Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos

Stephen S Gisselbrecht[1], Luis A Barrera[1–3], Martin Porsch[1,4], Anton Aboukhalil[1,5], Preston W Estep III[6], Anastasia Vedenko[1], Alexandre Palagi[1,7], Yongsok Kim[8], Xianmin Zhu[8], Brian W Busser[8], Caitlin E Gamble[8], Antonina Iagovitina[1,9], Aditi Singhania[8], Alan M Michelson[8] & Martha L Bulyk[1–3,9,10]

Transcriptional enhancers are a primary mechanism by which tissue-specific gene expression is achieved. Despite the importance of these regulatory elements in development, responses to environmental stresses and disease, testing enhancer activity in animals remains tedious, with a minority of enhancers having been characterized. Here we describe 'enhancer-FACS-seq' (eFS) for highly parallel identification of active, tissue-specific enhancers in *Drosophila melanogaster* embryos. Analysis of enhancers identified by eFS as being active in mesodermal tissues revealed enriched DNA binding site motifs of known and putative, previously uncharacterized mesodermal transcription factors. Naive Bayes classifiers using transcription factor binding site motifs accurately predicted mesodermal enhancer activity. Application of eFS to other cell types and organisms should accelerate the cataloging of enhancers and understanding how transcriptional regulation is encoded in them.

In metazoans, gene expression is regulated in a tissue-specific manner predominantly via noncoding genomic regions referred to as *cis*-regulatory modules (CRMs) that regulate the expression of typically the nearby gene(s)[1]. CRMs contain one or more DNA binding sites for one or more sequence-specific transcription factors that activate or repress gene expression. CRMs that activate gene expression are frequently referred to as transcriptional enhancers[2].

The fruit fly *D. melanogaster* has served as a powerful model organism for studies of transcriptional enhancers[2]. It has been estimated that there are ~50,000 enhancers in the *D. melanogaster* genome[3], yet to date the tissue-specific activities of only ~1,800 are known[4]. Technology for identifying enhancers active in particular cell types would aid in defining functional *cis*-regulatory elements and would facilitate computational identification of sequence features important for cell type–specific enhancer activity. Currently, regions identified by chromatin immunoprecipitation (ChIP) to be occupied by transcription factors are tested by low-throughput, traditional reporter assays[5,6]. Automated image analysis of reporter assays in embryos[3,7] requires vast infrastructure and resources. Although highly parallel reporter assays have been developed recently[8–13], none directly identify enhancer activity in a genomic context (integrated into the genome) in particular cell types of interest in a whole animal.

Our technology, termed 'enhancer-FACS-seq' (eFS), achieves highly parallel identification of active, tissue-specific transcriptional enhancers in whole *Drosophila* embryos (**Fig. 1a** and **Supplementary Fig. 1**). As with traditional enhancer assays, each candidate CRM (cCRM) is cloned upstream of a reporter gene. Our key innovation is the replacement of microscopy to screen for tissue-specific enhancers with FACS of dissociated cells. In each fly, one marker (here, rat CD2 cell-surface protein[14]) is used to label cells of a specific tissue for FACS, and the other marker (here, GFP) is used as a reporter of cCRM activity. Cells are sorted by tissue type and then by GFP fluorescence, allowing screening of hundreds of cCRMs in a time-efficient and cost-efficient manner.

RESULTS

Library of candidate *cis*-regulatory modules

We focused on embryonic mesoderm as our model system because it comprises a variety of cell types, the major regulatory factors governing mesoderm development are conserved between vertebrates and *Drosophila*[15], and many data sets are available for genomic features associated with active enhancers. We created a plasmid library of hundreds of reporter constructs for ~1-kb cCRMs (**Supplementary Note 1** and **Supplementary Table 1**) composed of sequences located next to mesodermally expressed

[1]Department of Medicine, Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]Harvard–Massachusetts Institute of Technology, Division of Health Sciences and Technology, Harvard Medical School, Boston, Massachusetts, USA. [3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts, USA. [4]Institute of Computer Science, Martin Luther University of Halle-Wittenberg, Halle, Germany. [5]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]TeloMe, Inc., Waltham, Massachusetts, USA. [7]Bioengineering Department at Polytech Nice Sophia, University of Nice Sophia Antipolis, Nice, France. [8]Laboratory of Developmental Systems Biology, Genetics and Developmental Biology Center, National Heart, Lung, and Blood Institute, US National Institutes of Health, Bethesda, Maryland, USA. [9]Systems Biology Graduate Program, Harvard University, Cambridge, Massachusetts, USA. [10]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

RECEIVED 19 FEBRUARY; ACCEPTED 3 JUNE; PUBLISHED ONLINE 14 JULY 2013; DOI:10.1038/NMETH.2558

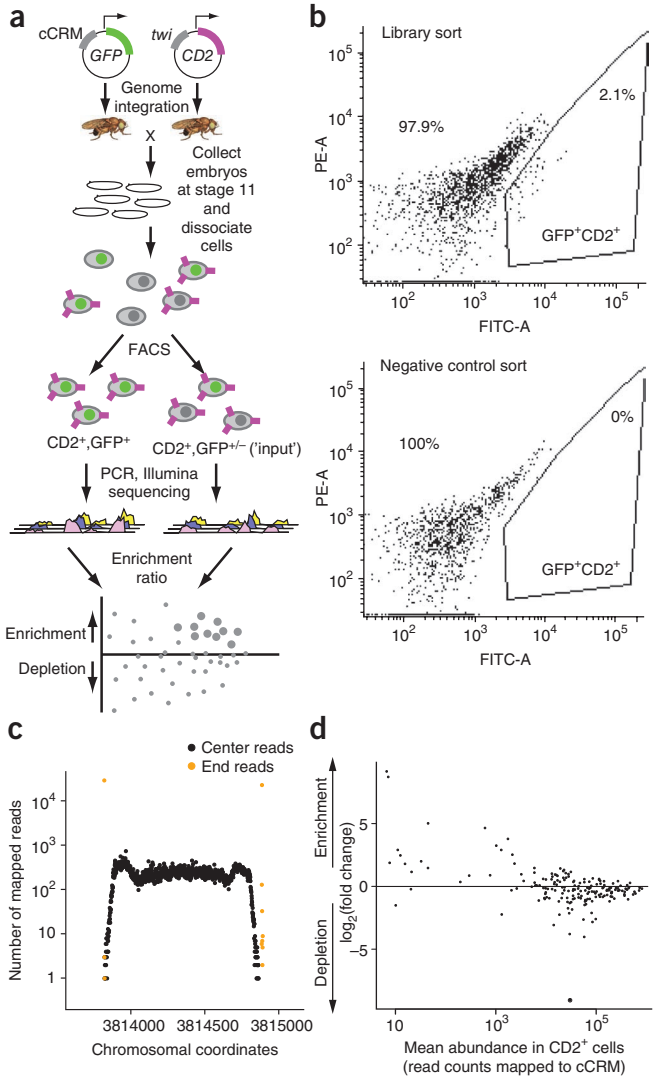


Figure 1 | eFS methodology. (**a**) Overall design of eFS. (**b**) FACS of GFP$^+$CD2$^+$ cells prepared from embryos resulting from a cross of *Mef2-I-E$_{D5}$:CD2* females to cCRM library transgenic males (top) and wild-type (GFP$^-$) males (bottom). Plotted is yellow ('PE-A') versus green ('FITC-A') fluorescence for cells that pass the CD2$^+$ gate out of 10^6 cells prepared from embryos. Percentages indicate fraction of cells called GFP$^+$ (in the depicted polygonal FACS gate) or GFP$^-$ (outside the depicted FACS gate). (**c**) Representative example of a cCRM, surrounded by native genomic flanking sequence, detected by eFS. (**d**) Enrichment ratios for cCRMs in *twi*:CD2$^-$ cells, as compared to *twi*:CD2$^+$ cells. Large points: $P_{adj} < 0.1$ (significantly enriched); small points: $P_{adj} > 0.1$.

genes and in addition having one of the following features: regions identified by ChIP[6] as bound by at least one of the somatic mesoderm transcription factors Twist (Twi), Tinman (Tin) or Myocyte enhancing factor 2 (Mef2); regions identified as bound by the transcriptional coactivator CREB binding protein (CBP)[16,17]; regions containing DNase I–hypersensitive sites (DHS)[18]; dense clusters of evolutionarily conserved motif occurrences for mesodermal transcription factors[19]; and additional regions surrounding mesodermal genes not covered by the aforementioned features (**Supplementary Note 2**).

eFS experiments

In our cCRM plasmid library, each cCRM flanked by *attL* sites was cloned into a vector that contains *attR* sites for Gateway cloning, the phiC31 *attB* site, the mini-white (mini-*w*) gene and a reporter cassette comprising the *Hsp70* minimal promoter driving expression of a nuclear localization signal–tagged *EGFP* gene with an SV40 polyadenylation sequence. We injected the library into two batches of *Drosophila* embryos carrying a single phiC31 *attP* site on the second chromosome. This strain of flies expresses a nuclear-localized phiC31 integrase under the control of the nanos (*nos*) promoter, which causes mRNA to be produced during oogenesis and deposited in the egg before fertilization. The recombination of an *attP* and an *attB* site, mediated by the phiC31 integrase, produces an *attL* and an *attR* site (distinct from and not cross-reacting with those used in the Gateway system), which are not themselves substrates for the integrase; thus, integration is nonreversible and one integration event destroys the *attP* site used, preventing any further events at that genomic locus. Each resulting embryo has one *GFP* reporter under the control of one cCRM integrated at the same genomic site by the phiC31 integrase[20]. Use of a site-specific integrase avoids artifacts that would result if more than one cCRM were present in a cell and also avoids potential position effects on enhancer activity. In the first batch, we injected ~3,500 embryos and crossed transformant males (selected by eye color) to females from two different *CD2* lines to identify enhancers active in distinct tissues: *twi:CD2* for whole mesoderm, and *I-E$_{D5}$:CD2* (*Mef2-I-E$_{D5}$:CD2*) for a subset[21] of largely fusion-competent myoblasts (FCMs). In the second batch, we injected ~4,500 embryos and crossed transformant males to *duf:CD2* females to identify activity in somatic mesoderm founder cells[22].

At developmental stages 11–12, we dissociated embryos and purified them by FACS. From the *twi:CD2* embryos, we collected ~315,000 GFP$^+$CD2$^+$ cells and ~198,000 GFP$^+$CD2$^-$ cells as well as 1×10^6 'input' cells regardless of GFP status (Online Methods, **Fig. 1b**, **Supplementary Fig. 2** and **Supplementary Table 2**). We collected fewer GFP$^+$CD2$^+$ cells from the *Mef2-I-E$_{D5}$:CD2* and *duf:CD2* embryos (**Supplementary Table 2**) because the *Mef2-I-E$_{D5}$* enhancer is active in ~50-fold fewer cells than the *twi* enhancer, which is active in roughly 50,000 cells at this stage, whereas the *duf* enhancer is active only in most of the 660 founder cells per embryo, nearly an order of magnitude fewer cells than for the *Mef2-I-E$_{D5}$* enhancer.

We extracted genomic DNA from the collected cells, amplified the cCRMs by PCR and sequenced the resulting amplicons on the Illumina platform. We mapped the sequencing reads (**Fig. 1c** and **Supplementary Table 3**) to the *D. melanogaster* genome using segemehl software[23] (**Supplementary Fig. 3**). We detected 213 and 400 cCRMs (false discovery rate (FDR) $< 5 \times 10^{-5}$; Online Methods) as having integrated into the fly genome from the first and second batches of injections, respectively. The greater number of cCRMs detected from the second batch was likely due to the fact that we collected transformant progeny from more injected embryos.

To evaluate the enhancer activity of the detected cCRMs, we calculated each cCRM's enrichment in a particular cell population as compared to the corresponding 'input' sample (**Fig. 1a**) using DESeq software[24]. The input sample provides information on the baseline read counts resulting from cCRM representation in the embryo populations. In control experiments CD2$^+$ and CD2$^-$

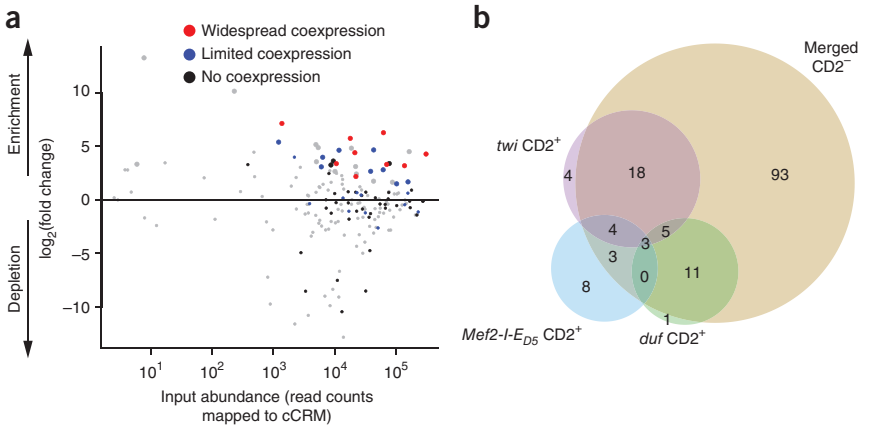


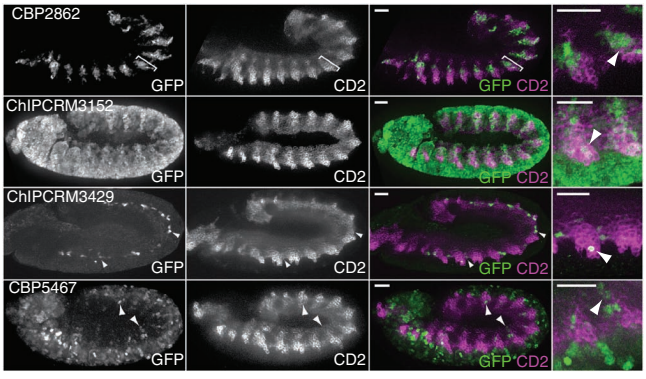
Figure 2 | Active enhancers identified from eFS data. (**a**) Enrichment ratios for cCRMs in *twi*:CD2$^+$GFP$^+$ cells, as compared to *twi:CD2* input cells. Large points: $P_{adj} < 0.1$ (significantly enriched); small points: $P_{adj} > 0.1$. Results from traditional reporter assays (**Supplementary Fig. 5**) revealed cCRMs whose GFP expression shows widespread, limited or no coexpression with *twi:CD2* expression. (**b**) Active enhancers ($P_{adj} < 0.1$) identified from different cell populations: *twi*:CD2$^+$; *Mef2-I-E$_{D5}$*:CD2$^+$; *duf*:CD2$^+$; nonredundant union of *twi*:CD2$^-$, *Mef2-I-E$_{D5}$*:CD2$^-$ and *duf*:CD2$^-$ (merged CD2$^-$).

cells exhibited no major differences in their cCRM content (**Fig. 1d**). Therefore, we used CD2$^+$ cells as input sample for so-called '*twi*:CD2$^+$GFP$^+$' cells, whereas for the rarer FCM and founder cell types, we used CD2$^-$ cells as input (**Supplementary Fig. 2**).

In total, by eFS we identified 150 of the detected cCRMs as being active enhancers (adjusted *P* value (P_{adj}) < 0.1) in at least one cell population. Of these, 57 were active mesodermal enhancers: 34 in whole mesoderm (**Fig. 2a**), 18 in FCMs and 20 in founder cells (**Supplementary Fig. 4**). Of these 57 active mesodermal cCRMs, 12 overlapped by at least 100 base pairs (bp) with a known mesodermal enhancer at an overlapping developmental time point in the REDfly database[25] (**Supplementary Table 4**), and the remaining 45 were putative new mesodermal enhancers, including 16 in FCMs and 14 in founder cells. Analysis of GFP$^+$CD2$^-$ cells collected from *twi:CD2*, *Mef2-I-E$_{D5}$:CD2* and *duf:CD2* embryos revealed 93 putative nonmesodermal enhancers (**Fig. 2b** and **Supplementary Table 4**). Comparing to enhancers identified from a recent screen of a genomic DNA library for enhancer activity in the S2 cell line and in cultured ovarian somatic cells[13], only 13 of the 57 mesodermal enhancers and 11 of the 93 nonmesodermal enhancers identified by eFS overlap by at least 100 bp with enhancers found in that study. This comparison highlights the value of eFS for identifying enhancers active in particular cell types of interest in whole embryos.

Validation of eFS results

To validate our eFS results, we performed traditional reporter assays in whole *Drosophila* embryos (Online Methods). For the *twi*:CD2$^+$ eFS data, we tested 69 of the cCRMs, including: 21 putative active mesodermal enhancers ($P_{adj} < 0.1$) and 48 putative inactive cCRMs ($P_{adj} > 0.1$). The specificity of eFS was excellent among significantly enriched cCRMs: 18 of the 21 tested



putative mesodermal enhancers drove expression in mesoderm at stages 11–12 (**Fig. 3** and **Supplementary Fig. 5**). eFS exhibited moderate sensitivity for significantly enriched enhancers that were active in relatively few mesodermal cells: nine enhancers had expression patterns that were manually assessed as 'widespread coexpression' (expression in a majority of strongly *twi*:CD2$^+$ cells) (for example, cCRMs named CBP2862 and ChIPCRM3152; **Fig. 3**), and the other nine drove 'limited coexpression' in smaller subsets of *twi*:CD2$^+$ cells (for example, ChIPCRM3429 and CBP5467; **Fig. 3**). Twelve of the 48 putative inactive cCRMs drove 'limited coexpression' (**Supplementary Fig. 5** and **Supplementary Table 4**). Some of these eFS false negatives drove expression in cells that expressed low levels of CD2 and might have been missed by our relatively stringent FACS gate for collecting *twi*:CD2$^+$ cells. In most cases, the observed expression domain was linked to an adjacent gene's expression (**Supplementary Table 5**). Although the data are slightly noisier for FCM and founder cell enhancers (6 of 9 tested putative FCM enhancers and 9 of 11 tested putative founder cell enhancers drove mesodermal expression; **Supplementary Fig. 5**), likely because we collected roughly 20-fold fewer CD2$^+$GFP$^+$ cells from the more specific *Mef2-I-E$_{D5}$:CD2* and *duf:CD2* lines, the results nevertheless demonstrate that eFS can identify enhancers active in rarer cell types. In addition, the majority of cCRMs identified by eFS as active in any of the three CD2$^-$GFP$^+$ cell collections (35 of 47 cCRMs tested) were indeed active at this developmental stage (**Supplementary Table 6**).

Comparisons of eFS data to other genomic data types

We examined the eFS-identified enhancers for enrichment of known enhancer-associated chromatin marks. Comparison to data from batch isolation of tissue-specific chromatin for immunoprecipitation (BiTS-ChIP) for mesodermal cells from stage 10–11 embryos[26] showed that acetylation of histone H3 on lysine 27 (H3K27ac), monomethylation of histone H3 on lysine 4

Figure 3 | Validations of enhancers identified as active ($P_{adj} < 0.1$) by eFS. Immunofluorescence micrographs showing GFP expression driven by the indicated cCRM (labels in upper left corner of leftmost images; coordinates provided in **Supplementary Table 1**) and mesodermal CD2 in stage 11–12 *Drosophila* embryos. Brackets indicate 'widespread coexpression' in somatic mesoderm for CBP2862. Coexpression (arrowheads) was observed as green and purple in the same cells, as the GFP in these embryos is nuclear and CD2 is expressed on the cell surface. Coexpression was assessed with the annotator being blind to the predicted activity of the cCRMs. Scale bars, 50 μm.

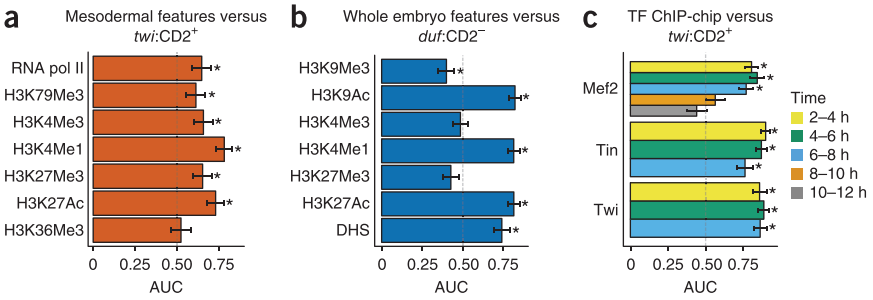


Figure 4 | Enrichment of various genomic marks among eFS-identified enhancers. Enrichment of the indicated genomic features (DHS, histone modifications, transcription factor (TF) ChIP binding) associated with active enhancers in whole mesoderm (**a**,**c**; *twi*:CD2$^+$) or approximately whole embryos (**b**; *duf*:CD2$^-$). *P < 0.05, Wilcoxon-Mann-Whitney U test. Error bars, 1 s.d. Developmental time points relative to egg deposition are indicated.

(H3K4me1), H3K4 trimethylation (H3K4me3), H3K79me3 and occupancy by RNA polymerase II[26–29] were enriched (area under receiver operating characteristic curve (AUC) ≥ 0.6, P < 0.05 by Wilcoxon-Mann-Whitney U test) among enhancers found to be active in mesoderm by eFS (**Fig. 4a**). However, in contrast to a prior report that H3K27me3 was depleted among active mesodermal enhancers[26], we found H3K27me3 to be enriched among mesodermal enhancers. We also observed enrichment of H3K27ac, H3K4me1 and H3K9ac when comparing modENCODE data for 4–8-h whole embryos[17] to active enhancers identified by eFS in *duf*:CD2$^-$ cells, which approximate whole embryo samples (**Fig. 4b** and **Supplementary Note 3**). Although H3K9ac is known as a mark of active transcription start sites[30], our observed enrichment of H3K9ac among active enhancers supports the observation of H3K9ac in the 'strong enhancer' chromatin state in human cells[31].

Our enhancer data allowed us to investigate which genomic data types[6,16–18] provide the greatest utility in identifying enhancers. Occupancy by sequence-specific transcription factors (Twi, Tin, Mef2, Bagpipe (Bap) and Biniou (Bin)) expressed specifically in the mesoderm was most enriched among active mesodermal enhancers (**Fig. 4c** and **Supplementary Fig. 6**). DHSs[18] were nearly as enriched as enhancer-associated histone modifications (**Fig. 4b**, **Supplementary Fig. 6**). Among enhancers found in whole mesoderm, we observed the greatest enrichment for regions bound by Tin at 2–4 h (stages 5–7), suggesting that Tin might be a pioneer factor[32] that premarks mesodermal enhancers that are active later in development. These same Tin-bound enhancers exhibited enhanced Tin binding at 4–6 h (stages 8–9; data not shown) and were consistent with *tin* being essential for specification of ventral founder cells[33] and also with *tin* activity and putative Tin binding sites being required for the activity in ventral muscle progenitors of an enhancer that does not become expressed until after Tin protein expression has become restricted to the dorsal mesoderm[34]. Our observed enrichment of Mef2, Twi and Tin occupancy at 4–6 h or 6–8 h (stages 10–11) among enhancers identified from *Mef2-I-E_{D5}*:CD2$^+$ cells supports the role of Mef2, Twi and Tin in regulating FCM genes coordinately with Lameduck (Lmd)[35].

Enrichment of transcription factor binding site motifs

We separately analyzed each of the three sets of eFS-identified mesodermal enhancers (whole mesoderm, FCMs or founder cells) for over-represented motifs and pairwise motif combinations that might be required for enhancer activity. We used the PhylCRM and Lever algorithms[19] to determine enrichment of matches, scored according to their evolutionary conservation, to 567 publicly available *Drosophila* transcription factor binding site motifs[6,35–38] (Online Methods). Many motifs were significantly enriched (AUC ≥ 0.65, FDR ≤ 0.1) either individually or in pairwise combination (**Fig. 5a**, **Supplementary Figs. 7** and **8**, and **Supplementary Table 7**) for the whole-mesoderm and FCM enhancers.

For each of these two sets of enhancers, we observed strong enrichment of the primary, known master regulator of that cell population: Twi for whole mesoderm[39] and Lmd for FCMs[21,40]. We found motifs for other known mesodermal regulators in enriched combinations, including Bap, Lola-PC and Mef2 in whole mesoderm, and Twi and Mef2 in FCMs. We also saw strong enrichment of motifs for sequence-specific DNA-binding proteins z, grh and Trl (also known as GAGA factor (GAF)) that participate in recruitment of chromatin-modifying PcG and trxG proteins[41], supporting prior findings of the enrichment of the z and/or Trl motifs among regions bound by Mef2, Twi or Tin in ChIP-microarray studies[42]. For the eFS-identified founder cell enhancers, no individual motifs or combinations thereof met our statistical significance criteria of AUC ≥ 0.65 and FDR ≤ 0.1, although a few combinations for known and candidate mesodermal regulators narrowly missed our thresholds (**Supplementary Table 7**).

FCM enhancers exhibited enrichment for a variety of motifs (among them Twi and Trl) in combination with a Lmd motif, supporting the previously observed enrichment of these motifs in Lmd ChIP-sequencing (ChIP-seq) peaks[35]. We also observed many significantly (AUC ≥ 0.65, FDR ≤ 0.1) enriched motif combinations (many involving the uncharacterized zinc-finger protein CG7928) not found in the Lmd ChIP-seq study[35]. As eFS data are not constrained by occupancy by a particular transcription factor, they allow for more unbiased identification of regulatory motifs. We also observed enrichment of many motif combinations comprising a master regulator and a factor with either ubiquitous or mesoderm-specific expression at the appropriate stage but no known role in mesoderm development (for example, schlank and Lola-PK), suggesting previously unidentified regulators of mesodermal expression (**Fig. 5a** and **Supplementary Fig. 7**).

Classifier to predict mesodermal enhancer activity

We developed a machine learning approach to model whether cCRMs will be active or inactive in the mesoderm or specifically in FCMs. We selected the mesodermal transcription factor binding site motifs, independently in tenfold cross-validation (we split cCRMs into ten equally sized sets, and in each of ten iterations we used 90% of the sets to learn discriminatory motifs and withheld the remaining 10% for subsequent testing as described below), that were most discriminatory in distinguishing active versus inactive cCRMs (Online Methods). We then trained a naive Bayes classifier[43] (**Fig. 5b**) based on the number and quality of

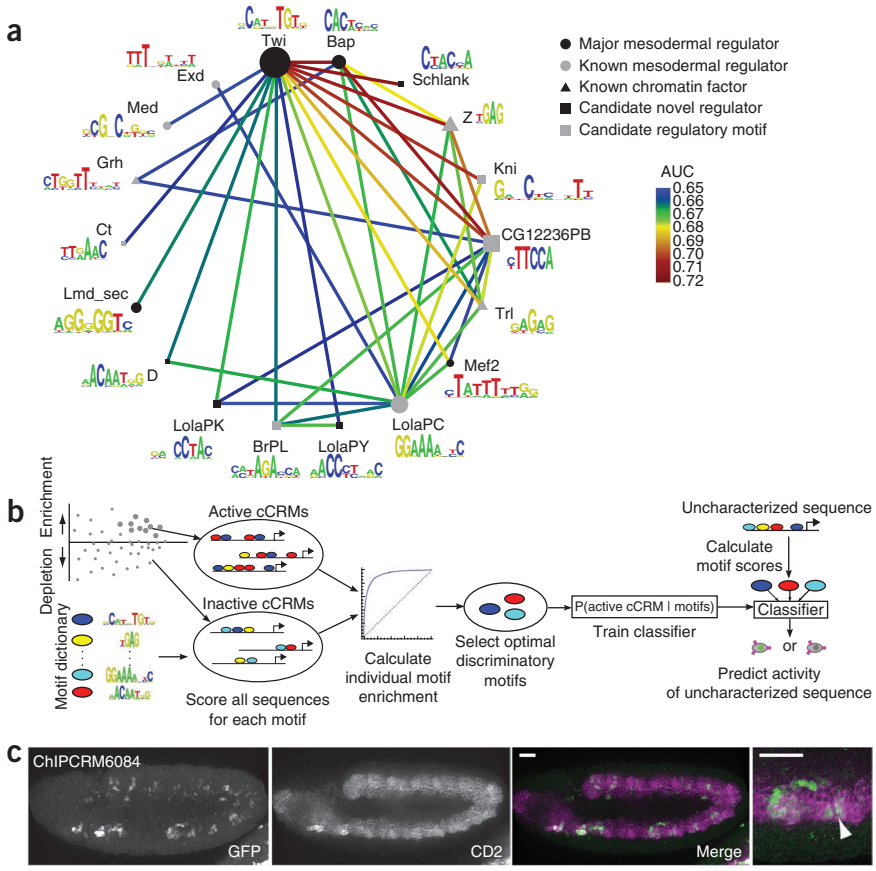


Figure 5 | Computational motif analysis of eFS-identified active enhancers. (**a**) Transcription factor binding site motifs or motif combinations significantly enriched (AUC ≥ 0.65, FDR ≤ 0.1) among eFS-identified active enhancers in *twi*:CD2$^+$ cells. Nodes in the shapes represent motifs for known or candidate transcription factors expressed in the mesoderm, sequence-specific DNA-binding proteins that target chromatin-modifying PcG and trxG complexes ('known chromatin factor'), or putative regulatory motifs for which the representative factors shown are not expressed in the embryonic mesoderm at the appropriate time but that may be recognized by other, mesodermally expressed *trans*-acting factors ('candidate regulatory motif'). Edges represent significant (AUC ≥ 0.65, FDR ≤ 0.1) pairwise 'and' combinations. Node diameter is proportional to $(\text{AUC} - 0.5)^2$ considering the Lever AUC for the individual motif. (**b**) Scheme of classifier analysis. (**c**) Maximum intensity projection of GFP expression driven by cCRM ChIPCRM6084 (chromosomal coordinates provided in **Supplementary Table 1**), correctly predicted to drive coexpression with *twi* CD2. Coexpression was observed and was assessed as described for **Figure 3**. Scale bars, 50 µm.

matches to the discriminatory motifs, independently for whole mesoderm, FCMs and founder cells.

We assessed the accuracy of our models by tenfold cross-validation (in each of the ten iterations, the 90% of the cCRMs that we used to learn discriminatory motifs were also used to train the classifier, and the remaining 10% were used to test the accuracy of the classifier). The whole-mesoderm model achieved an AUC of 0.74 ($P = 3.9 \times 10^{-4}$, Wilcoxon-Mann-Whitney test) using 12 discriminatory motifs, and the FCM-specific model achieved an AUC of 0.93 ($P = 1.2 \times 10^{-6}$, Wilcoxon-Mann-Whitney test) using 3 motifs. These models outperformed ones based solely on previously known *cis*-regulatory motifs for mesoderm and FCMs (AUC of 0.59 and 0.72, respectively; **Supplementary Note 2**). We found no statistically significant ($P < 0.05$) classifier for founder cells.

To demonstrate the practical utility of our models, we tested whether they could predict the activity of cCRMs whose activity had not been measured by eFS. We tested 39 classifier predictions by traditional reporter assays. Six of 10 cCRMs predicted to be active enhancers in mesoderm drove coexpression of GFP with CD2 (**Fig. 5c** and **Supplementary Fig. 9**), 19 of 29 cCRMs predicted to be inactive drove no expression in CD2$^+$ cells and 9 of the 10 remaining predicted negative cCRMs drove limited coexpression at stages 11–12 (**Supplementary Fig. 9**). Many of the *twi*:CD2$^+$ eFS-positive (DESeq $P_{adj} < 0.1$) enhancers in the training set exhibited 'widespread coexpression' with CD2 and fewer exhibited 'limited coexpression', and accordingly our classifier performed better in predicting the activity of cCRMs with 'widespread coexpression'.

DISCUSSION

Our results demonstrate the utility of eFS for highly parallel testing of cCRMs for tissue-specific enhancer activity. No single data type (sequence-specific transcription factor binding, histone modifications or DHS) was most enriched across all three tissues (**Supplementary Fig. 6**). Moreover, none of the different classes of genomic features that we used to prioritize cCRMs for testing by eFS (cCRMs identified by ChIP, CBP-bound regions and DHS) were significantly enriched ($P < 0.1$) among active cCRMs considering each of the three mesodermal cell populations or their nonredundant union (**Supplementary Table 8**). It is perhaps not surprising that these regions were not enriched in either the *Mef2-I-E*$_{D5}$:CD2$^+$GFP$^+$ or *duf*:CD2$^+$GFP$^+$ data, as FCMs and founder cells are relatively rare cell types and also because many of the putative regulatory regions might drive expression in other cell types as the adjacent genes are often expressed in additional cell types or at other time points.

Future studies will be needed to determine the regulatory functions of the putative mesodermal transcription factors suggested by the motif analysis results for eFS-identified enhancers in whole mesoderm and FCMs. The enrichment of binding sites for PcG and trxG recruitment factors, and combinations thereof with ubiquitously expressed and mesoderm-specific transcription factors, suggests that regulatory competence of enhancers requires binding sites of chromatin factors together with those of tissue-specific transcription factors.

Our classifier analysis results indicate that *cis* regulation in FCMs is specified by a smaller set of transcription factors than those used in regulation of a broader class of mesodermal genes expressed in a wider range of cell types, each of which might use different *cis*-regulatory codes[6,44] (**Supplementary Fig. 8b,c**). Likewise, the lack of a significant ($P < 0.05$) classifier for founder cells is likely due to

heterogeneity of founder cells and their associated enhancers[37,44]; eFS using CD2 driver lines specific to subsets or even unique founder cells should aid in the analysis of founder cell–specific *cis*-regulatory codes. Our results on enrichment of various histone modifications (**Supplementary Note 3**) are consistent with the model that there exist different classes of active enhancers that are enriched for different sets of histone modifications[26].

Here we applied the eFS technology to discover muscle enhancers. However, eFS can be used to test cCRMs in any cell type that has at least one known enhancer, by constructing CD2 driver lines using enhancers active in those cell types. eFS can be used to screen cCRMs without any prior experimental evidence (such as ChIP data). Moreover, eFS can be adapted for use in other organisms; the phiC31 integrase system has been used successfully in other species, including zebrafish[45], human and mouse cells[46], and mice[47]. In addition, eFS could be implemented using a different site-specific recombinase or other transformation method. Broader application of eFS should greatly expedite and expand the repertoire of well-defined enhancers and facilitate the development of a more comprehensive picture of their landscape and organization of CRMs across genomes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. Gene Expression Omnibus: GSE41503.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This project was supported in part by a US National Science Foundation Graduate Research Fellowship to L.A.B. and by grant R01 HG005287 from the US National Institutes of Health to M.L.B. We thank G. Losyev and C. Durkin for technical assistance, K.G. Guruharsha and K. VijayRaghavan for sharing coordinates of the *duf* enhancer before its publication, R.P. McCord, M. Markstein and O. Iartchouk for helpful discussion, and R. Gordân, M. Markstein and T. Siggers for critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

M.L.B. designed the study; S.S.G., P.W.E., A.M.M. and M.L.B. developed the eFS technology; S.S.G. and A.V. sorted flies; S.S.G., L.A.B., M.P. and A.A. performed computational data analysis; S.S.G., P.W.E., A.V., Y.K. and X.Z. performed PCRs; B.W.B., X.Z., A.S. and C.E.G. generated CD2 fly lines; S.S.G., A.V., A.P. and A.I. performed validation assays; S.S.G., L.A.B., M.P., A.A., B.W.B. and M.L.B. wrote **Supplementary Note 2**; S.S.G., L.A.B., M.P., A.A. and M.L.B. prepared figures and tables; and S.S.G. and M.L.B. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201 (2003).
2. Davidson, E. Inside the *cis*-regulatory module: control logic, and how regulatory environment is transduced into spatial patterns of gene expression. in *Genomic Regulatory Systems: Development and Evolution*, chapter 2, 25–62 (Academic Press, 2001).
3. Pfeiffer, B.D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **105**, 9715–9720 (2008).
4. 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**, D594–D598 (2008).
5. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21**, 436–449 (2007).
6. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65–70 (2009).
7. Jagalur, M., Pal, C., Learned-Miller, E., Zoeller, R.T. & Kulp, D. Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics* **8** (suppl. 10), S5 (2007).
8. Gertz, J., Siggia, E.D. & Cohen, B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–218 (2009).
9. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
10. Nam, J., Dong, P., Tarpine, R., Istrail, S. & Davidson, E.H. Functional cis-regulatory genomics for systems biology. *Proc. Natl. Acad. Sci. USA* **107**, 3930–3935 (2010).
11. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
12. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
13. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
14. Dunin-Borkowski, O.M. & Brown, N.H. Mammalian CD2 is an effective heterologous marker of the cell surface in *Drosophila*. *Dev. Biol.* **168**, 689–693 (1995).
15. Bate, M. The mesoderm and its derivatives. in *The development of Drosophila melanogaster* (eds., Bate, M. & Martinez-Arias, A.) 1013–1090 (Cold Spring Harbor Laboratory, 1993).
16. Roy, S. *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
17. Contrino, S. *et al.* modMine: flexible access to modENCODE data. *Nucleic Acids Res.* **40**, D1082–D1088 (2012).
18. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biol.* **12**, R43 (2011).
19. Warner, J. *et al.* Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods* **5**, 347–353 (2008).
20. Groth, A.C., Fish, M., Nusse, R. & Calos, M.P. Construction of transgenic *Drosophila* by using the site-specific integrase from phage phiC31. *Genetics* **166**, 1775–1782 (2004).
21. Duan, H., Skeath, J.B. & Nguyen, H.T. *Drosophila* Lame duck, a novel member of the Gli superfamily, acts as a key regulator of myogenesis by controlling fusion-competent myoblast development. *Development* **128**, 4489–4500 (2001).
22. Guruharsha, K.G., Ruiz-Gomez, M., Ranganath, H.A., Siddharthan, R. & Vijayraghavan, K. The complex spatio-temporal regulation of the *Drosophila* myoblast attractant gene duf/kirre. *PLoS ONE* **4**, e6960 (2009).
23. Hoffmann, S. *et al.* Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **5**, e1000502 (2009).
24. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
25. Gallo, S.M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res.* **39**, D118–D123 (2011).
26. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).
27. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
28. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
29. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
30. Kharchenko, P.V. *et al.* Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485 (2011).
31. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
32. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).

33. Azpiazu, N. & Frasch, M. tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of *Drosophila*. *Genes Dev.* **7**, 1325–1340 (1993).
34. 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).
35. Busser, B.W. *et al.* Integrative analysis of the zinc finger transcription factor Lame duck in the *Drosophila* myogenic gene regulatory network. *Proc. Natl. Acad. Sci. USA* **109**, 20768–20773 (2012).
36. Zhu, L.J. *et al.* FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39**, D111–D117 (2011).
37. Philippakis, A.A. *et al.* Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Comput. Biol.* **2**, e53 (2006).
38. Busser, B.W. *et al.* Molecular mechanism underlying the regulatory specificity of a *Drosophila* homeodomain protein that specifies myoblast identity. *Development* **139**, 1164–1174 (2012).
39. Thisse, B., el Messal, M. & Perrin-Schmitt, F. The twist gene: isolation of a *Drosophila* zygotic gene necessary for the establishment of dorsoventral pattern. *Nucleic Acids Res.* **15**, 3439–3453 (1987).
40. Furlong, E.E., Andersen, E.C., Null, B., White, K.P. & Scott, M.P. Patterns of gene expression during *Drosophila mesoderm* development. *Science* **293**, 1629–1633 (2001).
41. Grimaud, C., Negre, N. & Cavalli, G. From genetics to epigenetics: the tale of Polycomb group and trithorax group genes. *Chromosome Res.* **14**, 363–375 (2006).
42. Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.* **40**, e114 (2012).
43. Yuan, Y., Guo, L., Shen, L. & Liu, J.S. Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* **3**, e243 (2007).
44. Busser, B.W. *et al.* A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet.* **8**, e1002531 (2012).
45. Lister, J.A. Transgene excision in zebrafish using the phiC31 integrase. *Genesis* **48**, 137–143 (2010).
46. Thyagarajan, B., Olivares, E.C., Hollis, R.P., Ginsburg, D.S. & Calos, M.P. Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol. Cell Biol.* **21**, 3926–3934 (2001).
47. Hollis, R.P. *et al.* Phage integrases for the construction and manipulation of transgenic mammals. *Reprod. Biol. Endocrinol.* **1**, 79 (2003).

ONLINE METHODS

PCR amplification of cCRMs. The composition of our cCRM library is detailed in **Supplementary Note 2**. All cCRMs were chosen to be 900–1,100 bp long to avoid potential PCR bias. A two-step PCR amplification was used to include Gateway *attB* sites, and specific forward and reverse sequencing primers with Phusion enzyme (New England BioLabs) using *D. melanogaster* OreR genomic DNA as template, followed by amplification with common PCR primers (SEQ1 and SEQ2), as described in **Supplementary Note 2**.

Design of reporter vector pEFS-Dest. We created the vector for eFS, pEFS-Dest (**Supplementary Note 1**), by blunt-end cloning the 1.8 kb HindIII-SpeI fragment of pPelican[48] (containing a nuclear-localized *GFP* reporter construct with a *gypsy* insulator element upstream of the multiple cloning site (MCS) and minimal promoter) into pWattB, then replacing the MCS with a cassette providing *attR1* and *attR2* sites for Gateway cloning. pWattB was made by inserting (i) the phiC31 *attB* site from *Streptomyces lividans*[20] and (ii) the mini-white gene into the small cloning vector pSP73 (Promega). The reporter cassette comprises the *Hsp70* minimal promoter driving expression of a nuclear localization signal–tagged *EGFP* gene with an SV40 polyadenylation sequence[48].

Purification, normalization and cloning of cCRM library into eFS reporter vector. Aliquots of all PCRs were run on agarose gels with High DNA Mass Ladder (Invitrogen) and quantified using Quantity One software (Bio-Rad). Equal masses of each 900–1,100 bp band were pooled, precipitated, gel-purified and then cloned as a pool using Gateway BP Clonase II (Invitrogen) into pDONR221 (Invitrogen). Cloning reactions were transformed into *E. coli* Top10 cells (Invitrogen) and plated on LB agar with kanamycin. A plasmid pool was purified from the resulting colonies, from which the combined inserts were cloned using Gateway LR Clonase II (Invitrogen) into pEFS-Dest. Transformed cells were plated on LB agar with ampicillin, yielding colonies from which the final library plasmid pool was prepared for embryo injection.

Generation of CD2 vector pETWCD2. A minimal promoter was fused to rat *CD2* and subsequently cloned into P-element transformation vectors by PCR-amplifying the TATA box from pUAST-NTAP and *CD2* from *twi-CD2*[14]. These PCR products served as templates for an assembly PCR, the product of which was subcloned into pCR (Invitrogen), sequence-verified, digested with NheI and cloned into XbaI-digested pETWN[49], resulting in our CD2 vector pETWCD2. Primer sequences are provided in **Supplementary Note 2**.

Fly embryo injections and husbandry. The pooled plasmid cCRM library was injected posteriorly into syncytial embryos carrying the *nos-ϕC31\int.NLS* transgene[50] on the X chromosome and the *attP40* insertion[51] on the second chromosome. Surviving males were crossed to excess *y w* virgin females. Transformant male progeny were selected by eye color. We collected several thousand transformant males and, separately, several thousand virgin females from each tissue-specific CD2 line of interest. These flies were combined in population cages ~36 h before the beginning of embryo collections. Population cages were collected from twice 'prelays' to minimize the presence of older embryos due to retention of fertilized eggs by females, then two collections of 2 h (for *twi:CD2* sorting) or 2.5 h (for *Mef2-I-E_{D5}:CD2* and *duf:CD2* sorting) were performed. These plates were aged 10–11 h at 18 °C, after which embryos were collected and dechorionated, and single-cell suspensions were prepared for FACS.

Fluorescence-activated cell sorting. We modified a previous protocol for isolation of single cells for FACS from live *Drosophila* embryos at stage 11 (ref. 52) by incorporating a step in which dissociated cells are resuspended in *Drosophila* cell culture medium and incubated on ice with Alexa Fluor 647–conjugated anti–rat CD2 (AbD-Serotec, MCA154A647; 1:200), as described in **Supplementary Note 2**. After collection of cells by centrifugation, samples were filtered with Nytex mesh and supplemented with DAPI. Cells were washed, and then analyzed and separated by FACS (**Supplementary Note 2**).

cCRM insert amplifications from collected cells. Crude cell extracts were pooled according to sample where necessary to achieve sufficient numbers for accurate quantification of insert abundance (**Supplementary Fig. 2** and **Supplementary Note 2**), then split fivefold before nested PCR amplification to recover library inserts from genomic DNA (**Supplementary Fig. 2**). PCRs were performed using KAPA Hi-Fi HotStart ReadyMix (Kapa Biosystems), as described in **Supplementary Note 2**. PCR products were agarose gel–purified, quantified by NanoDrop and used for Illumina library preparation.

Illumina sequencing. Illumina sequencing libraries were prepared using minor modifications of standard protocols[53] and the Multiplexing Sample Preparation Oligonucleotide Kit (Illumina). Pooled PCR product was sonicated by Covaris S2 as described[53], and then end-repaired with the End-IT DNA End-Repair Kit (EpiCentre Biotechnologies) and A-tailed with Klenow exo$^-$ (New England BioLabs). Standard adaptors (Index PE Adaptor Oligo Mix) were ligated using Quick T4 DNA Ligase (New England BioLabs). Ligation products were size-selected from agarose gels, and quantified and checked for concentration and size distribution by Agilent 2200 TapeStation. Enrichment PCRs were performed using Phusion thermostable polymerase (New England BioLabs), as described in **Supplementary Note 2**. Purified enrichment PCR products were assessed by Agilent 2200 TapeStation and submitted to the Partners Center for Personalized Genetic Medicine for concentration measurement by PicoGreen fluorescence and quantitative (q)PCR, followed by equimolar index pooling and sequencing (50-base single-end read) on the Illumina HiSeq 2000.

Mapping Illumina sequencing reads. We used segemehl version 0.0.9.4 (version as of 15 August 2012 was 0.1.3)[23] with parameter settings -M 100 -E 5 -D 2 -A 80 to map Illumina sequencing reads to the *D. melanogaster* genome. For cCRM detection, we required: (i) ≥1 read from each of the 5′ and 3′ ends; (ii) ≥5 positions covered by center reads (that is, without the SEQ1 or SEQ2 primers); and (iii) ≥10 total reads. Where overlapping cCRM windows contributed indistinguishable reads to the same genomic regions, we used the unambiguous end reads as weights for dividing the reads that map to overlapping cCRM windows. Analysis of random

doi:10.1038/nmeth.2558

sets of genomic windows matched for length, sequence context (for example, intronic and intergenic) and G+C content to our foreground windows indicated that the FDR for cCRM detection was less than 5×10^{-5}.

Statistical analysis of eFS data. We collected the number of reads mapped to each cCRM for each replicate population and control 'input' population, and filtered out cCRMs not detected in any input sample replicate. Enrichment and statistical significance were calculated using DESeq[24] with standard parameters and size factor estimation, as described in **Supplementary Note 2**.

Statistical analysis of genomic features. For a given type of genomic feature, we calculated the scores for each cCRM as the weighted average of the score (for example, ChIP signal intensity) for feature intervals that overlap the peak as reported in the published Browser Extensive Data (BED) or Wiggle Track Format (WIG) file associated with that experiment. We defined the weights by the amount of overlap (in base pairs) between the cCRM and the feature's genomic coordinates. All comparisons of enrichment (or depletion) of various genomic marks were performed by calculating enrichment in the eFS-positive enhancers (DESeq $P_{adj} < 0.1$) as compared to an equally sized set of inactive cCRMs (DESeq $P_{adj} > 0.8$) chosen from the bottom of the ranked list (ranked by decreasing fold-enrichment value). The statistical significance of any such enrichment (or depletion) was determined as $P < 0.05$ by Wilcoxon-Mann-Whitney U test.

DNA sequence motif over-representation analysis. We compiled a dictionary of 567 publicly available *Drosophila* transcription factor binding site motifs[6,35–38]. Motifs were trimmed, redundant motifs were removed and motif exemplars were chosen, as described in **Supplementary Note 2**. To identify over-represented motifs in the *twi*:CD2$^+$GFP$^+$, *Mef2-I-E*$_{D5}$:CD2$^+$GFP$^+$ and *duf*:CD2$^+$GFP$^+$ foreground (FG) sequence sets, we used PhylCRM and Lever[19]. Lever calculates the over-representation of individual motifs or combinations thereof, according to their density and evolutionary conservation, as quantified by the PhylCRM scoring scheme[19], in each FG sequence set as compared to a random set of background (BG) sequences. BG sets were chosen to be about 20 times the size of the FG sets, and matched for length, G+C content and repeat content. All settings were as previously described[54], except that repeats were not masked and length correction was not used because all sequences were roughly the same length. Any motif that did not have occurrences in at least one-quarter of the FG sequences was removed from further consideration. We then used Lever to inspect the FG sets for over-representation of all single and pairwise combinations of the resulting 86-exemplar motif dictionary. Motif PWMs are provided in **Supplementary Table 7**.

Classifier analysis. For each cCRM, we generated a feature vector of scores that quantify the presence of motif matches for each PWM in the motif exemplar dictionary. The score for a particular PWM and a particular cCRM was defined as the sum of the log-odds ratios of PWM matches in the cCRM sequence that exceeded a permissive match threshold (log-odds ratio > 3.0). For classification we used the Gaussian naive Bayes implementation in the *scikit-learn* package[55] for Python (**Supplementary Note 2**). As for the motif over-representation analysis, positive cCRMs are those with DESeq $P_{adj} < 0.1$; here, negative cCRMs are those from an equally sized set chosen from the bottom of the cCRM list ranked by eFS P_{adj} value. To evaluate classification accuracy, we split the labeled cCRM feature vectors into training and test sets using stratified tenfold cross-validation. Feature selection was performed independently for each of the folds: in each, the k motifs with the highest individual AUC values in the training set were selected. The classifier was then trained using features corresponding only to those k motifs. We evaluated performance across multiple values of k and selected the value that maximized performance accuracy in cross-validation tests.

Traditional reporter assays. Homozygous or balanced heterozygous transformant males were crossed to homozygous *twi:CD2* females in small population cages, and broad collections (~2–17 h after egg deposition) of embryos were fixed and stained for immunofluorescence by standard protocols[49] (**Supplementary Note 2**). Stained embryos were imaged with a Zeiss Imager Z1 with Apotome in optical sectioning mode. Coexpression of GFP with CD2 (**Supplementary Table 5**) was evaluated in individual optical sections with the annotator being blind to the predicted activity of the cCRMs. Coexpression was observed as GFP and CD2 being present in the same cells because GFP in these embryos is nuclear and CD2 is expressed on the cell surface. For validations of CD2$^-$ eFS-positive cCRMs as being active enhancers, we assayed for activity anywhere in the embryo at this developmental stage.

48. Barolo, S., Carver, L.A. & Posakony, J.W. GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in *Drosophila*. *Biotechniques* **29**, 726–732 (2000).
49. 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).
50. Bischof, J., Maeda, R.K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. *Proc. Natl. Acad. Sci. USA* **104**, 3312–3317 (2007).
51. Markstein, M., Pitsouli, C., Villalta, C., Celniker, S.E. & Perrimon, N. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.* **40**, 476–483 (2008).
52. Estrada, B. *et al.* An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. *PLoS Genet.* **2**, e16 (2006).
53. Quail, M.A., Swerdlow, H. & Turner, D.J. Improved protocols for the illumina genome analyzer sequencing system. in *Current Protocols in Human Genetics* 18.2 (2009).
54. Aboukhalil, A. & Bulyk, M.L. LOESS correction for length variation in gene set-based genomic sequence analysis. *Bioinformatics* **28**, 1446–1454 (2012).
55. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).