Combinatorial binding predicts spatio-temporal *cis*-regulatory activity

Robert P. Zinzen[1]*, Charles Girardot[1]*, Julien Gagneur[1]*, Martina Braun[1] & Eileen E. M. Furlong[1]

Development requires the establishment of precise patterns of gene expression, which are primarily controlled by transcription factors binding to *cis*-regulatory modules. Although transcription factor occupancy can now be identified at genome-wide scales, decoding this regulatory landscape remains a daunting challenge. Here we used a novel approach to predict spatio-temporal *cis*-regulatory activity based only on *in vivo* transcription factor binding and enhancer activity data. We generated a high-resolution atlas of *cis*-regulatory modules describing their temporal and combinatorial occupancy during *Drosophila* mesoderm development. The binding profiles of *cis*-regulatory modules with characterized expression were used to train support vector machines to predict five spatio-temporal expression patterns. *In vivo* transgenic reporter assays demonstrate the high accuracy of these predictions and reveal an unanticipated plasticity in transcription factor binding leading to similar expression. This data-driven approach does not require previous knowledge of transcription factor sequence affinity, function or expression, making it widely applicable.

Gene expression states are established through the integration of signalling and transcriptional networks converging on enhancer elements, also known as *cis*-regulatory modules (CRMs)[1,2]. CRMs integrate the input of multiple transcription factors, leading to a specific spatio-temporal output of expression[3], and are therefore central to understanding gene regulation and metazoan development. Although there has been considerable progress in decoding individual CRM activities[4–15], it is difficult to scale these approaches to decipher entire transcriptional networks at a genomic level. Understanding global *cis*-regulatory networks requires a detailed knowledge of the location of all developmental CRMs, a comprehensive map of their combinatorial and temporal binding profiles and the ability to predict their spatio-temporal activity.

Chromatin immunoprecipitation followed by either microarray analysis (ChIP-on-chip) or sequencing enables the identification of active CRMs in an unbiased, genome-wide and systematic manner. For example, ChIP experiments with tissue-specific factors in the context of the entire embryo[16–20], or against factors in isolated tissues[21,22], have proven to be very accurate methods for identifying tissue-specific CRMs. Many of these studies focus on the occupancy of a single factor or a snapshot in time[16,19,23,24]; however, transcription factors rarely act in isolation, but rather bind to CRMs in a combinatorial and dynamic manner to regulate specific expression patterns. Thus, a global view of how the combinatorial binding of transcription factors translates into specific spatio-temporal expression patterns is still lacking.

Given the accuracy of ChIP approaches[16–22], the future challenge is to move beyond predicting the location of CRMs to rather predict their spatio-temporal activity. A number of sequence-based models have been applied to predict spatio-temporal activity of enhancers during development[14,25–27]. These methods are very accurate when tailored to individual CRMs[14,26] or even small numbers of regulatory modules[27]; however, they rely on detailed knowledge of the system, including estimates of transcription factor concentrations, their affinity for various sequence motifs, and cooperativity or competition between transcription factors[14,26,27]—data that are currently only available for a handful of CRMs. Here we present a complementary, data-driven approach to predict spatio-temporal CRM activity using only transcription factor binding and *in vivo* activity data as input to a machine-learning algorithm. We applied this approach to the transcriptional network governing the specification of the *Drosophila* mesoderm into different muscle primordia.

A high-resolution binding atlas of mesoderm development

The subdivision of the mesoderm requires the successive activation of transcription factors whose activities result in the specification of cardiac mesoderm (heart muscle), somatic muscle (analogous to vertebrate skeletal muscle) and visceral muscle (gut muscle) primordia. The basic helix–loop–helix factor Twist (Twi) acts as a 'master' regulator of mesoderm development (Fig. 1a)[28]. Twi directly regulates the expression of both Tinman (Tin)[29], which is essential for dorsal mesoderm specification[30], and Myocyte enhancing factor 2 (Mef2), which initiates muscle differentiation[31,32]. Tin in turn regulates Bagpipe (Bap) expression[13], which acts together with Biniou (Bin) to specify the visceral muscle[19,33]. Although many other transcription factors act to refine further the specification and subsequent differentiation of specific muscle types, these five key transcription factors (Fig. 1a) act as the high-level regulators of mesoderm development[34,35] and therefore serve as a good entry point to generate a global atlas of mesodermal CRMs. We previously examined the binding profiles of these transcription factors individually[16,18,19,36] but at a resolution too low to map combinatorial binding precisely or to model CRM activity. We therefore initiated this study by generating a high-resolution, genome-wide map of transcription factor occupancy for these five factors during multiple stages of mesoderm development.

ChIP-on-chip was performed on each transcription factor at consecutive time points spanning the majority of stages when each transcription factor is expressed, resulting in binding data for 15 developmental conditions (Fig. 1b). To minimize false positives owing to potential off-target effects, two independent antibodies for each transcription factor were used. The immunoprecipitated material was hybridized to genome-wide tiling arrays and enriched regions were identified as clusters of consecutive probes with significant

[1]European Molecular Biology Laboratory, D-69117 Heidelberg, Germany.
*These authors contributed equally to this work.

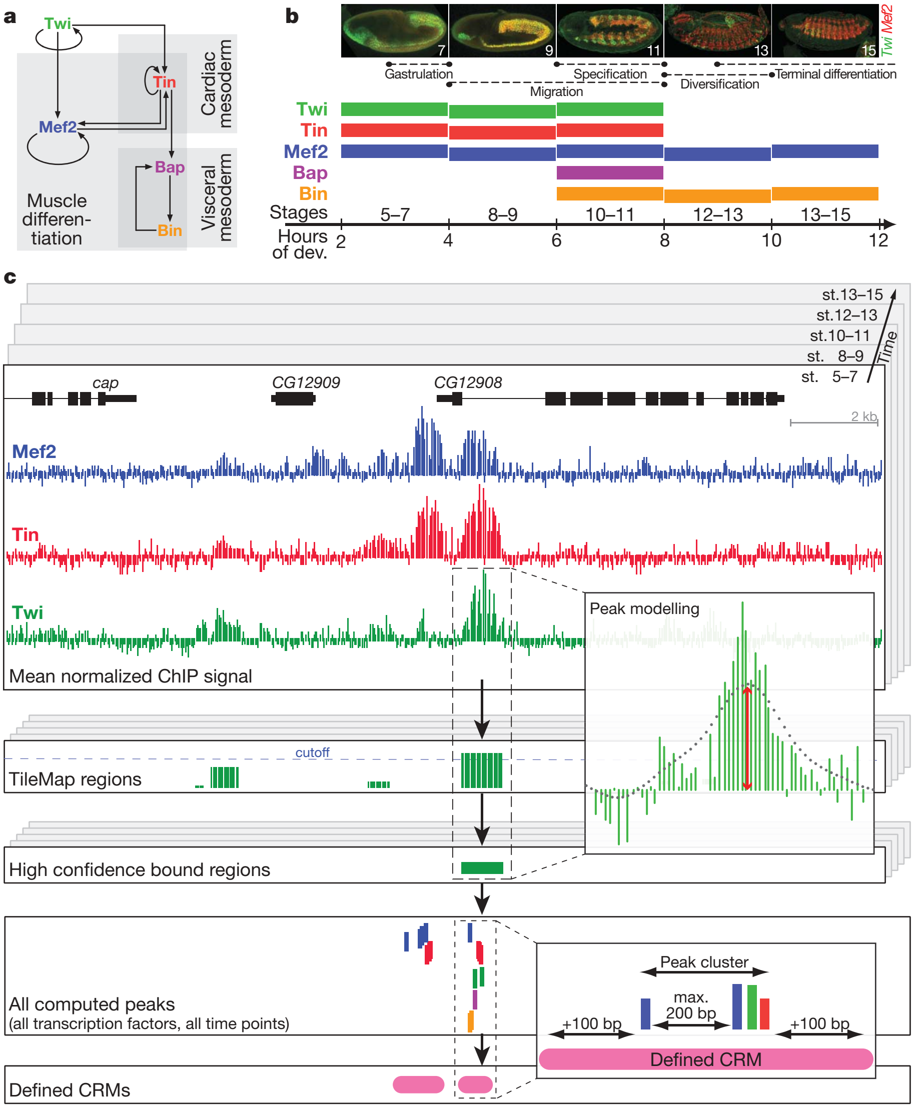

Figure 1 | Generating a high-resolution atlas of mesodermal CRMs. a, Myogenic network of five central transcription factors in mesoderm specification. **b**, Coloured rectangular boxes indicate consecutive 2-h time windows assayed by ChIP-on-chip for each transcription factor. Major events in mesoderm specification are indicated (top) during the developmental stages assayed (bottom). **c**, Data analysis workflow. Top panel shows normalized mean $\log_2$-ratios of ChIP-on-chip signal per transcription factor (horizontal tracks) and time point (stacked windows). Significantly bound regions are calculated for all transcription factors and time points using TileMap (shown for Twist at 2–4 h (stages 5–7)). Peaks were calculated as extrema (red arrow in 'peak modelling' inset) in selected regions. ChIP CRMs were computed based on peak clustering, indicated in pink.

signal (TileMap)[37]. These experiments identified thousands of high-confidence bound regions (Supplementary Table 1), providing a genome-wide map reflecting the location, temporal occupancy and combinatorial binding of mesodermal transcription factors *in vivo*.

Clustered transcription factor binding defines CRMs

Functional transcription factor binding sites (TFBSs) typically cluster together to form regulatory elements or CRMs; we used this property to search for genomic regions containing clusters of ChIP-binding peaks in close proximity. To determine an appropriate distance for proximity we first assessed the precision of our ChIP data. As TFBSs are generally located beneath ChIP enrichment peaks[20,38,39], we compared the relative location of computed ChIP peaks[40] (Fig. 1c and Methods) to previously characterized functional TFBSs, and more globally using predicted sites. Mutagenesis analyses in three CRMs regulating *Mef2* expression identified active TFBSs for Twi, Tin and Mef2[41–43] (Fig. 2a), providing a good test case to assess the ability of ChIP peaks to pinpoint functional sites. Computed ChIP peaks were, on average, within 50 bp of the functional TFBSs (Fig. 2a), demonstrating the high sensitivity and resolution of the data. More globally, the absolute distance of ChIP peaks for a given transcription factor to the closest predicted TFBS is significantly enriched within 100 bp (Supplementary Fig. 1).

We next assessed the relative distance of transcription factor binding events for different factors at the same stage of development. ChIP peaks for different transcription factors are positioned much closer to each other than expected by chance (Fig. 2b and Supplementary Fig. 2), indicating that these occupied binding sites cluster into regulatory modules. Given the ChIP precision of ±100 bp, we defined CRMs as 200-bp windows centred on ChIP peaks, and combined regions where peaks clustered into overlapping windows. The ChIP experiments from all 15 conditions identified 19,522 high-confidence ChIP peaks, which clustered into 8,008 regions; almost half of these (3,713) are multi-peak regions bound by more than one transcription factor, or by one transcription factor at more than one time point (Supplementary Fig. 3). These computed ChIP CRMs cover 2.17 Mb of genomic sequence, representing a 3.3-fold increase in resolution compared to TileMap enriched regions, and have an average length of 270 bp, which is significantly smaller than what is commonly tested in enhancer-reporter assays (Supplementary Fig. 4).

The accuracy of the ChIP-CRM atlas was assessed globally using TFBS conservation as an indicator of regulatory function and more locally using *in vivo* transgenic reporter assays. To assess conservation, we first optimized the position weight matrices (PWMs) for the five transcription factors (Supplementary Fig. 5, 5′ and Methods). The refined PWMs were used to search for all instances of these motifs within the *Drosophila* genome. As expected, these motifs are highly enriched within ChIP CRMs, being present within 100 bp of the ChIP peak in ~60–80% of CRMs (Supplementary Fig. 1, blue line). Two methods were used to globally assess TFBS conservation: first, the average PhastCons score[44] across predicted TFBSs within ChIP CRMs for a given transcription factor was compared to that of CRMs not bound by that transcription factor to minimize general sequence biases of CRMs (Fig. 2c)[20]. Second, the percentage of conserved TFBSs derived from pair-wise alignments of

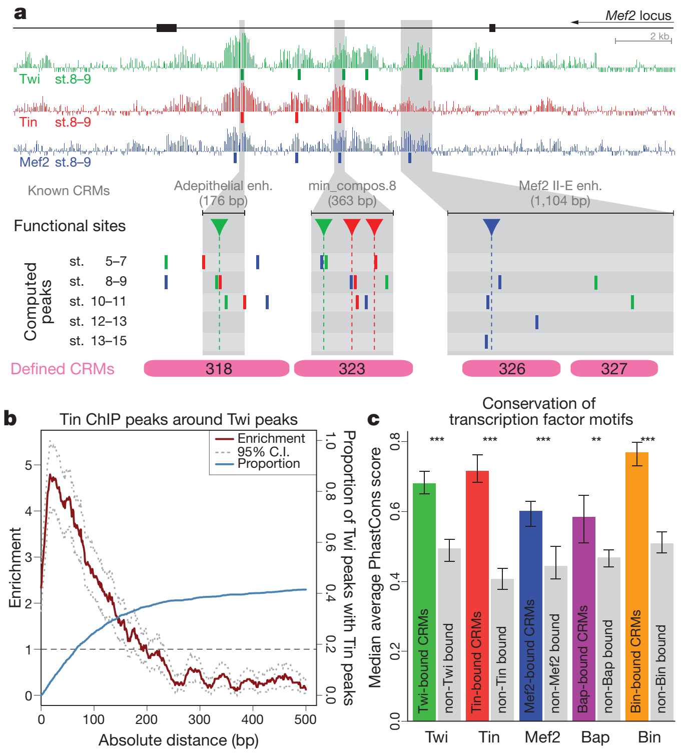

Figure 2 | ChIP peaks are within 100 bp of transcription factor motifs, which are globally conserved. a, Transcription factor binding in the *Mef2* locus. Three characterized CRMs (grey shading) with functional TFBSs (arrowheads, dotted lines): Twi, green; Tin, red; Mef2, blue. ChIP peaks (vertical bars) are within 31, 26, 101, 52 and 39 bp of verified TFBSs. ChIP CRMs are indicated in pink. **b,** Enrichment (red) and 95% confidence intervals (dotted grey) of Tin peaks at a given base-pair distance from a Twi peak. Blue line shows cumulative production of Twi peaks within Tin peaks. All at stage 8–9. **c,** Conservation of transcription factor motifs: median average PhastCons scores for motifs in bound (coloured bars) and non-bound CRMs for that transcription factor (grey bars). Error bars represent equi-tailed 95% confidence intervals of the median. $**P < 0.001$; $***P < 1.0 \times 10^{-6}$ (one-sided Wilcoxon's rank-sum test).

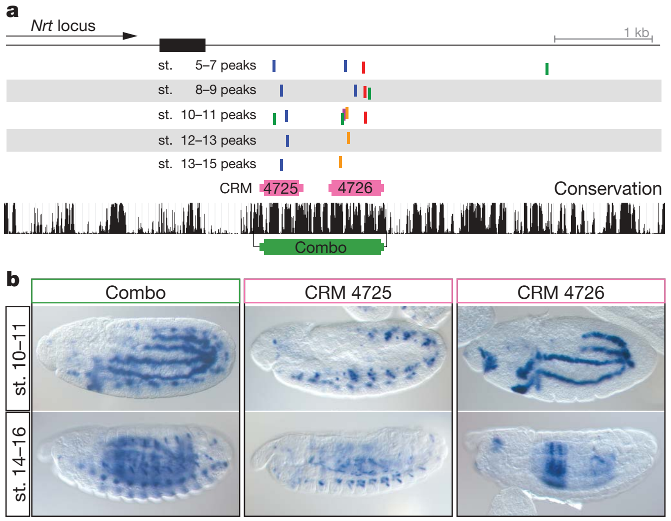

Figure 3 | ChIP CRMs act as discrete functional units. a, ChIP peaks for five transcription factors at multiple developmental stages within the first intron of *Neurotactin* (*Nrt*; top); peaks are indicated by vertical bars (colours as in Fig. 1b). Two distinct CRMs (numbers 4725 and 4726) were computed in close proximity, indicated in pink above a UCSC conservation track. **b,** *In situ* hybridization of the *lacZ* reporter driven by the 1.2-kb conserved region (green; Combo) and by the individual ChIP CRMs. The somatic muscle and visceral muscle activity (Combo) is modular and separable; CRM 4725 drives somatic muscle and 4726 visceral muscle expression. All embryos are oriented anterior-left, dorsal-up.

orthologous CRMs was compared between different *Drosophila* species (Supplementary Fig. 6 and Methods). Both analyses revealed significant TFBS conservation for all five transcription factors, suggesting that the majority of ChIP CRMs are likely to have regulatory function (Fig. 2c and Supplementary Fig. 6).

Notably, 35 out of 36 ChIP CRMs tested during this study are sufficient to function as discrete regulatory modules *in vivo*. These regions were selected based only on their transcription factor binding profiles (see below), irrespective of their evolutionary conservation or the identity of the neighbouring genes. A specific example is the *Neurotactin* locus (Fig. 3a), where several transcription factors bind within the first intron. Taking conservation as a guide, we tested a 1,200-bp region in transgenic reporter assays, which showed specific expression in a somatic muscle subset and in the visceral muscle (Fig. 3b). However, the ChIP-CRM atlas indicated two regulatory modules within this region: a 350-bp CRM containing 6 ChIP peaks (number 4725) and a 502-bp CRM containing 11 ChIP peaks (number 4726) (Fig. 3a). Testing the activities of these regions individually demonstrates that the somatic muscle and visceral muscle expression of the composite CRM are modular and separable, residing in CRM number 4725 and 4726, respectively (Fig. 3b). The high precision of the ChIP-CRM atlas can therefore define discrete regulatory units, facilitating a global analysis of spatio-temporal activity.

CAD, a CRM activity database

Our motivation for this study was to investigate whether combinatorial transcription factor binding is predictive of CRM activity during tissue differentiation. To facilitate this, we collected a reference data set of enhancers with characterized tissue-specific expression from *in vivo* transgenic reporter assays. The CRM Activity Database (CAD) (Fig. 4a) uses a controlled vocabulary to compile spatio-temporal information about the expression driven by published CRMs, using REDFly 2.0[45], literature surveys and our own experiments (Supplementary Methods). The annotation was manually reviewed on a CRM-by-CRM basis and overlapping regions were combined, split, or eliminated. This resulted in a collection of 525 largely non-redundant CRMs, 139 of which drive expression in mesoderm and/or muscle (Supplementary Tables 2 and 4).

As the majority of enhancers in CAD were identified in single gene studies, the 139 CRMs provide an independent resource to assess the sensitivity of the ChIP experiments. The ChIP-CRM atlas covers 77% of all known mesodermal-muscle CRMs, demonstrating high sensitivity and providing new insight into their temporal and combinatorial occupancy.

Using transcription factor binding to predict expression

The intersection of CAD and the ChIP-CRM atlas identified 310 ChIP CRMs for which both their combinatorial binding profile (ChIP atlas) and spatio-temporal expression (CAD) are known. We used these regions, referred to as the training set (Methods and Supplementary Table 8), to train a machine-learning algorithm to predict CRM spatio-temporal expression based on transcription factor binding profiles (Fig. 4). Five exclusive expression classes were defined as groups of CRMs with activity in specific domains, including three single-tissue classes—the early unspecified mesoderm ('Meso') and two muscle derivatives specified later in development ('somatic muscle' or 'visceral muscle')—and two complex classes combining two expression domains—'Meso & somatic muscle' or 'visceral muscle & somatic muscle' (Fig. 4b, Methods and Supplementary Table 3). For each expression class, a support vector machine (SVM) was trained to discriminate between members and non-members of the class given only transcription factor binding data. ChIP-peak heights were used as a quantitative estimate of transcription factor occupancy (Methods). Although differences in peak height may reflect several properties, it provides a biochemical quantification of the enrichment of

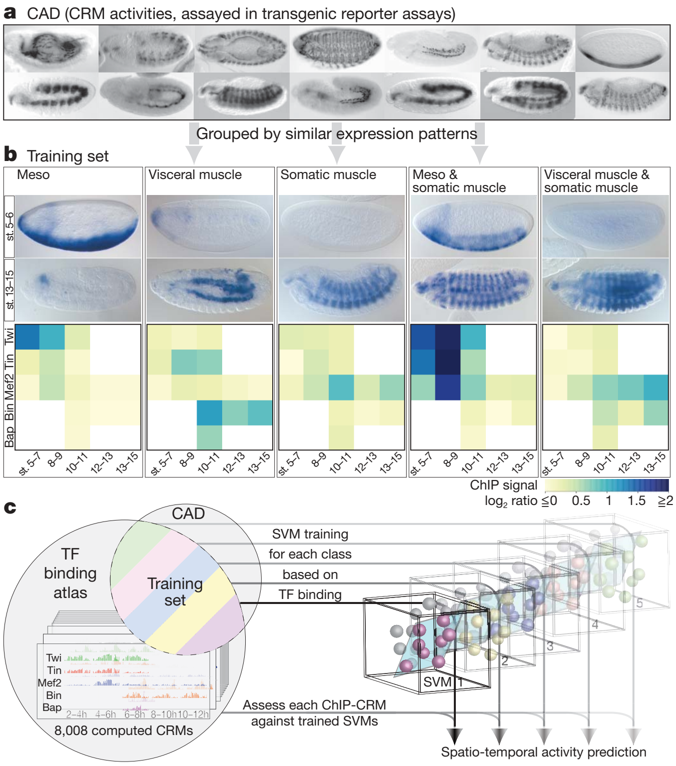
a CAD (CRM activities, assayed in transgenic reporter assays)
Grouped by similar expression patterns
b Training set
Meso
Visceral muscle
Somatic muscle
Meso & somatic muscle
Visceral muscle & somatic muscle
st. 5–6
st. 13–15
Twi
Tin
Mef2
Bin
Bap
st. 5–7
8–9
10–11
12–13
13–15
ChIP signal
log2 ratio ≤0 0.5 1 1.5 ≥2
c
CAD
TF binding atlas
Training set
SVM training for each class based on TF binding
Twi
Tin
Mef2
Bin
Bap
2–4h 4–6h 6–8h 8–10h 10–12h
8,008 computed CRMs
Assess each ChIP-CRM against trained SVMs
SVM 1
2
3
4
5
Spatio-temporal activity prediction

Figure 4 | Predicting CRM activity using a machine-learning approach. **a**, CAD contains a largely non-redundant collection of CRMs expression patterns. **b**, CRMs driving expression in the same tissue were grouped into five exclusive classes (Meso, visceral muscle, somatic muscle, Meso & somatic muscle, and visceral muscle & somatic muscle), which were used to train five SVM classifiers. The average transcription factor binding profile is shown as a heat map using peak height as a measure of transcription factor occupancy. **c**, The binding profile of each CRM within the training set was used to train an SVM, which was then run on all ChIP CRMs to predict their activity.

transcription factor binding and yields more accurate predictions than binary transcription factor binding information (Supplementary Fig. 7 and Methods). No other information was supplied to the SVM. The ability to accurately predict expression was assessed by a leave-one-out cross-validation procedure (ROC curves, Methods and Supplementary Fig. 8). SVMs with tuned parameters were trained on the complete training set and applied to all 8,008 ChIP CRMs for expression prediction (Fig. 4c, Methods and Supplementary Table 9).

Combinatorial binding predicts spatio-temporal activity

The spatio-temporal activity of at least six CRMs per expression class, selected with an SVM specificity score >95%, were tested *in vivo* using transgenic reporter assays. The ΦC31 integrase system[46] was used to integrate stably all constructs into a common genomic locus, eliminating positional effects. The expression pattern of the *lacZ* reporter driven by the CRMs was assessed by *in situ* hybridization (ISH) and CRM activity annotations were verified by fluorescent multiplex ISH using tissue-specific markers and by an independent expert (Fig. 5, Supplementary Figs 9a, b–13a, b, Supplementary Table 12a, b and Supplementary Methods).

CRMs were scored as 'correct' if they drove expression in the predicted tissue(s) (and in no other mesodermal tissue) and were therefore co-expressed with the appropriate tissue-specific markers (Supplementary Fig. 9b-13b); 'partial' if active in one of the predicted tissues, but other aspects of the mesodermal prediction did not hold true; or 'fail' if the CRM did not drive expression in any predicted tissue. Expression in non-mesodermal tissues was disregarded, as the SVM was not trained to discriminate against non-mesodermal expression. Using this stringent scoring system, the SVM predictions performed remarkably well (Supplementary Fig. 8): of the 35 individual CRMs tested, the spatio-temporal predictions of 71.4% (25) were correct, whereas 14.3% (5) worked partially and 14.3% (5) failed.

For example, 6 of the 7 CRMs predicted to drive expression in 'Meso' direct expression in the early unspecified mesoderm and not in its derived muscle tissues, even though many transcription factors within the ChIP-CRM atlas are expressed there (Fig. 5a and Supplementary Fig. 9a, b). Similarly, 7 out of 9 'visceral muscle' predictions tested regulate specific expression in visceral muscle and in no other mesodermal tissue (Fig. 5b and Supplementary Fig. 10a, b). Interestingly, a number of these modules drive expression in distinct visceral muscle subsets (Supplementary Fig. 10a, b), indicating input from additional factors to refine their expression.

Importantly, the SVM could also make accurate predictions of more complex spatio-temporal CRM expression involving tissue combinations. In the 'Meso & somatic muscle' class, 5 out of 6 CRMs tested direct expression as predicted in the unspecified mesoderm and somatic muscle, whereas the sixth CRM was partially correct, driving expression early in mesoderm, but not in somatic muscle (Fig. 5c and Supplementary Fig. 12a, b). Similarly, for the tissue combination class 'visceral muscle & somatic muscle', 5 out of 6 CRMs tested direct expression as predicted, whereas the sixth CRM was partially correct, driving expression in visceral muscle, but not in somatic muscle (Fig. 5d and Supplementary Fig. 13a, b).

The predictors of the somatic muscle class were less efficient at recovering training set members in leave-one-out cross validations (ROC plots, Supplementary Fig. 8). This was also reflected in the experimental validation (Supplementary Fig. 11a, b): whereas expression predictions for two out of seven CRMs were correct, predictions for another three CRMs were partially correct, showing expression in somatic muscle and the early mesoderm (CRMs 3775 and 6051) or visceral muscle (CRM 3775 and 6419). The poorer performance of this class probably reflects the inherent complexity of this tissue and the lack of known high-level regulators specific for somatic muscle development[35]. To investigate this further, we compared the binding signature of CRMs within the somatic muscle training set to that of the top SVM predictions (Supplementary Figs 14 and 15). This revealed enrichment in Mef2 binding and depletion in all other transcription factors as the most prominent binding signature for somatic-muscle-associated CRMs. As Mef2 is required for the differentiation of all muscle types[31] it serves as a weak predictor of this class. Nevertheless, even in the absence of binding data for a somatic muscle master regulator(s), the SVM can still recognize some 'somatic muscle signature' as only two CRMs predictions were complete 'fails' (Supplementary Fig. 11a), and the predictions for the 'visceral muscle & somatic muscle' and 'Meso & somatic muscle' classes were very accurate.

Intuitively, CRMs predicted to drive highly similar expression patterns were expected to have very similar transcription factor binding profiles. However, many CRMs with high-ranking predictions in the same tissue have an unexpected diversity in transcription factor occupancy (Fig. 5, Supplementary Fig. 14 and Supplementary Table 9). To demonstrate this, the combinatorial binding profiles of tested CRMs are displayed as 'binding matrices' (Fig. 5), where the intensity of blue represents the quantitative ChIP signal (Supplementary Fig. 14). For all classes, CRMs that drive similar spatio-temporal expression have heterogeneous transcription factor binding profiles in terms of (1) the identity of transcription factors occupying the CRM; (2) the duration of binding; and (3) the intensity of the ChIP signal. For example, three CRMs in the Meso class regulate similar expression in the early mesoderm from stage 5–10 but have quite divergent transcription factor binding profiles (Fig. 5a). This diversity of transcription factor binding, which is also reflected in the training set itself (Supplementary Fig. 15), argues against a stringent combinatorial binding code and probably reflects an unequal contribution of each transcription factor to CRM function. Although some transcription factors may act as the key 'switchers' regulating tissue-specific expression (for example, Twi in early mesoderm), other factors may serve as 'multipliers', fine-tuning the levels of CRM activity.

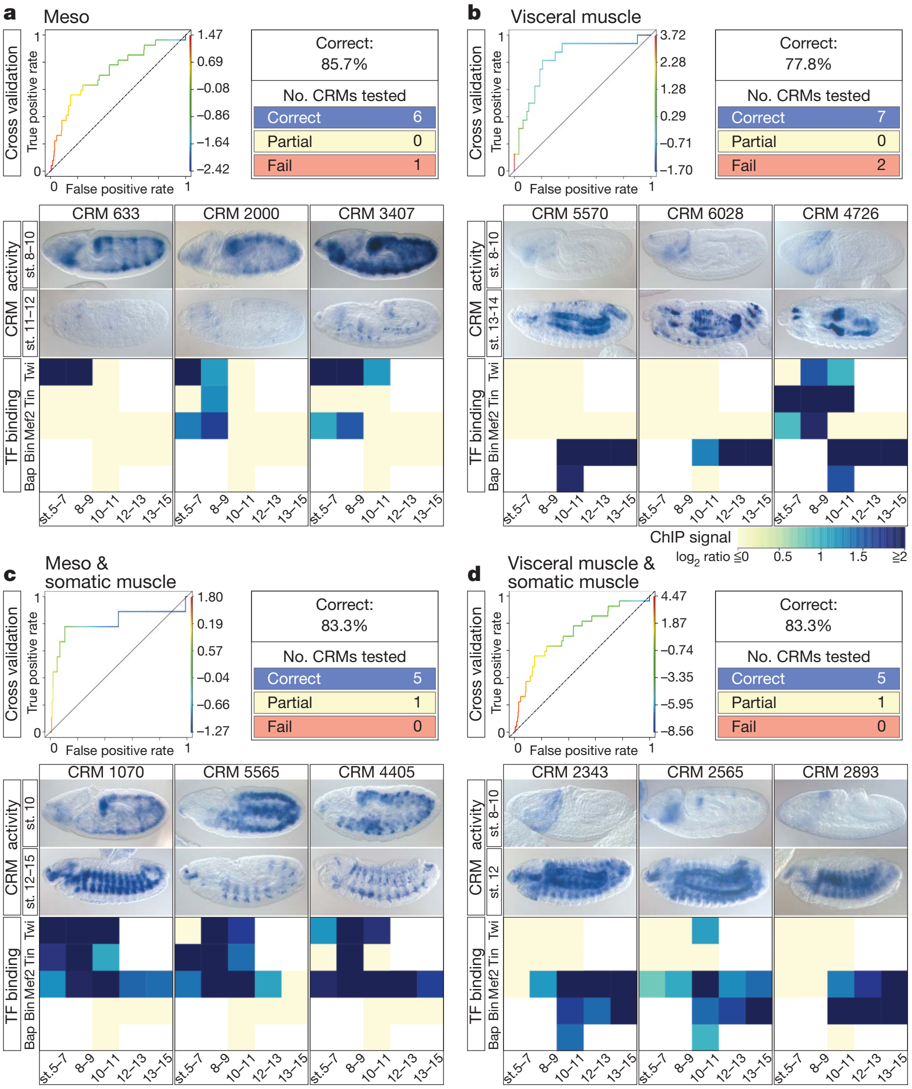

Figure 5 | Validation of CRM spatio-temporal predictions *in vivo*. Results shown for four expression classes: **a**, Meso; **b**, visceral muscle; **c**, Meso & somatic muscle; and **d**, visceral muscle & somatic muscle. **a**–**d**, Top left: ROC plots of SVM performance in leave-one-out cross validations. Line colour represents SVM score (right *y* axis). Top right: table summarizing all results of testing SVM predictions. Below: CRM activity in staged embryos, 3 CRM examples per class. Note the lack of somatic muscle or circular visceral muscle staining at stage 11–12 in Meso class (**a**) and lack of mesoderm staining at stage 8–10 in visceral muscle (**b**) and visceral muscle & somatic muscle (**d**) classes. The CRMs' transcription factor (TF) binding profiles are shown as heat maps using $\log_2$ ChIP-peak height (Supplementary Fig. 14). Diverse patterns of transcription factor binding lead to similar CRM expression. ISH data for all CRMs tested are shown in Supplementary Figs 9–13a, b.

Concluding remarks

Combinatorial binding data are often cited as revealing the regulatory logic of CRMs. However, although binding information is an essential prerequisite, few studies have attempted to bridge the gap from a 'binding code' to the actual regulatory activity. The studies that have modelled developmental CRM activity[14,26,27] used detailed physical models that relied on extensive prior knowledge, including the binding sites within specific CRMs, transcription factor concentrations and transcription factor-DNA affinity estimates. However, this level of information is only available for a very limited number of systems, such as the early patterning of the *Drosophila* embryo. Here we show that combinatorial transcription factor occupancy is sufficient to predict spatio-temporal CRM activity, without the need for prior knowledge of the transcription factors' expression or binding site affinities.

This data-driven approach performed with remarkable accuracy (>70%) given that these transcription factors are expressed in overlapping domains, and that the tissues involved share a developmental history (Fig. 5 and Supplementary Figs 9–13). In the majority of cases, the SVM performed as well or better than a specialist in the field. For example, although the SVM did not have genetic information about the function of these transcription factors, the primary binding signature of CRMs with predicted expression in 'Meso' and 'visceral muscle' was Twi and Bin, respectively, two key transcription factors essential for the development of these tissues (Supplementary Fig. 8). The observed diversity in the occupancy of CRMs regulating similar expression (Fig. 5) questions the generally assumed stringency of regulatory codes. A similar flexibility in regulatory architecture was observed in enhancers regulating a cohort of 19 co-expressed genes in *Ciona*[47] and thus may represent an inherent property of developmental *cis*-regulatory modules.

The predictive power of this approach will only improve as the activity of more CRMs is tested *in vivo* and more transcription factor occupancy data become available. Given the accumulation of such data in many developmental systems, including higher vertebrates[48,49], and the robustness of SVMs to accommodate heterogeneous inputs, this method represents a broadly applicable and accurate approach to predict spatio-temporal enhancer expression in complex developmental systems.

METHODS SUMMARY

ChIP was performed as described previously[50] and hybridized to whole-genome Affymetrix tiling arrays. CRMs were defined as neighbouring clusters of high-confidence transcription factor binding peaks. SVMs were trained with transcription factor binding profiles for five exclusive classes of CRMs with defined activity. SVM predictions were tested using *in vivo* transgenic reporter assays by *in situ* hybridization.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 July; accepted 22 September 2009.

1. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA* **102,** 4936–4942 (2005).
2. Ochoa-Espinosa, A. & Small, S. Developmental mechanisms and cis-regulatory codes. *Curr. Opin. Genet. Dev.* **16,** 165–170 (2006).

3. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94,** 890–898 (2005).
4. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. *EMBO J.* **11,** 4047–4057 (1992).
5. Studer, M., Popperl, H., Marshall, H., Kuroiwa, A. & Krumlauf, R. Role of a conserved retinoic acid response element in rhombomere restriction of Hoxb-1. *Science* **265,** 1728–1732 (1994).
6. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122,** 205–214 (1996).
7. Halfon, M. S. *et al.* Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell* **103,** 63–74 (2000).
8. Yuh, C. H., Bolouri, H. & Davidson, E. H. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128,** 617–629 (2001).
9. Knirr, S. & Frasch, M. Molecular integration of inductive and mesoderm-intrinsic inputs governs even-skipped enhancer activity in a subset of pericardial and dorsal muscle progenitors. *Dev. Biol.* **238,** 13–26 (2001).
10. Oliveri, P., Carrick, D. M. & Davidson, E. H. A regulatory gene network that directs micromere specification in the sea urchin embryo. *Dev. Biol.* **246,** 209–228 (2002).
11. Davidson, B. & Levine, M. Evolutionary origins of the vertebrate heart: Specification of the cardiac lineage in *Ciona intestinalis*. *Proc. Natl Acad. Sci. USA* **100,** 11469–11473 (2003).
12. Hadchouel, J. *et al.* Analysis of a key regulatory region upstream of the Myf5 gene reveals multiple phases of myogenesis, orchestrated at each site by a combination of elements dispersed throughout the locus. *Development* **130,** 3415–3426 (2003).
13. Lee, H. H. & Frasch, M. Nuclear integration of positive Dpp signals, antagonistic Wg inputs and mesodermal competence factors during *Drosophila* visceral mesoderm induction. *Development* **132,** 1429–1442 (2005).
14. Zinzen, R. P., Senger, K., Levine, M. & Papatsenko, D. Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr. Biol.* **16,** 1358–1365 (2006).
15. Rothbacher, U., Bertrand, V., Lamy, C. & Lemaire, P. A combinatorial code of maternal GATA, Ets and β-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development* **134,** 4023–4032 (2007).
16. Sandmann, T. *et al.* A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10,** 797–807 (2006).
17. Zeitlinger, J. *et al.* Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes Dev.* **21,** 385–390 (2007).
18. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21,** 436–449 (2007).
19. Jakobsen, J. S. *et al.* Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes Dev.* **21,** 2448–2460 (2007).
20. Li, X. Y. *et al.* Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* **6,** e27 (2008).
21. Vokes, S. A., Ji, H., Wong, W. H. & McMahon, A. P. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev.* **22,** 2651–2663 (2008).
22. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
23. Davidson, E. H. *The Regulatory Genome—Gene Regulatory Networks In Development and Evolution* 2nd edn (Elsevier Publishers, 2006).
24. MacArthur, S. *et al.* Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **10,** R80 (2009).
25. Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15,** 116–124 (2005).
26. Janssens, H. *et al.* Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene. *Nature Genet.* **38,** 1159–1165 (2006).
27. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451,** 535–540 (2008).
28. Baylies, M. K. & Bate, M. *twist*: a myogenic switch in *Drosophila. Science* **272,** 1481–1484 (1996).
29. Yin, Z., Xu, X. L. & Frasch, M. Regulation of the Twist target gene *tinman* by modular *cis*-regulatory elements during early mesoderm development. *Development* **124,** 4971–4982 (1997).
30. Azpiazu, N. & Frasch, M. tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of *Drosophila. Genes Dev.* **7** (7B), 1325–1340 (1993).
31. Bour, B. A. *et al. Drosophila* MEF2, a transcription factor that is essential for myogenesis. *Genes Dev.* **9,** 730–741 (1995).
32. Lilly, B., Galewsky, S., Firulli, A. B., Schulz, R. A. & Olson, E. N. D-MEF2: a MADS box transcription factor expressed in differentiating mesoderm and muscle cell lineages during *Drosophila* embryogenesis. *Proc. Natl Acad. Sci. USA* **91,** 5662–5666 (1994).
33. Zaffran, S., Kuchler, A., Lee, H. H. & Frasch, M. biniou (FoxF), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in *Drosophila. Genes Dev.* **15,** 2900–2915 (2001).
34. Furlong, E. E. Integrating transcriptional and signalling networks during muscle development. *Curr. Opin. Genet. Dev.* **14,** 343–350 (2004).
35. Sink, H. *Muscle Development in Drosophila* (Birkhäuser, 2006).
36. Liu, Y. H. *et al.* A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Dev. Cell* **16,** 280–291 (2009).
37. Ji, H. & Wong, W. H. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21,** 3629–3636 (2005).
38. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316,** 1497–1502 (2007).
39. Reiss, D. J., Facciotti, M. T. & Baliga, N. S. Model-based deconvolution of genome-wide DNA binding. *Bioinformatics* **24,** 396–403 (2008).
40. Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in *Drosophila melanogaster. Nature Genet.* **38,** 700–705 (2006).
41. Cripps, R. M. *et al.* The myogenic regulatory gene *Mef2* is a direct target for transcriptional activation by Twist during *Drosophila* myogenesis. *Genes Dev.* **12,** 422–434 (1998).
42. Cripps, R. M., Zhao, B. & Olson, E. N. Transcription of the myogenic regulatory gene *Mef2* in cardiac, somatic, and visceral muscle cell lineages is regulated by a Tinman-dependent core enhancer. *Dev. Biol.* **215,** 420–430 (1999).
43. Cripps, R. M., Lovato, T. L. & Olson, E. N. Positive autoregulation of the Myocyte enhancer factor-2 myogenic control gene during somatic muscle development in *Drosophila. Dev. Biol.* **267,** 536–547 (2004).
44. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
45. Halfon, M. S., Gallo, S. M. & Bergman, C. M. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila. Nucleic Acids Res.* **36** (Database issue), D594–D598 (2008).
46. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proc. Natl Acad. Sci. USA* **104,** 3312–3317 (2007).
47. Brown, C. D., Johnson, D. S. & Sidow, A. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317,** 1557–1560 (2007).
48. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35** (Database issue), D88–D92 (2007).
49. Choo, B. G. *et al.* Zebrafish transgenic Enhancer TRAP line database (ZETRAP). *BMC Dev. Biol.* **6,** 5 (2006).
50. Sandmann, T., Jakobsen, J. S. & Furlong, E. E. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nature Protocols* **1,** 2839–2855 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to M. Leptin for providing an independent assessment of the expression patterns driven by tested CRMs. We thank H. Gustafson for fly work, J. de Graaf for array hybridizations, S. Müller for embryo injections, and R. Bourgon for sharing code on signal peak identification. We thank all members of the Furlong laboratory for discussions and comments on the manuscript. This work was supported by a grant to E.E.M.F. and by a fellowship to R.P.Z. from the Human Frontiers Science Program.

Author Contributions M.B. performed ChIP experiments. R.P.Z., E.E.M.F. and C.G. generated CAD. R.P.Z. performed transgenic reporter experiments including *in situ* hybridizations and imaging. C.G. performed ChIP data analysis and motif analysis. J.G. devised the statistical and SVM analyses. E.E.M.F., R.P.Z., C.G. and J.G. formulated the hypotheses, designed experiments and wrote the manuscript.

Author Information All ChIP data are available in ArrayExpress under accession numbers E-TABM-648, E-TABM-649, E-TABM-650, E-TABM-651 and E-TABM-652, and the array design under A-AFFY-53. The CRM coordinates and transcription factor occupancy is available at http://furlonglab.embl.de/. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.E.M.F. (furlong@embl.de).

METHODS

ChIP-on-chip assays. For each transcription factor and time point, two independent chromatin immunoprecipitations (ChIPs) were performed using two different antibodies and compared to two independent mock ChIPs using rabbit pre-immune sera (Supplementary Methods). ChIPs were optimized by assaying for enrichment of a known binding site using real-time PCR as previously described[50], amplified and hybridized to Affymetrix GeneChip *Drosophila* Tiling array1.0R (Supplementary Methods).

Detection of transcription-factor-bound regions and peaks. All bioinformatics analyses were done using BDGP *Drosophila melanogaster* genome version 4 (UCSC dm2)[51] and the Flybase 4.3 genome annotation release[52]. Mapping of the Affymetrix GeneChip *Drosophila* Tiling 1.0R probes to the genome was obtained from the MAT website (http://liulab.dfci.harvard.edu/MAT/). Quantile normalization[53] was applied to the four data sets (two ChIP experiments, two mock controls) for each of the fifteen conditions. Significantly enriched genomic regions at each condition were identified as consecutive probes with significantly positive log-ratios of experiment over control using a Hidden Markov Model (HMM)-based algorithm (TileMap[37]). A cutoff on the probe-wise maximum *a posteriori* probability of each region returned by the HMM was determined manually for each data set (Supplementary Table 1). The top 5% regions below cutoff exhibiting more than 75% overlap with one or more above cutoff region (at any of the 15 conditions) were rescued and included in the final 'high-confidence' transcription factor binding profile. For each enriched region, probe intensity peaks were identified as extrema on a smoothed curve of the $\log_2$-ratio signal[40]. This 'peak height' was used as a quantitative measure of transcription factor enrichment.

Enrichment of ChIP peaks near motifs or other peaks. Enrichment for proximal genomic distances was assessed for (1) distances of transcription factor binding peaks to their closest PWM match (Supplementary Fig. 1) and (2) distances of transcription factor binding peaks in one condition to the closest peak in another condition (for example, distances of Twist binding peaks at 4–6 h to the closest Twist binding peak at 2–4 h, see Supplementary Fig. 2 and Fig. 2b). The background distribution of distances was derived assuming uniform distribution over the merged TileMap regions (that is, significantly enriched regions covered by the tiling array). Enrichment over background was defined as the ratio of the frequency in the data set over background frequency and was robustly estimated using a moving average with a 30-bp window. Equi-tailed 95% confidence intervals of the enrichment were estimated by re-sampling 1,000 times the observed distances, with replacement.

Computing ChIP CRMs from ChIP peaks. ChIP peaks across all conditions were clustered using a neighbour joining approach with a maximum distance of 200 bp between adjacent peaks. Peak cluster boundaries were extended by 100 bp past the terminal peak position on each side to account for peak position precision. Resulting genomic boundaries define the ChIP CRMs of the ChIP-CRM atlas (Supplementary Table 5).

Iterative position weight matrix optimization. Initial position weight matrices (PWMs) were gathered either from the literature or by *de novo* motif discovery using the software RSAT[54] and GAPWM[55] (Supplementary Methods). Positive genomic regions were defined as 200-bp regions centred on ChIP peaks. PWMs were iteratively updated, taking as binding sites the best hits predicted by the PWM of the previous iteration within relevant positive regions. In computing PWMs, binding sites were weighted by their peak height, giving more importance to peaks with stronger signal. To assess sensitivity and specificity, negative regions were randomly chosen from the repeat-masked genome (excluding exons and bound regions as defined by TileMap), matching the positive regions in number and length. Binding site predictions were generated using Patser[56]. For a given score cutoff, a region was regarded as 'positive' if it contains at least one match above cutoff (ROC curves, Supplementary Fig. 5). Final predictions were performed at a score cutoff corresponding to an estimated 40% false discovery rate.

Conservation of TFBS within bound regions. For each transcription factor, ChIP CRMs were split into two groups: a group containing ChIP CRMs bound by the considered transcription factor and a group containing ChIP CRMs not bound by that transcription factor. The latter was used as 'background' regions to control for general sequence biases in ChIP CRMs, such as GC content. In addition, random CRM sets matching the transcription-factor-bound regions in number and length were generated by sampling from the repeat-masked genome (excluding exons and bound regions as defined by TileMap). TFBSs were predicted as described above. Conservation of TFBSs was evaluated in two ways. First, the average of the PhastCons[44] score (dm2/phastCons15way from UCSC, http://hgdownload.cse.ucsc.edu/goldenPath/dm3/phastCons15way/) over the bases of the TFBS was computed for either the best scoring TFBS prediction (Fig. 2c) or all TFBS predictions in each CRM (Supplementary Fig. 6a). Second, pair-wise alignments between *D. melanogaster* (dm2) and *D. simulans* (droSim1), *D. yakuba* (droYak1), *D. ananassae* (droAna1), *D. pseudoobscura* (dp3) and *D. virilis* (droVir1) were obtained from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html#fruitfly). Best scoring TFBS predictions (Supplementary Fig. 6b–f) for each ChIP CRM and for each random CRM were used to extract the corresponding sequence from each pair-wise alignment (ungapped alignments only). A TFBS prediction was scored as 'conserved' in a particular species if its aligned sequence triggered a match scoring above cutoff (as defined earlier, see Supplementary Fig. 5), or was otherwise scored as 'not conserved' (unaligned TFBSs were also counted as 'not conserved').

Compiling the CRM Activity Database (CAD). The CRM Activity Database (CAD) compiles *in vivo Drosophila melanogaster* enhancer expression data from REDFly 2.0[45], literature surveys and our own experiments. CAD is the result of a semi-automated procedure whereby (1) each enhancer activity is manually reviewed and (2) redundancy within and across data sources is eliminated. We reviewed original literature and enhancer expression pattern images and added missing annotations (using the *Drosophila* gross anatomy ontology, http://www.obofoundry.org/), focusing on mesoderm and muscle subset expression. We then manually assessed redundancy of overlapping enhancers; in particular, we minimized enhancer boundaries where possible (overlapping enhancers of varying sizes with indistinguishable activities as assayed) and merged enhancers where necessary (extensively overlapping enhancers having distinct activities). Overlapping enhancers not falling into these two categories were left unmodified. Supplementary Table 4 provides all final CAD entries and references to the original enhancers.

Defining mesodermal and muscle expression patterns. While generating CAD, we analysed the expression patterns driven by previously characterized CRMs and annotated them using the *Drosophila* gross anatomy ontology (http://www.obofoundry.org/), describing the timing and location of expression. Particular attention was paid to expression in the presumptive and unspecified mesoderm (Meso, stage 5–9), and in 3 of its major derivatives (stage 10 onwards)—somatic musculature (SM), visceral musculature (VM), and heart musculature (CM)[35]. To define mesodermal and muscle expression classes, the anatomical vocabulary of CRMs in CAD was mapped to four high-level master terms: 'Meso', 'SM', 'VM' and 'CM' (Supplementary Table 3). VM-annotated CRMs show expression in circular trunk visceral muscle and/or foregut and hindgut visceral muscle. As the development of longitudinal visceral muscle (lVM) is regulated by transcription factors not included in our data sets, CRMs regulating lVM expression were not annotated as 'VM'.

The CRM training set. The training set consisted of 310 computed ChIP CRMs that overlap 250 annotated enhancers in CAD (Supplementary Table 8). We restricted this study to combinatorial expression classes that contained at least 9 positive CRMs: 'Meso' only, 'SM' only, 'VM' only, 'VM & SM' and 'Meso & SM'. These expression classes are mutually exclusive. To be included in the training set for 'Meso', 'VM', 'SM', 'Meso & SM' and 'VM & SM' the CRM had to drive expression in the specific tissue(s), or in subdomains of the tissue(s), at one or more stages of development. CRMs were included that drive expression in the defined expression class, regardless of expression in additional tissues (that is, outside Meso, visceral muscle, somatic muscle and heart musculature), such as ectoderm.

SVM parameter selection. We used support vector machines (SVM) with a radial basis function kernel, which requires setting two parameters, the penalization coefficient C and the kernel precision γ. Optimal parameter setting was achieved by evaluating the performance for several parameter values. Performance of each SVM was evaluated using the area under the curve (AUC) of receiver operating characteristic (ROC) curves obtained with a leave-one-out cross-validation scheme (Supplementary Information and Supplementary File 1). Sensitivity and specificity functions were smoothed using Gaussian kernel density estimates[57]. Once optimal parameters were identified, a final SVM was trained on the full training set and then applied to the full ChIP-CRM atlas. SVM scores and specificity levels for the 8,008 CRMs of the atlas, and for each expression class, are provided in Supplementary Table 9. For each ChIP CRM, we predicted expression if its SVM score corresponded to a specificity greater than 95% in a given expression class. In cases where CRMs scored with higher than 95% specificity in more than one class, we predicted expression corresponding to the class of higher specificity.

Transgenic reporter assays. To assay ChIP CRMs for enhancer activity, the genomic regions were placed in front of a minimal promoter driving a *lacZ* reporter gene in pDuo2n-attB (Supplementary Information). All constructs were targeted to chromosomal position 51C via attB/phiC31 mediated integration[46]. Transgenic lines were balanced, homozygosed and tested by *in situ* hybridization (ISH). *lacZ* ISH was performed either colorimetrically, or fluorescently while co-visualizing appropriate marker genes. No reporter gene activity was detected for an empty vector integrated into that genomic position.

Scoring CRM activity in transgenic reporter assays. Colorimetric *lacZ* ISH at all stages of embryogenesis were performed to annotate expression patterns driven by the CRMs (Fig. 5, Supplementary Figs 9a–13a and Supplementary Methods). Annotated CRM activities are compiled in Supplementary Table 12a, b. Two approaches were taken to confirm the tissue-specific activity of all CRMs.

First, we confirmed all *lacZ* reporter-based annotations using multiplex fluorescent ISH followed by confocal imaging. ISHs were performed using antisense probes against appropriate tissue-specific markers (Supplementary Methods), which were imaged in green (mesoderm and visceral muscle), or blue (somatic muscle), together with a probe directed against the *lacZ* reporter driven by the CRM (red). Overlapping *lacZ* and marker gene expression unequivocally demonstrates tissue-specific CRM activity (Supplementary Figs 9b–13b). The CRMs were annotated as driving expression in a tissue if the expression of the *lacZ* reporter overlapped with that of the marker gene specific for that tissue at any stage of development (Supplementary Table 12a).

Second, we asked an external expert on *Drosophila* mesoderm and muscle development to annotate the CRMs' expression independently (M. Leptin, Institute of Genetics, University of Cologne). M. Leptin was presented with the same embryo pictures shown in Supplementary Figs 9–13, but ordered by increasing CRM ID number and without any information about the respective predictions (Supplementary Information and Supplementary Table 12b).

CRMs were scored as correct SVM predictions if they drove expression within the predicted tissue (or a subregion of that tissue) at any stage of development and therefore show co-expression with the specific tissue marker. CRMs that drove expression in the predicted tissue, but also in another unpredicted mesoderm tissue, were scored as partial. For the two combinatorial expression classes (Meso & SM and VM & SM), CRMs that only drove expression in one of the two predicted tissues were also scored as partial. CRMs were scored as fail if their expression did not overlap marker gene expression as predicted by the SVM.

51. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**, RESEARCH0079 (2002).
52. Tweedie, S. *et al.* FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.* **37** (Database issue), D555–D559 (2009).
53. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
54. Thomas-Chollier, M. *et al.* RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* 36 (Web Server issue) W119–W127 (2008).
55. Li, L., Liang, Y. & Bass, R. L. GAPWM: a genetic algorithm method for optimizing a position weight matrix. *Bioinformatics* **23**, 1188–1194 (2007).
56. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
57. Lloyd, C. J. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *J. Am. Stat. Assoc.* **93**, 1356–1364 (1998).