8, 18, 21, 25, 27, 35\}} \text{stage}_k.\text{somite}_j / 6 \end{aligned}$$

15 where $i.\text{vs.}j$ corresponds to $I.\text{vs.}II$, $I.\text{vs.}III$ and $II.\text{vs.}III$. To recover possible stage-specific changes, we also defined contrasts on a per-stage basis:

$$\text{somite}_{i.\text{vs.}j}.\text{stage}_k = \text{somite}_i.\text{stage}_k - \text{somite}_j.\text{stage}_k$$

where k is one of the six stages and $i.\text{vs.}j$ the same as above. All three pairwise comparisons from each stage were tested at once. Thus, in these cases the p-value indicates whether the 20 gene is differentially expressed between at least a pair of somite ages.

We used a similar approach to test for differences in expression across development. Conserved differences between all somites irrespective of their maturity level were assessed by averaging somites I, II and III and testing each pairwise comparison between the six stages:

$$25 \quad \text{stage}_{k.\text{vs.}l} = \sum_{i \in \{I, II, III\}} \text{somite}_i.\text{stage}_k / 3 - \sum_{i \in \{I, II, III\}} \text{somite}_i.\text{stage}_l / 3$$

where $k.\text{vs.}l$ corresponds to all pairwise comparisons between the six stages. All contrasts were tested at once to avoid performing too many tests and, again, p-values indicate whether the gene is significantly different between at least a pair of stages. To check for any changes specific to a given somite age we repeated the analysis separately for somites I, II and III:

$$30 \quad \text{stage}_{k.\text{vs.}l}.\text{somite}_i = \text{stage}_k.\text{somite}_i - \text{stage}_l.\text{somite}_i$$

where i is any of the three somite ages and $k.\text{vs.}l$ the same as above. Genes were considered significantly differentially expressed if their adjusted p-value was lower than 0.05 (FDR < 5%)

and their absolute fold-change was greater than 1.5. Results from all differential expression analyses are available at <https://github.com/xibarrasoria/somitogenesis2022>.

ATAC-seq data alignment

5 Raw sequencing reads were aligned to the mouse reference genome (GRCm38) using bwa mem 0.7.12-r1039 (Li and Durbin 2009) with default parameters. On average, 93% of the total fragments were successfully aligned. The resulting SAM files were processed with samtools 1.5 (Li et al. 2009). One sample was sequenced to a disproportionately high depth compared to the rest (1.5 billion fragments compared to a median of 70.5 million). The mapped data for
10 this sample was downsampled to 15% of the total reads (samtools -s 0.15), and the resulting BAM file was used in the downstream processing steps.

Duplicated fragments were marked and removed using MarkDuplicates 1.103 from Picard tools (<http://broadinstitute.github.io/picard>) with option REMOVE_DUPLICATES=TRUE. We
15 further used samtools to remove any pairs that were not properly aligned (-f 0x02); supplementary alignments (-F 0x800); alignments with mapping quality lower than 30 (-q 30); and alignments outside the autosomes or chromosome X. The resulting BAM files represent the clean, good quality alignments used in all downstream analyses.

20 ATAC-seq quality control

To assess the quality of the libraries we used three different criteria: 1) the insert size distribution of the sequenced fragments; 2) the level of signal enrichment at the transcription start site (TSS) of expressed genes; and 3) the signal-to-noise ratio, assessed by the ability to call peaks (Figure S3 and Table S3).

25 To compute the insert size distribution of each library we used the getPESizes function from the csaw package (Lun and Smyth 2016). Each library's distribution was visually inspected and scored based on the number of nucleosomal peaks. Thus, a score of 0 implies that only short fragments were recovered, a score of 1 indicates presence of mononucleosomes, 2
30 corresponds to samples with both monomers and dimers, and so on. The maximum score assigned was 4, including samples with fragment sizes corresponding to nucleosome tetramers or larger (Figure S3A).

35 To estimate the enrichment of fragments at transcription start sites we applied the method recommended by the ENCODE standards for ATAC-seq data (<https://www.encodeproject.org/data-standards/terms/#enrichment>). Specifically, we restricted the analysis to genes with moderate to high expression as assessed from the RNA-seq data

(mean normalised counts per million greater than 10, corresponding to 8,025 genes). We then used the biomaRt package (Durinck et al. 2005) to extract the most 5' TSS for each gene and created a BED file of 2kb intervals centred at each TSS. We computed the coverage of such intervals using bedtools coverage 2.26.0 (Quinlan and Hall 2010) a BEDPE file containing the

5 Tn5 insertion sites inferred from the aligned fragments (by shifting the start/end coordinates by +5/-4 bp with an ad hoc perl script). To calculate the enrichment at the TSS we first computed the mean insertion counts at each base pair from all genes. We then used the mean of the first and last 100bp as an estimate of the background insertion rate. For each base pair, we computed the enrichment score as the fold-change against the background rate; this results
10 in an enrichment score of ~1 at the flanks of the 2kb interval which increases as it approaches the TSS (Figure S3B).

Finally, to assess the signal-to-noise ratio of each sample we used MACS2 2.1 (Zhang et al. 2008) to call peaks, with options callpeak -f BAMPE -g mm --keep-dup all --broad. Peaks
15 overlapping blacklisted regions (Amemiya, Kundaje, and Boyle 2019) (obtained from <https://github.com/Boyle-Lab/Blacklist/blob/master/lists/mm10-blacklist.v2.bed.gz>) were discarded. We calculated the fraction of reads in peaks (FRIP) as the total fragments overlapping called peaks over the total library size and used this, along with the total number
20 of peaks, as proxies for the signal-to-noise ratio (Figure S3C).

20 Libraries with an insert size distribution showing a good nucleosomal pattern generally had good TSS enrichment scores and signal-to-noise ratios. For each sample we defined a quality control pass if they had an insert size distribution score of 2 or higher; a fraction of reads in peaks of 3% or larger; at least 15,000 peaks; and a TSS enrichment score of 5 or higher
25 (Figure S3D). Samples satisfying at least three of these criteria were annotated as good quality and used in downstream analyses (50 of the 75 libraries). Importantly, insert size distribution scores were positively correlated with the experimentally measured DNA fragment sizes but showed no relation to sequencing depth, indicating that samples with poor quality control characteristics are not due to insufficient sequencing (Figure S3E).

30 **ATAC-seq peak calling**
To define a unified set of peaks for the whole dataset we combined the clean BAM files from the 50 samples that passed quality control and used them as input for MACS2 (same parameters as stated above). By calling peaks on the combined data from all samples, the
35 peak calling process becomes agnostic to the different conditions in our experimental design, which is important for downstream differential accessibility analyses. After removing peaks overlapping blacklisted regions, a total of 131,743 peaks were called, with a median width of

777 bp (interquartile range 418-1394 bp). We re-computed the FRiP for each sample using this common peak set.

It is possible that by merging all samples together, some low-enrichment stage-specific peaks
5 are lost. Thus, we repeated the peak calling procedure but on a stage-specific basis. When comparing the per-stage peak calls to the set obtained by using all 50 samples, around 96.6% of the peaks called in each stage were also called in the all-sample set (range 93.43-98.09%). The small proportion of peaks missed generally had small fold-changes and high q-values, and thus correspond to low significance calls. This provides confidence that we have not
10 missed stage-specific peaks by merging data from all samples.

ATAC-seq data normalisation

To normalise the ATAC-seq data we used the methods implemented in the csaw package (Lun and Smyth 2016). We generated MA plots comparing all pairs of samples by counting the
15 number of fragments in 10 kb windows tiling the genome. For high-abundance windows, which correspond to open chromatin, we observed a deviation of the \log_2 fold-change from the expected value of 0 that was correlated with the abundance level of the genomic region (Figure S4A). Conventional normalisation techniques used in the majority of ATAC-seq analyses compute a single size factor that captures systematic differences between samples; these
20 approaches fail to account for the trend observed in our data (Figure S4B). Thus, we instead used a loess-based approach to compute size factors specific to each abundance level. For this, we counted the number of fragments mapped to 150 bp windows, sliding along the genome by 50 bp, using the function windowCounts (with filter set to 75 and excluding any reads overlapping blacklisted regions). We then filtered out any windows that did not overlap
25 the common peak set or that had less than an average count of 4 fragments across samples. We finally used this set of windows to compute the size factors with the normOffsets function (with type=loess). This approach successfully removed the observed trend (Figure S4C). However, the first principal component estimated from the normalised counts of the 5000 most
30 variable windows was strongly correlated with samples' FRiP (Pearson's $r = -0.59$), suggesting other technical effects were still dominant in the data (Figure S4D).

To remove unwanted variation from the dataset we used the same strategy that we applied to the RNA-seq samples. That is, we obtained the residuals from a linear model fit of the interaction of the stage and somite age of each sample and applied PCA to capture the major
35 sources of variation. We retained the first 18 PCs, since these were deemed to explain significantly more variation than chance (as determined by the parallelPCA function), which significantly removed efficiency and batch effects (Figure S4E).

ATAC-seq differential accessibility analysis

To test for differences in accessibility across conditions we used the approach implemented in csaaw (Lun and Smyth 2016). We based the analysis on the window counts described above.

5 Window counts along with the corresponding size factors were converted into a DGEList object compatible with edgeR (Mark D. Robinson, McCarthy, and Smyth 2010) to perform differential analysis. The same approach as described for the RNA-seq data was used.

After each window was tested, we used the mergeWindows function to merge windows that

10 were no more than 150 bp apart, restricting the maximum width to 1.5 kb. Regions larger than 1.5 kb were broken into smaller overlapping regions of roughly equal size (+/- 100bp). We then computed a combined p-value for each of these regions with the combineTests function, using Simes' method. Correction for multiple testing was performed at the region level and regions were considered significantly different if their adjusted p-value was lower than 0.05 and their
15 absolute fold-change was greater than 1.5. Results from all differential expression analyses are available at <https://github.com/xibarrasoria/somitogenesis2022>.

Functional terms enrichment analysis

Gene ontology enrichment analysis was performed using the elim method from the TopGo

20 package (Alexa and Rahnenfuehrer 2022), as implemented in the topGOtable function from the PCAexplorer 2.18.0 package (Marini and Binder 2019). Enrichment of GO terms among differentially expressed genes was computed, using all genes expressed in somites as the background.

25 Enrichment analysis of GO terms and mouse KO phenotypes in differentially accessible chromatin regions was computed with GREAT (McLean et al. 2010), using the implementation from the rGREAT 1.24.0 package (Gu and Hübschmann 2022). All somite peaks were used as the background set.

30 Motif enrichment analysis

To determine transcription factor (TF) motifs enriched in the regions of open chromatin, we used Analysis of Motif Enrichment (McLeay and Bailey 2010) from the MEME suite (Bailey et al. 2009), with the human and mouse HOCOMOCOv11_full motif databases. Enrichment in

35 differentially accessible regions (either between somite ages, or between stages, split by the vertebral fate showing highest accessibility) was computed by comparing to a set of non-differentially accessible regions with a similar length distribution.

chromVAR

To estimate the accessibility dynamics of sites harbouring specific TF binding sites we used chromVAR 1.12.0 (Schep et al. 2017), on the normalised and corrected window counts described above. Following the authors' recommendations, we removed overlapping windows

5 with the filterPeaks function, and then scanned them for matches to the motif collection provided with the package (mouse_pwms_v2). Accessibility deviation scores were then computed with the computeDeviations function. For all plots, we use the z-scores returned by this function.

10 FigR

To infer peak-gene putative regulatory links we used FigR 0.1.0 (Kartha et al. 2022), restricted to the normalised and corrected counts of the 43 samples with both RNA and ATAC-seq profiles available. The function runGenePeakcorr was used to compute the correlation between each differentially expressed gene and all peaks within 100kb. This function uses

15 chromVAR to determine a set of 100 background peaks matched for accessibility and GC content levels, to determine if the observed correlation of the gene-peak of interest is significantly higher than the correlation of the gene to these unrelated background peaks.

chromVAR samples with replacement to define the background set. When there are not enough matching peaks, the number of *different* peaks in the background set can drop 20 substantially; in extreme cases this can result in a single peak repeated 100 times. Using this distribution as a null is non-informative and thus the p-values computed for such gene-peak pairs are misleading. To avoid these cases, we modified the code to check for the number of

25 different peaks included in the background and set the p-values to NA for any gene-peak pairs with background sets containing fewer than 50 different peaks. Downstream analyses were performed using gene-peak pairs with a p-value < 0.05 and a correlation score greater than 0.3. To infer regulatory interactions, we used the getDORCScores and runFigRGRN functions, on all genes with more than 5 linked peaks. TF-gene pairs with an absolute regulation score greater than 1.25 are considered putative interactions (Figure 4C-D).

30 Data availability

The raw and processed data from this study are available in the ArrayExpress repository and can be accessed through the BioStudies database (<https://www.ebi.ac.uk/biostudies/>) under accession numbers E-MTAB-12511 for the RNA-seq dataset and E-MTAB-12539 for the ATAC-seq dataset. These include the raw FASTQ files as well as raw, normalised and batch-corrected count tables. All code used for data processing and analysis are available at 35 <https://github.com/xibarrasoria/somitogenesis2022>.

Acknowledgements and funding

The authors would like to thank Aaron Lun for valuable suggestions on the analytical strategy, Pascale Gilardi-Hebenstreit for insightful comments on the manuscript and CRUK Cambridge Institute Bioinformatics and Biological Resources Unit Core facilities. This work was supported

5 by Cancer Research UK (ET, DTO: 20412, 22398), the Wellcome Trust (XIS: 108438/Z/15 to JCM; and 202878/Z/16/Z to DTO), the European Molecular Biology Laboratory (JCM), the European Research Council (DTO: 615584, 788937), and Helmholtz NCT funding (DTO: DKFZ Abteilung B270).

10 **Author contributions**

ET, XIS, JCM, DTO designed the study; ET performed experiments; AEM, GFM provided microdissection training; XIS analysed the data; XIS, JCM, DTO wrote the manuscript; XIS, JCM, DTO oversaw the work. All authors read and approved the final manuscript.

SUPPLEMENTAL INFORMATION

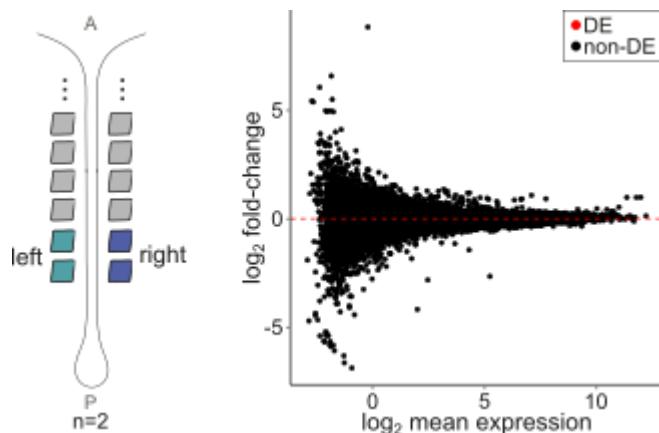


Figure S1 | The left and right somites are transcriptionally equivalent. We compared the transcriptomes of matched left and right somites by RNA-seq using the two most posterior somite pairs, as shown in the schematic. The scatter plot shows the average gene expression level on the x-axis and the corresponding fold-change between the left and right somites on the y-axis. No significantly differentially expressed (DE) genes were identified (FDR 5%), indicating that the transcriptomes of the two somites from the same pair are equivalent. A: anterior; P: posterior.

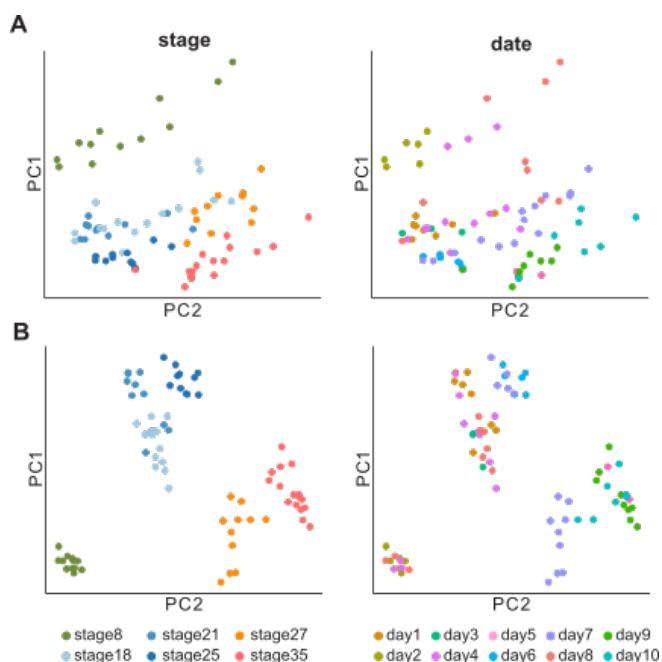


Figure S2 | Batch correction of RNA-seq data. **A)** PCA of the normalised counts of the thousand most variable genes across samples. There is clear separation by developmental stage (left). However, PC1 also separates samples based on their collection day (right). **B)** PCA after regressing out covariates capturing variation unrelated to the experimental design. Samples separate better by their developmental stage (left) and grouping by collection day is no longer evident (right).

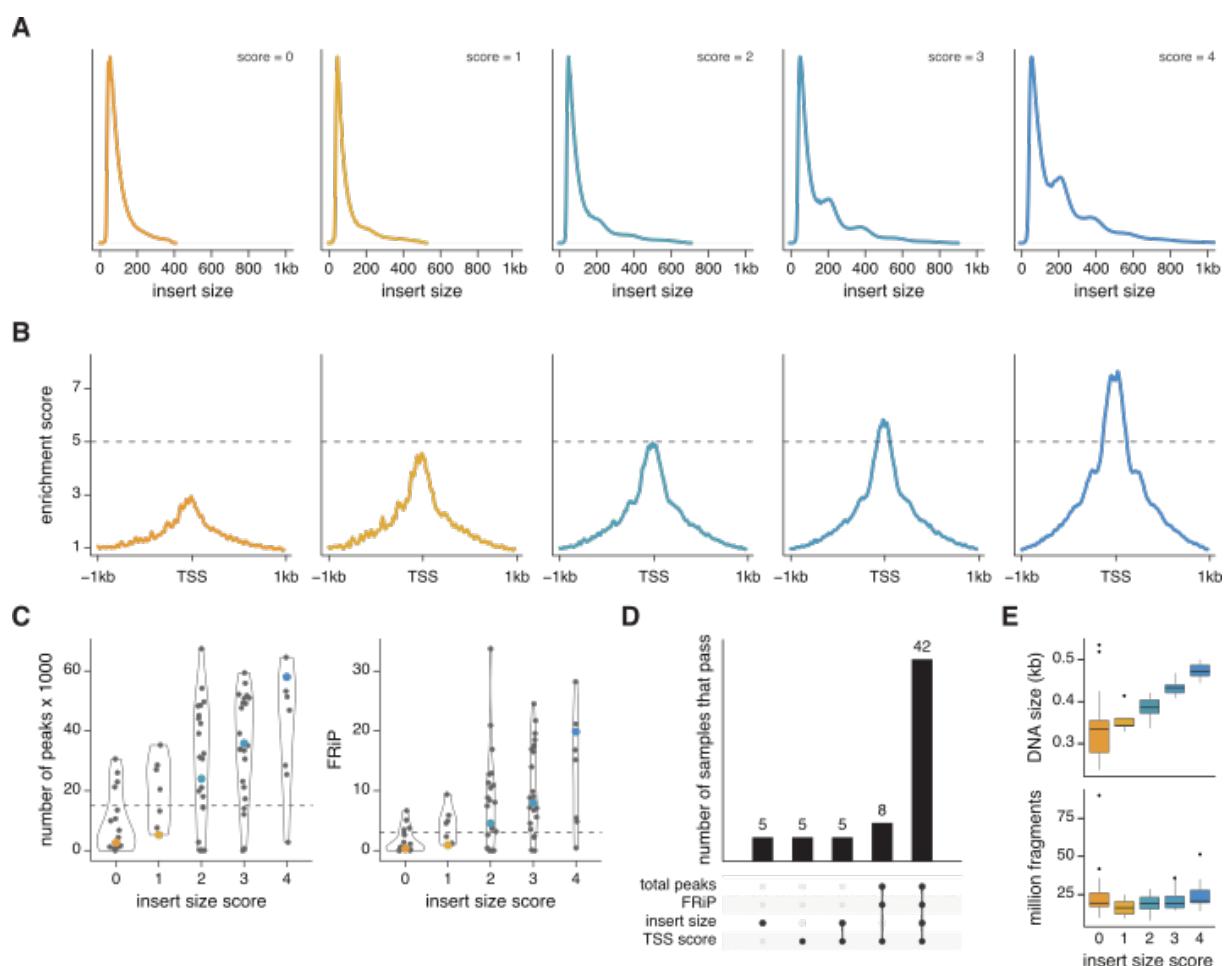


Figure S3 | Quality control of ATAC-seq libraries. Several metrics were used to determine if the ATAC-seq libraries were of good quality. **A)** Representative density plots of the fragment sizes of the sequencing data for each score group. Scores were determined based on the number of nucleosomal peaks. **B)** Representative plots of the cumulative signal for 2 kb intervals centred at the transcription start site (TSS) of expressed genes, for the same samples shown in A. An enrichment score larger than 1 indicates an excess of insertions relative to background. Signal is smoothed by taking the rolling median of 25 bp intervals. **C)** Violin plots depicting the number of total peaks called from each library (left) and the fraction of reads in peaks (FRIP; right), stratified by the sample's insert size distribution score. The samples depicted in A-B are highlighted by coloured points. **D)** Number of samples that pass each of the QC criteria. The 50 samples that passed three or four criteria were deemed of good enough quality for downstream analyses. Samples that failed all four criteria are not shown. **E)** Boxplots of the experimentally determined DNA fragment size (top) and the library size of the sequenced samples (bottom), stratified by the insert size distribution score. No relationship is observed between library size and size distribution score, indicating that the poor-quality samples are not due to insufficient sequencing.

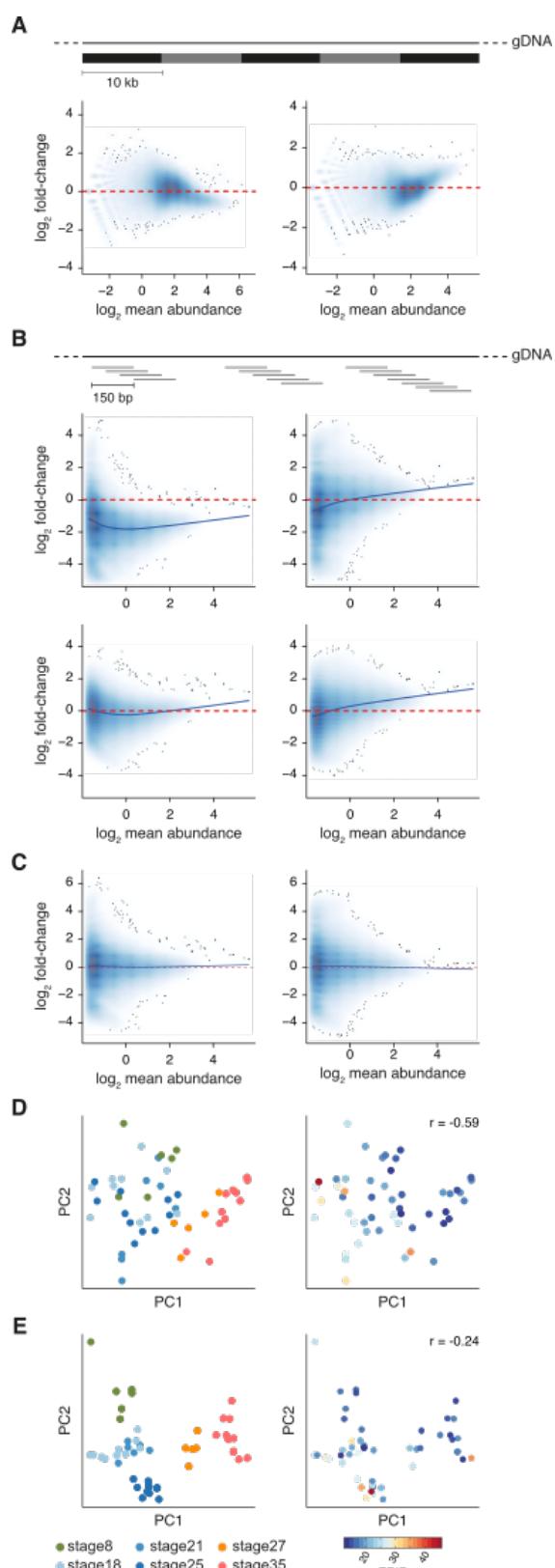


Figure S4 | Normalisation of ATAC-seq data. A) MA plots of two representative samples. On the x-axis is the \log_2 average number of sequencing fragments in 10kb bins covering the mouse genome. The y-axis corresponds to the \log_2 fold-change against a reference sample (same for both panels). Comparable samples should show \log_2 fold-changes centred around 0. High abundance bins (which contain open regions) show significant deviation from 0. This deviation shows a trend dependent on

mean abundance. **B)** To normalise the observed biases we focused on the regions of open chromatin. The MA plots now show on the x-axis the average counts in 150bp windows that slide 50bp, restricted to regions overlapping called peaks. At the top, the raw counts for the same samples in A. At the bottom, scaling normalisation is applied, which results in a shift towards a fold-change of 0. However, the observed trend dependent on mean abundance is still present in the data. **C)** MA plots for the same samples in B, but after applying loess-based normalisation, which computes a size factor for each abundance level. This successfully captured and removed the observed trend, with windows now centred around 0. **D)** PCA of the normalised counts of the 5000 most variable windows. On the left, samples are coloured by developmental stage while on the right they are coloured by their fraction of reads in peaks (FRiP). There is clear grouping of samples based on their FRiP. The Pearson correlation coefficient between FRiP and PC1 is noted. **E)** PCA after regressing out covariates capturing variation unrelated to the experimental design. Samples separate better by their developmental stage (left) and there is no longer a significant correlation with FRiP (right).

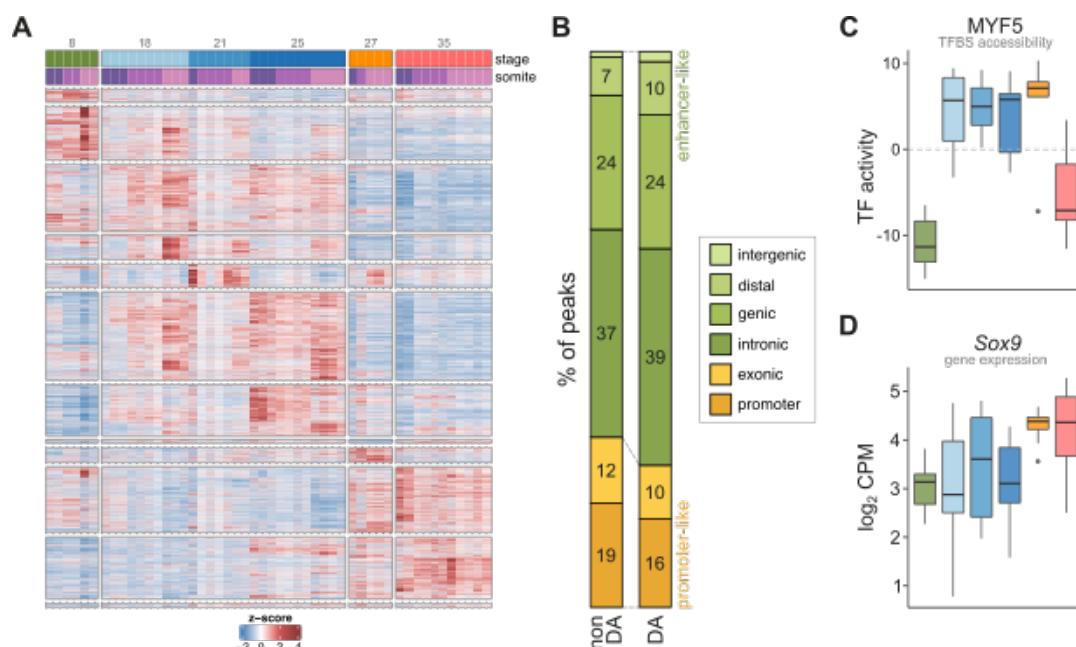


Figure S5 | Differentially accessible chromatin loci across development. **A)** Similar to Figure 3B but showing the accessibility levels of all differential peaks across development. Samples (columns) are ordered based on their observed somite number, and their stage and somite level are indicated at the top. Peaks (rows) are split into clusters by hierarchical clustering. **B)** Same as Figure 2E but for differentially accessible regions across development. **C)** Chromatin activity scores from chromVAR for the genome-wide binding sites (TFBS) of MYF5 across development. **D)** Gene expression levels for Sox9 across development.

File S1 | Photos of all the embryos profiled in this study.

Table S1 | Metadata of the samples collected and the RNA- and ATAC-seq libraries produced. The QC columns indicate whether the sample passed quality control (1) or not (0; highlighted in red); NA indicates that the sample did not yield a successful library for sequencing.

Table S2 | Quality control statistics from the RNA-seq libraries. Only one sample failed QC (in red) due to insufficient sequencing depth (totalFragments). The 'uniqueInExons' column indicates the number of fragments uniquely mapped to the genome that also can be assigned unambiguously to annotated exons. The 'numberGenes' column indicates the total number of genes with at least one count.

Table S3 | Quality control statistics from the ATAC-seq libraries. A third of the samples failed QC (in red). The 'unique' column corresponds to the number of fragments retained after removing PCR duplicates. The 'insertSizeDist', 'numberPeaks', 'readsInPeaks' and 'TSSenrichment' columns correspond to the metrics used to determine if a sample was of good quality (see Figure S3 for details).

REFERENCES

Alexa, A., and J. Rahnenfuhrer. 2022. 'TopGO: Enrichment Analysis for Gene Ontology'. *R Package Version 2.50.0*.

5 Amemiya, H. M., A. Kundaje, and A. P. Boyle. 2019. 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome'. *Sci Rep* 9 (1): 9354. <https://doi.org/10.1038/s41598-019-45839-z>.

Anderson, R. A., K. T. Schwalbach, S. R. Mui, E. E. LeClair, J. M. Topczewska, and J. Topczewski. 2020. 'Zebrafish Models of Skeletal Dysplasia Induced by Cholesterol Biosynthesis Deficiency'. *Dis Model Mech* 13 (6). <https://doi.org/10.1242/dmm.042549>.

10 Aoyama, Hirohiko. 1993. 'Developmental Plasticity of the Prospective Dermatome and the Prospective Sclerotome Region of an Avian Somite'. *Development, Growth & Differentiation* 35 (5): 507–19. <https://doi.org/10.1111/j.1440-169X.1993.00507.x>.

15 Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. 'MEME SUITE: Tools for Motif Discovery and Searching'. *Nucleic Acids Res* 37 (Web Server issue): W202–8. <https://doi.org/10.1093/nar/gkp335>.

Bénazéraf, B., and O. Pourquié. 2013. 'Formation and Segmentation of the Vertebrate Body Axis'. *Annu Rev Cell Dev Biol* 29: 1–26. <https://doi.org/10.1146/annurev-cellbio-101011-155703>.

20 Berti, Federica, Júlia Meireles Nogueira, Svenja Wöhrle, Débora Rodrigues Sobreira, Katarzyna Hawrot, and Susanne Dietrich. 2015. 'Time Course and Side-by-Side Analysis of Mesodermal, Pre-Myogenic, Myogenic and Differentiated Cell Markers in the Chicken Model for Skeletal Muscle Formation.' *Journal of Anatomy* 227 (3): 361–82. <https://doi.org/10.1111/joa.12353>.

25 Borman, W. H., and D. E. Yorde. 1994. 'Analysis of Chick Somite Myogenesis by in Situ Confocal Microscopy of Desmin Expression'. *J Histochem Cytochem* 42 (2): 265–72. <https://doi.org/10.1177/42.2.8288867>.

30 Campos, L. M., E. A. Rios, V. Midlej, G. C. Atella, S. Herculano-Houzel, M. Benchimol, C. Mermelstein, and M. L. Costa. 2015. 'Structural Analysis of Alterations in Zebrafish Muscle Differentiation Induced by Simvastatin and Their Recovery with Cholesterol'. *J Histochem Cytochem* 63 (6): 427–37. <https://doi.org/10.1369/0022155415580396>.

35 Chal, J., M. Oginuma, Z. Al Tanoury, B. Gobert, O. Sumara, A. Hick, F. Bousson, et al. 2015. 'Differentiation of Pluripotent Stem Cells to Muscle Fiber to Model Duchenne Muscular Dystrophy'. *Nat Biotechnol* 33 (9): 962–69. <https://doi.org/10.1038/nbt.3297>.

40 Chal, J., and O. Pourquié. 2017. 'Making Muscle: Skeletal Myogenesis in Vivo and in Vitro'. *Development* 144 (12): 2104–22. <https://doi.org/10.1242/dev.151035>.

Chalamalasetty, R. B., R. J. Garriock, W. C. Dunty Jr., M. W. Kennedy, P. Jailwala, H. Si, and T. P. Yamaguchi. 2014. 'Mesogenin 1 Is a Master Regulator of Paraxial Presomitic Mesoderm Differentiation'. *Development* 141 (22): 4285–97. <https://doi.org/10.1242/dev.110908>.

45 Chang, Yi-Cheng, Jan Manent, Jan Schroeder, Siew Fen Lisa Wong, Gabriel M. Hauswirth, Natalia A. Shylo, Emma L. Moore, et al. 2022. 'Nr6a1 Controls Hox Expression Dynamics and Is a Master Regulator of Vertebrate Trunk Development.' *Nature Communications* 13 (1): 7766. <https://doi.org/10.1038/s41467-022-35303-4>.

Christ, B., B. Brand-Saberi, M. Grim, and J. Wilting. 1992. 'Local Signalling in Dermomyotomal Cell Type Specification'. *Anat Embryol (Berl)* 186 (5): 505–10. <https://doi.org/10.1007/bf00185464>.

Christ, B., R. Huang, and M. Scaal. 2007. 'Amniote Somite Derivatives'. *Dev Dyn* 236 (9): 2382–96. <https://doi.org/10.1002/dvdy.21189>.

50 Christ, B., and C. P. Ordahl. 1995. 'Early Stages of Chick Somite Development'. *Anat Embryol (Berl)* 191 (5): 381–96. <https://doi.org/10.1007/bf00304424>.

Cooke, J., and E. C. Zeeman. 1976. 'A Clock and Wavefront Model for Control of the Number of Repeated Structures during Animal Morphogenesis.' *Journal of Theoretical Biology* 58 (2): 455–76. [https://doi.org/10.1016/s0022-5193\(76\)80131-2](https://doi.org/10.1016/s0022-5193(76)80131-2).

Corces, M. Ryan, Alejandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, et al. 2017. 'An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues.' *Nature Methods* 14 (10): 959–62. <https://doi.org/10.1038/nmeth.4396>.

5 Dequéant, M. L., E. Glynn, K. Gaudenz, M. Wahl, J. Chen, A. Mushegian, and O. Pourquié. 2006. 'A Complex Oscillating Network of Signaling Genes Underlies the Mouse Segmentation Clock'. *Science* 314 (5805): 1595–98. <https://doi.org/10.1126/science.1133141>.

10 Dequéant, M. L., and O. Pourquié. 2008. 'Segmental Patterning of the Vertebrate Embryonic Axis'. *Nat Rev Genet* 9 (5): 370–82. <https://doi.org/10.1038/nrg2320>.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. 'STAR: Ultrafast Universal RNA-Seq Aligner.' *Bioinformatics (Oxford, England)* 29 (1): 15–21. <https://doi.org/10.1093/bioinformatics/bts635>.

15 Draut, H., T. Liebenstein, and G. Begemann. 2019. 'New Insights into the Control of Cell Fate Choices and Differentiation by Retinoic Acid in Cranial, Axial and Caudal Structures'. *Biomolecules* 9 (12). <https://doi.org/10.3390/biom9120860>.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. 2005. 'BioMart and Bioconductor: A Powerful Link between Biological Databases and 20 Microarray Data Analysis'. *Bioinformatics* 21 (16): 3439–40. <https://doi.org/10.1093/bioinformatics/bti525>.

Gomez, C., E. M. Ozbudak, J. Wunderlich, D. Baumann, J. Lewis, and O. Pourquié. 2008. 'Control of Segment Number in Vertebrate Embryos'. *Nature* 454 (7202): 335–39. <https://doi.org/10.1038/nature07020>.

25 Green, J. D., V. Tollemar, M. Dougherty, Z. Yan, L. Yin, J. Ye, Z. Collier, et al. 2015. 'Multifaceted Signaling Regulators of Chondrogenesis: Implications in Cartilage Regeneration and Tissue Engineering'. *Genes Dis* 2 (4): 307–27. <https://doi.org/10.1016/j.gendis.2015.09.003>.

Gu, Z., and D. Hübschmann. 2022. 'RGREAT: An R/Bioconductor Package for Functional 30 Enrichment on Genomic Regions'. *Bioinformatics*, November. <https://doi.org/10.1093/bioinformatics/btac745>.

Jacob, H. J., B. Christ, and M. Jacob. 1974. '[Somitogenesis in the chick embryo. Experiments on the position development of the myotome]'. *Verh Anat Ges* 68: 581–89.

Jacob, M., B. Christ, and H. J. Jacob. 1975. '[Regional determination of the paraxial mesoderm in young chick embryos]'. *Verhandlungen der Anatomischen Gesellschaft* 69: 263–69.

Kartha, V. K., F. M. Duarte, Y. Hu, S. Ma, J. G. Chew, C. A. Lareau, A. Earl, et al. 2022. 'Functional Inference of Gene Regulation Using Single-Cell Multi-Omics'. *Cell Genom* 2 (9). <https://doi.org/10.1016/j.xgen.2022.100166>.

Kmita, M., and D. Duboule. 2003. 'Organizing Axes in Time and Space; 25 Years of Colinear Tinkering'. *Science* 301 (5631): 331–33. <https://doi.org/10.1126/science.1085753>.

40 Krol, A. J., D. Roellig, M. L. Dequéant, O. Tassy, E. Glynn, G. Hattem, A. Mushegian, A. C. Oates, and O. Pourquié. 2011. 'Evolutionary Plasticity of Segmentation Clock Networks'. *Development* 138 (13): 2783–92. <https://doi.org/10.1242/dev.063834>.

Krumlauf, R. 1994. 'Hox Genes in Vertebrate Development'. *Cell* 78 (2): 191–201. [https://doi.org/10.1016/0092-8674\(94\)90290-9](https://doi.org/10.1016/0092-8674(94)90290-9).

45 Lamouille, S., J. Xu, and R. Derynck. 2014. 'Molecular Mechanisms of Epithelial-Mesenchymal Transition'. *Nat Rev Mol Cell Biol* 15 (3): 178–96. <https://doi.org/10.1038/nrm3758>.

Li, H., and R. Durbin. 2009. 'Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform'. *Bioinformatics* 25 (14): 1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.

50 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25 (16): 2078–79. <https://doi.org/10.1093/bioinformatics/btp352>.

5 Lun, A. T., D. J. McCarthy, and J. C. Marioni. 2016. 'A Step-by-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data with Bioconductor'. *F1000Res* 5: 2122. <https://doi.org/10.12688/f1000research.9501.2>.

10 Lun, A. T., and G. K. Smyth. 2016. 'Csaw: A Bioconductor Package for Differential Binding Analysis of ChIP-Seq Data Using Sliding Windows'. *Nucleic Acids Res* 44 (5): e45. <https://doi.org/10.1093/nar/gkv1191>.

15 Mallo, M., D. M. Wellik, and J. Deschamps. 2010. 'Hox Genes and Regional Patterning of the Vertebrate Body Plan'. *Dev Biol* 344 (1): 7–15. <https://doi.org/10.1016/j.ydbio.2010.04.024>.

20 Marini, F., and H. Binder. 2019. 'PcaExplorer: An R/Bioconductor Package for Interacting with RNA-Seq Principal Components'. *BMC Bioinformatics* 20 (1): 331. <https://doi.org/10.1186/s12859-019-2879-1>.

25 Maschner, A., S. Krück, M. Draga, F. Pröls, and M. Scaal. 2016. 'Developmental Dynamics of Occipital and Cervical Somites'. *J Anat* 229 (5): 601–9. <https://doi.org/10.1111/joa.12516>.

30 McCarthy, D. J., Y. Chen, and G. K. Smyth. 2012. 'Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation'. *Nucleic Acids Res* 40 (10): 4288–97. <https://doi.org/10.1093/nar/gks042>.

35 McLean, C. Y., D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. 2010. 'GREAT Improves Functional Interpretation of Cis-Regulatory Regions'. *Nat Biotechnol* 28 (5): 495–501. <https://doi.org/10.1038/nbt.1630>.

40 McLeay, R. C., and T. L. Bailey. 2010. 'Motif Enrichment Analysis: A Unified Framework and an Evaluation on ChIP Data'. *BMC Bioinformatics* 11 (April): 165. <https://doi.org/10.1186/1471-2105-11-165>.

45 Mok, G. F., L. Folkes, S. A. Weldon, E. Maniou, V. Martinez-Heredia, A. M. Godden, R. M. Williams, et al. 2021. 'Characterising Open Chromatin in Chick Embryos Identifies Cis-Regulatory Elements Important for Paraxial Mesoderm Formation and Axis Extension'. *Nat Commun* 12 (1): 1157. <https://doi.org/10.1038/s41467-021-21426-7>.

50 Mok, Gi Fay, Rabeea Hazim Mohammed, and Dylan Sweetman. 2015. 'Expression of Myogenic Regulatory Factors in Chicken Embryos during Somite and Limb Development.' *Journal of Anatomy* 227 (3): 352–60. <https://doi.org/10.1111/joa.12340>.

55 Molkentin, J. D., B. L. Black, J. F. Martin, and E. N. Olson. 1995. 'Cooperative Activation of Muscle Gene Expression by MEF2 and Myogenic BHLH Proteins'. *Cell* 83 (7): 1125–36. [https://doi.org/10.1016/0092-8674\(95\)90139-6](https://doi.org/10.1016/0092-8674(95)90139-6).

60 Murtaugh, L. C., J. H. Chyung, and A. B. Lassar. 1999. 'Sonic Hedgehog Promotes Somitic Chondrogenesis by Altering the Cellular Response to BMP Signaling'. *Genes Dev* 13 (2): 225–37. <https://doi.org/10.1101/gad.13.2.225>.

65 Nimmagadda, S., P. Geetha Loganathan, R. Huang, M. Scaal, C. Schmidt, and B. Christ. 2005. 'BMP4 and Noggin Control Embryonic Blood Vessel Formation by Antagonistic Regulation of VEGFR-2 (Quek1) Expression'. *Dev Biol* 280 (1): 100–110. <https://doi.org/10.1016/j.ydbio.2005.01.005>.

70 Ordahl, C. P., and N. M. Le Douarin. 1992. 'Two Myogenic Lineages within the Developing Somite'. *Development* 114 (2): 339–53. <https://doi.org/10.1242/dev.114.2.339>.

75 Ozbudak, E. M., O. Tassy, and O. Pourquié. 2010. 'Spatiotemporal Compartmentalization of Key Physiological Processes during Muscle Precursor Differentiation'. *Proc Natl Acad Sci U S A* 107 (9): 4224–29. <https://doi.org/10.1073/pnas.0909375107>.

80 Palmeirim, I., D. Henrique, D. Ish-Horowicz, and O. Pourquié. 1997. 'Avian Hairy Gene Expression Identifies a Molecular Clock Linked to Vertebrate Segmentation and Somitogenesis'. *Cell* 91 (5): 639–48. [https://doi.org/10.1016/s0092-8674\(00\)80451-1](https://doi.org/10.1016/s0092-8674(00)80451-1).

85 Quinlan, A. R., and I. M. Hall. 2010. 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features'. *Bioinformatics* 26 (6): 841–42. <https://doi.org/10.1093/bioinformatics/btq033>.

90 Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. 2015. 'Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies'. *Nucleic Acids Res* 43 (7): e47. <https://doi.org/10.1093/nar/gkv007>.

Robinson, M. D., and A. Oshlack. 2010. 'A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data'. *Genome Biol* 11 (3): R25. <https://doi.org/10.1186/gb-2010-11-3-r25>.

5 Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. 'EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.' *Bioinformatics (Oxford, England)* 26 (1): 139–40. <https://doi.org/10.1093/bioinformatics/btp616>.

Saga, Y. 2012. 'The Mechanism of Somite Formation in Mice'. *Curr Opin Genet Dev* 22 (4): 10 331–38. <https://doi.org/10.1016/j.gde.2012.05.004>.

15 Sanchez, A. M., R. B. Candau, and H. Bernardi. 2014. 'FoxO Transcription Factors: Their Roles in the Maintenance of Skeletal Muscle Homeostasis'. *Cell Mol Life Sci* 71 (9): 1657–71. <https://doi.org/10.1007/s00018-013-1513-z>.

Schep, A. N., B. Wu, J. D. Buenrostro, and W. J. Greenleaf. 2017. 'ChromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data'. *Nat Methods* 14 (10): 975–78. <https://doi.org/10.1038/nmeth.4401>.

15 Soshnikova, N., and D. Duboule. 2009. 'Epigenetic Temporal Control of Mouse Hox Genes in Vivo'. *Science* 324 (5932): 1320–23. <https://doi.org/10.1126/science.1171468>.

20 Stottmann, R. W., A. Turbe-Doan, P. Tran, L. E. Kratz, J. L. Moran, R. I. Kelley, and D. R. Beier. 2011. 'Cholesterol Metabolism Is Required for Intracellular Hedgehog Signal Transduction in Vivo'. *PLoS Genet* 7 (9): e1002224. <https://doi.org/10.1371/journal.pgen.1002224>.

25 Vermot, J., and O. Pourquié. 2005. 'Retinoic Acid Coordinates Somitogenesis and Left-Right Patterning in Vertebrate Embryos'. *Nature* 435 (7039): 215–20. <https://doi.org/10.1038/nature03488>.

30 25 Vermot, Julien, Jabier Gallego Llamas, Valérie Fraulob, Karen Niederreither, Pierre Chambon, and Pascal Dollé. 2005. 'Retinoic Acid Controls the Bilateral Symmetry of Somite Formation in the Mouse Embryo'. *Science* 308 (5721): 563–66. <https://doi.org/10.1126/science.1108363>.

35 Weldon, S. A., and A. E. Münsterberg. 2022. 'Somite Development and Regionalisation of the Vertebral Axial Skeleton'. *Semin Cell Dev Biol* 127 (July): 10–16. <https://doi.org/10.1016/j.semcdb.2021.10.003>.

40 Wellik, D. M. 2007. 'Hox Patterning of the Vertebrate Axial Skeleton'. *Dev Dyn* 236 (9): 2454–63. <https://doi.org/10.1002/dvdy.21286>.

45 Wu, M., G. Chen, and Y. P. Li. 2016. 'TGF- β and BMP Signaling in Osteoblast, Skeletal Development, and Bone Formation, Homeostasis and Disease'. *Bone Res* 4: 16009. <https://doi.org/10.1038/boneres.2016.9>.

50 Xi, H., W. Fujiwara, K. Gonzalez, M. Jan, S. Liebscher, B. Van Handel, K. Schenke-Layland, and A. D. Pyle. 2017. 'In Vivo Human Somitogenesis Guides Somite Development from iPSCs'. *Cell Rep* 18 (6): 1573–85. <https://doi.org/10.1016/j.celrep.2017.01.040>.

40 Xu, J., K. Wang, Z. Zhang, D. Xue, W. Li, and Z. Pan. 2021. 'The Role of Forkhead Box Family in Bone Metabolism and Diseases'. *Front Pharmacol* 12: 772237. <https://doi.org/10.3389/fphar.2021.772237>.

45 Xu, S., and C. Tang. 2022. 'Cholesterol and Hedgehog Signaling: Mutual Regulation and Beyond'. *Front Cell Dev Biol* 10: 774291. <https://doi.org/10.3389/fcell.2022.774291>.

50 Yusuf, F., and B. Brand-Saberi. 2006. 'The Eventful Somite: Patterning, Fate Determination and Cell Division in the Somite'. *Anat Embryol (Berl)* 211 Suppl 1 (December): 21–30. <https://doi.org/10.1007/s00429-006-0119-8>.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, et al. 2008. 'Model-Based Analysis of ChIP-Seq (MACS)'. *Genome Biol* 9 (9): R137. <https://doi.org/10.1186/gb-2008-9-9-r137>.